Let Me Explain, Again: Multiplicity in Local Sufficient Explanations

Anonymous authors
Paper under double-blind review

Abstract

When asked to explain their decisions, humans can produce multiple complementary justifications. In contrast, several feature attribution methods for machine learning produce only one such attribution, despite the existence of multiple equally strong and succinct explanations. The explanations found by these methods thus offer an incomplete picture of model behavior. In this paper, we study the problem of explaining a machine learning model's prediction on a given input from the perspective of minimal feature subsets that are sufficient for the model's prediction, focusing on their non-uniqueness. We give a tour of perspectives on this non-uniqueness, in terms of Boolean logic, conditional independence, approximate sufficiency, and degenerate conditional feature distributions. To cope with the multiplicity of these explanations, we propose a wrapper methodology that can adapt and extend methods that find a single explanation into methods for finding multiple explanations of similar quality. Our experiments benchmark the proposed meta-algorithm, which we call Let Me Explain Again (LMEA), against two multi-explanation method baselines on synthetic and real-world multiple-instance learning problems for image classification and demonstrate the ability of LMEA to augment two single-explanation methods.

1 Introduction

Predictive machine learning (ML) models are increasingly used in high-stakes decision-making contexts such as healthcare (Shailaja et al., 2018), employment (Freire & de Castro, 2021), credit scoring (Thomas et al., 2017), and criminal justice (Rudin et al., 2020). As the consequences of their use grow, so does the importance of understanding the decision processes behind model predictions. However, a full understanding of complex models, such as deep neural networks, remains elusive: these models have been, and continue to be, widely regarded as "black boxes" (Alain & Bengio, 2016, p. 1).

These concerns have spurred the development of local feature attribution methods, which aim to explain the prediction of an ML model on a specific input by identifying the most "important" features. Among the many diverse methods, an increasingly popular family of approaches focuses on identifying a minimal subset of features that, once fixed, are sufficient to determine the prediction (Shih et al., 2018; Ignatiev et al., 2020; Darwiche & Hirth, 2020; Ribeiro et al., 2018; Wang et al., 2021; Amoukou & Brunel, 2022; Fong & Vedaldi, 2017; Fong et al., 2019; Luss & Dhurandhar, 2024). Formally, these methods aim to explain the prediction of a model $f : \mathcal{X} \to \mathcal{Y}$ on a fixed input $\mathbf{x} \in \mathcal{X} \subsetneq \mathbb{R}^d$ by identifying a minimal subset $\mathcal{S} \subseteq [d]$ such that fixing $\mathbf{x}_\mathcal{S}$ suffices to determine the model output $f(\mathbf{x})$ either exactly or approximately, and either deterministically or with high probability.

In this work, we study this class of explanations, which we refer to as *minimal sufficient subsets* (MSS). Although MSSs have been formally defined in various ways and their properties have been studied to different extents, a key aspect that has received little attention is their *multiplicity*. Empirically, MSSs have been shown to be not unique (Carter et al., 2019; Camburu et al., 2020), but the underlying theoretical reasons for this phenomenon remain largely unexplored. Even more concerning is that, despite this fact, several

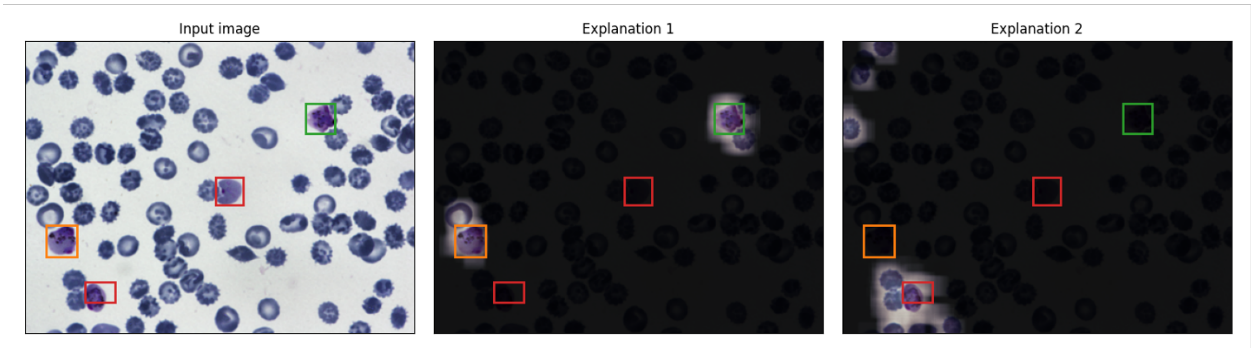


Figure 1: Multiple MSSs are common in the multiple instance learning (MIL) (Dietterich et al., 1997) setting. Pictured are example explanations for a classifier trained to predict the presence of certain malaria-infected cells (trophozoites) in images of blood smears from the BBBC041 dataset (Ljosa et al., 2012). The green box depicts the location of a trophozoite, while the red boxes depict the location of other kinds of infected cells. The orange box is labeled as "difficult," meaning it could not be classified by a human expert. *Left:* input image to be explained, for which the classifier predicts the presence of a trophozoite. *Middle:* an explanation recovered by a method for finding a single MSS. *Right:* an explanation found by running the same single-MSS method again using our proposed wrapper algorithm. Crucially, although both explanations are sufficient, the single-MSS approach misses the model's dependence on non-trophozoite cells, which our method reveals.

methods for finding MSSs (Fong & Vedaldi, 2017; Fong et al., 2019; MacDonald et al., 2019; Brinner & Zarrieß, 2023; Luss & Dhurandhar, 2024) return only one such subset.[1]

To understand why identifying only a single MSS can be problematic, consider a common downstream application of these methods: model debugging. *Post-hoc* explanations can be used to identify model dependence on semantically meaningless or irrelevant features (Carter et al., 2021; Wäldchen, 2022). In this context, methods that provide only one MSS will not necessarily identify the missing spurious features that influence a model's prediction (Chockler et al., 2023). This situation is depicted in Figure 1, which involves explaining a classifier for detecting infected cells (trophozoites) in blood smear images from the BBBC041 dataset (Ljosa et al., 2012). Seeing just the first explanation might lead the user to conclude that the non-trophozoite cells do not influence the classifier's prediction when, in fact, they do.

The multiplicity of MSSs raises important questions. Are there settings when MSSs are unique, or is this multiplicity inevitable (and when)? Furthermore, when multiple MSSs do exist, can we properly identify them all? In this work, we provide answers to these questions. To address the multiplicity of these explanations, we propose a meta-algorithm for extending methods that find a single (or a few) MSSs into methods capable of systematically identifying a representative subset of all MSSs. In the ideal case where we have access to an oracle (that returns a single MSS) this method provably recovers a subset of MSSs that intersects every MSS. In turn, our experiments demonstrate how augmenting existing methods that identify a single MSSs leads to more comprehensive explanations in real and synthetic computer vision settings.

1.1 Related work

MSSs have been formalized in diverse ways, drawing on ideas from probability (Amoukou & Brunel, 2022; Wang et al., 2021; Bharti et al., 2025), causality (Chockler et al., 2023), and logic (Ignatiev et al., 2019; 2020; Marques-Silva et al., 2020; Izza et al., 2020; Darwiche & Ji, 2022; Darwiche & Hirth, 2020). Across these formulations, it is often acknowledged that MSSs are generally non-unique. However, aside from works in logic-based explainability, few address this fact as a central concern. Among those that do, approaches can

[1] A notable exception to these statements is the growing set of methods based on logic (Ignatiev & Marques-Silva, 2021; Huang et al., 2021; Izza et al., 2020), for which the reasons behind MSS multiplicity are well understood. Enumeration of MSSs is possible for these methods in certain scenarios (Ignatiev et al., 2020); however, the problems these methods solve are often computationally hard, and even the most recent methods scale poorly to neural networks with high-dimensional inputs (Marques-Silva, 2023; Bassan et al., 2025).

be categorized along several axes; namely, whether they (a) assume black-box (Carter et al., 2019; Chockler et al., 2023) or gradient (Carter et al., 2021; Byra & Skibbe, 2025) access to the model, (b) identify disjoint (Carter et al., 2019; 2021) or overlapping (Shitole et al., 2021; Chockler et al., 2023; Byra & Skibbe, 2025) sets of MSSs and (c) possess (Carter et al., 2019) or lack (Chockler et al., 2023; Shitole et al., 2021; Amoukou & Brunel, 2022) theoretical guarantees on the subset of MSSs found.

Works related to ours are those by Shitole et al. (2021), Chockler et al. (2023), Amoukou & Brunel (2022), MacDonald et al. (2019), Fong & Vedaldi (2017), and Fong et al. (2019). Shitole et al. (2021) find overlapping sets of MSSs via forward beam search on a 7×7 image grid, and limit the overlap of the final MSS set by doing a greedy backward elimination on a submodular set intersection objective. Chockler et al. (2023) propose MultiReX, a method to identify multiple MSSs using principles from actual causality (Halpern, 2019). While empirically effective, the approach lacks theoretical guarantees about the subset of MSSs it recovers. Amoukou & Brunel (2022) propose a nonparametric algorithm for finding MSSs via a Random Forest (Breiman, 2001) and propose combining multiple explanations into a single attribution by exploring a subset of all MSSs. MacDonald et al. (2019) propose a method for finding minimal subsets of features that account for the prediction in a rate-distortion theoretic framework. Loosely speaking, their method finds a small set of features that approximately preserves the prediction when replacing the complement with Gaussian noise. Similarly, Fong & Vedaldi (2017); Fong et al. (2019) propose methods to find small, contiguous sets of features that maintain the prediction when the complement is replaced by background (e.g., zero or blurred input) values. Both (MacDonald et al., 2019) and (Fong et al., 2019) find a *single* MSS; we will show how each of these methods can be extended to find *multiple* MSSs in our experiments.

The works most relevant to ours are those of Bharti et al. (2025), Carter et al. (2019; 2021), and Byra & Skibbe (2025). Bharti et al. (2025) propose definitions for sufficiency and necessity of ML model explanations in the context of feature removal (Covert et al., 2021), connect these definitions to conditional independence (CI) and Shapley value explanations (Lundberg & Lee, 2017), and demonstrate via theory and experiments how augmenting the trade off between necessity and sufficiency leads to equally informative, yet different, explanations. In this work, we adopt a variation on their sufficiency definition. Carter et al. (2019), propose a backward-forward greedy algorithm that sequentially identifies multiple MSSs and demonstrate its effectiveness on MNIST digit classification, sentiment analysis, and genomics tasks. To improve the scalability of this algorithm, the authors extend their method to image classifiers in (Carter et al., 2021) by incorporating input gradients to guide the backward elimination, and show its application to discovering models' usage of spurious feature patterns (e.g., background pixels) in their predictions. Byra & Skibbe (2025) explore adapting (Fong et al., 2019) to recover multiple sufficient explanations using implicit neural representations. After finding an initial explanation, they solve the MSS problem again, preventing re-selection of features via a penalty term in their optimization objective. This approach involves reimplementation of the XP solver and determination of a penalty strength for which a suitable value is not obvious *a priori*.

While many works have noted the existence of multiple MSSs, several theoretical and qualitative reasons for the non-uniqueness of MSSs have not been made explicit. Furthermore, most of these works study the problem from the perspective of logic, and other perspectives on MSS non-uniqueness are comparatively under-explored. For example, while Bharti et al. (2025) study MSSs in a probabilistic framework, they do not address the reasons behind the non-uniqueness of MSS explanations in this probabilistic context. On the algorithmic side, most papers on the topic outside of the logic community propose specialized techniques to find multiple MSSs (e.g., (Carter et al., 2019; 2021; Shitole et al., 2021; Chockler et al., 2023)), and it is still unclear to what extent methods for finding a single MSS might be extended to recover multiple MSSs. While Byra & Skibbe (2025) take a first step toward answering this question, they only extend a single method.

1.2 Contributions

Our work addresses each of the aforementioned gaps by providing both a unified theoretical discussion of the MSS non-uniqueness problem and proposing a general framework for extending existing MSS-based techniques that find a single explanation. Concretely, our contributions are the following:

1. We present several mathematical and qualitative perspectives on multiple MSSs, including interpretations in terms of logic, symmetry, CI relationships, non-linearity, and approximation. Along

the way, we weave in intuition-building examples and unify logical and probabilistic viewpoints on the multi-MSS phenomenon. To the best of our knowledge, such a comprehensive treatment of the reasons behind the non-uniqueness of MSSs is lacking in the literature.

2. We propose a meta-algorithm, Let Me Explain Again (LMEA), that generalizes existing strategies for finding multiple MSSs (Carter et al., 2019; 2021; Byra & Skibbe, 2025) and can adapt explanation methods for finding a single MSS into methods for finding multiple. Given access to an oracle that finds a single MSS, we show that LMEA recovers a disjoint subset of MSSs whose union intersects every MSS.

3. We show how LMEA can be used to enhance the ability of two gradient-based methods—rate distortion explanations (RDE) (MacDonald et al., 2019) and extremal perturbations (XP) (Fong et al., 2019)—to retrieve MSSs. Our proposed method provides easy-to-set hyperparameters and allows the extension of gradient-based MSS explanation methods, like RDE and XP, with minimal to no modification of off-the-shelf implementations, enhancing the applicability of LMEA.

4. We showcase LMEA on a series of multiple instance learning (MIL) (Dietterich et al., 1997) image classification problems, comparing LMEA against two baselines for finding multiple MSSs, MultiReX and SIS. Our results demonstrate the ability of LMEA to extend single-MSS explanation methods into multiple-MSS explanation methods. Furthermore, our results show that extending RDE and XP with LMEA retrieves explanations that overlap human-annotated salient image regions at least as well as MultiReX and SIS.

2 Setting and background

Notation. Let $f : \mathcal{X} \to \mathcal{Y}$ be a model to be explained, where $\mathcal{X} \subseteq \mathbb{R}^d$ denotes the input space, and $\mathcal{Y}$ denotes the output space, which depends on the task. For instance, $\mathcal{Y} \subseteq \mathbb{R}^k$ in the regression setting, while $\mathcal{Y} = \mathcal{P}(\mathcal{L})$ or $\mathcal{Y} = \mathcal{L}$ in the classification setting, where $\mathcal{L}$ is a set of labels and $\mathcal{P}(\mathcal{L})$ is the set of probability distributions over $\mathcal{L}$. Sets will be written in calligraphic script, e.g., $\mathcal{S}$, and set complements will be denoted with a superscript c, e.g., $\mathcal{S}^c$. We will write $[d] \doteq \{1, 2, \ldots, d\}$. Random scalars will be written in uppercase, e.g., X, while realizations will be written in lowercase, e.g., x. Matrices will also be written in un-bolded uppercase, e.g., A; context will distinguish them from random variables. Vectors will be bolded, with random vectors in boldface uppercase font, e.g., $\mathbf{X}$, and deterministic vectors and realizations of random vectors in boldface lowercase font, e.g., $\mathbf{x}$. Subvectors will be denoted with subscripts, e.g., $\mathbf{x}_{\mathcal{S}} = (x_i)_{i \in \mathcal{S}}$. We will write $A_{\mathcal{S}}$ to denote the submatrix of A formed by the columns indexed by $\mathcal{S}$. Constant vectors and matrices of all zeros (resp. ones) will be written as $\mathbf{0}$ (resp. $\mathbf{1}$), with dimension determined by context. All random variables will be understood to be defined over an underlying sample space Ω with joint probability measure $\mathbb{P}$. Given a random variable Z, we will denote by $\mathbb{P}_Z$ the marginal distribution of Z, and p_Z its corresponding density. When clear from context, we will drop the density subscripts, e.g., $p(z)$ will be understood to mean $p_Z(z)$. Expectations will be denoted by $\mathbb{E}$. To avoid measure-theoretic issues, we will assume throughout that the distribution of $\mathbf{X}$ admits either a probability mass function (pmf) or probability density function (pdf), $p(\mathbf{x})$. Finally, for any measurable function φ, we will use the following notation for conditional expectations with respect to $p(\mathbf{x}'_{\mathcal{S}^c} \mid \mathbf{x}_{\mathcal{S}})$:

$$\mathop{\mathbb{E}}_{\mathbf{X}_{\mathcal{S}^c} \mid \mathbf{x}_{\mathcal{S}}} [\varphi(\mathbf{X}_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}})] \doteq \int p(\mathbf{x}'_{\mathcal{S}^c} \mid \mathbf{x}_{\mathcal{S}}) \varphi(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}) \, \mathrm{d}\mathbf{x}'_{\mathcal{S}^c}. \tag{1}$$

For discrete $\mathbf{X}$, one may replace the integral with a sum.

Problem Setting. The explanation problem we study is as follows: given a trained model $f : \mathcal{X} \to \mathcal{Y}$ that maps inputs in $\mathcal{X}$ to outputs in $\mathcal{Y}$, we seek a small *sufficient* set $\mathcal{S} \subseteq [d]$ such that the values of $\mathbf{x}$ on $\mathcal{S}$ approximately determine the prediction $f(\mathbf{x})$, in a sense made formal by Definition 1. When f is a classification model, we will usually take $\mathcal{Y}$ to be a space of distributions $\mathcal{P}(\mathcal{L})$ and $f_y(\mathbf{x})$ to approximate the conditional label probability $\mathbb{P}(Y = y \mid \mathbf{X} = \mathbf{x})$ for each $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{L}$. When f is a regression model, we will take f to approximate the conditional expectation $\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$. Throughout, we will consider a nonnegative function $D : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ with $D(\mathbf{y}, \mathbf{y}') = 0$ if and only if $\mathbf{y} = \mathbf{y}'$, but we do not require D to be symmetric. Choices of D can be, for instance, the squared ℓ_2 distance, KL divergence, or TV distance. To ensure that conditional expectations are well-defined, we will assume that the input to be explained, $\mathbf{x}$, has positive probability mass, or density, under $\mathbb{P}$, i.e., $p(\mathbf{x}) > 0$.

3 Minimal sufficient subset explanations

We will adopt the following definition of sufficiency, which strengthens the definition of Bharti et al. (2025) and takes inspiration from similar definitions of sufficiency (Chattopadhyay et al., 2022) and distortion (MacDonald et al., 2019).

Definition 1 (ε-Sufficiency)**.** Fix the model f, input instance $\mathbf{x}$, and D. The nonempty subset $\mathcal{S} \subseteq [d]$ is *ε-sufficient for f at $\mathbf{x}$ with respect to D if*

$$\mathop{\mathbb{E}}_{\mathbf{X}_{\mathcal{S}^c} \mid \mathbf{x}_{\mathcal{S}}} \left[D(f(\mathbf{X}_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}), f(\mathbf{x})) \right] \leq \varepsilon,$$

with $\varepsilon \geq 0$. When f, $\mathbf{x}$, or D are clear from context, we will omit the corresponding qualifiers.

By setting D to be an indicator, $D(\mathbf{u}, \mathbf{v}) = \mathbb{I}(\mathbf{u} \neq \mathbf{v})$, and taking $f : \mathbb{R}^d \to \mathcal{L}$ to be a classifier that outputs a class label, Definition 1 generalizes same-decision probability (Wang et al., 2021).

When $\varepsilon = 0$, we recover the logical definition of sufficiency, i.e. that $f(\mathbf{x}_{\mathcal{S}}, \mathbf{X}_{\mathcal{S}^c}) \overset{\text{a.s.}}{=} f(\mathbf{x})$, where $\mathbf{X}_{\mathcal{S}^c} \sim \mathbb{P}_{\mathbf{X}_{\mathcal{S}^c} \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}}}$. When $\varepsilon > 0$, on the other hand, Definition 1 recovers relaxed versions of this notion, with the choice of D playing a role in how these are formally quantified.

As in several prior works, we seek a *minimal* sufficient subset (MSS) that serves as a concise and informative explanation. Thus we present the following definition, which adopts the notion of minimality used in, e.g., prime implicant (PI) explanations (Shih et al., 2018).

Definition 2 (Minimal ε-Sufficient Subset)**.** A subset $\mathcal{S} \subseteq [d]$ is a minimal ε-sufficient subset (ε-MSS) if $\mathcal{S}$ is ε-sufficient and there exists no $\mathcal{T} \subsetneq \mathcal{S}$ that is ε-sufficient.

When $\varepsilon = 0$, we refer to a 0-MSS simply as an MSS. In what follows, terms like "smallest" or "minimal" will refer to minimality with respect to set inclusion.

4 The ubiquity of multiple minimal sufficient subsets

The non-uniqueness of MSSs is a multifaceted phenomenon. Here we present several perspectives on this non-uniqueness, which touch upon Boolean logic, application-specific considerations, CI, nonlinearity, approximate sufficiency, and degenerate conditional feature distributions. Throughout, we will discuss a number of instructive examples.

4.1 `OR` logic and multiple MSSs in vision problems

The following toy setup gives strong intuition on why multiple MSSs exist; it is an example of the widely noted multiplicity of (probabilistic) prime implicants in logic-based explanations (Wäldchen, 2022, p. 51), (Darwiche & Hirth, 2020; Ignatiev et al., 2020).[2]

Example 1 (`OR`)**.** Let $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \text{Ber}(1/2)$ and $Y = X_1 \vee X_2$, where $\vee$ is the `OR` operation. Then for $\mathbf{x} = (1, 1)$ and $f(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})$ the sets $\mathcal{S} = \{1\}$ and $\mathcal{S}' = \{2\}$ are both MSSs.

This `OR` logic, and generalizations thereof, are ubiquitous in machine learning. In the case of image classification, say, any time there are multiple objects of a particular target class in an image, there may be multiple sufficient explanations. For example, consider the problem of malaria detection via images of human blood smears (Ljosa et al., 2012), where the goal is to classify the entire image as "infected" or "not infected". A single infected cell suffices to determine the prediction, although multiple may be present. Symmetries also

[2] This example is also similar to (Carter et al., 2019, Example 2), albeit using a different definition of sufficiency.

easily give rise to multiple MSSs. If we consider classifying images based on the presence or absence of earrings in portraits, as we will do in Section 6, there will often be at least two earrings present, corresponding to the bilateral symmetry of the face. Seeing one (earring) should be all that a model needs to detect the presence of at least one (earring). A classifier trained to optimality on these tasks will therefore exhibit the type of `OR` logic outlined in Example 1, and will thus possess multiple explanations on certain inputs. While each of these sufficient explanations may not be strictly minimal, practically, we find that methods to find MSSs will often recover each of them.

4.2 The role of context-specific independence

Example 1 is intuitive and relates logical notions of sufficiency to our probabilistic Definition 1, but the degenerate conditional distribution $p(y \mid \mathbf{x})$ indicates a deterministic relationship between the features $\mathbf{X}$ and the label Y, which does not capture the more general situation where Y is only partially determined by $\mathbf{X}$. Here we present a more general, but related, reason for non-uniqueness of MSSs under our probabilistic definition of sufficiency by connecting MSSs to CI relationships.

Assuming the model f is the Bayes-optimal classifier, i.e., $f_y(\mathbf{x}) = \mathbb{P}(Y = y \mid \mathbf{X} = \mathbf{x})$, a subset $\mathcal{S}$ is an MSS if it is the minimal set satisfying the following CI relation (see Appendix A for proof):

$$\mathbf{X}_{\mathcal{S}^c} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}}. \tag{2}$$

In other words, $\mathcal{S}$ is the smallest set such that the remaining features $\mathbf{X}_{\mathcal{S}^c}$ are CI of $\mathbf{Y}$ given the observed features $\mathbf{x}_{\mathcal{S}}$. Due to the dependence on the specific values $\mathbf{x}_{\mathcal{S}}$ taken by $\mathbf{X}_{\mathcal{S}}$ (the *context*) in the conditioning event $\mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}}$, this CI relation is said to be *context-specific* (Boutilier et al., 1996). In characterizing an MSS this way, we can draw connections to the well-studied *Markov blankets* of $\mathbf{Y}$ (Pearl, 1988, p. 97), which are subsets $\mathcal{S}$ such that

$$\mathbf{X}_{\mathcal{S}^c} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}_{\mathcal{S}}. \tag{3}$$

The smallest such subset $\mathcal{S}$ is called the *Markov boundary* of $\mathbf{Y}$ (Pearl, 1988, p. 97). One might hope that standard conditions for Markov boundary uniqueness also imply the uniqueness of their context-specific counterpart (i.e., MSSs), thereby eliminating the issue of multiple explanations. For example, if the joint distribution $p(\mathbf{x}, \mathbf{y})$ satisfies strict positivity, i.e., $p(\mathbf{x}, \mathbf{y}) > 0$ for all values of $\mathbf{x}$ and $\mathbf{y}$, then the Markov boundary of $\mathbf{Y}$ is unique (Pearl, 1988, Theorem 1, Theorem 4). As we will shortly see, such conditions unfortunately do not carry over to the context-specific independence represented by MSSs and Equation 2.

Mathematically, Example 1 reflects that knowing that either $x_1 = 1$ or $x_2 = 1$ is sufficient because

$$\mathbb{P}(Y = 1 \mid \mathbf{X} = (1, 1)) = \mathbb{P}(Y = 1 \mid X_1 = 1) = \mathbb{P}(Y = 1 \mid X_2 = 1) = 1.$$

Since $\mathbb{P}(Y = 0 \mid \mathbf{X} = (1, 1)) = 0$, strict positivity of $p(\mathbf{x}, y)$ is violated. Revisiting the earlier discussion of strict positivity implying a unique Markov boundary, one might wonder whether “fixing” this degeneracy leads to a unique solution. Unfortunately, this is not the case. Consider the following example (adapted from (Klein & Shimony, 2004)), which, despite exhibiting a strictly positive joint distribution $p(\mathbf{x}, y)$, also exhibits multiple MSSs.

Example 2 (Leaky `OR`)**.** Let $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \text{Ber}(1/2)$, $N \sim \text{Ber}(\delta)$ with $0 < \delta < 1/2$, and $Y = (X_1 \vee X_2) \oplus N$, where $\oplus$ denotes `XOR`. Then, for $\mathbf{x} = (1, 1)$ and $f(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})$, the sets $\mathcal{S} = \{1\}$ and $\mathcal{S}' = \{2\}$ are both MSSs.

In Example 2, $p(\mathbf{x}, y) = p(\mathbf{x})p(y \mid \mathbf{x}) \geq \delta/4 > 0$ for all $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$, so strict positivity is satisfied. However,

$$\mathbb{P}(Y = 1 \mid \mathbf{X} = (1, 1)) = \mathbb{P}(Y = 1 \mid X_1 = 1) = \mathbb{P}(Y = 1 \mid X_2 = 1) = 1 - \delta,$$

so that $X_1 = 1$ and $X_2 = 1$ each determine Y up to the noise (Klein & Shimony, 2004).[3]

[3] This argument further reveals that the weaker intersection property of CI (Pearl, 1988, Theorem 1), which also suffices to guarantee a unique Markov boundary (Pearl, 1988, Theorem 4), fails to guarantee a unique MSS, since the intersection property follows from strict positivity (Drton et al., 2009, Proposition 3.1.3). The same reasoning shows that the intersection property of context-specific independence (Corander et al., 2019) also cannot guarantee a unique MSS.

4.3 Nonlinearity and approximate sufficiency

For both Example 1 and Example 2, the existence of multiple MSSs was a result of the relationship between $\mathbf{X}$ and Y. In general, the existence of multiple MSSs depends on the interplay between the joint feature distribution $p(\mathbf{x})$ and the model f. In contrast, for generalized linear models with certain nonlinearities, the MSS is unique *regardless* of the underlying feature distribution $p(\mathbf{x})$.

Proposition 1 (Generalized linear models have unique MSSs)**.** *Suppose that* $\mathbf{X}$ *has a continuous and strictly positive density, i.e.,* $p(\mathbf{x}) > 0$ *for all* $\mathbf{x} \in \mathcal{X} = \mathbb{R}^d$*, and let* D *be continuous. Let* $f(\mathbf{x}) = g(A\mathbf{x} + \mathbf{b})$*, where* $A \in \mathbb{R}^{m\times d}$, $\mathbf{b} \in \mathbb{R}^m$*, and* $g : \mathbb{R}^m \to \mathbb{R}^k$ *for some* m*. If* g *is continuous and injective on* $\operatorname{span}(A) + \mathbf{b}$*, where* $\operatorname{span}(A)$ *is the column span of* A*, then for all* $\mathbf{x} \in \mathcal{X}$ *the MSS for* f *at* $\mathbf{x}$ *with respect to* D *is unique and equals the nonzero column indices of* A*. Furthermore, if* $g = \operatorname{softmax}$*, then for all* $\mathbf{x} \in \mathcal{X}$*, the MSS is also unique and equals* $\{i \in [d] : \forall c \in \mathbb{R}, \mathbf{a}_i \neq c\mathbf{1}\}$*, where* $\mathbf{1}$ *is the constant vector of all ones and* $\mathbf{a}_i$ *is the* i^{th} *column of* A*.*

The proof is deferred to Appendix B. We remark that Proposition 1 encompasses linear regression models and linear classifiers (both binary and multi-class), since the logistic function is injective. The following corollary captures the role of nonlinearity in the multiplicity of MSSs more explicitly.

Corollary 1. *Suppose* $\mathbf{X}$ *has a continuous and strictly positive density, i.e.,* $p(\mathbf{x}) > 0$ *for all* $\mathbf{x} \in \mathcal{X} = \mathbb{R}^d$*. Let* f *and* D *be continuous. If there exist multiple MSSs for* f *at some* $\mathbf{x}^0 \in \mathbb{R}^d$*, then* f *is nonlinear.*

That is, under the assumptions of Proposition 1, nonlinearity is necessary for non-unique MSSs to occur. Proposition 1 stands in contrast with logic-based explanations, for which linear models may possess many MSSs (Marques-Silva et al., 2020). This difference, however, relies on the (strong) assumptions of Proposition 1. Relaxing these by allowing approximate ($\varepsilon > 0$) sufficiency and degenerate conditional distributions means that even linear models can have multiple MSSs, as Examples 3 and 4 demonstrate.

Example 3. Let $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, I)$, $\mathbf{X} \in \mathbb{R}^3$, and let $f(\mathbf{x}) = A\mathbf{x}$, where $A \in \mathbb{R}^{3\times 3}$ has normalized columns $\mathbf{a}_i$, $i \in [3]$. Write $A_\mathcal{S}$ to denote the submatrix of A formed by the columns indexed by $\mathcal{S}$. Setting D to the squared Euclidean metric in Definition 1, an ε-MSS $\mathcal{S}$ for f at $\mathbf{x} = \mathbf{0}$ satisfies

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{X}_{\mathcal{S}^c}|\mathbf{x}_\mathcal{S}} \left[\|A_{\mathcal{S}^c}\mathbf{X}_{\mathcal{S}^c} + A_\mathcal{S}\mathbf{x}_\mathcal{S} - A\mathbf{x}\|^2\right] &= \mathbb{E}\left[\|A_{\mathcal{S}^c}\mathbf{X}_{\mathcal{S}^c}\|^2\right] && (\mathbf{x} = \mathbf{0} \text{ and } \mathbf{X}_{\mathcal{S}^c} \perp\!\!\!\perp \mathbf{X}_\mathcal{S}) \\
&= \operatorname{Tr}(A_{\mathcal{S}^c}^\top A_{\mathcal{S}^c}) && (\operatorname{Tr} \text{ commutes with } \mathbb{E} \text{ and } \mathbb{E}[\mathbf{X}_{\mathcal{S}^c}\mathbf{X}_{\mathcal{S}^c}^\top] = I) \\
&= |\mathcal{S}^c| && ([A_{\mathcal{S}^c}^\top A_{\mathcal{S}^c}]_{ii} = \|\mathbf{a}_i\|_2^2 = 1) \\
&\leq \varepsilon.
\end{aligned}$$

Setting $\varepsilon = 1$, we have three MSSs: $\mathcal{S}_1 = \{2, 3\}$, $\mathcal{S}_2 = \{1, 3\}$, and $\mathcal{S}_3 = \{1, 2\}$.

In applications, it is natural to seek an ε-MSS with $\varepsilon > 0$: 0-MSSs may be so large that they lose interpretability (Wäldchen et al., 2021). Thus such $\varepsilon > 0$ settings are more relevant to practical applications of Definition 1 and, as Example 3 illustrates, unique MSSs should not be expected in this case.

4.4 Degenerate conditional feature distributions

So far, we have seen that MSS multiplicity depends on the properties of the model f. However, they may also arise due to the feature distribution $\mathbb{P}_\mathbf{X}$, independently of f, as the following example shows.

Example 4. Let $\mathbf{X}$ be discrete with uniform probability over the set $\{-1/2, 1/2\}^d \cup \{\mathbf{1}\}$. Then fixing any $X_i = 1$ fully determines the remaining $\mathbf{X}_{[d]\setminus\{i\}}$:

$$\mathbb{P}(\mathbf{X}_{[d]\setminus\{i\}} = \mathbf{1} \mid X_i = 1) = \frac{\mathbb{P}(\mathbf{X}_{[d]\setminus\{i\}} = \mathbf{1}, X_i = 1)}{\mathbb{P}(X_i = 1)} = 1,$$

and thus

$$\mathbb{E}[D(f(\mathbf{X}_{[d]\setminus\{i\}}, x_i), f(\mathbf{x})) \mid X_i = 1] = 0.$$

I.e., for any $i \in [d]$, $\{i\}$ is a MSS for any f at $\mathbf{x} = \mathbf{1}$.

We have discussed multiple perspectives on MSS non-uniqueness, demonstrating that disjunctive logic, symmetries and repetitions, intersection properties of context-specific independence, nonlinearity, approximate sufficiency, and degeneracies in the feature distribution all contribute to the presence of multiple MSSs. Taken together, these facts suggest that a unique MSS is the exception, rather than the rule. Therefore, rather than seeking a single MSS, it is better to seek multiple. Accordingly, in Section 5, we propose a simple method to extend techniques for finding a single MSS to strategies for finding several.

5 Let me explain, again

Since we do not know *a priori* when multiple minimal sufficient subsets will exist in a given scenario, it is sensible to look for multiple. However, some methods (e.g., (Fong & Vedaldi, 2017; Fong et al., 2019; MacDonald et al., 2019; Luss & Dhurandhar, 2024)) only find a single MSS. To extend such methods, we propose a strategy for identifying multiple MSSs. Similarly to prior work (Carter et al., 2019; 2021; Byra & Skibbe, 2025), this strategy finds explanations in an iterative fashion, constraining new explanations to be distinct from those found previously.

5.1 Meta-algorithm

The meta-algorithm we study here, which we term Let Me Explain Again (LMEA), finds multiple MSSs via repeated calls to a single-MSS explanation method, which we will call a *minimal sufficiency oracle (MSO).* For the discussion that follows, we will make the following assumption on the MSO.

Assumption 1. The MSO correctly identifies an ε-MSS $\mathcal{S}$ for f at $\mathbf{x}$, for all values of $\emptyset \neq \mathcal{A} \subseteq [d]$, provided an ε-MSS $\mathcal{S} \subseteq \mathcal{A}$ exists. If an ε-MSS does not exist, then the MSO returns the empty set $\emptyset$.

Under Assumption 1, LMEA provably finds a disjoint subset of all MSSs, which we formally state in Proposition 2. While Assumption 1 is strong, our results demonstrate the practical utility of LMEA in situations where we can expect the presence of multiple disjoint MSSs.

In words, the algorithm proceeds as follows. We start with the full active feature pool $\mathcal{A} = [d]$, and iteratively search for MSSs by querying the MSO, which returns a set $\mathcal{S}$. We then add $\mathcal{S}$ to a running set $\mathcal{E}$ of MSSs found so far and remove $\mathcal{S}$ from the active feature pool $\mathcal{A}$. This process is repeated until the MSO returns an empty set or a user-specified maximum number of iterations N is reached.

By substitution of various MSOs (as opposed to specific, fixed choices), LMEA generalizes the algorithms proposed by Carter et al. (2019; 2021), which operate via greedy search, and Byra & Skibbe (2025), which is similar to LMEA with extremal perturbations (XP) (Fong et al., 2019) as the MSO. We note that Algorithm 1 is not tied to our particular definition of sufficiency (Definition 1), and in Section 6, we show how LMEA can be applied to MSOs that optimize for different notions of sufficiency.

Algorithm 1 Let Me Explain Again (LMEA)

```
Require: Model $f$, input $\mathbf{x} \in \mathbb{R}^d$, sufficiency level $\varepsilon$,
  MSO, maximum explanations to return $N$
  $\mathcal{A} \leftarrow [d]$
  $\mathcal{E} \leftarrow \emptyset$
  for $k = 1, \ldots, N$ do
      $\mathcal{S} \leftarrow \mathsf{MSO}(f, \mathbf{x}, \mathcal{A})$
      if $\mathcal{S} = \emptyset$ then
          break
      end if
      $\mathcal{E} \leftarrow \mathcal{E} \cup \{\mathcal{S}\}$
      $\mathcal{A} \leftarrow \mathcal{A} \setminus \mathcal{S}$
  end for
  return $\mathcal{E}$
```

5.2 Termination and completeness

Here we show that under appropriate assumptions, Algorithm 1 terminates and returns a disjoint subset of the complete set of all MSSs. To streamline the presentation, we defer proofs to Appendix C. First, we note that LMEA terminates and has linear query complexity.

Lemma 1. *LMEA terminates after at most d calls to the MSO, where d is the dimension of* $\mathbf{x}$.

Under Assumption 1, LMEA inherits the correctness property of one of the algorithms it generalizes; the following result is analogous to an easy corollary of (Carter et al., 2019, Proposition 2) under certain assumptions.[4]

Proposition 2. *Under Assumption 1, LMEA with $N = d$ returns a set of explanations $\mathcal{E}$ whose union, $\mathcal{U} \doteq \bigcup_{\mathcal{S}\in\mathcal{E}} \mathcal{S}$, intersects each ε-MSS for f at $\mathbf{x}$: for all ε-MSSs $\mathcal{S}$, $\mathcal{S} \cap \mathcal{U} \neq \emptyset$.*

In words, Proposition 2 demonstrates that each ε-MSS will have at least one feature in common with some explanation in the returned set $\mathcal{E}$ of MSSs. When there are multiple ε-MSSs and they are all disjoint, Proposition 2 implies that LMEA will recover them all. However, ε-MSSs may overlap. In high-dimensional feature spaces (such as high-resolution images), the degree of this overlap may become perceptually negligible, as $\mathcal{U}$ may intersect each explanation $\mathcal{S} \in \mathcal{E}$ at one out of hundreds of features comprising $\mathcal{S}$. Furthermore, in the worst case, each MSS will intersect every other, and LMEA will only recover one MSS. Proposition 2 also makes a strong assumption on the MSO. However, when multiple disjoint ε-MSSs can be expected, LMEA performs well in practice, as we will show next.

6 Experiments

To demonstrate the practical applicability of LMEA, we present experiments on three multiple instance learning (MIL) (Dietterich et al., 1997) image classification tasks, for which the presence of a single object in the image suffices for a positive prediction, and for which there may be many such objects. For each of these tasks, we run LMEA on images from the test set to recover as many explanations as possible. The explanations output by LMEA are evaluated against ground truth labels (either bounding boxes or segmentation masks) that indicate the salient regions of the image for the prediction, implying sufficiency according to human annotators.

6.1 MSOs

Rate distortion explanations (RDE). One of our LMEA MSOs combines ideas from rate-distortion explanations (RDE) (MacDonald et al., 2019) and extremal perturbations (XP) (Fong et al., 2019). We choose RDE because its distortion criterion closely matches our sufficiency criterion in Definition 1, making it a good approximation to an MSO. More specifically, this MSO solves a problem of the form

$$\mathbf{s}^\star = h\left(\operatorname*{argmin}_{\mathbf{s}\in[0,1]^d} L(\mathbf{s})\right), \text{ with}$$
$$L(\mathbf{s}) \doteq \frac{1}{2}\,\mathbb{E}\left[\left\|f(h(\mathbf{s})\odot\mathbf{x} + (\mathbf{1}-h(\mathbf{s}))\odot\widetilde{\mathbf{X}}) - f(\mathbf{x})\right\|^2\right] + \lambda\|h(\mathbf{s})\|_1, \tag{4}$$

where $\widetilde{\mathbf{X}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for some choice of $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ (MacDonald et al., 2019), and h is a smoothing function for the mask $\mathbf{s}$. Empirically, this smoothing makes the resulting explanations more reliable, especially in later iterations of LMEA. Following Fong et al. (2019), we set h to be the smoothmax function. The solution $\mathbf{s}^\star$ to Equation (4) can be thought of as a soft version of the characteristic vector $\mathbf{1}_{\mathcal{S}^\star}$ corresponding to a small ε-sufficient set $\mathcal{S}^\star$.[5] We solve (4) for a number of λ's and pick the solution for the largest λ (and thus the sparsest mask $\mathbf{s}^\star$) that produces an ε-sufficient result, where ε is chosen as specified in Section 6.3.

Extremal perturbations (XP). We also study the use of extremal perturbations (XP) (Fong et al., 2019) as an MSO. XP solves the following optimization problem:

$$\mathbf{s}^\star = h\left(\operatorname*{argmax}_{\mathbf{s}\in[0,1]^d} f_{y^\star}(g(\mathbf{s},\mathbf{x})) - \lambda R_\alpha(\mathbf{s})\right),$$

where $y^\star = \operatorname{argmax}_y f_y(\mathbf{x})$ is the predicted class, $g(\mathbf{s},\mathbf{x})$ is a function that computes a perturbed version of the input $\mathbf{x}$ based on the mask $\mathbf{s}$, $R_\alpha(\mathbf{s})$ is a regularizer used to control the sparsity of the attribution

[4] Specifically, if we assume that each MSS $\mathcal{S}$ is such that every $\mathcal{T} \supseteq \mathcal{S}$ is also sufficient (with the definitions and context of (Carter et al., 2019)) then an analog of Proposition 2 follows from their Proposition 2 in the context of their paper.

[5] A characteristic vector $\mathbf{1}_\mathcal{S}$ is such that $(\mathbf{1}_\mathcal{S})_i = 1$ if $i \in \mathcal{S}$ and $(\mathbf{1}_\mathcal{S})_i = 0$ otherwise.

mask $\mathbf{s}^\star$, α is a sparsity constraint parameter, and λ is the corresponding regularization strength. The perturbation function g can be a blurred version of the original image with local spread of the Gaussian blurring kernel determined by the smoothed mask value (so that $h(\mathbf{s})_i = 1$ corresponds to no blur and $h(\mathbf{s})_i = 0$ corresponds to maximal blur), or the "fade-to-black" perturbation $g(\mathbf{s}, \mathbf{x}) = h(\mathbf{s}) \odot \mathbf{x}$, where h is the smoothmax function (Fong et al., 2019). Finally, for integer-valued αd, $R_\alpha(\mathbf{s}) = \|\text{vecsort}(\mathbf{s}) - \mathbf{1}_{[\alpha d]}\|^2$ is an area constraint regularizer that encourages the mask $\mathbf{s}$ to be nearly binary and αd-sparse. Although both XP and RDE find MSSs by solving continuous optimization problems, they differ in their (i) notions of sufficiency, (ii) methods of controlling sparsity, and (iii) perturbations of the input $\mathbf{x}$.

For the XP MSO, LMEA is similar to the method of Byra & Skibbe (2025) with an infinite Sørensen-Dice (Dice, 1945; Sørensen, 1948) penalty coefficient. In contrast to their approach, which re-implements XP using implicit neural representations, we instead re-use (without modification) the implementation of XP provided in the TorchRay package, using a trick outlined in Section 6.3 to restrict the active set $\mathcal{A}$.[6] Similarly to RDE, we pick the explanation with the smallest area parameter that is ε-sufficient, where ε is set as specified in Section 6.3.

6.2 Baselines

MultiReX. MultiReX (Chockler et al., 2023)[7] finds multiple explanations via a stochastic search procedure, based on a pre-computed ranking of pixels by causal responsibility for the classifier output. Our experiments test LMEA against this existing multi-MSS method, including both the number of explanations recovered and their quality. More details on MultiReX and its hyperparameters are given in Appendix D. For consistent comparison to LMEA, we set MultiReX's hyperparameters so that it returns a diverse set of explanations as measured by pairwise overlap. Specifically, we enforce a worst-case pairwise Sørensen-Dice score (Sørensen, 1948; Dice, 1945) between MSSs of 0.1.

Patch-wise SIS (PSIS). We also implement sufficient input subsets (SIS) (Carter et al., 2019). The original SIS paper proposed backward elimination on individual pixels. However, this approach requires $\mathcal{O}(d^2)$ forward passes through the model f (Carter et al., 2021) and results in explanations that lack spatial contiguity, making them harder to interpret. To compare to this method in our high-dimensional image classification setup, we therefore implement SIS to do backward selection on a 14×14 grid of image patches, similar to the grid used by Shitole et al. (2021). We refer to this method as patch-wise SIS (PSIS) and note that SIS can be interpreted as a special case of LMEA with a simple MSO and a slightly different notion of sufficiency. For fair comparison, we give SIS the same maximum number of explanations N as LMEA.

6.3 LMEA implementation details

Here we discuss two important implementation details of LMEA. Further implementation details are provided in Appendix G.1.

Sparsity-based postprocessing (SBP). While in Algorithm 1 the MSO returns $\mathcal{S} = \emptyset$ to indicate the absence of a MSS, in practice the solutions of MSOs must be checked to ensure sufficiency. If $N = \infty$ in Algorithm 1, then the MSO will eventually run out of informative features to select. In this case, the MSO may return a spurious mask; such masks are usually larger than those containing sets of informative features. Both PSIS and LMEA with the RDE MSO exhibit this problem. To mitigate this issue, we employ a sparsity-based postprocessing (SBP) strategy. Specifically, after LMEA has terminated, we compute the minimum explanation size $K_{\min} \doteq \min_{\mathcal{S}' \in \mathcal{E}} |\mathcal{S}'|$, and we keep only the explanations that fall within a certain multiple of this size, say those $\mathcal{S}$ with $|\mathcal{S}| > (1 + \delta)K_{\min}$. In the results that follow, we set $\delta = 1$. For the MultiReX and SIS baselines, we also apply this rule and report results both with and without SBP for consistent comparisons.

Specifying the active set $\mathcal{A}$. It is convenient to be able to re-use MSO implementations within the LMEA framework. However, note that it is not guaranteed that a given implementation, e.g., the XP implementation discussed in Section 6.1, will accept a restricted feature pool $\mathcal{A}$ as an argument as specified in Algorithm 1

[6] The TorchRay package is available at https://github.com/facebookresearch/TorchRay.

[7] We use the authors' implementation: https://github.com/ReX-XAI/ReX.

(aside from hyperparameters, such methods typically accept as input only the model f and features $\mathbf{x}$, since these methods find a *single* MSS). Therefore, we propose a technique to wrap existing implementations of MSOs that optimize soft attribution masks by solving similar optimization problems to Equation 4, making LMEA more generally applicable to off-the-shelf MSOs.

To illustrate the idea, assume that f is differentiable and that the MSO optimizes $L(\mathbf{s})$ in Equation (4) with $h(\mathbf{x}) = \mathbf{x}$ via projected gradient descent (PGD). Defining $\widetilde{\mathbf{X}}_{\mathbf{s}} \doteq \mathbf{s} \odot \mathbf{x} + (\mathbf{1} - \mathbf{s}) \odot \widetilde{\mathbf{X}}$ and applying the chain rule, the gradient of the n-sample Monte Carlo estimate of the loss $L(\mathbf{s}^t)$ at iteration t is

$$\nabla L(\mathbf{s}^t) = \frac{1}{2n} \sum_{n=1}^{n} \operatorname{diag}\big(\mathbf{x} - \widetilde{\mathbf{x}}_{\mathbf{s}^t}^{(n)}\big) J_f\big(\widetilde{\mathbf{x}}_{\mathbf{s}^t}^{(n)}\big)^\top \Big(f\big(\widetilde{\mathbf{x}}_{\mathbf{s}^t}^{(n)}\big) - f(\mathbf{x})\Big) + \lambda \mathbf{1},$$

where $J_f\big(\widetilde{\mathbf{x}}_{\mathbf{s}^t}^{(n)}\big)$ is the Jacobian of f evaluated at $\widetilde{\mathbf{x}}_{\mathbf{s}^t}^{(n)}$. If at each iteration t we intervene on the columns of the Jacobian $J_f\big(\widetilde{\mathbf{x}}_{\mathbf{s}^t}^{(n)}\big)$ corresponding to the already selected features $\mathcal{A}^c$ by setting these columns to zero, i.e., $[J_f\big(\widetilde{\mathbf{x}}_{\mathbf{s}^t}^{(n)}\big)]_{\mathcal{A}^c} \leftarrow \mathbf{0}$, we then obtain the PGD update

$$\begin{aligned} \mathbf{s}_{\mathcal{A}^c}^{t+1} &\leftarrow \operatorname{proj}_{[0,1]^d}\left(\mathbf{s}_{\mathcal{A}^c}^{t} - \eta\lambda\mathbf{1}\right) \\ \mathbf{s}_{\mathcal{A}}^{t+1} &\leftarrow \operatorname{proj}_{[0,1]^d}\left(\mathbf{s}_{\mathcal{A}}^{t} - \eta[\nabla L(\mathbf{s}^t)]_{\mathcal{A}}\right), \end{aligned}$$

where η is the learning rate. Thus $\lim_{t\to\infty} \mathbf{s}_{\mathcal{A}^c}^t = \mathbf{0}$, while the gradient updates to the coordinates in $\mathcal{A}$ are left undisturbed. The above idea can be easily implemented in PyTorch (Paszke et al., 2019) via a simple wrapper over the model f that disconnects certain variables from the computation graph (sample code is provided in Appendix G.1), rather than directly intervening in the optimization process. This approach allows obtaining new explanations simply by passing a wrapped model to the MSO, without altering the MSO implementation itself.[8] To adapt this idea to smoothing functions $h \neq \mathrm{Id}$ in our experiments, we perform a morphological dilation (Haralick et al., 1987) on each explanation $\mathcal{S}$ before removing it from $\mathcal{A}$ in Algorithm 1.

6.4 Evaluation

For the remainder of the paper, we will refer to LMEA, MultiReX, and PSIS as *multi-explanation methods (MEMs)*, since they each seek to produce multiple MSSs. To numerically assess both the number and quality of explanations found by each MEM, we introduce a number of metrics. These seek to measure the precision of the explanations produced, that is, the fraction of pixels in some MEM explanation overlap with the ground truth, and also the recall, or the fraction of ground truth object pixels that are recovered by the MEM. We also report the mean number of explanations recovered by each method, their sparsity, and the number of samples for which each method fails to find any explanation.

First, we report "Expl./Obj.," representing the number of explanations per image that each MEM produces, normalized by the number of salient objects per image. While we expect the number of explanations to roughly correspond to the number of salient objects, we do not know the correct number of explanations *a priori*, making comparisons of this metric difficult to interpret. We therefore refrain from specifying a desired ordering of this metric across MEMs or MSOs, but report it to convey the number of explanations typically recovered by each method. Since we seek minimal explanations, we report the sparsity of the binarized attribution masks, that is, the size of each MSS. For some samples, RDE and XP are unable to find any explanation. To quantify this, we also report "Missing," which is the number of instances for which each method failed to find any explanation. Ideally, this metric will be zero. Samples with missing explanations are excluded from metrics calculations, with the exception of "Expl./Obj."

Next, we introduce metrics to determine the degree to which explanations overlap with the known informative regions of the image for the given task. We report intersection-over-union (IoU) scores between the union of

[8] A similar argument applies to XP. XP uses an area constraint penalty and logarithmic parametrization of $\mathbf{s}^t$, so we are not guaranteed to drive $\mathbf{s}_{\mathcal{A}^c}^t$ to zero with this method; however, we observe that LMEA with the XP MSO rarely re-selects regions in our experiments.

MEM explanation mask regions $\mathcal{R}_E \doteq \bigcup_{\mathcal{S}\in\mathcal{E}} \mathcal{S}$ and the union of the ground truth segmentation mask regions $\mathcal{R}_G \doteq \bigcup_{\mathcal{T}\in\mathcal{M}} \mathcal{T}$, where $\mathcal{M}$ is the set of all ground truth segmentation masks for $\mathbf{x}$:

$$\mathrm{IoU}(\mathcal{R}_E, \mathcal{R}_G) \doteq \frac{|\mathcal{R}_E \cap \mathcal{R}_G|}{|\mathcal{R}_E \cup \mathcal{R}_G|}. \tag{5}$$

A score of 1 indicates perfect overlap, while a score of 0 indicates no overlap. Towards a notion of recall, we use the intersection-over-ground-truth (IoGT) metric, representing the same score instead normalized by the area of the ground truth segmentation:

$$\mathrm{IoGT}(\mathcal{R}_E, \mathcal{R}_G) \doteq \frac{|\mathcal{R}_E \cap \mathcal{R}_G|}{|\mathcal{R}_G|}. \tag{6}$$

If IoGT is 1, then all ground truth pixels belong to some explanation output by the MEM explanation, while an IoGT of 0 indicates that none were recovered. Similarly, we use the intersection-over-explanation (IoE) metric as a proxy for precision. The IoE is the intersection normalized by the explanation mask area:

$$\mathrm{IoE}(\mathcal{R}_E, \mathcal{R}_G) \doteq \frac{|\mathcal{R}_E \cap \mathcal{R}_G|}{|\mathcal{R}_E|}. \tag{7}$$

If IoE equals 1, then each pixel in each of the MEM explanations belongs to a ground truth object annotation. If it equals zero, then none belong to any ground truth object annotation. We report IoU and IoGT values without comparison across single-MSS methods (i.e., MSOs), since individual explanations may overlap a single salient object or some fraction thereof, which may constitute a small portion of $\mathcal{R}_G$.

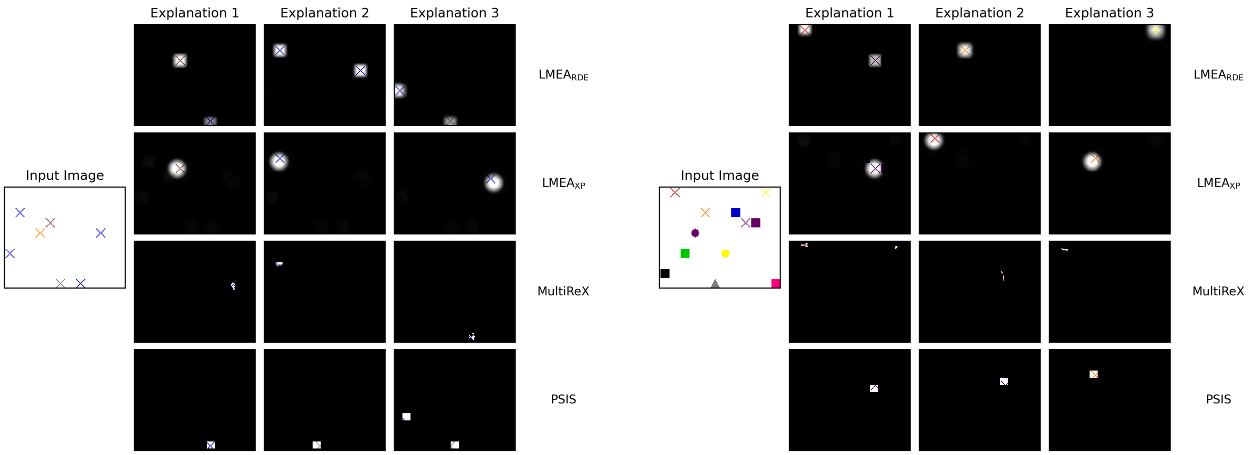
One may wonder why we do not set some thresholds on these intersection metrics and evaluate based on the precision/recall scores at these thresholds, as is done in, e.g., object detection (Padilla et al., 2021). Because we are explaining the prediction process of a classifier (as opposed to an object detection or segmentation model) it is unreasonable to expect that each explanation will overlap exactly with the ground truth annotations. For example, if an image is classified as "dog," the pixels corresponding to the head or tail may be sufficient for the classifier to make its decision, explanations which would have small IoGT and IoU values. By the same token, a classifier may make use of correlated features lying outside the ground truth annotations, which when included in an explanation reduce the IoE, IoGT, and IoU scores. That said, it is reasonable to expect that MSSs for accurate classifiers will overlap substantially with the objects in the image that correspond to the class label, so we use these scores as a measure of quality and reliability of the explanations produced by MEMs. Here, we simply report the mean scores and their associated standard deviations rather than deciding on a specific threshold value, which is difficult to justify.

For each dataset, we draw a held-out set from the training set (separate from the existing validation and test sets), which we use for MSO hyperparameter selection. For comparison between LMEA and MultiReX, we run each explanation on every sample from the test set that has a confident positive prediction, that is, an estimated positive probability of at least 0.8.

6.5 Synthetic dataset

First, we investigate the recovery rate of MSSs when the data generating process is fully known. To do this, we use Teneggi et al. (2022)'s synthetic shapes dataset, which consists of images of colorful shapes on a grid over a white background. Each shape (circle, square, triangle, "X") occupies a 10×10 patch of the size-100×120 image, and there is a random number of such shapes per image. The dataset is labeled such that $y^{(i)} = 1$ if there is at least one "X" shape present in image $\mathbf{x}^{(i)}$, else $y^{(i)} = 0$. We expect that the explanations recovered by each MEM should consist of some portion of an "X," and that the pixels of each "X" should be included in the union of all explanations output by each MEM.

Following a similar setup to Teneggi et al. (2022), we train a small convolutional neural network (a layerwise-downsized version of (Chollet, 2021, Listing 8.1)) on this dataset to 100% test accuracy. For the RDE MSO, we set the size of the smoothmax kernel to 11×11 pixels. For full details on the setup, including classifier training and RDE hyperparameters, see Appendix G.1. For this method, LMEA and PSIS are run for a



(a)

(b)

Figure 2: Sample runs of LMEA and baselines on images from the shapes dataset, where the model is trained to predict the presence of at least one "X". Black regions are unselected by each MEM. *Left*: without SBP. *Right*: with SBP. *Top row of each subfigure:* LMEA with RDE MSO. *Second row of each subfigure:* LMEA with XP MSO. *Bottom two rows of each subfigure:* MultiReX and PSIS. Further examples are provided in Appendix E.

maximum of ten iterations, meaning they can find up to ten explanations.[9] Randomly sampled runs of LMEA and MultiReX on this dataset are pictured in Figure 2.[10] The numerical results for this dataset are reported in Table 1. For this dataset, we use the provided bounding box annotations for evaluation. We find that, most of the time, MultiReX selects pieces of each "X" as an explanation, while LMEA with the RDE MSO selects an entire "X" or multiple "X"s as an explanation at each step, and LMEA with the XP MSO tends to select a single, complete "X" each time. PSIS tends to select patches of each "X". Because they select smaller regions, MultiReX and PSIS tend to find more explanations on this dataset. However, LMEA has a higher IoGT score, indicating that it does a better job of recovering all relevant pixels.

These results are expected, since MultiReX operates by masking images with a (scalar) background value, which in these experiments corresponds to the white background. That is, its naive masking strategy is a perfect match for this simple dataset. RDE, on the other hand, uses a mismatched background value corresponding to the empirical mean image computed over the training set. RDE and XP are also constrained to select larger regions due to their smoothing kernels. For this dataset, the SBP substantially improves MultiReX's IoE score. Aside from reducing the number of explanations found, it does not otherwise hurt performance. The results show that LMEA improves the ability of RDE and XP to find multiple explanations, as measured by Expl./Obj. and IoGT metrics.

6.6 BBBC041 dataset

Next, we examine the performance of LMEA on the task of detecting malaria-infected cells in blood smears. For this experiment we use the BBBC041 dataset (Ljosa et al., 2012).[11] The dataset comes with expert annotations for each cell type, including healthy cells (red blood cells, leukocytes) and infected cells (rings, trophozoites, schizonts, gametocytes). Following Teneggi et al. (2022), we consider the problem of classifying each blood smear image based on whether or not at least one trophozoite cell is present. We use Teneggi et al. (2022)'s dataset for this task which has labeled $y^{(i)} = 1$ if a trophozoite is present in the image $\mathbf{x}^{(i)}$, else $y^{(i)} = 0$, and we follow their model architecture and training setup. Specifically, we fine-tune a ResNet18

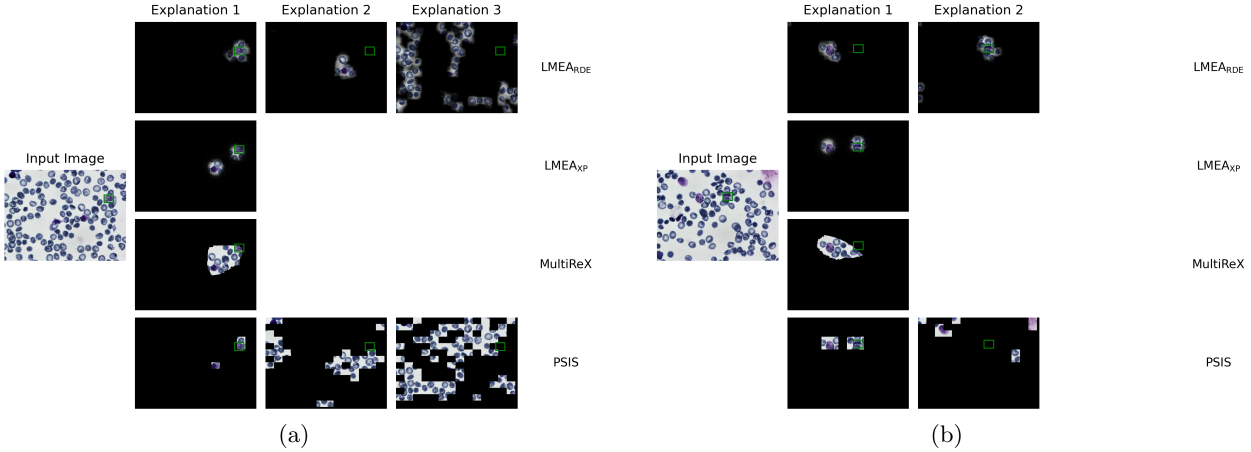
[9] MultiReX does not have an analagous hyperparameter. For details on MultiReX's hyperparameter settings, which are constant across experiments, see Appendix D.

[10] For details on this random sampling procedure, see Appendix F.

[11] The original dataset is available here: https://bbbc.broadinstitute.org/BBBC041, but we use a re-split version of the pre-processed dataset provided by Teneggi et al. (2022) as described in G.3.

Table 1: Results on the Shapes dataset. The table is grouped into single-MSS methods (MSOs) and multi-MSS methods (LMEA, MultiReX, and PSIS). The RDE and XP columns indicate the performance of those MSOs on their own, while the LMEA$_{\text{RDE}}$ and LMEA$_{\text{XP}}$ columns indicate the performance of LMEA with those choices of MSO. The table reports results both with and without the SBP procedure. Where applicable, each metric is reported alongside its standard deviation (in parentheses), and the result with the best mean value in each (single-/multi-MSS) category is bolded. The Expl./Obj. metric is not bolded, and the IoU and IoGT metrics are not bolded for single-MSS methods, for reasons explained in Section 6.4.

			Single-MSS		**Multi-MSS**			
	Metric		**RDE**	**XP**	**LMEA$_{\text{RDE}}$**	**LMEA$_{\text{XP}}$**	**MultiReX**	**PSIS**
Without SBP	Expl./Obj.		0.31 (0.27)	0.31 (0.27)	0.63 (0.16)	0.93 (0.10)	1.85 (0.74)	2.28 (1.52)
	IoU	↑	0.30 (0.15)	0.19 (0.08)	**0.55** (0.05)	0.36 (0.02)	0.14 (0.09)	0.31 (0.09)
	IoGT	↑	0.41 (0.29)	0.32 (0.26)	**0.95** (0.07)	0.92 (0.08)	0.34 (0.10)	0.72 (0.21)
	IoE	↑	**0.60** (0.09)	0.39 (0.06)	**0.57** (0.05)	0.37 (0.02)	0.23 (0.20)	0.39 (0.12)
	Sparsity	↓	**1.64** (0.54)	1.85 (0.05)	1.87 (0.59)	1.81 (0.22)	1.11 (2.42)	**0.69** (0.41)
	Missing	↓	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
With SBP	Expl./Obj.		0.31 (0.27)	0.31 (0.27)	0.59 (0.20)	0.93 (0.11)	1.01 (0.57)	1.81 (0.63)
	IoU	↑	0.29 (0.15)	0.19 (0.08)	**0.50** (0.10)	0.36 (0.02)	0.18 (0.10)	0.36 (0.07)
	IoGT	↑	0.41 (0.29)	0.32 (0.26)	0.84 (0.21)	**0.92** (0.09)	0.19 (0.12)	0.68 (0.19)
	IoE	↑	**0.60** (0.09)	0.39 (0.07)	0.57 (0.06)	0.37 (0.03)	**0.86** (0.12)	0.45 (0.07)
	Sparsity	↓	**1.60** (0.52)	1.84 (0.11)	1.76 (0.51)	1.81 (0.15)	**0.16** (0.07)	0.58 (0.20)
	Missing	↓	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**



(a)

(b)

Figure 3: Sample runs of LMEA and baselines on images from the BBBC041 dataset, where the classifier is trained to predict the presence of at least one trophozoite in the image. Green squares indicate the bounding box labels corresponding to infected (trophozoite) cells. *Left*: without SBP. *Right*: with SBP. *Top row of each subfigure:* LMEA with RDE MSO. *Middle row of each subfigure:* LMEA with XP MSO. *Bottom two rows of each subfigure:* MultiReX and PSIS. Further examples are provided in Appendix E.

Table 2: Results on the BBBC041 dataset. The table is grouped into single-MSS methods (MSOs) and multi-MSS methods (LMEA, MultiReX, and PSIS). The RDE and XP columns indicate the performance of those MSOs on their own, while the LMEA_{RDE} and LMEA_{XP} columns indicate the performance of LMEA with those choices of MSO. The table reports results both without and with the SBP procedure. Where applicable, each metric is reported alongside its standard deviation (in parentheses), and the result with the best mean value in each (single-/multi-MSS) category is bolded. Samples where LMEA/RDE failed to find any explanation were excluded from the corresponding metric calculations, with the exception of the Expl./Obj. metric. The Expl./Obj. metric is not bolded, and the IoU and IoGT metrics are not bolded for single-MSS methods, for reasons explained in Section 6.4.

			Single-MSS		**Multi-MSS**			
	Metric		**RDE**	**XP**	$\textbf{LMEA}_{\textbf{RDE}}$	$\textbf{LMEA}_{\textbf{XP}}$	**MultiReX**	**PSIS**
Without SBP	Expl./Obj.		0.50 (0.32)	0.50 (0.32)	1.26 (0.80)	0.77 (0.54)	0.71 (0.47)	1.49 (1.08)
	IoU	↑	0.27 (0.15)	0.27 (0.16)	0.11 (0.09)	**0.31** (0.14)	0.11 (0.12)	0.08 (0.07)
	IoGT	↑	0.57 (0.32)	0.58 (0.34)	**0.90** (0.12)	0.80 (0.21)	0.69 (0.30)	0.85 (0.11)
	IoE	↑	0.35 (0.18)	**0.39** (0.20)	0.12 (0.10)	**0.36** (0.17)	0.13 (0.17)	0.08 (0.09)
	Sparsity	↓	3.69 (0.96)	**3.63** (2.47)	9.21 (8.38)	**3.91** (2.43)	17.23 (14.29)	12.67 (15.13)
	Missing	↓	**0.00**	3.53	**0.00**	3.53	**0.00**	**0.00**
With SBP	Expl./Obj.		0.50 (0.32)	0.50 (0.32)	0.75 (0.49)	0.73 (0.57)	0.53 (0.31)	0.74 (0.52)
	IoU	↑	0.26 (0.15)	0.27 (0.16)	0.26 (0.12)	**0.29** (0.14)	0.13 (0.14)	0.28 (0.11)
	IoGT	↑	0.56 (0.33)	0.58 (0.34)	**0.74** (0.24)	0.72 (0.28)	0.54 (0.35)	0.61 (0.20)
	IoE	↑	0.35 (0.19)	**0.39** (0.20)	0.29 (0.15)	0.37 (0.18)	0.20 (0.24)	**0.39** (0.17)
	Sparsity	↓	3.67 (0.99)	**3.63** (2.47)	3.97 (1.15)	3.54 (2.16)	12.67 (10.37)	**2.74** (1.35)
	Missing	↓	**0.00**	3.53	**0.00**	3.53	**0.00**	**0.00**

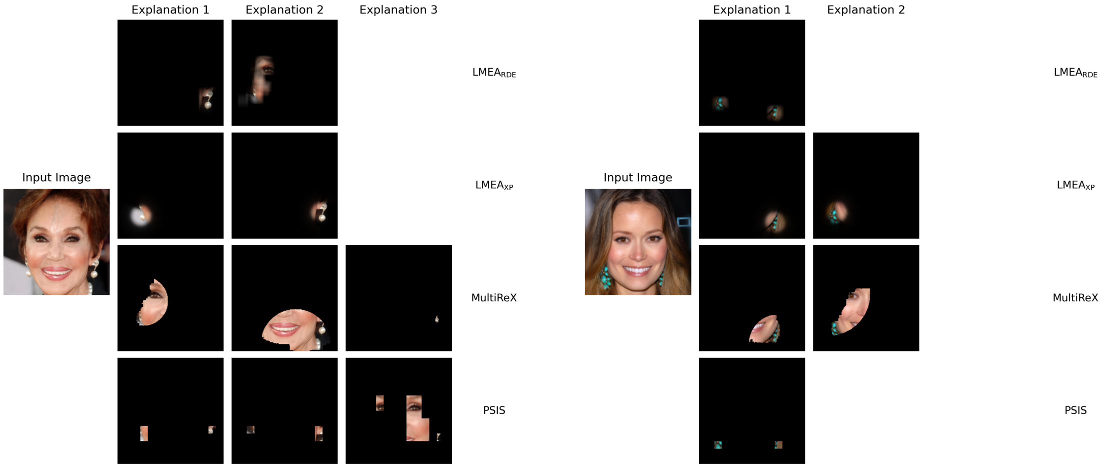
network (He et al., 2016) that was pre-trained on ImageNet (Deng et al., 2009). The trained model achieves a test accuracy of 93.42%. For further details on the experimental setup, see Appendix G.1.

For this dataset, LMEA and PSIS are run for a maximum of five iterations, meaning they can find up to five explanations. We provide examples of LMEA and MultiReX explanations in Figure 3. We find that both methods tend to select trophozoite cells, but often also include other cells in their explanations. These other cells are usually visually similar or belong to a neighborhood surrounding an infected cell. Quantitative results for this dataset are provided in Table 2. Without SBP, LMEA is competitive with baselines across overlap metrics. However, LMEA with the XP MSO finds explanations that are more specific to known salient ground truth regions, as measured by IoE. Also, without SBP, LMEA's explanations are much sparser on average than the baselines. With SBP, all multi-MSS methods perform similarly on IoU, IoGT, and IoE. However, MultiReX has much larger explanations than other methods. Overall, LMEA improves the number of explanations found, as measured by Expl./Obj. and IoGT metrics. The results also indicate that SBP allows LMEA to extend RDE to recover multiple explanations without sacrificing much specificity, as measured by IoE, and that SBP improves the sparsity and specificity of PSIS explanations.

6.7 CelebAMask-HQ dataset

We also apply LMEA to natural images, namely portraits of celebrity faces. We use the CelebAMask-HQ dataset (Lee et al., 2020) for this task. This dataset comes with segmentation annotations for a number of attributes, including physical characteristics and attire. Some of these, such as "wearing earrings" come in pairs. For this experiment, we choose this "wearing earrings" annotation and study whether LMEA can recover both earring regions: intuitively, each explanation should contain (a portion of) one earring if the classifier solves the task perfectly, i.e., if the explanation is an MSS.

Similar to the setup in Section 6.6, we fine-tune a ResNet18 (He et al., 2016) that was pre-trained on ImageNet (Deng et al., 2009). Due to label accuracy issues for the "wearing earrings" label (Wu et al., 2023),



(a)

(b)

Figure 4: Randomly sampled runs of LMEA and baselines without SBP on images from the CelebAMask-HQ dataset, where the classifier is trained to predict the presence of the "wearing earrings" label. *Left*: without SBP. *Right*: with SBP. *Top row of each subfigure:* LMEA with RDE MSO. *Middle row of each subfigure:* LMEA with XP MSO. *Bottom two rows of each subfigure:* MultiReX and PSIS. Further examples are provided in Appendix E.

we use the presence of earring segmentation masks as labels instead of the original dataset labels. The trained model achieves a test accuracy of 89.51%, again with respect to the labels derived from the presence/absence of earring segmentation masks. For full details on model training and MEM hyperparameters, see Appendix G.1. For this method, LMEA and PSIS are run for a maximum of three iterations, meaning they can find up to three explanations. We provide example runs of each MEM for this dataset in Figure 4, and the numerical results in Table 3. Without SBP, we can see that LMEA finds explanations that are more specific to the ground truth regions than baselines, as measured by IoU and IoE metrics. With SBP, LMEA is competitive with the baselines on overlap metrics while providing better coverage of the salient regions of the image, as measured by IoGT. As on the other datasets, the Expl./Obj. and IoGT metrics show that LMEA improves the ability of XP and RDE to find multiple MSSs (although the IoGT improvement is less pronounced on this dataset). Furthermore, SBP improves the IoU and IoE scores of PSIS without worsening other metrics besides Expl./Obj.

6.8 Summary

The experiments show that LMEA recovers more MSSs than the wrapped MSO, and that the explanations output by LMEA are similar in quality to the MSO explanation in terms of sparsity and overlap with ground truth labels, when SBP is applied. We also find that LMEA is competitive with two multiple-MSS explanation baselines, MultiReX and PSIS, in terms of number and quality of explanations recovered. Our results further demonstrate that SBP improves the specificity of both LMEA and PSIS without sacrificing other metrics besides the number of explanations found.

7 Limitations

The main advantage of LMEA is its simplicity, but this simplicity comes with certain limitations. It is not guaranteed that ε-MSSs will be disjoint, and, as discussed in Section 5.2, LMEA's performance will likely suffer in settings where ε-MSSs overlap substantially. However, for MSOs based on optimizing a smooth attribution mask, such as XP and RDE, overlapping explanations can be obtained by reducing the morphological dilation radius mentioned in Section 6.3, or omitting the dilation step entirely. The diversity

Table 3: Results on the CelebAMask-HQ dataset. The table is grouped into single-MSS methods (MSOs) and multi-MSS methods (LMEA, MultiReX, and PSIS). The RDE and XP columns indicate the performance of those MSOs on their own, while the LMEA_{RDE} and LMEA_{XP} columns indicate the performance of LMEA with those choices of MSO. The table reports results both without and with the SBP procedure. Where applicable, each metric is reported alongside its standard deviation (in parentheses), and the result with the best mean value in each (single-/multi-MSS) category is bolded. Samples where LMEA/RDE failed to find any explanation were excluded from the corresponding metric calculations, with the exception of the "Expl./Obj." metric. The Expl./Obj. metric is not bolded, and the IoU and IoGT metrics are not bolded for single-MSS methods, for reasons explained in Section 6.4.

			Single-MSS		**Multi-MSS**			
	Metric		**RDE**	**XP**	**LMEA_{RDE}**	**LMEA_{XP}**	**MultiReX**	**PSIS**
Without SBP	Expl./Obj.		0.62 (0.22)	0.62 (0.22)	1.11 (0.46)	1.15 (0.43)	1.24 (0.50)	1.45 (0.52)
	IoU	↑	0.14 (0.11)	0.12 (0.08)	0.10 (0.09)	**0.12** (0.10)	0.03 (0.03)	0.06 (0.06)
	IoGT	↑	0.55 (0.30)	0.51 (0.28)	0.71 (0.25)	**0.76** (0.24)	0.38 (0.30)	0.69 (0.26)
	IoE	↑	**0.19** (0.16)	0.16 (0.14)	0.12 (0.12)	**0.13** (0.12)	0.03 (0.04)	0.07 (0.08)
	Sparsity	↓	**4.21** (1.78)	4.25 (2.37)	6.19 (4.39)	**5.24** (3.59)	8.60 (8.32)	10.52 (11.36)
	Missing	↓	0.88	**0.22**	0.88	0.22	**0.00**	**0.00**
With SBP	Expl./Obj.		0.62 (0.22)	0.62 (0.22)	0.88 (0.39)	0.94 (0.30)	0.81 (0.33)	0.95 (0.45)
	IoU	↑	0.14 (0.11)	0.12 (0.08)	0.13 (0.10)	**0.14** (0.10)	0.03 (0.04)	0.10 (0.09)
	IoGT	↑	0.55 (0.30)	0.51 (0.28)	0.66 (0.28)	**0.72** (0.27)	0.26 (0.28)	0.47 (0.29)
	IoE	↑	**0.19** (0.16)	0.16 (0.13)	0.16 (0.15)	**0.17** (0.13)	0.06 (0.11)	0.14 (0.15)
	Sparsity	↓	**4.20** (1.79)	4.23 (2.36)	4.56 (2.11)	**4.07** (2.24)	6.02 (4.82)	5.68 (5.39)
	Missing	↓	0.88	**0.22**	0.88	0.22	**0.00**	**0.00**

of the resulting explanations can then be controlled by a postprocessing pruning step, as in (Shitole et al., 2021).

Algorithm 1 assumes that there is a way to specify the active set $\mathcal{A}$ to the MSO, but this is atypical. In this work, we present a practical and effective approach to avoid this limitation in Section 6.3 for two gradient-based MSOs; however, in general some reimplementation of the MSO may be required to restrict the active feature pool $\mathcal{A}$. Furthermore, for some MSOs, there may be more natural ways to find multiple MSSs. For instance, some logic-based explanations such as (Ignatiev et al., 2020) already support enumeration of MSSs.

Finally, our experiments involve MIL datasets which we know *a priori* to have several non-overlapping MSSs, and these may not be reflective of other machine learning scenarios. That said, in other structured domains such as text, we also anticipate that LMEA will succeed in extending existing single-MSS methods, and these settings are left as a matter of future work.

8 Conclusion

In this work, we presented several complementary perspectives on the non-uniqueness of MSSs, including the multiplicity of (probabilistic) prime implicants, the inability of intersection properties of (context-specific) CI to guarantee uniqueness, the breakdown of the MSS uniqueness property for linear models once exact sufficiency ($\varepsilon = 0$) is relaxed to approximate sufficiency ($\varepsilon > 0$), and the potential for degenerate distributions to cause multiple MSSs regardless of the model f.

Motivated by the ubiquitous nature of MSS multiplicity, we proposed a meta-algorithm, LMEA, that generalizes previously studied approaches (Carter et al., 2019; 2021; Byra & Skibbe, 2025) and showed that like (Carter et al., 2019), it recovers a representative set of all ε-MSSs. Experiments on three MIL datasets benchmarked LMEA against prior approaches and demonstrated the utility of LMEA for extending two MSOs that operate via gradient descent on attribution masks.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Salim I Amoukou and Nicolas Brunel. Consistent sufficient explanations and minimal local rules for explaining the decision of any classifier or regressor. *Advances in Neural Information Processing Systems*, 35: 8027–8040, 2022.

Shahaf Bassan, Yizhak Yisrael Elboher, Tobias Ladner, Matthias Althoff, and Guy Katz. Explaining, fast and slow: Abstraction and refinement of provable explanations. *arXiv preprint arXiv:2506.08505*, 2025.

Beepul Bharti, Paul Yi, and Jeremias Sulam. Sufficient and necessary explanations (and what lies in between). In *The Second Conference on Parsimony and Learning (Proceedings Track)*, 2025. URL `https://openreview.net/forum?id=H43BmpeJII`.

Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller. Context-specific independence in bayesian networks. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI '96, pp. 115–123, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Marc Brinner and Sina Zarrieß. Model interpretability and rationale extraction by input mask optimization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13722–13744, 2023.

Michal Byra and Henrik Skibbe. Generating visual explanations from deep networks using implicit neural representations. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3310–3319, 2025. doi: 10.1109/WACV61041.2025.00327.

Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. The struggles of feature-based explanations: Shapley values vs. minimal sufficient subsets, 2020.

Brandon Carter, Jonas Mueller, Siddhartha Jain, and David Gifford. What made you do this? understanding black-box decisions with sufficient input subsets. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 567–576. PMLR, 2019.

Brandon Carter, Siddhartha Jain, Jonas W Mueller, and David Gifford. Overinterpretation reveals image classification model pathologies. *Advances in Neural Information Processing Systems*, 34:15395–15407, 2021.

Aditya Chattopadhyay, Stewart Slocum, Benjamin D Haeffele, René Vidal, and Donald Geman. Interpretable by design: Learning predictors by composing interpretable queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7430–7443, 2022.

Hana Chockler, David A Kelly, and Daniel Kroening. Multiple different black box explanations for image classifiers. *arXiv preprint arXiv:2309.14309*, 2023.

François Chollet. *Deep learning with Python.* Simon and Schuster, 2021.

Kai Lai Chung. *A Course in Probability Theory.* Elsevier, 3 edition, 2001.

Jukka Corander, Antti Hyttinen, Juha Kontinen, Johan Pensar, and Jouko Väänänen. A logical approach to context-specific independence. *Annals of Pure and Applied Logic*, 170(9):975–992, 2019.

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.

Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In *ECAI 2020*, pp. 712–720. IOS Press, 2020.

Adnan Darwiche and Chunxi Ji. On the computation of necessary and sufficient explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5582–5591, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on algebraic statistics*. Springer, 2009.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2950–2958, 2019.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437, 2017.

Mauricio Noris Freire and Leandro Nunes de Castro. e-recruitment recommender systems: a systematic review. *Knowledge and Information Systems*, 63:1–20, 2021.

Aditya Grover, Eric Wang, Aaron Zweig, and Stefano Ermon. Stochastic optimization of sorting networks via continuous relaxations. *arXiv preprint arXiv:1903.08850*, 2019.

Joseph Y Halpern. *Actual Causality*. MIT Press, 2019.

Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang. Image analysis using mathematical morphology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(4):532–550, 1987.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On efficiently explaining graph-based classifiers. *arXiv preprint arXiv:2106.01350*, 2021.

Alexey Ignatiev and Joao Marques-Silva. Sat-based rigorous explanations for decision lists. In *Theory and Applications of Satisfiability Testing–SAT 2021: 24th International Conference, Barcelona, Spain, July 5-9, 2021, Proceedings 24*, pp. 251–269. Springer, 2021.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On relating explanations and adversarial examples. *Advances in neural information processing systems*, 32, 2019.

Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and Joao Marques-Silva. From contrastive to abductive explanations and back again. In *International Conference of the Italian Association for Artificial Intelligence*, pp. 335–355. Springer, 2020.

Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On explaining decision trees. *arXiv preprint arXiv:2010.11034*, 2020.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Assaf Klein and Solomon Eyal Shimony. Discovery of context-specific markov blankets. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 4, pp. 3833–3838. IEEE, 2004.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637, 2012.

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

Ronny Luss and Amit Dhurandhar. When stability meets sufficiency: Informative explanations that do not overwhelm. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=8JNXOB6FtW`.

Jan MacDonald, Stephan Wäldchen, Sascha Hauch, and Gitta Kutyniok. A rate-distortion framework for explaining neural network decisions. *arXiv preprint arXiv:1905.11092*, 2019.

Joao Marques-Silva. Disproving xai myths with formal methods–initial results. In *2023 27th International Conference on Engineering of Complex Computer Systems (ICECCS)*, pp. 12–21. IEEE, 2023.

Joao Marques-Silva, Thomas Gerspacher, Martin Cooper, Alexey Ignatiev, and Nina Narodytska. Explaining naive bayes and other linear classifiers with polynomial time and delay. *Advances in Neural Information Processing Systems*, 33:20590–20600, 2020.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

Rafael Padilla, Wesley L Passos, Thadeu LB Dias, Sergio L Netto, and Eduardo AB Da Silva. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*, 10(3):279, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf`.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 1558604790.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1):1, 2020.

Walter Rudin. *Principles of Mathematical Analysis.* McGraw Hill, 3 edition, 1976.

K Shailaja, Banoth Seetharamulu, and MA Jabbar. Machine learning in healthcare: A review. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 910–914. IEEE, 2018.

Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. *arXiv preprint arXiv:1805.03364*, 2018.

Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. One explanation is not enough: structured attention graphs for image classification. *Advances in Neural Information Processing Systems*, 34:11352–11363, 2021.

Thorvald Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, 5:1–34, 1948.

Jacopo Teneggi, Alexandre Luster, and Jeremias Sulam. Fast hierarchical games for image explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4494–4503, 2022.

Lyn Thomas, Jonathan Crook, and David Edelman. *Credit scoring and its applications.* SIAM, 2017.

Stephan Wäldchen, Jan Macdonald, Sascha Hauch, and Gitta Kutyniok. The computational complexity of understanding binary classifier decisions. *Journal of Artificial Intelligence Research*, 70:351–387, 2021.

Eric Wang, Pasha Khosravi, and Guy Van den Broeck. Probabilistic sufficient explanations. *arXiv preprint arXiv:2105.10118*, 2021.

Haiyu Wu, Grace Bezold, Manuel Günther, Terrance Boult, Michael C King, and Kevin W Bowyer. Consistency and accuracy of celeba attribute values. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3258–3266, 2023.

Stephan Wäldchen. *Towards Explainable Artificial Intelligence: Interpreting Neural Network Classifiers with Probabilistic Prime Implicants.* PhD thesis, Technischen Universität Berlin, 2022.

A Proof that MSSs correspond to context-specific CI relationships

In Section 4.2, we claimed that, if f represents the true conditional probability mass function (pmf),

$$f_y(\mathbf{x}) = f_y^\star(\mathbf{x}) := p(y \mid \mathbf{x}),$$

then any MSS $\mathcal{S}$ for f at $\mathbf{x}$ satisfies $\mathbf{X}_{\mathcal{S}^c} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}}$. While this result is straightforward, the formal statement and proof are included here for completeness.

The following result, which is a simple consequence of the properties of D and the expectation operator $\mathbb{E}_{\mathbf{X}_{\mathcal{S}^c}|\mathbf{x}_{\mathcal{S}}}$, formalizes what is meant by the above CI notation.

Claim 1. *Let $\mathcal{S} \subseteq [d]$ be sufficient for $f = f^\star$ at $\mathbf{x}$. Then $\mathbf{X}_{\mathcal{S}^c} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}}$, that is,*

$$p(y \mid \mathbf{x}) \overset{\text{a.s.}}{=} p(y \mid \mathbf{X}_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}),$$

where "a.s." denotes almost-sure equality with $\mathbf{X}_{\mathcal{S}^c} \sim \mathbb{P}_{\mathbf{X}_{\mathcal{S}^c}|\mathbf{X}_{\mathcal{S}}=\mathbf{x}_{\mathcal{S}}}$.

Proof. Define $\Delta \doteq \Delta(\mathbf{X}_{\mathcal{S}^c}) \doteq D(f(\mathbf{X}_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}), f(\mathbf{x}))$, where $\mathbf{X}_{\mathcal{S}^c} \sim \mathbb{P}_{\mathbf{X}_{\mathcal{S}^c}|\mathbf{X}_{\mathcal{S}}=\mathbf{x}_{\mathcal{S}}}$. Recall that we assume $\mathbf{X}$ has a density or pmf $p(\mathbf{x})$, so that $\mathbb{E}_{\mathbf{X}_{\mathcal{S}^c}|\mathbf{x}_{\mathcal{S}}}[\Delta]$ is a well defined real number. By our assumption that D is nonnegative in Section 2, we have $\Delta \geq 0$. Sufficiency of $\mathcal{S}$ implies $\mathbb{E}_{\mathbf{X}_{\mathcal{S}^c}|\mathbf{x}_{\mathcal{S}}}[\Delta] = 0$. Thus $\Pr(\Delta = 0 \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}}) = 1$ (Chung, 2001, Exercise 3.2.1, p. 46). Recall from Section 2 that $D(\mathbf{u}, \mathbf{v}) = 0 \iff \mathbf{u} = \mathbf{v}$, so that

$$\Pr\left(D(f(\mathbf{X}_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}), f(\mathbf{x})) = 0 \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}}\right) = \Pr\left(f(\mathbf{X}_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}) = f(\mathbf{x}) \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}}\right) = 1.$$

By hypothesis, $f_y(\mathbf{x}) = p(y \mid \mathbf{x})$ for all $y \in \mathcal{Y}$. Thus, taking $\mathbf{X}_{\mathcal{S}^c} \sim \mathbb{P}_{\mathbf{X}_{\mathcal{S}^c}|\mathbf{X}_{\mathcal{S}}=\mathbf{x}_{\mathcal{S}}}$,

$$p(y \mid \mathbf{x}) \overset{\text{a.s.}}{=} p(y \mid \mathbf{X}_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}) \quad \forall y \in \mathcal{Y}.$$

□

B Proof of Proposition 1

First, we will need the following simple lemma.

Lemma 2. *Suppose that $\mathcal{X} = \mathbb{R}$ and that $\mathbf{X}$ admits a continuous, strictly positive density, i.e., $p(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathbb{R}^d$. Further suppose that f and D are continuous. Then the set $\mathcal{S}$ is sufficient for f at $\mathbf{x}$ if and only if $f(\mathbf{x}_{\mathcal{S}}, \mathbf{x}'_{\mathcal{S}^c}) = f(\mathbf{x})$ for all $\mathbf{x}'_{\mathcal{S}^c} \in \mathbb{R}^{|\mathcal{S}^c|}$.*

Proof. Showing that $f(\mathbf{x}_{\mathcal{S}}, \mathbf{x}'_{\mathcal{S}^c}) = f(\mathbf{x})$ for all $\mathbf{x}'_{\mathcal{S}^c} \in \mathbb{R}^{|\mathcal{S}^c|}$ implies $\mathcal{S}$ is sufficient for f at $\mathbf{x}$ is trivial.

For the other direction, we will use continuity of f and D to first show that $D(f(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}), f(\mathbf{x})) = 0$ for all $\mathbf{x}'_{\mathcal{S}^c}$. From the assumption of sufficiency of $\mathcal{S}$ for f at $\mathbf{x}$, we have

$$\mathop{\mathbb{E}}_{\mathbf{X}_{\mathcal{S}^c}|\mathbf{x}_{\mathcal{S}}}[D(f(\mathbf{X}_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}), f(\mathbf{x}))] = 0.$$

By the assumption of a positive, continuous density $p(\mathbf{x})$, we have

$$\begin{aligned}
&\int_{\mathbb{R}^{|\mathcal{S}^c|}} p(\mathbf{x}'_{\mathcal{S}^c} \mid \mathbf{x}_{\mathcal{S}}) D(f(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}), f(\mathbf{x}))\, \mathrm{d}\mathbf{x}'_{\mathcal{S}^c} = \int_{\mathbb{R}^{|\mathcal{S}^c|}} \frac{p(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}})}{p(\mathbf{x}_{\mathcal{S}})} D(f(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}), f(\mathbf{x}))\, \mathrm{d}\mathbf{x}'_{\mathcal{S}^c} = 0 \\
\implies &\int_{\mathbb{R}^{|\mathcal{S}^c|}} \underbrace{p(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}) D(f(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}), f(\mathbf{x}))}_{\doteq \psi(\mathbf{x}'_{\mathcal{S}^c})}\, \mathrm{d}\mathbf{x}'_{\mathcal{S}^c} = 0.
\end{aligned}$$

Note that ψ is continuous, since $\mathbf{x}'_{\mathcal{S}^c} \mapsto D(f(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}), f(\mathbf{x}))$ and $\mathbf{x}'_{\mathcal{S}^c} \mapsto p(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}})$ are continuous. We have

$$\int_{\mathbb{R}^{|\mathcal{S}^c|}} \psi(\mathbf{z})\, \mathrm{d}\mathbf{z} = 0.$$

By continuity and nonnegativity of ψ, this implies $\psi(\mathbf{z}) = 0$ for all $\mathbf{z} \in \mathbb{R}^{|\mathcal{S}^c|}$.[12] Therefore, for all $\mathbf{x}'_{\mathcal{S}^c} \in \mathbb{R}^{|\mathcal{S}^c|}$, we have

$$\begin{aligned}
&\psi(\mathbf{x}'_{\mathcal{S}^c}) = p(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}) D(f(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}), f(\mathbf{x})) = 0 \\
\implies &D(f(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}), f(\mathbf{x})) = 0 \\
\implies &f(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}) = f(\mathbf{x}),
\end{aligned}$$

where the last step follows from our assumption on D that $D(\mathbf{u}, \mathbf{v}) = 0 \iff \mathbf{u} = \mathbf{v}$. □

We are now ready to prove Proposition 1.

Proof of Proposition 1. If $\mathcal{S}$ is a MSS for f at $\mathbf{x}$, then

$$\mathop{\mathbb{E}}_{\mathbf{X}_{\mathcal{S}^c}|\mathbf{x}_{\mathcal{S}}}[D(f(\mathbf{X}_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}), f(\mathbf{x}))] = 0,$$

and by Lemma 2, this occurs if and only if

$$f(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}) = f(\mathbf{x}), \quad \forall \mathbf{x}'_{\mathcal{S}^c} \in \mathbb{R}^{|\mathcal{S}^c|}. \tag{8}$$

We will define $\mathcal{S} \doteq \{i \in [d] : \mathbf{a}_i \neq \mathbf{0}\}$. We will prove the result in two parts.

Part 1: If g is injective on $\operatorname{span}(A) + \mathbf{b}$, then the MSS is unique and equals the nonzero column indices of A.

By Equation (8),

$$\begin{aligned}
&f(\mathbf{x}'_{\mathcal{S}^c}, \mathbf{x}_{\mathcal{S}}) = f(\mathbf{x}), \quad \forall \mathbf{x}'_{\mathcal{S}^c} \in \mathbb{R}^{|\mathcal{S}^c|} \\
\iff &g(A_{\mathcal{S}}\mathbf{x}_{\mathcal{S}} + A_{\mathcal{S}^c}\mathbf{x}'_{\mathcal{S}^c} + \mathbf{b}) = g(A\mathbf{x} + \mathbf{b}), \quad \forall \mathbf{x}'_{\mathcal{S}^c} \in \mathbb{R}^{|\mathcal{S}^c|} \\
\iff &A_{\mathcal{S}}\mathbf{x}_{\mathcal{S}} + A_{\mathcal{S}^c}\mathbf{x}'_{\mathcal{S}^c} + \mathbf{b} = A_{\mathcal{S}}\mathbf{x}_{\mathcal{S}} + A_{\mathcal{S}^c}\mathbf{x}_{\mathcal{S}^c} + \mathbf{b}, \quad \forall \mathbf{x}'_{\mathcal{S}^c} \in \mathbb{R}^{|\mathcal{S}^c|} \\
\iff &A_{\mathcal{S}^c}(\mathbf{x}'_{\mathcal{S}^c} - \mathbf{x}_{\mathcal{S}^c}) = \mathbf{0}, \quad \forall \mathbf{x}'_{\mathcal{S}^c} \in \mathbb{R}^{|\mathcal{S}^c|} \\
\iff &A_{\mathcal{S}^c}\mathbf{u} = \mathbf{0}, \quad \forall \mathbf{u} \in \mathbb{R}^{|\mathcal{S}^c|} \\
\iff &A_{\mathcal{S}^c} = \mathbf{0}.
\end{aligned}$$

[12] The one-dimensional version of this statement is a common exercise in real analysis (see, e.g., (Rudin, 1976, Chapter 6, Exercise 2)).

Therefore, $\mathcal{S}^c$ is the unique largest set for which Equation (8) holds: any superset thereof is not the complement of a sufficient set. Conclude that $\mathcal{S}$ is the unique MSS.

Part 2: If $g = \text{softmax}$, then the MSS corresponds to the indices of columns of A not proportional to $\mathbf{1}$. We note that for any $c \in \mathbb{R}$, and any $\mathbf{z} \in \mathbb{R}^d$, $g(\mathbf{z} + c\mathbf{1}) = g(\mathbf{z})$. Denote $Q \in \mathbb{R}^{d \times d-1}$ a matrix with orthonormal columns forming a basis for $\text{span}\{\mathbf{1}\}^\perp$. It follows that $I - QQ^\top$ is an orthonormal projection onto $\text{span}\{\mathbf{1}\}$, and therefore for any $\mathbf{z} \in \mathbb{R}^d$, there exists c such that $(I - QQ^\top)\mathbf{z} = c\mathbf{1}$. Hence

$$g(A\mathbf{x} + \mathbf{b}) = g\big(QQ^\top(A\mathbf{x} + \mathbf{b}) + (I - QQ^\top)(A\mathbf{x} + \mathbf{b})\big) = g\big(QQ^\top(A\mathbf{x} + \mathbf{b})\big) = \widetilde{g}\big(\widetilde{A}\mathbf{x} + \widetilde{\mathbf{b}}\big)\big),$$

where $\widetilde{g}(\mathbf{z}) = g(Q\mathbf{z})$, $\widetilde{A} = Q^\top A$, and $\widetilde{b} = Q^\top \mathbf{b}$. We will now show that $\widetilde{g}$ admits a left inverse, and is thus injective. By construction, Q is injective as a linear map, so it suffices to show that there exists a map h such that for any $\mathbf{q} \perp \mathbf{1}$, $(h \circ g)(\mathbf{q}) = \mathbf{q}$, demonstrating injectivity of g on the space orthogonal to $\mathbf{1}$. By construction as an orthogonal projection, $QQ^\top \mathbf{q} = \mathbf{q}$. Denote by $\log(\mathbf{x})_i = \log(x_i)$ the elementwise natural logarithm, and set $\beta \doteq \sum_{i \in [d]} \exp(q_i)$. Observe that

$$\log(g(\mathbf{q}))_i = \log(\exp(q_i)/\beta) = q_i - \log(\beta),$$

and so, $QQ^\top \log(g(\mathbf{q})) = QQ^\top(\mathbf{q} - \log(\beta)\mathbf{1}) = \mathbf{q}$. Hence $h(\mathbf{v}) = QQ^\top \log(\mathbf{v})$ is a left inverse of $g|_{\text{span}\{\mathbf{1}\}^\perp}$. Therefore, $\widetilde{g}$ is a continuous and injective nonlinearity. From part 1 of the proof, it follows that a MSS for $f(\mathbf{x}) = g(A\mathbf{x} + \mathbf{b}) = \widetilde{g}(\widetilde{A}\mathbf{x} + \widetilde{\mathbf{b}})$ is the set of nonzero column indices of $\widetilde{A}$, which are precisely the $i \in [d]$ such that $Q^\top \mathbf{a}_i \neq \mathbf{0}$, i.e., the set of $i \in [d]$ such that $\mathbf{a}_i$ is not a multiple of $\mathbf{1}$, $\mathcal{S} = \{i \in [d] : \forall c \in \mathbb{R}, \mathbf{a}_i \neq c\mathbf{1}\}$. □

C Proofs of Lemma 1 and Proposition 2

Proof of Lemma 1. First, it is trivial to see that if $N \leq d$ the statement is true. Now consider the case where $N > d$. Take the size of the active feature pool $|\mathcal{A}|$ as a progress measure. The sufficient $\mathcal{S}$ found by the MSO at each iteration satisfies $\mathcal{S} \cap \mathcal{A}^c = \emptyset$. Furthermore, by Definition 1, the empty set is not sufficient and thus $\mathcal{S} \neq \emptyset$. At each iteration, $\mathcal{S}$ is removed from $\mathcal{A}$. Therefore $|\mathcal{A}|$ decreases by at least one at each iteration, and LMEA runs for at most $d < \infty$ iterations. □

Proof of Proposition 2. By Lemma 1, LMEA terminates, so a set of explanations $\mathcal{E}$ is returned. Seeking contradiction, suppose that there is some ε-MSS $\mathcal{S}$ such that for all $\mathcal{S}' \in \mathcal{E}$, $\mathcal{S}' \cap \mathcal{S} = \emptyset$. Then $\mathcal{S} \subseteq \mathcal{A}$ in the final iteration of LMEA, since $\mathcal{A} = [d] \setminus \mathcal{U}$. With $N = d$, there are two possibilities: either (i) LMEA ran for d iterations before terminating, or (ii) the MSO returned $\emptyset$ in the last iteration executed. In case (i), $\mathcal{U} = [d]$, so that $\mathcal{A} = [d] \setminus [d] = \emptyset$ which implies $\mathcal{S} \subseteq \mathcal{A} = \emptyset$. By Definition 1, sufficient sets cannot be empty, so we have reached a contradiction. In case (ii), we also have a contradiction: by Assumption 1 and Definition 1, the MSO returns $\emptyset$ if and only if $\mathcal{A}$ contains no sufficient sets. □

D MultiReX: background and hyperparameters

MultiReX operates on an objective landscape defined by a ranking $\boldsymbol{\phi}$ of each pixel's causal contribution toward the classification. This causal responsibility ranking is computed via an iterative refinement procedure with number of refinement steps n_{iter}. Given $\boldsymbol{\phi}$, MultiReX first finds a global explanation $\mathcal{T}_0$, before proceeding to a stochastic local search procedure for further explanations.

For this local search, MultiReX initializes a number s of circles $\mathcal{C}_i \subseteq [d]$, $i \in [s]$, each of radius r, which serve as initial candidate explanations to be refined. We will first describe the execution of MultiReX for a single circle $\mathcal{C}_i$. After initializing $\mathcal{C}_i$, MultiReX checks whether it is sufficient. If not, then $\mathcal{C}_i$ is expanded (its radius increased) by a factor of $1 + \alpha$, and its sufficiency is re-checked. If the expanded circle $(1 + \alpha)\mathcal{C}_i$ is sufficient, then it is added to a set of candidate explanations, and the search for circle i terminates. Otherwise, this process is repeated a maximum of n_{exp} times. If after all of these expansions, $\mathcal{C}_i$ is still not sufficient, it is re-initialized in a neighboring position via an accept-reject method that moves to areas with higher mean responsibility $\frac{1}{|\mathcal{C}_i|} \sum_{i=1}^{|\mathcal{C}_i|} \phi_i$. This process is repeated until a sufficient $\mathcal{C}_i$ is found, or a maximum number of

Table 4: MultiReX hyperparameter settings used in the Shapes and CelebAMask-HQ experiments. The MultiReX package's default values for each hyperparameter are shown in parentheses. A ↑ indicates that a larger value is more generous, while a ↓ indicates that a smaller value is more generous.

$s \uparrow$	$n_{\text{step}} \uparrow$	$r \downarrow$	$\alpha \downarrow$	$n_{\text{exp}} \uparrow$	$\tau \uparrow$	$n_{\text{iter}} \uparrow$
20 (10)	80 (40)	5 (25)	0.2 (0.2)	25 (4)	0.1 (0.0)	100 (20)

steps n_{step} is reached. If the final $\mathcal{C}_i$ is sufficient, then it is pruned to an explanation $\mathcal{T}_i$. The aforementioned procedure is executed for all $\mathcal{C}_i$, leading to a set of at most $s+1$ candidate explanations (including the global explanation) $\{\mathcal{T}_i\}_{i=0}^{k}$, $k \leq s$. Finally, this set of candidate explanations is pruned to minimize overlap; the only hyperparameter for this step is the allowed Sørensen-Dice coefficient (Dice, 1945; Sørensen, 1948) threshold τ between any two explanations. The output of this final stage is a set of MSSs $\{\mathcal{S}_i\}_{i=0}^{k'}$, $k' \leq s$.

The implementation of MultiReX[13] allows the configuration of many hyperparameters. We set each of these to increase the computational budget over the default configuration[14] for a generous comparison. Table 4 reports the hyperparameters used in our experiments, which are held fixed across datasets. We note that, due to rounding in the implementation of the circle expansion step, α is the smallest allowable radius expansion parameter for $r = 5$. MultiReX returns multiple sets of explanations satisfying the overlap bound τ. We pick the one that has the smallest total area, which yields the smallest sparsity metric in our experiments.

E Further examples

F Generation of example figures

To generate the examples in Figures 2, 3, 4, 5, 6, 7, 8, 9, and 10 we used the following procedure. For each experiment, after generating explanations for each image in the test set, we filtered the set of candidate images to those that had at least one explanation for each method (to rule out atypical examples where one of the methods fails to find an explanation). Next, we filtered to those images for which at least one of $\{\text{LMEA}_{\text{RDE}}, \text{LMEA}_{\text{XP}}, \text{MultiReX}, \text{SIS}\}$ had at least two explanations. Finally, we took a random permutation of the images that remained and took the first seven images for each dataset for each of the figures mentioned above. We then picked representatives of these random sets for Figures 2, 3, and 4 in the main body of the paper.

G Experimental details

G.1 Further LMEA implementation details

Explanation binarization. Both RDE and XP MSOs return soft masks $\mathbf{s} \in [0,1]^d$. To convert these into MSSs $\mathcal{S}$, we threshold them according to Otsu's method (Otsu, 1979).

Termination criterion. All of our experiments involve explaining binary classifiers. Therefore, to recover the most explanations possible with LMEA, we dynamically set the sufficiency criterion to $\varepsilon = |f_1(\mathbf{x}) - 1/2|$, where $f_1(\mathbf{x})$ is the model's estimated probability that $Y = 1$ given $\mathbf{X} = \mathbf{x}$. After each call to the MSO, which yields a soft attribution mask $\mathbf{s} \in [0,1]^d$, we additionally check the sufficiency of the solution according to

$$\|f(\mathbf{s} \odot \mathbf{x} + (\mathbf{1} - \mathbf{s}) \odot \mathbf{b}_{\text{suff}}) - f(\mathbf{x})\|_{\text{TV}} \leq \varepsilon,$$

where $\mathbf{b}_{\text{suff}}$ is a background value, which will be set to zero or the empirical mean image computed over the training set, depending on the dataset. These are standard choices of background inputs in the feature attribution literature (Fong et al., 2019; Teneggi et al., 2022). This setup causes LMEA to terminate once the predicted label changes: $1 = f(\mathbf{x}) \neq \text{argmax}_{y \in \mathcal{L}}\, f_y(\mathbf{s} \odot \mathbf{x} + (\mathbf{1} - \mathbf{s}) \odot \mathbf{b}_{\text{suff}})$. If the above condition does

[13] https://github.com/ReX-XAI/ReX

[14] The default arguments reported are those in this file.

not hold, then LMEA exits early. A similar sufficiency check is performed on the active set using its binary characteristic mask vector $\mathbf{1}_{\mathcal{A}}$, as specified in Algorithm 1. We also terminate early whenever an explanation is found that overlaps with previously found explanations, as this tends to correspond to an absence of informative features in $\mathcal{A}$, and LMEA with the XP MSO may find MSSs that overlap with explanations from previous iterations.

Masked stop-gradient. The following simple Python wrapper, based on a similar trick in the NeuralSort (Grover et al., 2019) repository (https://github.com/ermongroup/neuralsort) suffices to prevent re-selection of previously selected features in LMEA for RDE and XP MSOs.

Listing 1 Masked stop-gradient wrapper.

```
class MaskedStopGradient(nn.Module):
    def __init__(self, model, mask):
        super().__init__()
        self._model = model
        self._mask = mask

    def forward(self, x):
        return self._model((1 - self._mask) * x + (self._mask * x).detach())
```

The mask in Listing 1 represents the coordinates in $\mathcal{A}^c$ (`1 - active_set`) that should not be re-selected. Passing `MaskedStopGradient(model, 1 - active_set)` to the MSO means that gradients corresponding to $\mathcal{A}^c$ will not backpropagate, as described in Section 6.3.

G.2 Synthetic dataset

Model training. We trained the simple convolutional network described in section 6.5 using Adam (Kingma, 2014) with learning rate 10^{-3} and batch size of 64 for 40 epochs, keeping the best model checkpoint based on validation (cross-entropy) loss. The images are preprocessed during training by subtracting the background $\mathbf{b} = \mathbf{1}$ and dividing by the approximate channel-wise standard deviations $\widehat{\boldsymbol{\sigma}} = (0.1323, 0.165, 0.17)$.

LMEA. For LMEA sufficiency checks, we used the background value $\mathbf{b}_{\text{suff}} = \mathbf{0}$, described in Section 6.3, which due to feature normalization corresponds to the white background of the shapes images.

RDE. In RDE, we set the smoothmax temperature to $T = 1$. We set the perturbation distribution to $\mathcal{N}(\widehat{\boldsymbol{\mu}}, \frac{1}{8}\widehat{\boldsymbol{\Sigma}})$, where $\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)}$ is the empirical mean image and $\widehat{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}^{(i)} - \widehat{\boldsymbol{\mu}})(\mathbf{x}^{(i)} - \widehat{\boldsymbol{\mu}})^\top$ is the corresponding empirical covariance image computed over the set of training images $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$. To obtain the RDE explanations, we solved Equation (4) on the grid $\lambda \in \{1, 5, 10, 15, 20\}$ via 3000 iterations of Adam with learning rate 10^{-3}.

XP. We use the implementation of XP in the TorchRay package.[15], solving the "preservation" (Fong et al., 2019) (sufficiency) objective with a "fade-to-black" (Fong et al., 2019) (zero) background. Due to feature normalization, this corresponds to setting the unselected pixels to white in the corresponding shapes images. We use the contrastive reward $r(\mathbf{x}) \doteq a_1(\widetilde{\mathbf{x}}) - a_0(\widetilde{\mathbf{x}})$, where $a_1(\widetilde{\mathbf{x}})$ is the activation (logit) corresponding to class one for input $\widetilde{\mathbf{x}}$ and $a_0(\widetilde{\mathbf{x}})$ is the activation (logit) corresponding to class zero for perturbed input $\widetilde{\mathbf{x}}$. This is because

$$
\begin{aligned}
p_1 &= \frac{\exp(r(\widetilde{\mathbf{x}}) + a_0(\widetilde{\mathbf{x}}))}{\exp(a_0(\widetilde{\mathbf{x}})) + \exp(r(\widetilde{\mathbf{x}}) + a_0(\widetilde{\mathbf{x}}))} \\
&= \frac{\exp(r(\widetilde{\mathbf{x}}))}{1 + \exp(r(\widetilde{\mathbf{x}}))} \\
&= \sigma(r(\widetilde{\mathbf{x}})),
\end{aligned}
$$

[15] https://github.com/facebookresearch/TorchRay

where σ is the sigmoid functon. Since σ is monotonically increasing, maximizing r over perturbed inputs $\widetilde{\mathbf{x}}$ is equivalent to maximizing the model's probability of the positive class $f_1(\widetilde{\mathbf{x}})$, which is also equivalent to minimizing the sufficiency objective $\|f(\widetilde{\mathbf{x}}) - \widehat{\mathbf{y}}_{\text{conf}}\|_{\text{TV}}$, where $\widehat{\mathbf{y}}_{\text{conf}} = (1, 0)$ is the perfectly confident positive prediction and $\|\cdot\|_{\text{TV}}$ denotes the TV norm between distributions.

We run the XP solver for 1500 iterations on the grid of area constraint parameters $\alpha \in \{0.05, 0.1, 0.15, 0.2\}$, check ε-sufficiency according to Section 6.3, and choose the ε-sufficient solution with the smallest α. For this dataset, we set the mask smoothing parameter `sigma` to 7, and we also disable the jitter option. All the other parameters are left at their default values.

G.3 BBBC041 dataset

Model training. We fine-tune the pretrained ResNet18 model via 30 epochs of Adam with a learning rate of 10^{-4}, a batch size of 4, and a learning rate decay factor of 0.2 applied after each 10 epochs, similarly to Teneggi et al. (2022) (we train for five more epochs than Teneggi et al. (2022)). In addition to Teneggi et al. (2022)'s random horizontal flip augmentations, we use JPEG compression and random vertical flip augmentations during training. Due to heterogeneity between training, validation, and test sets, we had difficulty training a model to perform well on the test set provided by Teneggi et al. (2022).[16] As a result, we randomly re-split the training and validation subsets of the dataset into training, validation, and test sets. As in the shapes experiment, we keep the best-performing model checkpoint according to the validation loss.

LMEA. For LMEA sufficiency checks, we used the background value $\mathbf{b}_{\text{suff}} = \widehat{\boldsymbol{\mu}}$.

RDE. For this dataset, we set the perturbation distribution to $\mathcal{N}(\widehat{\boldsymbol{\mu}}, \frac{1}{4}\widehat{\boldsymbol{\Sigma}})$, where once again the mean and covariance are estimated from the training set. We use the same grid for the sparsity penalty as in Section 6.5, and solve RDE for each λ via 500 iterations of Adam with learning rate 10^{-2}.

XP. For this dataset, we use the "blur" background, which is a blurred version of the original input image, and we leave the mask smoothing parameter `sigma` at its default value of 21. The rest of the details are the same as in Appendix G.2.

G.4 CelebAMask-HQ dataset

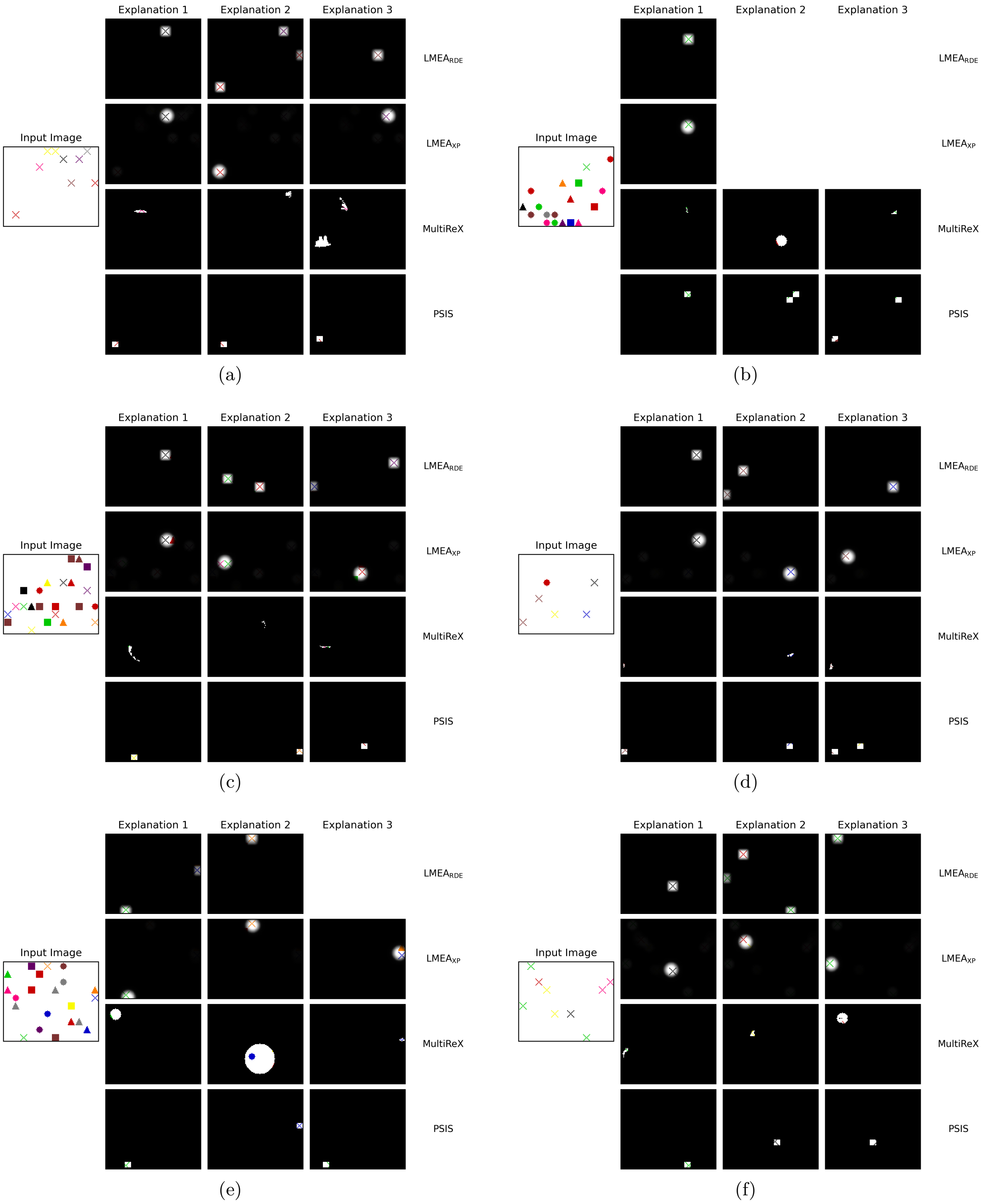
Model training. The ImageNet-pretrained ResNet18 model was fine-tuned for 25 epochs of Adam using a learning rate of 10^{-4} and a batch size of 64. For consistency with the pretrained model's preprocessing pipeline, we resize each image to 256 × 256, center crop to 224 × 224, and normalize by subtracting $\widetilde{\boldsymbol{\mu}} = (0.485, 0.456, 0.406)$ from R, G, and B channels, respectively, and dividing by $\widetilde{\boldsymbol{\sigma}} = (0.229, 0.224, 0.225)$, again, channel-wise. We keep the best model checkpoint according to validation loss.

LMEA. For LMEA sufficiency checks, we used the background value $\mathbf{b}_{\text{suff}} = \widehat{\boldsymbol{\mu}}$.

RDE. We set the perturbation distribution to $\mathcal{N}(\widehat{\boldsymbol{\mu}}, \frac{1}{4}\widehat{\boldsymbol{\Sigma}})$, with empirical (training set) mean and covariance, as before. We again use the same grid for the sparsity penalty as in Section G.4, and solve RDE for each λ via 500 iterations of Adam with learning rate 10^{-2}.

XP. We use the same parameters as in Appendix G.3.

[16] We use the trophozoite dataset provided by Teneggi et al. (2022): https://zenodo.org/records/5914342.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5: Randomly sampled runs of LMEA and baselines without SBP on images from the shapes dataset, where the model is trained to predict the presence of at least one "X". Black regions are unselected by each MEM. *Top row of each subfigure:* LMEA with RDE MSO. *Middle row of each subfigure:* LMEA with XP MSO. *Bottom two rows of each subfigure:* MultiReX and PSIS.

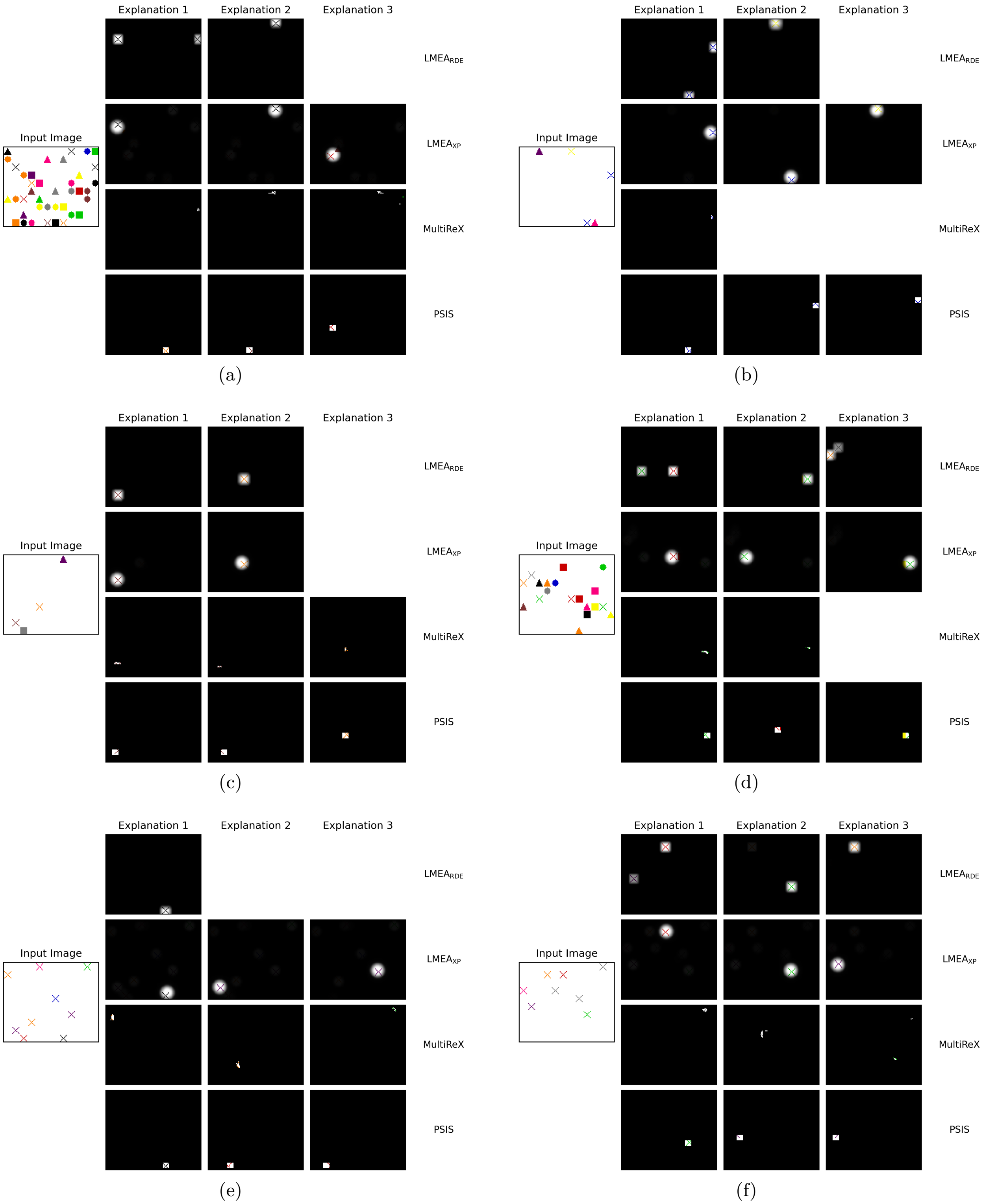
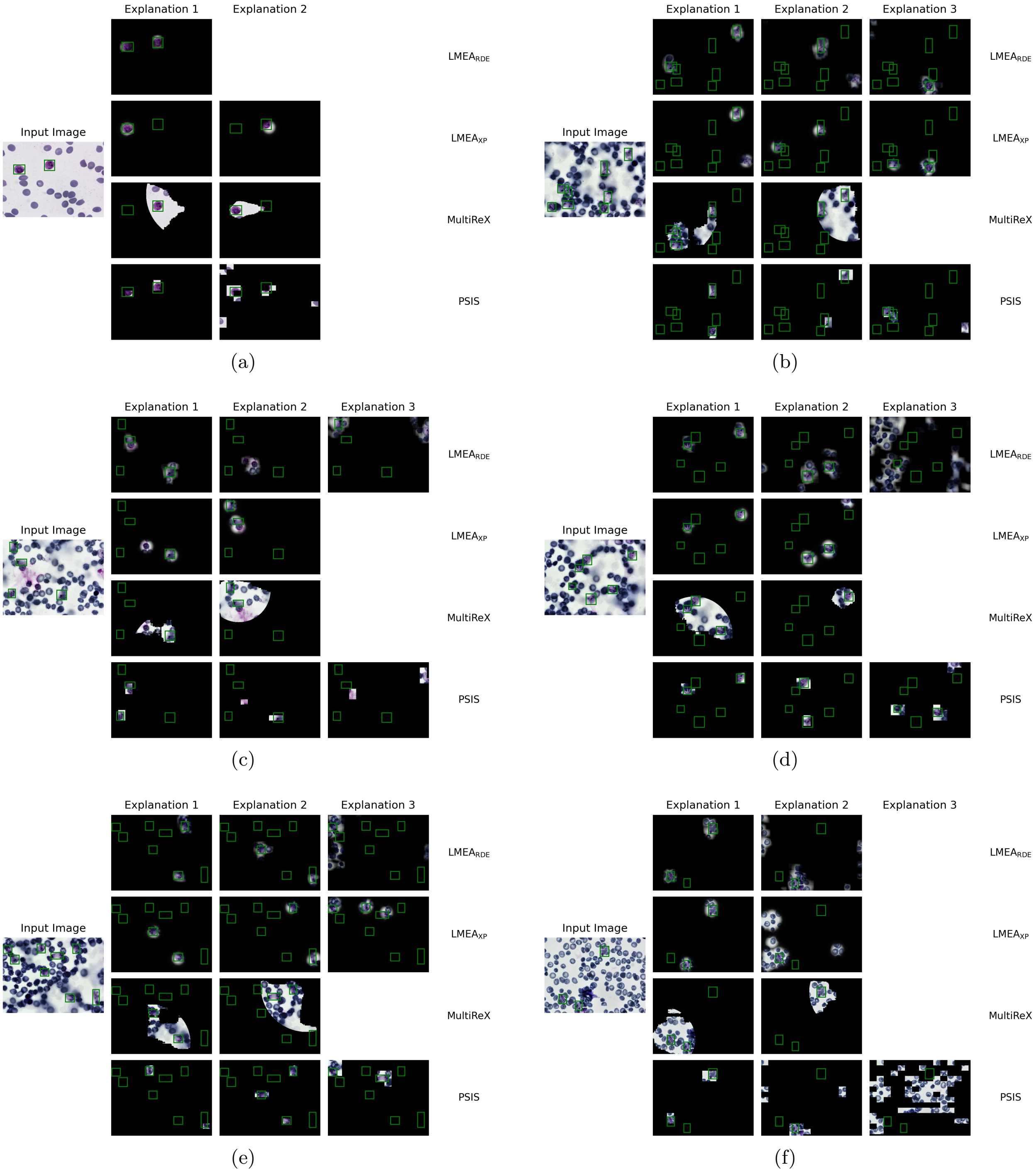


Figure 6: Randomly sampled runs of LMEA and baselines with SBP on images from the shapes dataset, where the model is trained to predict the presence of at least one "X". Black regions are unselected by each MEM. *Top row of each subfigure:* LMEA with RDE MSO. *Middle row of each subfigure:* LMEA with XP MSO. *Bottom two rows of each subfigure:* MultiReX and PSIS.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 7: Randomly sampled runs of LMEA and baselines without SBP on images from the BBBC041 dataset, where the classifier is trained to predict the presence of at least one trophozoite in the image. Green squares indicate the bounding box labels corresponding to infected (trophozoite) cells. *Top row of each subfigure:* LMEA with RDE MSO. *Middle row of each subfigure:* LMEA with XP MSO. *Bottom two rows of each subfigure:* MultiReX and PSIS.

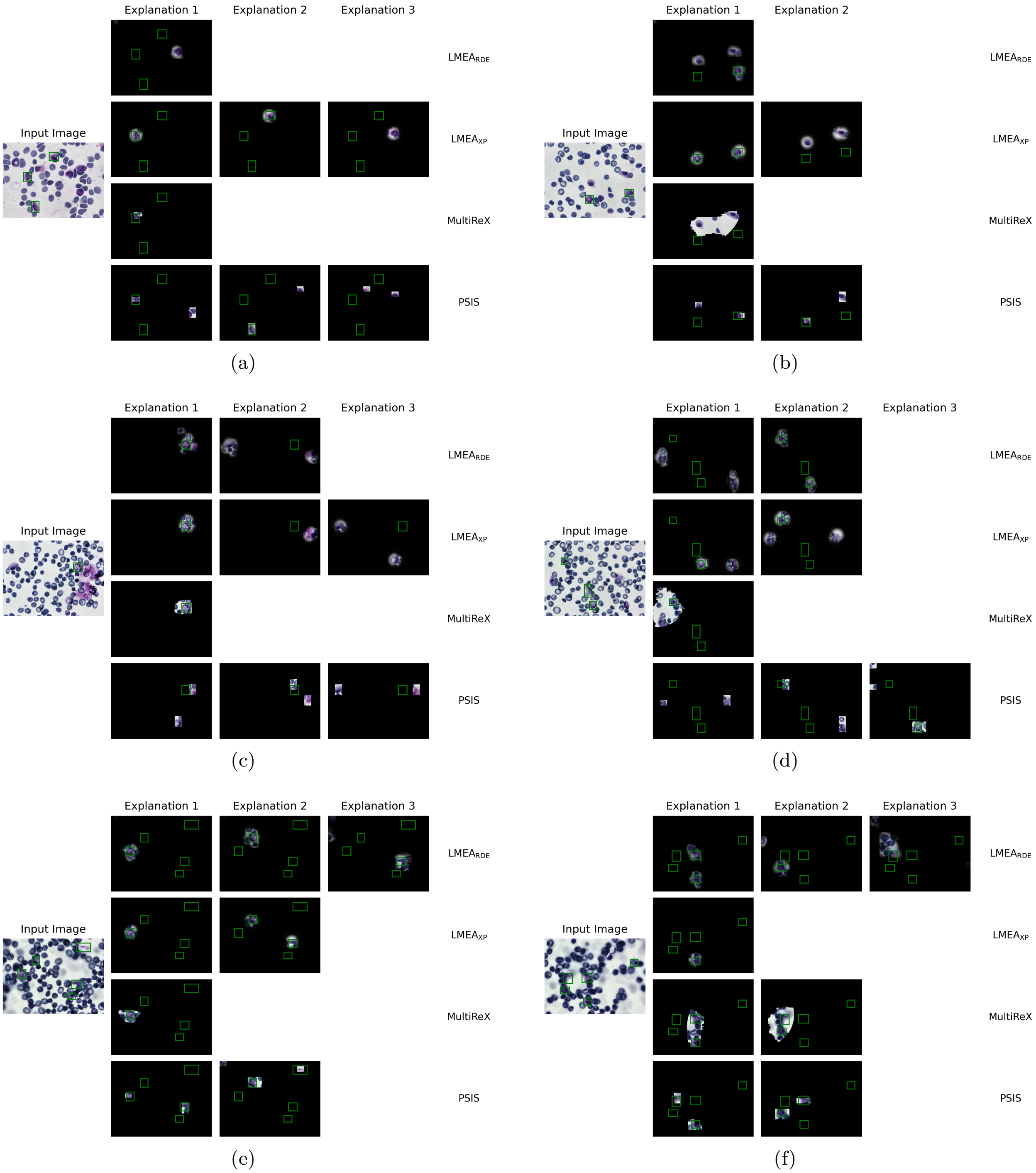
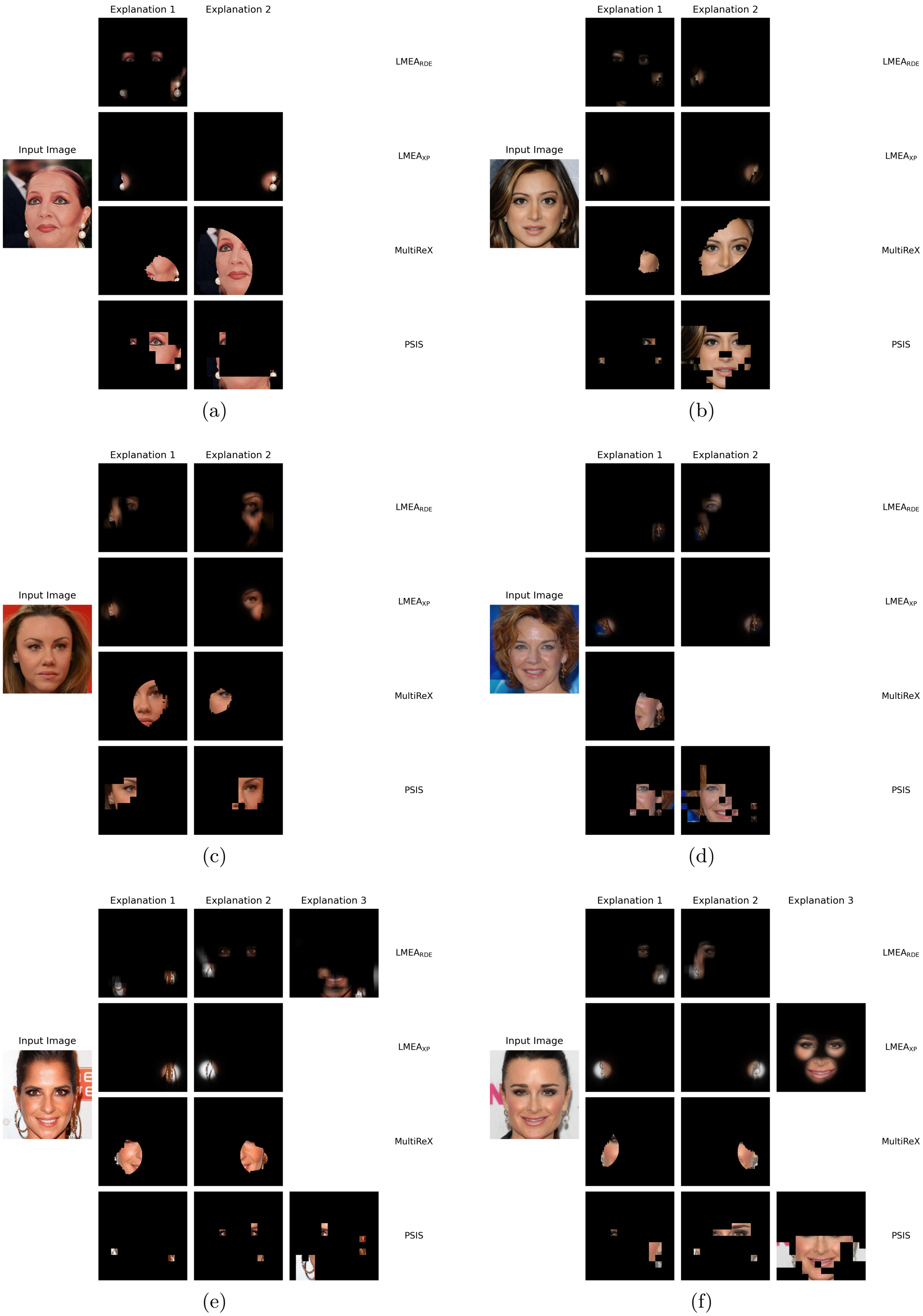
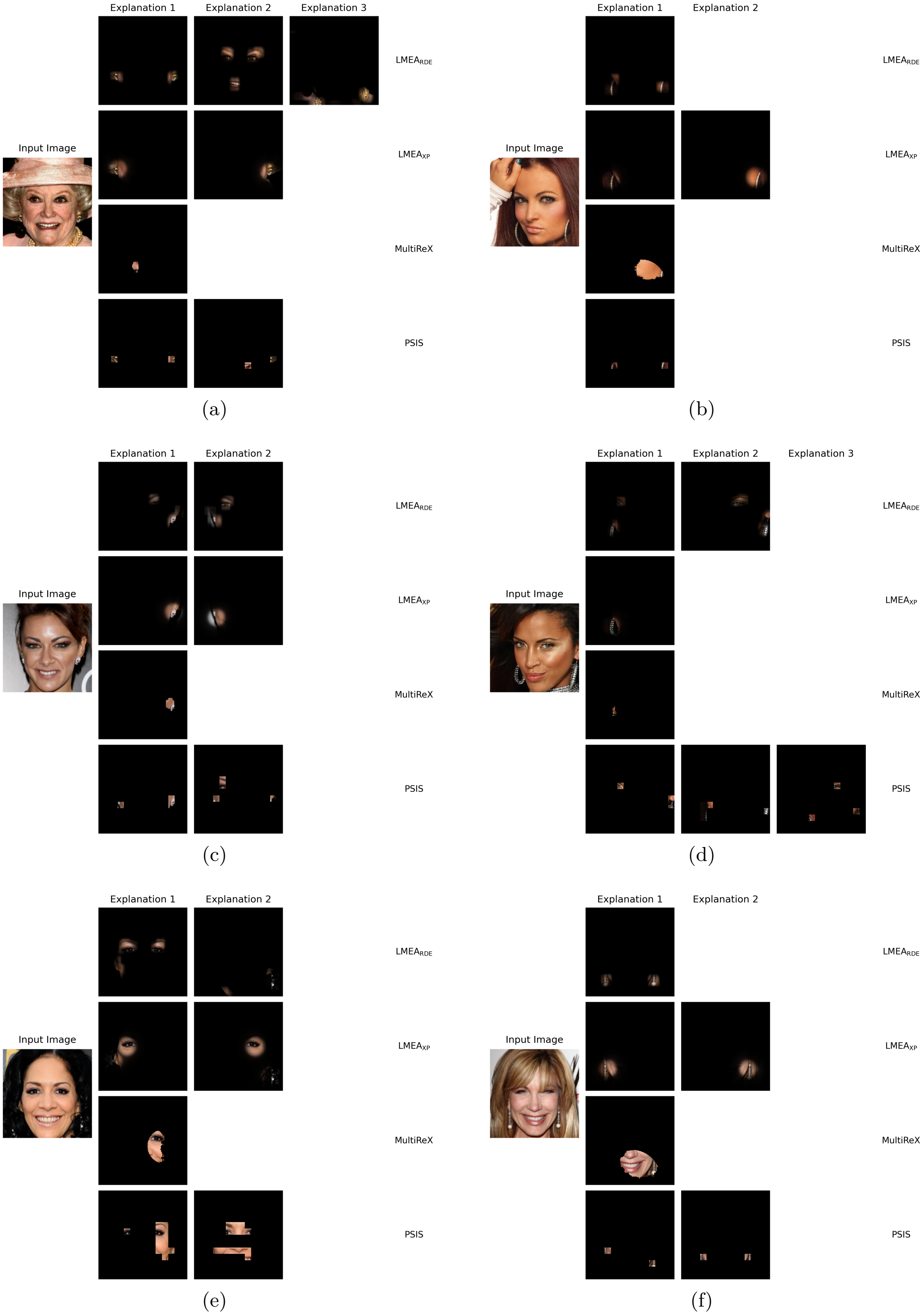


Figure 8: Randomly sampled runs of LMEA and baselines with SBP on images from the BBBC041 dataset, where the classifier is trained to predict the presence of at least one trophozoite in the image. Green squares indicate the bounding box labels corresponding to infected (trophozoite) cells. *Top row of each subfigure:* LMEA with RDE MSO. *Middle row of each subfigure:* LMEA with XP MSO. *Bottom two rows of each subfigure:* MultiReX and PSIS.



(a) (b) (c) (d) (e) (f)

Figure 9: Randomly sampled runs of LMEA and baselines without SBP on images from the CelebAMask-HQ dataset, where the classifier is trained to predict the presence of the “wearing earrings” label. *Top row of each subfigure:* LMEA with RDE MSO. *Middle row of each subfigure:* LMEA with XP MSO. *Bottom two rows of each subfigure:* MultiReX and PSIS.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 10: Randomly sampled runs of LMEA and baselines with SBP on images from the CelebAMask-HQ dataset, where the classifier is trained to predict the presence of the “wearing earrings” label. *Top row of each subfigure:* LMEA with RDE MSO. *Middle row of each subfigure:* LMEA with XP MSO. *Bottom two rows of each subfigure:* MultiReX and PSIS.