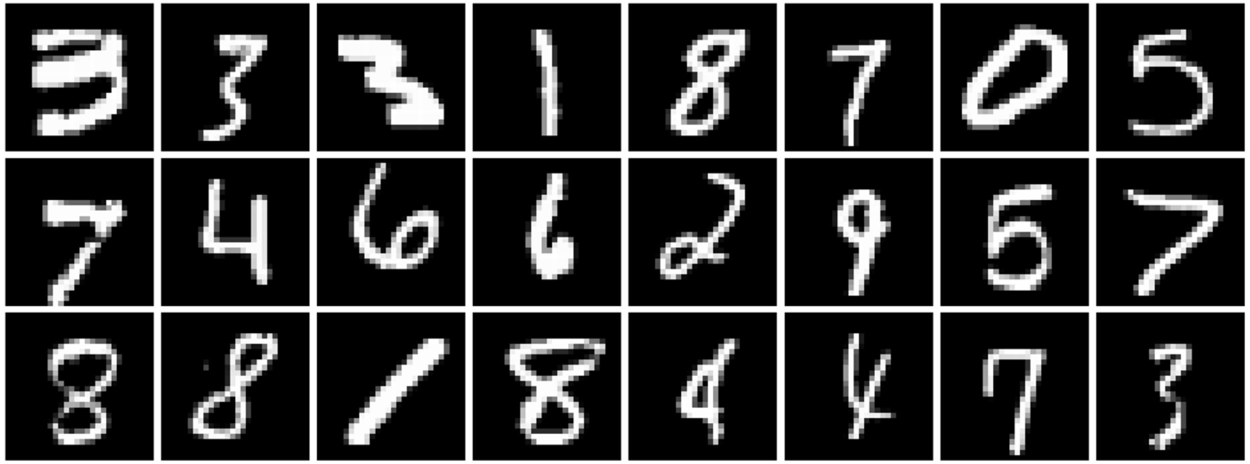
Understanding Bias in Large-Scale Visual Datasets

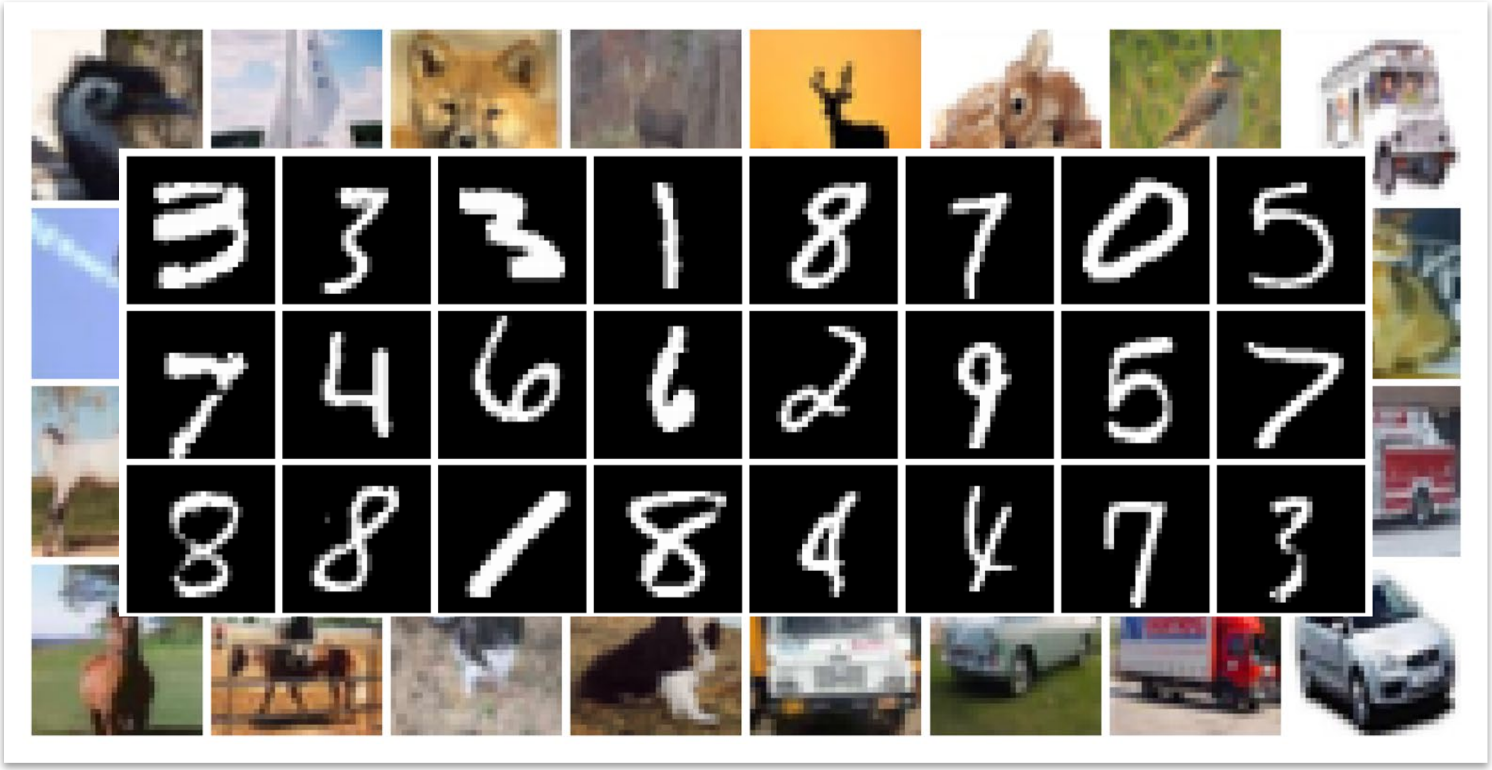
Boya Zeng*, Yida Yin*, Zhuang Liu
University of Pennsylvania, UC Berkeley, Meta FAIR
(*equal contribution)

Scaling Data



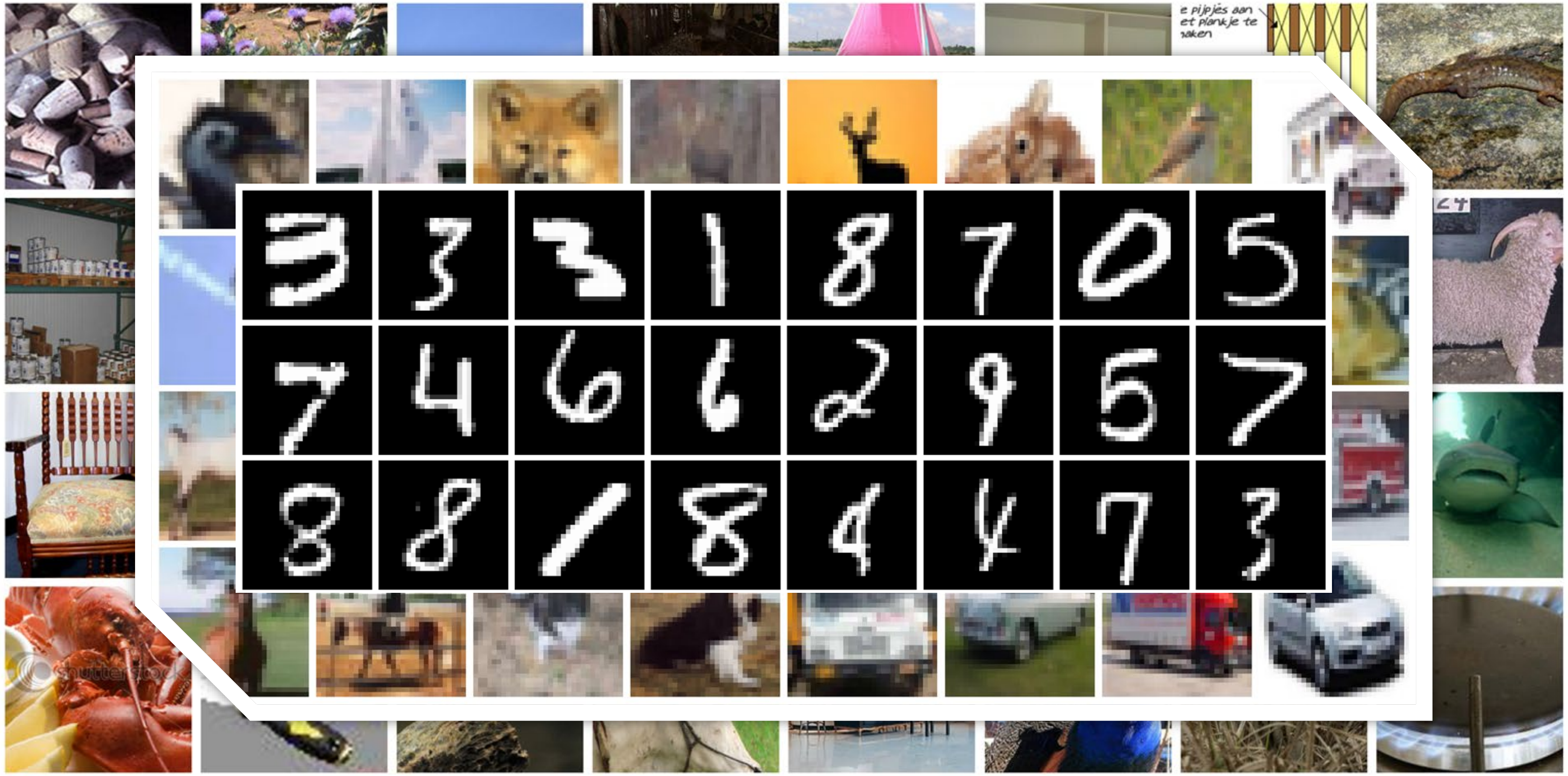
MNIST

Scaling Data



CIFAR-10

Scaling Data



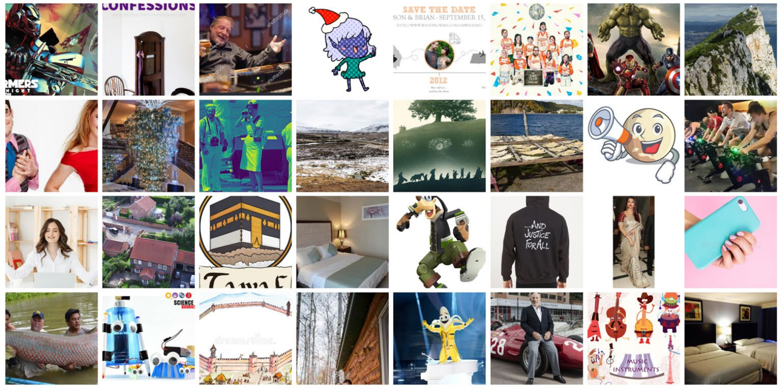
ImageNet

Modern Large Visual Datasets

YFCC, 100M

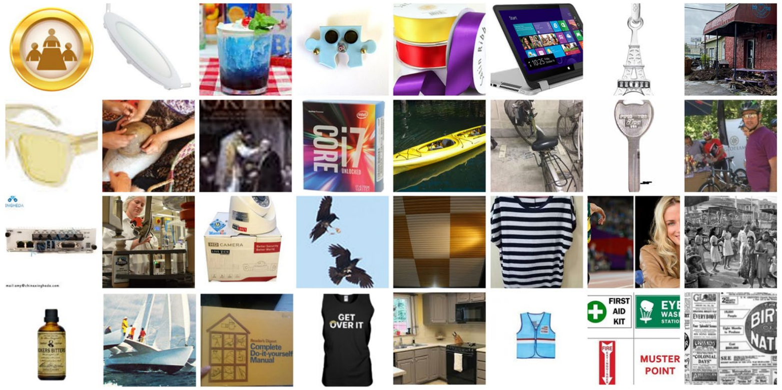
from Flickr

CC,12M



filtered web images and text

DataComp, 1B



aligned image-text pairs from Common Crawl

Thomee et al., ’16; Changpinyo et al., ’21; Gadre et al., ’23

Datasets are general enough?

YFCC = CC = DataComp =  ?

Dataset Classification, Revisited

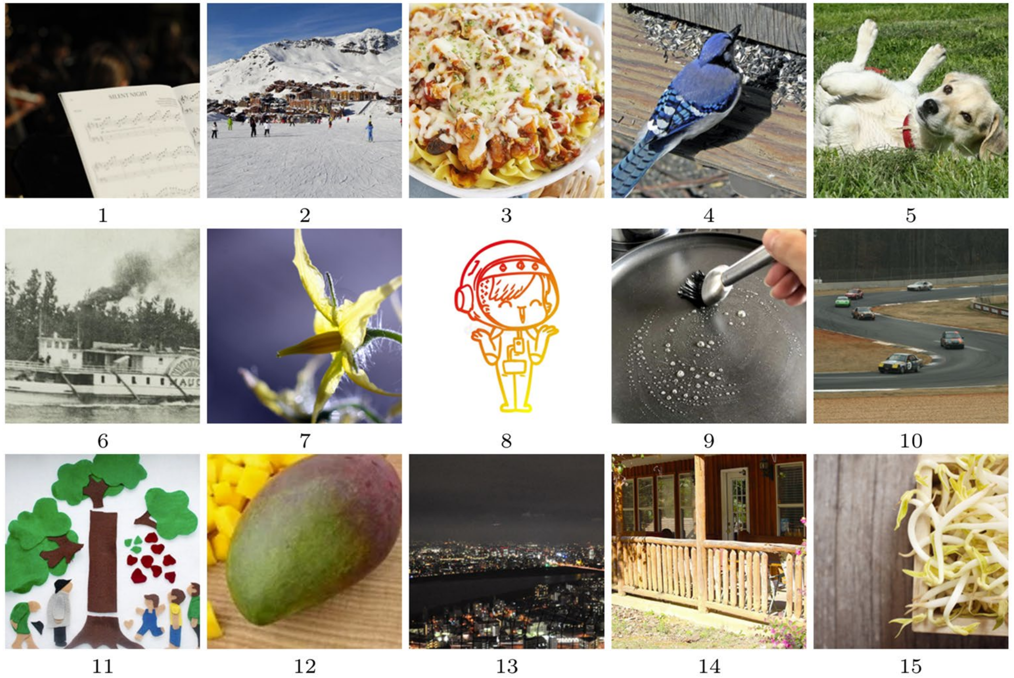
2011

Caltech-101, COIL-100, ..., or LabelMe?

39% for SVM

Larger
More diverse
More representative

2024



YFCC, CC, or DataComp?

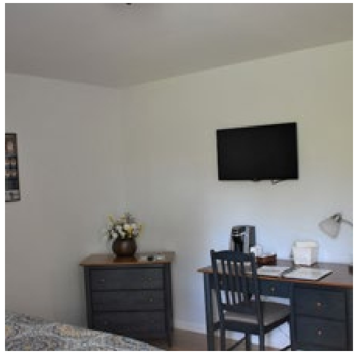
82% for modern neural network

Liu & He, A Decade's Battle on Dataset Bias: Are We There Yet? '24

What are the concrete forms of bias?

YFCC CC DataComp



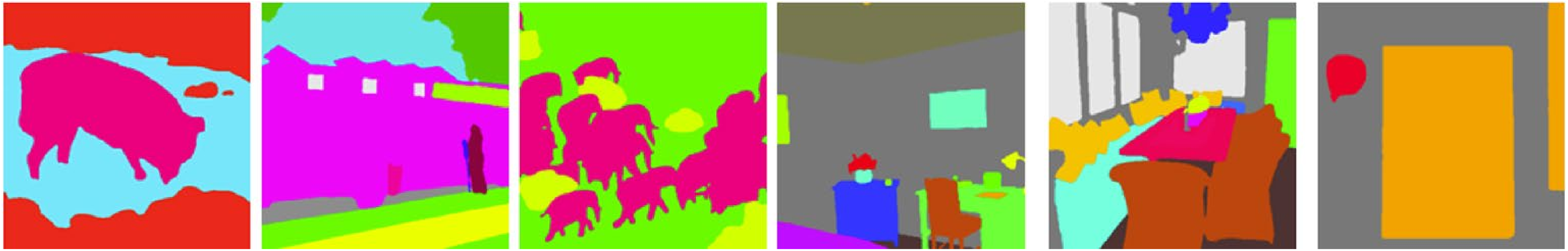



82% Biased!

Semantics? Spatial? Structure? Color? Frequency? …

Semantics: Semantic Segmentation

YFCC CC DataComp

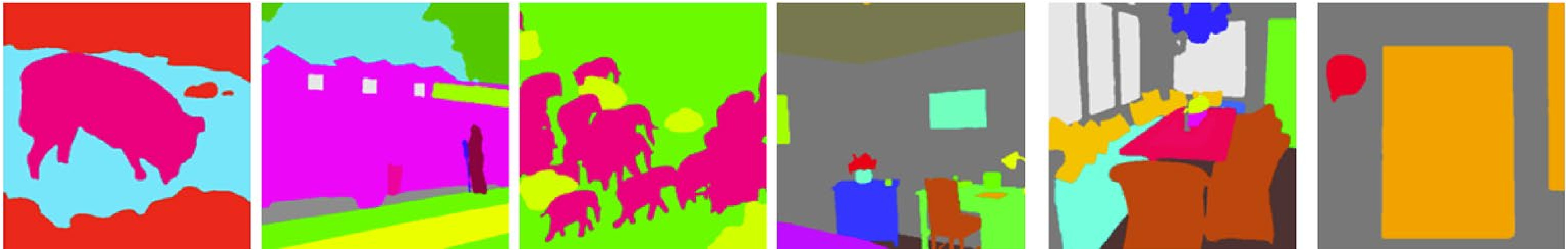


fine-grained semantic annotation

Guess?

Semantics: Semantic Segmentation

YFCC CC DataComp



fine-grained semantic annotation

67.6%

Semantics: Image Captioning

YFCC		CC		DataComp	
A black and white dog is standing in a stream of water.	A food truck is parked on the side of the road.	A herd of elephants walking through a grassy field.	A room with a bed, a desk, a chair, a TV, and a vase with flowers.	A large wooden dining table with wicker chairs and a chandelier above it.	A notebook with polka dots and a pink and blue book on a table.

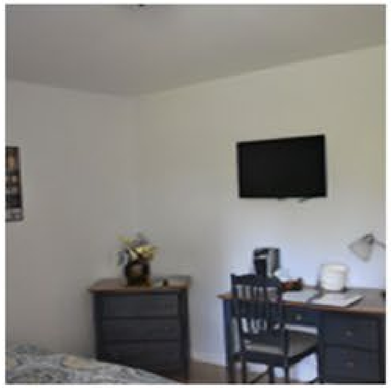
semantic representations with no visual information

63.8% (short) / 66.1% (long)

Semantics: Variational Autoencoder

YFCC


CC

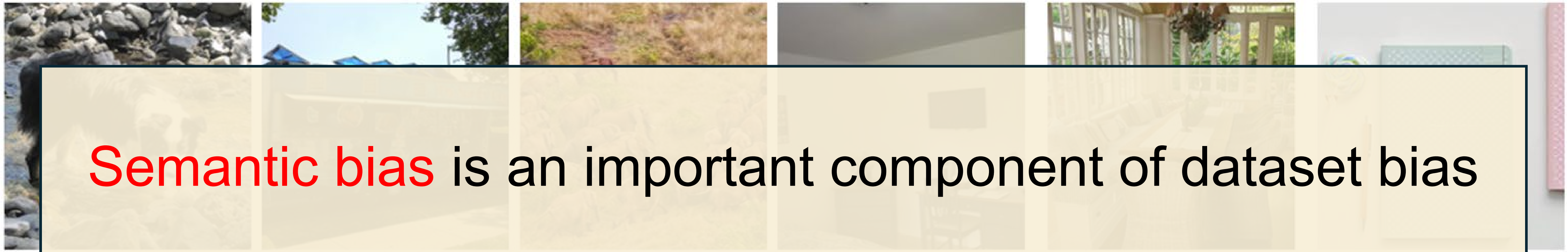
DataComp


may encode semantic information and suppress low-level signatures

77.4%

Semantics: Variational Autoencoder

YFCC CC DataComp



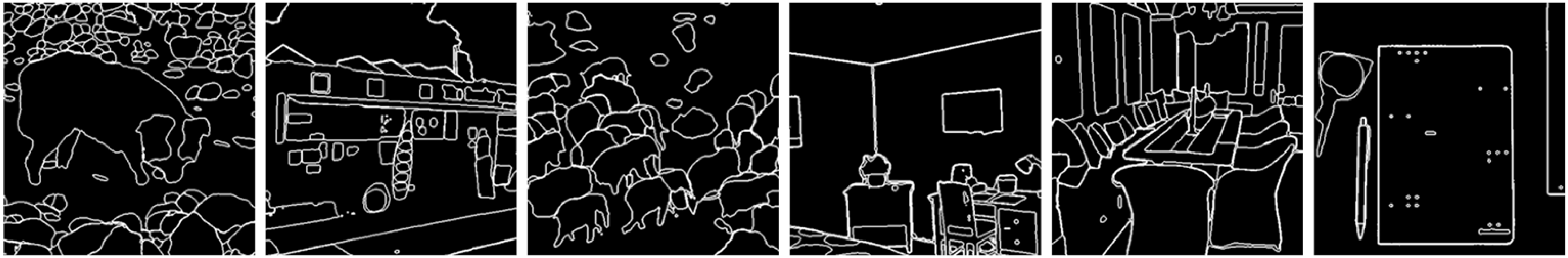
Semantic bias is an important component of dataset bias

may encode semantic information and suppress low-level signatures

77.4%

Structures: Segment Anything Model

YFCC CC DataComp

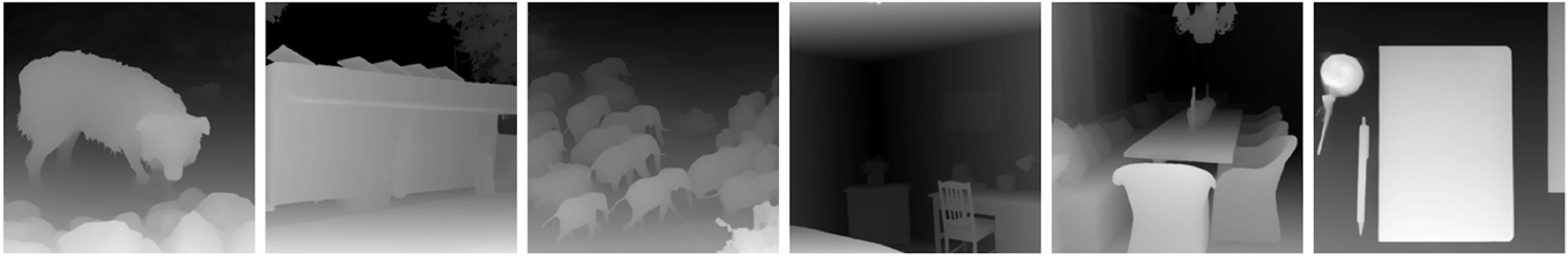


high-quality class-agnostic object segmentation masks

73.2%

Structures: Depth

YFCC CC DataComp

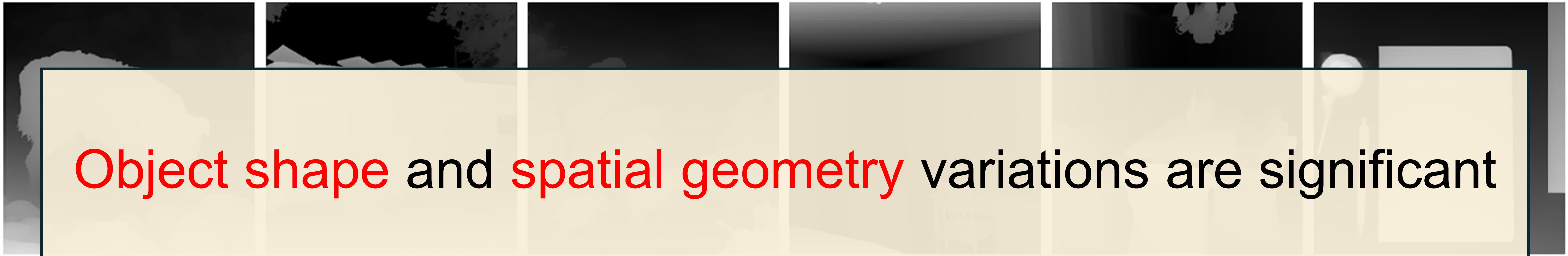


fine-grained spatial context and relative object positioning

73.1%

Structures: Depth

YFCC CC DataComp



Object shape and spatial geometry variations are significant

fine-grained spatial context and relative object positioning

73.1%

Spatial Permutations: Pixel Shuffling

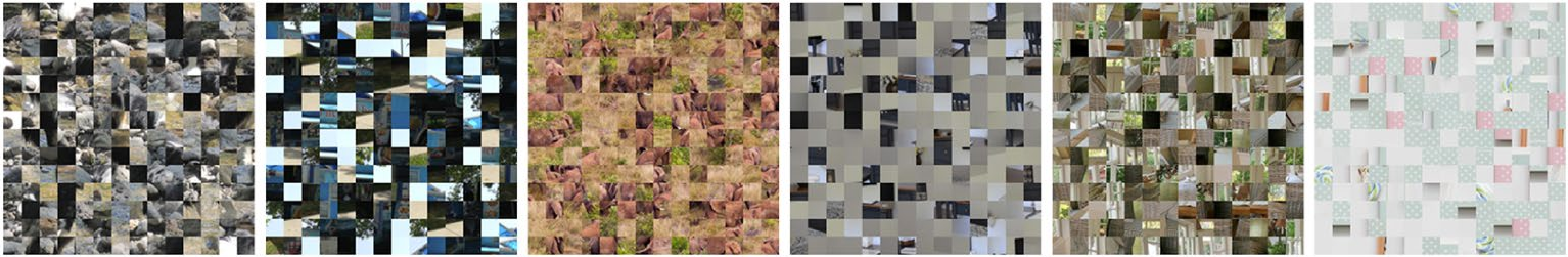
YFCC CC DataComp

color distribution of pixels

52.2% (random) / 58.5% (fixed)

Spatial Permutations: Patch Shuffling

YFCC CC DataComp

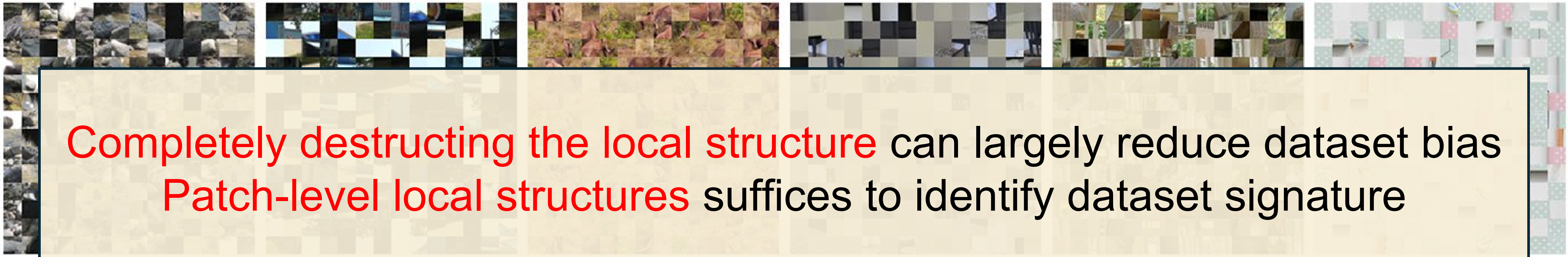


preserves more local spatial information

80.1% (random) / 81.2% (fixed)

Spatial Permutations: Patch Shuffling

YFCC CC DataComp



Completely destructing the local structure can largely reduce dataset bias
Patch-level local structures suffices to identify dataset signature

preserves more local spatial information

80.1% (random) / 81.2% (fixed)

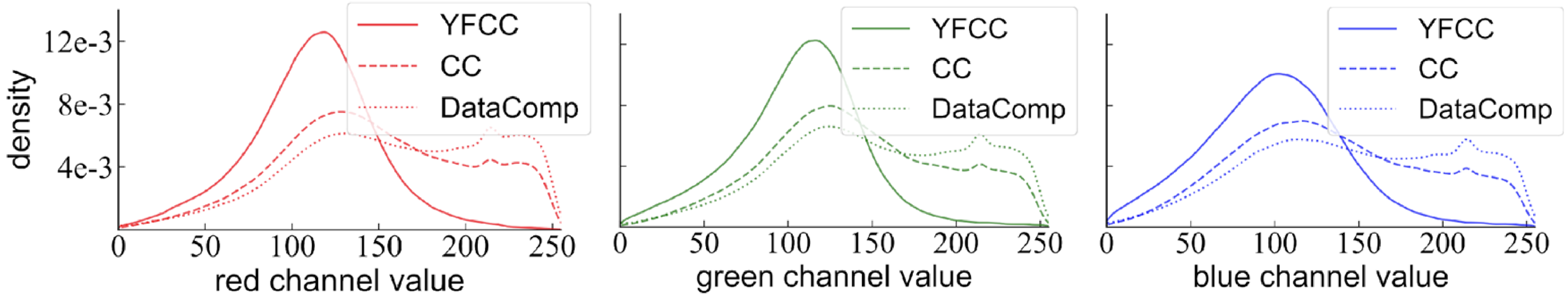
Mean RGB

YFCC CC DataComp

abstracts the pixel details into a constant RGB color map

48.5%

Mean RGB



YFCC is much darker than CC and DataComp

Synthetic Image: Unconditional Generation

YFCC

CC

DataComp

trains an unconditional diffusion model on each dataset

77.6%

Synthetic Image: Unconditional Generation

YFCC CC DataComp

Synthetic images can inherit the bias in diffusion model training images

trains an unconditional diffusion model on each dataset

77.6%

Synthetic Image: Text-to-Image

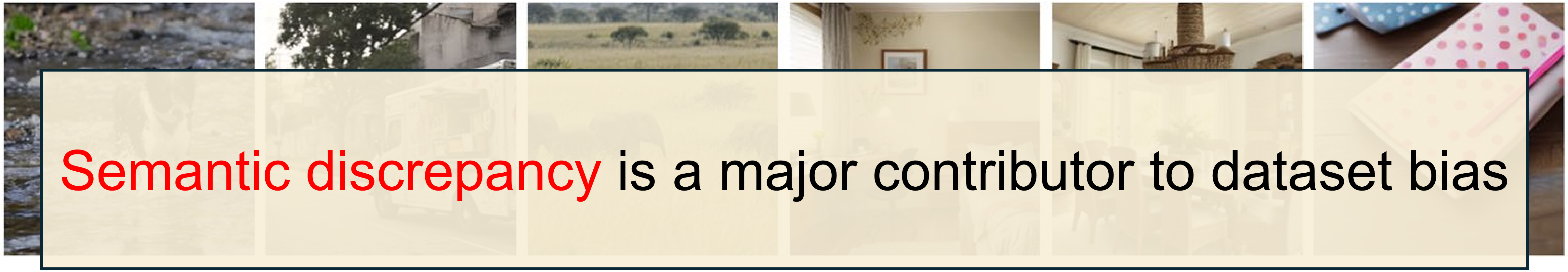
YFCC CC DataComp

potentially preserves the semantic bias in the original images

58.1%

Synthetic Image: Text-to-Image

YFCC CC DataComp



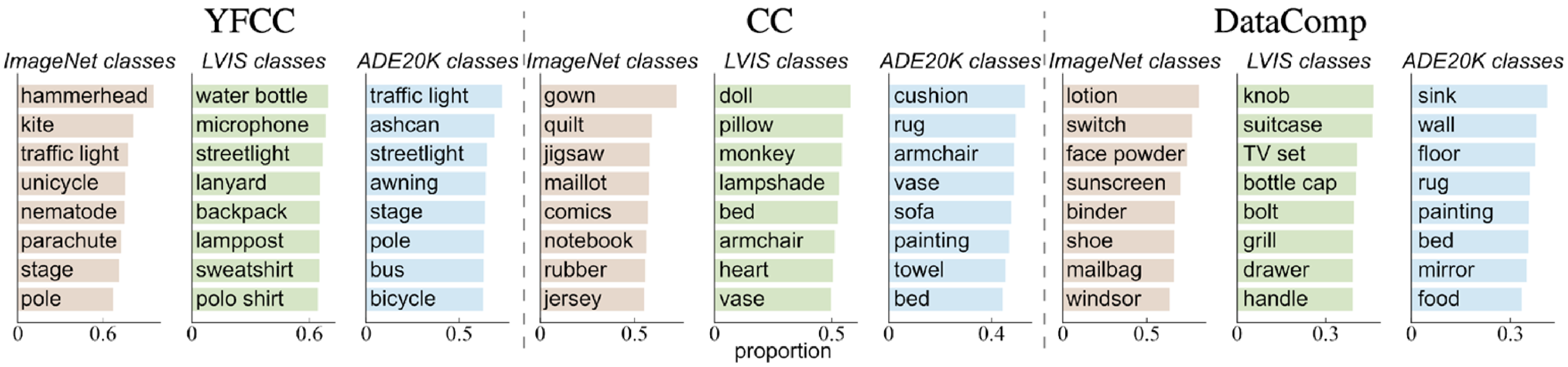
potentially preserves the semantic bias in the original images

58.1%

Explaining Semantic Bias: Objects

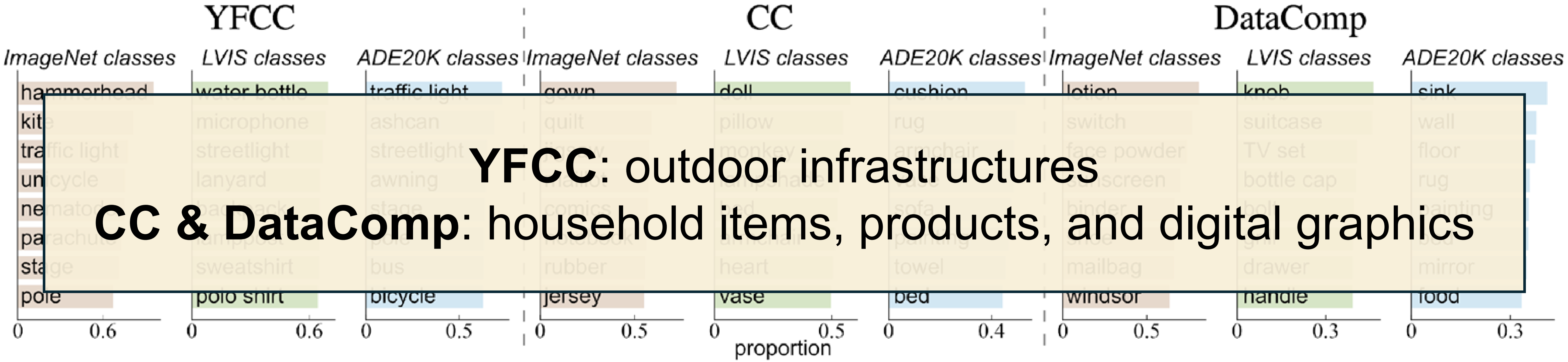
ImageNet	Dining Table
LVIS	Armchair, Pillow, Cushion…
ADE20K	Chandelier, Armchair, Floor...

Dataset Proportions within Object Class



considerable imbalance in object-level distribution across datasets

Dataset Proportions within Object Class



YFCC: outdoor infrastructures

CC & DataComp: household items, products, and digital graphics

can identify objects that are more balanced across datasets

Explaining Semantic Bias: Language

A large wooden dining table with wicker chairs and a chandelier above it.

explicit semantic representations

Unsupervised Topic Discovery

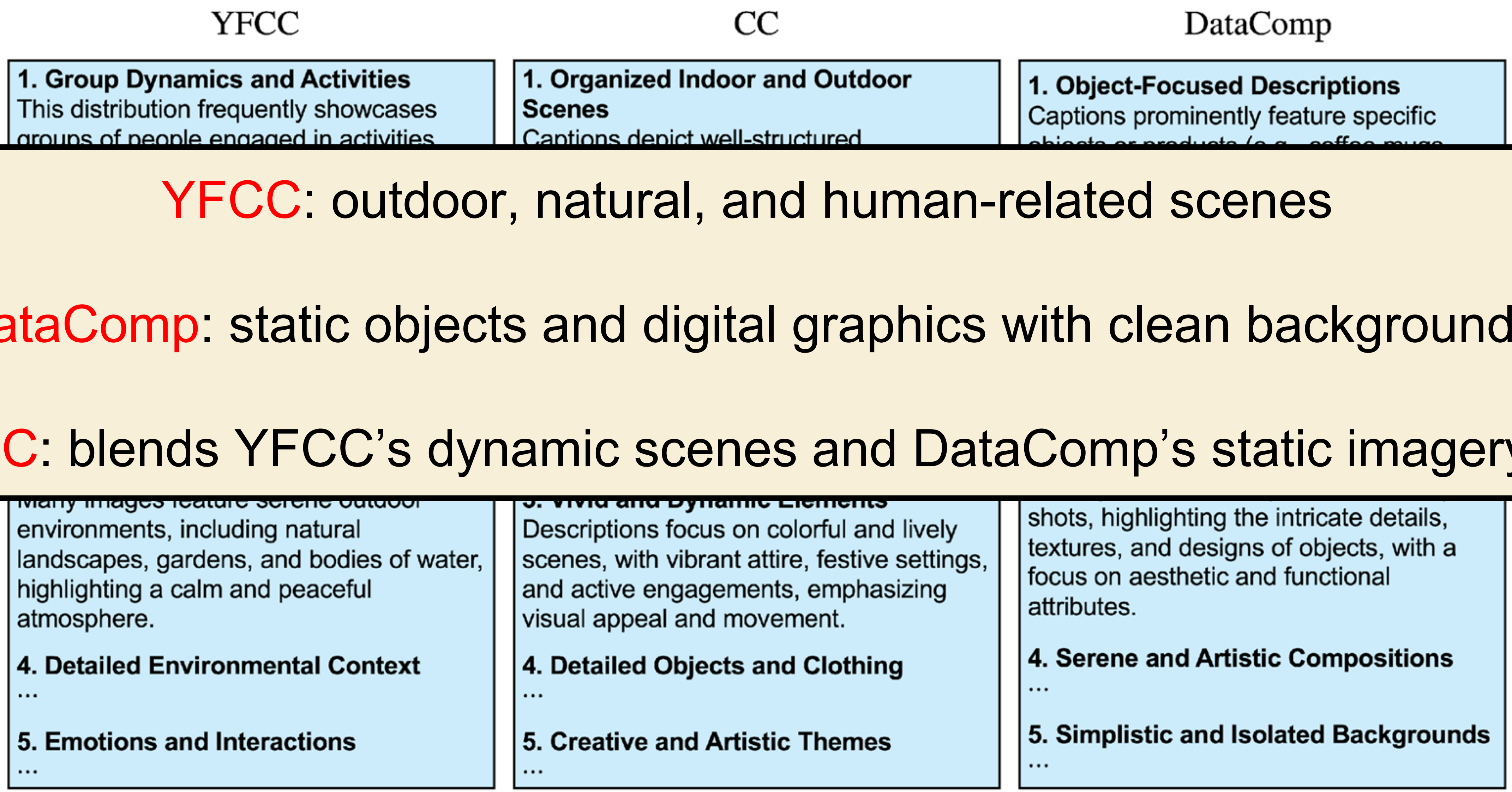
YFCC	CC	DataComp
[building, scene, street, car, sign]	[room, table, chairs, chair, dining]	[logo, background, design, book, colors]
[scene, field, game, person, dog]	[design, background, colors, logo, display]	[scene, table, room, building, atmosphere]
[table, room, dining, scene, items]	[woman, man, shirt, scene, dress]	[car, scene, truck, background, kitchen]
[people, man, woman, scene, group]	[scene, people, water, group, atmosphere]	[background, table, design, box, bottle]
[scene, water, sky, trees, tree]	[scene, building, car, person, dog]	[man, woman, scene, shirt, person]

YFCC: outdoor scenes
CC & DataComp: digital graphics

LLM Summarization

YFCC	CC	DataComp
1. Group Dynamics and Activities This distribution frequently showcases groups of people engaged in activities such as playing music, attending events, or participating in sports, emphasizing social interactions and communal settings. **2. Urban and Social Settings** Captions often describe dynamic environments filled with people and activity in urban or public settings, such as busy city streets, transportation hubs, and social events. **3. Serene Natural Settings** Many images feature serene outdoor environments, including natural landscapes, gardens, and bodies of water, highlighting a calm and peaceful atmosphere. **4. Detailed Environmental Context** ... **5. Emotions and Interactions** ...	**1. Organized Indoor and Outdoor Scenes** Captions depict well-structured environments, including cozy bedrooms, dining areas, cityscapes, and architectural landmarks, emphasizing the arrangement and detail. **2. Human Interactions and Social Events** Emphasis on social and formal gatherings like weddings, concerts, and ceremonies, highlighting attire, decor, and the lively atmosphere. **3. Vivid and Dynamic Elements** Descriptions focus on colorful and lively scenes, with vibrant attire, festive settings, and active engagements, emphasizing visual appeal and movement. **4. Detailed Objects and Clothing** ... **5. Creative and Artistic Themes** ...	**1. Object-Focused Descriptions** Captions prominently feature specific objects or products (e.g., coffee mugs, toys, cars), often isolated against minimalistic backgrounds to highlight their characteristics. **2. Vibrant and Playful Visuals** Scenes frequently include vibrant, colorful, and playful elements, focusing on visually appealing and lively imagery that captures attention. **3. Close-Up and Detailed Views** Descriptions often emphasize close-up shots, highlighting the intricate details, textures, and designs of objects, with a focus on aesthetic and functional attributes. **4. Serene and Artistic Compositions** ... **5. Simplistic and Isolated Backgrounds** ...

LLM Summarization



YFCC: outdoor, natural, and human-related scenes

DataComp: static objects and digital graphics with clean backgrounds

CC: blends YFCC’s dynamic scenes and DataComp’s static imagery

Summary

Framework to study the bias in large-scale datasets

Use transformation to quantify bias in each type of information

Decompose semantic bias with object and language analysis

Showcase on YFCC, CC, and DataComp

Paper & Code
boyazeng.github.io/understand_bias