DOMINO: DISCOVERING SYSTEMATIC ERRORS WITH CROSS-MODAL EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning models that achieve high overall accuracy often make systematic errors on important subgroups (or *slices*) of data. When working with high-dimensional inputs (*e.g.* images, audio) where important slices are often unlabeled, identifying underperforming slices is a challenging task. In order to address this issue, recent studies have proposed automated slice discovery methods (SDMs), which leverage learned model representations to mine input data for slices on which a model performs poorly. To be useful to a practitioner, these methods must identify slices that are both underperforming and *coherent* (*i.e.* united by a human-understandable concept). However, no quantitative evaluation framework currently exists for rigorously assessing SDMs with respect to these criteria. Additionally, prior qualitative evaluations have shown that SDMs often identify slices that are incoherent. In this work, we address these challenges by first designing a principled evaluation framework that enables a quantitative comparison of SDMs across 1,235 slice discovery settings in three input domains (natural images, medical images, and time-series data). Then, motivated by the recent development of powerful cross-modal representation learning approaches, we present *Domino*, an SDM that leverages cross-modal embeddings and a novel error-aware mixture model to discover and describe coherent slices. We find that Domino accurately identifies 36% of the 1,235 slices in our framework – a 12 percentage-point improvement over prior methods. Further, Domino is the first SDM that can generate natural language descriptions of identified slices, correctly outputting the exact name of the slice in 35% of settings.

1 INTRODUCTION

Machine learning models often make systematic errors on important subsets (or *slices*) of data. For instance, models trained to detect collapsed lungs in chest x-rays have been shown to make predictions based on the presence of chest drains, a device typically used during treatment (Oakden-Rayner et al., 2019). As a result, these models frequently make prediction errors on cases without chest drains, a critical data slice where false negative predictions could be life-threatening. Similar slice performance gaps have been observed in radiograph classification (Badgeley et al., 2019a; Zech et al., 2018a; DeGrave et al., 2021a), melanoma detection (Winkler et al., 2019a), natural language processing (Orr et al., 2020; Goel et al., 2021), and object detection (de Vries et al., 2019b), among others. If underperforming slices can be accurately identified and labeled, we can then improve model robustness by either addressing biases in the training dataset or using robust optimization techniques (Zhang et al., 2018; Sagawa et al., 2020).

However, identifying underperforming slices is difficult in practice. When working with high-dimensional inputs (*e.g.* images, time-series data, video) where individual features (*e.g.* pixels) have little semantic meaning, slices are often "hidden", meaning that they cannot easily be extracted from the inputs and are not captured in metadata (Oakden-Rayner et al., 2019). In this setting, we must perform *slice discovery*: the task of mining unstructured data for semantically meaningful subgroups on which the model performs poorly.

In modern machine learning workflows, practitioners commonly perform slice discovery with a combination of feature-based interpretability methods (e.g. GradCAM, LIME) and manual inspection (Selvaraju et al., 2017; Ribeiro et al., 2016). However, these approaches are time-consuming

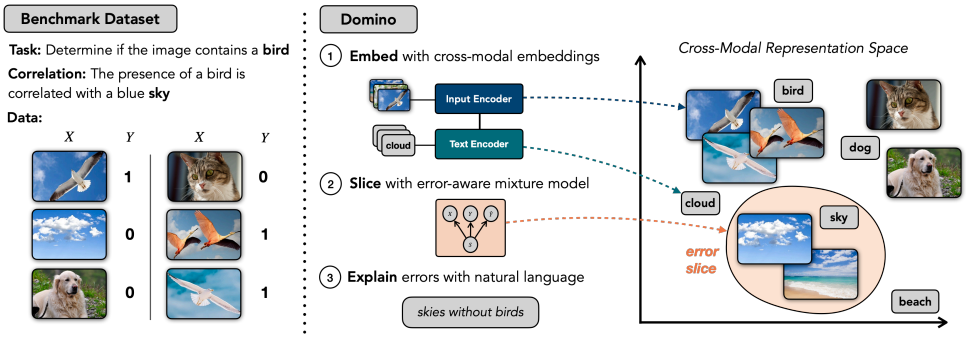

Figure 1: **The Domino Slice Discovery Method**. *Domino* discovers and describes underperforming slices. (Left) In this example, a model is trained and evaluated on the task of detecting birds in images. The dataset shows that a correlation exists between the presence of birds and the presence of the sky. (Right) As a result, the classifier makes false positive predictions on skies without birds. *Domino* uses cross-modal embeddings to identify and describe the error slice.

and susceptible to confirmation bias (Adebayo et al., 2018). As a result, recent works have proposed automated *slice discovery methods* (SDMs), which utilize learned input representations to identify semantically meaningful slices where the model makes prediction errors (d'Eon et al., 2021; Yeh et al., 2020; Sohoni et al., 2020; Kim et al., 2018). An ideal SDM should automatically identify data slices that fulfill two desiderata: (a) slices should contain examples on which the model *underperforms*, or has a high error rate and (b) slices should contain examples that are *coherent*, or align closely with a human-understandable concept. An SDM that is able to reliably satisfy these desiderata across a wide range of settings has yet to be demonstrated for two reasons:

Issue 1: *No quantitative evaluation framework exists for measuring performance of SDMs with respect to these desiderata.* Existing SDM evaluations are either qualitative (d'Eon et al., 2021), performed on purely synthetic data (Yeh et al., 2020), or consider only a small selection of tasks and slices (Sohoni et al., 2020). A comprehensive evaluation framework should be quantitative, use realistic data, cover a broad range of contexts, and evaluate both underperformance and coherence. Currently, no datasets or frameworks exist to support such an evaluation, making it difficult to evaluate the tradeoffs among prior SDMs.

Issue 2: *Prior qualitative evaluations have demonstrated that existing SDMs often identify slices that are incoherent.* A practically useful SDM should discover coherent slices that are understandable by a domain expert. For example, in the chest x-ray setting described earlier, the slice "patients without chest drains" is meaningful to a physician. Slice coherence has previously been evaluated qualitatively by requiring users to manually inspect examples and identify common attributes (d'Eon et al., 2021; Yeh et al., 2020). Such evaluations have shown that discovered slices often do not align with concepts understandable to a domain expert. Additionally, even if slices do align well with concepts, it may be difficult for humans to identify the shared concept. Thus, an ideal SDM would not only output coherent slices, but also identify the concept connecting examples in each slice.

In this work, we address both of these issues by (1) developing a framework to quantitatively evaluate the effectiveness of slice discovery methods at scale and (2) leveraging this framework to demonstrate that a powerful class of recently-developed cross-modal embeddings can be used to create an SDM that satisfies the above desiderata. Our approach – *Domino* – identifies coherent slices and generates automated slice descriptions.

After formally describing the slice discovery problem in Section 2, we introduce a benchmark evaluation framework for rigorously assessing SDM performance in Section 3. We curate a set of over 1235 slice discovery settings, each consisting of a real-world classification dataset, a trained model, and one or more "ground truth" slices corresponding to a meaningful concept in the domain. During evaluation, the SDM is provided with the dataset and the model, and we measure if the labeled slices can be successfully identified. We find that existing methods identify "ground truth" slices in no more than 23% of these settings.

Motivated by the recent development of large cross-modal representation learning approaches (e.g. CLIP) that embed inputs and text in the same latent representation space, in Section 4 we proceed to

design *Domino*, a novel SDM that uses cross-modal embeddings to identify coherent slices. Cross-modal representations incorporate semantic meaning from text into input embeddings, which we demonstrate can improve slice coherence and enable the generation of automated slice descriptions. *Domino* embeds inputs alongside natural language with cross-modal representations, identifies coherent slices with an error-aware Gaussian mixture model, and generates natural language descriptions for discovered slices. In Section 5, we use our evaluation framework to show that Domino identifies 36% of the "ground truth" coherent slices across three input domains (natural images, medical images, and time-series) – a 12 percentage-point improvement over existing methods.

2 Slice Discovery Preliminaries

We consider a standard classification setting with input $X \in \mathcal{X}$ (*e.g.* an image, time-series, or graph) and label $Y \in \mathcal{Y} = \{1, 2, ..., C\}$ over C classes. Additionally, we assume there exists a set of k *slices* $\mathbf{S} = \{S^{(j)}\}_{j=1}^{k} \in \{0, 1\}^k$ that partition the data into *coherent* (potentially overlapping) subgroups, where each subgroup captures a distinct concept or attribute that would be familiar to a domain expert. The slices, inputs, and labels vary jointly according to a probability distribution $P(X, Y, \mathbf{S})$ over $\mathcal{X} \times \mathcal{Y} \times \{0, 1\}^k$. We assume that training, validation and test data are drawn i.i.d. from this distribution. A model $h_\theta : \mathcal{X} \to \mathcal{Y}$ exhibits degraded performance with respect to a slice $S^{(j)}$ and metric $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ if

$$\mathbb{E}_{X,Y|S^{(j)}=1}[\ell(h_\theta(X), Y)] < \mathbb{E}_{X,Y|S^{(j)}=0}[\ell(h_\theta(X), Y)]. \tag{1}$$

Assuming that a trained classifier $h_\theta : \mathcal{X} \to \mathcal{Y}$ exhibits degraded performance on each of the k slices in $\mathbf{S}$, we define the *slice discovery problem* as follows:

- **Inputs**: a trained classifier h_θ and labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ drawn from $P(X, Y)$.
- **Output**: a set of $\hat{k}$ slicing functions $\Psi = \{\psi^{(j)} : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}\}_{j=1}^{\hat{k}}$ that partition the data into $\hat{k}$ subgroups.

We consider an output to be successful if, for each ground truth slice $S^{(u)}$, a slicing function $\psi^{(v)}$ predicts $S^{(u)}$ with precision above some threshold β:

$$\forall u \in [k]. \quad \exists v \in [\hat{k}]. \quad P(S^{(u)} | \psi^{(v)}(X, Y) = 1) > \beta. \tag{2}$$

A *slice discovery method* (SDM), $M(\mathcal{D}, h_\theta) \to \Psi$, aims to solve the slice discovery problem.

3 Slice Discovery Evaluation Framework

We propose a framework that measures the performance of each SDM with respect to the two desiderata outlined in Section 1: (a) the model h_θ should exhibit degraded performance on slices predicted by the SDM and (b) slices predicted by the SDM should be coherent. Specifically, we frame this problem as an information retrieval task, where we label coherent and underperforming slices in datasets and measure the performance of SDMs at retrieving "ground truth" slices.

Our evaluations are fueled by large sets of slice discovery *settings*, each consisting of: (1) a labeled **dataset** $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, (2) a **model** h_θ trained on $\mathcal{D}$, and (3) **slice annotations** $\{\mathbf{s}_i\}_{i=1}^{n}$ for one or more coherent slices $\mathbf{S}$ on which the model h_θ exhibits degraded performance. (1) and (2) correspond to the inputs to the SDM, while (3) corresponds to the expected output. Using these slice discovery settings, we can quantitatively evaluate an SDM $M(\mathcal{D}, f_\theta) \to \Psi$. The procedure is detailed in Algorithm 1.

In the remainder of Section 3, we propose a process for generating representative slice discovery settings across a variety of different slice categories in Section 3.1 and describe the evaluation metrics L and datasets that we use to assess SDM performance in In Section 3.2.

3.1 Generating Slice Discovery Settings

To obtain accurate estimates of SDM performance, the settings fueling our evaluation should ideally be both representative of real-world slice discovery settings and large in number. Unfortunately,

Algorithm 1 SDM Evaluation Process

for $(\mathcal{D}, \mathbf{s}, h_\theta) \in$ settings **do**
  $\Psi \leftarrow M(\mathcal{D}_{\text{valid}}, h_\theta)$ ▷ Fit the SDM on the validation set, yielding a set of slicing functions Ψ
  for $j \in [\hat{k}]$ **do**
    $\hat{\mathbf{s}}^{(\mathbf{j})} \leftarrow \psi^{(j)}(\mathcal{D}_{\text{test}})$ ▷ Apply the slicing functions to the test set, yielding $\hat{s} \in [0,1]^{n_{\text{test}}}$.
  end for
  metrics $\leftarrow \{\max_{j \in [\hat{k}]} L(\mathbf{s}^{(i)}, \hat{s}^{(j)})\}_{i=1}^{k}$ ▷ Compute metric L comparing $\hat{\mathbf{s}}$ and $\mathbf{s}$
end for

such slice discovery settings are usually unavailable. Many public machine learning datasets come with pretrained models, making it straightforward to source parts (1) and (2); however, relatively few datasets specify slices on which those models perform poorly, making part (3) difficult to obtain.

Here, we propose a repeatable process for programmatically generating a large number of realistic slice discovery settings. We begin with a base dataset $\mathcal{D}_{\text{base}}$ that has either a hierarchical label structure (*e.g.* ImageNet) or rich metadata accompanying each example (*e.g.*. CelebA). We select a target variable Y and slice variable $\mathbf{S}$, each defined in terms of the metadata or class structure. This allows us to derive target and slice labels $\{(y_i, \mathbf{s}_i)\}_{i=1}^{n}$ directly from the dataset. In addition, because the slice $\mathbf{S}$ is defined in terms of meaningful annotations, we know that the slice is coherent. After selecting a target variable and slice variable, we (1) generate a dataset $\mathcal{D}$ and (2) generate a model h_θ that exhibits degraded performance with respect to $\mathbf{S}$.

3.1.1 Dataset Generation

We categorize each slice discovery setting based on the underlying reason that the model h_θ exhibits degraded performance on the slices $\mathbf{S}$. We survey the literature for examples of underperforming slices in the wild, which we document in Section A.6. Based on our survey and prior work (Oakden-Rayner et al., 2019), we identify three popular slice types - rare slices, correlation slices, and noisy label slices. We provide descriptions in Section A.1, and describe how we generate them below:

Rare slice. To generate settings with rare slices, we construct $\mathcal{D}$ such that for a given class label Y, elements in subclass C occur with proportion α, where $0.01 < \alpha < 0.1$.

Correlation slice. To generate settings with correlation slices, we construct $\mathcal{D}$ such that a linear correlation α exists between the target variable and other class labels, where $0.2 < \alpha < 0.8$. (implementation details in Section A.2).

Noisy label slice. To generate settings with noisy labels, we construct $\mathcal{D}$ such that for each class label Y, the elements in subclass C exhibit label noise with probability α, where $0.01 < \alpha < 0.3$.

3.1.2 Model Generation

Each slice discovery setting includes a model h_θ that exhibits degraded performance with respect to a set of slices $\mathbf{S}$. We generate two classes of models: (a) trained models, which are trained on our generated datasets D, and (b) synthetic models, which are used to simulate model predictions.

Trained Models. We train a distinct model h_θ across each of our generated datasets D. Our model h_θ is valid if it exhibits a statistically significant degradation in performance with respect to the slices $\mathbf{S}$; in this case, the model h_θ, dataset D, and slices $\mathbf{S}$ will comprise a valid slice discovery setting. It is important to note that real-world datasets often contain a large number of unlabeled slices, so trained models could identify underperforming, coherent slices that are not explicitly labeled as "ground truth" in our evaluation framework. This may complicate the interpretation of our results.

Synthetic Models. We also create settings with simulated model predictions $\bar{h} : [0,1]^k \rightarrow [0,1]$. This allows us to address the limitation highlighted above by explicitly controlling the presence of underperforming slices. We simulate model predictions by sampling from a beta distribution.

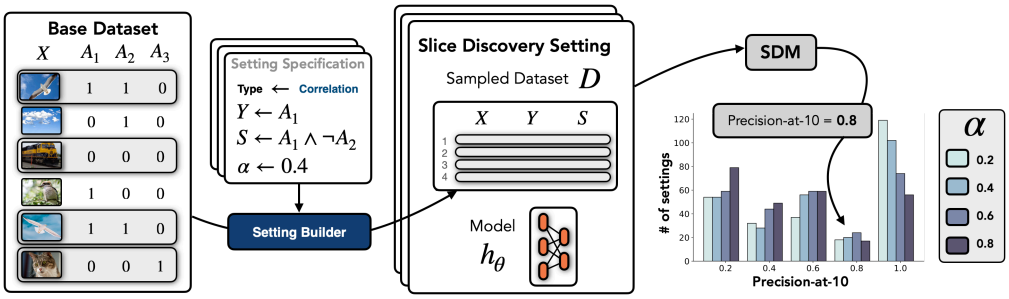

Figure 2: **Evaluation Framework**. We design a programmatic evaluation framework for rigorously evaluating SDMs across a wide range of settings, which is adaptable to any dataset and task.

3.2 Evaluation Approach

We instantiate our evaluation framework by curating 1235 slice discovery settings across a number of different tasks, applications, and base datasets. Detailed statistics on our settings are provided in the Appendix (Table 1). We utilize the following three domains to generate slice discovery settings:

Natural Images (CelebA and ImageNet): We use two natural image datasets. The CelebFaces Attributes Dataset (CelebA) includes over 200k images of celebrities with 40 labeled attributes (Liu et al., 2015). ImageNet is a large image dataset with over 1.2 million images across 1000 labeled classes organized in a hierarchical structure via WordNet (Deng et al., 2009; Fellbaum, 1998).

Medical Images (MIMIC-CXR): The MIMIC Chest X-Ray (MIMIC-CXR) dataset includes 377,110 chest x-rays collected from the Beth Israel Deaconess Medical Center. Annotations indicate the presence or absence of fourteen conditions (Johnson et al., 2019; 2020).

Medical Time-Series Data (EEG): In addition to our image modalities, we also explore time-series data. We obtain a dataset of short 12 second electroencephalography (EEG) signals, which have been used in prior work for predicting the onset of seizures. Additional information on our dataset is included in (blinded citation).

We evaluate SDM performance with *precision-at-k*, which measures the proportion of the top k elements in the discovered slice that are in the ground truth slice. We use $k = 10$ in this work.

4 Domino

In this section, we introduce *Domino*, an SDM that uses cross-modal embeddings to identify coherent slices and describe them in natural language using a three-step procedure (visualized in Fig. 1):

1. **Embed:** We encode the inputs $\{x_i\}_{i=1}^n$ in a cross-modal embedding space via a function $\psi_{\text{input}} : \mathcal{X} \to \mathbb{R}^d$. We learn this embedding function ψ_{input} jointly with an embedding function $\psi_{\text{text}} : \mathcal{T} \to \mathbb{R}^d$ that embeds text in the same space as the inputs.
2. **Slice:** We identify underperforming regions in the cross-modal embedding space using an error-aware mixture model fit on the input embeddings $\mathbf{Z}_{\text{input}} := \{\mathbf{z}_i := \psi_{\text{input}}(x_i)\}_{i=1}^n$, model predictions $\{\hat{y}_i := h_\theta(x_i)\}_{i=1}^n$, and true class labels $\{y_i\}_{i=1}^n$. This yields $\hat{k}$ slicing functions of the form $\psi_{\text{slice}}^{(j)} : \mathbb{R}^d \times \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$.
3. **Describe:** Finally, we use the text embedding function ψ_{text} learned in step (1) to generate a set of $\hat{k}$ natural language descriptions of the discovered slices. Note that the ability to perform this step without human intervention is a particular advantage of our approach.

4.1 Encoding Inputs with Cross-Modal Embeddings

Given input-text pairs (such as images and captions), cross-modal embeddings can be generated using metric learning techniques, where the inputs and text are passed to separate encoders and the output representations are aligned based on the semantic similarity between the pair. Formally, given a set of inputs $N \in \mathcal{X}$ and text descriptions $T \in \mathcal{T}$ expressed in pairs as $\{n_i, t_i\}_{i=1}^m$, we learn two

embedding functions $\psi_{\text{input}} : \mathcal{X} \to \mathbb{R}^d$ and $\psi_{\text{text}} : \mathcal{T} \to \mathbb{R}^d$ such that the distances between pairs of embeddings $dist(\psi_{\text{input}}(n_i), \psi_{\text{text}}(t_i))$ reflect the semantic similarity between n_i and t_i.

Ultimately, this joint training procedure enables the creation of semantically meaningful input embeddings that incorporate information from text. In this work, our key insight is that input representations generated from cross-modal learning techniques encode the semantic knowledge necessary for identifying coherent slices. Our method relies on the assumption that we have access to either (a) pretrained cross-modal embedding functions or (b) a dataset with paired input-text data that can be used to learn cross-modal embedding functions. In practice, we find that this assumption can generally be satisfied since many input types naturally coexist with textual descriptions.

Domino uses four types of cross-modal embeddings to enable slice discovery across our input domains: CLIP (Radford et al., 2021), ConVIRT (Zhang et al., 2020), MIMIC-CLIP, and EEG-CLIP. We adapt CLIP and ConVIRT from prior work, and we train MIMIC-CLIP and EEG-CLIP on large datasets with paired inputs and text (implementation details are provided in Section A.3).

4.2 Clustering Embeddings with Error-Aware Mixture Model

With cross-modal embeddings in hand, we proceed to the second step in the Domino pipeline: slicing. Recall from Section 2 that our goal is to find a set of $\hat{k}$ slicing functions that partition our data into subgroups. This step resembles a standard unsupervised clustering problem but differs in an important way: we are specifically interested in finding *underperforming* clusters where the model makes systematic prediction errors.

Taking inspiration from the recently developed Spotlight algorithm d'Eon et al. (2021), we propose a simple mixture model that jointly models the input embeddings, class labels, and model predictions. This encourages clusters that are homogeneous with respect to error type (*e.g.* all false positives). The model assumes the following generative process (visualized in Figure 1): given the slice S, the embeddings are normally distributed $Z|S \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ with parameters mean $\mu \in \mathbb{R}^d$ and $\mathbf{\Sigma} \in \mathbb{S}^d_{++}$ (the set of symmetric positive definite $d \times d$ matrices), the labels vary as a categorical $Y|S \sim Cat(\mathbf{p})$ with parameter $\mathbf{p} \in \{\mathbf{p} \in \mathbb{R}^c_+ : \sum_{i=1}^c p_i = 1\}$, and the model predictions also vary as a categorical $\hat{Y}|S \sim Cat(\hat{\mathbf{p}})$ with parameter $\hat{\mathbf{p}} \in \{\mathbf{p} \in \mathbb{R}^c_+ : \sum_{i=1}^c p_i = 1\}$. Critically, this assumes that the embedding, label, and prediction are all independent of one another conditioned on the slice.

The mixture model is parameterized by $\phi = [\mu, \Sigma, \mathbf{p}, \hat{\mathbf{p}}]$. The log-likelihood over the n examples in the validation dataset D_v is given as follows and maximized using expectation-maximization:

$$\ell(\phi) = \sum_{i=1}^{n} \log \sum_{s=1}^{k} P(S = s)P(X = x_i|S = s)P(Y = y_i|S = s)P(\hat{Y} = h_\theta(x_i)|S = s), \quad (3)$$

4.3 Generating Natural Language Explanations for Underperforming Slices

Unlike previous SDMs, Domino provides actionable natural language statements that describe characteristics shared between examples in the discovered slices. To generate slice descriptions, we begin by sourcing a corpus of candidate natural language phrases $\mathcal{D}_{\text{text}} = \{t_j\}_{j=1}^{n_{\text{text}}}$. Critically, this text data does not have to be paired with the inputs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ and can be sourced independently. For natural images, we include the $n_{\text{text}} = 10{,}000$ most frequently used words on English Wikipedia (Semenov & Arefin, 2019). For our medical image and time-series datasets, we use corpora of physician reports sourced from MIMIC-CXR (Johnson et al., 2019) and (blinded citation) with $n_{\text{text}} = 159{,}830$ and $n_{\text{text}} = 41{,}258$, respectively.

We generate an embedding for each phrase in the corpus $\mathcal{D}_{\text{text}}$ using the cross-modal embedding function ψ_{text}, yielding $\{\mathbf{z}_{\text{text}}^{(i)} := \psi_{\text{text}}(t_i)\}_{i=1}^{n_{\text{text}}}$. Then, we compute a prototype embedding for each of the discovered slices by taking the weighted average of the input embeddings in the slice, $\{\mathbf{z}_{\text{slice}}^{(j)} := \psi_{\text{slice}}^{(j)}(\mathbf{Z}_{\text{input}})^\top \mathbf{Z}_{\text{input}}\}_{j=1}^{\hat{k}}$. We also compute a prototype embedding for each class $\{\mathbf{z}_{\text{class}}^{(c)} := \mathbf{1}[\mathbf{y} = c]^\top \mathbf{Z}_{\text{input}}\}_{c=1}^{C}$. To distill the slice prototypes, we subtract out the prototype of the most common class in the slice $\mathbf{z}_{\text{slice}}^{(j)} - \mathbf{z}_{\text{class}}^{(c)}$. Finally, to find text that describes each slice, we compute the dot product between the distilled slice prototypes and the text embeddings and return the phrase with

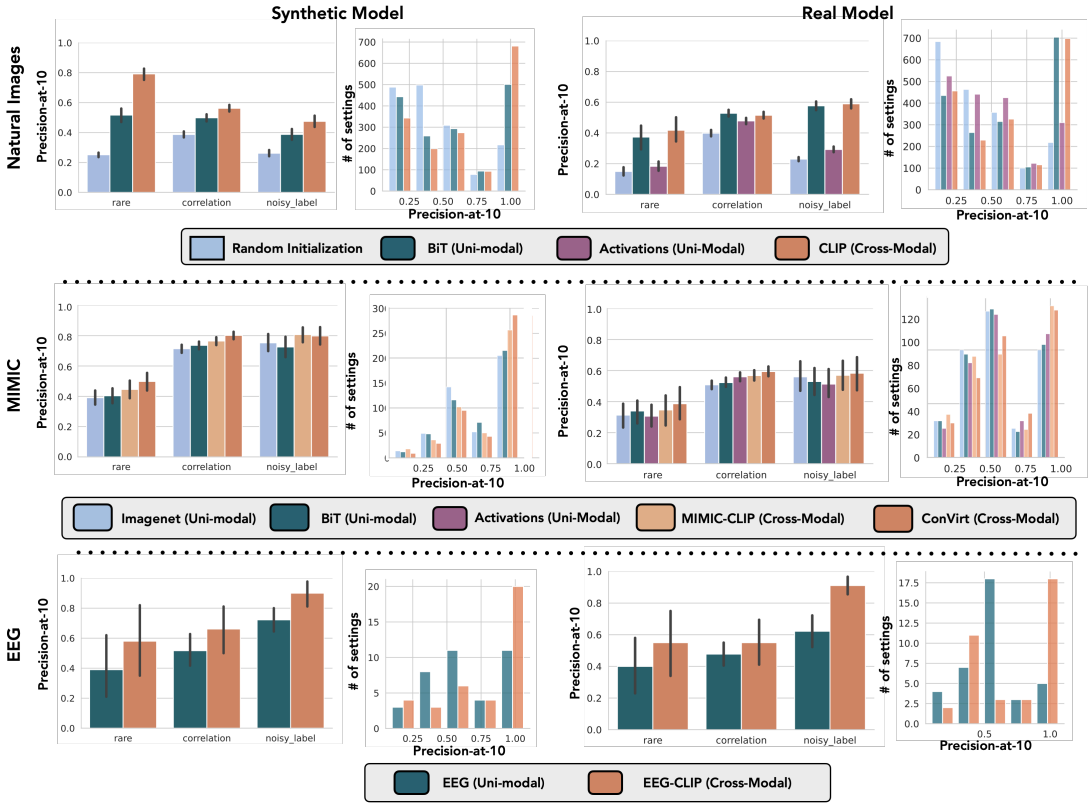

Figure 3: **Cross-modal embeddings enable accurate slice discovery**. Using our evaluation framework, we demonstrate that the use of cross-modal embeddings leads to consistent improvements in slice discovery across three datasets and two input modalities, evaluated over 1,235 settings.

the highest value

$$\text{argmax}_{i \in [n_{\text{text}}]} \mathbf{z}_{\text{text}}^{(i)\top} (\mathbf{z}_{\text{slice}}^{(j)} - \mathbf{z}_{\text{class}}^{(c)}). \tag{4}$$

5 EXPERIMENTS

We use the evaluation framework developed in Section 3 to systematically assess Domino, comparing it to existing SDMs across 1,235 slice discovery settings. Our experiments validate the three core design choices behind Domino: (1) the use of cross-modal embeddings, (2) the use of a novel error-aware mixture model, and (3) the generation of natural language descriptions for slices.

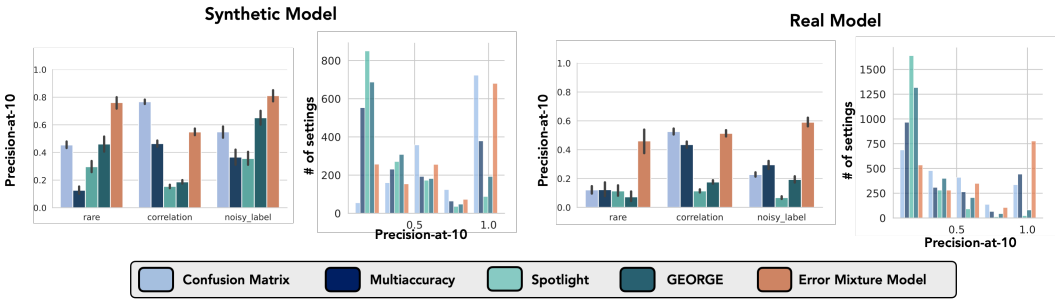

Figure 4: **Error-aware mixture model enables accurate slice discovery**. We show that when cross-modal embeddings are provided as input, our error-aware mixture model often outperforms previously-designed SDMs. Similar results on MIMIC and EEG are detailed in Section A.4.

5.1 CROSS-MODAL EMBEDDINGS IMPROVE SDM PERFORMANCE

To measure how the choice of embedding affects performance, we hold step 2 (clutering algorithm) of Domino fixed and vary the embeddings used in step 1.

Natural Images. In slice discovery settings with natural images, we compare four embeddings: the final-layer activations of a randomly-initialized ResNet-50, the final-layer activations of the trained classifier h_θ, BiT (Kolesnikov et al., 2019), and CLIP (Radford et al., 2021). Out of the four embedding types, CLIP is the only cross-modal embedding. Results are shown in Figure 3.

When evaluating with synthetic models, we find that using CLIP embeddings results in a mean precision-at-10 of 0.570 (0.554,0.586), a 9 percentage-point increase over BiT embeddings and a 23-percentage point increase over untrained activations. Note that synthetic models do not have activations, so we cannot compare to the final-layer activations in this setting.

When evaluating with trained models, we find no difference between using CLIP embeddings and BiT embeddings. However, both outperform using the activations of the trained classifier h_θ by nearly 15-points in mean precision-at-10. This finding is of particular interest given that classifier activations are a popular embedding choice in existing SDMs (d'Eon et al., 2021; Sohoni et al., 2020). Notably, the gap between CLIP and h_θ activations is much smaller in settings with correlation slices. This makes sense because a model that relies on a correlate to make predictions will likely capture information about the correlate in its activations (Sohoni et al., 2020).

Medical Images. On MIMIC-CXR, we compare five embeddings: the final-layer activations of a ResNet-50 pretrained on ImageNet, the final-layer activations of the trained classifier h_θ, BiT (Kolesnikov et al., 2019), and domain-specific cross-modal embeddings that we trained using two different methods: CLIP and ConVIRT. For synthetic models, cross-modal ConVIRT embeddings enable a mean precision-at-10 of 0.765 (0.747,0.784), a 7-point improvement over the best unimodal embeddings (BiT) with mean precision-at-10 of 0.695 (0.674,0.716). With trained models, we again find that although h_θ activations are the worst performing embeddings on both rare and noisy label slices, they are competitive with the multi-modal embeddings on correlation slices.

Medical Time-Series. For our EEG dataset, we compare two embeddings: the final-layer activations of a pretrained seizure classifier and a CLIP-style cross-modal embedding trained on EEG-report pairs. When evaluating with synthetic models, we find that cross-modal embeddings recover coherent slices with a mean precision-at-10 of 0.697 (0.605,0.784). This represents a 17-point gain over using unimodal embeddings 0.532 (0.459,0.608). This demonstrates that cross-modal embeddings can aid in recovering coherent slices even in input modalities other than images.

5.2 ERROR-AWARE MIXTURE MODEL IMPROVES SDM PERFORMANCE

In order to understand the how the choice of clustering algorithm affects performance, we hold constant the first step of the Domino pipeline (cross-modal embeddings) and vary the algorithm used in step 2.

We compare the error-aware mixture model to four prior SDMs: George (Sohoni et al., 2020), Multiaccuracy (Kim et al., 2018), Spotlight (d'Eon et al., 2021), and an important baseline we call ConfusionSDM, which outputs slicing functions that partition data into the cells of the confusion matrix. Additional implementation details are in Section A.2. We provide cross-modal embeddings as input to all five SDMs. On noisy and rare slices, the error-aware mixture model recovers underperforming, coherent slices with a mean precision-at-10 of 0.639 (0.617,0.660) – this represents a 105% improvement over the next-best method, George. Interestingly, on correlation slices, the error-aware mixture model outperforms all methods except the simple ConfusionSDM baseline.

5.3 DOMINO PROVIDES NATURAL LANGUAGE DESCRIPTIONS OF DISCOVERED SLICES.

Domino is the first SDM that can generate natural language descriptions for identified slices. For natural images, we provide a quantitative analysis of these descriptions. Specifically, since Domino returns a ranking over all phrases in the corpus $\mathcal{D}_{\text{text}}$ (specified by the dot product computed in Equation 4), we can compute the percentage of settings in which the name of the "ground truth" slice (or a synonym) appears in the top-m words returned by Domino. In Figure 5, we plot this

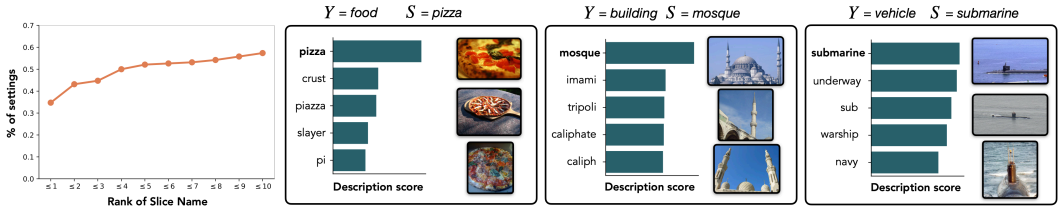

Figure 5: **Domino produces natural language explanations for discovered slices**. (Left) The fraction of natural image settings with rare slices where Domino includes the exact name of the slice in the top-k slice descriptions. (Right) Three slice discovery settings randomly selected from the set of the 85 rare, natural image settings where Domino includes the exact name of the slice in its top 3 slice descriptions. The chart shows the confidence scores for the top 5 descriptions. The images represent the top 3 images that Domino associates with the discovered slice.

percentage for $m = 1$ to $m = 10$ and show three randomly sampled examples. We find that for 34.7% of slices, Domino ranks the name of the slice (or a synonym) first out of the 10,000 words in our corpus. In 57.4% of slices, Domino ranks it in the top ten. For our medical domains, we provide the top-3 returned descriptions in Section A.5, Table 2. For MIMIC, Domino correctly describes *cardiomegaly* (i.e. enlarged heart), a correlate that the model incorrectly relied on to predict a pleural condition. For EEG, Domino correctly describes patient age, a correlate that the model incorrectly relied on to predict seizure.

6 Related Work

Our work builds on prior efforts for slice discovery, cross-modal training, and SDM evaluation.

Slice discovery methods. There have been several recent studies that have proposed SDMs (d'Eon et al., 2021; Yeh et al., 2020; Sohoni et al., 2020; Kim et al., 2018). However, the conditions under which these SDMs succeed at identifying coherent, underperforming slices remain unclear.

Cross-modal training. Large-scale cross-modal representation learning approaches yield powerful embeddings that have contributed to large performance improvements across information retrieval and classification tasks (Radford et al., 2021). Cross-modal models that have inspired our work include CLIP for natural images (Radford et al., 2021), ConVIRT for medical images (Zhang et al., 2020), and WikiSatNet (Uzkent et al., 2019) and Tile2Vec (Jean et al., 2019) for satellite imagery.

Slice discovery datasets. Recently, several benchmark datasets have been proposed for evaluating the performance of models on shifting data distributions. These benchmarks are valuable because they provide labels specifying important slices of data. However, for the purpose of evaluating SDMs, they do not suffice because they either only annotate a small number of slices (Koh et al., 2021) or do not provide pretrained models that underperform on the slices (He et al., 2021; Khosla et al., 2011; Hendrycks & Dietterich, 2019; Liang & Zou, 2021). In Section A.6, we provide a comprehensive survey of underperforming slices in the wild.

7 Conclusion

In this work, we analyze the slice discovery problem. First, we observe that existing approaches for evaluating SDM performance do not allow for large-scale, quantitative evaluations. We address this challenge by introducing a programmable approach to measure SDM performance across two axes: underperformance and coherence. Second, we propose Domino, a novel SDM that combines cross-modal representations with an error-aware mixture model. Using our evaluation framework, we demonstrate that the embedding and slicing steps of Domino outperform those of existing SDMs. We also show for the first time that using cross-modal embeddings for slice discovery can enable the generation of semantically meaningful slice descriptions. Notably, Domino requires only blackbox access to models, and can thus be broadly useful in settings where users have API access to models. We hope that our approach is useful not only for improving slice discovery in practical settings, but also for allowing models to be trained in a manner that directly incorporates human understanding of the task at hand.

8 REPRODUCIBILITY STATEMENT

We provide a robust, open-source implementation of our evaluation framework at `https://github.com/<blindedurl>/domino`. The implementations will enable researchers to reproduce the results described here as well as run their own evaluations on additional datasets. The implementation also includes scripts for preprocessing the publicly available datasets used in this study.

9 ETHICS STATEMENT

Domino is a tool for identifying systematic model errors. No matter how effective it is at this task, there may still be error-modes Domino will not catch. There is a legitimate concern that model debugging tools like Domino could give practitioners a false sense of security, when in fact their models are failing on important slices not recovered by Domino. It is critical that practitioners still run standard evaluations on accurately-labeled, representative test sets in addition to using Domino for auditing models. Additionally, because Domino uses embeddings trained on image-text pairs sourced from the web, it may reflect societal biases when identifying and describing slices. Future work should explore the impacts of using biased embeddings to identify errors in models. What kinds of error modes might we miss? Are certain underrepresented groups or concepts less likely to be identified as an underperforming slice?

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 9505–9515. Curran Associates, Inc., 2018.

Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2(1):1–10, April 2019a.

Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ digital medicine*, 2(1):1–10, 2019b.

Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (de) constructing bias on skin lesion datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 52–59, 2019a.

Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019b.

Alex J DeGrave, Jose Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, May 2021a.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, pp. 1–10, 2021b.

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, June 2009.

Greg d'Eon, Jason d'Eon, James R. Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models, 2021.

Arjun D Desai, Francesco Caliva, Claudia Iriondo, Aliasghar Mortazi, Sachin Jambawalikar, Ulas Bagci, Mathias Perslev, Christian Igel, Erik B Dam, Sibaji Gaj, et al. The international workshop on osteoarthritis imaging knee mri segmentation challenge: a multi-institute evaluation and analysis framework on a standardized dataset. *Radiology: Artificial Intelligence*, 3(3):e200078, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

Karan Goel, Nazneen Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caiming Xiong, Mohit Bansal, and Christopher Ré. Robustness Gym: Unifying the NLP Evaluation Landscape. *arXiv:2101. 04840 [cs]*, January 2021.

Yue He, Zheyan Shen, and Peng Cui. Towards non-i.i.d. image classification: A dataset and baselines. *Pattern Recognit.*, 110:107383, February 2021.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. March 2019.

Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3967–3974, 2019.

A Johnson, L Bulgarelli, T Pollard, S Horng, LA Celi, and R Mark. Mimic-iv (version 1.0), 2020.

Alistair E W Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. *arXiv:1805. 12317 [cs, stat]*, August 2018.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A Earnshaw, Imran S Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv:2012. 07421 [cs]*, March 2021.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. December 2019.

Weixin Liang and James Zou. Metadataset: A dataset of datasets for evaluating distribution shifts and training conflicts. In *ICML2021 ML4data Workshop*, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. September 2019.

Laurel Orr, Megan Leszczynski, Simran Arora, Sen Wu, Neel Guha, Xiao Ling, and Christopher Re. Bootleg: Chasing the tail with self-supervised named entity disambiguation. October 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. February 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. Chrononet: a deep recurrent neural network for abnormal eeg identification. In *Conference on Artificial Intelligence in Medicine in Europe*, pp. 47–56. Springer, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *arXiv:1911. 08731 [cs, stat]*, April 2020.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Ilya Semenov and Shamsul Arefin. Wikipedia word frequency. https://github.com/IlyaSemenov/wikipedia-word-frequency, 2019.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1500–1519, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19339–19352. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e0688d13958a19e087e123148555e4b4-Paper.pdf.

Burak Uzkent, Evan Sheehan, Chenlin Meng, Zhongyi Tang, Marshall Burke, David Lobell, and Stefano Ermon. Learning to interpret satellite images using wikipedia. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, and Holger A Haenssle. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatol.*, 155(10):1135, October 2019a.

Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of

a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155 (10):1135–1141, 2019b.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20554–20565. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/ecb287ff763c169694f682af52c1f309-Paper.pdf.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.*, 15(11):e1002683, November 2018a.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018b.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. *Association for the Advancement of Artificial Intelligence (AAAI)*, January 2018.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *CoRR*, abs/2010.00747, 2020. URL https://arxiv.org/abs/2010.00747.

A APPENDIX

A.1 SLICE DISCOVERY SETTINGS

Domain	**Rare Settings**	**Correlation Settings**	**Noisy Label Settings**	**Total settings**
Natural Images	177 (trained) 197 (synthetic)	520 (trained) 520 (synthetic)	287 (trained) 95 (synthetic)	984 (trained) 812 (synthetic)
MIMIC	15 (trained) 55 (synthetic)	176 (trained) 352 (synthetic)	30 (trained) 55 (synthetic)	221 (trained) 462 (synthetic)
EEG	10 (trained) 10 (synthetic)	10 (trained) 10 (synthetic)	10 (trained) 10 (synthetic)	30 (trained) 30 (synthetic)

Table 1: **Overview of evaluation framework.** We evaluate SDMs across 1235 slice discovery settings across three domains, four slice categories, and five slice parameters (α).

A.1.1 EXPANDED SLICE DESCRIPTIONS

Rare slice. Consider a slice $S \in \{0, 1\}$ (*e.g.* patients with a rare disease, photos taken at night) that occurs infrequently in the training set (*i.e.* $P(S = 1) < \alpha$ for some small α). Since a rare slice will not significantly affect model loss during training, the model may fail to learn to classify examples within the slice. To generate settings with rare slices, we use base datasets with hierarchical label schema (*e.g.* ImageNet (Deng et al., 2009)). We construct dataset $\mathcal{D}$ such that for a given class label Y, the elements in subclass C occur with proportion α, where α ranges between 0.01 and 0.1.

Correlation slice. If the target variable Y (*e.g.* pneumothorax) is correlated with another variable C (*e.g.* chest tubes), the model may learn to rely on C to make predictions. This will induce a slice $S = \mathbf{1}[C \neq Y]$ (*e.g.* pneumothorax without chest tubes and normal with chest tubes) with degraded performance. To generate settings with correlation slices, we use base datasets with metadata annotations (*e.g.* CelebA (Liu et al., 2015)). We construct dataset $\mathcal{D}$ such that a linear correlation α exists between the target variable and other class labels, where α ranges between 0.2 and 0.8.

Noisy label slice. Errors in labels are not always distributed uniformly across the training distribution. A slice of data $S \in \{0, 1\}$ may exhibit higher label error rates than the rest of the training distribution. This could be due to a number of different factors: classes may be ambiguous (*e.g.* annotators labeling sandwiches may disagree whether to include hot dogs), labeling heuristics may fail on important slices (e.g. medical imaging heuristics that only activate on images from one scanner type), or human annotators may lack expertise on certain subsets (e.g. annotators from one country labeling street signs in another). A model h_θ trained on these labels will likely exhibit degraded performance on S. To generate settings with noisy label slices, we can use a base dataset with metadata annotations (*e.g.* CelebA (Liu et al., 2015)). We construct dataset D such that for each class label Y, the elements in subclass C exhibit label noise with probability α, where α ranges between 0.01 and 0.3.

A.1.2 CORRELATION SLICES

Assume we two binary attributes $A, B \in \{0, 1\}$ (Bernoulli random variables) and a dataset of $\hat{\mathcal{D}} = \{(a_i, b_i)\}_{i=1}^{\hat{n}}$. Given a target correlation ρ, we would like to subsample the dataset $\hat{\mathcal{D}}$ such that the resulting dataset $D = (a_i, b_i)_{i=1}^{n}$of size n exhibits a sample correlation between A and B of ρ.

The population correlation between A and B is given by $\rho(A, B) = \frac{\text{cov}(A,B)}{\sigma_A, \sigma_B}$ and $\text{cov}(A, B) = \mathbb{E}[AB] - E[A]E[B]$. For a sample, the unbiased estimator of the covariance is:

$$\text{cov}(\mathbf{a}, \mathbf{b}) = \frac{1}{n-1} \sum_{i=1}^{n} (a_i - \bar{a})(b_i - \bar{b}) = \frac{1}{n-1} \left(\sum_{i=1}^{n} a_i b_i - \bar{b}\bar{a} \right)$$

Since we know A and B are Bernoulli random variables, we can express this in terms of variables like $n_{a=1,b=1}$ (*i.e.* the number of samples i where $a_i = 1$ and $b_i = 1$) and $n_{a=1}$(*i.e.* the number of samples i where a_i=1).

$$\text{cov}(\mathbf{a}, \mathbf{b}) = \frac{1}{n-1}\left(n_{a=1,b=1} - \frac{n_{a=1}n_{b=1}}{n}\right)$$

The correlation coefficient can then be expressed in terms of this covariance and the sample standard deviations s_a and s_b:

$$r = \frac{n_{a=1,b=1} - \frac{n_{a=1}n_{b=1}}{n}}{(n-1)s_a s_b}$$

In addition to supplying a target correlation ρ, assume we also have target means μ_a and μ_b (these could be the sample means in the original dataset $\mathcal{D}$ for example) and a target sample size $\hat{n}$. We can then derive simple formulas for computing the desired $n_{a=1,b=1}$, etc.:

$$n_{a=1} = \mu_a n \qquad n_{b=1} = \mu_b n \qquad n_{a=1,b=1} = \rho(n-1)s_a s_b + \frac{n_{a=1}n_{b=1}}{n}$$

$$n_{a=1,b=0} = n_{a=1} - n_{a=1,b=1} \qquad n_{a=0,b=1} = n_{b=1} - n_{a=1,b=1} \qquad n_{a=0,b=0} = n - (n_{a=1} + n_{b=1} - n_{a=1,b=1})$$

where $s_a = \mu_a(1-\mu_a)$ and $s_b = \mu_b(1-\mu_b)$ (since they are Bernoulli).

A.2 SDM Implementations

The Spotlight. d'Eon et al. (2021) search for an underperforming slice by maximizing the expected loss with respect to a multivariate Gaussian with spherical covariance. We use the authors' implementation provided at `https://github.com/gregdeon/spotlight`. We enforce a minimum spotlight size equal to 2% of the validation data as recommended in the paper. We use a initial learning rate of 1×10^{-3} (the default in the implementation) and apply the same annealing and barriers as the authors. We optimize for 1,000 steps per spotlight (the default in the implementation).

GEORGE. Sohoni et al. (2020) propose identifying underperforming slices by doing class conditional clustering on a U-MAP reduction of the embeddings. We use the implementation provided at `https://github.com/HazyResearch/hidden-stratification`.

Multiaccuracy Boost. To identify slices where the model h_θ is systematically making mistakes, Kim et al. (2018) use ridge regression to learn a function $f : \mathbb{R}^d \to \mathbb{R}_+$ mapping from an example's embedding $z_i \in \mathbb{R}^d$ to the partial derivative of the cross entropy loss with respect to the prediction

$$\frac{\partial \ell(h_\theta(x_i), y_i)}{\partial h_\theta(x_i)} = \frac{1}{1 - h_\theta(x_i) - y_i}. \tag{5}$$

Because this function grows as the absolute value of the residual $|h_\theta(x_i) - y_i|$ grows, a good f should correlate with the residual.

In order to discover multiple distinct slices, the authors repeat this process $\hat{k}$ times updating the model predictions on each iteration according to

$$h_\theta^{(j+1)}(x_i) \propto e^{-\eta f(z_i)} h_\theta^{(j)}(x_i), \tag{6}$$

where η is a hyperparameter defining the step size for the update.

We use an implementation of Multiaccuracy Boost based on the authors', which was released at `https://github.com/amiratag/MultiAccuracyBoost`. We use $\eta = 0.1$ as in the authors' implementation. We fit f on 70% of the validation data and use the remaining 30% for evaluating the correlation with the residual. The authors' use the same ratio in their implementation.

Confusion SDM. A simple, embedding-agnostic way to identify underperforming subgroups is to simply inspect the cells of the confusion matrix. We include this important baseline to determine when more complicated slice discovery techniques are actually useful.

A.3 CROSS-MODAL EMBEDDINGS

Here, we provide implementation details for the four cross-modal embeddings utilized in this work: CLIP, ConVIRT, MIMIC-CLIP, and EEG-CLIP. Domino relies on the assumption that we have access to either (a) pretrained cross-modal embedding functions or (b) a dataset with paired input-text data that can be used to learn embedding functions.

Large-scale pretrained cross-modal embeddings can be utilized to generate accurate representations of input examples. For instance, if our inference dataset consists of natural images, we can utilize a pre-trained CLIP model as embedding functions ψ_{input} and ψ_{text} to obtain image embeddings that lie in the same latent representation space as word embeddings.

However, pre-trained cross-modal embeddings are only useful if the generated representations accurately represent the inputs in the inference dataset. For example, if the inference dataset consists of images from specialized domains (i.e. X-Rays) or non-image inputs, CLIP is likely to generate poor representations.

If pre-trained cross-modal embeddings are not available or cannot effectively represent the inference dataset, we require access to a separate dataset that can be used to learn the cross-modal embedding functions ψ_{input} and ψ_{text}; we will refer to this dataset as the *CM-training dataset* for the remainder of this work. The CM-training dataset must consist of paired input-text data (e.g. image-caption pairs or radiograph-report pairs). Further, we assume that the text data provides a sufficient description of the input and includes information about potential correlates, such as object attributes or subject demographics. Note that paired input-text data is only required for the CM-training dataset; we make no such assumptions about the slice discovery dataset.

We utilize the following four cross-modal embeddings in our analysis:

- **CLIP (Natural Images)**: CLIP embeddings are cross-modal representations generated from a large neural network trained on 400 million image-text pairs (Radford et al., 2021).
- **ConVIRT (Medical Images)**: ConVIRT embeddings are generated from pairs of chest X-rays and radiologist reports in the MIMIC-CXR dataset. We create a CM-training set with 70% of the subjects in MIMIC-CXR, ensuring that no examples in our training set occur in validation or test data at slice discovery time. We then replicate the training procedure detailed in Zhang et al. (2020), which uses contrastive learning to align embeddings.
- **MIMIC (Medical Images)**: We generate a separate set of cross-modal embeddings for the MIMIC-CXR dataset using a variant of the CLIP training procedure (Radford et al., 2021). We use the same CM-training set as detailed above, with $89{,}651$ image and report pairs. In order to generate text representations, we extract the findings and impressions sections from radiologist reports, which then serve as input to a BERT-based transformer initialized with weights from CheXBert (Smit et al., 2020). Chest x-rays are passed as input to a visual transformer (ViT) pre-trained on ImageNet-21k and ImageNet 2012 at an image resolution of 224×224 (Dosovitskiy et al., 2021). All images are resized to 224×224 and normalized using mean and standard deviation values from ImageNet. Text representations are extracted from the output of the [CLS] token and image representations are extracted from the output of the final model layer. Image representation and text representations are separately passed through projection layers consisting of fully-connected layers and non-linear activation functions as detailed in an open-source implementation of CLIP[1]. Finally, we align text and image representations using the InfoNCE loss function with in-batch negatives as detailed in Radford et al. (2021). We train our implementation for 30 epochs with a learning rate of 10^{-4}, a batch size of 64, and an embedding dimension of 256. The training process comes to an early stop if the loss fails to decrease for ten epochs.
- **EEG (Medical Time-Series Data)**: In order to generate cross-modal embeddings for EEG readings and associated neurologist reports, we modify the CLIP training procedure to work with time-series data (Radford et al., 2021)(blinded citation). We create a CM-training set with $6{,}254$ EEG signals and associated neurologist reports. In order to represent text representations, we extract the findings and narrative from the reports, which are then served as input to a BERT-based transformer initialized with weights from CheXBert (Smit et al.,

[1] https://github.com/moein-shariatnia/OpenAI-CLIP

2020). Since the EEG signals are sampled at 200 Hz, and the EEG clip length is 12 seconds, with 19 EEG electrodes, the input EEG has shape 19×2400. The EEG encoder is a densely connected inception convolution neural network (Roy et al., 2019). Text and EEG representations are aligned using the InfoNCE loss function with in-batch negatives as described in (Radford et al., 2021). We also add a binary cross-entropy classification loss for seizure classification, where we weigh the InfoNCE loss by 0.9, and the cross-entropy loss by 0.1. The cross-modal model is trained with a learning rate of 10^{-6}, an embedding dimension of 128, and a batch size of 32 for 200 epochs.

A.4 Additional Experiments

Additional experimental results with our error-aware mixture model are shown in Figure 6.

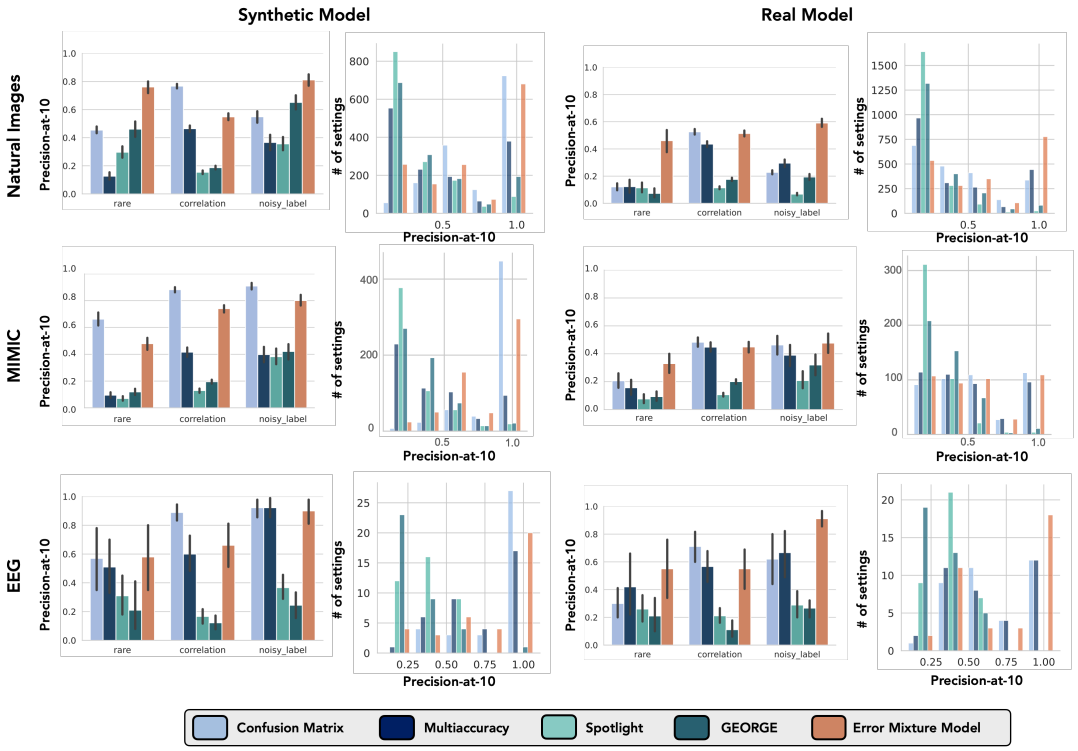

Figure 6: **Error-aware mixture model enables accurate slice discovery**. We show that when cross-modal embeddings are provided as input, our error-aware mixture model often outperforms previously-designed SDMs.

A.5 Natural Language Explanations

In this section, we provide examples of natural language explanations on our medical image and medical time-series datasets. Refer to Table 2 for examples of explanations generated by Domino.

A.6 Survey of Slices Naturally Occurring in the Wild

Skin lesion classification (correlation). Malignant lesions with bandaids in ISIC. A computer vision model trained to classify skin lesions performs poorly on images of malignant lesions with color charts (i.e. colored bandages) next to them because their low prevalence causes the model to leverage a spurious relationship between these bandages and malignancy (Bissoto et al., 2019).

Skin lesion classification (correlation). A CNN was trained on the ISIC dataset to detect malignant skin lesions (a binary classification task). In the course of clinical practice, dermotologists routinely mark suspicious lesions with a gentian violet surgical skin marker. Winkler et al. gathered images of

Dataset	Slice Type	Slice Setting	Description 1	Description 2	Description 3
MIMIC	Correlation with $\alpha = 0.8$	Pleural condition correlated with cardiomegaly	*moderate cardiomegaly* persists, but no evidence of pneumonia or pulmonary edema	heart size top normal. no pulmonary edema or mediastinal abnormality.	mild *enlargement of the cardiac silhouette*
EEG	Correlation with $\alpha = 0.8$	Young patient with seizure	15 month old boy with recent febrile seizure	Intractable epilepsy - stiffening, eyes rolling up, breathing stopping	17 yo girl ex-35 weeker with IVH, cerebral palsy, and brief seizures

Table 2: **Domino produces natural language explanations for discovered slices**: Since Domino is supported by cross-modal representations, which embed natural language in the same latent space as the input samples, we are able to generate textual descriptions for identified slices. In this table, we list some examples of datasets, slice discovery settings, and the top-3 generated explanations. Note that in our medical examples, the vocabulary consists of physician reports in the training set; however, we are unable to provide the full-text reports due to patient privacy concerns, so the table includes relevant phrases extracted from the reports.

130 melanocytic lesions (107 benign nevi and 23 melanomas). The images of the benign nevi were gathered prspectively, allowing the authors to image the lesion both before and after the surgical skin markings were applied. Unmarked images of the melanoma were pulled retrospectively from a library of images (presumably because they occur less frequently in clinical practice, so prospective collection is difficult). To create a set of corresponding marked melanomas the authors used photoshop. Using this dataset, the authors show that the skin markings significanlty increased the mean probability of melanoma output by the model (0.16 ->0.54 on nevi and 0.94 ->1.0 in melanomas) (Winkler et al., 2019b).

Pneumothorax Detection (correlation). Identifying pneumothorax on chest radiographs (Oakden-Rayner et al., 2019).

Knee MR segmentation (rare). Segment key tissues (articular cartilage + meniscus) in the knee (Desai et al., 2021).

Hip-fracture detection (rare). Cervical fractures in the Adelaide dataset. A computer vision model trained to detect hip-fractures achieves low sensitivity on cervical fractures. Presumably, the model has learned that fractures in the cervix are rare, so the threshold for abnormality is higher (Oakden-Rayner et al., 2019).

Hip-fracture detection (correlation). A model was trained to detect hip fracture from x-ray on a dataset of 23,557 radiographs from several clinical sites across at least three different departments (inpatient, outpatient, ED, and unknown) at Mount Sinai A total of twelve different devices were used across time and clinical sites. The model could predict fracture on an IID test set moderately well (AUC=0.79). However, when evaluating on a test dataset subsampled to uncorrelate fracture from patient variables (”age”, ”gender”, ”bmi”, ”pain”, ”fall”) and hospital variables (”scanner model”, ”department”, ”radiation”, ”radiologist name”, ”order time”) – model performance drops to barely better than random. This suggest that the model is relying primarily on patient and hospital variables to make predictions (Badgeley et al., 2019b).

Gender classification in images (rare). Commercial gender classification performance is lower on some intersectional subgroups (e.g. darker faces, female faces). Three commercial systems were found to perform worse on darker faces (Buolamwini & Gebru, 2018).

Chest X-ray COVID-19 detection (correlation). Diagnose CVOID-19 from chest radiographs. They use saliency maps and find that they highlight regions outside the lung fields. They also use a GAN to try to identify these confounds: ”However, the generative networks frequently add or re-

move laterality markers and annotations (Fig. 2b, solid red boxes), reinforcing our observation from saliency maps that these spurious confounds also enable ML models to differentiate the COVID-19-positive and COVID-19-negative radiographs." (DeGrave et al., 2021b).

Chest X-ray pneuomonia detection (correlation). DDiagnose pneumonia from chest radiographs. Differences in prevalance of pneumonia across hospital systems (e.g. NIH, Mt. Sinai) led the model to rely on confounding variables. They don't explore precisely what the confounders are, but they do show saliency maps that focus on laterality markers (Zech et al., 2018b).

Speech recognition (rare). By analyzing a large corpus of sociolinguistic interviews with white and African American speakers, we demonstrate large racial disparities in the performance of five popular commercial ASR systems (Koenecke et al., 2020).

Object detection (rare). The paper analyzes the accuracy of publicly available object-recognition systems on a geographically diverse dataset. This dataset contains household items and was designed to have a more representative geographical coverage than commonly used image datasets in object recognition. We find that the systems perform relatively poorly on household items that commonly occur in countries with a low household income (de Vries et al., 2019a).