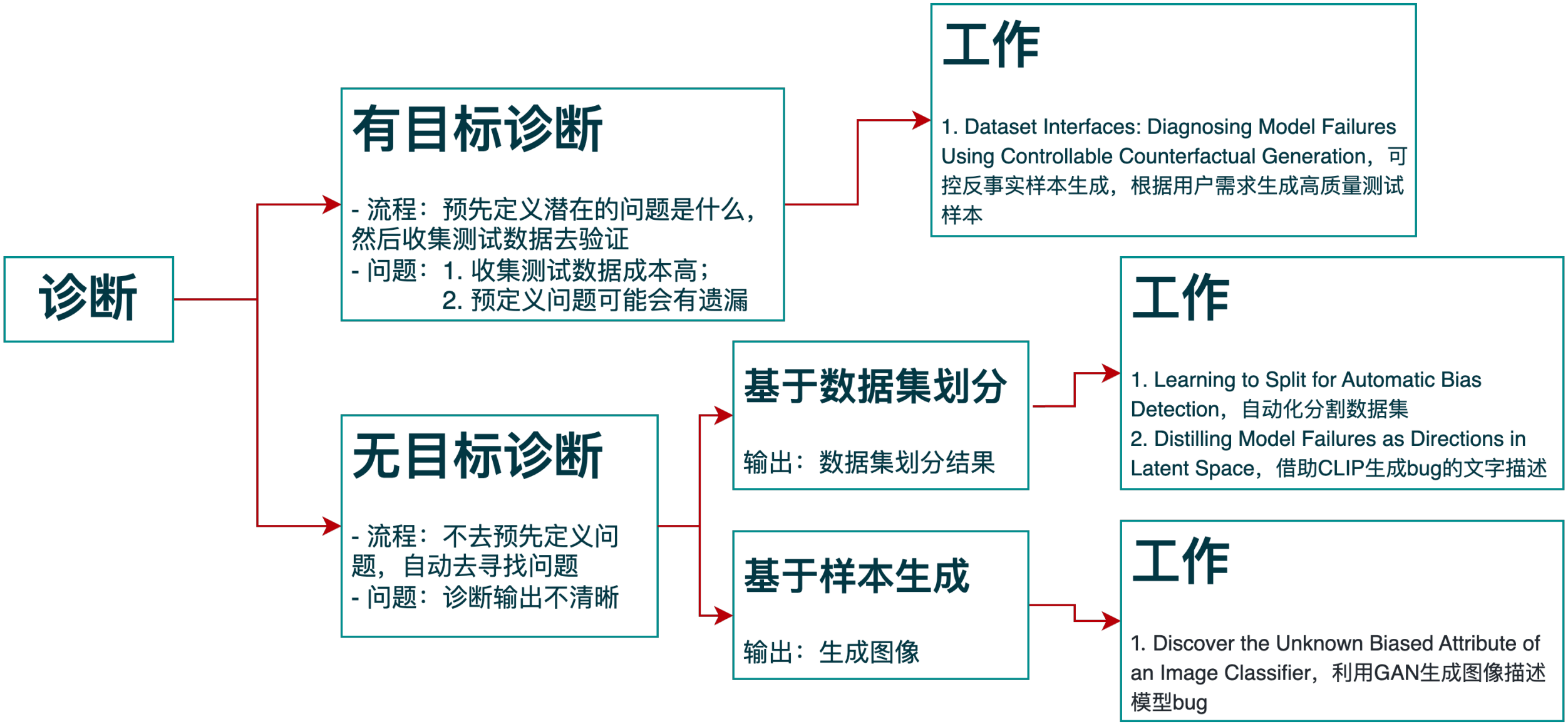
非解释性诊断方法

胡锐

2023.3.3


诊断
有目标诊断
- 流程：预先定义潜在的问题是什么，然后收集测试数据去验证
- 问题：1. 收集测试数据成本高；
2. 预定义问题可能会有遗漏
工作
1. Dataset Interfaces: Diagnosing Model Failures Using Controllable Counterfactual Generation，可控反事实样本生成，根据用户需求生成高质量测试样本
无目标诊断
- 流程：不去预先定义问题，自动去寻找问题
- 问题：诊断输出不清晰
基于数据集划分
输出：数据集划分结果
工作
1. Learning to Split for Automatic Bias Detection，自动化分割数据集
2. Distilling Model Failures as Directions in Latent Space，借助CLIP生成bug的文字描述
基于样本生成
输出：生成图像
工作
1. Discover the Unknown Biased Attribute of an Image Classifier，利用GAN生成图像描述模型bug

有目标诊断

一般分为三个步骤

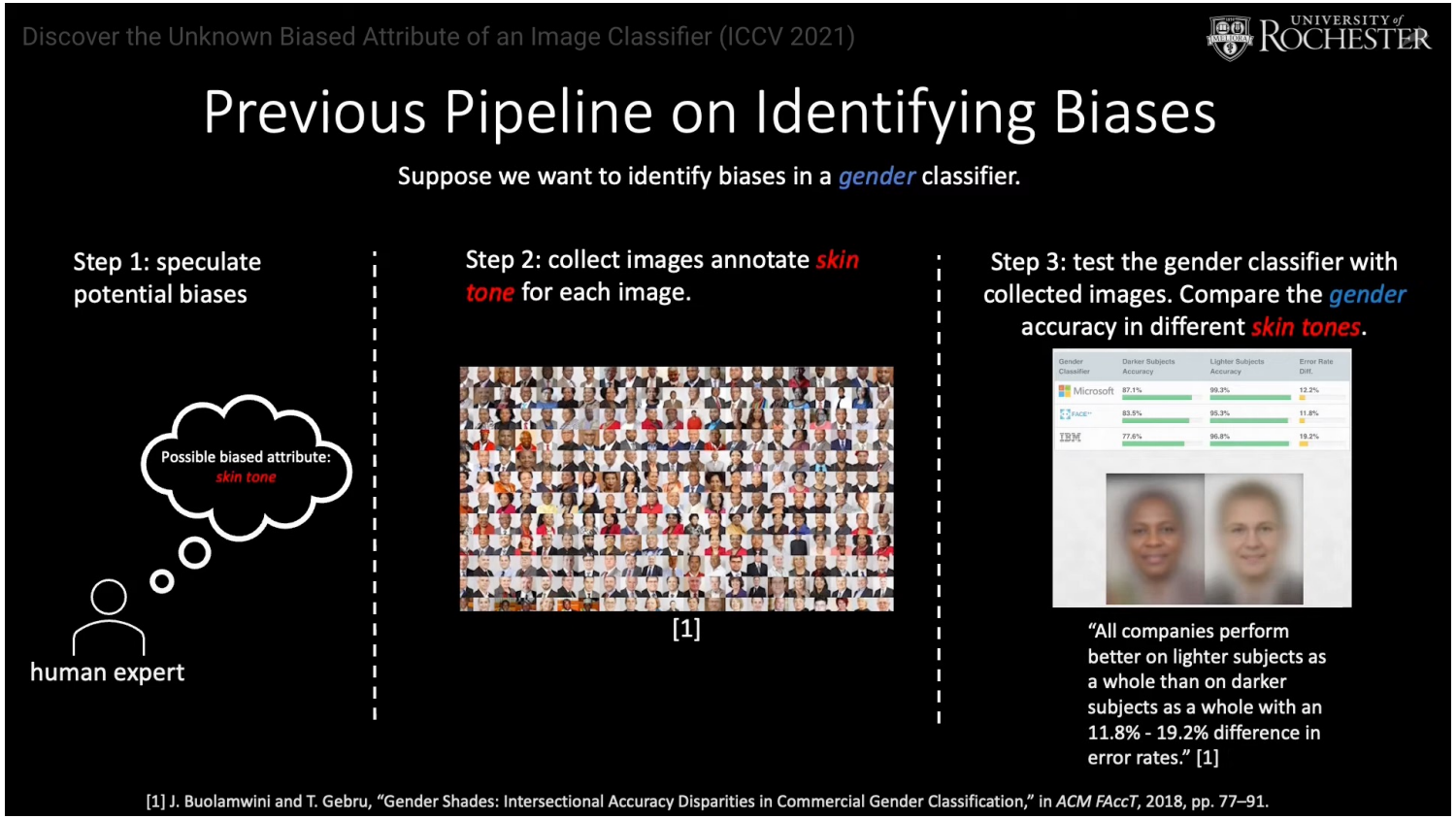
step1: 预定义模型潜在问题；

step2: 根据预定义收集目标测试集；

step3: 测试模型并分析模型性能；

问题：

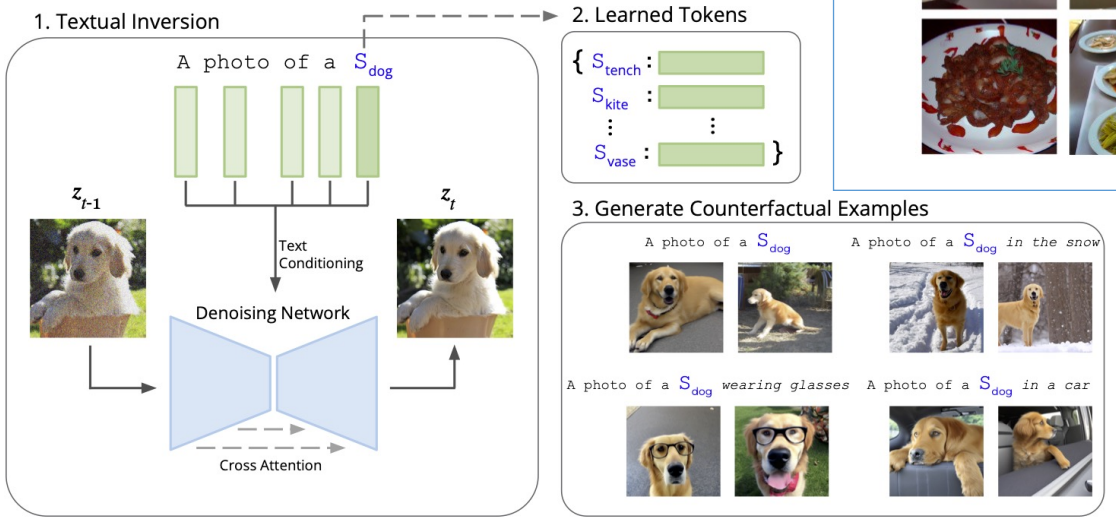
- 预定义潜在问题可能有遗漏 -> 无目标诊断
- 收集目标测试集成本/难度大 -> **自动生成测试集**



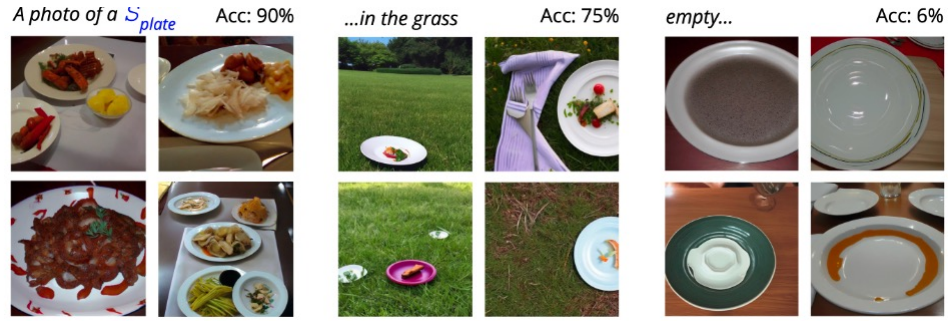
Dataset Interfaces: Diagnosing Model Failures Using Controllable Counterfactual Generation
可控反事实生成

- 提出了一种基于Stable diffusion的反事实样本生成方法，可以细粒度地控制生成图片的属性，方便做对照实验从而有效诊断模型bug
- 现有生成模型生成图片不可控：想要准确的诊断，需要准确控制测试集之间的变化

背景和内容和发生变化，无法确定原因



学习能代表原始测试集类别信息的embedding



通过控制实验进行诊断

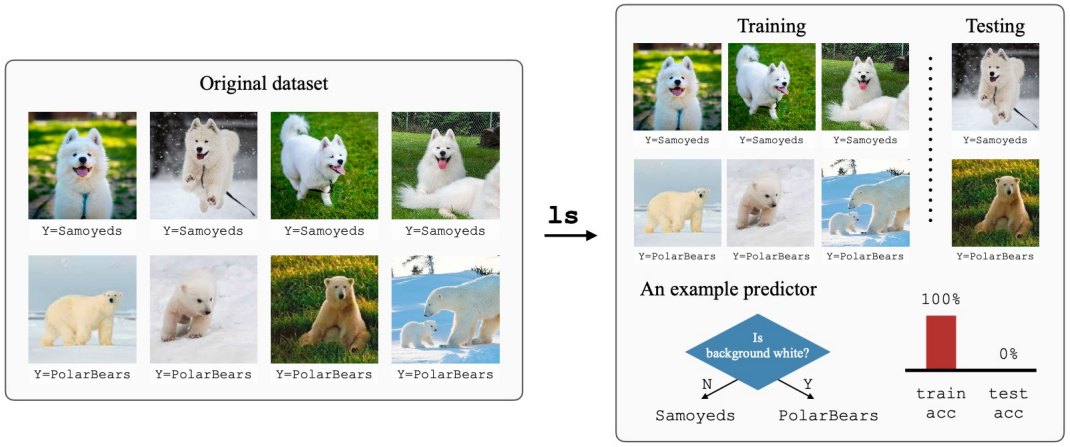
无目标诊断

- 基于数据集划分的诊断：输出是数据集划分结果
- 基于生成的诊断：输出是生成的样本

Learning to Split for Automatic Bias Detection

learn to split 功能介绍

- ls可以学习**自动将数据集分割为训练集和测试集**，使得在训练集上训练的模型**无法泛化到测试集**
- 产生的分割很有价值，因为它们可以帮助我们调试数据集，从而获得更鲁棒的模型
  - 例如数据集中 minority group 是什么?
  - 有没有标注错误?



训练方法

- ls有两个关键组成部分：分割器和预测器
  - 分割器：是一个二分类模型，对于每一个样本，决定分到测试集或训练集
  - 预测器：就是目标任务分类器，并用于输出分割器所分割的训练集和测试集的指标Gap（例如准确率差）
- 优化目标：**最大化分割Gap**

Splitter

Generate a train-test split of the dataset.

Predictor

Minimize the loss on the training set.*

Estimate the generalization error on the testing set.

Maximize the generalization error.

使用ls诊断预训练模型步骤

1. 收集少量目标任务数据（目的是为了诊断模型错误，所以少量数据就可以了）
2. 训练ls，其中使用微调预训练模型当做Predictior

3. 观察分割结果，诊断预训练模型应用到该目标任务时的潜在bug

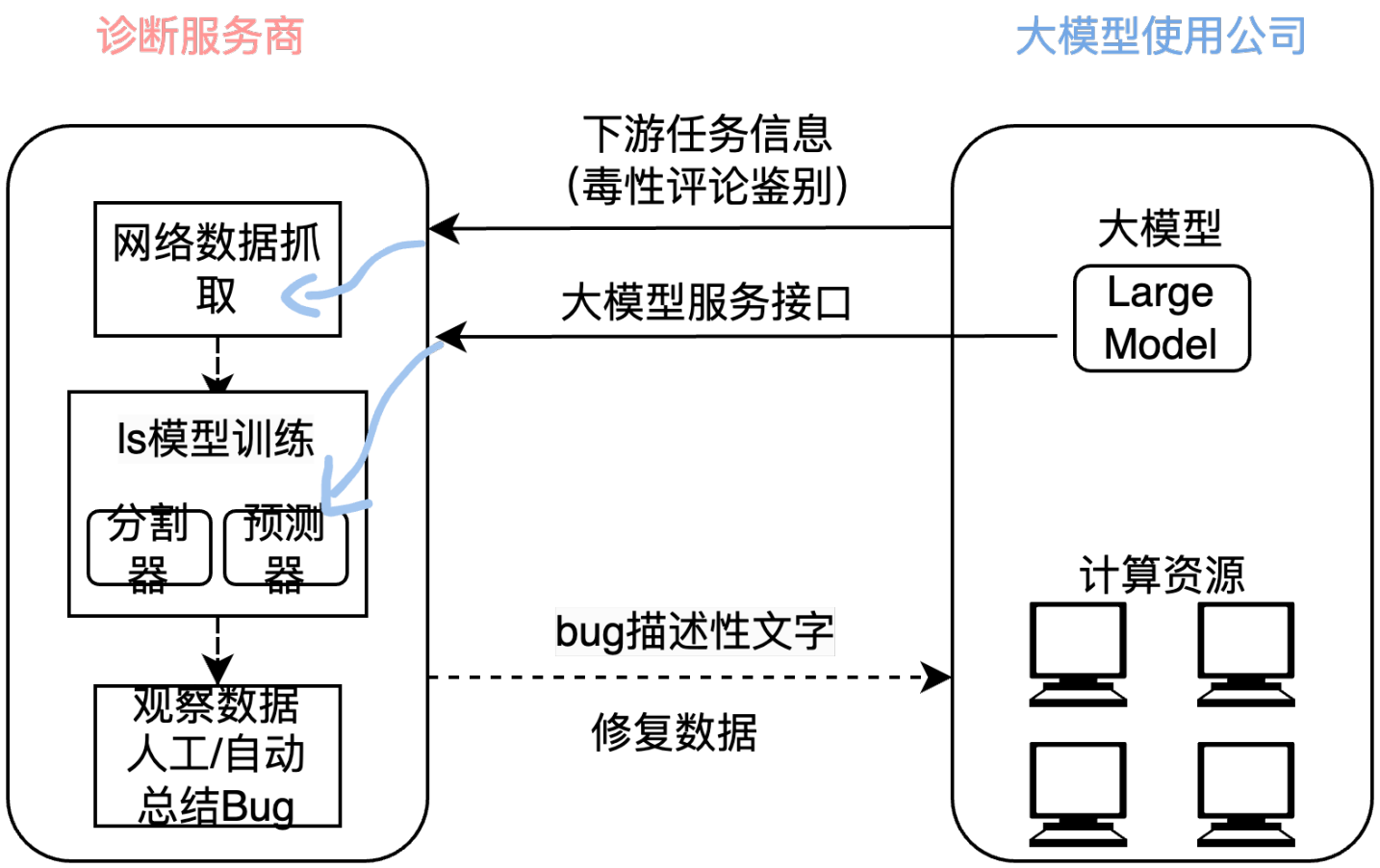
Quickstart

You can directly use the `ls.learning_to_split()` interface to generate challenging splits on PyTorch dataset object. Here is a quick example using the Tox21 dataset:

```
>>> import ls

# Load the Tox21 dataset.
>>> data = ls.datasets.Tox21()

# Learning to split the Tox21 dataset.
# Here we use a simple mlp as our model backbone and use roc_auc as the evaluation metric.
>>> train_data, test_data = ls.learning_to_split(data, model={'name': 'mlp'}, metric='roc_auc')

Best split:
ls outer loop 9 @ 23:51:42 2022/10/17
| generalization gap 64.31 (val 98.97, test 34.65)
| train count 72.7% (7440)
| test  count 27.3% (2800)
| train label dist {0: 7218, 1: 222}
| test  label dist {0: 2627, 1: 173}
```

By default, `learning_to_split` will output the split status for each outer loop iteration (see tox21.log for the full log). In this example, we see that `ls` converged after 9 iterations. It identified a very challenging train/test split (generalization gap = 64.31%).

ls可能的应用模式



大模型使用公司

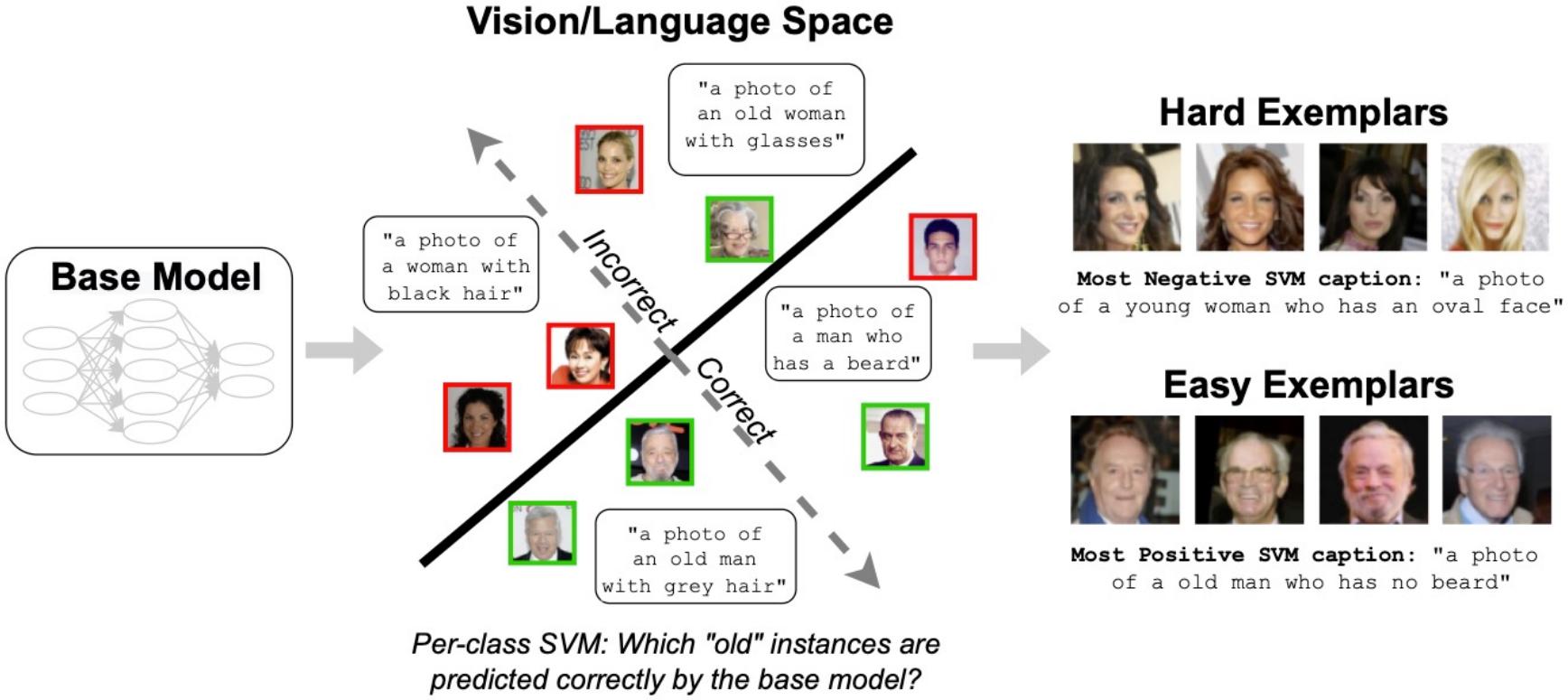
- 定位：小微企业，付费使用大模型接口，用于自己的下游任务；
- 核心竞争力：业务数据；
- 需求：在业务数据上微调后的大模型风险评估；

诊断服务商

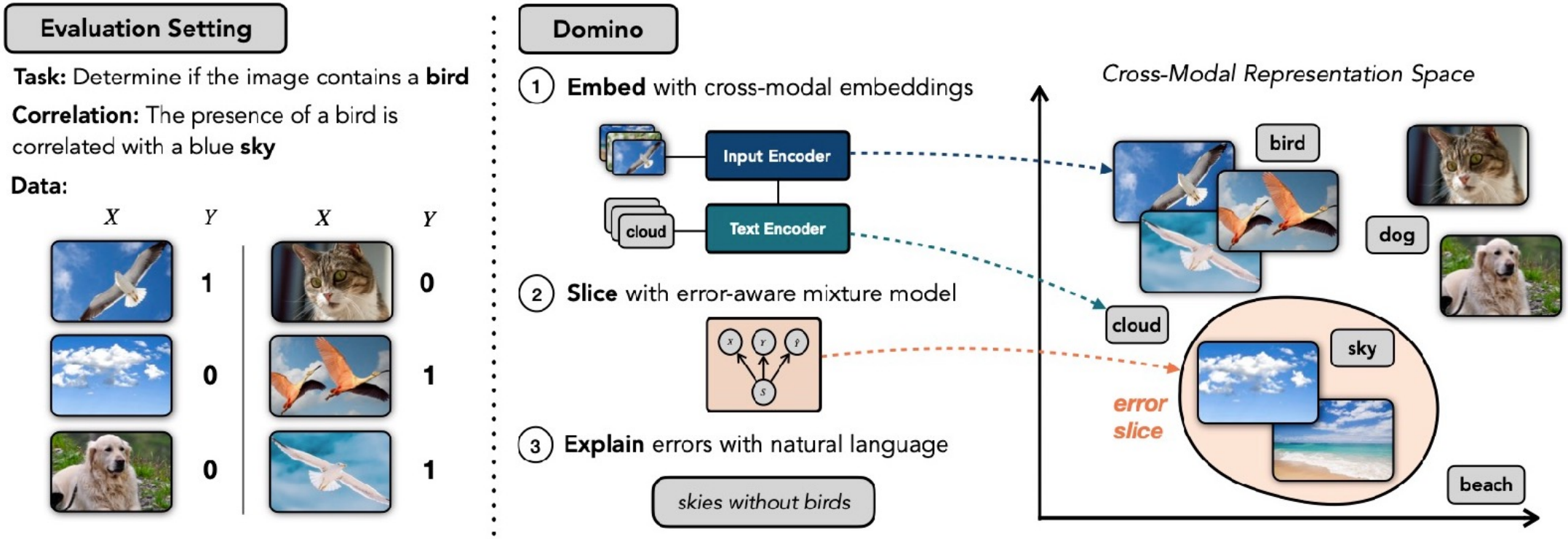
- 定位：提供自动化黑盒诊断服务，风险提示；
- 核心竞争力：自动化诊断技术（爬取数据->模型训练->风险总结）；
- 优势：成本小，不需要太多计算资源，容易生存（分割器很小的网络就可以，预测器调用客户接口）；

Distilling Model Failures as Directions in Latent Space

- 将某个类别（例如老年人）的样本放入CLIP的嵌入空间，再在该嵌入空间训练SVM二分类器来判断目标模型是否会将这些样本分类正确，就可以得到代表模型行为的向量方向，从而将样本分为困难样本和简单杨样本
- 同时借助CLIP的图像文本表征能力，对困难样本进行caption，就可以获得模型潜在bug的描述

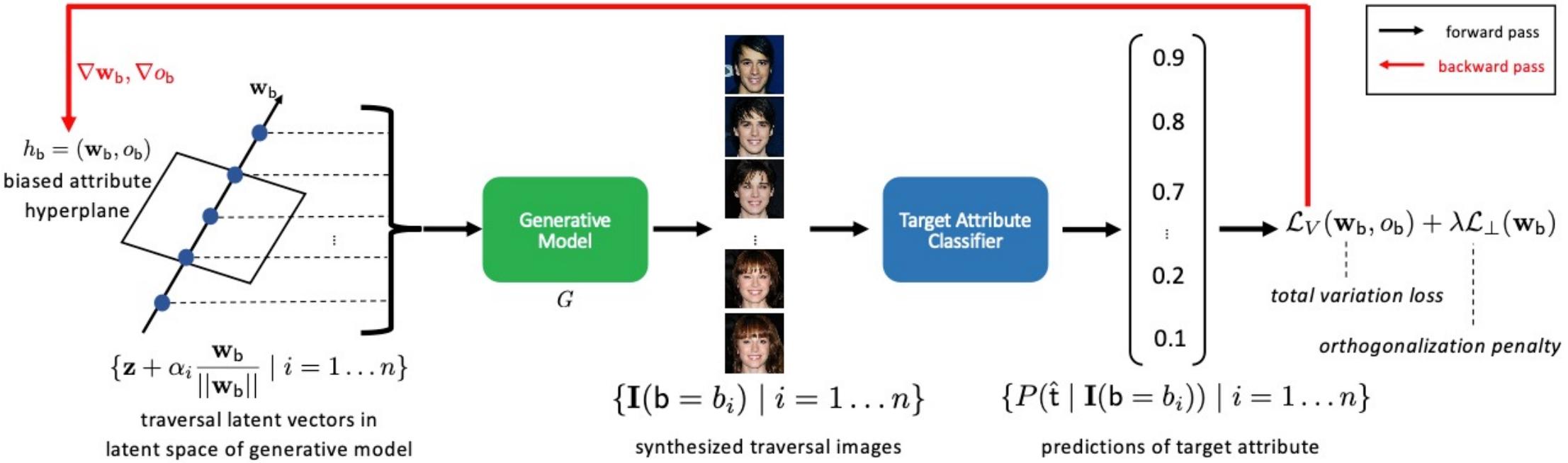


DOMINO: DISCOVERING SYSTEMATIC ERRORS WITH CROSS-MODAL EMBEDDINGS



Discovered Unknown Biased Attributes

- 借助生成模型（GAN或者VAE的解码器），来建立bias属性的超平面，使得等间隔的法向量上的点，经过生成模型G以后，目标模型输出结果也是渐变的，即由分类正确变为分类错误



Classifier	Classfier's Target Prediction	Discovered Unknown Biased Attribute		
ResNet-18 Trained on ImageNet [1]	Cat	shade of fur color (light → dark)		
ResNet-18 Trained on Places365 [2]	Bedroom	number of beds (1 → 2)		
ResNet-18 Trained on Places365 [2]	Bridge	buildings in the background (no building → building)		
ResNet-18 Trained on Places365 [2]	Conference Room	layout of conference room (table → hollow square table / no table)		
ResNet-18 Trained on Places365 [2]	Tower	is Eiffel Tower (Eiffel tower → other towers)		

总结

诊断分为有目标诊断和无目标诊断

- 有目标诊断
  - 重心在于提升测试数据生成的质量，通过分析测试结果即可诊断出模型问题；
- 无目标诊断
  - 常见做法是切分数据集，找到Hard Slice（困难样本切片），通过分析困难样本来总结模型问题；
  - 也有通过生成模型来观察和分析合成样本渐变过程来诊断模型问题；

- 不管是有目标诊断还是无目标诊断，目前都需要有人的参与
  - 有目标诊断需要人类专家指出潜在的问题目标
  - 无目标诊断在于最终结果需要人类观察分析并总结

->未来趋势：**诊断流程自动化** -> eg. 对 Hard Slice/渐变过程 自动归纳总结，输出人类可以理解的诊断结果