ConceptScope: Characterizing Dataset Bias via Disentangled Visual Concepts

NeurIPS 2025

Jinho Choi

KAIST AI
Kim Jaechul Graduate School

Hyesu Lim

KAIST AI
Kim Jaechul Graduate School

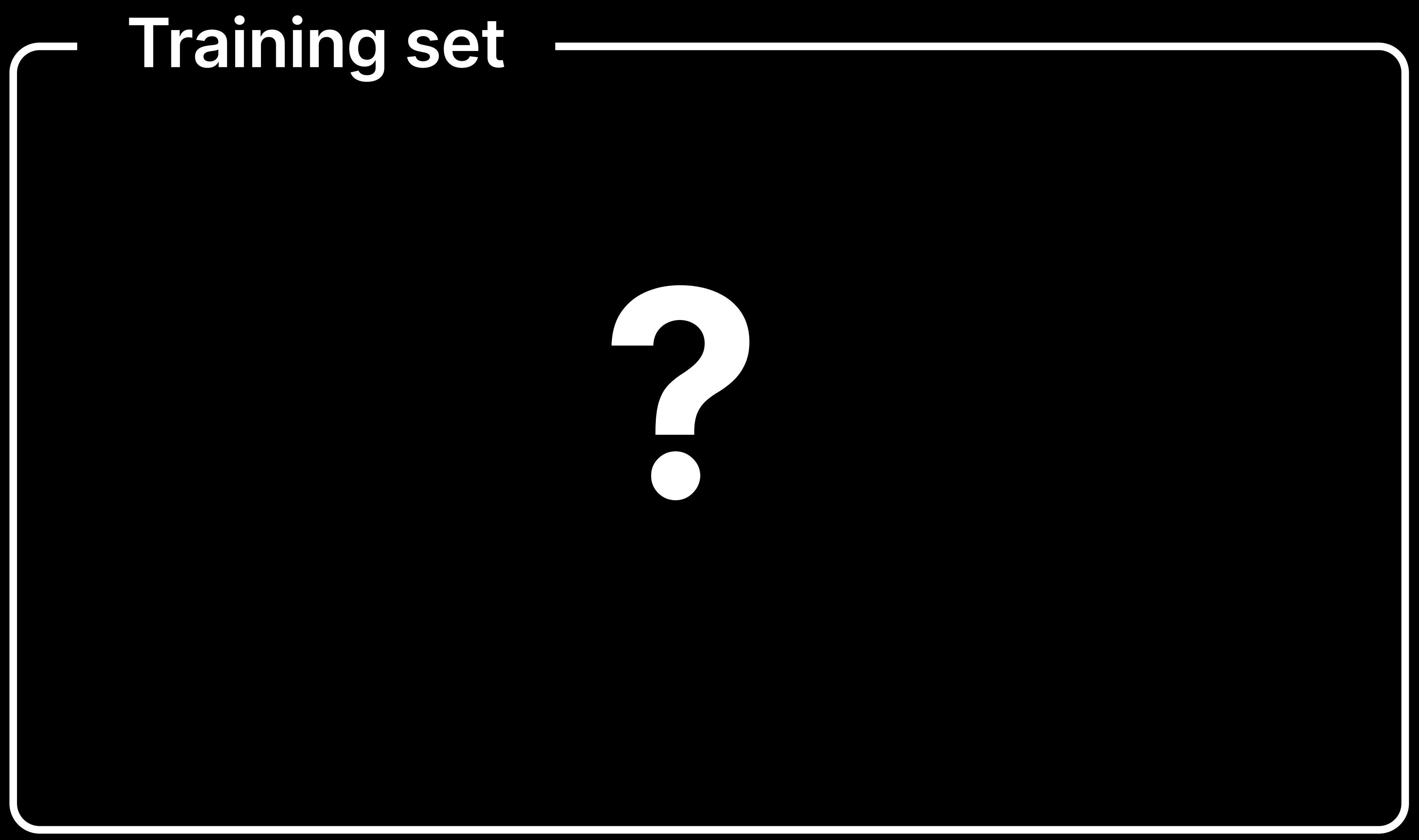
Steffen Schneider

HELMHOLTZ MUNICH

Jaegul Choo

KAIST AI
Kim Jaechul Graduate School

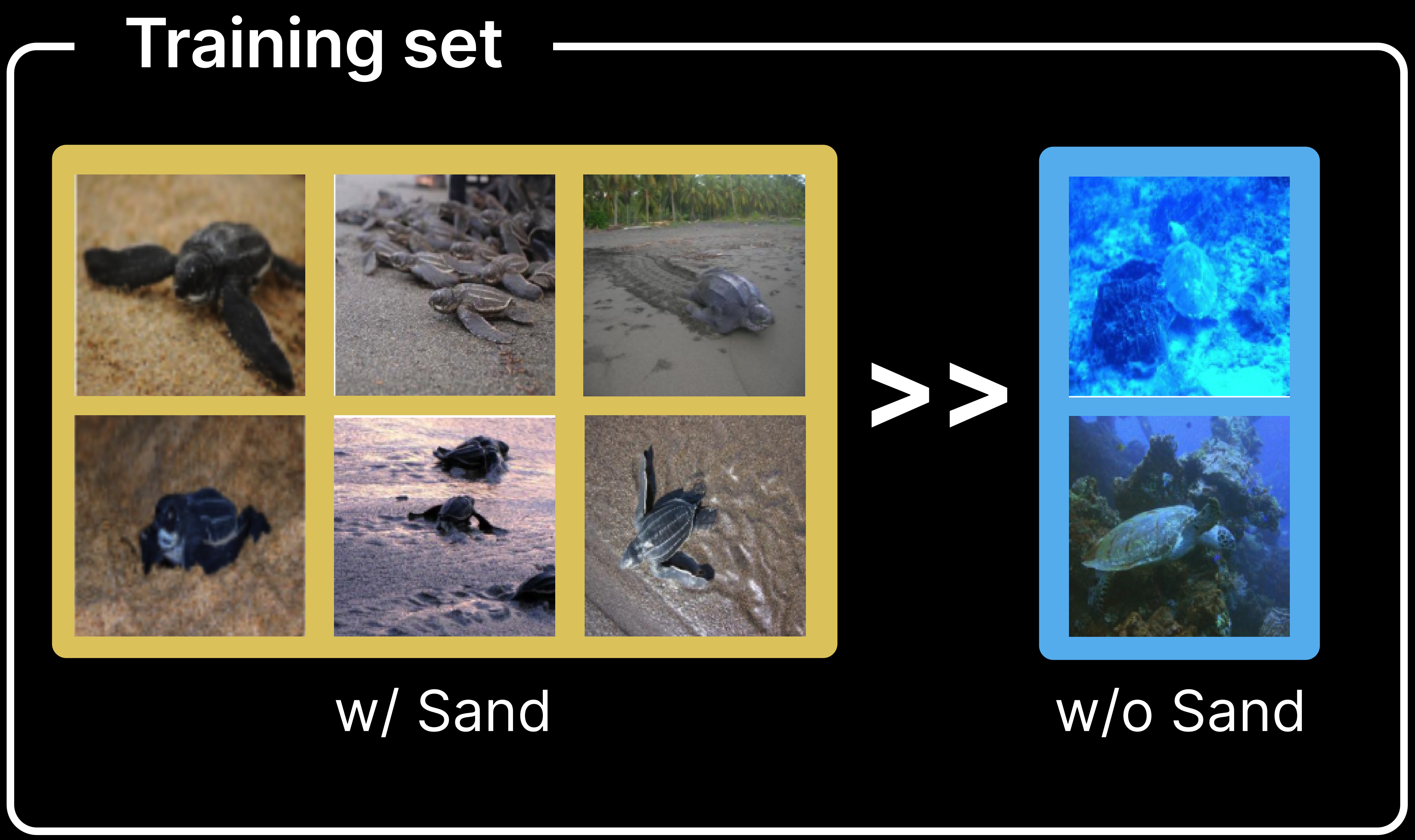
Motivation



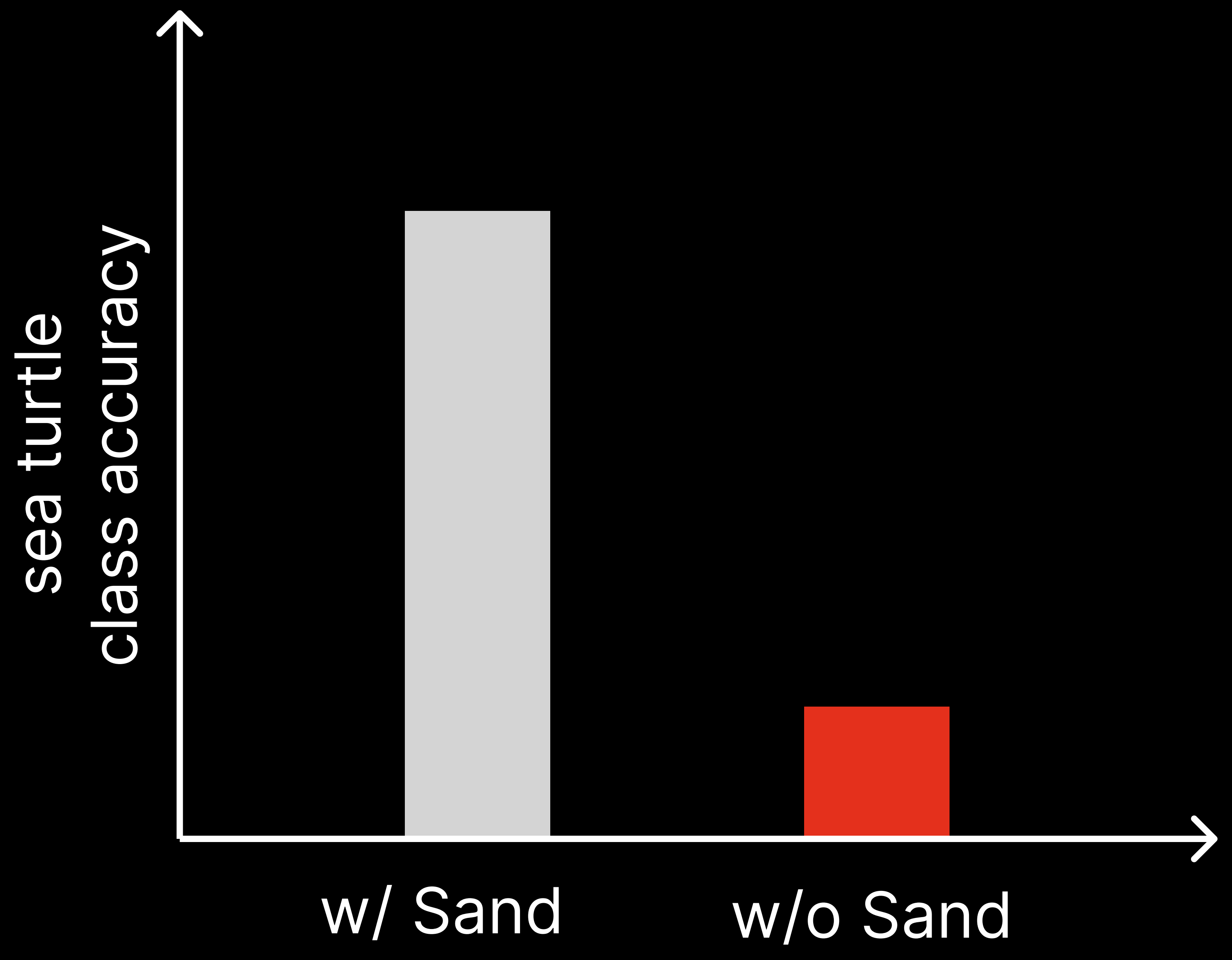
Training set

?

Motivation

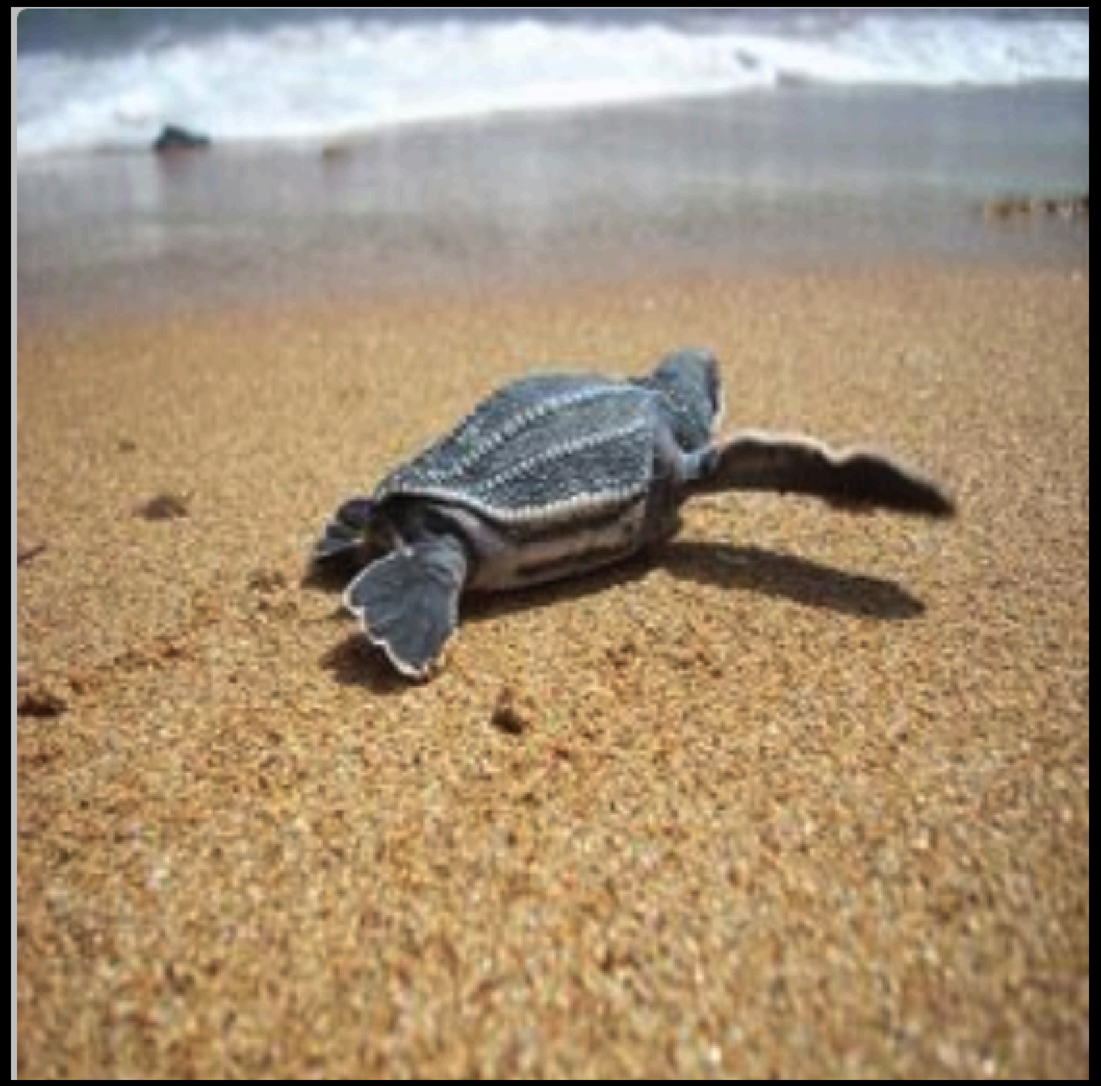


Collection bias exits in datasets



Dataset bias lead to model bias

Existing Approach



👨‍💻

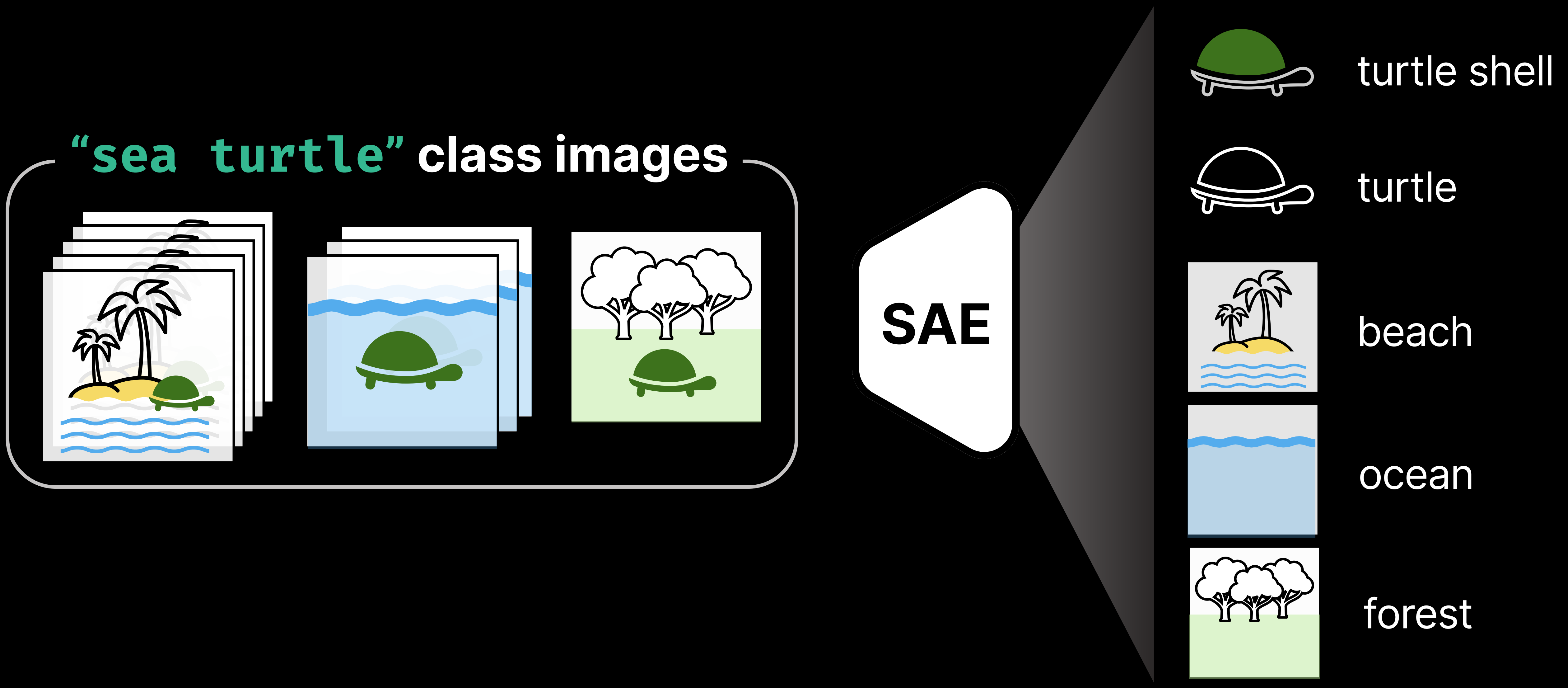
A small sea turtle crawls across the sandy beach toward the ocean waves

👩‍💻

A newly hatched leatherback turtle makes a dash for the water

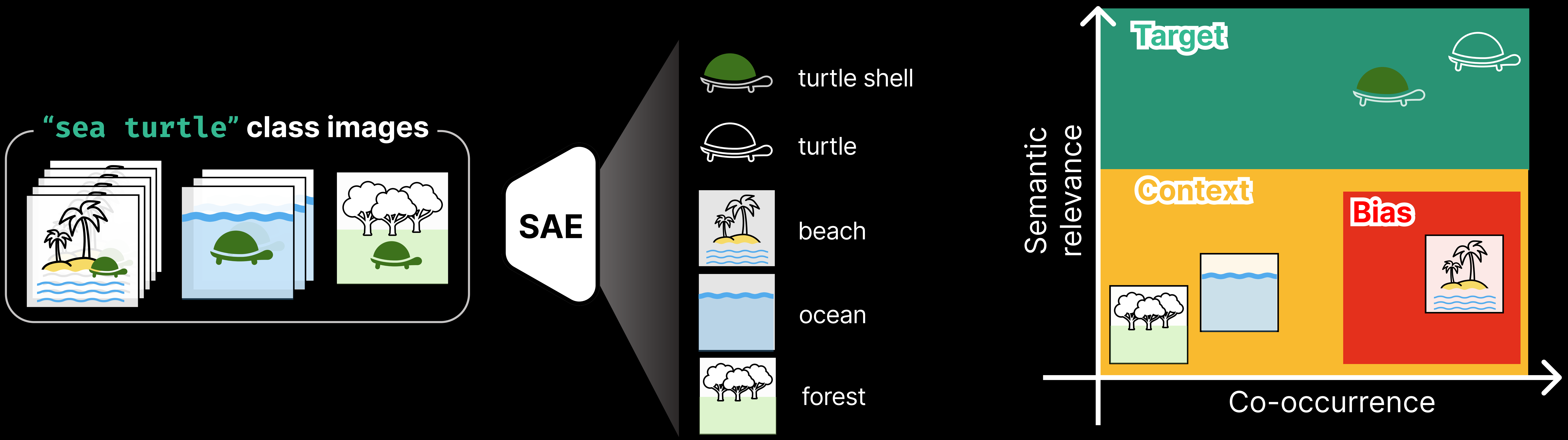
Our Approach: ConceptScpe

Sparse Autoencoder (SAE) as a concept extractor



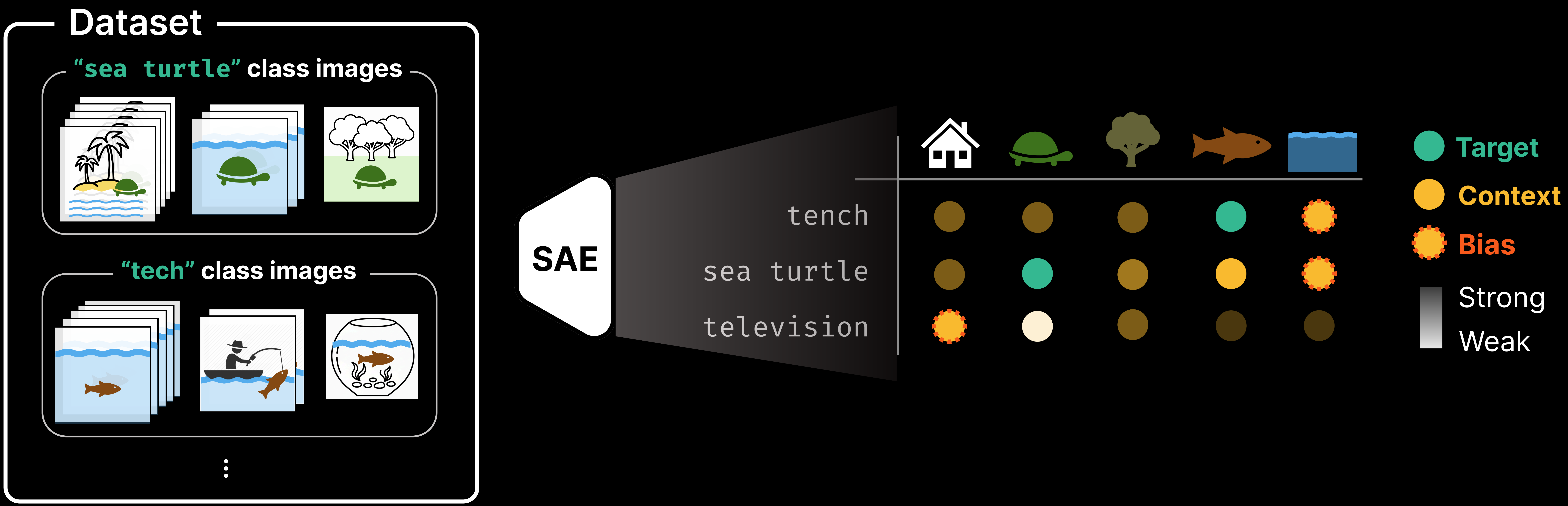
Our Approach: ConceptScpe

Sparse Autoencoder (SAE) as a concept extractor



Our Approach: ConceptScpe

Sparse Autoencoder (SAE) as a concept extractor



Dataset

“sea turtle” class images

“tech” class images

⋮

SAE

tench

sea turtle

television

● Target

● Context

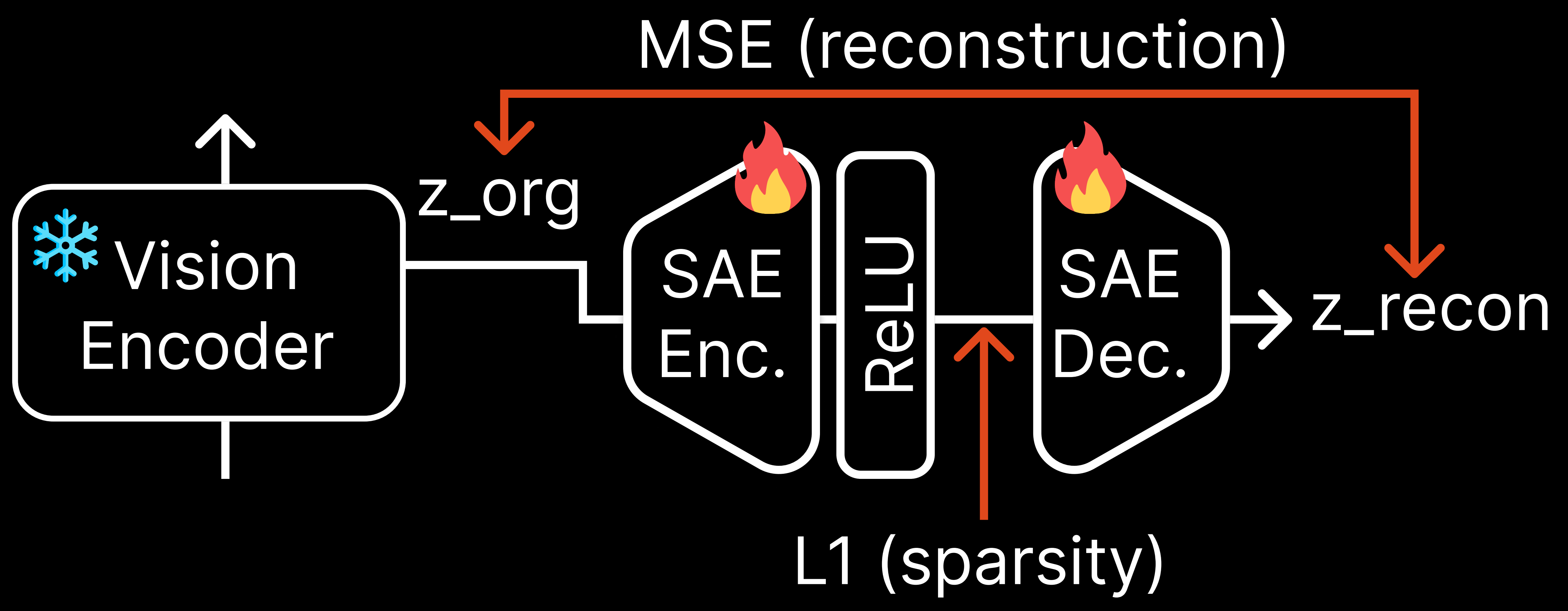
● Bias

Strong

Weak

ConceptScpe: Characterizing dataset bias via visual concepts

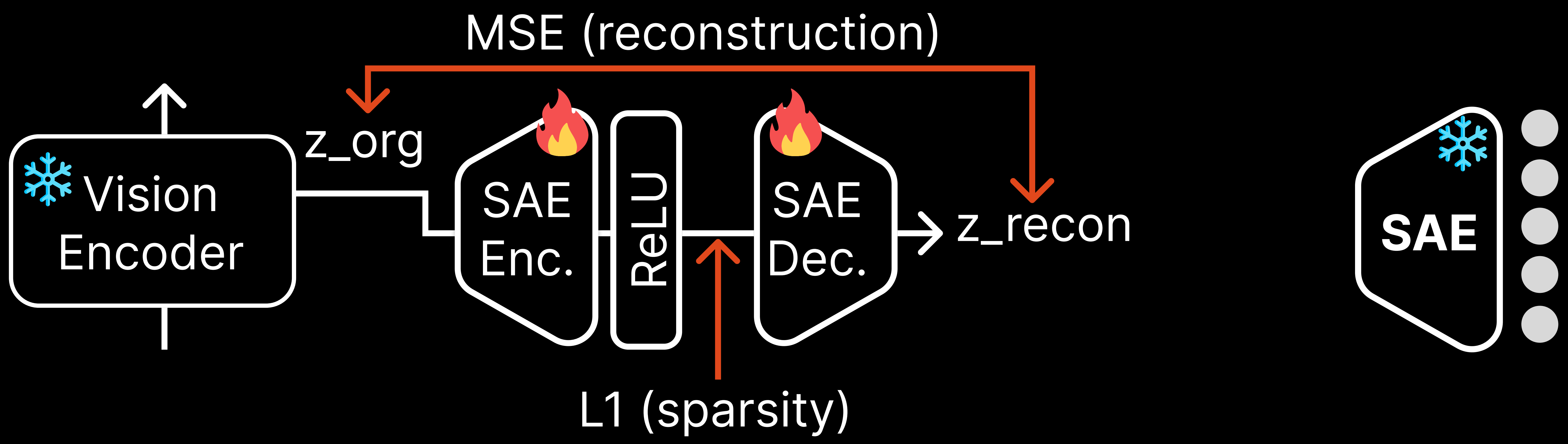
Training SAE and Interpreting Concepts



Lim, Hyesu, et al. "Sparse autoencoders reveal selective remapping of visual concepts during adaptation." (ICLR 2025)

ConceptScpe: Characterizing dataset bias via visual concepts

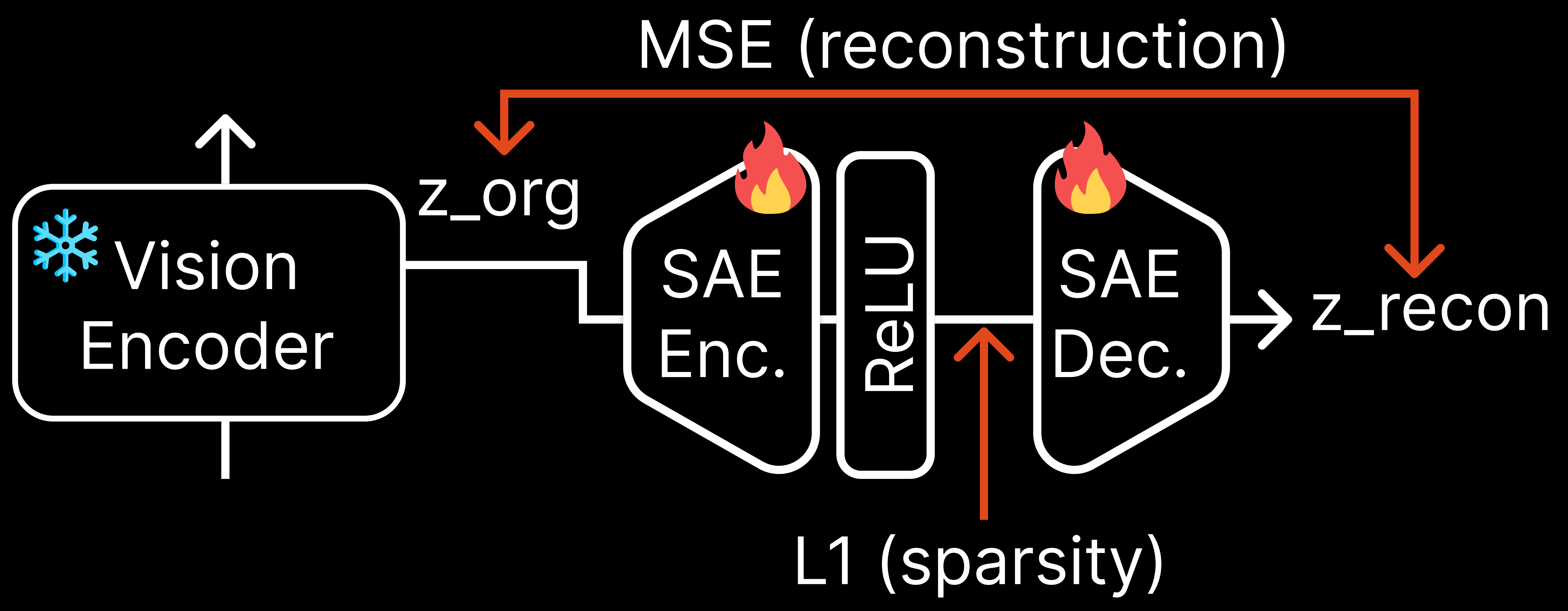
Training SAE and Interpreting Concepts

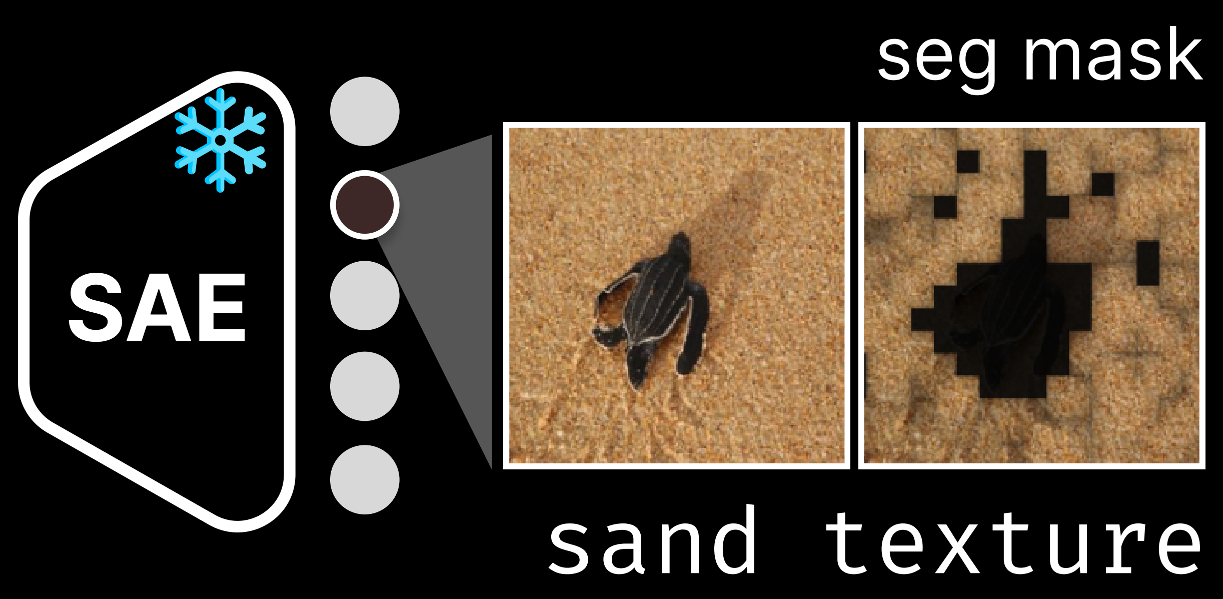


Lim, Hyesu, et al. "Sparse autoencoders reveal selective remapping of visual concepts during adaptation." (ICLR 2025)

ConceptScpe: Characterizing dataset bias via visual concepts

Training SAE and Interpreting Concepts

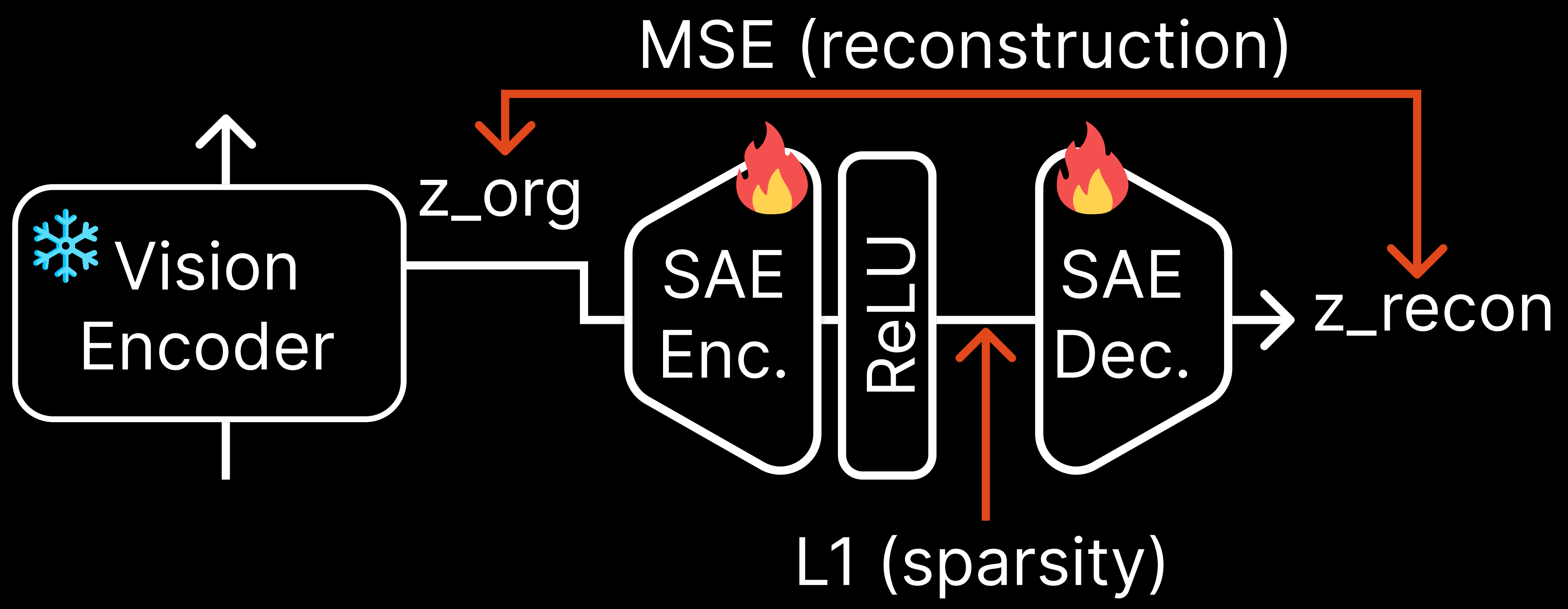


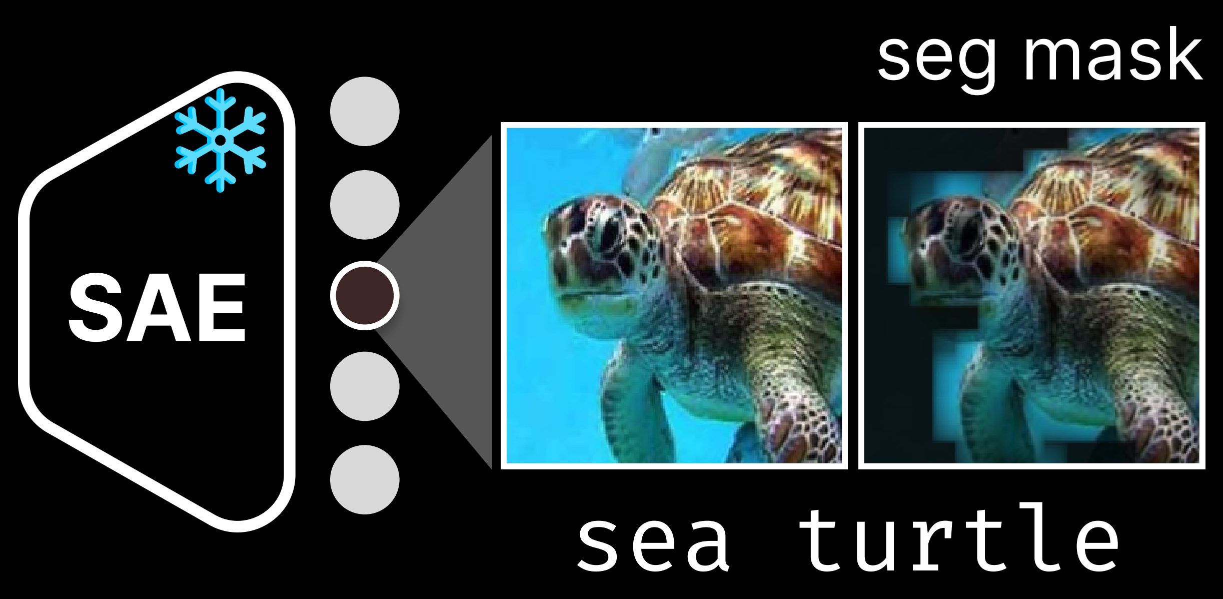


Lim, Hyesu, et al. "Sparse autoencoders reveal selective remapping of visual concepts during adaptation." (ICLR 2025)

ConceptScpe: Characterizing dataset bias via visual concepts

Training SAE and Interpreting Concepts

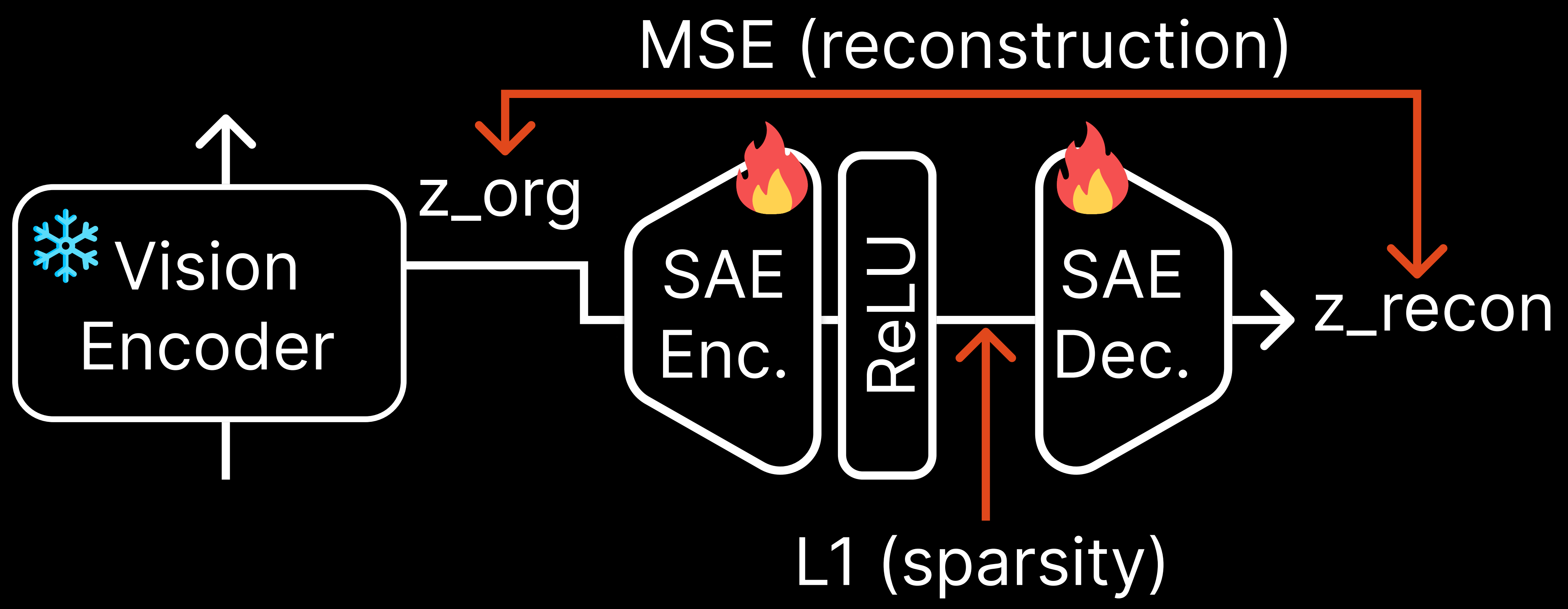


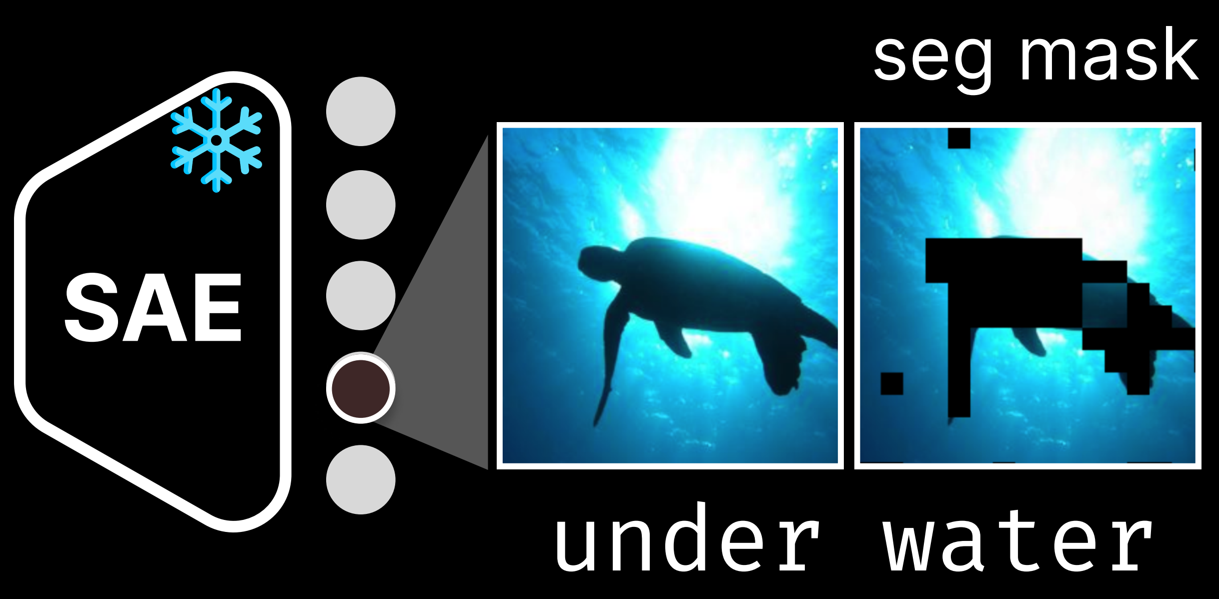


Lim, Hyesu, et al. "Sparse autoencoders reveal selective remapping of visual concepts during adaptation." (ICLR 2025)

ConceptScpe: Characterizing dataset bias via visual concepts

Training SAE and Interpreting Concepts

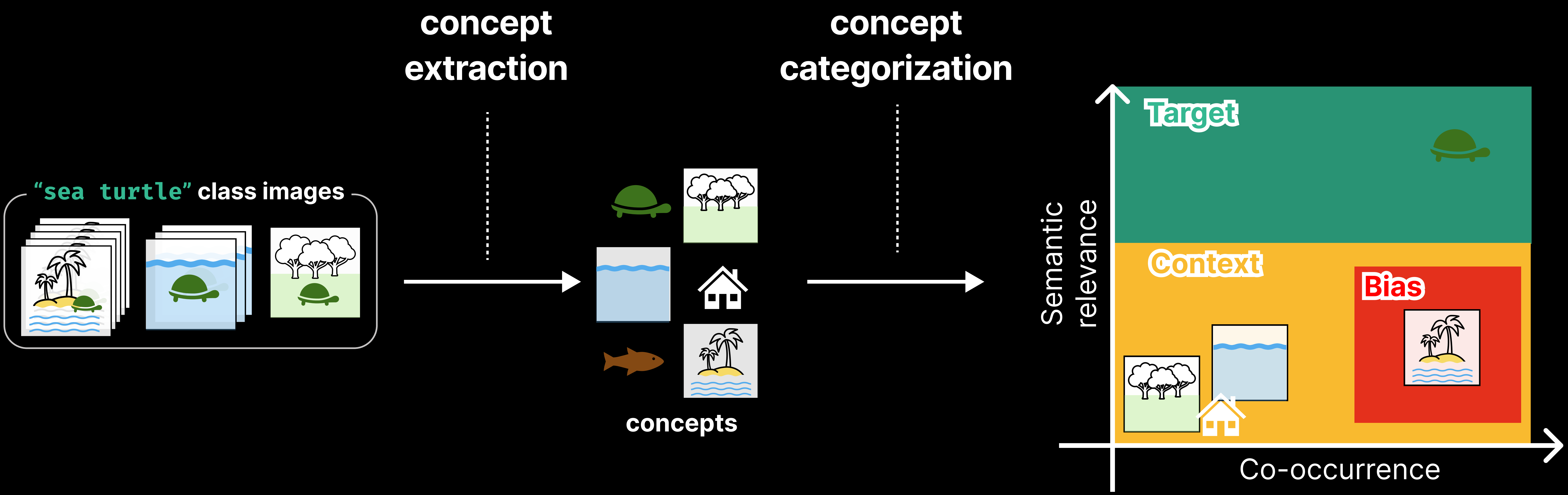




Lim, Hyesu, et al. "Sparse autoencoders reveal selective remapping of visual concepts during adaptation." (ICLR 2025)

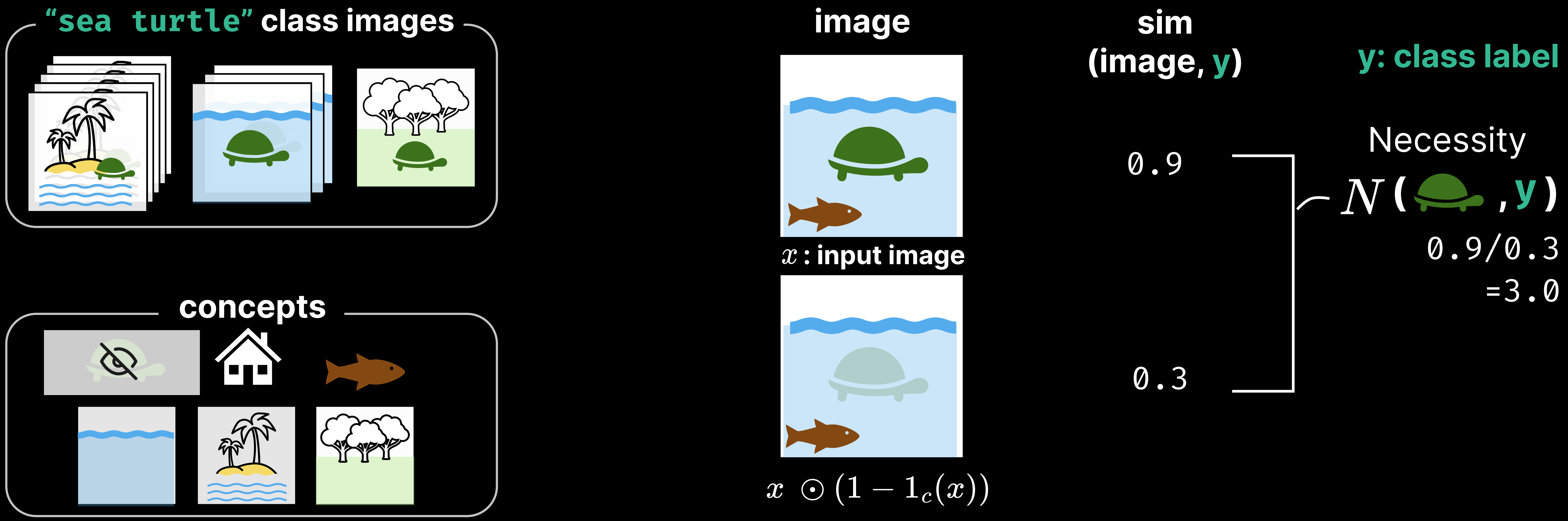
ConceptScpe: Characterizing dataset bias via visual concepts

Categorizing Concepts



ConceptScpe: Characterizing dataset bias via visual concepts

Computing alignment score



“sea turtle” class images

concepts

image

x : input image

$x \odot (1 - 1_c(x))$

sim (image, y)

0.9

0.3

y: class label

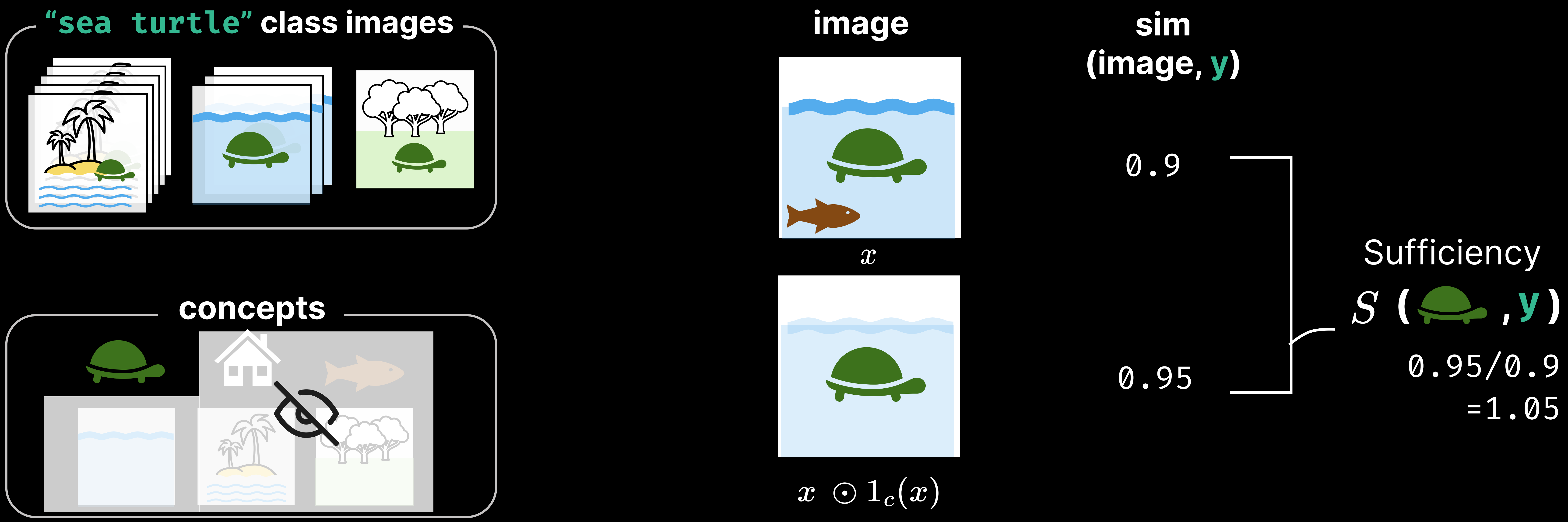
Necessity

N(🐢, y)

0.9/0.3 =3.0

ConceptScpe: Characterizing dataset bias via visual concepts

Computing alignment score



"sea turtle" class images

concepts

image

x

$x \odot 1_c(x)$

sim (image, y)

0.9

0.95

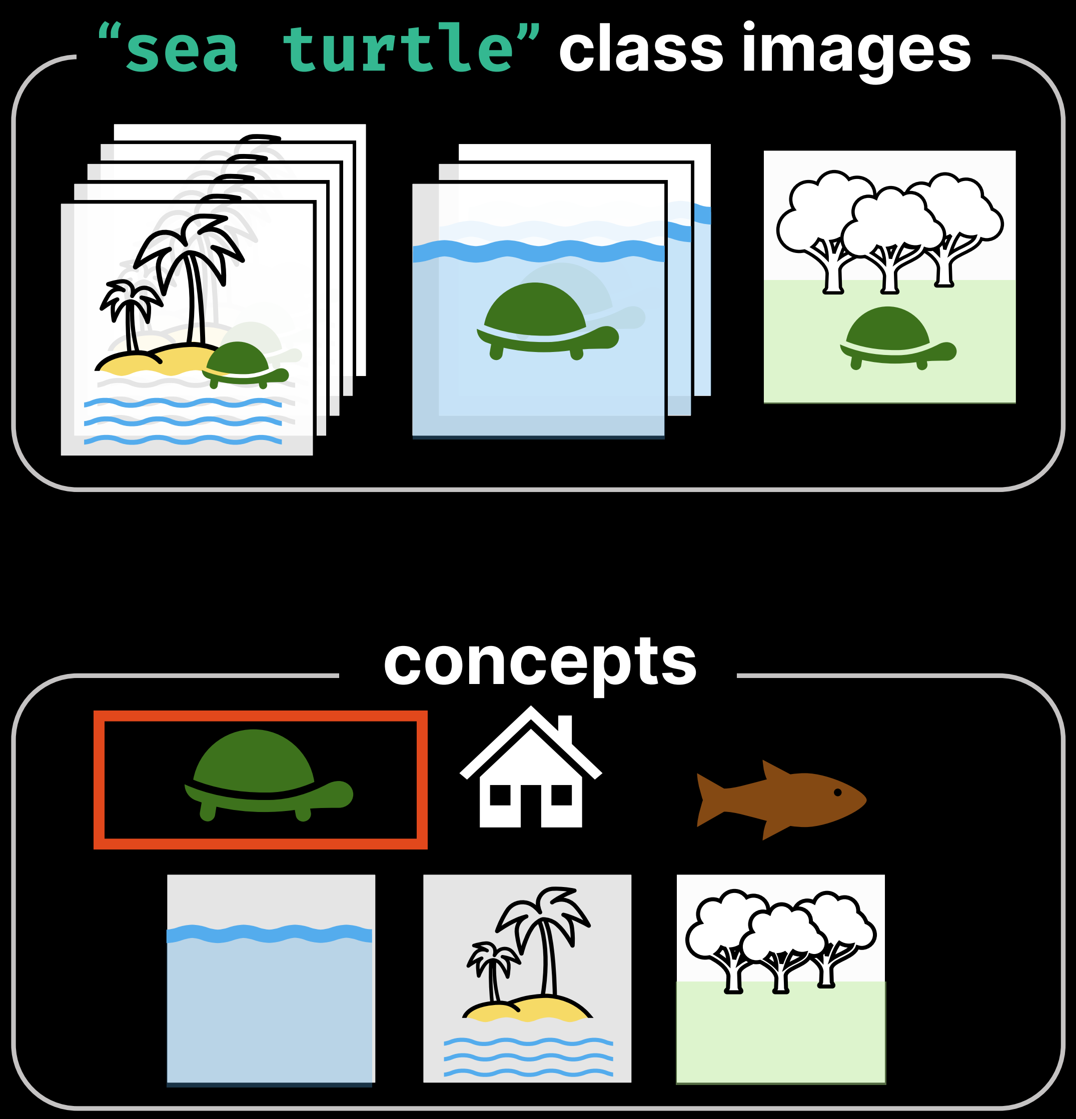
Sufficiency

S (, y)

0.95/0.9
=1.05

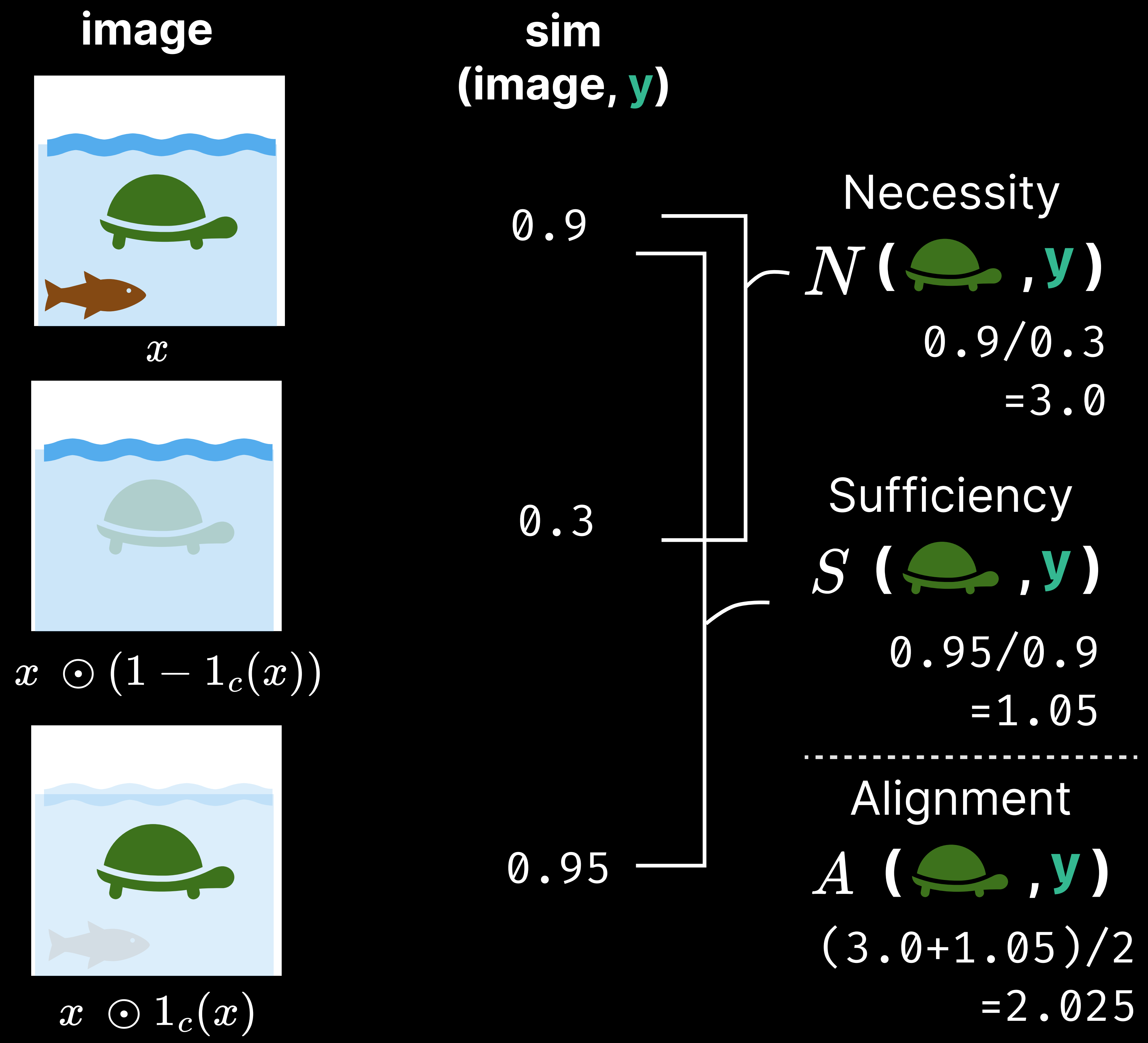
ConceptScpe: Characterizing dataset bias via visual concepts

Computing alignment score



“sea turtle” class images

concepts



image

sim (image, y)

0.9

x

0.3

$x \odot (1 - 1_c(x))$

0.95

$x \odot 1_c(x)$

Necessity

$N($🐢$, y)$

0.9/0.3 =3.0

Sufficiency

$S($🐢$, y)$

0.95/0.9 =1.05

Alignment

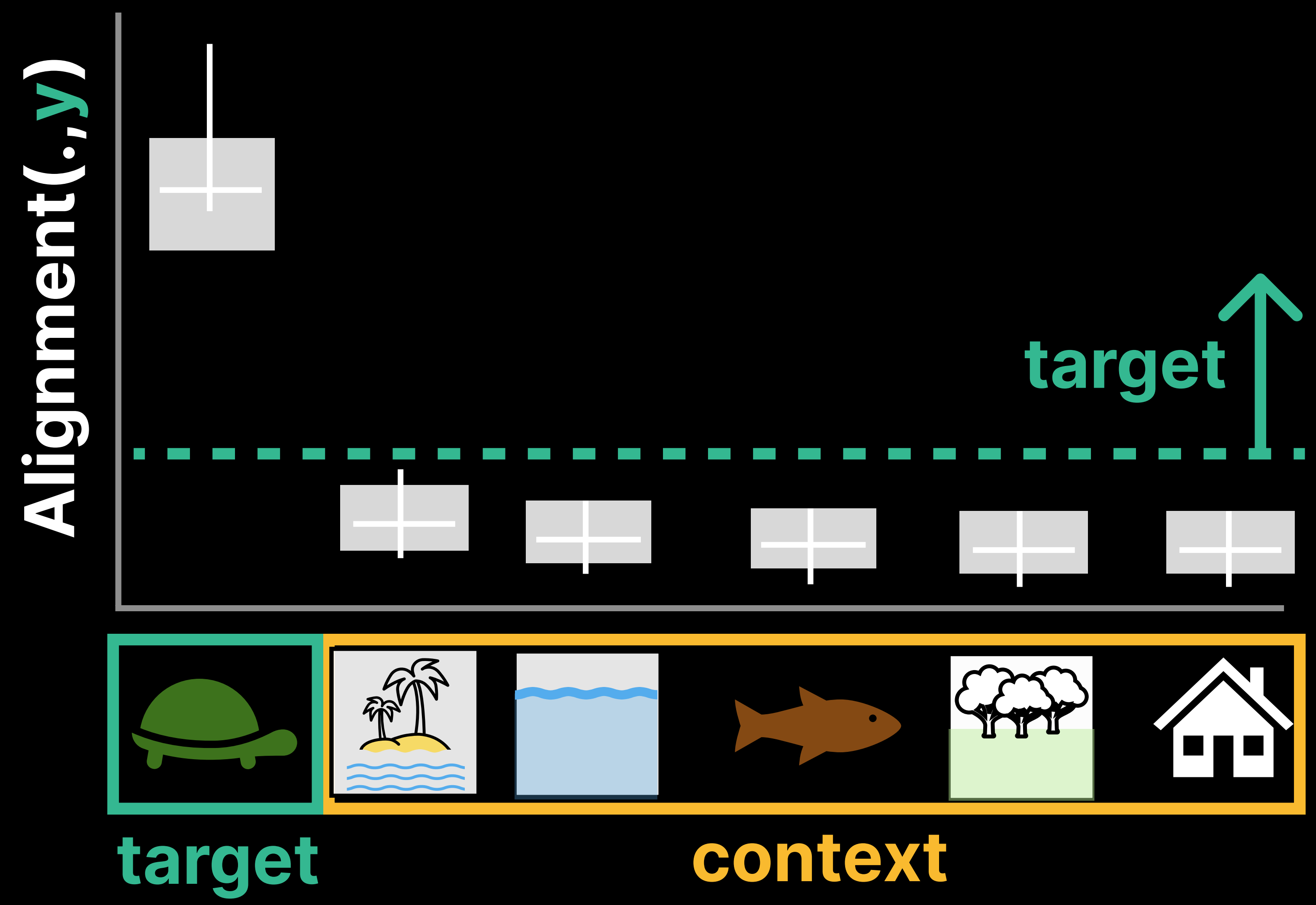
$A($🐢$, y)$

(3.0+1.05)/2 =2.025

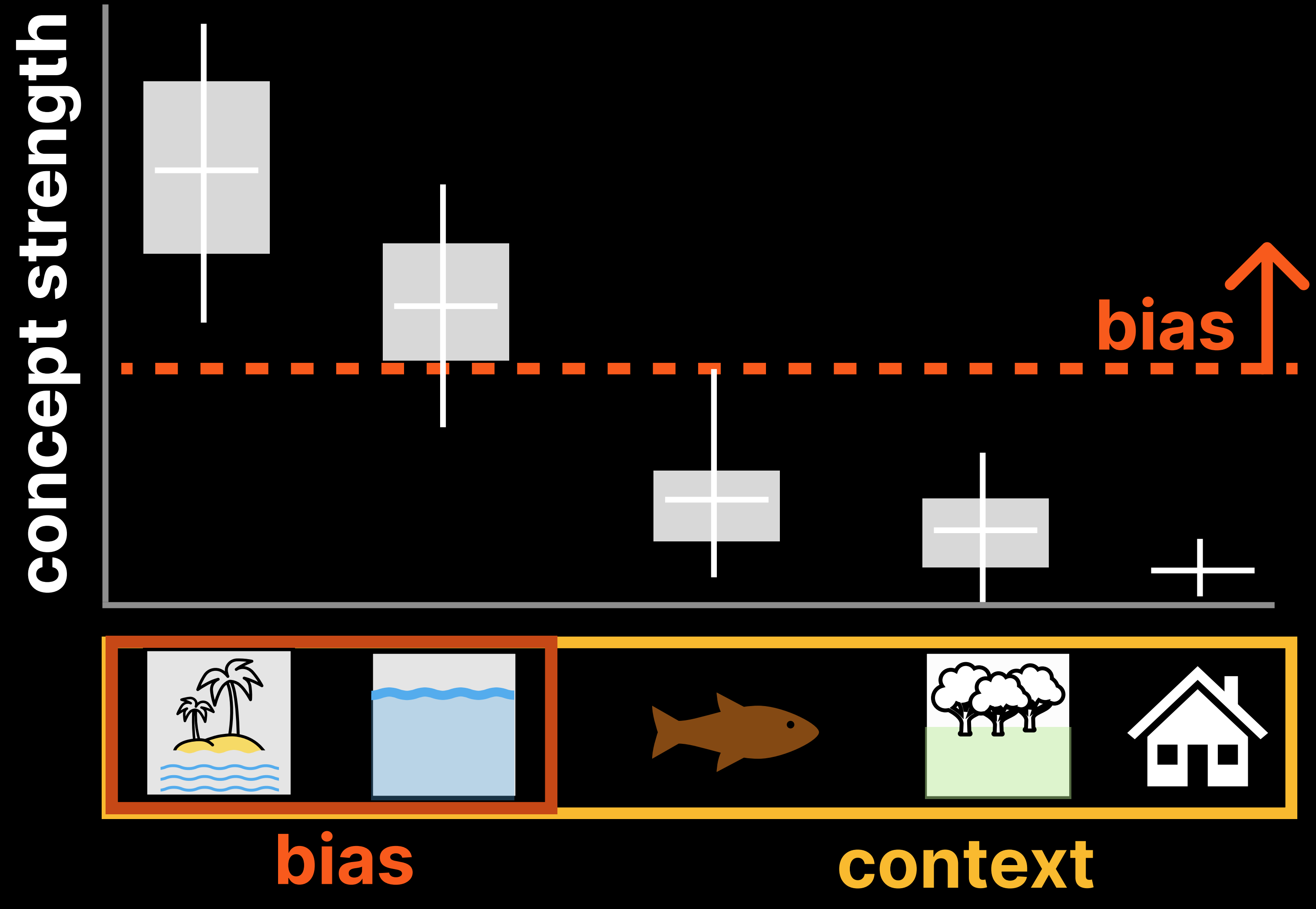
ConceptScpe: Characterizing dataset bias via visual concepts

Categorizing Concepts

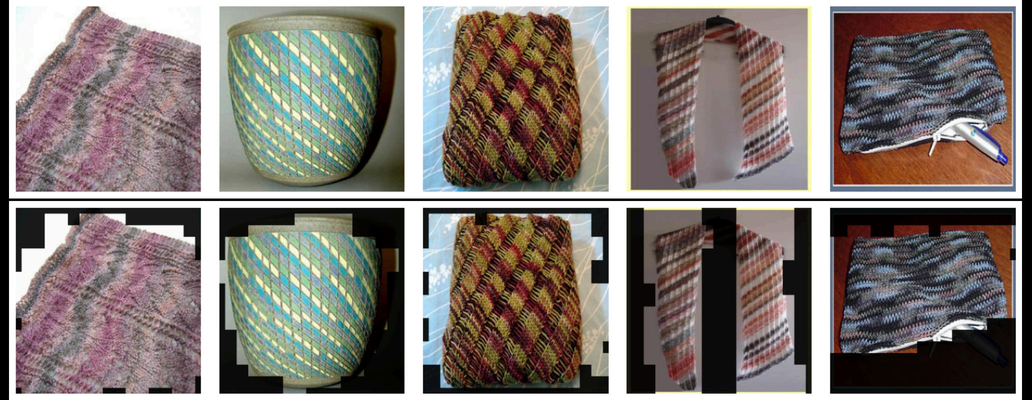
Target / context discrimination



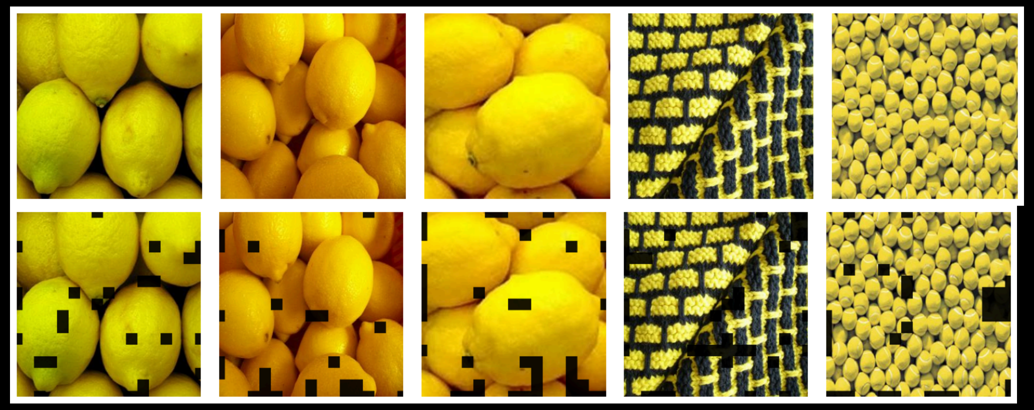
Bias concept identification



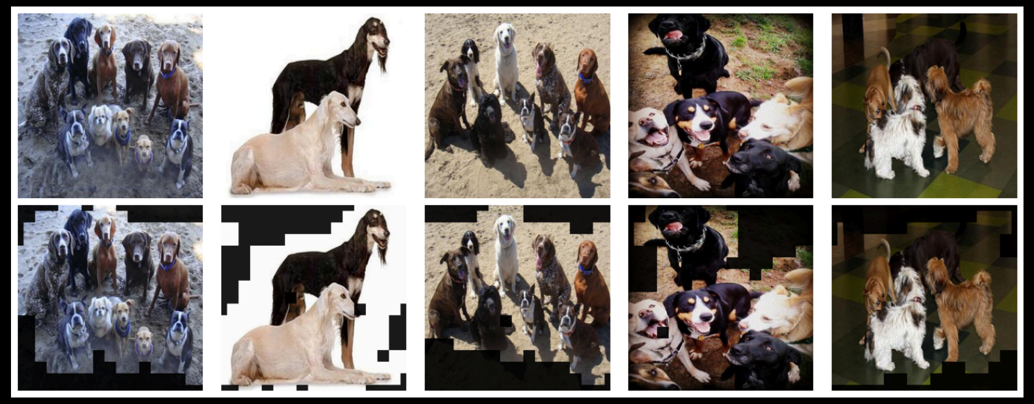
Results: SAEs can discover a wide range of visual concepts



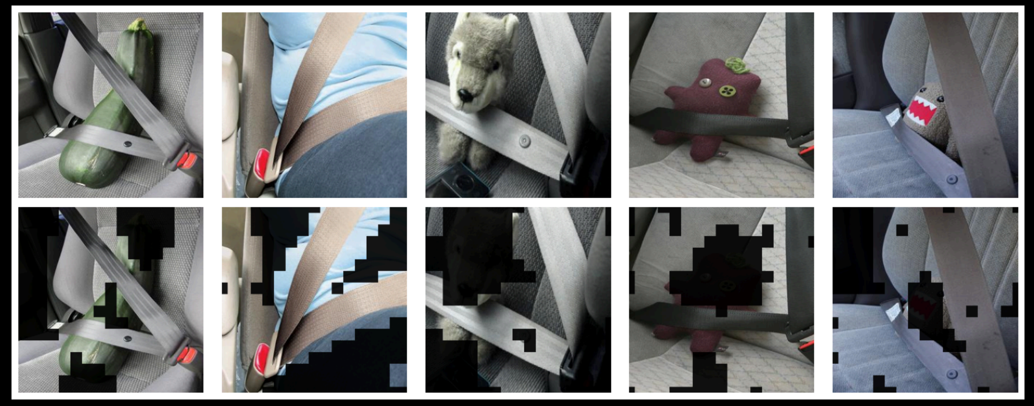
knitted fabric



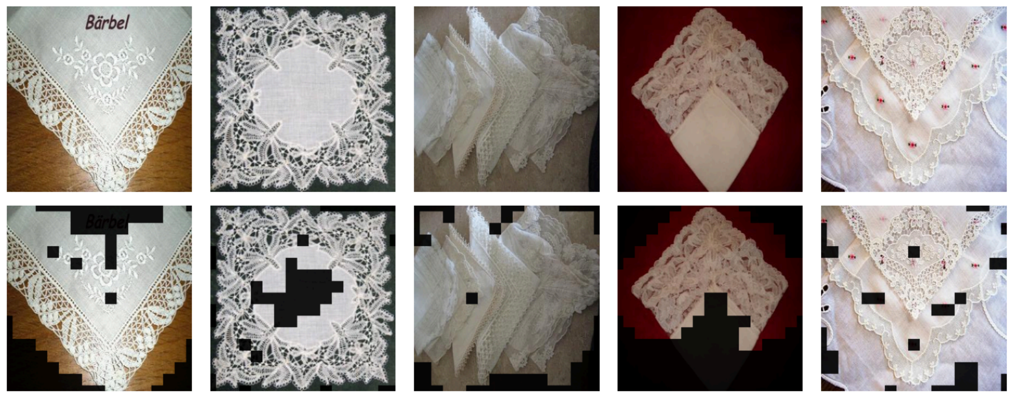
yellow objects



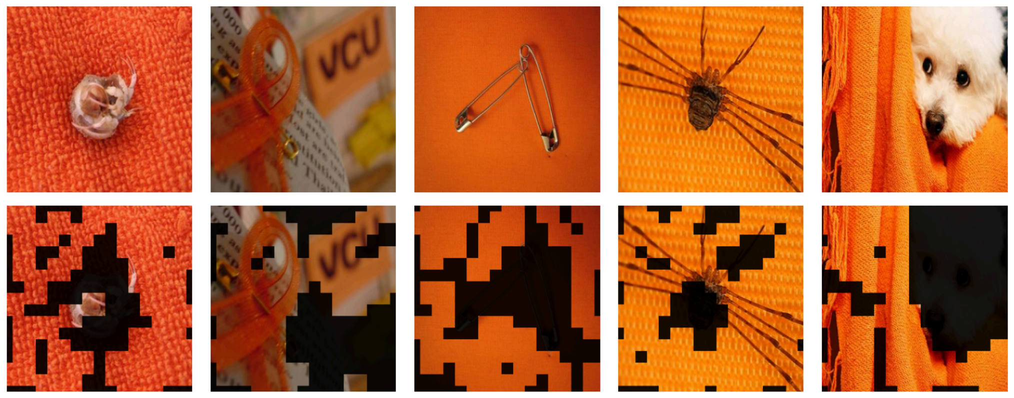
group of dogs



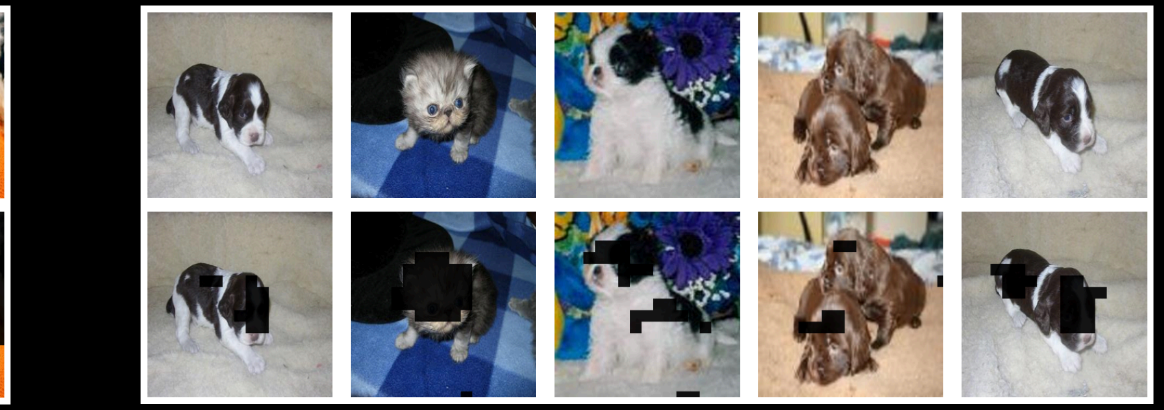
seat belts



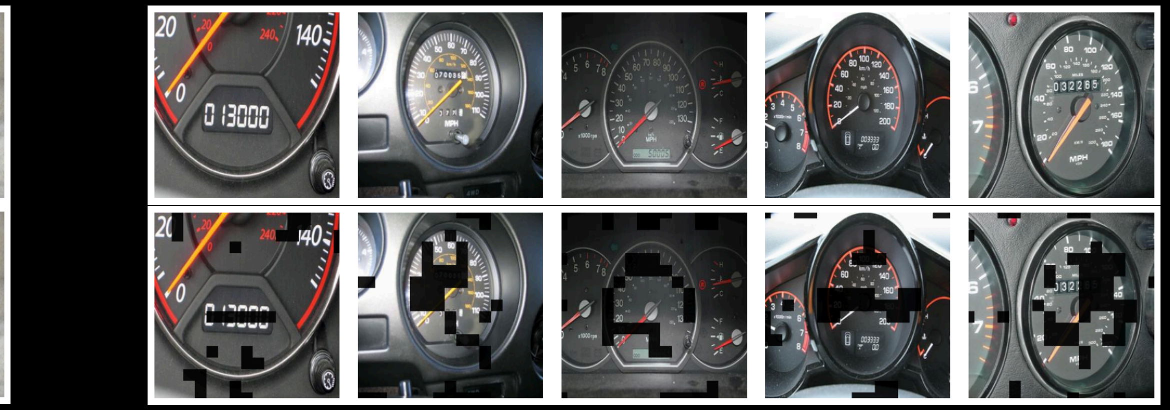
lace patterns



orange objects



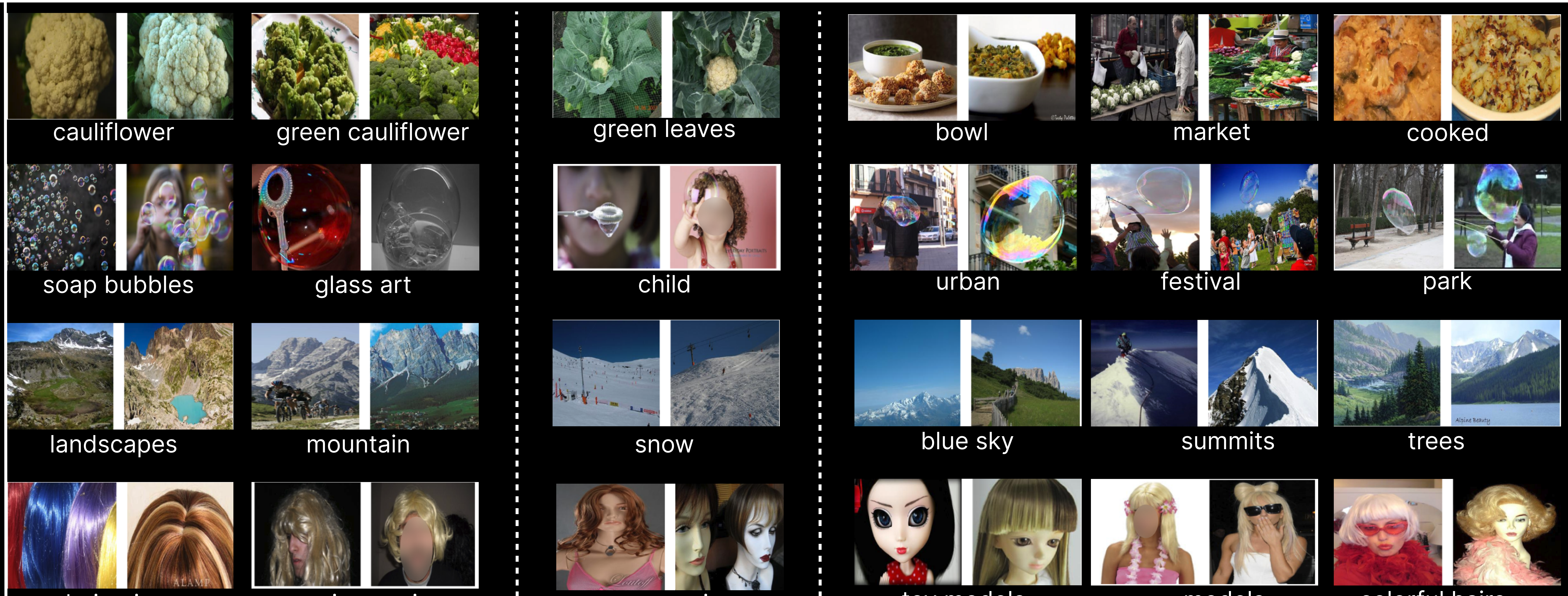
puppies



car gauges

Method	Metric	Caltech101 (Objects)	DTD (Textures)	Waterbird (Backgrounds)	CelebA (Facial Attr.)	RAF-DB (Emotions)	Stanford40 (Actions)	Average
BLIP-2	F_1	0.64±0.35	0.38±0.25	0.37±0.10	0.27±0.24	0.24±0.17	0.66±0.18	0.43
LLaVA-NeXT	F_1	0.61±0.35	0.40±0.21	0.57±0.12	0.62±0.24	0.45±0.18	0.80±0.16	0.58
ConceptScope	F_1	0.83±0.21	0.57±0.20	0.78±0.07	0.81±0.11	0.55±0.18	0.78±0.13	0.72
	AUPRC	0.89±0.19	0.57±0.23	0.83±0.09	0.85±0.13	0.59±0.21	0.82±0.15	0.76

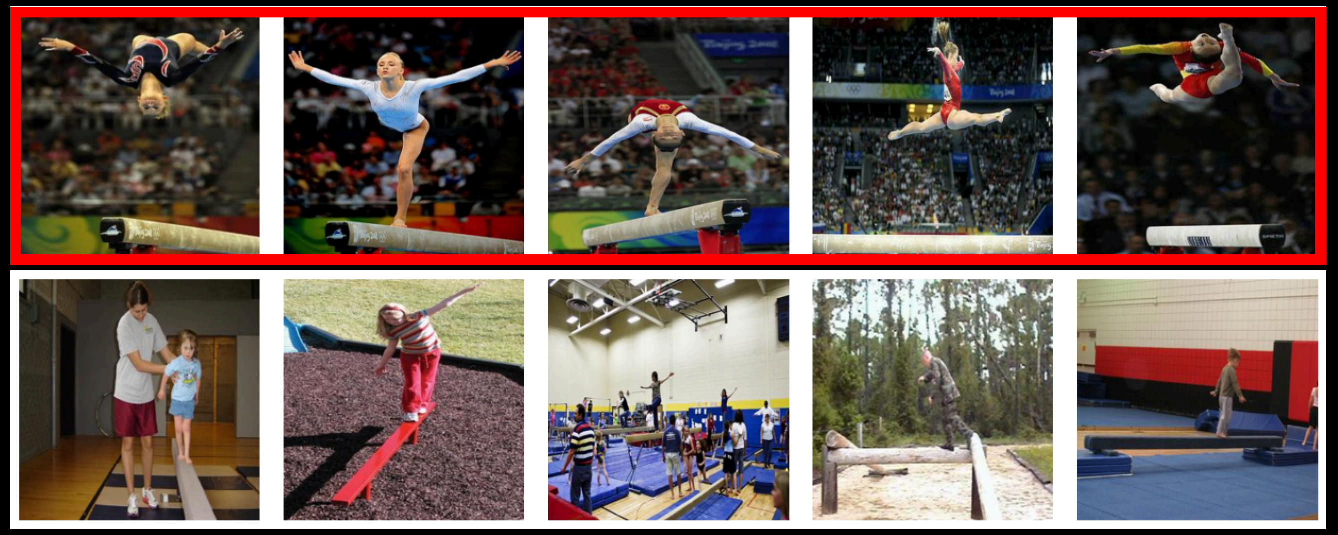
Results: ConceptScope captures diverse visual states within each class



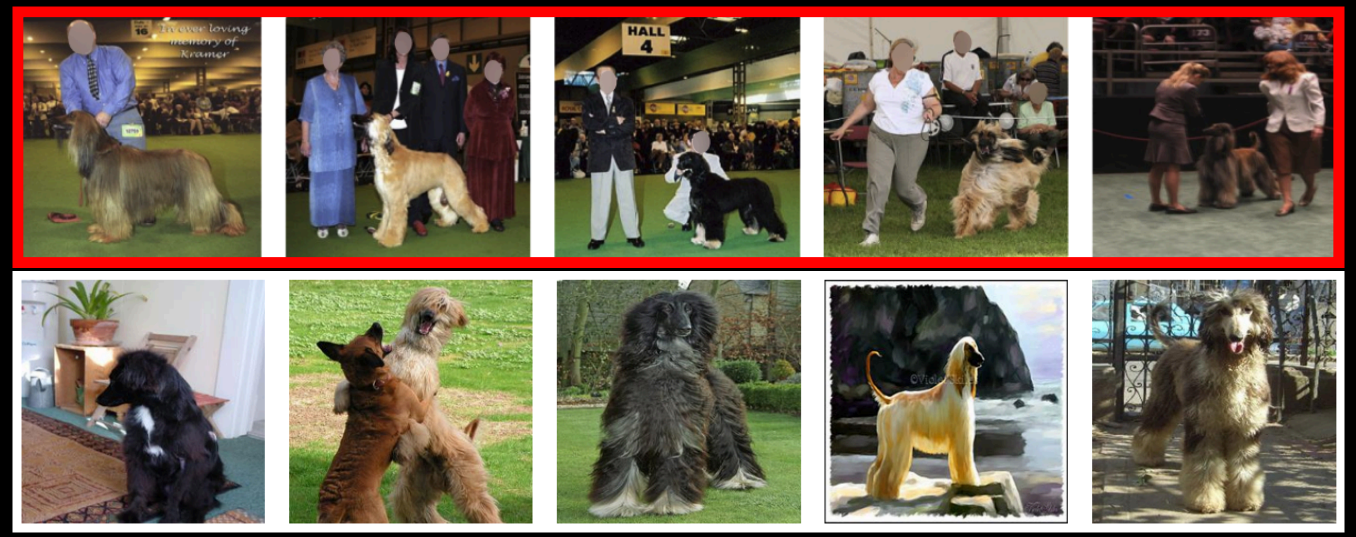
class name	target		bias	context		
cauliflower	cauliflower	green cauliflower	green leaves	bowl	market	cooked
bubble	soap bubbles	glass art	child	urban	festival	park
mountain	landscapes	mountain	snow	blue sky	summits	trees
hair wig	hair wig	wearing a wig	mannequin	toy models	models	colorful hairs

Results: ConceptScpe discovers real-world dataset bias

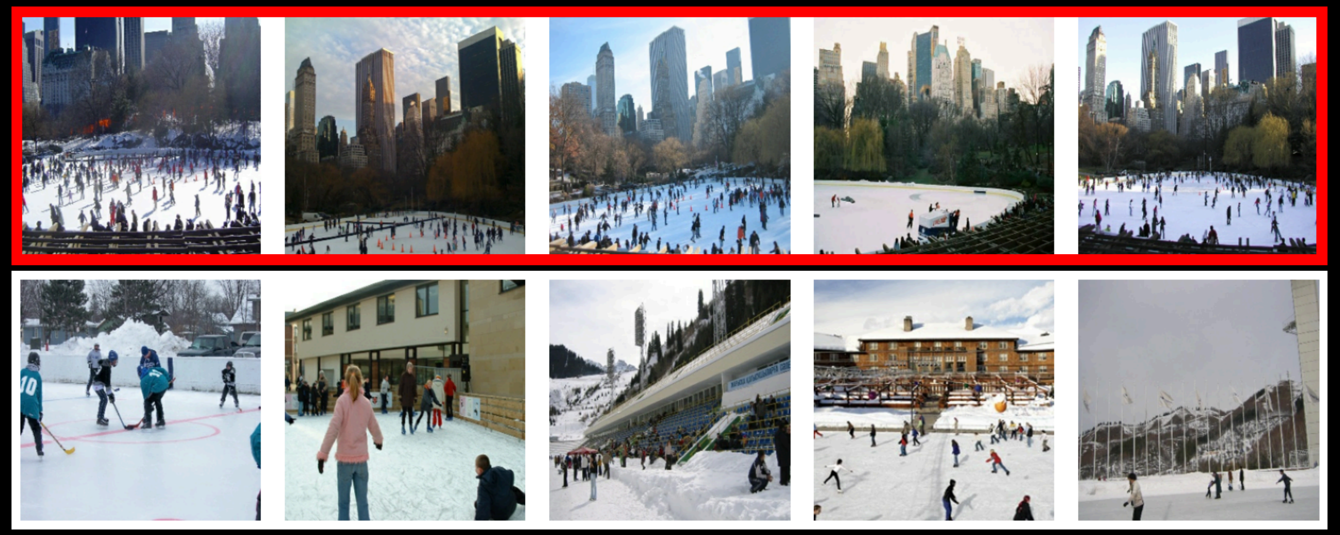
High bias ▭ Low bias ▭



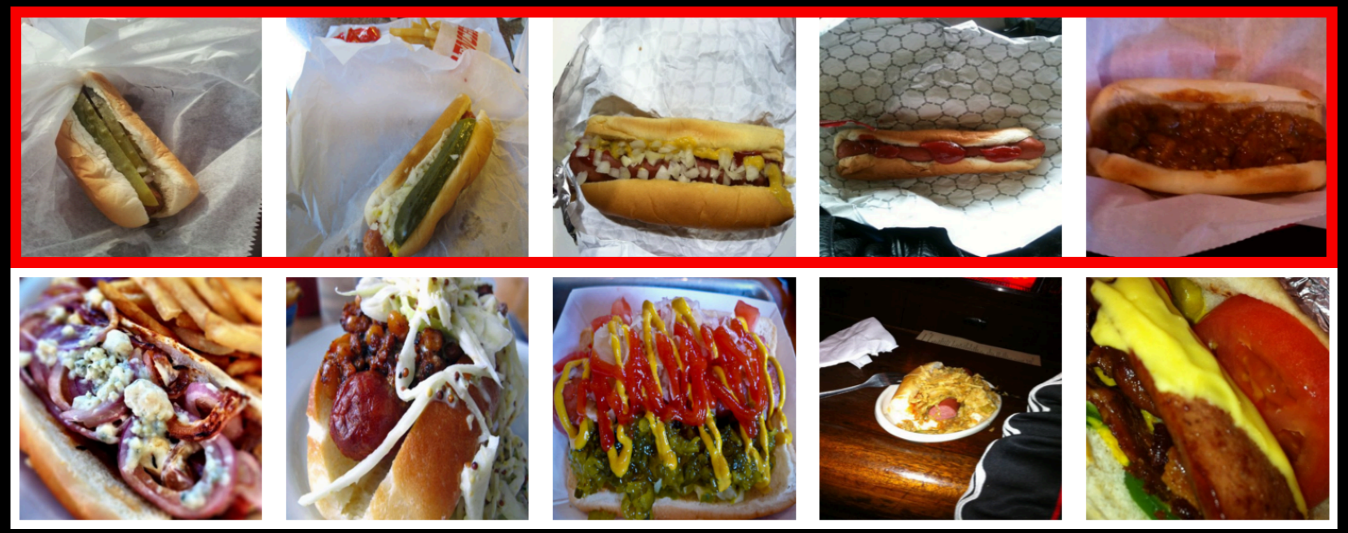
ImageNet - "**balance beam**"
biased to "**competition**"



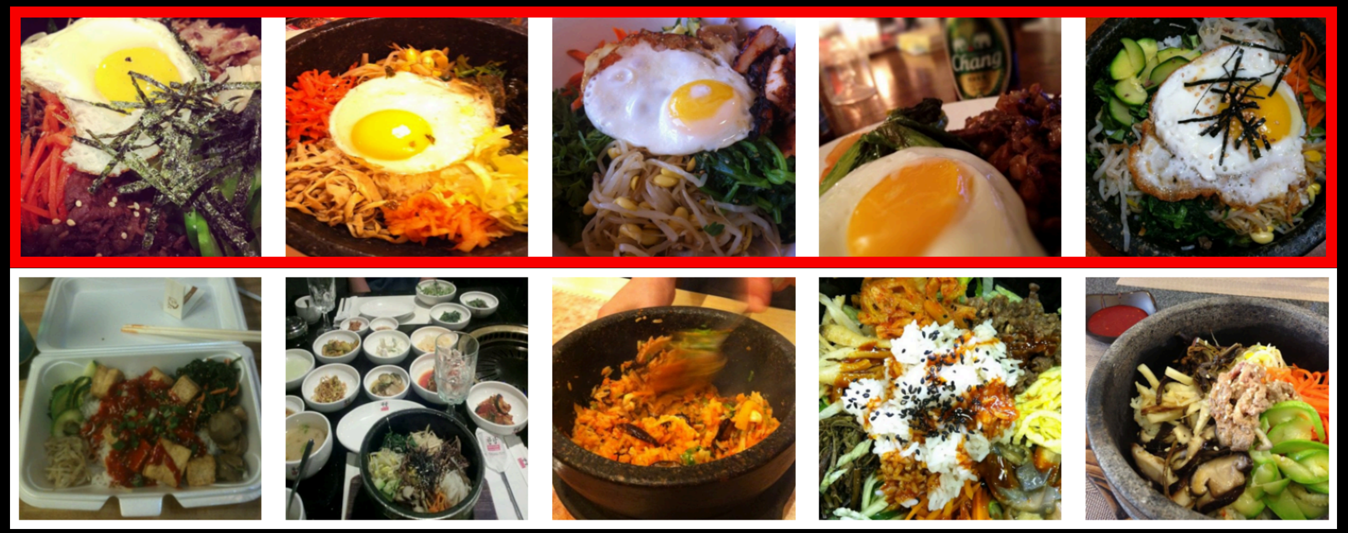
ImageNet - "**afghan hound**"
biased to "**dog show**"



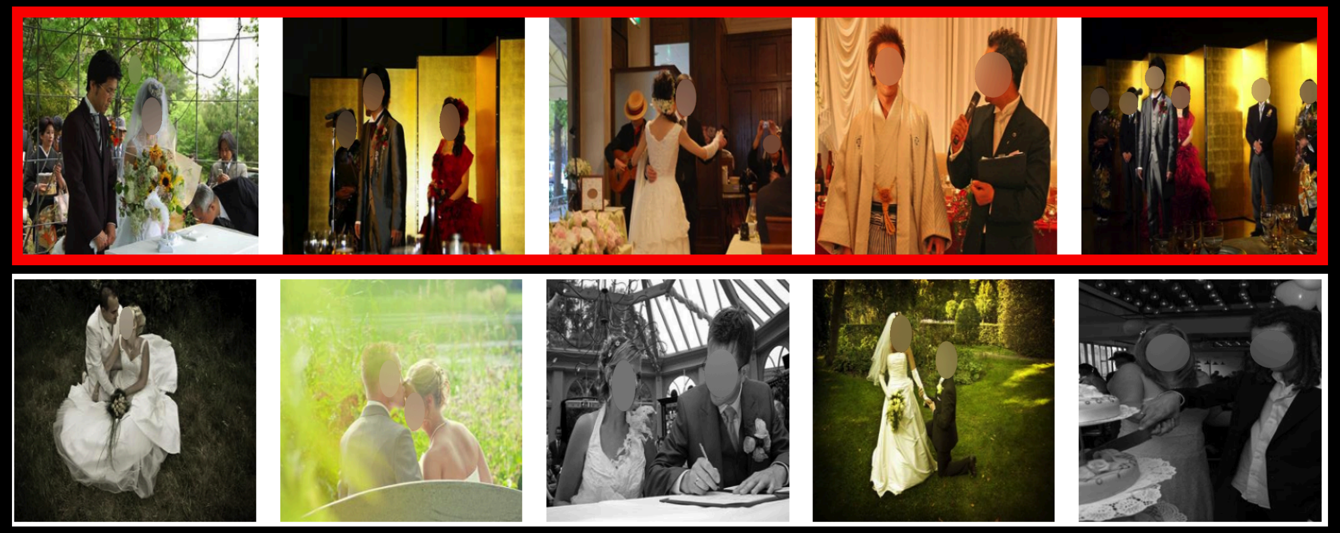
SUN397 - "**ice skating rink**" class
biased to "**New York**"



Food101 - "**hotdog**"
biased to " **food wrappers**"



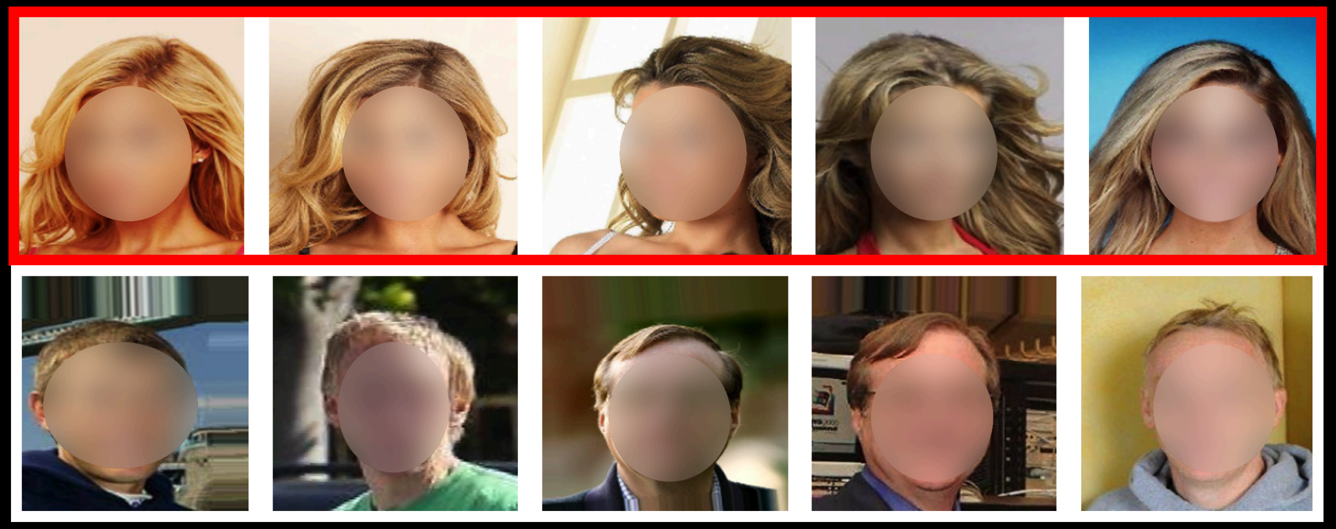
Food101 - "**bibmbap**"
biased to " **fried eggs**"



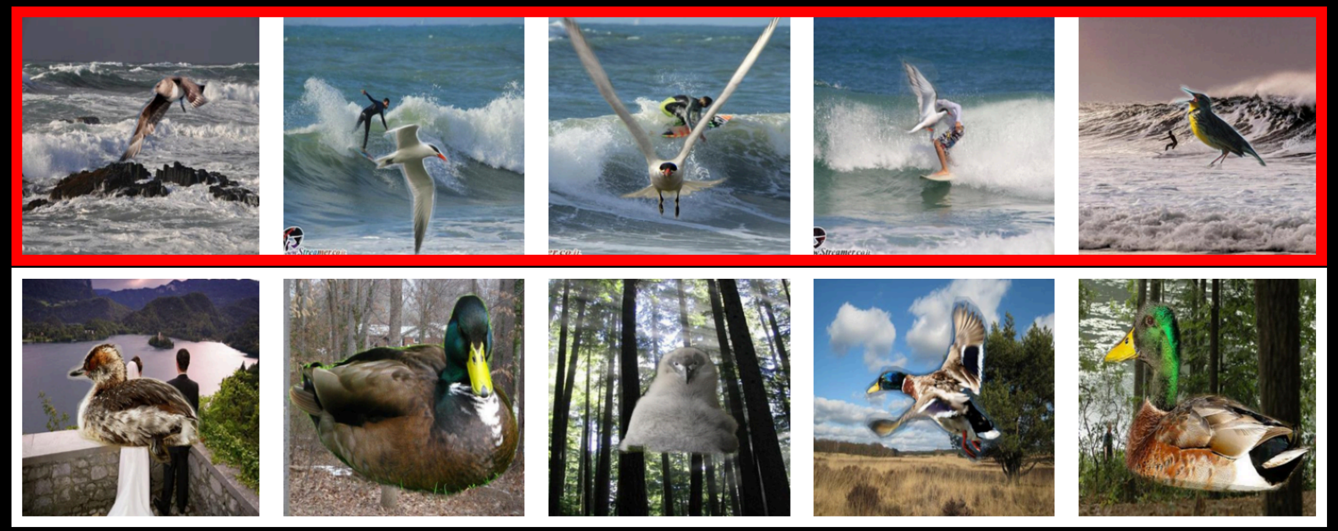
ImageNet - "**bridgeroom**" class
biased to "**east asian culture**"

Results: ConceptScpe discovers real-world dataset bias

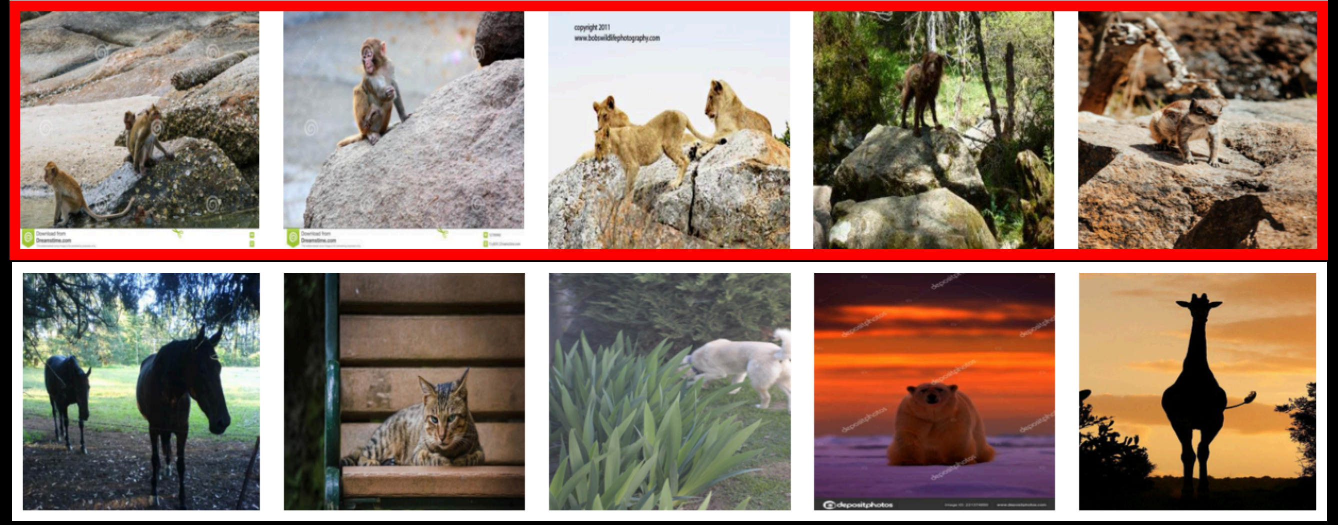
High bias □ Low bias □



CelebA - “**blond hair**”
biased to “**female**”



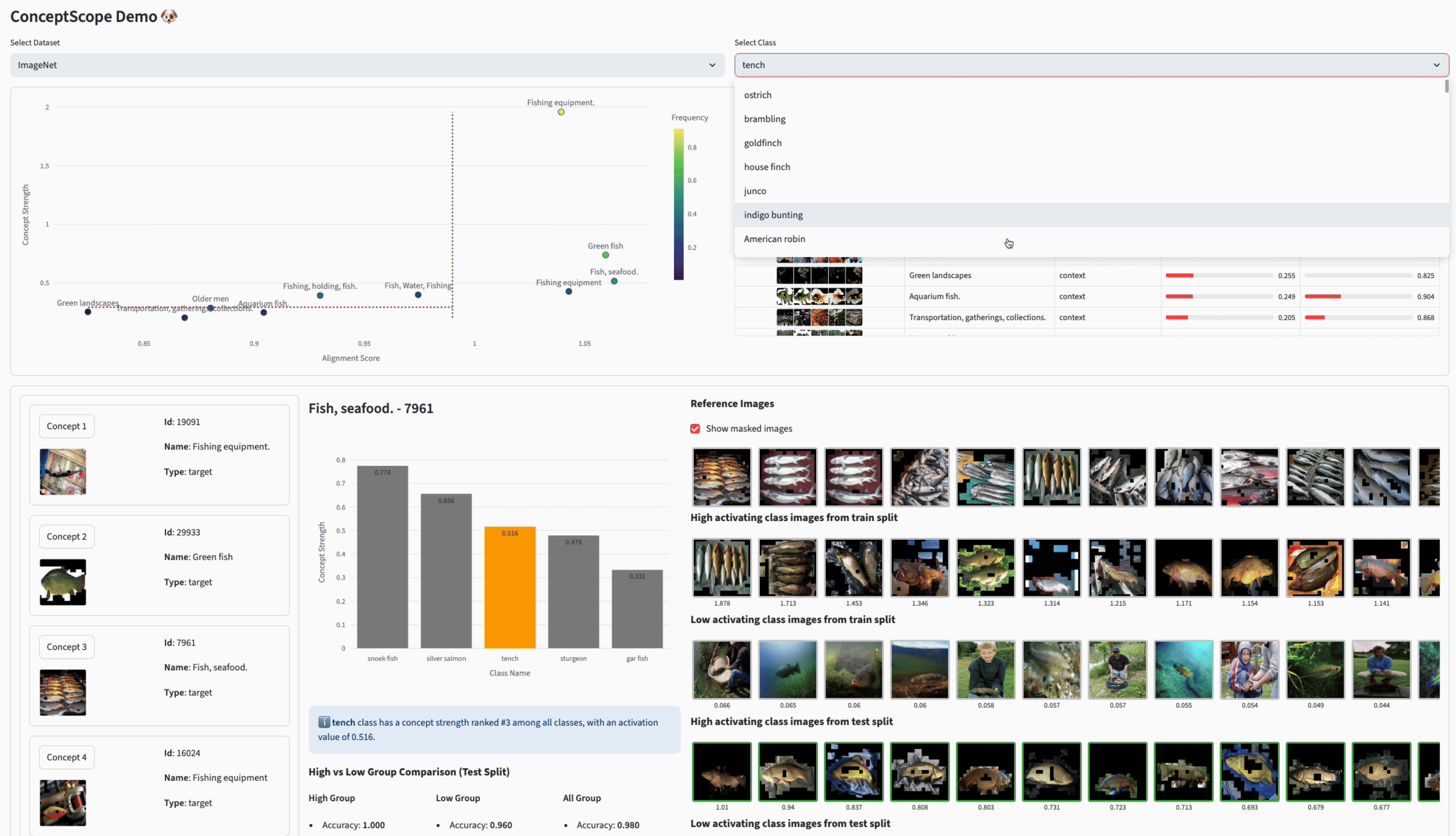
Waterbirds - “waterbirds”
biased to “**ocean**”



Nico++ - “**mammal**” class
biased to “**rock**”

Method	Waterbirds	CelebA	Nico++(75)	Nico++ (90)	Nico++ (95)
DOMINO	90.0%	87.0%	24.0%	24.0%	24.0%
FACTS	100.0 %	100.0%	55.0%	60.8%	61.0%
ViG-Bias	100.0%	100.0%	60.0%	66.7%	65.0%
ConceptScope (Ours)	**100.0%**	**100.0%**	**72.9%**	**73.1%**	**74.0%**

ConceptScpe: Characterizing dataset bias via visual concepts



ConceptScope Demo 🐶

Select Dataset: ImageNet

Select Class: tench

Green landscapes	context	0.255	0.825
Aquarium fish.	context	0.249	0.904
Transportation, gatherings, collections.	context	0.205	0.868

Concept 1
Id: 19091
Name: Fishing equipment.
Type: target

Concept 2
Id: 29933
Name: Green fish
Type: target

Concept 3
Id: 7961
Name: Fish, seafood.
Type: target

Concept 4
Id: 16024
Name: Fishing equipment
Type: target

Fish, seafood. - 7961

ℹ️ **tench** class has a concept strength ranked #3 among all classes, with an activation value of 0.516.

High vs Low Group Comparison (Test Split)

High Group
- Accuracy: **1.000**

Low Group
- Accuracy: **0.960**

All Group
- Accuracy: **0.980**

Reference Images

☑ Show masked images

High activating class images from train split

1.878 1.713 1.453 1.346 1.323 1.314 1.215 1.171 1.154 1.153 1.141

Low activating class images from train split

0.066 0.065 0.06 0.06 0.058 0.057 0.057 0.055 0.054 0.049 0.044

High activating class images from test split

1.01 0.94 0.837 0.808 0.803 0.731 0.723 0.713 0.693 0.679 0.677

Low activating class images from test split

Project page

https://jjho-choi.github.io/ConcepScope-projectpage/

Code & Demo

https://github.com/jjho-choi/ConceptScope