Explanations that reveal all through the definition of encoding

Aahlad Puli[*], **Nhi Nguyen**[*], **Rajesh Ranganath**
New York University

Abstract

Feature attributions attempt to highlight what inputs drive predictive power. Good attributions or explanations are thus those that produce inputs that retain this predictive power; accordingly, evaluations of explanations score their quality of prediction. However, evaluations produce scores better than what appears possible from the values in the explanation for a class of explanations, called encoding explanations. Probing for encoding remains a challenge because there is no general characterization of what gives the extra predictive power. We develop a definition of encoding that identifies this extra predictive power via conditional dependence and show that the definition fits existing examples of encoding. This definition implies, in contrast to encoding explanations, that non-encoding explanations contain all the informative inputs used to produce the explanation, giving them a "what you see is what you get" property, which makes them transparent and simple to use. Next, we prove that existing scores (ROAR, FRESH, EVAL-X) do not rank non-encoding explanations above encoding ones, and develop STRIPE-X which ranks them correctly. After empirically demonstrating the theoretical insights, we use STRIPE-X to show that despite prompting an LLM to produce non-encoding explanations for a sentiment analysis task, the LLM-generated explanations encode.

1 Introduction

Artificial intelligence can unlock information in data that was previously unknown. In medicine, for example, using AI, researchers have shown that electrocardiograms are predictive of structural heart conditions [1] or new-onset diabetes [2]. Good predictions often lead one to ask what in the input is important for a prediction; this question is a driving factor behind research in interpretability and explainability [3, 4]. One primary direction in interpretability seeks to produce explanations that are subsets of the input that retain the predictability of the label. These types of explanations and interpretations are called feature attributions and have been used to find factors associated with debt defaults [5], to demonstrate that detecting COVID-19 from chest radiographs can rely on non-physiological signals [6], and to discover a new class of antibiotics [7].

Several methods exist for producing feature attributions or explanations. While some methods compute functions of model gradients [8] or look at predictability after removing features [9], other methods attribute scores to different inputs by treating them as players in a game [4, 10] or amortize their explanations by learning a single model to select subsets for each instance [11]. Choosing one from the many feature attribution methods requires an evaluation. There are, however, many approaches to evaluation itself: qualitative ones [12, 13, 14], which are limited to cases where humans have precise knowledge about the inputs relevant to prediction, and quantitative ones [2, 4, 15, 16, 17, 18, 19, 20], which do not require human knowledge.

Intuitively, a good evaluation method for feature attributions should assign higher scores to explanations that select inputs that are more predictive of the label. However, evaluations that score explanations based on the predictability of the label from the explanation face one major challenge: *encoding*. Informally, an encoding explanation is one where the explanation predicts the label beyond what seems plausible from the values of the inputs themselves. The top left panel of Figure 1 shows an explanation that predicts the label of dog or cat depending on whether the

[*]Equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

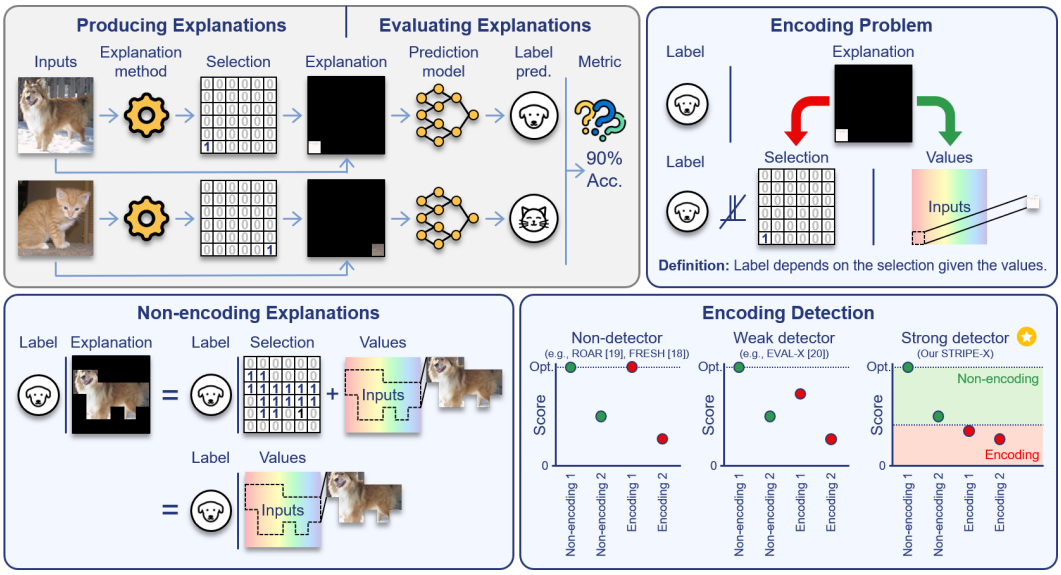


Figure 1: Overview of the paper. Explanations are produced to find inputs that are relevant to predicting a label. However, explanations can predict the label well due to the selection being predictive of the label beyond the explanation's values. Such explanations are called encoding. In contrast, predicting instead from a non-encoding explanation is equivalent to predicting from the values in the explanation. When explanations are evaluated purely based on the quality of prediction, encoding can go undetected. We classify existing evaluations into non-detectors and weak detectors and develop a strong detector, called STRIPE-X.

explanation is a pixel on the right half or left half of the image respectively. Many explanation methods fit the description of encoding [20, 21]. Further, given that many evaluations only look at the quality of prediction, encoding can go undetected, rendering the evaluations ineffective at picking explanations. In contrast, non-encoding explanations predict the label well only when the values in the explanation do, making them easy to reason about.

In addressing encoding, this work makes the following contributions:

- **Develops a simple statistical definition of encoding** via a conditional dependence property.
- Confirms the introduced definition captures all existing ad hoc encoding instances.
- Shows that **non-encoding explanations are easy to use** because they retain all the predictive inputs used to build them, meaning that predictive non-encoding explanations reveal inputs that predict the label to their users, **and thus have a "what you see is what you get" property**.
- Formalizes evaluations' sensitivity to encoding as *weak detection* (optimal scoring explanations are non-encoding) and *strong detection* (non-encoding explanations score above encoding ones).
- Demonstrates that the evaluations ROAR [19] and FRESH [18] do not weakly detect encoding.
- Proves that EVAL-X [20] weakly detects encoding, but does not strongly detect encoding.
- Develops **STRIPE-X** and proves that it **strongly detects encoding**.
- Uses STRIPE-X to show that despite prompting an LLM to produce non-encoding explanations for a sentiment analysis task, the LLM-generated explanations encode.

Figure 1 provides an overview of this paper.

2 Evaluating explanations

We focus on explanation methods where the goal is to produce subsets of the input that predict the label [22, 23]. Explanation methods of this form, also called feature attributions, saliency methods [4, 8, 24], or just "explanations," include thresholded rankings from Shapley values [25, 26], LIME [24], and REAL-X [20]. With $\mathbf{y}$ as the label and $\mathbf{x} \in \mathbf{R}^d$ as the inputs, let $q(\mathbf{y}, \mathbf{x})$ be the joint distribution over them. An explanation method e maps the inputs $\mathbf{x}$ to a binary selection mask $e(\mathbf{x})$ over the inputs: $e : \mathbf{R}^d \to \{0,1\}^d$. The explanation $\mathbf{x}_{e(\mathbf{x})}$ is a pair: the *selection* $e(\mathbf{x})$ and the vector of explanation's *values*. For example, if $\mathbf{x} = [a, b, c]$ is three-dimensional and $e(\mathbf{x}) = [0, 1, 1]$, $\mathbf{x}_{e(\mathbf{x})}$ consists of the binary mask $e(\mathbf{x})$ and the values associated with the inputs that correspond to the indices in $e(\mathbf{x})$ with value 1:

$$\mathbf{x}_{e(\mathbf{x})} = (e(\mathbf{x}), [b, c]).$$

We keep track of the indices because the same value can lead to different predictions depending on the index it appears at; for example, in predicting mortality from patient vital signs, a heart rate above 110 can occur in healthy patients but a temperature of 110°F is almost always fatal. Equivalently, like in existing work [15, 16, 17, 20], one can choose $\mathbf{x}_{e(\mathbf{x})}$ to retain the values in the explanation in the same position and mask out those not selected: $\mathbf{x}_{e(\mathbf{x})} = e(\mathbf{x}) \times \mathbf{x} + (1 - e(\mathbf{x})) \times \texttt{mask-token}$. For concision, we overload the word "explanation" to mean the explanation method instead of the random variable $\mathbf{x}_{e(\mathbf{x})}$ when it is clear from context.

Choosing between explanation methods requires evaluation. Explanation methods seek to return inputs that predict the label, so existing evaluations consider how well the explanation $\mathbf{x}_{e(\mathbf{x})}$ predicts the label $\mathbf{y}$ [15, 16, 17, 20]. To score explanations based on predictive power, an evaluation method $\alpha(\cdot)$ takes as arguments both the explanation $e(\mathbf{x})$ and the joint distribution $q(\mathbf{y}, \mathbf{x})$: $\alpha(q, e)$. Without loss of generality let higher be better.

2.1 Encoding: A disconnect between the predictiveness of explanations and the predictiveness of their values

We give a simple example of encoding to build intuition for the disconnect between predicting the label from the explanation and predicting the label from the explanation's values. Imagine that the goal is to explain which set of vital signs signal bacterial pneumonia as the diagnosis compared to the common cold. Consider the explanation method that selects the patient's height when the true probability of pneumonia is high given the whole set of observables (including labs, symptoms, and vital signs) and otherwise selects the patient's hair color. Physiologically, height and hair color do not indicate that the patient has pneumonia, meaning that this explanation should not be highly predictive of the label. However, by construction, pneumonia is likely exactly when the explanation selects height, and predicting the label from the explanation achieves the same accuracy as predicting with the full conditional $\arg\max_{y \in \{\text{pneumonia, cold}\}} q(\mathbf{y} = y \mid \mathbf{x})$. Thus, despite the explanation method only selecting physiologically irrelevant inputs, the explanation predicts the label well.

Encoding examples such as the one above are neither contrived nor unique. For example, Jethani et al. [20] show that certain procedures that learn to explain, when applied to MNIST digit classification, yield explanations that select a background, black pixel that predicts the label at an accuracy $> 90\%$; (see Figure 1 in [20]). Other examples of encoding explanations that predict better than what is expected from the explanation's values exist [20, 21]. Encoding explanations should not score optimally under a good evaluation because the explanation selects inputs that do not appear to predict the label. However, without a general characterization of the discrepancy in predictive power for encoding, finding explanations whose values predict well remains a challenge. The next section develops a definition of encoding.

3 Formalizing encoding

Intuitively, encoding is a phenomenon where the information about the label in the explanation $\mathbf{x}_{e(\mathbf{x})}$ exceeds what is known from the *explanation's values*. As the input $\mathbf{x}$ determines the explanation $\mathbf{x}_{e(\mathbf{x})}$, the quality of predicting the label $\mathbf{y}$ from the explanation relies on the information about the label transmitted from $\mathbf{x}$ to $\mathbf{x}_{e(\mathbf{x})}$. There are two pathways for this transmission; we elaborate below.

Denoting the values in a subset $\mathbf{v}$ by $\mathbf{x}_\mathbf{v}$, compare the event this subset takes the values $\mathbf{a}$, i.e. $\mathbf{x}_\mathbf{v} = \mathbf{a}$ to the event that the explanation's selection is $\mathbf{v}$ and that the explanation's values are $\mathbf{a}$, i.e., $\mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})$.

1. Knowing that the explanation is $\mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})$ implies not only that the values in the explanation are determined as $\mathbf{x}_\mathbf{v} = \mathbf{a}$, but also that the selection is determined as $e(\mathbf{x}) = \mathbf{v}$.
2. In reverse, knowing that the values of a subset of inputs are $\mathbf{x}_\mathbf{v} = \mathbf{a}$ and knowing the selection $e(\mathbf{x}) = \mathbf{v}$ implies that the explanation are $\mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})$.

Putting these two points together yields an equality between events:

$$\{\mathbf{x} : \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})\} = \{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\} \cap \{\mathbf{x} : \mathbf{x}_\mathbf{v} = \mathbf{a}\}. \tag{1}$$

Thus, the two pathways for information between $\mathbf{x}$ and the explanation $\mathbf{x}_{e(\mathbf{x})}$ are the selection $e(\mathbf{x})$ and explanation's values $\mathbf{x}_\mathbf{v}$; see Figure 2. Existing work makes similar intuitive observations but stops short of formalizing the additional predictive power in an explanation $\mathbf{x}_{e(\mathbf{x})}$ [20, 21].

To formalize this extra predictive power, define the explanation indicator $\mathbf{E_v} = \mathbb{1}[e(\mathbf{x}) = \mathbf{v}]$. A little algebra in Appendix A.1 shows the explanation indicator $\mathbf{E_v}$ provides the extra information:

$$q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})) = q(\mathbf{y} \mid \mathbf{x_v} = \mathbf{a}, \underbrace{\mathbf{E_v} = 1}_{\text{extra information in } \mathbf{x}_{e(\mathbf{x})}}) \neq q(\mathbf{y} \mid \mathbf{x_v} = \mathbf{a}).$$

Building on this insight, we define encoding as a conditional *dependence*:

Definition 1 (Encoding). *The explanation $e(\mathbf{x})$ is encoding if there exists an $\mathbf{S}$ where $q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{S}) > 0$ such that for every $(\mathbf{v}, \mathbf{a}) \in \mathbf{S}$:*

$$\mathbf{y} \not\perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v} = \mathbf{a}. \tag{2}$$

An example mathematical construction of an encoding explanation is provided in Appendix B.1. The dependence in Def: Encoding means that for encoding explanations, there is a disconnect between how well the explanation $\mathbf{x}_{e(\mathbf{x})}$ predicts the label versus only the explanation's values $\mathbf{x_v}$. This disconnect means that evaluations that score explanations based on predictions from the explanation or their transformations [15, 16, 17, 20] can favor explanation methods that select inputs whose values have little relevance to predicting the label.

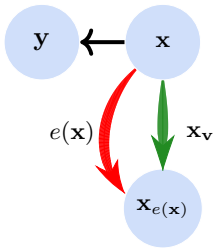


Figure 2: Intuition for encoding: There are two ways the information in the inputs $\mathbf{x}$ about the label $\mathbf{y}$ is transmitted to the explanation $\mathbf{x}_{e(\mathbf{x})}$: (1) through the values in the explanation and (2) the selection $e(\mathbf{x})$ (in red). When the latter happens, the explanation is said to be *encoding*.

Beyond the disconnect in prediction, encoding explanations are undesirable as they conceal predictive inputs that nevertheless affect the explanation. This concealment can lead to incorrect conclusions, such as that inputs outside the selection are irrelevant, or bewilderment because predictive inputs outside the explanation drive changes in the selection in ways that cannot be understood from the explanation itself. An example is provided in Appendix B.2.

Non-encoding explanations. Conversely, for a non-encoding explanation, there exists no positive measure set of explanations $\mathbf{x}_{e(\mathbf{x})}$, where the explanation indicator has conditional dependence given the explanation's values. That is, for a set $\mathbf{A}$ where $q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{A}) = 1$, then for all $(\mathbf{v}, \mathbf{a}) \in \mathbf{A}$

$$\mathbf{y} \perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v} = \mathbf{a}$$

which in turn guarantees

$$q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})) = q(\mathbf{y} \mid \mathbf{x_v} = \mathbf{a}, \mathbf{E_v} = 1) = q(\mathbf{y} \mid \mathbf{x_v} = \mathbf{a}).$$

Appendix A.2 shows this. A simple example of a non-encoding explanation is a constant explanation that always picks the same subset of inputs, since a constant $\mathbf{E_v}$ is independent of any variable. The information for predicting the label in a non-encoding explanation lives in the explanation's values. Evaluations based on predictions from the explanation $\mathbf{x}_{e(\mathbf{x})}$ of non-encoding explanations will yield explanations where the input values $\mathbf{x_v}$ predict the label. *In other words, non-encoding explanations reveal all the informative inputs they depend on, and* ***"what you see is what you get"*** *in the explanation.*

3.1 Encoding explanations in the wild

Def: Encoding encompasses examples in the existing literature beyond the example in Section 2.1. In that example, the information about $\mathbf{y}$ lies in the positions in the selection $e(\mathbf{x})$, which motivates the name position-based encoding (POSI). This section describes two other informal examples from the literature of encoding explanations, prediction-based encoding (PRED) [21] and marginal encoding (MARG) [20], and explains the intuition behind why they encode. In the appendix, we develop formalizations of these types of encoding and show that these formulations meet Def: Encoding.

Prediction-based encoding (PRED). To understand how prediction-based encoding occurs, consider the task of sentiment analysis from movie reviews. Assume that reviews can either be of type "My day was terrible, but the movie was [ADJ1]." and "The movie was [ADJ2], but the day was not great." where adjective ADJ1 can be "good" or "not great" and adjective ADJ2 can be "not great" or "terrible". Due to common English parlance, "terrible" indicates bad sentiment more often than "not great". Then, in the example setup above, only seeing that the fourth word is "terrible" yields bad sentiment with higher probability than when only seeing that the phrase is "not great". However, the fourth word does not always describe the movie. An explanation can look at "not great" describing the movie as bad but then selects "terrible" to encode the bad sentiment. This explanation encodes because the selected word may not describe the movie but the selection predicts the sentiment.

Marginal encoding (MARG). This type of encoding occurs when some inputs determine which other inputs determine the label. For example, in Figure 3, the color determines whether the top right patch produces the label or the bottom right patch. Inputs that *control* where the label comes from are named *control flow inputs*. For a real-world example, consider the following example from Jethani et al. [20], where the goal is to predict mortality for patients with chest pain. A lab value that checks for heart injury and acts like a control flow input is troponin. Abnormal troponin indicates that cardiac issues exist and cardiac imaging would inform mortality. Normal troponin on the other hand can indicate that chest pain is unrelated to cardiac health and a chest X-ray would instead inform mortality. Selecting one image or the other, but not the control flow input, conceals information about why the image was relevant to the label.

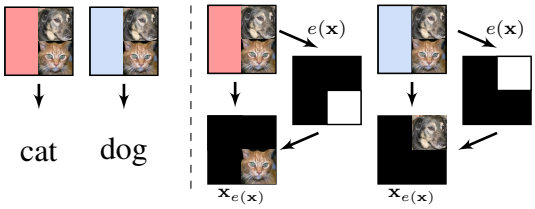


Figure 3: Left: Consider data where the color in the left half determines whether the label "cat", "dog") is produced from the top or bottom image on the right. **Right:** A MARG encoding explanation that produces only the top or the bottom animal image based on the color. The animal image alone says less about the label than knowing the animal image and the color. Knowing the selection determines the color and thus provides additional information about the label.

Formalization. In Appendix B, we provide mathematical formulations of each informal example and show that they fall under the definition of encoding in Def: Encoding: position-based encoding (Appendix B.3), prediction-based encoding (Appendix B.4), and marginal encoding (Appendix B.5). The key intuition behind all of these is that the explanation $e(\mathbf{x})$ varies with inputs other than the selected ones, and these additional inputs provide information about the label beyond the selected ones. Next, we turn to detecting encoding via quantitative evaluations.

4 Detecting encoding in explanations

This section develops notions of sensitivity to encoding for evaluation methods, and uses the mathematical definition of encoding developed in the previous section to establish which methods detect encoding and which do not. Hsia et al. [21] suggest that evaluation methods like EVAL-X can be gamed to produce high scores for encoding explanations by optimizing the evaluation. To study this case, we introduce the notion of *weak detection.* If the optimal score of an evaluation of explanations does not permit encoding, then that evaluation is said to weakly detect encoding:

Definition 2 (Weak detection of encoding). *An evaluation $\alpha(q, e)$ of explanations weakly detects encoding if the optimal explanations e^*, i.e. $\alpha(q, e^*) = \max_e \alpha(q, e)$, are non-encoding.*

Weak detection provides a recipe for finding non-encoding explanations: find the explanation that achieves the maximum score of a weak detector. However, such a recipe would only work when optimizing without constraints because weak detection does not require non-encoding explanations to have a better score than any encoding one. Requiring this leads to the definition of *strong detection*.

Definition 3 (Strong detection of encoding). *An evaluation $\alpha(q, e)$ strongly detects encoding if for any encoding explanation e and non-encoding explanation e', $\alpha(q, e') > \alpha(q, e)$.*

Evaluations that are not weak detectors cannot be strong detectors because they score some encoding explanation optimally.

4.1 Do existing evaluation methods detect encoding?

Here, we consider whether several techniques for evaluating explanations: ROAR [19], FRESH [18], and EVAL-X [20] can detect encoding. We analyze these evaluations on the following distribution q

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3] \sim \mathcal{B}(0.5)^{\otimes 3}, \qquad \mathbf{y} = \begin{cases} \mathbf{x}_1 & \text{w.p. } 0.9 \quad \text{else} \quad 1 - \mathbf{x}_1 \quad \text{if } \mathbf{x}_3 = 1, \\ \mathbf{x}_2 & \text{w.p. } 0.9 \quad \text{else} \quad 1 - \mathbf{x}_2 \quad \text{if } \mathbf{x}_3 = 0. \end{cases} \tag{3}$$

Consider the explanation $e_{\text{encode}}(\mathbf{x}) = \xi_1 = [1, 0, 0]$ if $\mathbf{x}_3 = 1$ and $\xi_2 = [0, 1, 0]$ otherwise; this encodes because $\mathbf{x}_3$ is used to create the explanation and $\mathbf{x}_3$ predicts the label conditional on $\mathbf{x}_1$ when $\mathbf{E}_{\xi_1} = 1$. This is a MARG explanation (see Section 3.1).

ROAR and FRESH do not weakly detect encoding. ROAR evaluates explanations by predicting the label from the inputs not selected by the explanation, denoted as $\mathbf{x}_{-e(\mathbf{x})}$; ROAR scores explanations optimally if the predictions from the remaining covariates are as random as predicting without any covariates at all. In other words, ROAR checks how informative $\mathbf{x}_{-e(\mathbf{x})}$ is of $\mathbf{y}$ and provides the highest score when $\mathbf{y} \perp\!\!\!\perp \mathbf{x}_{-e(\mathbf{x})}$. In contrast, FRESH evaluates explanations by predicting the label from the explanation after removing all other inputs, denoted as $\texttt{val}(\mathbf{x}_{e(\mathbf{x})})$. For example, assume we are given an input $\mathbf{x}$ = "Visually stunning. My favorite movie ever" and an explanation $e(\mathbf{x})$ that selects the words "stunning" and "favorite". Then, the explanation is $\mathbf{x}_{e(\mathbf{x})} = ([0, 1, 0, 1, 0, 0], [\text{"stunning"}, \text{"favorite"}])$, whereas $\texttt{val}(\mathbf{x}_{e(\mathbf{x})}) = [\text{"stunning"}, \text{"favorite"}, \texttt{pad-token}, \texttt{pad-token}, \texttt{pad-token}, \texttt{pad-token}]$, which drops the information about where the selected words are in the input. See Appendix B.6 for a formal definition of $\texttt{val}(\mathbf{x}_{e(\mathbf{x})})$. FRESH checks how predictive $q(\mathbf{y} \mid \texttt{val}(\mathbf{x}_{e(\mathbf{x})}))$ is and assigns an optimal score if the prediction is as good as that of $q(\mathbf{y} \mid \mathbf{x})$. These conditions hold for $e_{\text{encode}}(\mathbf{x})$ in eq. (3):

Proposition 1. *For the data generating process (DGP) in eq. (3), ROAR and FRESH assign their respective optimal scores to the encoding explanation $e_{encode}(\mathbf{x})$.*

The proof is in Appendix B.6. The intuition is that the encoding explanation $e_{\text{encode}}(\mathbf{x})$ always selects the input that informs the label given the control flow $\mathbf{x}_3$; removing the only conditionally informative input means that $\mathbf{x}_{-e_{\text{encode}}(\mathbf{x})}$ has no information about $\mathbf{y}$. In turn, ROAR scores an encoding explanation $\mathbf{x}_{-e_{\text{encode}}(\mathbf{x})}$ optimally, meaning it does not even weakly detect encoding. In addition, $\texttt{val}(\mathbf{x}_{e(\mathbf{x})})$ provides the exact same information about the label regardless of which position it came from. As a result, $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \texttt{val}(\mathbf{x}_{e(\mathbf{x})})$, so FRESH scores $e_{\text{encode}}(\mathbf{x})$ optimally. Even though FRESH attempts to drop the information about the selection $\mathbf{v} = e(\mathbf{x})$ during evaluation, $\texttt{val}(\mathbf{x}_{e(\mathbf{x})})$ remains a function of $\mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})$, so extra information can still be transmitted through the selection $\mathbf{v}$. Thus, ROAR and FRESH are not weak detectors of encoding.

EVAL-X weakly detects encoding but not strongly. EVAL-X [10, 26] is an evaluation method and is sometimes called the surrogate model score. The EVAL-X score with log-probabilities is

$$\text{EVAL-X}(q, e) := \mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})})} \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v},\mathbf{a}))} \left[\log q\left(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}} = \mathbf{a}\right)\right]. \tag{4}$$

This score measures the expected log-likelihood of the labels given the input values chosen by the explanation method e and is grounded in the sampling distribution q. Log-likelihoods are maximized by matching the true distribution, this leads to EVAL-X's weak detection:

Theorem 1. *If $e(\mathbf{x})$ is EVAL-X optimal, then $e(\mathbf{x})$ is not encoding.*

Appendix A.4 gives a proof. The proof shows that at optimality, the prediction from the *values* of explanation has to match the prediction from the full inputs. In turn, given the values there is no additional information in $\mathbf{x}$ about $\mathbf{y}$, which means the explanation indicator $\mathbf{E}_{\mathbf{v}}$ is independent of $\mathbf{y}$; this violates Def: Encoding, which proves the non-encoding nature of EVAL-X-optimal explanations.

To test strong detection for EVAL-X, we consider explanations constrained to select one input. Such reductive constraints appear in practice because the goal of producing an explanation is often to aid humans who benefit from reduced complexity. Such constraints prohibit explanations from reaching EVAL-X's optimal score. Compare $e_{encode}(\mathbf{x})$ with a non-encoding constant explanation:

Proposition 2. *Let $e_{\mathbf{c}}(\mathbf{x}) = \xi_3$. Then, for the DGP in eq. (3), $\text{EVAL-X}(q, e_{encode}) > \text{EVAL-X}(q, e_{\mathbf{c}})$.*

Thus, EVAL-X is not a strong detector. The intuition is that the first two coordinates $\mathbf{x}_1, \mathbf{x}_2$ predict the label when selected by e_{encode}, while the control flow feature does not predict the label. EVAL-X not being a strong detector means that optimizing EVAL-X over a reductive set may yield an encoding explanation. In this case, e_{encode} is one of the EVAL-X-optimal reductive explanations (Lemma 6).

4.2 STRIPE-X: a strong detector of encoding

Encoding explanations induce the dependence between the label $\mathbf{y}$ and the identity of the selection $\mathbf{E}_{\mathbf{v}} = \mathbb{1}[e(\mathbf{x}) = \mathbf{v}]$ given the values in the explanation $\mathbf{x}_{\mathbf{v}}$ (Def: Encoding). This dependence can be tested for by building on conditional independence tests [27, 28, 29]. Rather than testing, direct quantification of dependence can be useful for when combining with other scores, which can be done using instantaneous conditional mutual information:

$$\phi_q(e) := \mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})})} \mathbf{I}\left(\mathbf{E}_{\mathbf{v}}; \mathbf{y} \mid \mathbf{x}_{\mathbf{v}} = \mathbf{a}\right) \quad \text{(ENCODE-METER)}. \tag{5}$$

ENCODE-METER is 0 only when Def: Encoding does not hold:

Proposition 3. ENCODE-METER $\phi_q(e) = 0$ *if and only if* e *is not encoding.*

The proof is in Appendix A.5. Combining EVAL-X with ENCODE-METER weighed by α yields a method we call the strongly information-penalized evaluator (STRIPE-X):

$$\text{STRIPE-X}_\alpha(q, e) := \text{EVAL-X}(q, e) - \alpha\phi_q(e). \tag{6}$$

For a large enough α, the added penalty term pushes down the scores of encoding explanations below that of all non-encoding ones, meaning that STRIPE-X is a strong detector of encoding:

Theorem 2. *With finite* $\mathbf{H}(\mathbf{y} \mid \mathbf{x})$ *and* $\mathbf{H}(\mathbf{y})$*, for any explanation that encodes* e *and any that does not encode* e'*, there exists an* α^* *such that* $\forall \alpha > \alpha^*$ $\text{STRIPE-X}_\alpha(q, e') > \text{STRIPE-X}_\alpha(q, e)$.

Method	Weak	Strong
ROAR [19]	✗	✗
FRESH [18]	✗	✗
EVAL-X [20]	✓	✗
STRIPE-X	✓	✓

Table 1: The weak and strong detection properties of different evaluation methods. Existing scores like ROAR [19] and FRESH [18], are not weak detectors, which in turn means they are not strong detectors either.

The proof is in Appendix A.5. The intuition behind the proof is that for a large enough α, the STRIPE-X scores for any encoding explanations will be dominated by the information term, and thus will become smaller than any non-encoding explanation whose score is lower bounded by the negative marginal entropy, $-\mathbf{H}_q(\mathbf{y})$. Table 1 summarizes the weak and strong detection properties of different evaluations.

Estimating STRIPE-X. The first component of STRIPE-X is EVAL-X. Computing EVAL-X (eq. (4)) requires an estimate of the predictive distribution of the label $\mathbf{y}$ given $\mathbf{x_v}$, $q(\mathbf{y} \mid \mathbf{x_v})$ [20]. Estimation can be done in two ways. The first way makes use of a surrogate model trained to predict the label from different random subsets using masked tokens [9, 20]. The second way to compute EVAL-X (eq. (4)) relies on conditional generative models [30, 31]. Both hyperparameters and a combination of the estimators can be chosen to maximize the average log-likelihood on a held-out validation set across random input subsets.

To estimate the second part of STRIPE-X, the ENCODE-METER, first expand the mutual information terms in ENCODE-METER, $\phi_q(e)$, in terms of expected $\mathbf{KL}$:

$$\phi_q(e) = \mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})})}\mathbb{E}_{\mathbf{y}\sim q(\mathbf{y} \mid \mathbf{x_v}=\mathbf{a})}\mathbf{KL}\left[q(\mathbf{E_v} \mid \mathbf{x_v} = \mathbf{a}, \mathbf{y}) \parallel q(\mathbf{E_v} \mid \mathbf{x_v} = \mathbf{a})\right]. \tag{7}$$

The outer expectation can be estimated using samples from the data and the inner expectation over $\mathbf{y}$ can be estimated using the EVAL-X model $q(\mathbf{y} \mid \mathbf{x_v})$. The distributions over $\mathbf{E_v}$ can be estimated using a classifier of $\mathbf{E_v}$ that randomly masks the label and masks different subsets of the inputs. Further details and a generative way to estimate STRIPE-X are in Appendix C.1 and Appendix C.3; full algorithms are given in Appendix D.

STRIPE-X in practice. Using STRIPE-X to choose between explanations is straightforward: pick the one with the larger score. However, like other evaluations that use learned models, misestimation can pose a problem. With large α, non-encoding explanations with misestimated ENCODE-METER will have bad STRIPE-X scores, while with small α some encoding explanations can have good scores. Across all experiments, we set $\alpha = 20$, which yielded STRIPE-X scores for known encoding explanations worse than known non-encoding explanations.

5 Experiments

This section consists of two parts. The first part demonstrates the weak and strong detection capabilities of the evaluations ROAR, EVAL-X, and STRIPE-X in a simulated setting and on an image recognition task. To demonstrate these capabilities, we run these evaluations on instantiations of POSI, PRED, and MARG. Additionally, we evaluate an existing method that learns to explain under a reductive constraint, called REAL-X [20]. The second part shows how STRIPE-X enables discovering encoding explanations in the wild, without specific knowledge of the DGP or the method that produced the explanation. We employ STRIPE-X to uncover encoding in explanations generated by a large language model (LLM) for predicting sentiments from movie reviews.

5.1 Empirically studying the detection of encoding in a simulated setting

We construct two examples with binary labels $\mathbf{y}$: one discrete input $\mathbf{x}$ and one that is a hybrid of continuous and discrete components. Both use one binary input in $\mathbf{x} \in \{0, 1\}^5$ as a control flow

POSI	PRED	MARG
$e(\mathbf{x}) = \begin{cases} \xi_4 \text{ if } \pi(\mathbf{x}) > 0.5, \\ \xi_5 \text{ else.} \end{cases}$	$e(\mathbf{x}) = \begin{cases} \arg\max_{M:\lvert M\rvert\leq 1} \pi(\mathbf{x}_M) \text{ if } \pi(\mathbf{x}) > 0.5, \\ \arg\max_{M:\lvert M\rvert\leq 1} 1-\pi(\mathbf{x}_M) \text{ else.} \end{cases}$	$e(\mathbf{x}) = \begin{cases} \xi_1 \text{ if } \mathbf{x}_3 = 1, \\ \xi_2 \text{ else.} \end{cases}$

Table 2: Here, $\pi(\mathbf{x}) = q(\mathbf{y} = 1 \mid \mathbf{x})$. Different encoding explanation methods that we consider.

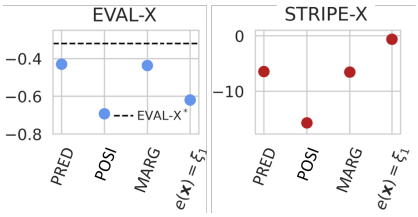
variable and switch the inputs that $\mathbf{y}$ depends on. In both DGPs, $\mathbf{y}$ only depends on $\mathbf{x}_1$ if $\mathbf{x}_3 = 1$, and only on $\mathbf{x}_2$ if $\mathbf{x}_3 = 0$; this means that $\mathbf{x}_4, \mathbf{x}_5$ are purely noise. For both DGPs, $\mathbf{y}$ is sampled per the following distribution where $\mathbf{x}_3$ determines the subset the $\mathbf{y}$ depends on

$$q(\mathbf{y} = 1 \mid \mathbf{x}) = \mathbb{1}[\mathbf{x}_3 = 1]q(\mathbf{y} \mid \mathbf{x}_1, \mathbf{x}_3) + \mathbb{1}[\mathbf{x}_3 = 0]q(\mathbf{y} \mid \mathbf{x}_2, \mathbf{x}_3). \tag{8}$$

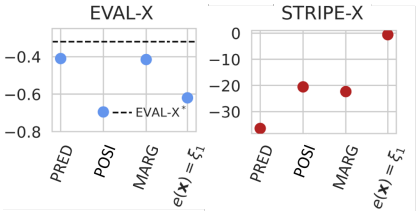
Thus, EVAL-X* is achieved by an explanation of size 2: $e(\mathbf{x}) = \xi_1 + \xi_3$ if $\mathbf{x}_3 = 1$ else $e(\mathbf{x}) = \xi_2 + \xi_3$. See Appendix C.4 for details; the exact DGPs are given in eq. (36) and eq. (37).

Encoding explanations. Table 2 describes the encoding explanations we consider for this setting. In Appendix C.4, we check that Def: Encoding holds for these explanations in the discrete DGP by estimating the role of the unselected inputs in affecting the explanation and the role of $\mathbf{E}_\mathbf{v}$ in predicting $\mathbf{y}$ beyond $\mathbf{x}_\mathbf{v}$; a characterization of Def: Encoding to support this check is in Lemma 1.

ROAR and FRESH fails to weakly detect encoding. To empirically test the analysis about ROAR and FRESH, we study whether the two evaluations weakly detect encoding. In this study, we compare each evaluation's score on the all-inputs explanation, which is optimal, to the score assigned to MARG. MARG ignores $\mathbf{x}_3$ which is required to produce the label $\mathbf{y}$ in eq. (8). ROAR log-likelihoods for MARG and the all-inputs explanation are approximately $-\mathbf{H}(\mathbf{y}) = -0.69$ for both DGPs. In addition, the value of the input that MARG selects alone contains all the information about the label regardless of whether MARG selects $\mathbf{x}_1$ or $\mathbf{x}_2$. Thus, FRESH log-likelihoods for MARG and the all-inputs explanation are both approximately -0.29 for both DGPs. This result validates that ROAR and FRESH are not weak detectors because they do not separate the optimal explanation from all encoding explanations.



(a) Results: discrete DGP.



(b) Results: hybrid DGP.

Figure 4: EVAL-X and STRIPE-X scores of the 3 encoding constructions and the non-encoding constant explanation ($e(\mathbf{x}) = \xi_1$), for both DGPs. EVAL-X, being only a weak detector, assigns suboptimal scores to all encoding explanations ($<$), but scores some encoding explanations above the constant explanation. On the other hand, STRIPE-X, being a strong detector, pushes down the scores of all the encoding explanations below that of the non-encoding constant explanation that always selects $\mathbf{x}_1$.

EVAL-X is a weak detector of encoding but not a strong detector. EVAL-X log-likelihood scores are given in blue in Figures 4a and 4b. EVAL-X, being a weak detector, scores the encoding constructions (POSI, PRED, and MARG) strictly lower than the log-likelihood of the optimal explanation EVAL-X*. However, the EVAL-X score for the MARG explanation is -0.4, which is above the score of -0.6 achieved by a non-encoding explanation $e(\mathbf{x}) = \xi_1$; *thus,* EVAL-X *is not a strong detector.*

Strong detector STRIPE-X prices out all the encoding explanations. Figures 4a and 4b report STRIPE-X scores for the same set of explanations as above; STRIPE-X scores are shown in red. *Strong detector* STRIPE-X *scores the non-encoding explanations above the negative entropy* $-\mathbf{H}_q(\mathbf{y}) = -0.69$ *and scores every encoding construction under that threshold.*

5.2 Detecting encoding on images of dogs and cats

The goal of this section is to study the encoding detection capabilities of ROAR, EVAL-X, and STRIPE-X on real data. We consider an image recognition task like the one in Figure 3 with labels and images from the `cats_vs_dogs` dataset from the Tensorflow package [32]. We break images of size 64×64 into 4 patches each of size 32×32. In left-right then top-down order, let $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ be the upper left, upper right, bottom left, and bottom right patches respectively; $\mathbf{x}_1, \mathbf{x}_3$ capture color, and $\mathbf{x}_2, \mathbf{x}_4$ are the animal images. With `annot(image)` denoting the annota-

tion in the `cats_vs_dogs` dataset the `image` having a `dog` or a `cat`, the label is assigned as:

$$\mathbf{y} = \mathbb{1}[\mathbf{x}_1 = \texttt{blue}] \times \mathbb{1}[\texttt{annot(image } \mathbf{x}_2\texttt{)=dog}] + \mathbb{1}[\mathbf{x}_1 = \texttt{red}] \times \mathbb{1}[\texttt{annot(image } \mathbf{x}_4\texttt{)=dog}]$$

We consider three encoding explanations (POSI, PRED, MARG) and two non-encoding ones: 1) `optimal`, which selects the color and the patch that produces the label as dictated by the color, and 2) denoted `fixed`, which always outputs the bottom right patch $\mathbf{x}_4$. Appendix C.6 gives details.

We report the scores assigned to each explanation by ROAR, EVAL-X, and STRIPE-X in Table 3. ROAR scores two encoding explanations PRED and MARG as high as the `optimal` explanation, meaning it is not even a weak detector. POSI, PRED, and MARG all score worse than the `optimal` explanation under both the weak detector EVAL-X and the strong detector STRIPE-X. However, EVAL-X scores one non-encoding explanation (`fixed`) worse than two encoding ones, meaning it is not a strong detector. Being a strong detector, STRIPE-X scores the `fixed` explanation above the negative marginal entropy $-\mathbf{H}_q(\mathbf{y}) = -0.69$ and scores every encoding construction under that threshold.

	ROAR	FRESH	EVAL-X	STRIPE-X
`opt`	**0.69**	**−0.23**	**−0.27**	**−0.31**
`fixed`	0.59	−0.64	−0.64	−0.64
POSI	0.51	−0.69	−0.70	−5.98
PRED	**0.69**	**−0.23**	−0.51	−1.40
MARG	**0.69**	**−0.23**	−0.53	−1.02

Table 3: ROAR, FRESH, EVAL-X, and STRIPE-X scores for the image recognition experiment. Higher is better. ROAR and FRESH score two encoding explanations PRED and MARG as high as the `optimal` explanation, meaning they are not even weak detectors. EVAL-X being only a weak detector scores POSI, PRED, and MARG all worse than the `optimal` explanation under both EVAL-X but not the non-encoding constant explanation ($e(\mathbf{x}) = \xi_4$), denoted `fixed`. STRIPE-X being a strong detector scores the non-encoding explanations above the negative marginal entropy $-\mathbf{H}_q(\mathbf{y}) = -0.69$ and scores every encoding construction under that threshold.

Evaluating explanations produced by REAL-X [20]. We ran REAL-X to learn explanations for the simulated setting and the image recognition task. In the simulated setting, REAL-X is run to select one input; Appendix C.4 gives details. In the image recognition task, REAL-X is run to select one of the four patches as an explanation; Appendix C.6 gives details. In both the simulated setting (see Figures 4a and 4b) and the image recognition task (see Table 3), REAL-X fails to achieve the optimal EVAL-X score while achieving a STRIPE-X score below the threshold of negative marginal entropy $-\mathbf{H}_q(\mathbf{y}) = -0.69$. Upon investigation, we found that REAL-X produced an explanation that matched the MARG construction on at least 80% of the inputs in the simulated setting. On the image recognition task, REAL-X explanation matched the MARG explanation on the whole dataset. In both cases, STRIPE-X, being a strong detector, correctly alerts that the REAL-X explanation encodes.

5.3 Encoding in LLM-generated explanations

One can detect encoding in any explanation by checking if the STRIPE-X score falls below the negative marginal entropy. Recent work uses LLMs to produce explanations; e.g. [33] prompt an LLM to generate explanations for reasoning tasks which are later used to improve smaller models. If the LLM explanation encodes, the smaller model can falsely ignore the informative inputs the larger model's explanation depends on and yet does not reveal. In this section, we evaluate explanations generated by an LLM, Llama 3, for a sentiment analysis task. We consider reviews that take one of two forms: with `ADJ1` and `ADJ2` as adjectives, the review is

- `'My day was <ADJ1> and the movie was <ADJ2>. that is it'` or
- `'My day was <ADJ1> and the movie was <ADJ2>. oh wait, reverse the adjectives'.`

The second sentence in the review acts as a "control flow" input and determines whether `ADJ1` or `ADJ2` describes the sentiment about the movie. We prompt Llama 3 (see Appendix C.8) to predict the sentiment and select a few words from the review that were important for that sentiment; the selected parts form the generated explanation. To discourage encoding, the prompt explicitly instructs the LLM to select all the words that the LLM based the selection on; such an explanation, by Lemma 1, would be non-encoding. On the 5 most common selections $e(\mathbf{x})$ generated by the LLM, we compute the EVAL-X score and the ENCODE-METER $\phi_q(e)$. The resulting STRIPE-X score is -2.78, falling short of the negative entropy $-\mathbf{H}_q(\mathbf{y}) = -0.69$, meaning the LLM encodes. We investigated why.

As an example, consider the review `'My day was resplendent and the movie was hollow. that is it.'`; the LLM selects only `hollow` in the explanation. However, the LLM instead selects `resplendent` when `that is it` is switched to `oh wait, reverse the adjectives`. Such occurrences are common. On $> 70\%$ of the data, the LLM selects the word that describes the movie

but does not select the second sentence in the review which controls which adjective describes the movie; this is akin to MARG encoding. Thus, the LLM-generated explanation encodes by looking at the control flow input in the second sentence to find the correct adjectives, but failing to select the control flow input. Such an explanation falsely indicates that only the adjectives are relevant to predicting the label. In contrast, a non-encoding explanation would, in addition to the adjective that describes the movie, reveal control flow words that indicate which adjective predicts the label.

In summary, despite being instructed to include all the words that were looked at when producing the explanation, the LLM encodes. Building non-encoding explanations with LLMs may require an extensive search over prompts or finetuning guided by scores from STRIPE-X.

6 Discussion

When an explanation is encodes, predictions from the explanation become disconnected from predictions from the values in the explanations. Such explanations can select values with little relevance to the label and yet score highly on the many existing predictive evaluations. We develop a simple statistical definition of encoding. Inverting this definition shows that when non-encoding explanations predict the label, users know the values of those inputs selected in the explanation predict the label. We then show that existing evaluations are either non-detectors (ROAR[19],FRESH [18]) or only weak detectors (EVAL-X [20]). Motivated by this, we introduce a new strong detector, STRIPE-X. After empirically demonstrating the detection capabilities (or lack thereof) of said evaluations, we use STRIPE-X to discover encoding in LLM-generated explanations.

More related work. Other investigations into evaluating explanations focused on label leakage [26, 34] and faithfulness [18, 35, 36, 37, 38]. Label leakage is similar to encoding in that additional information is in the explanation, but focuses on explanations that have access to both the inputs and the observed label; we leave extending Def: Encoding to leakage to the future. Faithfulness, intuitively, asks that the explanation reflect the process of how a label is predicted from the inputs; a formalization does not exist. Jacovi and Goldberg [38] note the need to define faithfulness formally. Encoding explanations are not faithful to the process of making an explanation because predictive inputs outside those selected by the explanation control the explanation.

Limitations and the future. Using misestimated models in evaluations (like EVAL-X) may lead to mistakes (see Appendix B.8 for an example). The retinal fundus experiment from Jethani et al. [26] is an example where misestimation leads to reductive explanations scoring higher than using the full input. Misestimation can be due to poor uncertainty or due to dependence on shortcut features. One fruitful direction is to use better uncertainty estimates, like conformal inference [39] or calibration [40], or employ robustness methods [41, 42] to ameliorate errors due to misestimation. Another direction is use tricks like REINFORCE-style gradients to construct non-encoding explanations by optimizing STRIPE-X. Explanations that output subsets may not always help humans interpret the mechanism of the prediction. For example, imagine one wants to understand why a model correctly answers the question "Who won the ski halfpipe at the X-games 3 years after her debut in 2021?" with "Eileen Gu". A subset explanation may return "3 years after her debut in 2021" and "ski halfpipe", but that does not help a human interpret how the model predicts. A better interpretation would be to make the model output, "3 years after 2021 is 2024. Eileen Gu won in 2024, and debuted in 2021." Such explanations can also encode information about the prediction in the text produced as a rationale [43]. An important direction here would be to extend the definitions of weak and strong detectors of encoding to evaluations of free-text rationales.

Data versus Model Explanations. Even with the formal definitions of explanation methods, there is a question about what is being explained: the data or the model. These two concepts often get blended together in the literature [11, 20]. We clarify this point and abstract the choice away as two different ways to produce the joint distribution $q(\mathbf{y}, \mathbf{x})$. In *data explanation*, the distribution under which a feature attribution method seeks to output a subset of inputs that predict the label should be the population distribution of the data [23]. If, instead, the goal is *model explanation*, the goal should not be to highlight inputs that predict the label well in samples of the data; rather it should be *to predict the label well in samples from the model.* Formally, a model with parameters $\boldsymbol{\theta}$ is a conditional distribution, $p_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$. To target a model explanation, a feature attribution method would aim to output a subset of inputs that predict the label under the distribution $F(\mathbf{x})p_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$.

Acknowledgements

This work was partly supported by the NIH/NHLBI Award R01HL148248, NSF Award 1922658 NRT-HDR:FUTURE Foundations, Translation, and Responsibility for Data Science, NSF CAREER Award 2145542, NSF Award 2404476, ONR N00014-23-1-2634, Google DeepMind, and Apple. The authors would like to thank Yoav Wald, the NeurIPS 2024 reviewers and the NeurIPS 2024 area chair for helpful feedback.

References

[1] Pierre Elias, Timothy J Poterucha, Vijay Rajaram, Luca Matos Moller, Victor Rodriguez, Shreyas Bhave, Rebecca T Hahn, Geoffrey Tison, Sean A Abreau, Joshua Barrios, et al. Deep learning electrocardiographic analysis for detection of left-sided valvular heart disease. *Journal of the American College of Cardiology*, 80(6):613–626, 2022.

[2] Neil Jethani, Aahlad Puli, Hao Zhang, Leonid Garber, Lior Jankelson, Yindalon Aphinyanaphongs, and Rajesh Ranganath. New-onset diabetes assessment using artificial intelligence-enhanced electrocardiography. *arXiv preprint arXiv:2205.02900*, 2022.

[3] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Visualising image classification models and saliency maps. *Deep Inside Convolutional Networks*, 2, 2014.

[4] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[5] Kim Long Tran, Hoang Anh Le, Thanh Hien Nguyen, and Duc Trung Nguyen. Explainable machine learning for financial distress prediction: evidence from vietnam. *Data*, 7(11):160, 2022.

[6] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

[7] Felix Wong, Erica J Zheng, Jacqueline A Valeri, Nina M Donghia, Melis N Anahtar, Satotaka Omori, Alicia Li, Andres Cubillos-Ruiz, Aarti Krishnan, Wengong Jin, et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, pages 1–9, 2023.

[8] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[9] Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

[10] Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. Fastshap: Real-time shapley value estimation. In *International Conference on Learning Representations*, 2022.

[11] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Invase: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2018.

[12] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 59–67, 2019.

[13] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10): 867–878, 2022.

[14] Jonathan Crabbé, Alicia Curth, Ioana Bica, and Mihaela van der Schaar. Benchmarking heterogeneous treatment effect models through the lens of interpretability. *Advances in Neural Information Processing Systems*, 35:12295–12309, 2022.

[15] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

[16] V Petsiuk, A Das, and K Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[17] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30, 2017.

[18] Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. Learning to faithfully rationalize by construction. *arXiv preprint arXiv:2005.00115*, 2020.

[19] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

[20] Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In *International Conference on Artificial Intelligence and Statistics*, pages 1459–1467. PMLR, 2021.

[21] Jennifer Hsia, Danish Pruthi, Aarti Singh, and Zachary C Lipton. Goodhart's law applies to nlp's explanation benchmarks. *arXiv preprint arXiv:2308.14272*, 2023.

[22] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[23] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International conference on machine learning*, pages 883–892. PMLR, 2018.

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016.

[25] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665, 2014.

[26] Neil Jethani, Adriel Saporta, and Rajesh Ranganath. Don't be fooled: label leakage in explanation methods and the importance of their quantitative evaluation. In *International Conference on Artificial Intelligence and Statistics*, pages 8925–8953. PMLR, 2023.

[27] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011.

[28] Mukund Sudarshan, Wesley Tansey, and Rajesh Ranganath. Deep direct likelihood knockoffs. *Advances in neural information processing systems*, 33:5036–5046, 2020.

[29] Mukund Sudarshan, Aahlad Puli, Wesley Tansey, and Rajesh Ranganath. Diet: Conditional independence testing with marginal dependence measures of residual information. In *International Conference on Artificial Intelligence and Statistics*, pages 10343–10367. PMLR, 2023.

[30] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. *URL https://arxiv.org/abs/2204.06125*, 7, 2022.

[32] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/. Software available from tensorflow.org.

[33] Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*, 2022.

[34] Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? *arXiv preprint arXiv:2010.04119*, 2020.

[35] Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. " will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification. *arXiv preprint arXiv:2111.07367*, 2021.

[36] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9623–9633, 2022.

[37] Yiming Ju, Yuanzhe Zhang, Zhao Yang, Zhongtao Jiang, Kang Liu, and Jun Zhao. Logic traps in evaluating attribution scores. *arXiv preprint arXiv:2109.05463*, 2021.

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

[39] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[40] Mark Goldstein, Xintian Han, Aahlad Puli, Adler Perotte, and Rajesh Ranganath. X-cal: Explicit calibration for survival analysis. *Advances in neural information processing systems*, 33: 18296–18307, 2020.

[41] Aahlad Manas Puli, Lily H Zhang, Eric Karl Oermann, and Rajesh Ranganath. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. *ICLR 2022*, 2021.

[42] Aahlad Manas Puli, Lily Zhang, Yoav Wald, and Rajesh Ranganath. Don't blame dataset shift! shortcut learning due to gradients and cross entropy. *Advances in Neural Information Processing Systems*, 36, 2023.

[43] Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. *arXiv preprint arXiv:2305.18029*, 2023.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification:
   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification:
   Guidelines:
   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
3. **Theory Assumptions and Proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification:
   Guidelines:
   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification:
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [No]
   Justification: We only use public data. We use standard existing training techniques, describe the hyperparameters in detail, and provided the Llama 3 prompts we used in our experiments.
   Guidelines:
   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
   Answer: [Yes]
   Justification:
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment Statistical Significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [No]
   Justification: The error bars in 5.1 and 5.2 are negligible because we estimate means of metrics over datasets of 5000 samples. For 5.3, error bars are irrelevant because there is no comparison against an existing method.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification:
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).
9. **Code Of Ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
   Answer: [Yes]
   Justification: We do not use human subjects and only use public data and public models. Our work focuses on explainable machine learning and does not pose additional ethical harms beyond what is standard for the field.
   Guidelines:
   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).
10. **Broader Impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    Answer: [NA]
    Justification: We investigate a problem with explanations and fix it. There is no societal impact of our work that exceed that of the usage of explanations.
    Guidelines:
    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**
    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
    Answer: [NA]
    Justification: We do not release any data or models.
    Guidelines:
    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.
12. **Licenses for existing assets**
    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
    Answer: [Yes]
    Justification:
    Guidelines:
    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.
13. **New Assets**
    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
    Answer: [NA]
    Justification:
    Guidelines:
    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.
14. **Crowdsourcing and Research with Human Subjects**
    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]
Justification:
Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
Answer: [NA]
Justification:
Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

A Theoretical Details

A.1 Expressing $q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})})$ in terms of the values and the identity of the explanation

To express $q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})})$, we use the following equivalence of events from eq. (1)

$$\{\mathbf{x} : \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})\} = \{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\} \cap \{\mathbf{x} : \mathbf{x}_{\mathbf{v}} = \mathbf{a}\}. \tag{eq. (1)}$$

Then, intuitively, conditioning on the event that $\mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})$ gives you the same information as conditioning on the events $e(\mathbf{x}) = \mathbf{v}$ and $\mathbf{x}_{\mathbf{v}} = \mathbf{a}$ simultaneously. We make this formal below.

For discrete $\mathbf{x}$, for any $\mathbf{v}, \mathbf{a}$ such that the probability $q(\mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})) > 0$, define the LHS and RHS of eq. (1) as $B_{\mathbf{v}}, C_{\mathbf{v}}$ respectively. Then, the conditionals $q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a}))$ and $q(\mathbf{y} \mid \mathbf{E}_{\mathbf{v}} = 1, \mathbf{x}_{\mathbf{v}} = \mathbf{a})$ exist and can be written as follows:

$$q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})) = q(\mathbf{y} \mid \mathbf{x} \in B_{\mathbf{v}}) \qquad q(\mathbf{y} \mid \mathbf{E}_{\mathbf{v}} = 1, \mathbf{x}_{\mathbf{v}} = \mathbf{a}) = q(\mathbf{y} \mid \mathbf{x} \in C_{\mathbf{v}}).$$

These two conditionals are equal because $B_{\mathbf{v}} = C_{\mathbf{v}}$.

The same kind of result holds for general random vectors (discrete or continuous) $\mathbf{x}$ but is a little more involved because $B_{\mathbf{v}}$ may be non-empty while $q(\mathbf{x} \in B_{\mathbf{v}}) = 0$ and the equality of conditional densities/probabilities need to be written via measure theory. Assume the regular conditional probabilities $q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})})$ and $q(\mathbf{y}, \mathbf{E}_{\mathbf{v}} \mid \mathbf{x}_{\mathbf{v}}), q(\mathbf{y} \mid \mathbf{E}_{\mathbf{v}}, \mathbf{x}_{\mathbf{v}})$ and $q(\mathbf{E}_{\mathbf{v}} \mid \mathbf{x}_{\mathbf{v}})$ are defined almost everywhere in their respective probability measures. Take any $\mathbf{S}_{\mathbf{v}} \subseteq \{\mathbf{x}_{\mathbf{v}} : e(\mathbf{x}) = \mathbf{v}\}$ where $q(\mathbf{x}_{\mathbf{v}} \in \mathbf{S}_{\mathbf{v}}) > 0$ and $q(e(\mathbf{x}) = \mathbf{v}) > 0$. Consider any measurable sets $\mathbf{Y}$ over $\mathbf{y}$ and $\mathbf{B}_{\mathbf{v}}(\mathbf{S}_{\mathbf{v}}) := \{(\mathbf{v}, \mathbf{a}) : \mathbf{a} \in \mathbf{S}_{\mathbf{v}}\}$ over $\mathbf{x}_{e(\mathbf{x})}$. Now, by definition of regular conditional probability measures, joint probabilities are obtained by taking the expectation of the conditional with respect to marginal distributions over the conditioning set:

$$\begin{aligned} q(\mathbf{y} \in \mathbf{Y}, \mathbf{x}_{e(\mathbf{x})} \in \mathbf{B}_{\mathbf{v}}(\mathbf{S}_{\mathbf{v}})) &= \int_{\mathbf{B}_{\mathbf{v}}(\mathbf{S}_{\mathbf{v}})} q(\mathbf{y} \in \mathbf{Y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})) q(d\mathbf{x}_{e(\mathbf{x})}) \\ &= \int_{\mathbf{S}_{\mathbf{v}}} q(\mathbf{y} \in \mathbf{Y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})) q(\mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})) d\mathbf{a} \end{aligned} \tag{9}$$

and

$$\begin{aligned} q(\mathbf{y} \in \mathbf{Y}, \mathbf{E}_{\mathbf{v}} = 1, \mathbf{x}_{\mathbf{v}} \in \mathbf{S}_{\mathbf{v}}) &= \int_{\mathbf{S}_{\mathbf{v}}} q(\mathbf{y} \in \mathbf{Y}, \mathbf{E}_{\mathbf{v}} = 1 \mid \mathbf{x}_{\mathbf{v}} = \mathbf{a}) q(\mathbf{x}_{\mathbf{v}} = \mathbf{a}) d\mathbf{a} \\ &= \int_{\mathbf{S}_{\mathbf{v}}} q(\mathbf{y} \in \mathbf{Y} \mid \mathbf{E}_{\mathbf{v}} = 1, \mathbf{x}_{\mathbf{v}} = \mathbf{a}) q(\mathbf{E}_{\mathbf{v}} = 1, \mathbf{x}_{\mathbf{v}} = \mathbf{a}) d\mathbf{a}. \end{aligned} \tag{10}$$

Due to eq. (1), the LHS terms of the two equations above are equal and so are the probability measures over the integrating variables in eqs. (9) and (10). Letting $\mathbf{x}_{\mathbf{v}}$ be defined on a Borel sigma algebra, these two integrals eqs. (9) and (10) are equal if and only if for any Borel set $\mathbf{S}_{\mathbf{v}}$, for almost every $\mathbf{a} \in \mathbf{S}_{\mathbf{v}}$

$$q(\mathbf{y} \in \mathbf{Y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})) = q(\mathbf{y} \in \mathbf{Y} \mid \mathbf{E}_{\mathbf{v}} = 1, \mathbf{x}_{\mathbf{v}} = \mathbf{a}).$$

That is, in more plain terms, the conditional distributions are equal $q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})) = q(\mathbf{y} \mid \mathbf{E}_{\mathbf{v}} = 1, \mathbf{x}_{\mathbf{v}} = \mathbf{a})$.

A.2 With non-encoding explanations "what you see is what you get"

Def: Encoding says that an explanation $e(\mathbf{x})$ is encoding if there exists an $\mathbf{S}$ where $q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{S}) > 0$ such that for every $(\mathbf{v}, \mathbf{a}) \in \mathbf{S}$, :

$$\mathbf{y} \not\perp\!\!\!\perp \mathbf{E}_{\mathbf{v}} \mid \mathbf{x}_{\mathbf{v}} = \mathbf{a}. \tag{11}$$

For a non-encoding explanation, Def: Encoding does not hold. Here, we derive the implications of violating Def: Encoding. Define the set $\mathbf{A}$ to contain all $(\mathbf{v}, \mathbf{a})$ where eq. (11) is violated:

$$\mathbf{A} = \{(\mathbf{v}, \mathbf{a}) : \mathbf{y} \perp\!\!\!\perp \mathbf{E}_{\mathbf{v}} \mid \mathbf{x}_{\mathbf{v}} = \mathbf{a}\}. \tag{12}$$

By definition, the complement $\mathbf{A}^C$ is such that

$$\forall(\mathbf{v}, \mathbf{a}) \in \mathbf{A}^C, \quad \mathbf{y} \not\perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v} = \mathbf{a}.$$

Such a set cannot have positive measure when Def: Encoding is violated which means

$$q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{A}^C) = 0.$$

In turn,

$$q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{A}) = 1 - q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{A}^C) = 1.$$

Thus, $\mathbf{A}$ is such that $q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{A}) = 1$, and by eq. (12) for all $(\mathbf{v}, \mathbf{a}) \in \mathbf{A}$

$$\mathbf{y} \perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v} = \mathbf{a},$$

which in turn guarantees

$$q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})) = q(\mathbf{y} \mid \mathbf{x_v} = \mathbf{a}, \mathbf{E_v} = 1) = q(\mathbf{y} \mid \mathbf{x_v} = \mathbf{a}).$$

A.3 Helpful Lemmas and their proofs

A.3.1 Alternate conditions equivalent to Def: Encoding

The dependence in Def: Encoding occurs due to two reasons, understanding which sheds more light on the definition. First, for some selection $e(\mathbf{x}) = \mathbf{v}$, the explanation's values $\mathbf{x_v}$ do not provide enough information to reveal that the explanation should select the inputs denoted by $\mathbf{v}$. In other words, the indicator of the selection is variable even after fixing the explanation's values themselves. Second, this indicator is predictive of the label for the data with the explanation $\mathbf{v}$. These two properties provide intuition on the definition of encoding:

Lemma 1. *Def: Encoding holds for an explanation $e(\mathbf{x})$ if and only if there exists a selection $\mathbf{v}$ such that $q(e(\mathbf{x}) = \mathbf{v}) > 0$ and a set $\mathbf{S_v} \subseteq \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\}$ such that $q(\mathbf{x_v} \in \mathbf{S_v}) > 0$ where both of the following conditions hold for almost every $\mathbf{a} \in \mathbf{S_v}$:*

Unpredictability of Explanation	$q(\mathbf{E_v} = 1 \mid \mathbf{x_v} = \mathbf{a}) \neq 1;$
Additional Information from Explanation	$q(\mathbf{y} \mid \mathbf{x_v} = \mathbf{a}, \mathbf{E_v} = 1) \neq q(\mathbf{y} \mid \mathbf{x_v} = \mathbf{a}, \mathbf{E_v} = 0).$

Proof. First, Lemma 2 shows the Def: Encoding holds if only if there exists a selection $\mathbf{v}$ such that $q(e(\mathbf{x}) = \mathbf{v}) > 0$ and a set $\mathbf{S_v} \subseteq \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\}$ such that $q(\mathbf{x_v} \in \mathbf{S_v}) > 0$ where

$$\forall \mathbf{a} \in \mathbf{S_v}, \quad \mathbf{y} \not\perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v}.$$

We use this alternate definition in what follows.

Given a non-measure zero set $\mathbf{S_v}$, by Lemma 3, almost everywhere in $\mathbf{S_v}$ it holds that $q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) > 0$.

Conditional dependence implies Unpredictability and Additional information (the only if part). If $q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) = 1$ almost everywhere (under $q(\mathbf{x})$), then $\mathbf{E_v}$ is constant given $\mathbf{x_v}$, and therefore independent of any variable given $\mathbf{x_v}$:

$$q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) = 1 \implies \mathbf{y} \perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v}.$$

Then, it follows that conditional dependence implies the unpredictability property

$$\mathbf{y} \not\perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v} \implies q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) < 1.$$

Second, with the result from Lemma 3, we have $q(\mathbf{E_v} = 1 | \mathbf{x_v}) \in (0, 1)$. Thus $q(\mathbf{y} | \mathbf{x_v}, \mathbf{E_v} = 1)$ and $q(\mathbf{y} | \mathbf{x_v}, \mathbf{E_v} = 0)$ exist almost every where in $\mathbf{S_v}$. Then, by definition of conditional dependence, there is additional information about the label in the explanation:

$$\mathbf{y} \not\perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v} \implies q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 1) \neq q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 0).$$

This shows that Def: Encoding implies the additional information property.

Conditional dependence implied by Unpredictability and Additional information (the if part). Now, if $q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) \in (0, 1)$, then the following two conditional distributions exist almost everywhere in $\mathbf{S_v}$

$$q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 1) \qquad , \qquad q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 0).$$

Then, by definition of dependence almost everywhere $\mathbf{S_v}$:

$$q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 1) \neq q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 0) \implies \mathbf{y} \not\perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v}.$$

Thus, the unpredictability and the additional information properties imply Def: Encoding.

□

Lemma 2. *Def: Encoding holds for an explanation $e(\mathbf{x})$ if and only if there exists a selection $\mathbf{v}$ such that $q(e(\mathbf{x}) = \mathbf{v}) > 0$ and a set $\mathbf{S_v} \subseteq \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\}$ such that $q(\mathbf{x_v} \in \mathbf{S_v}) > 0$ where*

$$\forall \mathbf{a} \in \mathbf{S_v}, \quad \mathbf{y} \not\perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v} = \mathbf{a}. \tag{13}$$

Proof. Def: Encoding says that the explanation $e(\mathbf{x})$ is encoding if there exists an $\mathbf{S}$ where $q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{S}) > 0$ such that for every $(\mathbf{v}, \mathbf{a}) \in \mathbf{S}$ eq. (13) holds. This proof works by showing that $\mathbf{S}$ having a positive measure implies the existence of $\mathbf{v}$ and $\mathbf{S_v}$ as in Lemma 2 such that eq. (13) holds.

Decompose $q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{S})$ by introducing an expectation over $\mathbf{v} \sim q(e(\mathbf{x}))$,

$$q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{S}) = \mathbb{E}_{\mathbf{v} \sim q(e(\mathbf{x}))} q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{S} \mid e(\mathbf{x}) = \mathbf{v}).$$

As there are only finitely many $\mathbf{v}$,

$$q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{S}) > 0 \quad \iff \quad \exists \mathbf{v} \quad \text{s.t.} \quad q(e(\mathbf{x}) = \mathbf{v}) > 0 \quad \text{and} \quad q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{S} \mid e(\mathbf{x}) = \mathbf{v}) > 0.$$

The "only if" direction. Pick any $\mathbf{v}$ such that the RHS above holds and define $\mathbf{S_v} = \{\mathbf{a} : (\mathbf{v}, \mathbf{a}) \in \mathbf{S}\}$. By definition,

$$\mathbf{S_v} = \{\mathbf{x_v} : (\mathbf{v}, \mathbf{x_v}) \in \mathbf{S}\} \cap \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\} \subseteq \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\}.$$

This proves that $\mathbf{S_v}$ has positive measure:

$$q(\mathbf{x_v} \in \mathbf{S_v}) = q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{S}, e(\mathbf{x}) = \mathbf{v}) = q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{S} \mid e(\mathbf{x}) = \mathbf{v}) * q(e(\mathbf{x}) = \mathbf{v}) > 0.$$

Finally, as $\mathbf{a} \in \mathbf{S_v} \implies (\mathbf{v}, \mathbf{a}) \in \mathbf{S}$, eq. (13) holds:

$$\forall \mathbf{a} \in \mathbf{S_v}, \quad \mathbf{y} \not\perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v} = \mathbf{a}.$$

This completes the "only if" direction.

The "if" direction. Assume that there exists $\mathbf{v}$ such that $q(e(\mathbf{x}) = \mathbf{v}) > 0$ and $\mathbf{S_v} \subseteq \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\}$ such that $q(\mathbf{x_v} \in \mathbf{S_v}) > 0$ where

$$\forall \mathbf{a} \in \mathbf{S_v} \qquad \mathbf{y} \not\perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v} = \mathbf{a}.$$

Define $\mathbf{S} = \{(\mathbf{v}, \mathbf{a}) : \mathbf{a} \in \mathbf{S_v}\}$. By this construction, $\mathbf{S}$ has positive measure:

$$\begin{aligned} q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{S}) &= q((\mathbf{v}, \mathbf{x_v}) \in \mathbf{S}) \\ &= q(e(\mathbf{x}) = \mathbf{v}) q((\mathbf{v}, \mathbf{x_v}) \in \mathbf{S} \mid e(\mathbf{x}) = \mathbf{v}) \\ &= q(e(\mathbf{x}) = \mathbf{v}) q(\mathbf{x_v} \in \mathbf{S_v} \mid e(\mathbf{x}) = \mathbf{v}) \\ &= q(e(\mathbf{x}) = \mathbf{v}) q(\mathbf{x_v} \in \mathbf{S_v}) \qquad \{ \text{ as } \mathbf{S_v} \subseteq \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\} \\ &> 0, \end{aligned}$$

where the last inequality holds because by assumption

$$q(e(\mathbf{x}) = \mathbf{v}) > 0 \qquad q(\mathbf{x_v} \in \mathbf{S_v}) > 0.$$

Finally, as $(\mathbf{v}, \mathbf{a}) \in \mathbf{S} \implies \mathbf{a} \in \mathbf{S_v}$, eq. (13) holds:

$$\forall (\mathbf{v}, \mathbf{a}) \in \mathbf{S}, \quad \mathbf{y} \not\perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v} = \mathbf{a}.$$

This completes the "if" directions and with that the proof. □

Lemma 3. *For any set* $\mathbf{S_v} \subseteq \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\}$ *such that* $q(\mathbf{x_v} \in \mathbf{S_v}) > 0$*, then for almost every* $\mathbf{a} \in \mathbf{S_v}$*,* $q(\mathbf{E_v} = 1 \mid \mathbf{x_v} = \mathbf{a}) > 0$*.*

Proof. Define the set $\mathbf{A_v} = \{\mathbf{a} : q(\mathbf{E_v} = 1 \mid \mathbf{x_v} = \mathbf{a}) = 0\}$. Next compute the joint probability

$$q(\mathbf{x_v} \in \mathbf{A_v} \cap \mathbf{S_v}) = q(\mathbf{x_v} \in \mathbf{A_v}) q(\mathbf{x_v} \in \mathbf{S_v} \mid \mathbf{x_v} \in \mathbf{A_v}).$$

Now, noting that $\mathbf{S_v}$ is a subset of $\{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\}$, which is equivalent to $\{\mathbf{x_v} : \mathbf{E_v} = 1\}$, thus

$$\begin{aligned}
&q(\mathbf{x_v} \in \mathbf{S_v} \mid \mathbf{x_v} \in \mathbf{A_v}) \\
&= \int q(\mathbf{x_v} \in \mathbf{S_v} \mid \mathbf{x_v} = \mathbf{a}, \mathbf{x_v} \in \mathbf{A_v}) q(\mathbf{x_v} = \mathbf{a} \mid \mathbf{x_v} \in \mathbf{A_v}) d\mathbf{a} \\
&= \int q(\mathbf{x_v} \in \mathbf{S_v} \mid \mathbf{x_v} = \mathbf{a}) q(\mathbf{x_v} = \mathbf{a} \mid \mathbf{x_v} \in \mathbf{A_v}) d\mathbf{a} \\
&\leq \int q(\mathbf{E_v} = 1 \mid \mathbf{x_v} = \mathbf{a}) q(\mathbf{x_v} = \mathbf{a} \mid \mathbf{x_v} \in \mathbf{A_v}) d\mathbf{a} \\
&= \int q(\mathbf{E_v} = 1 \mid \mathbf{x_v} = \mathbf{a}) q(\mathbf{x_v} = \mathbf{a} \mid \mathbf{x_v} \in \{\mathbf{a} : q(\mathbf{E_v} = 1 \mid \mathbf{x_v} = \mathbf{a}) = 0\}) d\mathbf{a} \\
&= 0.
\end{aligned}$$

The probability $q(\mathbf{x_v} \in \mathbf{S_v} \mid \mathbf{x_v} \in \mathbf{A_v})$ is non-negative, so it must be zero. Plugging this conditional back into the joint gives $q(\mathbf{x_v} \in \mathbf{A_v} \cap \mathbf{S_v}) = 0$. Then expanding yields

$$0 = q(\mathbf{x_v} \in \mathbf{A_v} \cap \mathbf{S_v}) = q(\mathbf{x_v} \in \mathbf{A_v} \mid \mathbf{x_v} \in \mathbf{S_v}) q(\mathbf{x_v} \in \mathbf{S_v}).$$

Since $q(\mathbf{x_v} \in \mathbf{S_v}) > 0$, $q(\mathbf{x_v} \in \mathbf{A_v} \mid \mathbf{x_v} \in \mathbf{S_v})$ must be zero and thus, $q(\mathbf{x_v} \notin \mathbf{A_v} \mid \mathbf{x_v} \in \mathbf{S_v}) = 1$, where expanding out the definition of $\mathbf{A_v}$ gives the desired result that $q(\mathbf{E_v} = 1 \mid \mathbf{x_v} = \mathbf{a}) > 0$ for almost $\mathbf{a} \in \mathbf{S_v}$:

$$\begin{aligned}
1 &= q(\mathbf{a} \notin \mathbf{A_v} \mid \mathbf{a} \in \mathbf{S_v}) \\
&= q(\mathbf{a} \notin \{\mathbf{a} : q(\mathbf{E_v} = 1 \mid \mathbf{x_v} = \mathbf{a}) = 0\} \mid \mathbf{a} \in \mathbf{S_v}) \\
&= q(\mathbf{a} \in \{\mathbf{a} : q(\mathbf{E_v} = 1 \mid \mathbf{x_v} = \mathbf{a}) > 0\} \mid \mathbf{a} \in \mathbf{S_v}).
\end{aligned}$$

□

A.3.2 Optimal value and the optimal gap under EVAL-X

Lemma 4. *The* EVAL-X *optimality gap value for* $e(\cdot)$ *is an averaged* $\mathbf{KL}$ *between* $q(\mathbf{y} \mid \mathbf{x})$ *and* $q(\mathbf{y} \mid \mathbf{x_v})$*:* $\sum_{\mathbf{v} \in \mathcal{V}} q(e(\mathbf{x}) = \mathbf{v}) \mathbb{E}_{q(\mathbf{x} \mid e(\mathbf{x}) = \mathbf{v})} \mathbf{KL}(q(\mathbf{y} \mid \mathbf{x}) \| q(\mathbf{y} \mid \mathbf{x_v}))$*. This gap is zero, i.e.* $e(\mathbf{x})$ *is optimal with the score* EVAL-X$^* = \mathbb{E}_q[\log q(\mathbf{y} \mid \mathbf{x})]$ *if for all* $\mathbf{v}$ *such that* $q(e(\mathbf{x}) = \mathbf{v}) > 0$*,*

$$q(\mathbf{y} \mid \mathbf{x}) = q(\mathbf{y} \mid \mathbf{x_v}) \quad \textit{a.e. in} \quad \{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\}.$$

Proof. Let p be a generic conditional distribution and let $\mathbf{x}_{-\mathbf{v}}$ be the values outside the explanation.

$$\begin{aligned}
\max_e \text{EVAL-X}(q, e) &= \max_e \mathbb{E}_{(\mathbf{v},\mathbf{a}) \sim q(\mathbf{x}_{e(\mathbf{x})})} \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v},\mathbf{a}))} \log q(\mathbf{y} \mid \mathbf{x_v} = \mathbf{a}) \\
&= \max_e \mathbb{E}_{(\mathbf{v},\mathbf{a}) \sim q(\mathbf{x}_{e(\mathbf{x})})} \mathbb{E}_{q(\mathbf{x}_{-\mathbf{v}} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v},\mathbf{a}))} \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v},\mathbf{a}), \mathbf{x}_{-\mathbf{v}})} \log q(\mathbf{y} \mid \mathbf{x_v} = \mathbf{a}) \\
&= \max_e \mathbb{E}_{(\mathbf{v},\mathbf{a}) \sim q(\mathbf{x}_{e(\mathbf{x})})} \mathbb{E}_{q(\mathbf{x} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v},\mathbf{a}))} \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x})} \log q(\mathbf{y} \mid \mathbf{x_v} = \mathbf{a}) \\
&\leq \max_p \mathbb{E}_{(\mathbf{v},\mathbf{a}) \sim q(\mathbf{x}_{e(\mathbf{x})})} \mathbb{E}_{q(\mathbf{x} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v},\mathbf{a}))} \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x})} [\log p(\mathbf{y} \mid \mathbf{x})] \\
&\leq \max_p \mathbb{E}_{q(\mathbf{x})} \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x})} [\log p(\mathbf{y} \mid \mathbf{x})] \\
&= \max_p -\mathbb{E}_{q(\mathbf{x})} \mathbf{KL}(q(\mathbf{y} \mid \mathbf{x}) || p(\mathbf{y} \mid \mathbf{x})) + \mathbb{E}_q[\log q(\mathbf{y} \mid \mathbf{x})] \\
&= \mathbb{E}_q[\log q(\mathbf{y} \mid \mathbf{x})].
\end{aligned}$$

This upper bound is achievable by an explanation that selects all inputs, so the maximum EVAL-X denoted as EVAL-X$^* = \mathbb{E}_q \log q(\mathbf{y} \mid \mathbf{x})$.

As in the math above, the EVAL-X score for an explanation method can be expanded as

$$\begin{aligned}
\text{EVAL-X}^e &= \mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})})}\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})}=(\mathbf{v},\mathbf{a}))} \log q\left(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}} = \mathbf{a}\right) \\
&= \mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})})}\mathbb{E}_{q(\mathbf{x} \mid \mathbf{x}_{e(\mathbf{x})}=(\mathbf{v},\mathbf{a}))}\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x})} \log q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}} = \mathbf{a}) \\
&= \mathbb{E}_{\mathbf{v}\sim q(e(\mathbf{x}))}\mathbb{E}_{\mathbf{a}\sim q(\mathbf{x}_{\mathbf{v}} \mid e(\mathbf{x})=\mathbf{v})}\mathbb{E}_{q(\mathbf{x} \mid \mathbf{x}_{\mathbf{v}}=\mathbf{a}, e(\mathbf{x})=\mathbf{v}))}\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x})} \log q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}} = \mathbf{a}) \\
&= \mathbb{E}_{\mathbf{v}\sim q(e(\mathbf{x}))}\mathbb{E}_{q(\mathbf{x} \mid e(\mathbf{x})=\mathbf{v}))}\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x})} \log q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}),
\end{aligned}$$

where in the last step, we dropped $\mathbf{a}$ because it equals $\mathbf{x}_{\mathbf{v}}$ almost surely. Similarly, the optimal score EVAL-X* expands to

$$\mathbb{E}_{\mathbf{v}\sim q(e(\mathbf{x}))}\mathbb{E}_{q(\mathbf{x} \mid e(\mathbf{x})=\mathbf{v}))}\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x})} \log q(\mathbf{y} \mid \mathbf{x}).$$

Let $\mathcal{V}$ be the set of values that explanations can take on, then taking the difference from optimality

$$\begin{aligned}
&\text{EVAL-X}^* - \text{EVAL-X}^e \\
&= \mathbb{E}_{\mathbf{v}\sim q(e(\mathbf{x}))}\mathbb{E}_{q(\mathbf{x} \mid e(\mathbf{x})=\mathbf{v}))}\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x})} \log \frac{q(\mathbf{y} \mid \mathbf{x})}{q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}} = \mathbf{x})} \\
&= \mathbb{E}_{\mathbf{v}\sim q(e(\mathbf{x}))}\mathbb{E}_{q(\mathbf{x} \mid e(\mathbf{x})=\mathbf{v}))}\mathbf{KL}\left[q(\mathbf{y} \mid \mathbf{x}) \parallel q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}})\right] \\
&= \sum_{\mathbf{v}\in\mathcal{V}} q(e(\mathbf{x}) = \mathbf{v})\mathbb{E}_{q(\mathbf{x} \mid e(\mathbf{x})=\mathbf{v}))}\mathbf{KL}\left[q(\mathbf{y} \mid \mathbf{x}) \parallel q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}})\right].
\end{aligned}$$

As each $\mathbf{KL}$ term is non-negative, each term in the sum being set to 0 simultaneously achieves the optimum, which happens when for all $\mathbf{v}$ such that $q(e(\mathbf{x}) = \mathbf{v}) > 0$,

$$q(\mathbf{y} \mid \mathbf{x}) = q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}) \quad \text{for almost every} \quad \{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\}.$$

□

In Appendix A.4, we use the results from Lemma 1 and Lemma 4 to prove that the optimal score of EVAL-X can only be achieved by non-encoding explanations.

A.4 Proof of Theorem 1

Theorem 1. *If $e(\mathbf{x})$ is* EVAL-X *optimal, then $e(\mathbf{x})$ is not encoding.*

Proof. Note only $q(e(\mathbf{x}) = \mathbf{v}) > 0$ are of interest, since $q(e(\mathbf{x}) = \mathbf{v}) = 0$ implies that $\mathbf{E}_{\mathbf{v}} = 0$ almost surely and thus $\mathbf{y} \perp\!\!\!\perp \mathbf{E}_{\mathbf{v}} \mid \mathbf{x}_{\mathbf{v}}$.

Then if $e(\mathbf{x})$ achieves EVAL-X*, then by Lemma 4, for all $\mathbf{v}$ such that $q(e(\mathbf{x}) = \mathbf{v}) > 0$,

$$q(\mathbf{y} \mid \mathbf{x}) = q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}) \quad \text{for almost every} \quad \{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\}.$$

First, this optimality criteria can incorporate $\mathbf{E}_{\mathbf{v}} = 1$ on the lefthand side by first conditioning on $e(\mathbf{x})$ and then noting that the equality holds for $\mathbf{x}$ where $e(\mathbf{x}) = \mathbf{v}$.

$$\begin{aligned}
& q(\mathbf{y} \mid \mathbf{x}) = q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}) && \text{for almost every} && \{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\} \\
\iff\ & q(\mathbf{y} \mid \mathbf{x}, e(\mathbf{x})) = q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}) && \text{for almost every} && \{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\} \\
\iff\ & q(\mathbf{y} \mid \mathbf{x}, e(\mathbf{x}) = \mathbf{v}) = q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}) && \text{for almost every} && \{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\} \\
\iff\ & q(\mathbf{y} \mid \mathbf{x}, \mathbf{E}_{\mathbf{v}} = 1) = q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}) && \text{for almost every} && \{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\}.
\end{aligned}$$

To understand if the optimality criterion disallows encoding, integrate the left and right-hand sides of this optimality criterion with the respect to complement of the inputs in $\mathbf{x}_{\mathbf{v}}$, $q(\mathbf{x}_{\mathbf{v}}^c \mid \mathbf{x}_{\mathbf{v}}, \mathbf{E}_{\mathbf{v}} = 1)$ yields

$$\begin{aligned}
&\int q(\mathbf{y} \mid \mathbf{x}, \mathbf{E}_{\mathbf{v}} = 1) q(\mathbf{x}_{\mathbf{v}}^c \mid \mathbf{x}_{\mathbf{v}}, \mathbf{E}_{\mathbf{v}} = 1) d\mathbf{x}_{\mathbf{v}}^c = \int q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}) q(\mathbf{x}_{\mathbf{v}}^c \mid \mathbf{x}_{\mathbf{v}}, \mathbf{E}_{\mathbf{v}} = 1) d\mathbf{x}_{\mathbf{v}}^c \\
&\qquad\qquad \text{for almost every } \{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\} \\
\iff\ &\int q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}^c, \mathbf{x}_{\mathbf{v}}, \mathbf{E}_{\mathbf{v}} = 1) q(\mathbf{x}_{\mathbf{v}}^c \mid \mathbf{x}_{\mathbf{v}}, \mathbf{E}_{\mathbf{v}} = 1) d\mathbf{x}_{\mathbf{v}}^c = q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}) \int q(\mathbf{x}_{\mathbf{v}}^c \mid \mathbf{x}_{\mathbf{v}}, \mathbf{E}_{\mathbf{v}} = 1) d\mathbf{x}_{\mathbf{v}}^c \\
&\qquad\qquad \text{for almost every } \{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\} \\
\iff\ & q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}, \mathbf{E}_{\mathbf{v}} = 1) = q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}) \quad \text{for almost every } \{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\} \\
\iff\ & q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}, \mathbf{E}_{\mathbf{v}} = 1) = q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}}) \quad \text{for almost every } \{\mathbf{x}_{\mathbf{v}} : e(\mathbf{x}) = \mathbf{v}\}.
\end{aligned}$$

Now expanding the right-hand side gives

$$q(\mathbf{y} \mid \mathbf{x_v}) = q(\mathbf{y}, \mathbf{E_v} = 1 \mid \mathbf{x_v}) + q(\mathbf{y}, \mathbf{E_v} = 0 \mid \mathbf{x_v}).$$

Combing the two equations gives

$$q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 1) = q(\mathbf{y}, \mathbf{E_v} = 1 \mid \mathbf{x_v}) + q(\mathbf{y}, \mathbf{E_v} = 0 \mid \mathbf{x_v}) \quad \text{for almost every } \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\}. \tag{14}$$

We show that this equality implies that $\mathbf{y} \perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v}$ by splitting the analysis into cases based on $q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) = 1$ and $q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) < 1$. In turn, the condition in Def: Encoding is violated and the explanation $e(\cdot)$ is not encoding.

Case 1: EVAL-X optimality holds when the explanation is predictable. The first case to consider is when the event that the explanation takes the value $\mathbf{v}$ is determined by $\mathbf{x_v}$ for all samples with the explanation $\mathbf{v}$. That is, $q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) = 1$:

$$q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) = 1 \iff q(\mathbf{E_v} = 0 \mid \mathbf{x_v}) = 0.$$

Then expanding this marginal into the joint shows that the joint $q(\mathbf{y}, \mathbf{E_v} = 0 \mid \mathbf{x_v})$ has to be zero as well.

$$q(\mathbf{E_v} = 0 \mid \mathbf{x_v}) = \int q(\mathbf{y}, \mathbf{E_v} = 0 \mid \mathbf{x_v}) d\mathbf{y} = 0,$$

because an integral of non-negative terms being zero implies that each term itself is zero almost surely.

Then, we can show that the determinism condition $q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) = 1$ is sufficient for the optimality criterion eq. (14):

$$\begin{aligned} q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 1) &= q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 1) \times 1 \\ &= q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 1) q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) \\ &= q(\mathbf{y}, \mathbf{E_v} = 1 \mid \mathbf{x_v}) \\ &= q(\mathbf{y}, \mathbf{E_v} = 1 \mid \mathbf{x_v}) + q(\mathbf{y}, \mathbf{E_v} = 0 \mid \mathbf{x_v}) \quad \text{for almost every } \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\}. \end{aligned}$$

This shows that the EVAL-X optimality criteria is satisfied when the $q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) = 1$, thus the explanation is completely predictable from the explanation for examples with that explanation. By Lemma 1, we have

$$q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) = 1 \implies \mathbf{y} \perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v},$$

which violates Def: Encoding. So there is no encoding in this case.

Case 2: When the explanation is unpredictable, EVAL-X optimality requires that the explanation provide no extra information. Now consider the alternative case, $q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) < 1$. Here the explanation does not determine the explanation opening the possibility that the EVAL-X-optimal explanation method can encode information in the explanation.

Because $q(e(\mathbf{x}) = \mathbf{v}) > 0$, we have $q(\mathbf{x_v} \in \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\}) > 0$. Thus, by Lemma 3, for almost every $\{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\}$ it holds that $q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) > 0$. Putting this result together with alternative case ($q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) < 1$) gives: $0 < q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) < 1$ for almost every $\{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\}$.

Now, expanding out the optimality criterion eq. (14):

$$\begin{aligned} & q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 1) = q(\mathbf{y}, \mathbf{E_v} = 1 \mid \mathbf{x_v}) \times 1 + q(\mathbf{y}, \mathbf{E_v} = 0 \mid \mathbf{x_v}) \times 1 \\ & \qquad\qquad \text{for almost every } \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\} \\ \iff & q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 1) = q(\mathbf{y} \mid \mathbf{E_v} = 1, \mathbf{x_v}) q(\mathbf{E_v} = 1 \mid \mathbf{x_v}) \\ & \qquad + q(\mathbf{y} \mid \mathbf{E_v} = 0, \mathbf{x_v})(1 - q(\mathbf{E_v} = 1 \mid \mathbf{x_v})) \\ & \qquad\qquad \text{for almost every } \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\} \\ \iff & (1 - q(\mathbf{E_v} = 1 \mid \mathbf{x_v})) q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 1) = (1 - q(\mathbf{E_v} = 1 \mid \mathbf{x_v})) q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 0) \\ & \qquad\qquad \text{for almost every } \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\} \\ \iff & q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 1) = q(\mathbf{y} \mid \mathbf{x_v}, \mathbf{E_v} = 0) \quad \text{for almost every } \{\mathbf{x_v} : e(\mathbf{x}) = \mathbf{v}\}. \end{aligned}$$

This equality says for all samples with the explanation $\mathbf{v}$, knowing $\mathbf{E_v}$ does not change the distribution of the label $\mathbf{y}$. By Lemma 1, this equality implies that the independence $\mathbf{y} \perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v}$ holds which violates Def: Encoding. □

A.5 Proof of Proposition 3 and Theorem 2

Proposition 3. ENCODE-METER $\phi_q(e) = 0$ *if and only if* e *is not encoding.*

Proof. First, Def: Encoding is violated if and only if there exists a set $\mathbf{A}$ such that $q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{A}) = 1$ and

$$\forall (\mathbf{v}, \mathbf{a}) \in \mathbf{A} \qquad \mathbf{y} \perp\!\!\!\perp \mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v} = \mathbf{a}.$$

For the if direction, note that if ENCODE-METER $\phi_q(e) = 0$,

$$\mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})})} \mathbf{I}(\mathbf{y}; \mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v} = \mathbf{a}) = 0.$$

To show the forward direction, the above equality means that if $\phi_q(e) = 0$, almost surely for every $(\mathbf{v}, \mathbf{a}) \sim q(\mathbf{x}_{e(\mathbf{x})})$, the instantaneous mutual information is 0 which implies the desired conditional independence

$$\mathbf{I}(\mathbf{y}; \mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v} = \mathbf{a}) = 0 \implies \mathbf{y} \perp\!\!\!\perp \mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v} = \mathbf{a}.$$

By definition of almost surely, there exists a set $\mathbf{A}$ such that $q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{A}) = 1$ the independence above holds; this completes the "if" direction.

To show the reverse direction, let there exist a set $\mathbf{A}$ such that $q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{A}) = 1$, for every $(\mathbf{v}, \mathbf{a}) \in \mathbf{A}$,

$$\mathbf{y} \perp\!\!\!\perp \mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v}.$$

In turn, for all $(\mathbf{v}, \mathbf{a}) \in \mathbf{A}$,

$$\mathbf{I}(\mathbf{y}; \mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v} = \mathbf{a}) = 0.$$

Then, the fact that $q(\mathbf{x}_{e(\mathbf{x})} \in \mathbf{A}) = 1$ implies that expectations with respect to $q(\mathbf{x}_{e(\mathbf{x})})$ over the whole support equal expectations over $q(\mathbf{x}_{e(\mathbf{x})} \mid \mathbf{x}_{e(\mathbf{x})} \in \mathbf{A})$, which is $q(\mathbf{x}_{e(\mathbf{x})})$ restricted to $\mathbf{A}$:

$$\mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})})} \mathbf{I}(\mathbf{y}; \mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v} = \mathbf{a}) = \mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})} \mid \mathbf{x}_{e(\mathbf{x})} \in \mathbf{A})} \mathbf{I}(\mathbf{y}; \mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v} = \mathbf{a}) = 0.$$

This completes the "only if" direction. □

Theorem 2. *With finite* $\mathbf{H}(\mathbf{y} \mid \mathbf{x})$ *and* $\mathbf{H}(\mathbf{y})$*, for any explanation that encodes* e *and any that does not encode* e'*, there exists an* α^* *such that* $\forall \alpha > \alpha^*$ STRIPE-X$_\alpha(q, e')$ $>$ STRIPE-X$_\alpha(q, e)$.

Proof. Recall that STRIPE-X is

$$\text{STRIPE-X}_\alpha(q, e) := \mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})})} \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v},\mathbf{a}))} [\log q(\mathbf{y} \mid \mathbf{x}_\mathbf{v} = \mathbf{a})] - \alpha \phi_q(e),$$

where the ENCODE-METER

$$\phi_q(e) := \mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})})} \mathbf{I}(\mathbf{E}_\mathbf{v}; \mathbf{y} \mid \mathbf{x}_\mathbf{v} = \mathbf{a}).$$

We first show bounds for the first term in STRIPE-X and then derive the STRIPE-X scores for encoding and non-encoding explanations.

Bounds on EVAL-X scores. We lower bound the EVAL-X score, which is the first term in STRIPE-X, for non-encoding explanations.

For non-encoding explanations, almost surely over $\mathbf{v}, \mathbf{a} \sim q(\mathbf{x}_{e(\mathbf{x})})$

$$q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})) = q(\mathbf{y} \mid \mathbf{x}_\mathbf{v} = \mathbf{a}).$$

Then,

$$\begin{aligned}
&\text{EVAL-X}(q, e) - \mathbb{E}_{q(\mathbf{y}, \mathbf{x}_\mathbf{v} = \mathbf{a})} \log q(\mathbf{y}) \\
&= \mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})})} \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v},\mathbf{a}))} \log q(\mathbf{y} \mid \mathbf{x}_\mathbf{v} = \mathbf{a}) - \mathbb{E}_{q(\mathbf{y}, \mathbf{x}_{e(\mathbf{x})})} \log q(\mathbf{y}) \\
&= \mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})})} \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v},\mathbf{a}))} [\log q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a}))] - \mathbb{E}_{q(\mathbf{y}, \mathbf{x}_{e(\mathbf{x})})} \log q(\mathbf{y}) \\
&= \mathbb{E}_{q(\mathbf{y}, \mathbf{x}_{e(\mathbf{x})})} [\log q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})})] - \mathbb{E}_{q(\mathbf{y}, \mathbf{x}_{e(\mathbf{x})})} \log q(\mathbf{y}) \\
&= \mathbb{E}_{q(\mathbf{y}, \mathbf{x}_{e(\mathbf{x})})} \log \frac{q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})})}{q(\mathbf{y})} \\
&= \mathbf{I}(\mathbf{y}; \mathbf{x}_{e(\mathbf{x})}).
\end{aligned}$$

The above inequality implies that

$$\begin{aligned}&\text{EVAL-X}(q,e) - \mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})})}\mathbb{E}_{\mathbf{y}\sim q(\mathbf{y}\mid \mathbf{x}_{e(\mathbf{x})}=(\mathbf{v},\mathbf{a}))}\log q(\mathbf{y}) = \mathbf{I}(\mathbf{y};\mathbf{x}_{e(\mathbf{x})}) \geq 0\\ &\implies \text{EVAL-X}(q,e) + \mathbf{H}_q(\mathbf{y}) \geq 0\\ &\implies \text{EVAL-X}(q,e) \geq -\mathbf{H}_q(\mathbf{y}).\end{aligned}$$

Every inequality in the derivation above becomes strict when the explanation selects inputs that are predictive of the label because

$$\mathbf{I}(\mathbf{y};\mathbf{x}_{e(\mathbf{x})}) > 0.$$

Thus, non-encoding explanations have EVAL-X scores that are at least $-\mathbf{H}_q(\mathbf{y})$.

The optimal EVAL-X score for any explanation (see Lemma 4) equals the negative conditional entropy which is upper bounded by some finite number:

$$\mathbb{E}_q\left[\log q(\mathbf{y}\mid\mathbf{x})\right] = -\mathbf{H}_q(\mathbf{y}\mid\mathbf{x}) = C.$$

Comparing explanations via STRIPE-X. For any encoding explanation, by Proposition 3, for some $c > 0$

$$\phi_q(e) > c.$$

Now, consider $\alpha^* = \frac{\mathbf{I}(\mathbf{y};\mathbf{x})}{c} \geq 0$, which is finite because each term in the ratio is finite. Then, for all $\alpha > \alpha^*$

$$\alpha\phi_q(e) > \alpha^*\phi_q(e) > \mathbf{I}(\mathbf{y};\mathbf{x}).$$

Thus,

$$-\alpha\phi_q(e) < -\mathbf{I}(\mathbf{y};\mathbf{x}).$$

As EVAL-X scores are below $C = -\mathbf{H}(\mathbf{y}\mid\mathbf{x})$ for any encoding explanation,

$$\begin{aligned}\text{STRIPE-X}_\alpha(q,e) &= \text{EVAL-X}(q,e) - \alpha\phi_q(e)\\ &< -\mathbf{H}(\mathbf{y}\mid\mathbf{x}) - \mathbf{I}(\mathbf{y};\mathbf{x})\\ &< -\mathbf{H}(\mathbf{y}\mid\mathbf{x}) - (\mathbf{H}_q(\mathbf{y}) - \mathbf{H}(\mathbf{y}\mid\mathbf{x}))\\ &= -\mathbf{H}_q(\mathbf{y}).\end{aligned}$$

Finally, for any non-encoding explanation, $\phi_q(e') = 0$ by Proposition 3, STRIPE-X scores equal EVAL-X scores, which are lower bounded at $-\mathbf{H}_q(\mathbf{y})$.

Together, for every non-encoding explanation $e'(\mathbf{x})$ and encoding explanation $e(\mathbf{x})$, it holds that

$$\text{STRIPE-X}_\alpha(q,e') \geq -\mathbf{H}_q(\mathbf{y}) > \text{STRIPE-X}_\alpha(q,e).$$

This proves that STRIPE-X is a strong detector of encoding. □

B Encoding examples, non-detection of ROAR,FRESH, and non-strong detection of EVAL-X

B.1 An illustrative DGP for Def: Encoding

With $\mathcal{B}(0.5)$ being a Bernoulli distribution, consider the following example

$$\begin{aligned}&\mathbf{y}\sim\mathcal{B}(0.5) \quad , \quad \mathbf{z}\sim\mathcal{B}(0.5) \quad , \quad \epsilon_1,\epsilon_2,\epsilon_3\sim\mathcal{N}(0,\mathbf{I}),\\ &\mathbf{x} \quad = \quad \begin{cases}[\mathbf{y}+\epsilon_1, \quad \epsilon_3, 0, \epsilon_2] & \text{if } \mathbf{z}=0,\\ [\quad \epsilon_3, \mathbf{y}+\epsilon_1, 1, \epsilon_2] & \text{if } \mathbf{z}=1.\end{cases}\end{aligned}$$

For this problem, if the third coordinate $\mathbf{x}_3 = 0$, all the information between the label and the covariates is in the first coordinate $\mathbf{x}_1$, and if $\mathbf{x}_3 = 1$, the information is between the label and the second coordinate $\mathbf{x}_2$. The corresponding explanation function is $e(\mathbf{x}) = \mathbb{1}[\mathbf{x}_3 = 0]\xi_1 + \mathbb{1}[\mathbf{x}_3 = 1]\xi_2$. This explanation is encoding because neither explanation's values $\mathbf{x}_1$ nor $\mathbf{x}_2$ determine the explanation function because it depends on $\mathbf{x}_3$. Formally

$$q(\mathbf{y}=1\mid\mathbf{x}_1,\mathbf{x}_3=1) \neq q(\mathbf{y}=1\mid\mathbf{x}_1,\mathbf{x}_3=0) \implies \mathbf{y}\not\perp\!\!\!\perp\mathbf{x}_3\mid\mathbf{x}_1 \implies \mathbf{y}\not\perp\!\!\!\perp\mathbf{E}_\mathbf{v}\mid\mathbf{x}_1,$$

which meets Def: Encoding. Consider an alternate non-encoding explanation function $e(\mathbf{x}) = [\mathbb{1}[\mathbf{x}_4 > 0], \mathbb{1}[\mathbf{x}_4 \leq 0], 0, 0]$; $\mathbf{x}_1, \mathbf{x}_2$ do not determine $e(\mathbf{x})$ that depends on the noise ϵ_2 in $\mathbf{x}_4$. That means the unpredictability property in Lemma 1 holds. However, by construction,

$$(\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2) \perp\!\!\!\perp \epsilon_2 \implies (\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2) \perp\!\!\!\perp \mathbf{E}_\mathbf{v} \implies \mathbf{y} \perp\!\!\!\perp \mathbf{E}_\mathbf{v} \mid \mathbf{x}_1 \quad \text{and} \quad \mathbf{y} \perp\!\!\!\perp \mathbf{E}_\mathbf{v} \mid \mathbf{x}_2.$$

So no additional information about the label is encoded:

$$\mathbf{y} \perp\!\!\!\perp \mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v}.$$

The additional information property in Lemma 1 avoids such cases where the explanations keeps additional information that is irrelevant to the label.

B.2 Encoding explanations conceal predictive inputs that affect the explanation

Consider the following DGP

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3] \sim \mathcal{B}(0.5)^{\otimes 3}, \qquad \mathbf{y} = \begin{cases} \mathbf{x}_1 & \text{if } \mathbf{x}_3 = 1, \\ \mathbf{x}_2 & \text{if } \mathbf{x}_3 = 0. \end{cases} \tag{15}$$

Let e be an encoding explanation that selects the first coordinate if $\mathbf{x}_3 = 1$ and the second coordinate otherwise. We never observe $\mathbf{x}_3$ when looking only at the explanation. Table 4 shows all possible values of this explanation. Notice that in the third and fourth rows, the value of $\mathbf{x}_{e(\mathbf{x})}$ changes to match the label $\mathbf{y}$ exactly, even though the values of the first two coordinates that we can observe stay constant. It is impossible to understand the perfect predictiveness of $\mathbf{x}_{e(\mathbf{x})}$, as the encoding explanation conceals the control flow feature $\mathbf{x}_3$ that determines which of the first two features should be picked to predict the label.

Table 4: Possible values of the inputs, label, and explanation for the DGP in eq. (15)

				$\mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})$	
$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{y}$	$\mathbf{v}$	$\mathbf{a}$
0	0	0	0	[0, 1, 0]	0
0	0	1	0	[1, 0, 0]	0
0	1	0	1	[0, 1, 0]	1
0	1	1	0	[1, 0, 0]	0
1	0	0	0	[0, 1, 0]	0
1	0	1	1	[1, 0, 0]	1
1	1	0	1	[0, 1, 0]	1
1	1	1	1	[1, 0, 0]	1

B.3 Position-based encoding fits Def: Encoding

Recall the perceptual task that classifying images of dogs versus classifying images of cats, and consider the encoding explanation $e_{\text{position}}(\mathbf{x})$ that is

$$e_{\text{position}}(\mathbf{x}) = \xi_1 \quad \text{if} \quad q(\mathbf{y} = \text{dog} \mid \mathbf{x}) = 1,$$
$$e_{\text{position}}(\mathbf{x}) = \xi_2 \quad \text{if} \quad q(\mathbf{y} = \text{cat} \mid \mathbf{x}) = 1.$$

Assume that the inputs in the top leftmost pixels are always background, meaning that the values of these inputs provide no information about the label $\mathbf{y} \perp\!\!\!\perp \mathbf{x}_1, \mathbf{x}_2$. Now we check if this intuitively-defined position-encoded explanation meets the definition for encoding (Def: Encoding). To condition on $\mathbf{x}_{\xi_1}, \mathbf{E}_{\xi_1} = 0$, we need $q(\mathbf{y} = \text{dog}) \neq 1$. Note that

$$q(\mathbf{E}_{\xi_1} = 1 \mid \mathbf{x}_{\xi_1}) = q(\mathbf{y} = \text{dog} \mid \mathbf{x}_{\xi_1}) = q(\mathbf{y} = \text{dog}) \neq 1.$$

Def: Encoding holds because the indicator of which explanation was chosen $\mathbf{E}_{\xi_1}$ determines the label.

$$q(\mathbf{y} = \text{dog} \mid \mathbf{x}_{\xi_1}, \mathbf{E}_{\xi_1} = 1) = 1 \neq 0 = q(\mathbf{y} = \text{dog} \mid \mathbf{x}_{\xi_1}, \mathbf{E}_{\xi_1} = 0) \implies \mathbf{y} \not\perp\!\!\!\perp \mathbf{E}_{\xi_1} \mid \mathbf{x}_{\xi_1}.$$

This example shows how the encoding definition Def: Encoding captures the informally described position-based encoding from the literature.

B.4 Prediction-based encoding fits Def: Encoding

The informal example of prediction-based encoding from Section 3.1 selects a single input that makes the prediction from all of the input have the highest confidence when given the single input. One way to mathematically express such a selection is as follows:

$$\begin{aligned} e_{\text{prediction}}(\mathbf{x}) &= \xi_{\text{argmax}_i\, q(\mathbf{y}=1 \mid \mathbf{x}_i)} \text{ if } q(\mathbf{y}=1 \mid \mathbf{x}) > 0.5, \\ e_{\text{prediction}}(\mathbf{x}) &= \xi_{\text{argmin}_i\, q(\mathbf{y}=1 \mid \mathbf{x}_i)} \text{ if } q(\mathbf{y}=1 \mid \mathbf{x}) \leq 0.5. \end{aligned} \tag{16}$$

Here, we describe one set of conditions on the distribution $q(\mathbf{y}, \mathbf{x})$ for which the explanation in eq. (16) fits the definition of encoding in Def: Encoding. Assume that there exists a non-measure-zero set $\mathbf{U} \subseteq \{\mathbf{x} : q(\mathbf{y}=1 \mid \mathbf{x}) > 0.5\}$ and an index k such that

$$\mathbf{x} \in \mathbf{U} \implies q(\mathbf{y}=1 \mid \mathbf{x}) \geq \rho, \tag{17}$$

$$\mathbf{x} \in \mathbf{U} \implies \forall i \quad q(\mathbf{y}=1 \mid \mathbf{x}_{\xi_i}) < q(\mathbf{y}=1 \mid \mathbf{x}_{\xi_k}), \tag{18}$$

$$\mathbf{x} \notin \mathbf{U} \implies q(\mathbf{y}=1 \mid \mathbf{x}) < \rho, \tag{19}$$

$$\mathbf{x} \notin \mathbf{U} \implies \exists i, j \quad q(\mathbf{y}=1 \mid \mathbf{x}_{\xi_i}) > q(\mathbf{y}=1 \mid \mathbf{x}_{\xi_k}) > q(\mathbf{y}=1 \mid \mathbf{x}_{\xi_j}). \tag{20}$$

Further, assume that $\mathbf{x}_{\xi_k}$ alone does not determine $\mathbf{x} \in \mathbf{U}$:

$$0 < \mathbb{E}\left[\mathbb{1}[\mathbf{x} \in \mathbf{U}] \mid \mathbf{x}_{\xi_k}\right] < 1. \tag{21}$$

The assumptions above imply the facts below about $e_{\text{prediction}}(\mathbf{x})$:

1. By eqs. (18) and (20)
$$\mathbf{x} \in \mathbf{U} \Leftrightarrow e_{\text{prediction}}(\mathbf{x}) = \xi_k.$$
Define the explanation indicator $\mathbf{E}_{\mathbf{v}} = \mathbb{1}[e_{\text{prediction}}(\mathbf{x}) = \mathbf{v}]$.
2. By eqs. (18) and (20) and the definition of $\mathbf{E}_{\mathbf{v}}$
$$\mathbf{x} \in \mathbf{U} \Leftrightarrow \mathbf{E}_{\xi_k} = 1. \tag{22}$$
3. By eqs. (21) and (22)
$$0 < q(\mathbf{E}_{\xi_k} = 1 \mid \mathbf{x}_{\xi_k}) = q(\mathbf{x} \in \mathbf{U} \mid \mathbf{x}_{\xi_k}) < 1. \tag{23}$$
4. By eq. (23), $q(\mathbf{y}=1 \mid \mathbf{x}_{\xi_k}, \mathbf{E}_{\xi_k}=1)$ and $q(\mathbf{y}=1 \mid \mathbf{x}_{\xi_k}, \mathbf{E}_{\xi_k}=0)$ are well defined. Then, by eqs. (17) and (19), for $\mathbf{x} \in \mathbf{U}$

$$\begin{aligned} &q(\mathbf{y}=1 \mid \mathbf{x}_{\xi_k}, \mathbf{E}_k = 1) \\ &\quad = \mathbb{E}_{q(\mathbf{x} \mid \mathbf{x}_{\xi_k}, \mathbf{E}_k=1)} q(\mathbf{y} \mid \mathbf{x}) \\ &\quad \geq \mathbb{E}_{q(\mathbf{x} \mid \mathbf{x}_{\xi_k}, \mathbf{E}_k=1)} \rho \qquad \{\text{as } \mathbf{E}_k = 1 \implies \mathbf{x} \in \mathbf{U}\} \\ &\quad = \rho. \end{aligned}$$

$$\begin{aligned} &q(\mathbf{y}=1 \mid \mathbf{x}_{\xi_k}, \mathbf{E}_k = 0) \\ &\quad = \mathbb{E}_{q(\mathbf{x} \mid \mathbf{x}_{\xi_k}, \mathbf{E}_k=0)} q(\mathbf{y} \mid \mathbf{x}) \\ &\quad < \mathbb{E}_{q(\mathbf{x} \mid \mathbf{x}_{\xi_k}, \mathbf{E}_k=0)} \rho \qquad \{\text{as } \mathbf{E}_k = 0 \implies \mathbf{x} \notin \mathbf{U}\} \\ &\quad = \rho. \end{aligned}$$

Thus, for all elements of $\{\mathbf{x}_{\xi_k} : \mathbf{x} \in \mathbf{U}\}$

$$q(\mathbf{y}=1 \mid \mathbf{x}_{\xi_k}, \mathbf{E}_{\xi_k}=1) > q(\mathbf{y}=1 \mid \mathbf{x}_{\xi_k}, \mathbf{E}_{\xi_k}=0). \tag{24}$$

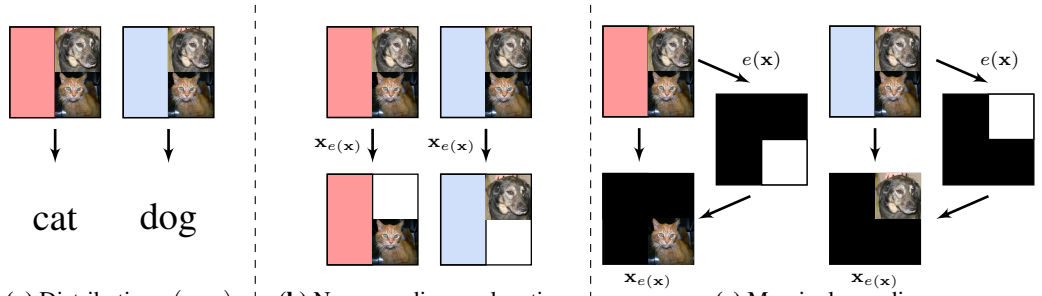
By Lemma 1, the properties in eqs. (23) and (24) imply that $\forall \mathbf{a} \in \{\mathbf{x}_{\xi_k} : \mathbf{x} \in \mathbf{U}\}$

$$\mathbf{y} \not\perp\!\!\!\perp \mathbf{E}_{\xi_k} \mid \mathbf{x}_{\xi_k} = \mathbf{a}.$$

Finally, the set $\mathbf{U}_k = \{\mathbf{x}_{\xi_k} : \mathbf{x} \in \mathbf{U}\}$ is non-measure-zero: as $\mathbb{1}[\mathbf{x} \in \mathbf{U}] = 1 \implies \mathbb{1}[\mathbf{x}_{\xi_k} \in \mathbf{U}_k] = 1$, accumulating $q(\mathbf{x})$ with the restriction $\mathbf{x}_{\xi_k} \in \mathbf{U}_k$ leads to at least as much mass as accumulating with the stricter restriction $\mathbf{x} \in \mathbf{U}$:

$$q(\mathbf{x}_{\xi_k} \in \mathbf{U}_k) = \int q(\mathbf{x}) \mathbb{1}[\mathbf{x}_{\xi_k} \in \mathbf{U}_k] d\mathbf{x} \geq \int q(\mathbf{x}) \mathbb{1}[\mathbf{x} \in \mathbf{U}] d\mathbf{x} = q(\mathbf{x} \in \mathbf{U}) > 0.$$

Together, the last two equations implies that Def: Encoding holds for $e_{\text{prediction}}(\mathbf{x})$ from eq. (16).



(a) Distribution $q(\mathbf{x}, \mathbf{y})$ **(b)** Non-encoding explanation **(c)** Marginal encoding

Figure 5: Example DGP and MARG encoding. **(a)** The color determines whether the label is produced from the top or bottom image. **(b)** An explanation that correctly reveals that the label is generated based on both the color and, as dictated by the color, the top or the bottom image. The label is deterministic given the value of the explanation which means the label can be predicted perfectly. **(c)** An encoding explanation would be one that produces only the top or the bottom animal image based on the color being red of blue respectively. This returned animal image does not indicate the fact that the data generating process depends on color. Now, the animal image selected by the explanation alone is insufficient to dictate the label because the color determines which image determines the label. The identity of the image, whether top or bottom, provides additional information about the label beyond the values explanation, as captured in Def: Encoding.

B.5 MARG explanations are encoding

We provide an illustrative example of a MARG explanation for the DGP in Figure 5. Here, we show how a mathematical formulation of MARG satisfied Def: Encoding.

Consider a generic DGP with a Bernoulli control flow input denoted $\mathbf{x}_c$: for some distinct sets U, V that do not include c and let no combination of $\mathbf{x}_c, \mathbf{x}_U, \mathbf{x}_V$ determine the rest

$$\begin{aligned} \mathbf{x}_c = 1 &\implies q(\mathbf{y} \mid \mathbf{x}) = q(\mathbf{y} \mid \mathbf{x}_U), \\ \mathbf{x}_c = 0 &\implies q(\mathbf{y} \mid \mathbf{x}) = q(\mathbf{y} \mid \mathbf{x}_V). \end{aligned}$$

Further, assume that the two subsets leads to different distributions over $\mathbf{y} = 1$ such that on a non-zero measure subset $\mathbf{S}_U \subseteq \{\mathbf{x}_U : \mathbf{x} \text{ such that } \mathbf{x}_c = 1\}$

$$q(\mathbf{y} = 1 \mid \mathbf{x}_U, \mathbf{x}_c = 1) \neq q(\mathbf{y} = 1 \mid \mathbf{x}_U, \mathbf{x}_c = 0). \tag{25}$$

Now, consider a MARG explanation that looks at $\mathbf{x}_c$ and outputs the corresponding sets U, V:

$$e(\mathbf{x}) = U \quad \text{if} \quad \mathbf{x}_c = 1 \quad \text{else} \quad e(\mathbf{x}) = V.$$

By definition the explanation only depends on the control flow input, not by $\mathbf{x}_U$ or $\mathbf{x}_V$. Next, as $\mathbf{E}_U = 1$ is the same event as $\mathbf{x}_c = 1$, $e(\mathbf{x})$ is encoding because the assumption from eq. (25) implies:

$$q(\mathbf{y} \mid \mathbf{x}_U, \mathbf{E}_U = 1) \neq q(\mathbf{y} \mid \mathbf{x}_U, \mathbf{E}_U = 0).$$

Then, this inequality holds for all elements of the non-measure-zero set $\mathbf{S}_U$, by Lemma 1, MARG is encoding.

B.6 Proof of Proposition 1

Definition 4. *We denote* $\texttt{val}(\mathbf{x}_{e(\mathbf{x})})$ *as the function that maps explanation* $\mathbf{x}_{e(\mathbf{x})} = (\mathbf{v}, \mathbf{a})$ *to the values the inputs take at the selected indices, right-padded to have the same dimension as the input* $\mathbf{x} \in \mathbf{R}^d$*:*

$$\texttt{val}(\mathbf{x}_{e(\mathbf{x})})_j = \begin{cases} \mathbf{a}_j & \text{if } 1 \leq j \leq \sum_{i=1}^{d} \mathbf{v}_i \\ \texttt{pad-token} & \text{if } \sum_{i=1}^{d} \mathbf{v}_i < j \leq d \end{cases}$$

For example, if $\mathbf{x} = [\alpha, \beta, \gamma]$ *and* $e(\mathbf{x}) = [0, 1, 0]$*, then* $\mathbf{x}_{e(\mathbf{x})} = ([0, 1, 0], [\beta])$ *and*

$$\texttt{val}(\mathbf{x}_{e(\mathbf{x})}) = [\beta, \texttt{pad-token}, \texttt{pad-token}].$$

Table 5: Probability table for the DGP in eq. (3). Conditional on the explanation $\mathbf{x}_3$, does predict the label. For example, given knowing $\mathbf{x}_1 = 1$, if $\mathbf{x}_3 = 1$ implies $p(\mathbf{y} = 1) = 0.9$ but if $\mathbf{x}_3 = 0$, $p(\mathbf{y} = 1) = 0.5$. The probability table in Table 5 shows this.

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$p(\mathbf{y} = 1 \mid \mathbf{x})$	$e(\mathbf{x})$	$\texttt{val}(\mathbf{x}_{e(\mathbf{x})})$
0	0	0	0.1	ξ_2	[0, pad-token, pad-token]
0	0	1	0.1	ξ_1	[0, pad-token, pad-token]
0	1	0	0.9	ξ_2	[1, pad-token, pad-token]
0	1	1	0.1	ξ_1	[0, pad-token, pad-token]
1	0	0	0.1	ξ_2	[0, pad-token, pad-token]
1	0	1	0.9	ξ_1	[1, pad-token, pad-token]
1	1	0	0.9	ξ_2	[1, pad-token, pad-token]
1	1	1	0.9	ξ_1	[1, pad-token, pad-token]

Proposition 1. *For the* DGP *in eq. (3),* ROAR *and* FRESH *assign their respective optimal scores to the encoding explanation* $e_{encode}(\mathbf{x})$.

Proof. First, note that

$$\mathbf{x}_3 \perp\!\!\!\perp \mathbf{y}.$$

See Table 5 for the probability table for why this is true.

For this proof let $e(\mathbf{x}) = e_{\text{encode}}(\mathbf{x})$. In the example eq. (3), masking out the inputs selected by $e(\mathbf{x})$ would mean that

$$\mathbf{x}_3 = 1 \implies \mathbf{x}_{-e(\mathbf{x})} = (\mathbf{1} - e(\mathbf{x}), [\mathbf{x}_2, \mathbf{x}_3]), \qquad \mathbf{x}_3 = 0 \implies \mathbf{x}_{-e(\mathbf{x})} = (\mathbf{1} - e(\mathbf{x}), [\mathbf{x}_1, \mathbf{x}_3]).$$

In turn, by the construction in eq. (3)

$$\begin{matrix}\mathbf{x}_{-e(\mathbf{x})} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}_3 = 1, \\ \mathbf{x}_{-e(\mathbf{x})} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}_3 = 0.\end{matrix} \implies \mathbf{x}_{-e(\mathbf{x})} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}_3 \implies (\mathbf{x}_{-e(\mathbf{x})}, \mathbf{x}_3) \perp\!\!\!\perp \mathbf{y} \implies \mathbf{x}_{-e(\mathbf{x})} \perp\!\!\!\perp \mathbf{y}, \tag{26}$$

where the conditional independence in the second step turns into the joint independence in the third step due to $\mathbf{x}_3 \perp\!\!\!\perp \mathbf{y}$.

ROAR scores an explanation highly if $\mathbf{x}_{-e(\mathbf{x})}$ predicts the label poorly. So if $\mathbf{x}_{-e(\mathbf{x})}$ is independent of $\mathbf{y}$, then $e(\mathbf{x})$ would be scored optimally. Then, due to eq. (26), ROAR scores an encoding explanation optimally.

FRESH scores an explanation highly if the selected value $\texttt{val}(\mathbf{x}_{e(\mathbf{x})})$ (as defined in Definition 4) predicts the label well. From Table 5, we see that $p(\mathbf{y} = 1 \mid \mathbf{x}) = 0.9$ for all cases with $\texttt{val}(\mathbf{x}_{e(\mathbf{x})}) = [1, \texttt{pad-token}, \texttt{pad-token}]$ and $p(\mathbf{y} = 1 \mid \mathbf{x}) = 0.1$ for all cases with $\texttt{val}(\mathbf{x}_{e(\mathbf{x})}) = [0, \texttt{pad-token}, \texttt{pad-token}]$. Thus, if $\texttt{val}(\mathbf{x}_{e(\mathbf{x})}) = [1, \texttt{pad-token}, \texttt{pad-token}]$ then

$$\begin{aligned} p(\mathbf{y} = 1 \mid \mathbf{x}) &= 0.9 \\ &= \mathbb{E}_{\mathbf{x}' \sim p(\mathbf{x}' \mid \texttt{val}(\mathbf{x}_{e(\mathbf{x})}) = [1, \texttt{pad-token}, \texttt{pad-token}])}[0.9] \\ &= \mathbb{E}_{\mathbf{x}' \sim p(\mathbf{x}' \mid \texttt{val}(\mathbf{x}_{e(\mathbf{x})}) = [1, \texttt{pad-token}, \texttt{pad-token}])}[p(\mathbf{y} = 1 \mid \mathbf{x}')] \\ &= p(\mathbf{y} = 1 \mid \texttt{val}(\mathbf{x}_{e(\mathbf{x})}) = [1, \texttt{pad-token}, \texttt{pad-token}]), \end{aligned}$$

and if $\texttt{val}(\mathbf{x}_{e(\mathbf{x})}) = [0, \texttt{pad-token}, \texttt{pad-token}]$ then

$$\begin{aligned} p(\mathbf{y} = 1 \mid \mathbf{x}) &= 0.1 \\ &= \mathbb{E}_{\mathbf{x}' \sim p(\mathbf{x}' \mid \texttt{val}(\mathbf{x}_{e(\mathbf{x})}) = [0, \texttt{pad-token}, \texttt{pad-token}])}[0.1] \\ &= \mathbb{E}_{\mathbf{x}' \sim p(\mathbf{x}' \mid \texttt{val}(\mathbf{x}_{e(\mathbf{x})}) = [0, \texttt{pad-token}, \texttt{pad-token}])} p(\mathbf{y} = 1 \mid \mathbf{x}') \\ &= p(\mathbf{y} = 1 \mid \texttt{val}(\mathbf{x}_{e(\mathbf{x})}) = [0, \texttt{pad-token}, \texttt{pad-token}]). \end{aligned}$$

Therefore, in all cases, $p(\mathbf{y} = 1 \mid \mathbf{x}) = p(\mathbf{y} = 1 \mid \texttt{val}(\mathbf{x}_{e(\mathbf{x})}))$. Thus, predicting label from the selected value alone is as good as predicting from the whole input. As a result, FRESH scores this explanation optimally.

□

B.7 Showing that e_{encode} is the optimal reductive explanation for eq. (3) and scores better than a constant explanation under EVAL-X

We repeat the DGP in eq. (3) here

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3] \sim \mathcal{B}(0.5)^{\otimes 3},$$
$$\mathbf{y} = \begin{cases} \mathbf{x}_1 & \text{w.p. } 0.9 \quad \text{else} \quad 1 - \mathbf{x}_1 \quad \text{if } \mathbf{x}_3 = 1, \\ \mathbf{x}_2 & \text{w.p. } 0.9 \quad \text{else} \quad 1 - \mathbf{x}_2 \quad \text{if } \mathbf{x}_3 = 0. \end{cases}$$

Lemma 5. *In the* DGP *in eq. (3),*

$$q(\mathbf{y} = 1 \mid \mathbf{x}_1 = 1) = 0.7$$
$$q(\mathbf{y} = 1 \mid \mathbf{x}_2 = 1) = 0.7,$$
$$q(\mathbf{y} = 1 \mid \mathbf{x}_1 = 0) = 0.3,$$
$$q(\mathbf{y} = 1 \mid \mathbf{x}_2 = 0) = 0.3.$$
$$q(\mathbf{y} = 1 \mid \mathbf{x}_3) = 0.5.$$

Proof. We can compute these values from Table 5. □

Proposition 2. *Let* $e_{\mathbf{c}}(\mathbf{x}) = \xi_3$. *Then, for the* DGP *in eq. (3),* $\text{EVAL-X}(q, e_{encode}) > \text{EVAL-X}(q, e_{\mathbf{c}})$.

Proof. By Lemma 5, we have

$$\text{EVAL-X}(q, e_c) = \mathbb{E}_{q(\mathbf{y},\mathbf{x})} \log q(\mathbf{y} \mid \mathbf{x}_3) = \mathbb{E}_{q(\mathbf{y},\mathbf{x})} \log 0.5 \approx -0.69.$$

Now, denote e_{encode} as e_e for ease of reading

$$\begin{aligned}
\text{EVAL-X}(q, e_e) &= \mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q(\mathbf{x}_{e(\mathbf{x})})} \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})}=(\mathbf{v},\mathbf{a}))} [\log q(\mathbf{y} \mid \mathbf{x}_{\mathbf{v}} = \mathbf{a})] \\
&= q(\mathbf{x}_3 = 1) \mathbb{E}_{q(\mathbf{x}_1)} \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_1, \mathbf{x}_3 = 1)} \log q(\mathbf{y} \mid \mathbf{x}_1) \\
&\quad + q(\mathbf{x}_3 = 0) \mathbb{E}_{q(\mathbf{x}_2)} \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_2, \mathbf{x}_3 = 0)} \log q(\mathbf{y} \mid \mathbf{x}_2) \\
&= 0.5 * 0.5 * (0.9 * -\log 0.7 + 0.1 * -\log 0.3) * 2 \\
&\quad + 0.5 * 0.5 * (0.9 * -\log 0.7 + 0.1 * -\log 0.3) * 2 \\
&\approx -0.44.
\end{aligned}$$

This concludes that $\text{EVAL-X}(q, e_{\text{encode}}) > \text{EVAL-X}(q, e_c)$. □

Lemma 6. *In the* DGP *in eq. (3),* $e_{encode}(\mathbf{x})$ *is an* EVAL-X*-optimal reductive explanation and is encoding.*

Proof. First, the following properties show for the DGP because when $\mathbf{x}_3 = 1$, $\mathbf{y}$ only depends on $\mathbf{x}_1$, and if $\mathbf{x}_3 = 0$, $\mathbf{y}$ only depends on $\mathbf{x}_2$:

$$\mathbf{x}_3 \perp\!\!\!\perp \mathbf{y} \quad , \quad \mathbf{y} \perp\!\!\!\perp \mathbf{x}_2 \mid \mathbf{x}_3 = 1 \quad , \quad \mathbf{y} \perp\!\!\!\perp \mathbf{x}_1 \mid \mathbf{x}_3 = 0.$$

These independencies imply that

$$q(\mathbf{y} \mid \mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, 1]) = q(\mathbf{y} \mid \mathbf{x}_1, \mathbf{x}_3 = 1) \quad , \quad q(\mathbf{y} \mid \mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, 0]) = q(\mathbf{y} \mid \mathbf{x}_2, \mathbf{x}_3 = 0).$$

Then, the optimal explanation function that achieves EVAL-X^* is $e(\mathbf{x}) = [1, 0, 1]$ if $\mathbf{x}_3 = 1$ and $[0, 1, 1]$ otherwise.

Reductive explanations of size 1. If the explanation is forced to have fewer than 2 inputs, the optimal reductive explanation $e(\mathbf{x})$ is only allowed to be one of ξ_1, ξ_2, ξ_3:

$$\max_{e:|e(\mathbf{x})|\leq 1} \mathbb{E}_{q(\mathbf{y},\mathbf{x})} \sum_{i\in\{1,2,3\}} \mathbb{1}[e(\mathbf{x}) = \xi_i] q(\mathbf{y} \mid \mathbf{x}_i).$$

Rewriting this expression to split the support of $\mathbf{x}$ based on $\mathbf{x}_3 = 1$ or 0:

$$\begin{aligned}
&\mathbb{E}_{q(\mathbf{y},\mathbf{x})} \sum_{i\in\{1,2,3\}} \mathbb{1}[e(\mathbf{x}) = \xi_i] \log q(\mathbf{y} \mid \mathbf{x}_i) \\
&= q(\mathbf{x}_3 = 1)\mathbb{E}_{q(\mathbf{x}_2 \mid \mathbf{x}_3=1)}\mathbb{E}_{q(\mathbf{y},\mathbf{x}_1 \mid \mathbf{x}_3=1)} \sum_{i\in\{1,2,3\}} \mathbb{1}[e(\mathbf{x}) = \xi_i] \log q(\mathbf{y} \mid \mathbf{x}_i) \\
&\quad + q(\mathbf{x}_3 = 0)\mathbb{E}_{q(\mathbf{x}_1 \mid \mathbf{x}_3=0)}\mathbb{E}_{q(\mathbf{y},\mathbf{x}_2 \mid \mathbf{x}_3=0)} \sum_{i\in\{1,2,3\}} \mathbb{1}[e(\mathbf{x}) = \xi_i] \log q(\mathbf{y} \mid \mathbf{x}_i) \\
&= 0.5\,\mathbb{E}_{q(\mathbf{x}_2)}\mathbb{E}_{q(\mathbf{y},\mathbf{x}_1 \mid \mathbf{x}_3=1)} \sum_{i\in\{1,2,3\}} \mathbb{1}[e(\mathbf{x}) = \xi_i] \log q(\mathbf{y} \mid \mathbf{x}_i) \\
&\quad + 0.5\,\mathbb{E}_{q(\mathbf{x}_1)}\mathbb{E}_{q(\mathbf{y},\mathbf{x}_2 \mid \mathbf{x}_3=0)} \sum_{i\in\{1,2,3\}} \mathbb{1}[e(\mathbf{x}) = \xi_i] \log q(\mathbf{y} \mid \mathbf{x}_i) \\
&= \frac{1}{2}\Bigg(\mathbb{E}_{q(\mathbf{x}_2)}\Bigg[\mathbb{E}_{q(\mathbf{y},\mathbf{x}_1 \mid \mathbf{x}_3=1)}\,\mathbb{1}[e([\mathbf{x}_1,\mathbf{x}_2,1]) = \xi_1] \log q(\mathbf{y} \mid \mathbf{x}_1) \\
&\qquad\qquad + \mathbb{E}_{q(\mathbf{y},\mathbf{x}_1 \mid \mathbf{x}_3=1)} \sum_{i\in 2,3} \mathbb{1}[e([\mathbf{x}_1,\mathbf{x}_2,1]) = \xi_i] \log q(\mathbf{y} \mid \mathbf{x}_i)\Bigg] \\
&\quad + \mathbb{E}_{q(\mathbf{x}_1)}\Bigg[\mathbb{E}_{q(\mathbf{y},\mathbf{x}_2 \mid \mathbf{x}_3=0)}\,\mathbb{1}[e([\mathbf{x}_1,\mathbf{x}_2,0]) = \xi_2] \log q(\mathbf{y} \mid \mathbf{x}_2) \\
&\qquad\qquad + \mathbb{E}_{q(\mathbf{y},\mathbf{x}_2 \mid \mathbf{x}_3=0)} \sum_{i\in 1,3} \mathbb{1}[e([\mathbf{x}_1,\mathbf{x}_2,0]) = \xi_i] \log q(\mathbf{y} \mid \mathbf{x}_i)\Bigg]\Bigg) \\
&= \frac{1}{2}\mathbb{E}_{q(\mathbf{x}_1,\mathbf{x}_2 \mid \mathbf{x}_3=1)}\Bigg[\mathbb{1}[e([\mathbf{x}_1,\mathbf{x}_2,1]) = \xi_1]\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_1,\mathbf{x}_3=1)} \log q(\mathbf{y} \mid \mathbf{x}_1) \\
&\qquad\qquad + \sum_{i\in 2,3} \mathbb{1}[e([\mathbf{x}_1,\mathbf{x}_2,1]) = \xi_i]\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_1,\mathbf{x}_3=1)} \log q(\mathbf{y} \mid \mathbf{x}_i)\Bigg] \qquad (27) \\
&\quad + \frac{1}{2}\mathbb{E}_{q(\mathbf{x}_1,\mathbf{x}_2 \mid \mathbf{x}_3=0)}\Bigg[\mathbb{1}[e([\mathbf{x}_1,\mathbf{x}_2,0]) = \xi_2]\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_2,\mathbf{x}_3=0)} \log q(\mathbf{y} \mid \mathbf{x}_2) \\
&\qquad\qquad + \sum_{i\in 1,3} \mathbb{1}[e([\mathbf{x}_1,\mathbf{x}_2,0]) = \xi_i]\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_2,\mathbf{x}_3=0)} \log q(\mathbf{y} \mid \mathbf{x}_i)\Bigg]. \qquad (28)
\end{aligned}$$

We will now focus on the three terms within each of eq. (27) and eq. (28). Due to the following equality

$$\begin{aligned}
q(\mathbf{y} = 1 \mid \mathbf{x}_1 = 1, \mathbf{x}_3 = 1) &= q(\mathbf{y} = 0 \mid \mathbf{x}_1 = 0, \mathbf{x}_3 = 1) \\
&= q(\mathbf{y} = 1 \mid \mathbf{x}_2 = 1, \mathbf{x}_3 = 0) = q(\mathbf{y} = 0 \mid \mathbf{x}_2 = 0, \mathbf{x}_3 = 0) = 0.9,
\end{aligned}$$

the expectations in the first terms in each of eq. (27) and eq. (28) are

$$\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_1,\mathbf{x}_3=1)} \log q(\mathbf{y} \mid \mathbf{x}_1) = \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_2,\mathbf{x}_3=1)} \log q(\mathbf{y} \mid \mathbf{x}_2) = 0.9 \log 0.7 + 0.1 \log 0.3 \approx -0.44.$$

Next we turn to setting $i = 1$ term in eq. (27). Due that $q(\mathbf{y} = 1 \mid \mathbf{x}_2 = 1) = q(\mathbf{y} = 0 \mid \mathbf{x}_2 = 0)$,

$$\mathbf{x}_1 = \mathbf{x}_2 \implies \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_1,\mathbf{x}_3=1)} \log q(\mathbf{y} \mid \mathbf{x}_2) = (0.9 \log 0.7 + 0.1 \log 0.3) \approx -0.44,$$

$$\mathbf{x}_1 \neq \mathbf{x}_2 \implies \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_1, \mathbf{x}_3=1)} \log q(\mathbf{y} \mid \mathbf{x}_2) = (0.9 \log 0.3 + 0.1 \log 0.7) \approx -1.12.$$

The same equalities hold for the $i = 2$ term in eq. (28) $\mathbb{E}_{q(\mathbf{y}, \mid \mathbf{x}_2, \mathbf{x}_3=0)} \log q(\mathbf{y} \mid \mathbf{x}_1)$. Finally, regardless of $\mathbf{x}_1, \mathbf{x}_2$, the $i = 3$ terms in both eq. (27) and eq. (28) can be expressed as follows:

$$\mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_1, \mathbf{x}_3=1)} \log q(\mathbf{y} \mid \mathbf{x}_2) \mathbb{E}_{q(\mathbf{y} \mid \mathbf{x}_2, \mathbf{x}_3=0)} \log q(\mathbf{y} \mid \mathbf{x}_1) = (0.9 \log 0.5 + 0.1 \log 0.5) \approx -0.69.$$

Now we can maximize the sum of eq. (27) and eq. (28), over $e(\mathbf{x})$ such that $|e(\mathbf{x})| = 1$.

Notice that setting $\mathbb{1}[e(\mathbf{x}) = \xi_1] = 1$ when $\mathbf{x}_3 = 1$ and $\mathbb{1}[e(\mathbf{x}) = \xi_2] = 1$ when $\mathbf{x}_3 = 0$ achieves the highest score -0.44 in each of eq. (27) and eq. (28). This implies that one optimal reductive explanation is $\xi_1 = [1, 0, 0]$ if $\mathbf{x}_3 = 1$ and $\xi_2 = [0, 1, 0]$ otherwise. This is an encoding explanation as we show below. Due to $\mathbf{E}_{\xi_1} = \mathbf{x}_3$,

$$q(\mathbf{y} = 1 \mid \mathbf{x}_1, \mathbf{E}_{\xi_1} = 1) = 0.9 \neq q(\mathbf{y} = 1 \mid \mathbf{x}_1, \mathbf{E}_{\xi_1} = 0) = 0.5.$$

In turn, $\mathbf{y} \not\perp\!\!\!\perp \mathbf{E}_{\xi_1} \mid \mathbf{x}_{\xi_1}$ for $\{\mathbf{x} : e(\mathbf{x}) = \xi_1\}$ and Def: Encoding holds, meaning that $e(\mathbf{x})$ is encoding.

□

B.8 An example of misestimation of EVAL-X

Consider the following example where

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4],$$
$$\mathbf{x}_2, \mathbf{x}_3 \sim \mathcal{B}(0.5)^{\otimes 2}, \ \mathbf{x}_1, \mathbf{x}_4 \sim \mathcal{N}(0, \mathbf{I}), \quad \mathbf{y} = \mathbf{x}_2 \oplus \mathbf{x}_3.$$

Assume that the misestimated EVAL-X model satisfies these equalities

$$q_{\xi_1}^{\text{misestimated}}(\mathbf{y} = 1 \mid \mathbf{x}_1) = 1 \quad \text{for all } \mathbf{x}_1,$$
$$q_{\xi_4}^{\text{misestimated}}(\mathbf{y} = 0 \mid \mathbf{x}_4) = 1 \quad \text{for all } \mathbf{x}_4.$$

There exists a bad explanation that scores optimally under the misestimated EVAL-X:

$$e(\mathbf{x}) = \begin{cases} \xi_1 & \text{if } \mathbf{x}_2 \oplus \mathbf{x}_3 = 1, \\ \xi_4 & \text{if } \mathbf{x}_2 \oplus \mathbf{x}_3 = 0. \end{cases}$$

Then the EVAL-X score of this explanation under this particular misestimation is

$$\begin{aligned}
&\text{EVAL-X}^{\text{misestimated}}(q, e) = \mathbb{E}_q[\log q_{e(\mathbf{x})}^{\text{misestimated}}(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})})] \\
&= q(\mathbf{y} = 1)\mathbb{E}_q[\log q_{e(\mathbf{x})}^{\text{misestimated}}(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})}) \mid \mathbf{y} = 1] + q(\mathbf{y} = 0)\mathbb{E}_q[\log q_{e(\mathbf{x})}^{\text{misestimated}}(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})}) \mid \mathbf{y} = 0] \\
&= 0.5 \cdot \mathbb{E}_q[\log q_{\xi_1}^{\text{misestimated}}(\mathbf{y} \mid \mathbf{x}_1) \mid \mathbf{y} = 1] + 0.5 \cdot \mathbb{E}_q[\log q_{\xi_4}^{\text{misestimated}}(\mathbf{y} \mid \mathbf{x}_4) \mid \mathbf{y} = 0] \\
&= 0.
\end{aligned}$$

Since $\mathbf{y}$ is deterministic given $\mathbf{x}$ so the maximum value of the EVAL-X score is also 0. So the bad explanation scores optimally due to misestimation. Deterministic $\mathbf{y} \mid \mathbf{x}$ is not necessary for estimation error to affect explanation quality. Here, with this incorrectly estimated EVAL-X, inputs that are pure noise, independent of everything, will be chosen.

B.9 Attention map explanations be encode.

Here, treating each of the coordinates of $\mathbf{x}$ as tokens, we consider a cross-attention based predictive model of the following form: with $\gamma(\mathbf{a})$ as softmax function over a vector $\mathbf{a}$, W as a matrix, $\boldsymbol{\alpha}, \boldsymbol{\beta}$ as vectors, and σ as the sigmoid function, and κ as the temperature, the predictive model $f(\cdot)$ is

$$f(\mathbf{x}) = \sigma\left(\sum_i \boldsymbol{\beta}_i \left[\sum_j \gamma(\kappa \mathbf{x}_i * W\mathbf{x})_j \boldsymbol{\alpha}_j \mathbf{x}_j\right]\right).$$

We then show that using the highest attention score as the explanation produces an encoding explanation. For this example, we consider the following DGP:

$$\begin{aligned}
\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 &\sim \mathcal{B}(0.5)^{\otimes 3}, \\
\mathbf{z}^+ &= [\mathbf{z}_1 + 1, \quad 0, \quad +1], \\
\mathbf{z}^- &= [0 \quad , -\mathbf{z}_2 - 1, \quad -1], \\
\mathbf{x} &= \begin{cases} \mathbf{z}^+ \text{ if } \mathbf{z}_3 = 1, \\ \mathbf{z}^- \text{ if } \mathbf{z}_3 = 0, \end{cases} \\
\mathbf{y} &\sim \mathcal{B}(\rho) \quad \text{where} \quad \rho = \begin{cases} \sigma(\mathbf{x}_1) & \text{if } \mathbf{x}_3 = 1, \\ \sigma(-\mathbf{x}_2) & \text{if } \mathbf{x}_3 = -1. \end{cases}
\end{aligned} \tag{29}$$

The following setting of parameters in $f(\mathbf{x})$ produces a function of $\mathbf{x}$ that converges to ρ as $\kappa \to \infty$:

$$\boldsymbol{\alpha} = [1, -1, 0], \qquad \boldsymbol{\beta} = [1, 1, 0], \qquad W = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{30}$$

By definition of W

$$\begin{aligned}
W\mathbf{x} &= \begin{cases} [0, 0, 1] \text{ if } \mathbf{x}_3 = 1 \\ [0, 0, -1] \text{ if } \mathbf{x}_3 = -1 \end{cases} \\
\implies \mathbf{x}_1 W\mathbf{x} &= \begin{cases} [\mathbf{z}_1 + 1, 0, 0] \text{ if } \mathbf{x}_3 = 1 \\ [0, 0, 0] \text{ if } \mathbf{x}_3 = -1 \end{cases} \implies \gamma(\mathbf{x}_1 W\mathbf{x}) \overset{\kappa\to\infty}{\longrightarrow} \begin{cases} [1, 0, 0] \text{ if } \mathbf{x}_3 = 1 \\ [0, 0, 0] \text{ if } \mathbf{x}_3 = -1 \end{cases} \\
\implies \mathbf{x}_2 W\mathbf{x} &= \begin{cases} [0, 0, 0] \text{ if } \mathbf{x}_3 = 1 \\ [0, \mathbf{z}_2 + 1, 0] \text{ if } \mathbf{x}_3 = -1 \end{cases} \implies \gamma(\mathbf{x}_2 W\mathbf{x}) \overset{\kappa\to\infty}{\longrightarrow} \begin{cases} [0, 0, 0] \text{ if } \mathbf{x}_3 = 1 \\ [0, 1, 0] \text{ if } \mathbf{x}_3 = -1 \end{cases}.
\end{aligned}$$

Then, $\boldsymbol{\beta}_3 = 0$, the inner sum for $i = 3$ does not appear in the function $f(\mathbf{x})$ Then, as $\alpha_3 = 0$, $\alpha_1 = 1, \alpha_2 = -1$,

$$\sum_j \gamma(\kappa \mathbf{x}_1 * W\mathbf{x})_j \boldsymbol{\alpha}_j \mathbf{x}_j \overset{\kappa\to\infty}{\longrightarrow} \begin{cases} \mathbf{x}_1 \text{ if } \mathbf{x}_3 = 1 \\ 0 \text{ if } \mathbf{x}_3 = -1 \end{cases}, \tag{31}$$

$$\sum_j \gamma(\kappa \mathbf{x}_2 * W\mathbf{x})_j \boldsymbol{\alpha}_j \mathbf{x}_j \overset{\kappa\to\infty}{\longrightarrow} \begin{cases} 0 \text{ if } \mathbf{x}_3 = 1 \\ -\mathbf{x}_2 \text{ if } \mathbf{x}_3 = -1 \end{cases}. \tag{32}$$

In turn, as $\beta_1 = \beta_2 = 1$

$$\sum_{i \in 1,2} \beta_i \left[\sum_j \gamma(\kappa \mathbf{x}_i * W\mathbf{x})_j \boldsymbol{\alpha}_j \mathbf{x}_j \right] \overset{\kappa\to\infty}{\longrightarrow} \begin{cases} \mathbf{x}_1 \text{ if } \mathbf{x}_3 = 1 \\ -\mathbf{x}_2 \text{ if } \mathbf{x}_3 = -1 \end{cases},$$

$$f(\mathbf{x}) = \sigma\left(\sum_{i \in 1,2} \beta_i \left[\sum_j \gamma(\kappa \mathbf{x}_i * W\mathbf{x})_j \boldsymbol{\alpha}_j \mathbf{x}_j \right] \right) \overset{\kappa\to\infty}{\longrightarrow} \begin{cases} \sigma(\mathbf{x}_1) \text{ if } \mathbf{x}_3 = 1 \\ \sigma(-\mathbf{x}_2) \text{ if } \mathbf{x}_3 = -1 \end{cases}. \tag{33}$$

So, as $\kappa \to \infty$, the function $f(\mathbf{x})$, with the parameters in eq. (30), converges to $\rho(\mathbf{x})$, meaning that this model will achieve the population log-likelihood optimum under the DGP in eq. (29).

Now, the attention map as an explanation selects $\mathbf{x}_1$ if $\mathbf{x}_3 = 1$ and $\mathbf{x}_2$ otherwise; this comes from eq. (31) and eq. (32). This is an encoding explanation because $\mathbf{E}_{\xi_1} = 1$ if $\mathbf{x}_3 = 1$ which gives

$$q(\mathbf{y} \mid \mathbf{x}_1) \neq q(\mathbf{y} \mid \mathbf{x}_1, \mathbf{x}_3 = 1) \implies \mathbf{y} \not\perp\!\!\!\perp \mathbf{E}_{\xi_1} \mid \mathbf{x}_{\xi_1}.$$

C Experimental details

C.1 Estimating STRIPE-X

To compute the $\mathbf{KL}$ term in ENCODE-METER, we estimate $q(\mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v}, \mathbf{y})$ and $q(\mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v})$. To estimate these, we train a single model — to predict $\mathbf{E}_\mathbf{v}$ from $\mathbf{x}_\mathbf{v}$ and a new variable ℓ that can equal

the label $\mathbf{y}$ or a dummy value `null` that is outside the support of $\mathbf{y}$ — in the following way:

$$\arg\max_{\theta} \mathbb{E}_{\mathbf{x},\mathbf{y}\sim q(\mathbf{x},\mathbf{y})}\Big[\log p_\theta(\mathbf{E}_\mathbf{v} = \mathbb{1}[e(\mathbf{x}) = \mathbf{v}] \mid \mathbf{x}_\mathbf{v}, \ell = \mathbf{y}) + \log p_\theta(\mathbf{E}_\mathbf{v} = \mathbb{1}[e(\mathbf{x}) = \mathbf{v}] \mid \mathbf{x}_\mathbf{v}, \ell = \texttt{null})\Big]. \tag{34}$$

As log-likelihood is a proper scoring rule and $q(\mathbf{y} = \texttt{null}) = 0$, the maximum above is achieved when

$$p_\theta(\mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v}, \ell = \mathbf{y}) = q(\mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v}, \mathbf{y} = \mathbf{y}) \qquad p_\theta(\mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v}, \ell = \texttt{null}) = q(\mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v}).$$

In summary, to estimate ENCODE-METER, solve eq. (34), use its solution to estimate the $\mathbf{KL}$ term from the RHS in eq. (7) for each $\mathbf{x}_\mathbf{v}, \mathbf{y}$, and then average this $\mathbf{KL}$ term over samples of $\mathbf{x}_\mathbf{v}$ from the data such that $e(\mathbf{x}) = \mathbf{v}$ and samples of $\mathbf{y}$ from the EVAL-X model for $q(\mathbf{y} \mid \mathbf{x}_\mathbf{v})$.

In practice, one does not need train a model for each $\mathbf{v}$. We describe how to estimate ENCODE-METER with a single model in Appendix C.2. We give the full STRIPE-X estimation procedure in Algorithm 2 in Appendix D.

C.2 Estimating the encoding cost in STRIPE-X with categorical predictive models

STRIPE-X consists of the EVAL-X score and a cost of encoding measured by ENCODE-METER. Define $\mathcal{V}$ to be the set of possible explanations and let $\mathcal{V}[j]$ denote the jth element of $\mathcal{V}$. The EVAL-X model $p_\gamma(\mathbf{y} \mid \mathbf{x}_\mathbf{v})$ is trained to predict the label $\mathbf{y}$ from subsets $\mathbf{x}_\mathbf{v}$ where $\mathbf{v}$ is uniformly sampled from $\mathcal{V}$ [20]. Next is computing the ENCODE-METER $\phi_q(e)$ that is used in the encoding cost term in STRIPE-X. For each explanation, let $\mathbf{F}$ be the categorical variable (instead of an indicator $\mathbf{E}_\mathbf{v}$) that denotes, for each sample, which inputs were selected by the explanation $e(\mathbf{x})$: $\mathbf{F} = j$ if $\mathbf{E}_{\mathcal{V}[j]} = \mathbb{1}[e(\mathbf{x}) = \mathcal{V}[j]] = 1$. Let $q(j)$ be the distribution over j induced by $q(e(\mathbf{x}))$. We train a model $p_\theta(\mathbf{F} \mid \mathbf{x}_\mathbf{v}, \ell, \mathbf{v})$ with a modification of eq. (34) that averages over $\mathbf{v} \sim q(e(\mathbf{x}))$:

$$\arg\max_{\theta} \mathbb{E}_{\mathbf{v}\sim q(e(\mathbf{x}))}\mathbb{E}_{\mathbf{x},\mathbf{y}\sim q(\mathbf{x},\mathbf{y})} \sum_{\mathcal{V}[j]\in\mathcal{V}} \Big(\mathbb{1}[e(\mathbf{x}) = \mathcal{V}[j]]\big[\log p_\theta(\mathbf{F} = j \mid \mathbf{x}_\mathbf{v}, \ell = \mathbf{y}, \mathbf{v}) + \log p_\theta(\mathbf{F} = j \mid \mathbf{x}_\mathbf{v}, \ell = \texttt{null}, \mathbf{v})\big]\Big). \tag{35}$$

The variable ℓ takes values in $\{-1, 0, 1\}$ where 0 and 1 correspond to $\mathbf{y} = 0$ and $\mathbf{y} = 1$ respectively and -1 corresponds to the `null` value. For a flexible enough model p_θ that achieves the population maximum of eq. (35), for any $\mathbf{v} = \mathcal{V}[j] \in \mathcal{V}$,

$$p_\theta(\mathbf{F} = j \mid \mathbf{x}_\mathbf{v}, \ell = \mathbf{y}, \mathbf{v}) = q(\mathbf{E}_\mathbf{v} = 1 \mid \mathbf{x}_\mathbf{v}, \mathbf{y} = \mathbf{y}),$$
$$p_\theta(\mathbf{F} = j \mid \mathbf{x}_\mathbf{v}, \ell = \texttt{null}, \mathbf{v}) = q(\mathbf{E}_\mathbf{v} = 1 \mid \mathbf{x}_\mathbf{v}).$$

This fact indicates how one can use the model p_θ to estimate ENCODE-METER. First, construct the explanation dataset $D_e = \{(\mathbf{y}, \mathbf{x}_{e(\mathbf{x})})\}$ from D_t. Define q_{D_e} to be the uniform distribution over D_e. Define $\mathcal{E}_{(\mathbf{v},\mathbf{a})}$ as the uniform distribution over K samples of $\mathbf{y}$ from the EVAL-X model:

$$\mathcal{E}_{(\mathbf{v},\mathbf{a})} = \mathbf{U}\left[\{\hat{\mathbf{y}}\}_{k\le K}\right] \quad \{\text{ where } \hat{\mathbf{y}}^k \sim p_\gamma(\mathbf{y} \mid \mathbf{x}_\mathbf{v} = \mathbf{a})\}.$$

Then, estimate ENCODE-METER as follows:

$$\hat{\phi}(q, e) = \mathbb{E}_{(\mathbf{v},\mathbf{a})\sim q_{D_e}(\mathbf{x}_{e(\mathbf{x})})}\mathbb{E}_{\hat{\mathbf{y}}\sim\mathcal{E}_{(\mathbf{v},\mathbf{a})}}\left(p_\theta(\mathbf{F} = j \mid \mathbf{x}_\mathbf{v}, \ell = \hat{\mathbf{y}}, \mathbf{v}) \log \frac{p_\theta(\mathbf{F} = j \mid \mathbf{x}_\mathbf{v}, \ell = \hat{\mathbf{y}}, \mathbf{v})}{p_\theta(\mathbf{F} = j \mid \mathbf{x}_\mathbf{v}, \ell = \texttt{null}, \mathbf{v})} + p_\theta(\mathbf{F} \neq j \mid \mathbf{x}_\mathbf{v}, \ell = \hat{\mathbf{y}}, \mathbf{v}) \log \frac{p_\theta(\mathbf{F} \neq j \mid \mathbf{x}_\mathbf{v}, \ell = \hat{\mathbf{y}}, \mathbf{v})}{p_\theta(\mathbf{F} \neq j \mid \mathbf{x}_\mathbf{v}, \ell = \texttt{null}, \mathbf{v})}\right).$$

C.3 Estimating ENCODE-METER with a generative model

When estimating the STRIPE-X score with procedure above for many different explanations, the maximization in eq. (34) repeated for every explanation, which can be computationally expensive.

Table 6: Position-based, prediction-based, and marginal explanation schemes are all encoding. For samples in the set $\{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\}$ for one of the selections $\mathbf{v}$ that e produces, accuracy < 1 and the $\mathbf{KL}$ being non-zero means these explanations are all encoding per Lemma 1.

Encoding	Acc. $\mathbf{E}_\mathbf{v}$ ($\uparrow$)	$\mathbf{KL}$ ($\downarrow$)
POSI	0.61	0.88
PRED	0.51	0.18
MARG	0.51	0.20

This motivates a second procedure to estimate ENCODE-METER that avoids having to retrain models for each explanation by using generative model for $q(\mathbf{x} \mid \mathbf{x}_\mathbf{v}, \mathbf{y})$. Formally, with $\mathbf{x}_\mathbf{v}$ fixed, the conditional mutual information term in eq. (5) can be computed as the marginal dependence between N samples of $\mathbf{y}$ from $q(\mathbf{y} \mid \mathbf{x}_\mathbf{v})$ and $q(\mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v}, \mathbf{y})$. The model for the former is available from EVAL-X estimation. Simulating from the later, namely $q(\mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v}, \mathbf{y})$, is done by sampling from the generative model $\mathbf{x} \mid \mathbf{x}_\mathbf{v}, \mathbf{y}$ and then computing the indicator $\mathbf{E}_\mathbf{v}$ as $\mathbb{1}[e(\mathbf{x}) = \mathbf{v}]$. Mechanically, with an estimator of mutual information from samples $(\{\mathbf{a}^i\}_{i \le N}, \{\mathbf{b}^i\}_{i \le N})$ denoted $\texttt{MI}(\{\mathbf{a}^i\}, \{\mathbf{b}^i\})$ and with samples $\{\mathbf{a}^i\}$ produced conditionally on values $\mathbf{c}^i$ denoted by a subscript of the conditioned value $\{\mathbf{a}^i\}_{\mathbf{c}^i}$, one can estimate ENCODE-METER as follows: sample $\mathbf{y}^i_{(\mathbf{v},\mathbf{a})} \sim \mathbf{y} \mid \mathbf{x}_\mathbf{v} = \mathbf{a}$ and $\mathbf{x}^i_{\mathbf{v},\mathbf{a},\mathbf{y}^i} \sim q(\mathbf{x} \mid \mathbf{x}_\mathbf{v} = \mathbf{a}, \mathbf{y} = \mathbf{y}^i)$ repeatedly N times and compute

$$\mathbb{E}_{(\mathbf{v},\mathbf{a}) \sim q(\mathbf{x}_{e(\mathbf{x})})} \texttt{MI}\left(\{\mathbf{y}^i\}_{(\mathbf{v},\mathbf{a})}, \{\mathbb{1}[e(\mathbf{x}^i) = \mathbf{v}]\}_{\mathbf{v},\mathbf{a},\mathbf{y}^i}\right).$$

We give the full procedure in Algorithm 1.

C.4 Experimental details from the simulated study.

The data-generating processes from the experiments. Let $\mathcal{N}$ be the standard normal distribution and let $\mathcal{B}(\alpha)$ be the Bernoulli distribution with 1 occurring with probability α. With $\rho = 0.9$, the discrete DGP is:

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5] \sim \mathcal{B}(0.5)^{\otimes 5}, \qquad \mathbf{y} = \begin{cases} \mathbf{x}_1 & \text{w.p. } \rho \quad \text{else} \quad 1 - \mathbf{x}_1 & \text{if } \mathbf{x}_3 = 1, \\ \mathbf{x}_2 & \text{w.p. } \rho \quad \text{else} \quad 1 - \mathbf{x}_2 & \text{if } \mathbf{x}_3 = 0. \end{cases} \tag{36}$$

The hybrid DGP is as follows: with $\gamma = 5$ and $\sigma(x) = \frac{1}{1+\exp(-x)}$ as the sigmoid function

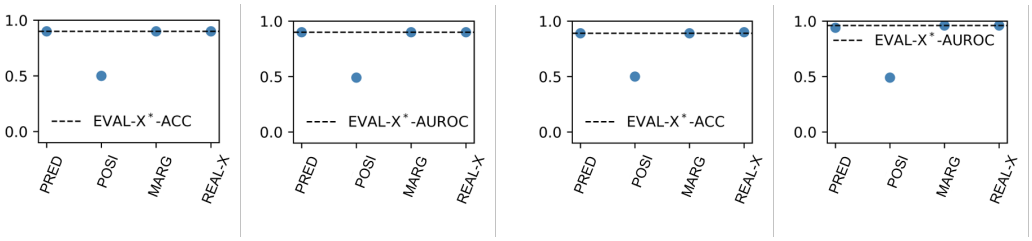
$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5] \sim \mathcal{N}(0.5)^{\otimes 4}, \mathbf{x}_3 \sim \mathcal{B}(0.5), \qquad \rho = \begin{cases} \sigma(\gamma\, \mathbf{x}_1) & \text{if } \mathbf{x}_3 = 1, \\ \sigma(\gamma\, \mathbf{x}_2) & \text{if } \mathbf{x}_3 = 0, \end{cases} \qquad \mathbf{y} \sim \mathcal{B}(\rho). \tag{37}$$

Computing accuracy and KL to show that POSI, PRED, MARG are encoding. For each encoding type, we build two decision trees from 1000 samples from eq. (36): the first decision tree learns $q(\mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v})$ and the second learns $q(\mathbf{y} \mid \mathbf{x}_\mathbf{v}, \mathbf{E}_\mathbf{v} = b)$ for $b \in \{0, 1\}$. We set the maximum depth to be 6. Trees of this depth learn any function of 6 binary digits; $\mathbf{x}$ with $\mathbf{E}_\mathbf{v}$ as an additional column amounts to 6 binary digits. These decision trees are used to compute the accuracy of predicting $\mathbf{E}_\mathbf{v}$ with $q(\mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v})$ and the $\mathbf{KL}$ between $q(\mathbf{y} \mid \mathbf{x}_\mathbf{v}, \mathbf{E}_\mathbf{v} = 1)$ and $q(\mathbf{y} \mid \mathbf{x}_\mathbf{v}, \mathbf{E}_\mathbf{v} = 0)$. Within a set $\{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\}$ that is all $\mathbf{x}$ that have one of the possible selections $\mathbf{v}$, Table 6 report the accuracy of predicting $\mathbf{E}_\mathbf{v}$ with $q(\mathbf{E}_\mathbf{v} \mid \mathbf{x}_\mathbf{v})$ and the $\mathbf{KL}$ between $q(\mathbf{y} \mid \mathbf{x}_\mathbf{v}, \mathbf{E}_\mathbf{v} = 1)$ and $q(\mathbf{y} \mid \mathbf{x}_\mathbf{v}, \mathbf{E}_\mathbf{v} = 0)$, averaged only over samples in $\{\mathbf{x} : e(\mathbf{x}) = \mathbf{v}\}$.

EVAL-X. To estimate EVAL-X for the DGPs in eq. (36) and eq. (37), we compute conditionals $q(\mathbf{y} = 1 \mid \mathbf{x}_\mathbf{v})$ via Monte Carlo approximation. Due to the different coordinates of $\mathbf{x}$ being independent, one can compute $q(\mathbf{y} = 1 \mid \mathbf{x}_\mathbf{v})$ as a marginal expectation over the inputs except those in $\mathbf{v}$:

$$q(\mathbf{y} \mid \mathbf{x}_\mathbf{v}) = \mathbb{E}_{q(\mathbf{x}^c_\mathbf{v} \mid \mathbf{x}_\mathbf{v})} q(\mathbf{y} \mid \mathbf{x}_\mathbf{v}, \mathbf{x}^c_\mathbf{v}) = \mathbb{E}_{q(\mathbf{x}^c_\mathbf{v})} q(\mathbf{y} \mid \mathbf{x}).$$

We Monte Carlo estimate the RHS of this equation over 500 resamples of $\mathbf{x}^c_\mathbf{v}$. We take 5000 samples from each DGP to estimate EVAL-X scores with respect to $q(\mathbf{y}, \mathbf{x})$. In Appendix C.5 we also show experiment results where we use the EVAL-X accuracy and AUROC as the score instead.



(a) Results: Discrete DGP. **(b)** Results: hybrid DGP.

Figure 6: The EVAL-X-ACC and AUROC scores for the different explanations for the discrete DGP are on the left and the scores for the hybrid DGP are in the right. In both, multiple encoding explanations (PRED, MARG, and the reductive one from REAL-X) all achieve the same score as the corresponding EVAL-X* score. Thus, ranking metrics like accuracy and AUROC are not sensitive to encoding explanations like EVAL-X log-likelihoods, and can fail to even weakly detect encoding. This stems from the fact that accuracy and AUROC only depend on the ranking of the datapoint, and therefore are not sensitive to differences in log probabilities that do not change ranks.

REAL-X. We solve REAL-X for any specified explanation size K as follows. In the case of the discrete DGP, for each possible value of $x \in \text{supp}(q(\mathbf{x}))$ (of which there are finitely many), we make $e(\mathbf{x})$ output the subset of at most size K that achieves that maximum averaged log-likelihood over the samples that equal said value $\mathbf{x} = x$. This produces the optimally-scoring explanation $e(\mathbf{x})$ that maps each finite value in the support of $q(\mathbf{x})$ to one subset of the coordinates of $\mathbf{x}$. To do the same in the continuous DGP in eq. (37), we round $\mathbf{x}$ to integers and then use the same type of optimization as in the discrete case.

STRIPE-X. In estimating ENCODE-METER, the model $p_\theta(\mathbf{E_v} \mid \mathbf{x_v}, \ell)$ is a decision tree of depth at most 5, which is then used to estimate averaged $\mathbf{KL}$ in the RHS of eq. (7) with a single sample from $\mathbf{y} \mid \mathbf{x_v}$. The process is repeated for each $\mathbf{v}$ and averaged to produce the ENCODE-METER. The simulated experiments were done on a CPU with the whole runtime around 10 minutes.

C.5 Experiments with EVAL-X accuracy and EVAL-X AUROC

Here, instead of EVAL-X log-likelihoods, we use the accuracy and AUROC of the EVAL-X model as the score. We call these EVAL-X-ACC and EVAL-X-AUROC scores. These metrics only depend on the ranking of the datapoints, and therefore are not sensitive to differences in log probabilities that do not change ranks. Figure 6 shows that, due to this insensitivity, multiple encoding explanations (PRED, MARG, and the excessively reductive one) all achieve the same score as the corresponding EVAL-X* score. In summary, ranking metrics like accuracy and AUROC are not sensitive to encoding explanations like EVAL-X log-likelihoods.

C.6 Classifying dogs and cats.

The POSI explanation selects the upper or the lower color patch depending on whether $q(\mathbf{y} = 1 \mid \mathbf{x}) > 0.5$ or not. The PRED explanation selects the patch predicting from which best matches the prediction from $q(\mathbf{y} = 1 \mid \mathbf{x})$. The MARG explanation selects the top or the bottom image patch based on the color as in the DGP in Figure 5. We consider two non-encoding explanations. The first explanation, denoted `optimal`, selects exactly the features that occur in the DGP: $\{\mathbf{x}_1, \mathbf{x}_2\}$ if the color patch $\mathbf{x}_1$ is `blue` and $\{\mathbf{x}_1, \mathbf{x}_4\}$ otherwise. As $\mathbf{y}$ is determined by the explanation, meaning $\mathbf{y} \perp\!\!\!\perp \mathbf{E_v} \mid \mathbf{x_v}$ for all $\mathbf{v}$ and values $\mathbf{x_v}$, this explanation is non-encoding. The second one, denoted `fixed`, always outputs the bottom right patch $\mathbf{x}_4$; this explanation is constant which violates the first criterion in Lemma 1 meaning there is no encoding.

We also run the REAL-X method from [20] to produce an explanation. REAL-X was run over explanations that select one of the four quarter patches and exact marginalization over the selections.

The base cat and dog images were obtained from the `cats_vs_dogs` dataset from the Tensorflow datasets package. To construct images like in Figure 5, the color and the two images are sampled independently. The color being blue/red determines that the label associated with the top/bottom image becomes the label for the constructed image. The training, validation, and test dataset consist of 8000, 1000, and 1000 samples respectively.

We follow the procedure in Appendix C.2 to estimate STRIPE-X. The EVAL-X model is a pre-trained 18-layer residual network. The model $p_\theta(\mathbf{F} \mid \mathbf{x_v}, \ell, \mathbf{v})$ used in computing the ENCODE-METER term in STRIPE-X (eq. (6)) are 34-layer Residual neural networks. The EVAL-X model is trained for 100 epochs with a batch size of 100 with the Adam optimizer, with the learning rate and weight decay parameters set to 10^{-3} and 0 respectively. The $p_\theta(\mathbf{F} \mid \mathbf{x_v}, \ell, \mathbf{v})$ model is trained for 50 epochs with a batch size of 200 with the Adam optimizer, with the learning rate and weight decay parameters set to 5×10^{-5} and 1 respectively. The p_θ model sees variable ℓ through an entire extra channel where all the pixels take the value ℓ. We used validation loss as the metric to early stop. The EVAL-X and STRIPE-X scores are computed on the test dataset. The cats vs. dogs experiment were done on an A100 GPU where the whole training and evaluation ran in less than 20 minutes.

Remark on the gap between FRESH and EVAL-X scores for the optimal explanation. As the optimal explanation selects features sufficient to produce the label, meaning $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x_v}$ or $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \texttt{val}(\mathbf{x}_{e(\mathbf{x})})$, FRESH and EVAL-X log-likelihoods should be the same as predicting from the full feature set. One potential reason there is a gap between the two scores in table 3 is that the FRESH and EVAL-X scores are computed with ResNet18 models that solve prediction problems of different levels of difficulty On one hand, FRESH is computed with a model trained for a single prediction task: predict $\mathbf{y}$ from $\texttt{val}(\mathbf{x}_{e(\mathbf{x})})$. On the other hand, EVAL-X is computed with a model trained for a more complicated task: for a range possible $\mathbf{v}$, predict $\mathbf{y}$ from $\mathbf{x_v}$. Using large models with appropriate regularization, such as weight decay, should mitigate the gap in scores.

C.7 LLM experiment details

We generate $10,000$ reviews of the following type: with `ADJ1` and `ADJ2` as adjectives, the review is

- `'My day was <ADJ1> and the movie was <ADJ2>. that is it'` or
- `'My day was <ADJ1> and the movie was <ADJ2>. oh wait, reverse the adjectives'.`

The second sentence in the review acts as a "control flow" input and determines whether `ADJ1` or `ADJ2` describes the sentiment about the movie. We prompted Llama 3 to predict the sentiment and select words relevant to predicting the sentiment. In Appendix C.8, we give the prompts we used to make Llama 3 produce explanations from. For this problem, the inputs $\mathbf{x}$ are the reviews and Llama 3 produces explanations $e(\mathbf{x})$ that select a subset of words in the review. The summaries and explanations were generated for all $10,000$ samples but to estimate STRIPE-X, we only used data from the 5 most common explanations (we restricted to inputs whose explanations $\mathbf{v}$ had high $q(e(\mathbf{x}) = \mathbf{v})$). This resulted in a dataset of size 8136, which we split into a training, validation, and test datasets of sizes $6102, 1017$, and 1017 respectively.

Both the EVAL-X model and the model for $p_\theta(\mathbf{F} \mid \mathbf{x_v}, \mathbf{v}, \ell)$ (see Appendix C.2) used in estimating the ENCODE-METER term in STRIPE-X were finetuned GPT-2 models. For the EVAL-X model, we used the AdamW optimizer with a batch size of 100 and trained for 50 epochs with the learning rate set to $5e-5$, weight decay set to 0, and a Cosine learning rate scheduler with the number of cycles set to 1. For the p_θ model used in estimating ENCODE-METER, we used the AdamW optimizer with a batch size of 50 and trained for 25 epochs with the learning rate set to $5e-5$, weight decay set to 0, and a Cosine learning rate scheduler with number of cycles set to 1. The p_θ model sees variable ℓ through the following word added to the input sequence of words: `positive` if $\ell = \mathbf{y} = 1$, `negative` if $\ell = \mathbf{y} = 0$, and nothing if $\ell = \texttt{null}$. We used validation loss to early stop. We follow the procedure in Appendix C.2 to compute ENCODE-METER with $p_\theta(\mathbf{F} \mid \mathbf{x_v}, \mathbf{v}, \ell)$ on the test data with the averaging over $\mathbf{y} \mid \mathbf{x_v}$ estimated using a 5 samples per value of $\mathbf{x_v}$. All training and inference for this experiment was done on an A100. The explanation step and the estimation for both parts of STRIPE-X together took under 2 hours. The LLM-generated explanations achieves an EVAL-X score of -0.497 and an ENCODE-METER value was 0.114.

C.8 Prompts used to predict sentiment and produce explanation

In Figure 7, we provide the prompt we used to predict the sentiment from a review and generate an explanation for that prediction.

Figure 7: Llama 3 prompt used to predict sentiment and generate an explanation for that prediction.

```
System: You are a helpful and honest assistant. Please, respond concisely and
    truthfully.

You are asked to summarize movie reviews of the form "first sentence. second
    sentence".
The following are examples along with the reasoning.


Consider ‘My day was moving and the movie was overblown. that is it.’
The second sentence means the second adjective ’overblown’ describes the movie.
Due to this description, the sentiment is negative.


Consider ‘My day was moving and the movie was overblown. oh wait, reverse the
    adjectives.’
The second sentence means the first adjective ’moving’ describes the movie.
Due to this description, the sentiment is positive.


These are all examples.

user: What is the sentiment about the movie in this review ‘<REVIEW>‘?

Think step-by-step about this latest review. If the second sentence instructs it,
    switch the adjectives and then based on the new descriptor of the movie, answer
     either ’positive’ or ’negative’.

Explain why you chose those this sentiment by selecting as few words as possible
    from the review. Include all the words that you looked at.

Use this helpful format: "the sentiment is <positive/negative> and the explanation
    is <words, ...>. END. "
```

D Algorithms

Algorithm 1 describes an alternate way to estimate the ENCODE-METER component of STRIPE-X with a conditional generative model. Algorithm 2 describes the predictive version of STRIPE-X estimation, which we used in our experiments.

Algorithm 1: ENCODE-METER, generative version.

Input: Training data $D \sim q(\mathbf{y}, \mathbf{x})$ and test data $D_t \sim q(\mathbf{y}, \mathbf{x})$, explanation function $e(\mathbf{x})$, penalty weight λ. EVAL-X model $p_\gamma(\mathbf{y} \mid \mathbf{x_v})$. Conditional generative model $p_\theta(\mathbf{x} \mid \mathbf{x_v}, \mathbf{y})$ and mutual information estimator that takes two sets as arguments $\texttt{MI}[\{c_i\}, \{d_i\}]$;

Result: Return estimate of STRIPE-X :

Define $\mathbf{J}_{(\mathbf{v},\mathbf{a})}$ as the set of K random samples of $\mathbf{y}$ from the EVAL-X model:

$$\mathbf{J}_{(\mathbf{v},\mathbf{a})} = \{\hat{\mathbf{y}}^k\}_{k \le K} \quad \{ \text{ where } \hat{\mathbf{y}}^k \sim p_\gamma(\mathbf{y} \mid \mathbf{x_v} = \mathbf{a})\}$$

Define $\mathbf{L}_{(\mathbf{v},\mathbf{a},\mathbf{J}_{(\mathbf{v},\mathbf{a})})}$ as the set of K random samples of $\mathbf{x}$ from p_θ conditioned on $\mathbf{a}$ and $\hat{\mathbf{y}}$:

$$\mathbf{L}_{(\mathbf{v},\mathbf{a},\mathbf{J}_{(\mathbf{v},\mathbf{a})})} = \{\mathbb{1}[e(\hat{\mathbf{x}}^k) = \mathbf{v}]\}_{k \le K} \quad \{ \text{ where } \hat{\mathbf{x}}^k \sim p_\theta(\mathbf{x} \mid \mathbf{x_v} = \mathbf{a}, \mathbf{y} = \hat{\mathbf{y}}^k)\}$$

Construct the explanation dataset $D_e = \{(\mathbf{x}_{e(\mathbf{x})} = (e(\mathbf{x}), \mathbf{a}))\}$ from D_t.

Define q_{D_e} to be the uniform distribution over D_e.

Compute the following averaging of estimated mutual information between, $\mathbf{J}, \mathbf{L}$

$$\hat{\phi}_q(e) = \mathbb{E}_{(\mathbf{v},\mathbf{a}) \sim q_{D_e}(\mathbf{x}_{e(\mathbf{x})})} \texttt{MI}\left[\mathbf{J}_{(\mathbf{v},\mathbf{a})}, \mathbf{L}_{(\mathbf{v},\mathbf{a},\mathbf{J}_{(\mathbf{v},\mathbf{a})})}\right]$$

Return $\hat{\phi}(q, e)$ as the ENCODE-METER estimate.

Algorithm 2: STRIPE-X, predictive version.

Input: Training data $D \sim q(\mathbf{y}, \mathbf{x})$ and test data $D_t \sim q(\mathbf{y}, \mathbf{x})$, explanation function $e(\mathbf{x})$, penalty weight λ. Specifications for the models $p_\gamma(\mathbf{y} \mid \mathbf{x_v})$ and $p_\theta(\mathbf{F} \mid \mathbf{x_v}, \ell)$.

Result: Return estimate of STRIPE-X :

Define q_D to be the uniform distribution over D

Construct the explanation dataset $D_e = \{(\mathbf{y}, \mathbf{x}_{e(\mathbf{x})})\}$ from D_t

Define q_{D_e} to be the uniform distribution over D_e.

Estimate `EVAL-X()`

Solve the following minimization problem to learn $p_\gamma(\mathbf{y} \mid \mathbf{x_v})$

$$\arg\max_\gamma \mathbb{E}_{\mathbf{v} \sim q_D(e(\mathbf{x}))} \mathbb{E}_{\mathbf{x},\mathbf{y} \sim q_D(\mathbf{x},\mathbf{y})} \big[\log p_\gamma(\mathbf{y} \mid \mathbf{x_v}) \big]$$

Output: p_γ

Estimate `ENCODE-METER()`

Construct the set of possible selections $\mathcal{V} = \{\mathbf{v} : q(e(\mathbf{x}) = \mathbf{v}) > 0\}$

Construct data of the form $(\mathbf{x}, \mathbf{F})$ where $\mathbf{F} = j$ if $\mathbf{E}_{\mathcal{V}[j]} = 1$.

Fit the model $p_\theta(\mathbf{F} \mid \mathbf{x_v}, \ell)$ via the following log-likelihood maximization:

$$\arg\max_\theta \mathbb{E}_{\mathbf{v} \sim q_D(e(\mathbf{x}))} \mathbb{E}_{\mathbf{x},\mathbf{y} \sim q_D(\mathbf{x},\mathbf{y})} \sum_{\mathcal{V}[j] \in \mathcal{V}} \Bigg(\mathbb{1}[e(\mathbf{x}) = \mathcal{V}[j]] \big[\log p_\theta(\mathbf{F} = j \mid \mathbf{x_v}, \ell = \mathbf{y}, \mathbf{v}) + \log p_\theta(\mathbf{F} = j \mid \mathbf{x_v}, \ell = \texttt{null}, \mathbf{v}) \big] \Bigg)$$

Define $\mathcal{E}_{(\mathbf{v},\mathbf{a})}$ as the uniform distribution over K samples of $\mathbf{y}$ from the EVAL-X model:

$$\mathcal{E}_{(\mathbf{v},\mathbf{a})} = \mathbf{U}\left[\{\hat{\mathbf{y}}\}_{k \le K}\right] \quad \{ \text{ where } \hat{\mathbf{y}}^k \sim p_\gamma(\mathbf{y} \mid \mathbf{x_v} = \mathbf{a}) \}$$

Output: The following nested expectation over q_{D_e} and $\mathcal{E}(\cdot)$:

$$\mathbb{E}_{(\mathbf{v},\mathbf{a}) \sim q_{D_e}(\mathbf{x}_{e(\mathbf{x})})} \mathbb{E}_{\hat{\mathbf{y}} \sim \mathcal{E}_{(\mathbf{v},\mathbf{a})}} \Bigg(p_\theta(\mathbf{F} = j \mid \mathbf{x_v}, \ell = \hat{\mathbf{y}}, \mathbf{v}) \log \frac{p_\theta(\mathbf{F} = j \mid \mathbf{x_v}, \ell = \hat{\mathbf{y}}, \mathbf{v})}{p_\theta(\mathbf{F} = j \mid \mathbf{x_v}, \ell = \texttt{null}, \mathbf{v})} + p_\theta(\mathbf{F} \neq j \mid \mathbf{x_v}, \ell = \hat{\mathbf{y}}, \mathbf{v}) \log \frac{p_\theta(\mathbf{F} \neq j \mid \mathbf{x_v}, \ell = \hat{\mathbf{y}}, \mathbf{v})}{p_\theta(\mathbf{F} \neq j \mid \mathbf{x_v}, \ell = \texttt{null}, \mathbf{v})} \Bigg)$$

Learn the EVAL-X model $p_\gamma \leftarrow$`EVAL-X()`.

Estimate ENCODE-METER as the $\hat{\phi}_q(e) \leftarrow$`ENCODE-METER()`.

Return the following as the STRIPE-X estimate:

$$\mathbb{E}_{(\mathbf{v},\mathbf{a}) \sim q_{D_e}(\mathbf{x}_{e(\mathbf{x})})} \mathbb{E}_{\mathbf{y} \sim q_{D_e}(\mathbf{y} \mid \mathbf{x}_{e(\mathbf{x})} = (\mathbf{v},\mathbf{a}))} \left[\log p_\gamma \left(\mathbf{y} = \mathbf{y} \mid \mathbf{x_v} = \mathbf{a} \right) \right] - \lambda \hat{\phi}_q(e)$$