ICML
International Conference
On Machine Learning

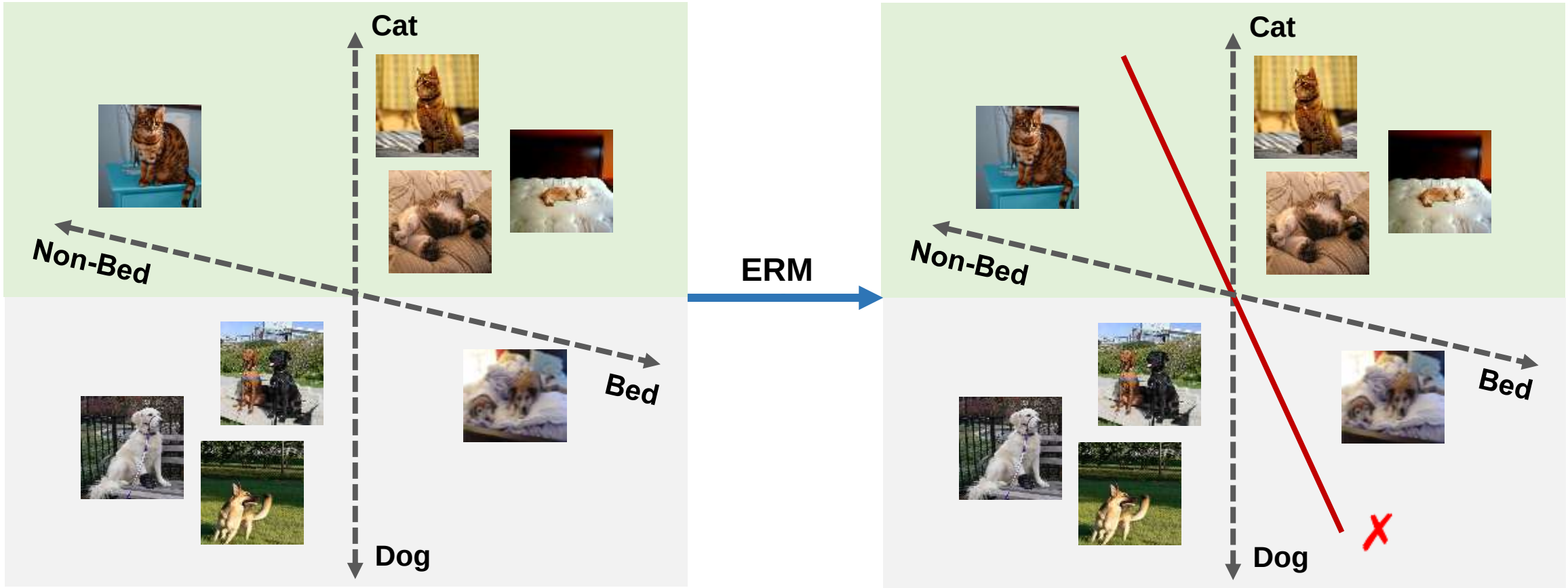
RUTGERS

Discover and Cure: Concept-aware Mitigation of Spurious Correlation

Shirley Wu
Stanford University

Lead-in

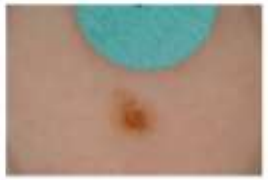
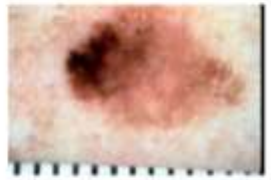
—— What's the problem?



Lead-in

—— Why is it important?

Target: benign / melanoma skin lesions
Spurious features: dark corners, hair, gel borders, gel bubbles, ruler, ink markings/staining, patches.

Image:			...		
Target $y \in \{0,1\}$**:**	0 (benign)	0 (benign)	...	1 (malignant)	1 (malignant)
Spurious s**:**	patch, gel border	ink, hair	...	dark corner, gel bubble	ruler, dark corner

Target: one of 62 building or land use categories, e.g., park, shopping mall, dam, stadium, airport.
Spurious features: Unknown (not explicitly given by the data source).

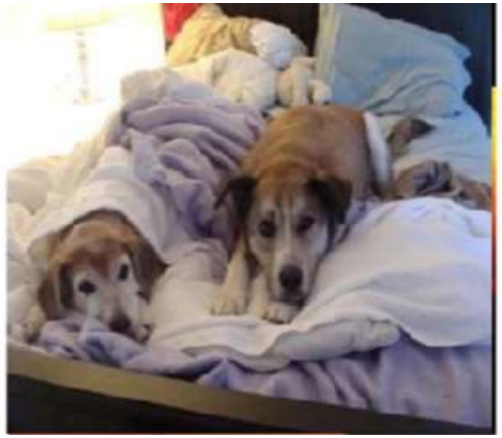
Image:					
Group g**:**	Europe	Asia	Americas	Africa	Oceania

Lead-in

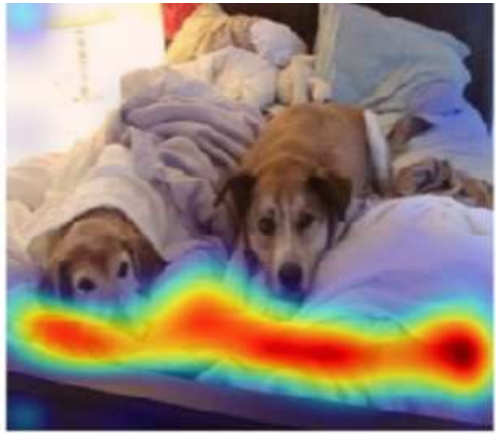
—— Why is it hard?

Spurious features are **hard** to find without any annotations.

Prediction: Cat



Grad-CAM

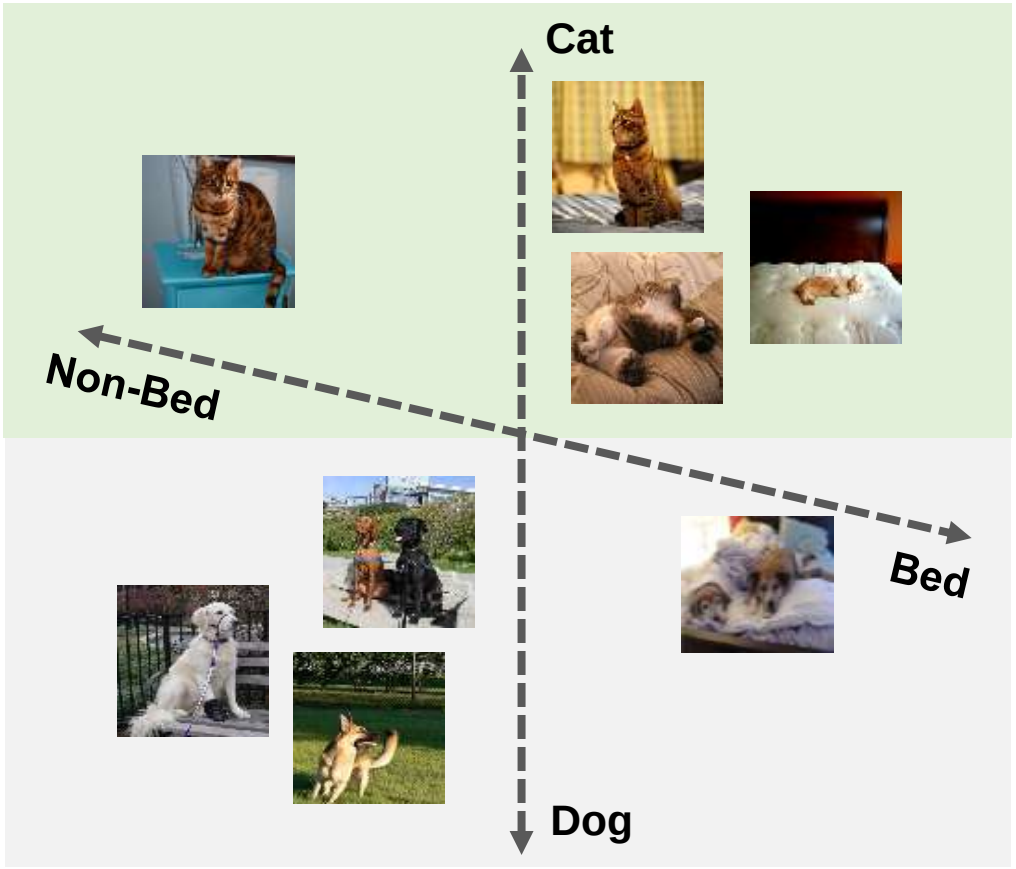


Some explainability methods can generate feature maps for instances with wrong prediction, but they are **ambiguous** for people to understand.

Other approaches to unlearn the spurious correlations:
① Simply removing spurious features could introduce more noise or cause over-fitting.
② Upweight images without spurious features, but it is difficult when multiple spurious features exist.

The key intuitions behind DISC

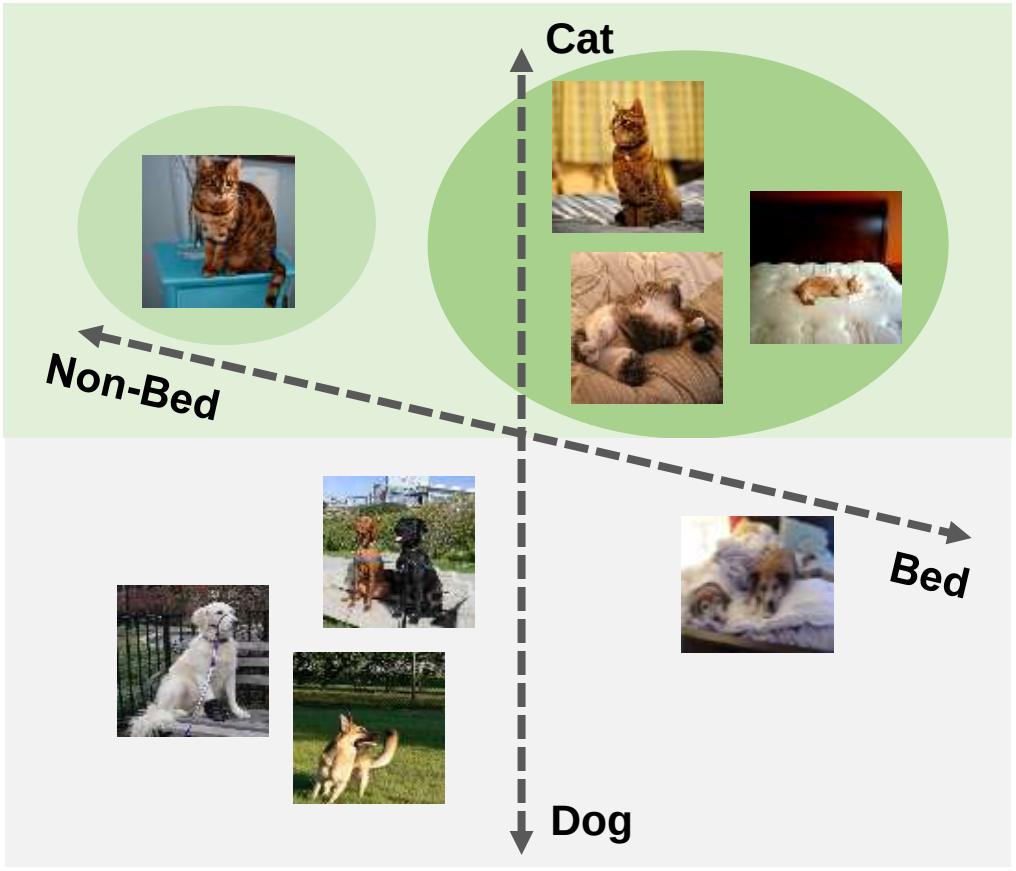
—— What causes models to learn from spurious correlation?



[Intuition I] The spurious correlation of models is caused by the **distribution imbalance** of the spurious feature over different classes.

The key intuitions behind DISC

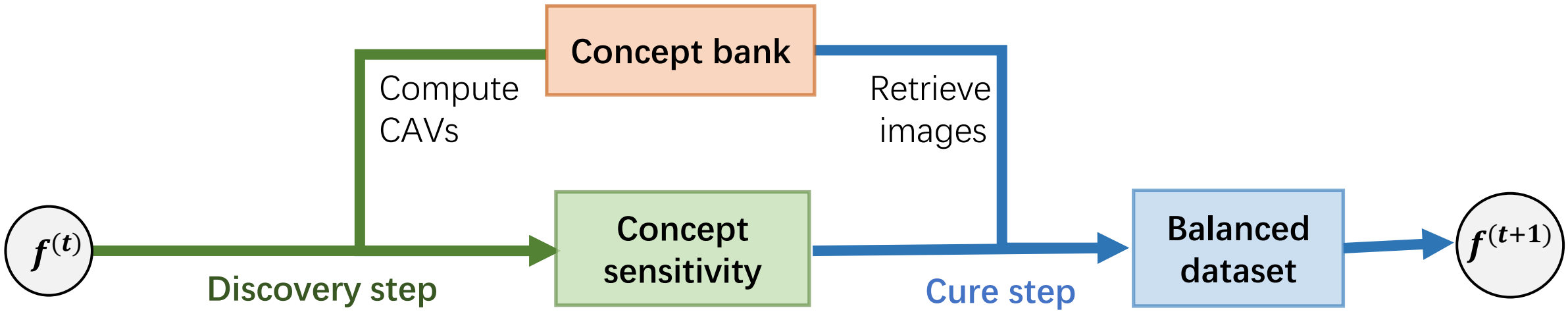
—— What is the property of spurious correlation?



[Intuition II] Spurious features tend to be present in **heterogeneous subsets** of the data and their correlations with the label are also heterogeneous.

Overview of DISC

Intuition II (Inconsistancy property) → Discover spurious concepts (discovery step)
Intuition I (Distributional imbalance) → Remove the spurious correlation (cure step)



Concept bank

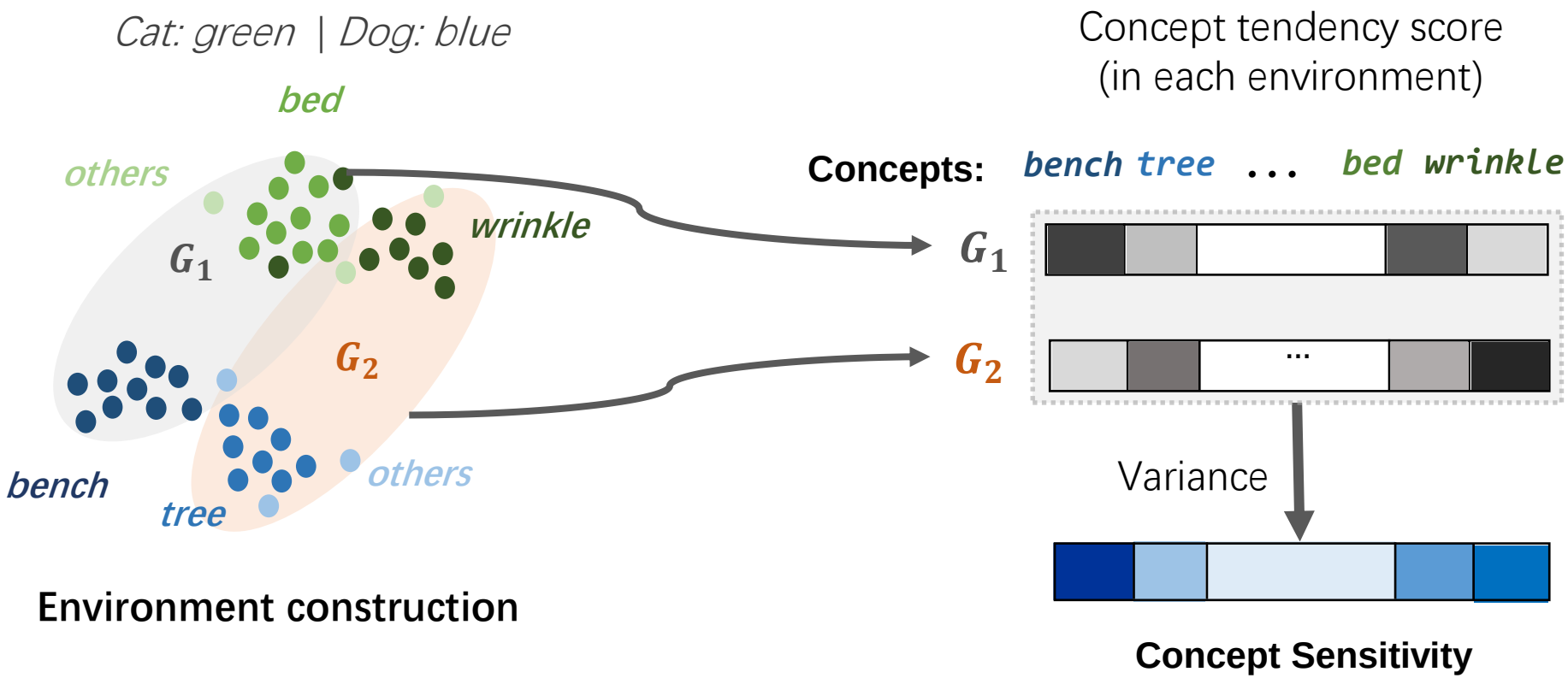
$$\{c_1 : \boldsymbol{\mathcal{P}}_{c_1}, \ldots, c_m : \boldsymbol{\mathcal{P}}_{c_m}\}$$

Concept	Category	Example of Concept Images
Blueness	Color	
Stripes	Texture	
Tree	Nature	
Streets	City	

Query operation

A Concept Activation Vector (CAV) is the direction in the hidden space representing the existence of a concept.

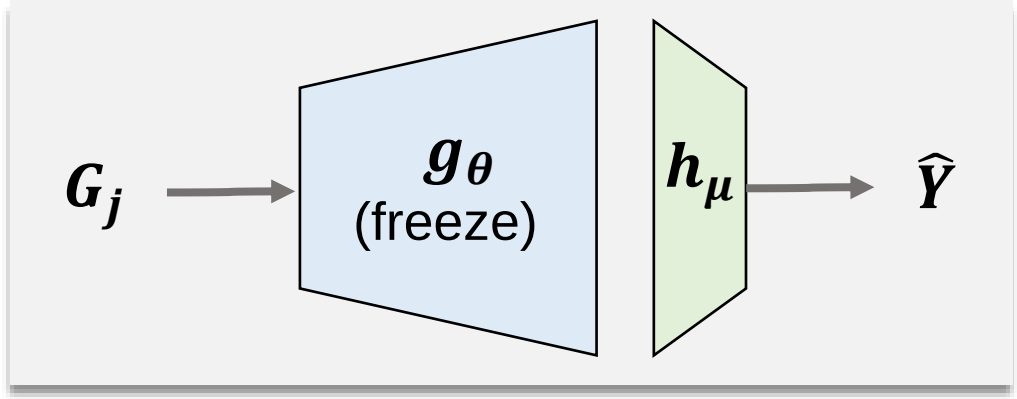
On discovering spurious concepts



On discovering spurious concepts

—— Compute concept tendency score

$$X \rightarrow Z \in R^m \rightarrow \hat{Y} \in R^c$$



$$M_j = \lambda \frac{\nabla L(\hat{Y}, Y)}{\nabla \mu} \in R^{m \times c}$$

Gradient matrix

m: hidden dim
k: number of concepts
c : number of classes

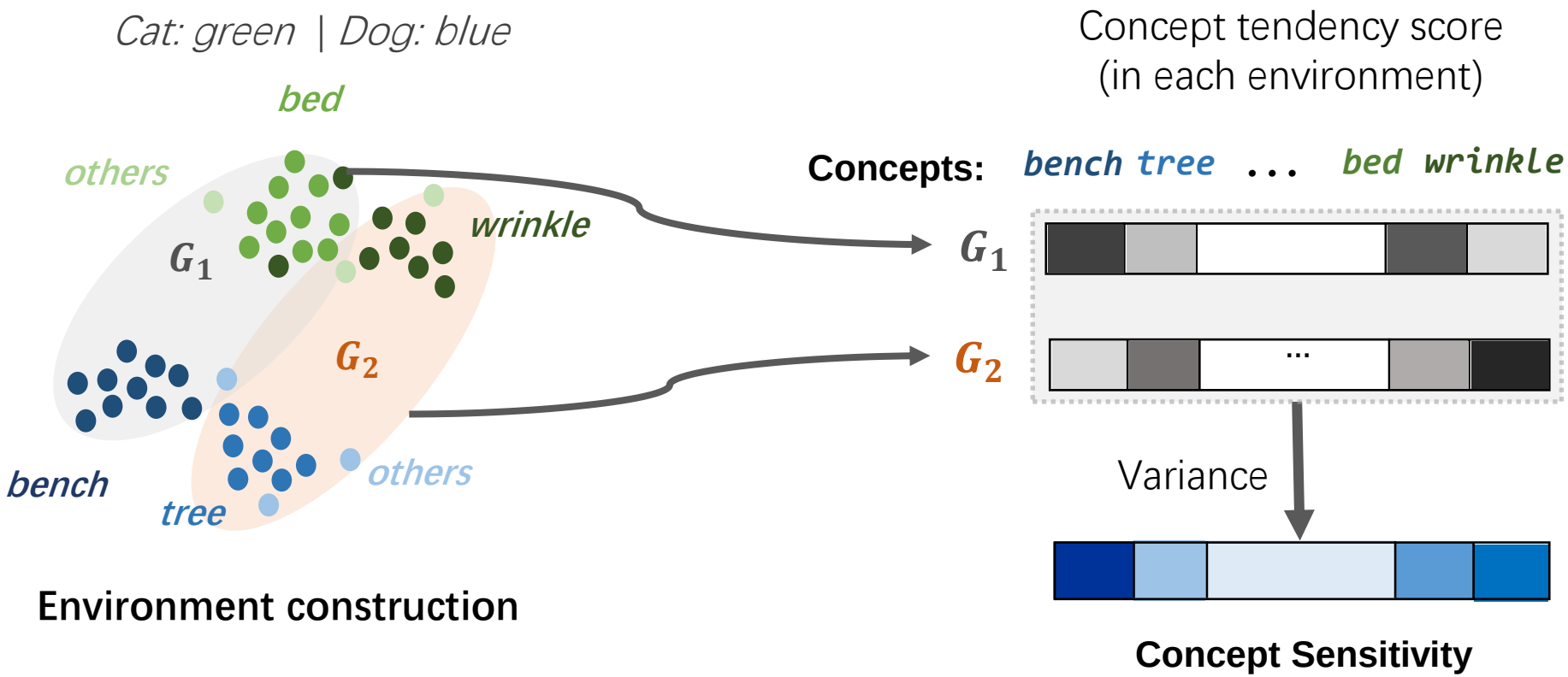
$v_i \in R^m$
Concept activation vector

$v_i \cdot M_j \in R^c$
Concept tendency score

On discovering spurious concepts

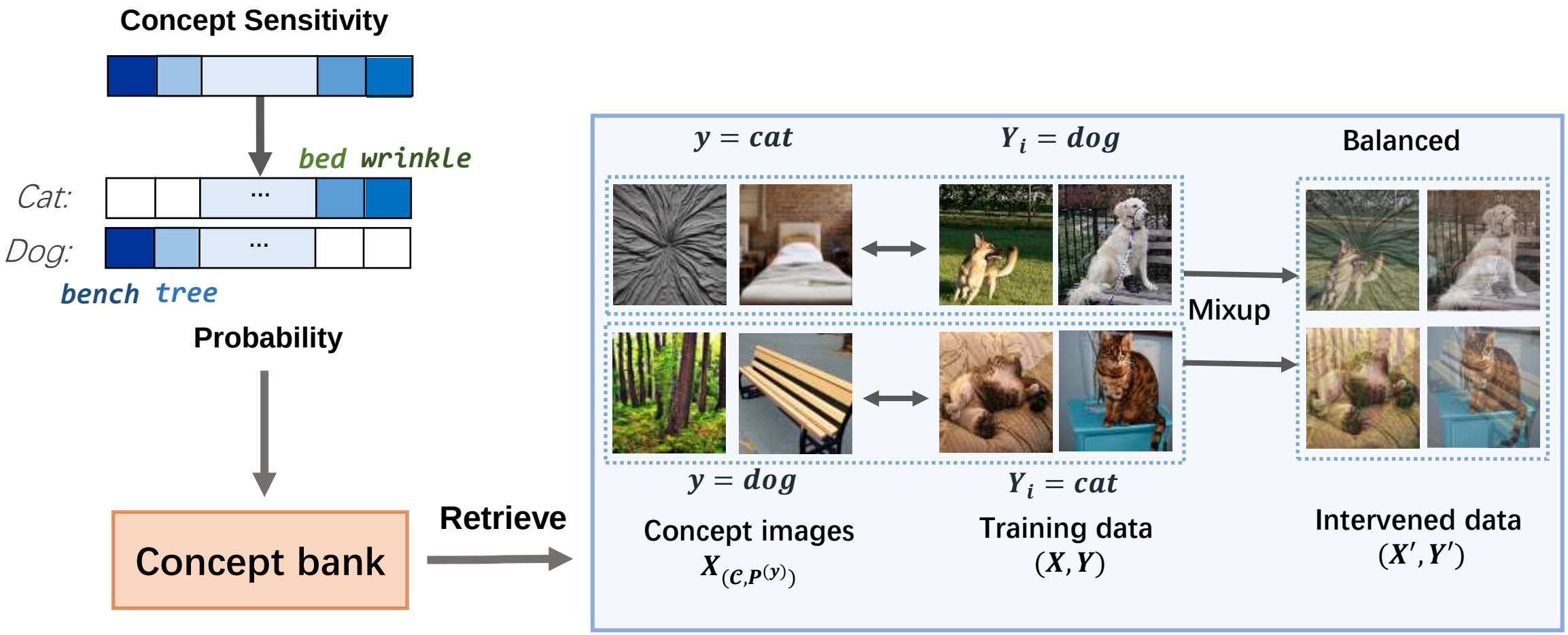
—— Summary

"The importance of each concept for the classes in each environment"



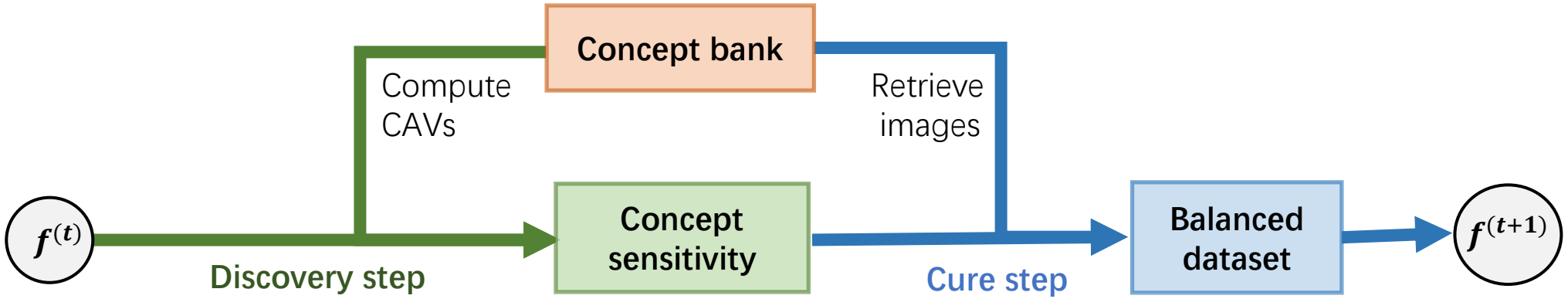
"To what extent does the importance of each concept vary across different environments"

Concept-aware intervention

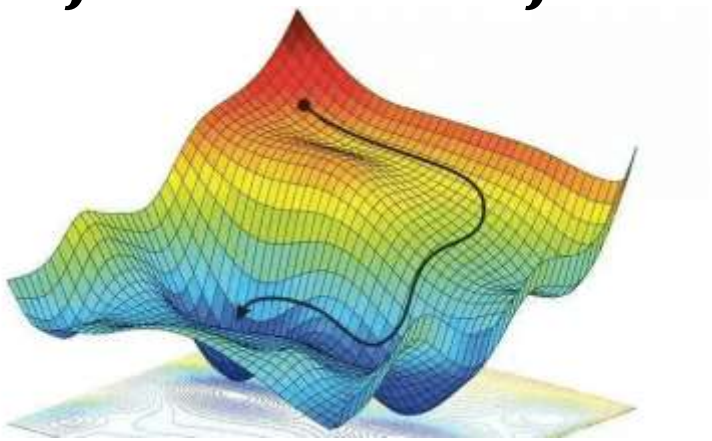


Training process

—— Final state



$$f^{(0)} \longrightarrow \dots \quad f^{(t)} \longrightarrow f^{(t+1)} \dots \longrightarrow f^{(T)}$$



Training on different subsets / Group stratification won't change the concept importance!
→ Necessary condition of an "oracle" model

Experiments

—— How effective is DISC on tasks with spurious correlations?

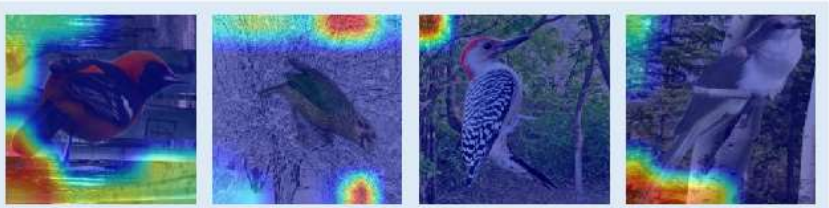
Table 1. Overall experimental results. The best results are **bold** and the second best results are underlined.

	MetaShift		Waterbirds		FMoW		ISIC
	Avg. Acc.	Worst Acc.	Avg. Acc.	Worst Acc.	Avg. Acc.	Worst Acc.	Avg. AUROC
ERM	72.9 ± 1.4%	62.1 ± 4.8%	97.0 ± 0.2%	63.7 ± 1.9%	53.0 ± 0.6%	32.3 ± 1.3%	36.4 ± 0.7%
ERM+aug	75.5 ± 1.7%	65.7 ± 3.3%	87.4 ± 0.5%	76.4 ± 2.0%	55.5 ± 0.4%	35.7 ± 0.3%	38.9 ± 1.5%
UW	72.1 ± 0.9%	60.5 ± 3.8%	96.3 ± 0.3%	76.2 ± 1.4%	52.5 ± 0.5%	30.7 ± 1.5%	39.2 ± 0.6%
IRM	73.9 ± 0.8%	64.7 ± 2.1%	87.5 ± 0.7%	75.6 ± 3.1%	50.8 ± 0.1%	30.0 ± 1.4%	45.5 ± 3.6%
IB-IRM	74.8 ± 0.2%	65.6 ± 1.1%	88.5 ± 0.9%	76.5 ± 1.2%	49.5 ± 0.5%	28.4 ± 0.9%	38.6 ± 1.5%
V-REx	72.7 ± 1.7%	60.8 ± 5.5%	88.0 ± 1.4%	73.6 ± 0.2%	48.0 ± 0.6%	27.2 ± 0.8%	24.5 ± 6.4%
CORAL	73.6 ± 0.4%	62.8 ± 2.7%	90.3 ± 1.1%	79.8 ± 1.8%	50.5 ± 0.4%	31.7 ± 1.2%	37.9 ± 0.7%
Fish	64.4 ± 2.0%	53.2 ± 4.5%	85.6 ± 0.4%	64.0 ± 0.3%	51.8 ± 0.3%	34.6 ± 0.2%	42.0 ± 0.8%
GroupDRO	73.6 ± 2.1%	66.0 ± 3.8%	91.8 ± 0.3%	**90.6 ± 1.1%**	52.1 ± 0.5%	30.8 ± 0.8%	36.4 ± 0.9%
JTT	74.4 ± 0.6%	64.6 ± 2.3%	93.3 ± 0.3%	86.7 ± 1.5%	52.5 ± 0.3%	33.4 ± 0.9%	33.8 ± 0.0%
DM-ADA	74.0 ± 0.8%	65.7 ± 1.4%	76.4 ± 0.3%	53.0 ± 1.3%	51.6 ± 0.2%	34.2 ± 0.8%	35.8 ± 1.0%
LISA	70.0 ± 0.7%	59.8 ± 2.3%	91.8 ± 0.3%	88.5 ± 0.8%	52.8 ± 0.9%	35.5 ± 0.7%	38.0 ± 1.3%
DISC	75.5 ± 1.1%	**73.5 ± 1.4%**	93.8 ± 0.7%	88.7 ± 0.4%	53.9 ± 0.4%	**36.1 ± 1.8%**	**55.1 ± 2.3%**

Experiments

—— How effective is DISC in discovering spurious concepts?

(a) Grad-CAM

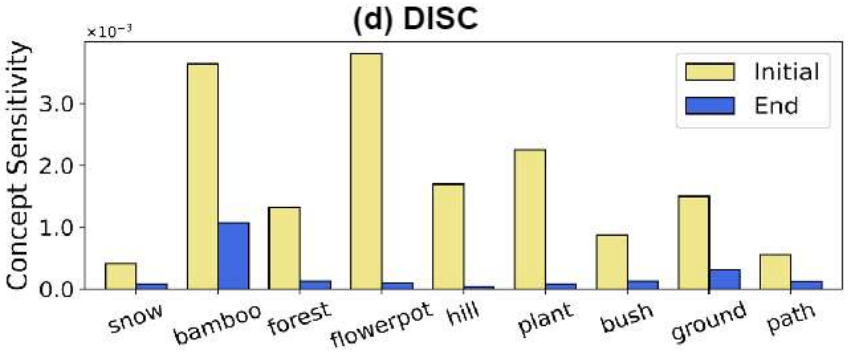


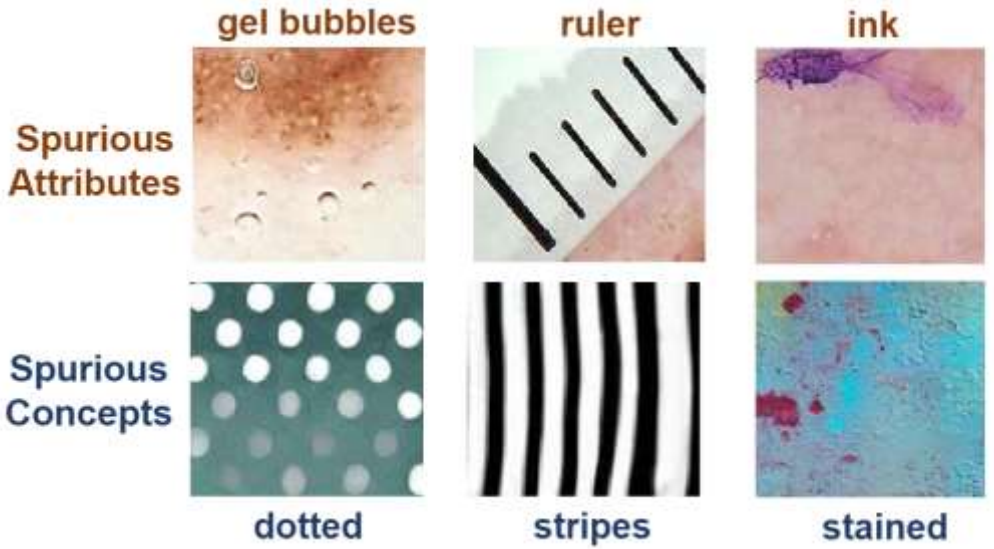
(b) Failure-Directions

Word	Forest	Tree	Wild	Branch
Score (normalized)	1.0	0.72	0.21	0.09

(c) CCE

(d) DISC



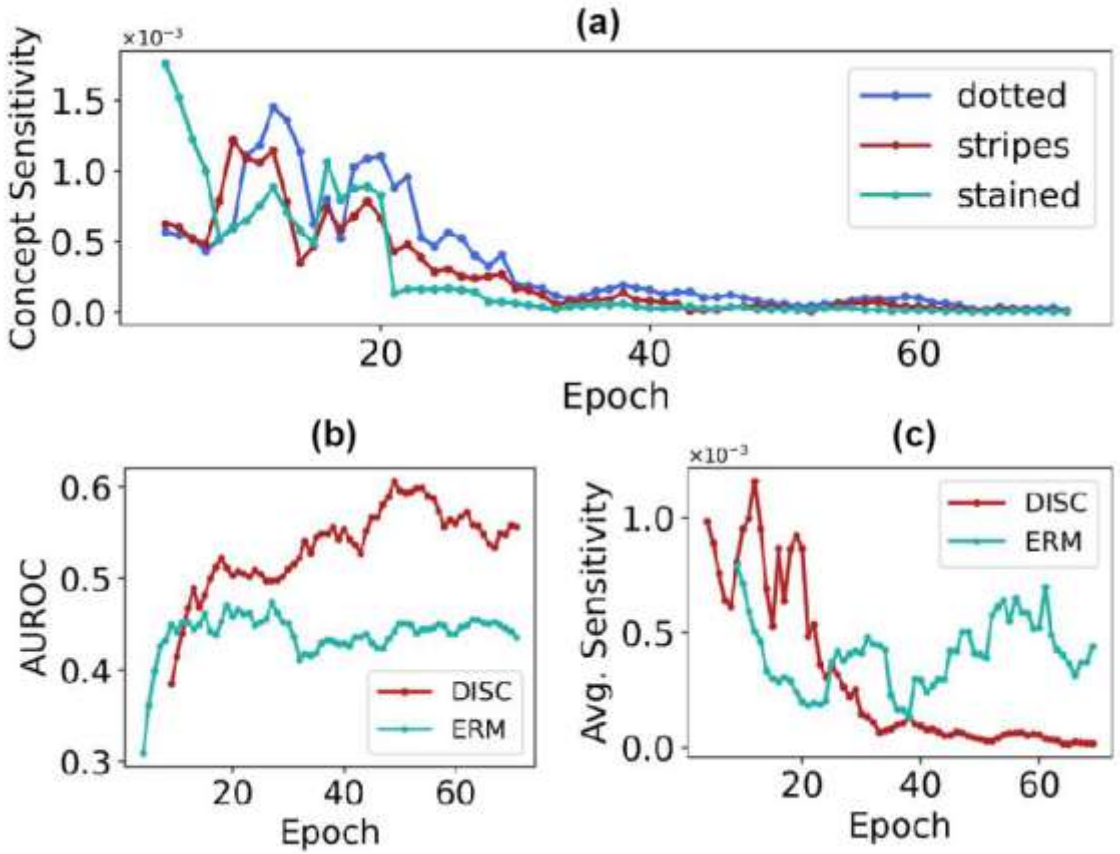


① The concept sensitivity faithfully reflects the spurious correlations in the training data.

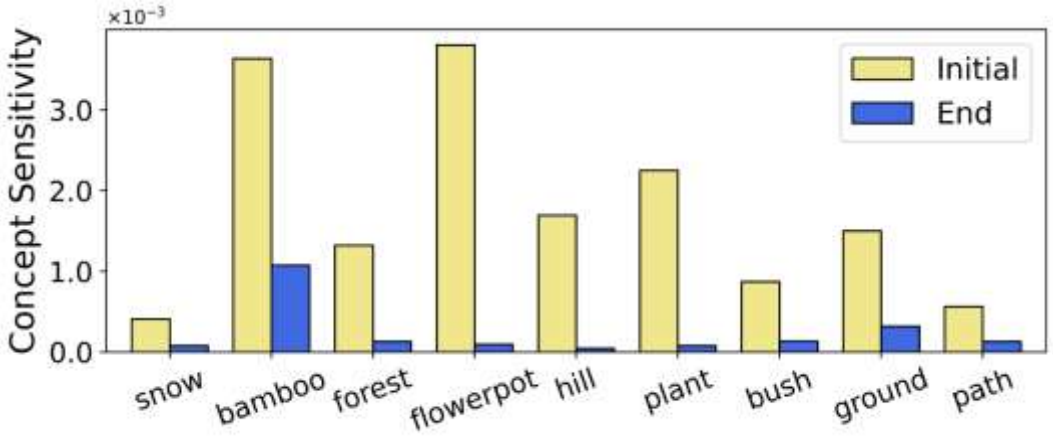
② The interpretations are also robust when the ground truth spurious concepts doesn't exist in the concept bank.

Experiments

—— The training dynamics of DISC



ISIC dataset



Waterbirds dataset

① The concept sensitivity reflects the extent of a model being affected by spurious bias, which helps probe the current model state

② The reduction of average sensitivity indicates that the model weight has reached a “sweet spot”

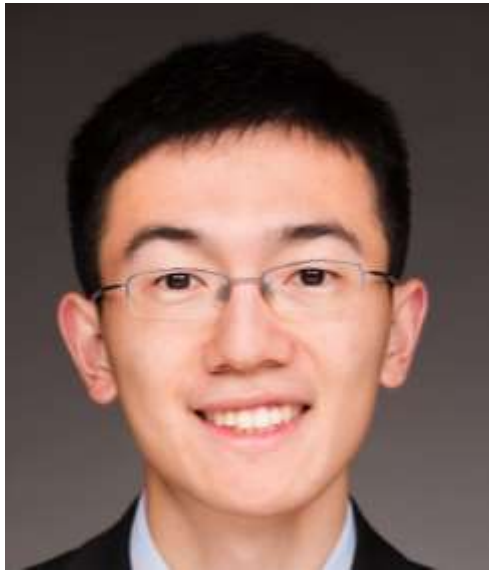
Limitations and future works

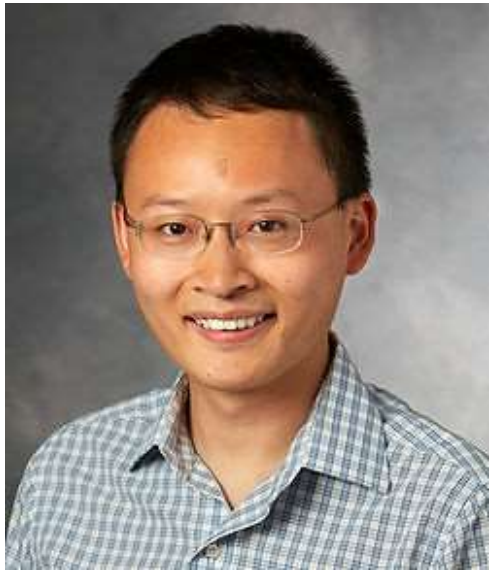
- **Limitations**
  - Computing the CAVs requires extra computational resource
  - The generative model may have its own bias

- **Future Works**
  - Use generative models for OOD generalization
  - More faithful concept bank
  - DISC for NLP tasks or more complicated image tasks

- **The Future: Multimodality**
  - How to leverage multimodal models for a specific task?
  - How to leverage multimodal models for interpretability or fairness?
  - More modalities : video, tabular data, scientific data (molecules, time series, physics), …

Thanks!





Shirley Wu
shirwu@cs.stanford.edu

Mert Yuksekgonul

Linjun Zhang

James Zou

DISC

https://github.com/Wuyxin/DISC

https://arxiv.org/abs/2305.00650

We also thank Serina Chang, Zhi Huang, Lingjiao Chen, Boyang Deng, Ruocheng Wang, Yang Zheng at Stanford University and the anonymous reviewers at ICML2023 conference.