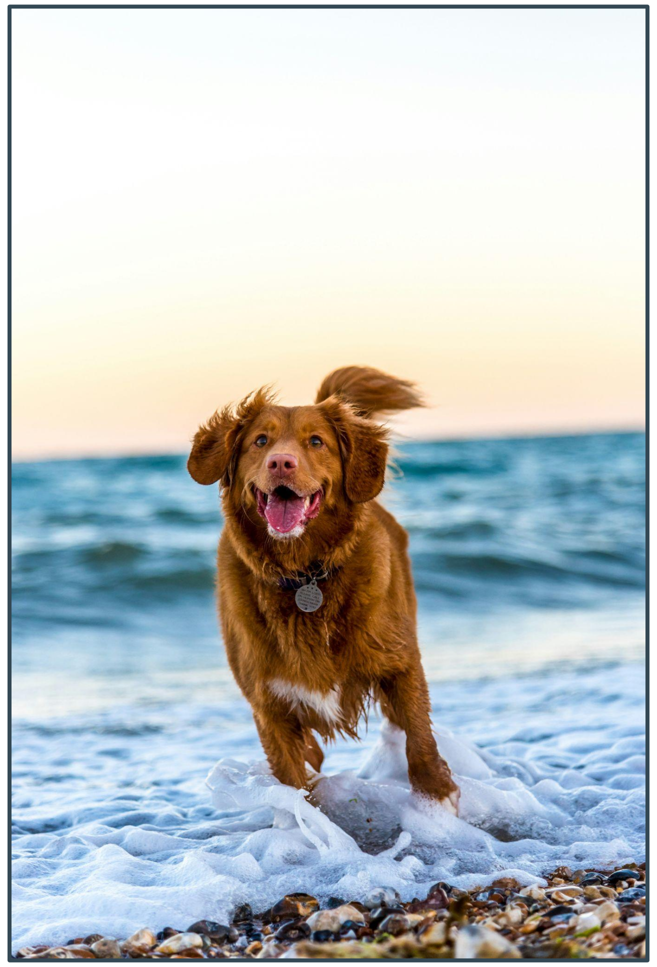
UniBench: Visual Reasoning Requires Rethinking Vision-Language Beyond Scaling

Paper: [link]
Codebase: https://github.com/facebookresearch/unibench/

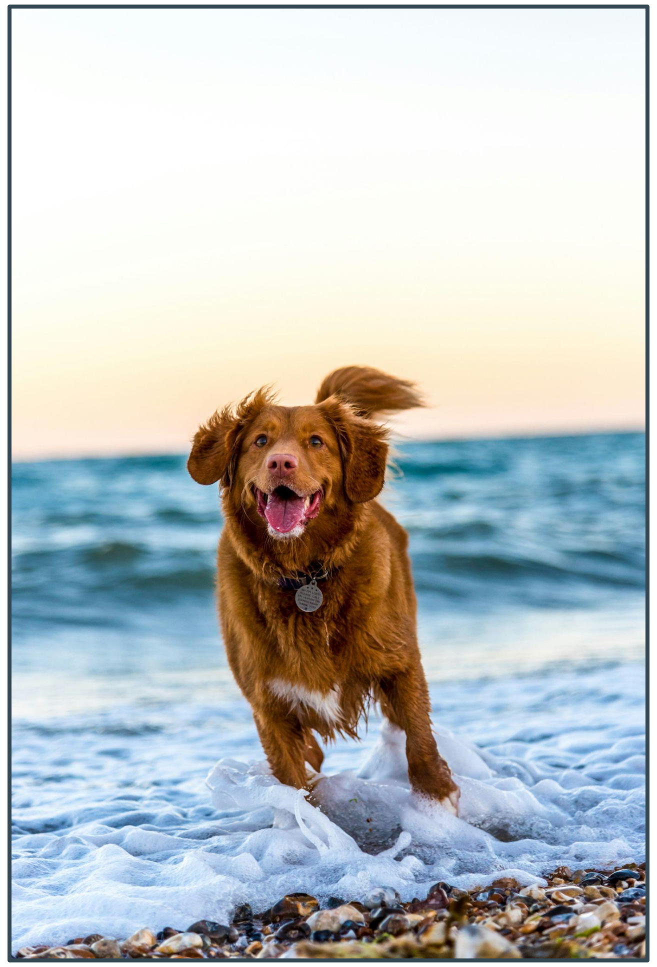
01 Introduction

Traditional Supervised Vision Model



(Source: unsplash)

Traditional Supervised Vision Model

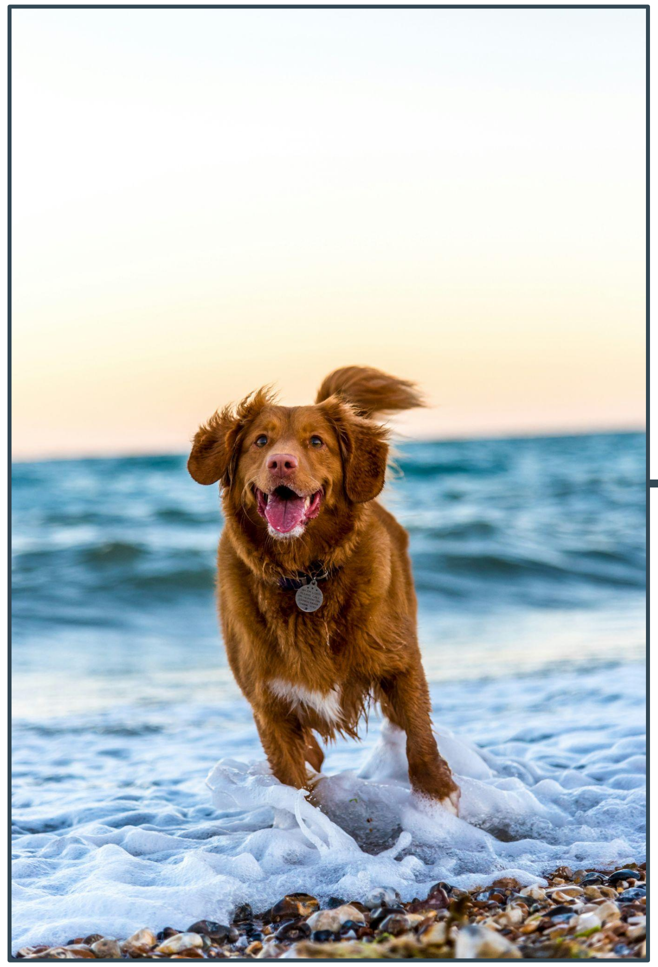


(Source: unsplash)

<u>Predefined Classes</u>

Dog
Cat
Bird
Car

Traditional Supervised Vision Model



(Source: unsplash)

Supervised

<u>Predefined Classes</u>

Dog
Cat
Bird
Car

Traditional Supervised Vision Model


(Source: unsplash)

Supervised

<u>Predefined Classes</u>

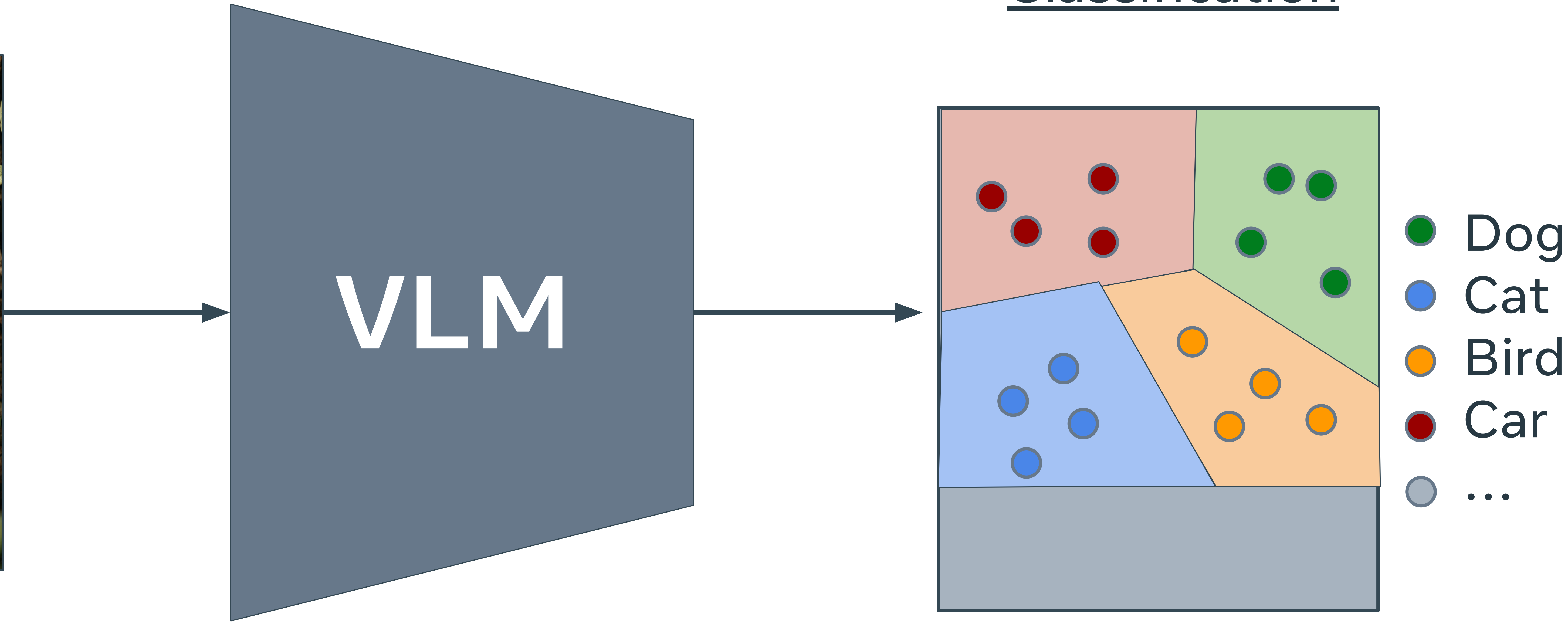
Dog
Cat
Bird
Car

Monkey

Properties of Vision-Language Models

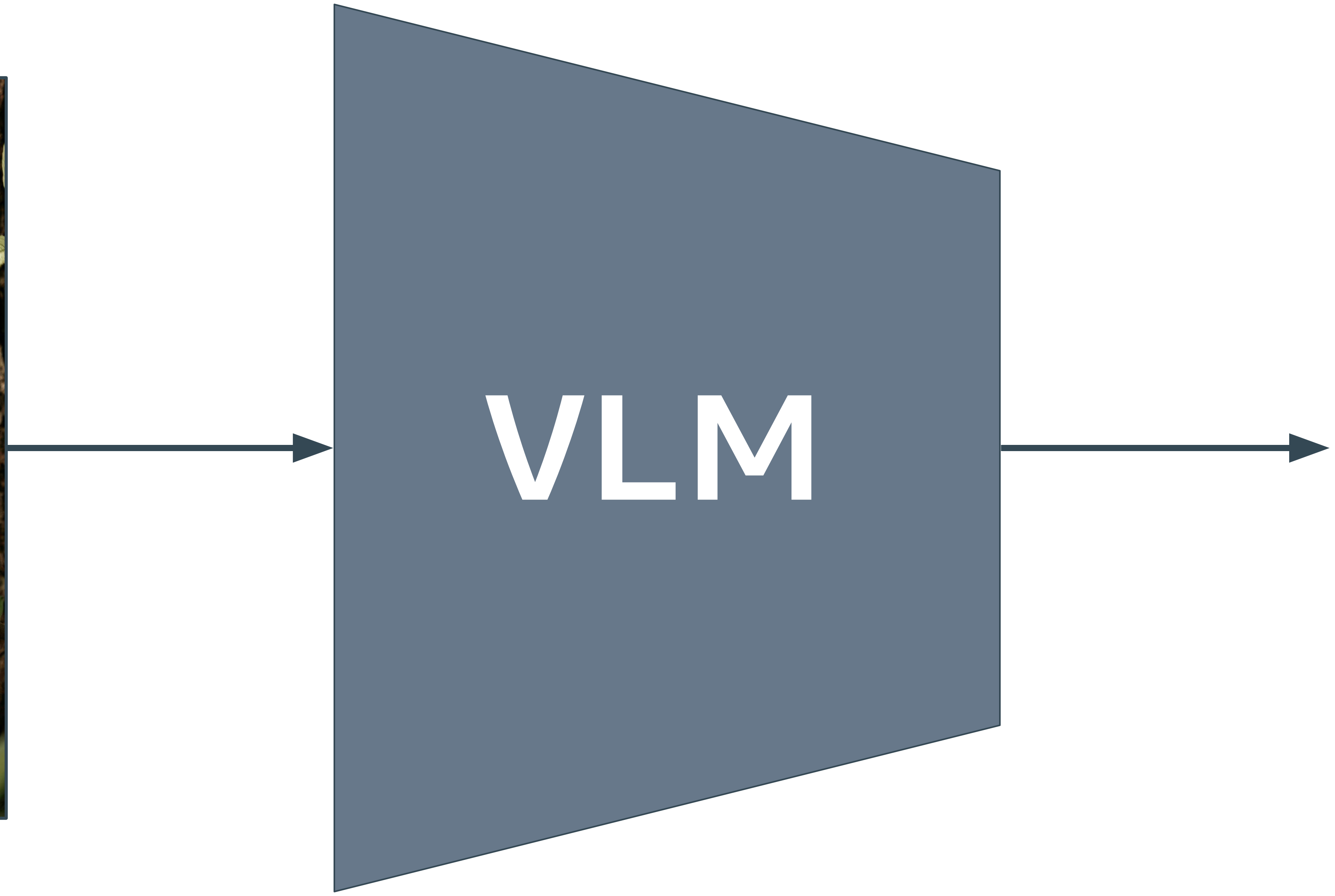
(Source: unsplash)

Zero-shot Image Classification



Properties of Vision-Language Models

(Source: unsplash)



Zero-shot Image Classification

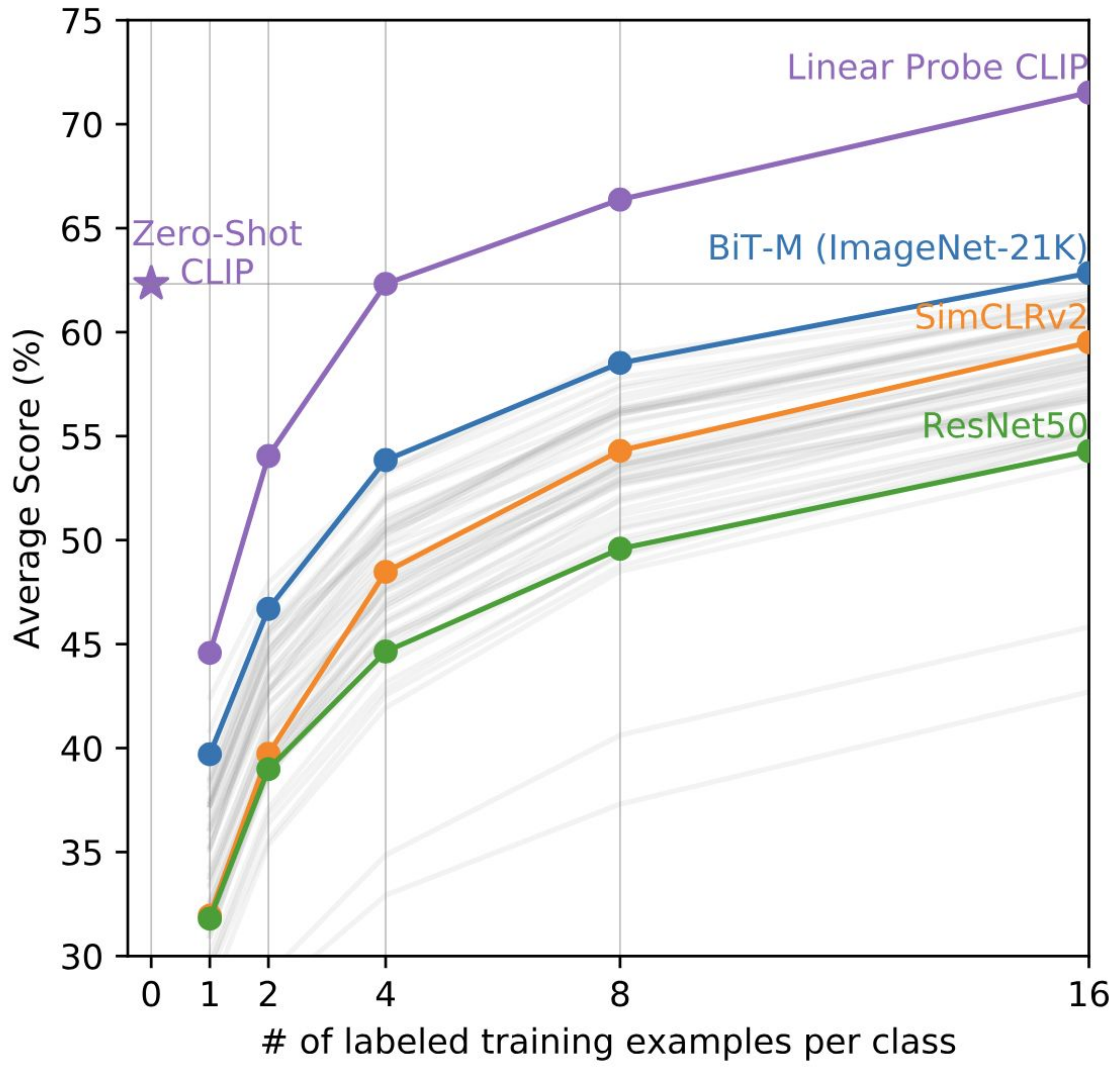
Is the Monkey in the

1) Foreground

or

2) Background

Properties of Vision-Language Models



(Radford, Alec et al., 2021)

Models & Benchmarks

Models & Benchmarks

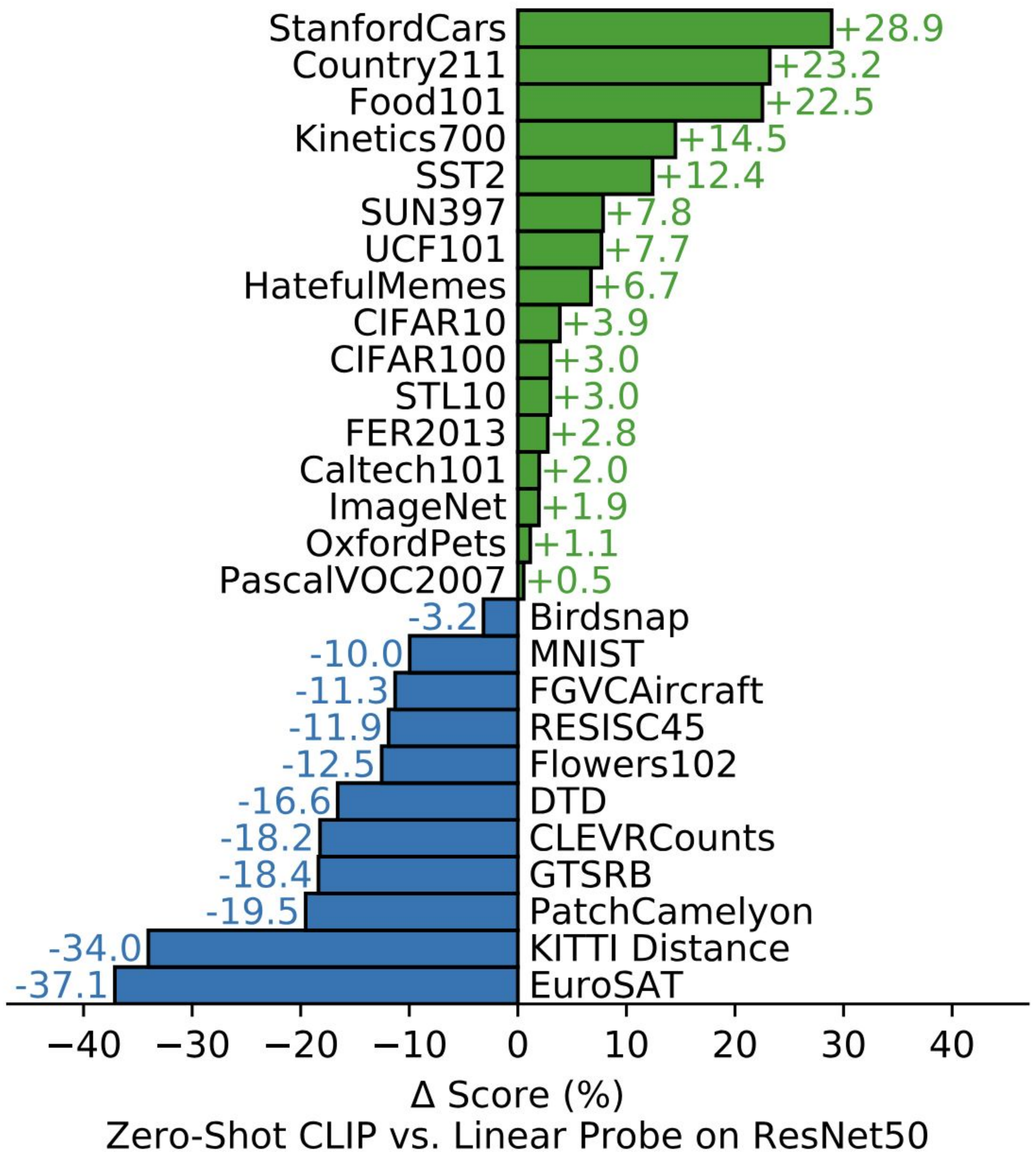
- CLIP (Radford, Alec et al., 2021)
- LiT (Zhai, Xiaohua et al., 2021)
- FLAVA (Singh, Amanpreet et al., 2021)
- CoCa (Yu, Jiahui et al., 2022)
- BLIP (Li, Junnan et al., 2022)
- EVA (Fang, Yuxin et al., 2023)
- CLIPA (Li, Xianhang et al., 2023)
- SigLIP (Zhai, Xiaohua et al., 2023)
- MetaCLIP (Xu, Hu et al., 2023)
- ...

80 VLMs

Models & Benchmarks

- CLIP (Radford, Alec et al., 2021)
- LiT (Zhai, Xiaohua et al., 2021)
- FLAVA (Singh, Amanpreet et al., 2021)
- CoCa (Yu, Jiahui et al., 2022)
- BLIP (Li, Junnan et al., 2022)
- EVA (Fang, Yuxin et al., 2023)
- CLIPA (Li, Xianhang et al., 2023)
- SigLIP (Zhai, Xiaohua et al., 2023)
- MetaCLIP (Xu, Hu et al., 2023)
- ...

80 VLMs



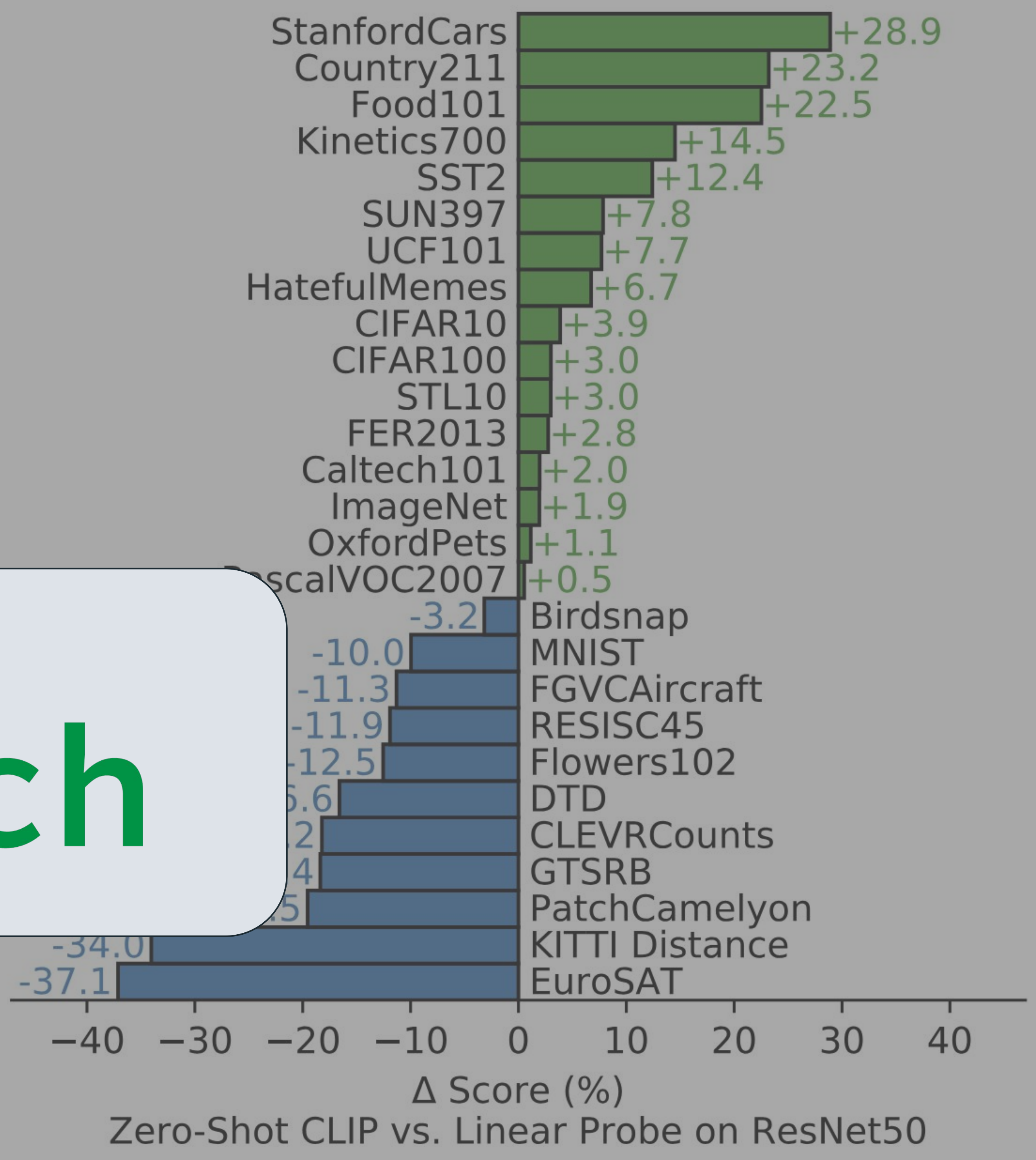
50+ Benchmarks

Models & Benchmarks

- CLIP (Radford, Alec et al., 2021)
- LiT (Zhai, Xiaohua et al., 2021)
- FLAVA (Singh, Amanpreet et al., 2021)
- CoCa (Yu, Jiahui et al., 2022)
- BLIP (Li, Junnan et al., 2022)
- EVA (Fang, Yuxin et al.,
- CLIPA (Li, Xianhang et
- SigLIP (Zhai, Xiaohua e
- MetaCLIP (Xu, Hu et al.
- …

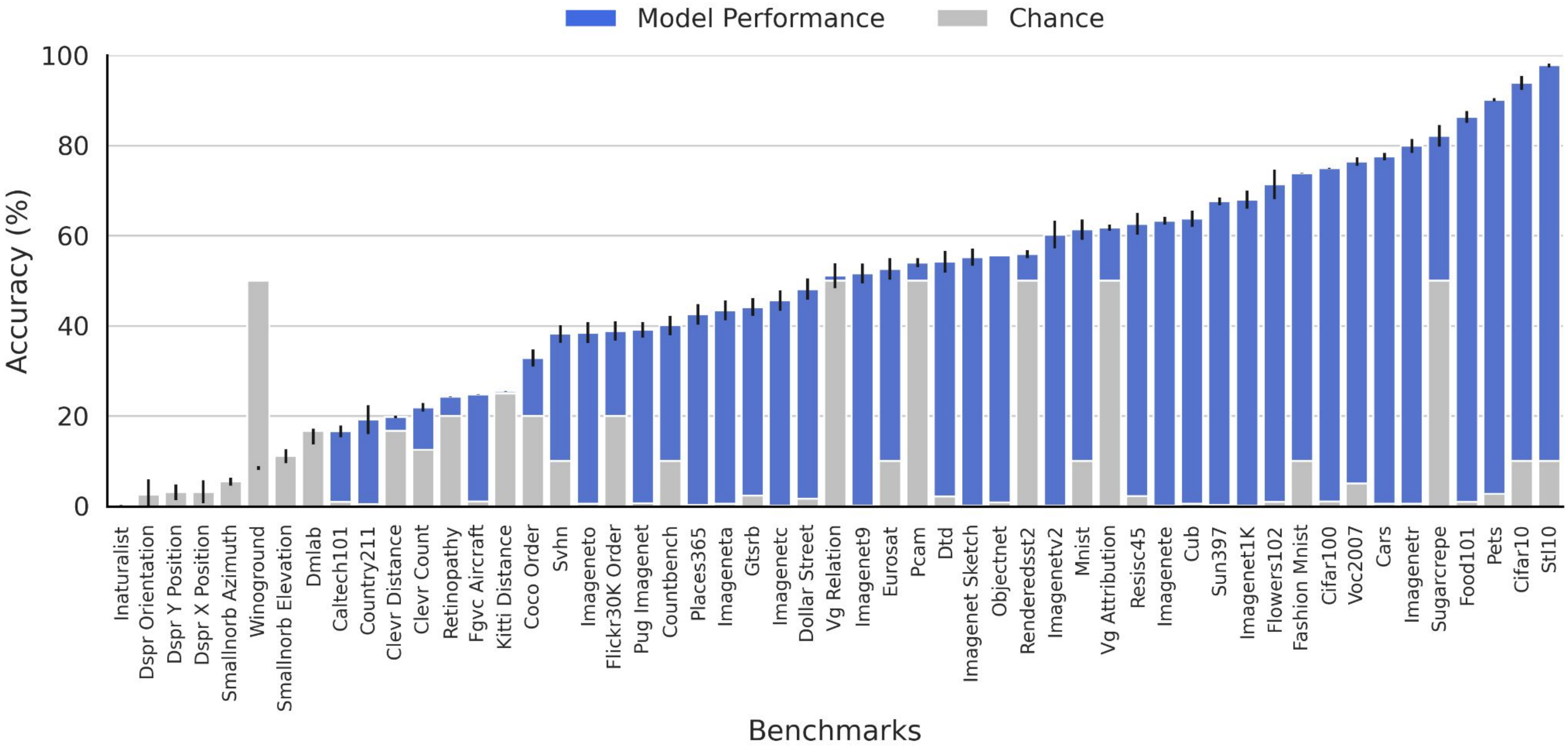
UniBench

80 VLMs



50+ Benchmarks

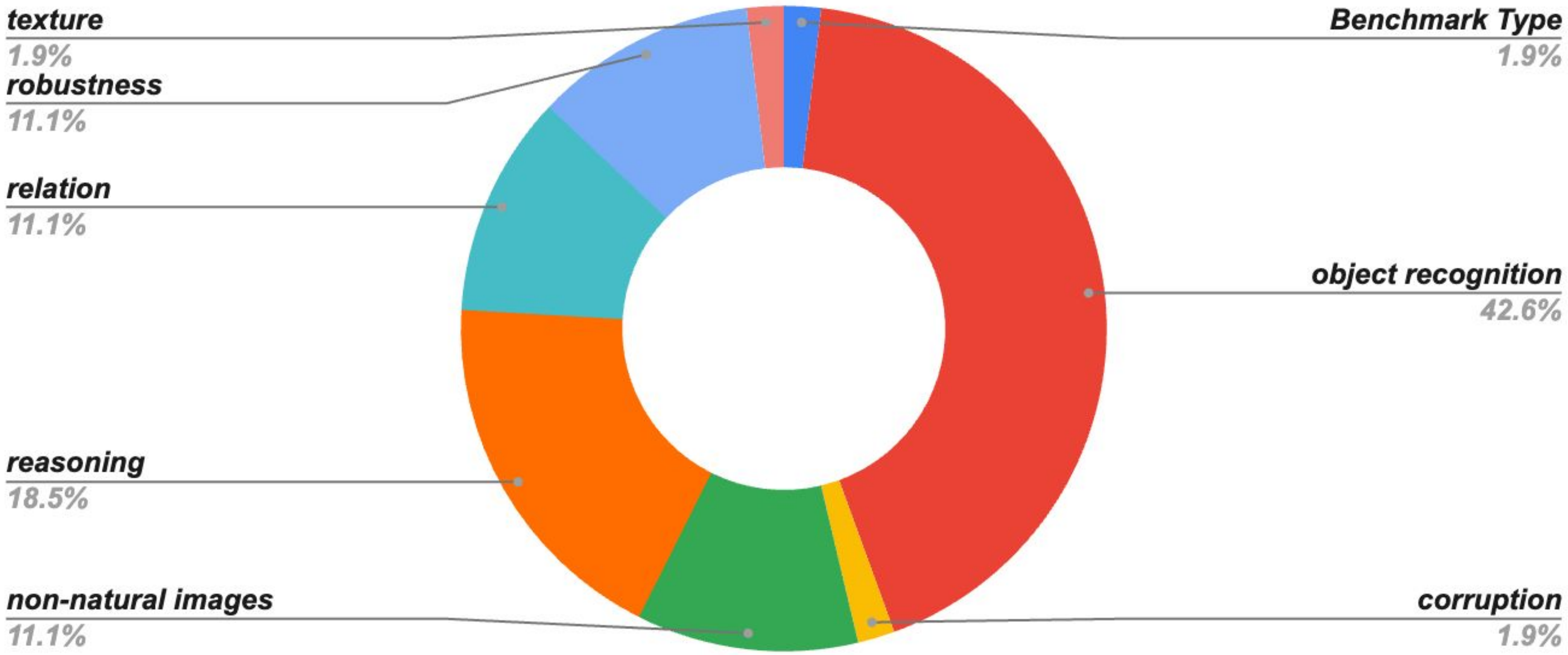
VLMs have strengths and weaknesses



Benchmark Types

Benchmark Type	Blur/Noise	Robustness	Object Recognition	Non-Natural Images	Texture	Relation	Reasoning
Image							
Target	Black Swan	Dog	Cat	Bird	Zig Zagged	White cabinet is above black cabinet	two red pingpong rackets

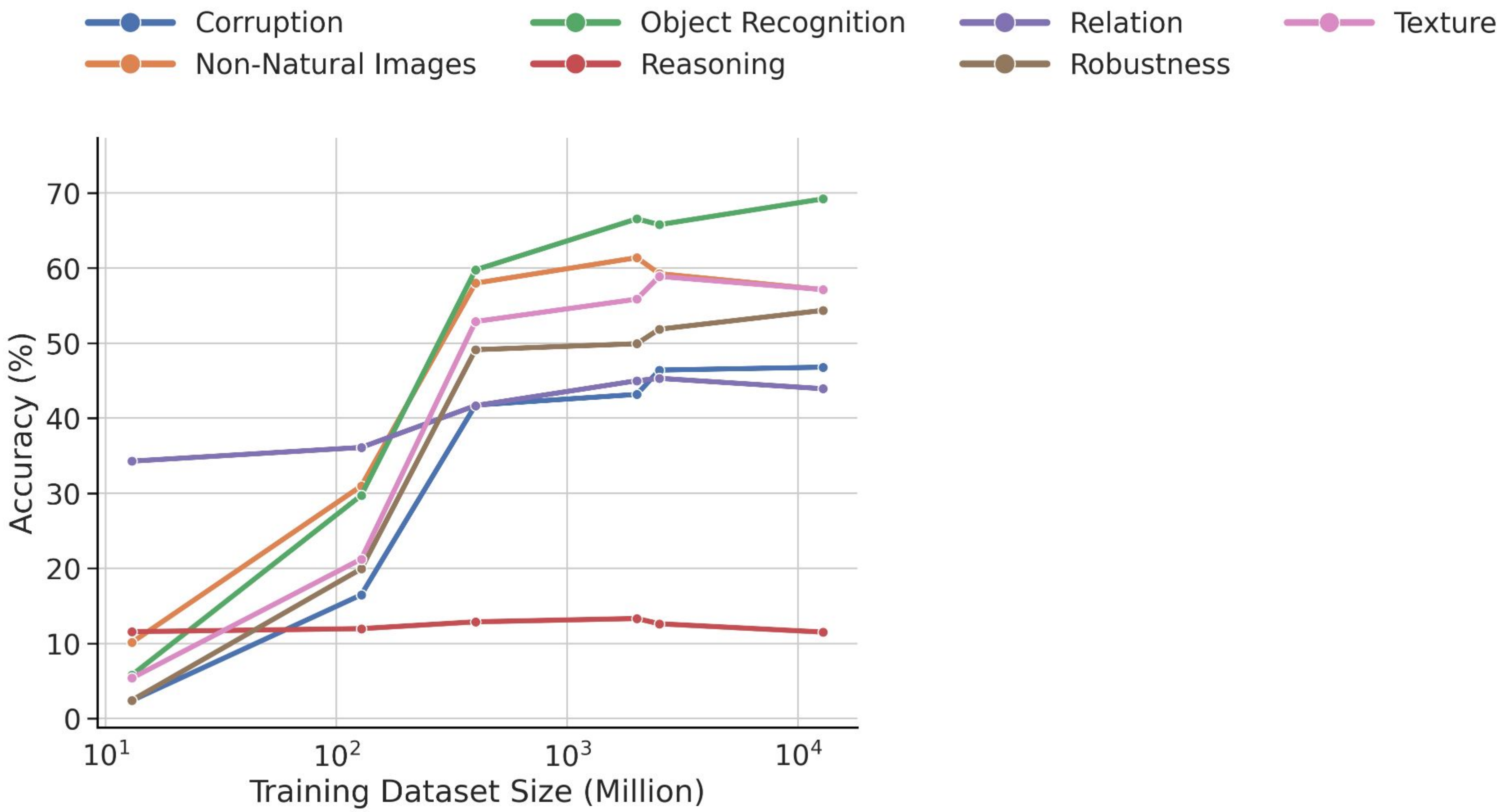
Benchmark Type Distribution



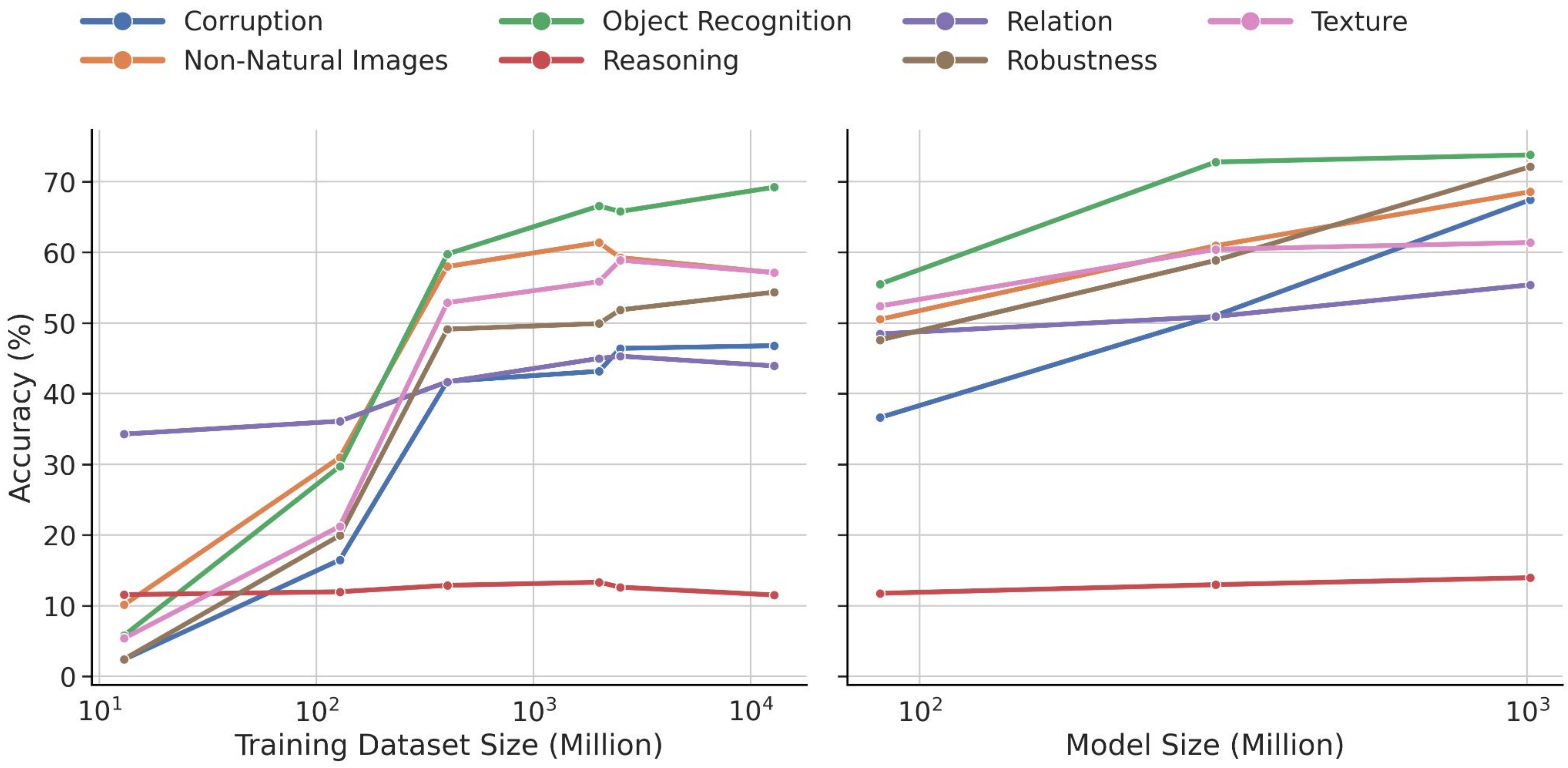
Outline

- Does Scaling **Models** helps in performance?
- How do I select a **Model** for my task?
- Do we need to evaluate on all these **Benchmarks**?

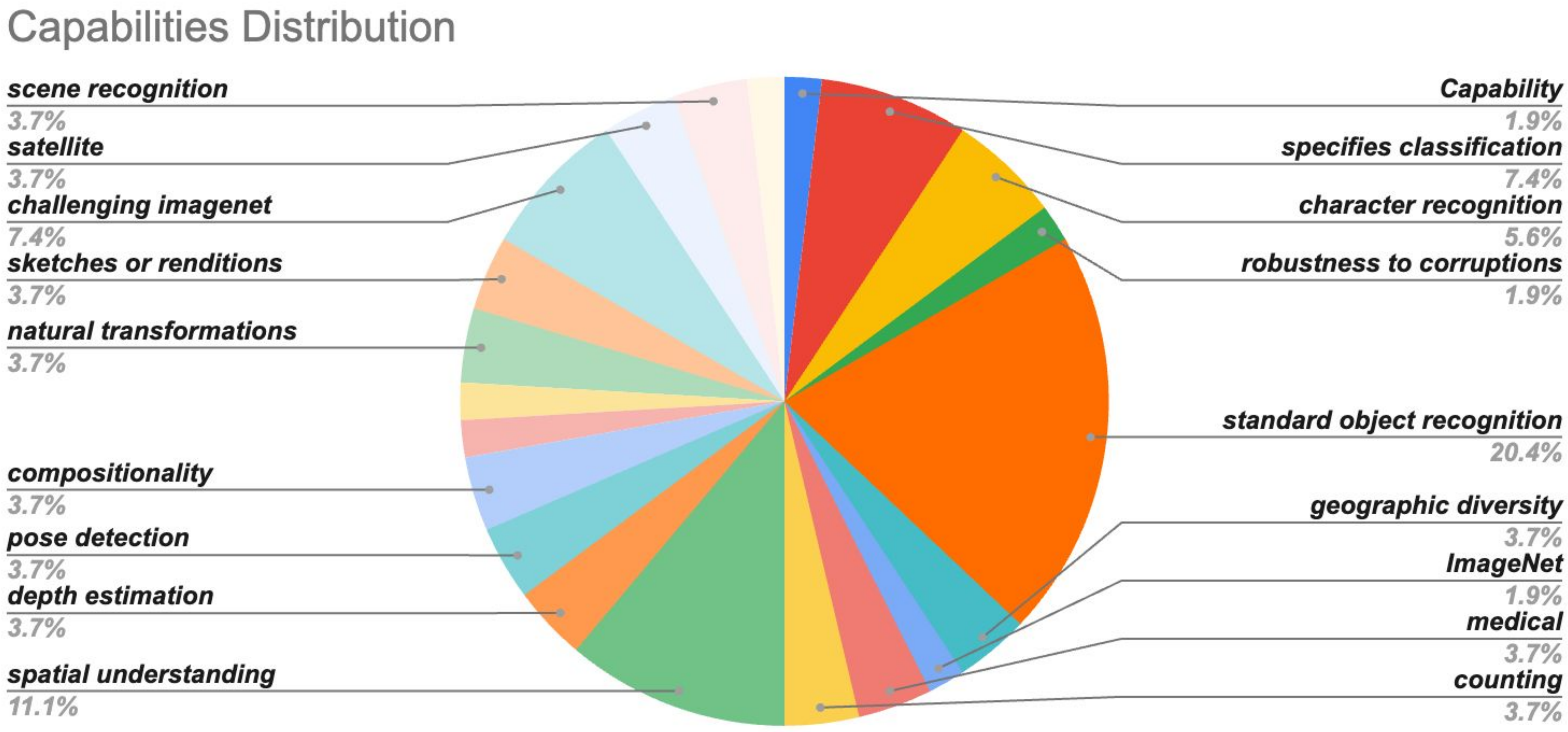
Scaling Training Dataset and Model Size Hardly Helps for Reasoning and Relations



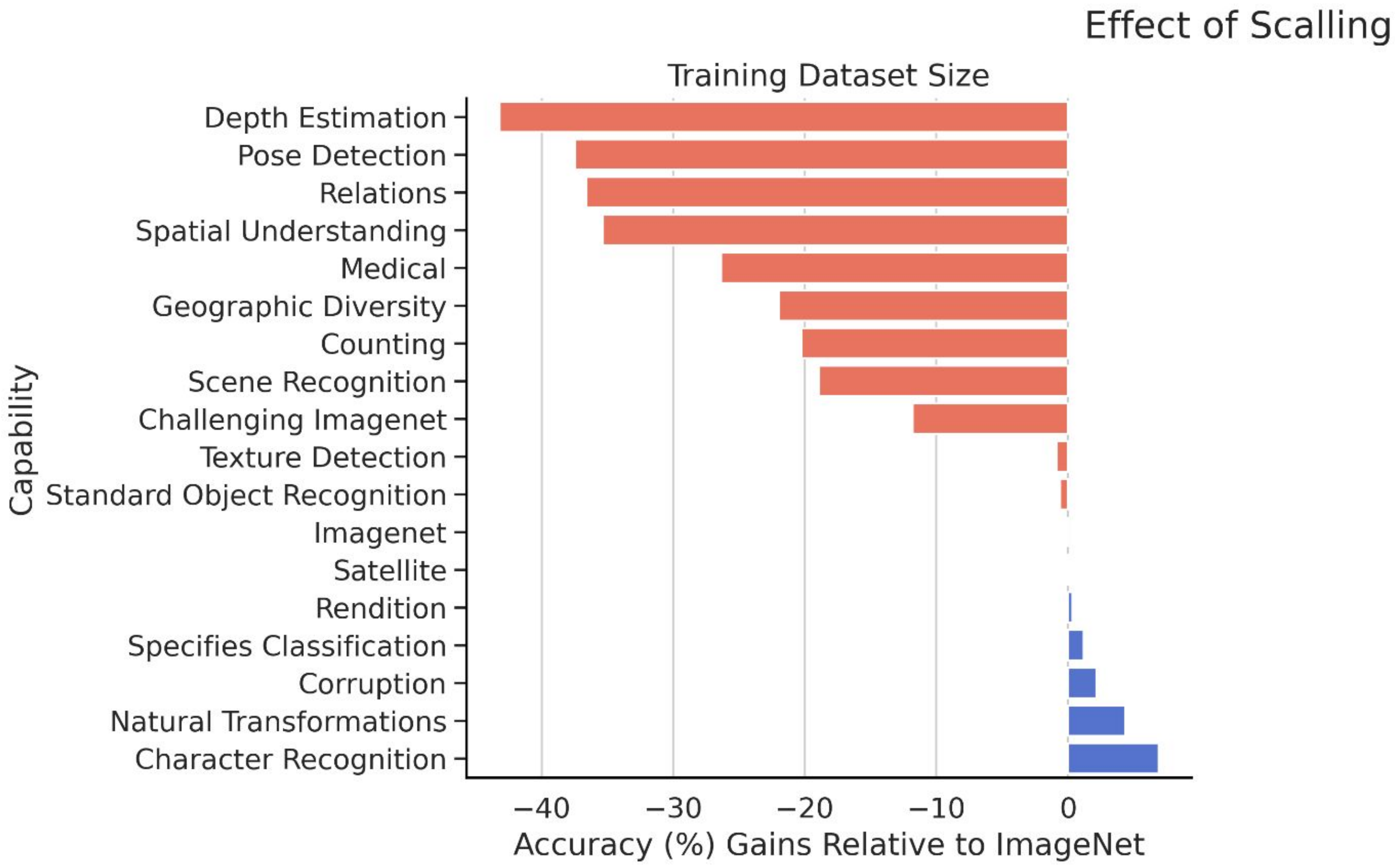
Scaling Training Dataset and Model Size Hardly Helps for Reasoning and Relations



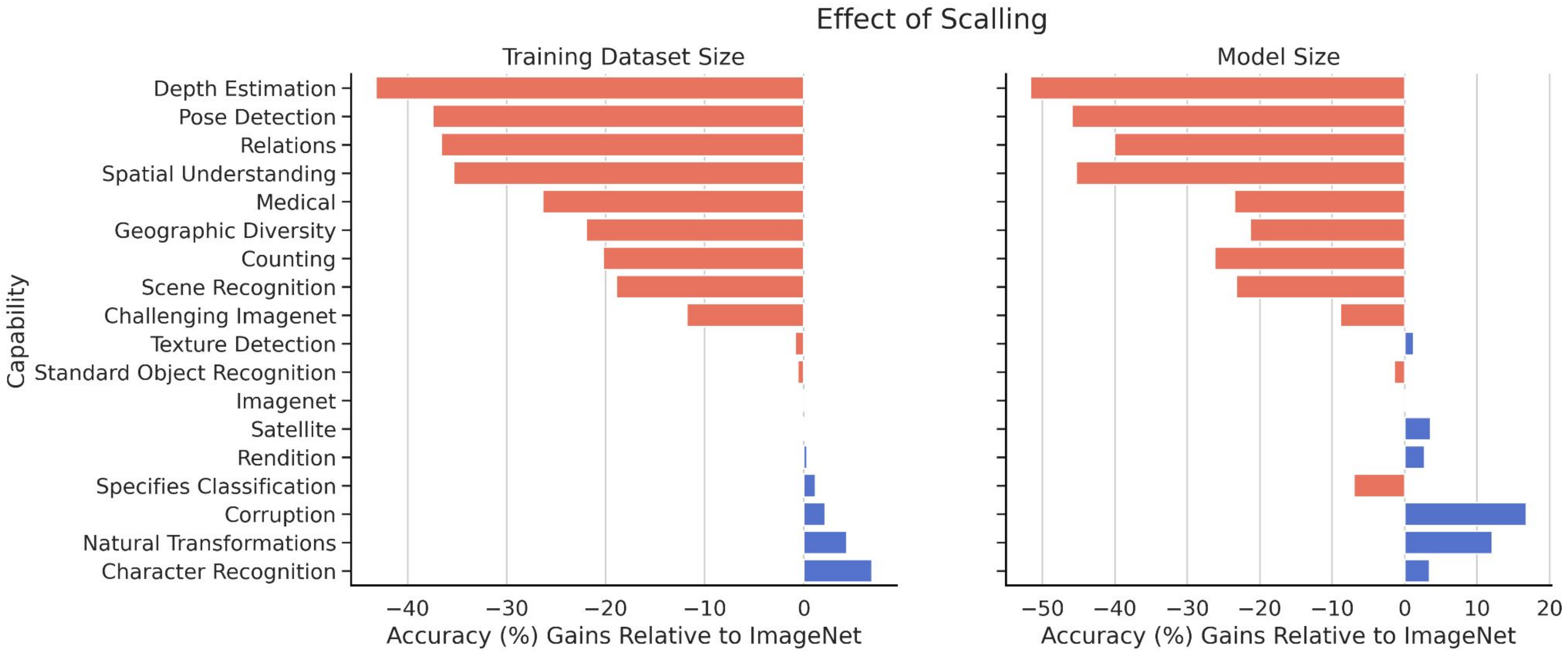
Benchmarks


Capabilities Distribution
scene recognition
3.7%
satellite
3.7%
challenging imagenet
7.4%
sketches or renditions
3.7%
natural transformations
3.7%
compositionality
3.7%
pose detection
3.7%
depth estimation
3.7%
spatial understanding
11.1%
Capability
1.9%
specifies classification
7.4%
character recognition
5.6%
robustness to corruptions
1.9%
standard object recognition
20.4%
geographic diversity
3.7%
ImageNet
1.9%
medical
3.7%
counting
3.7%

Scaling Training Dataset and Model Size Hardly Helps on Fine-grain Tasks



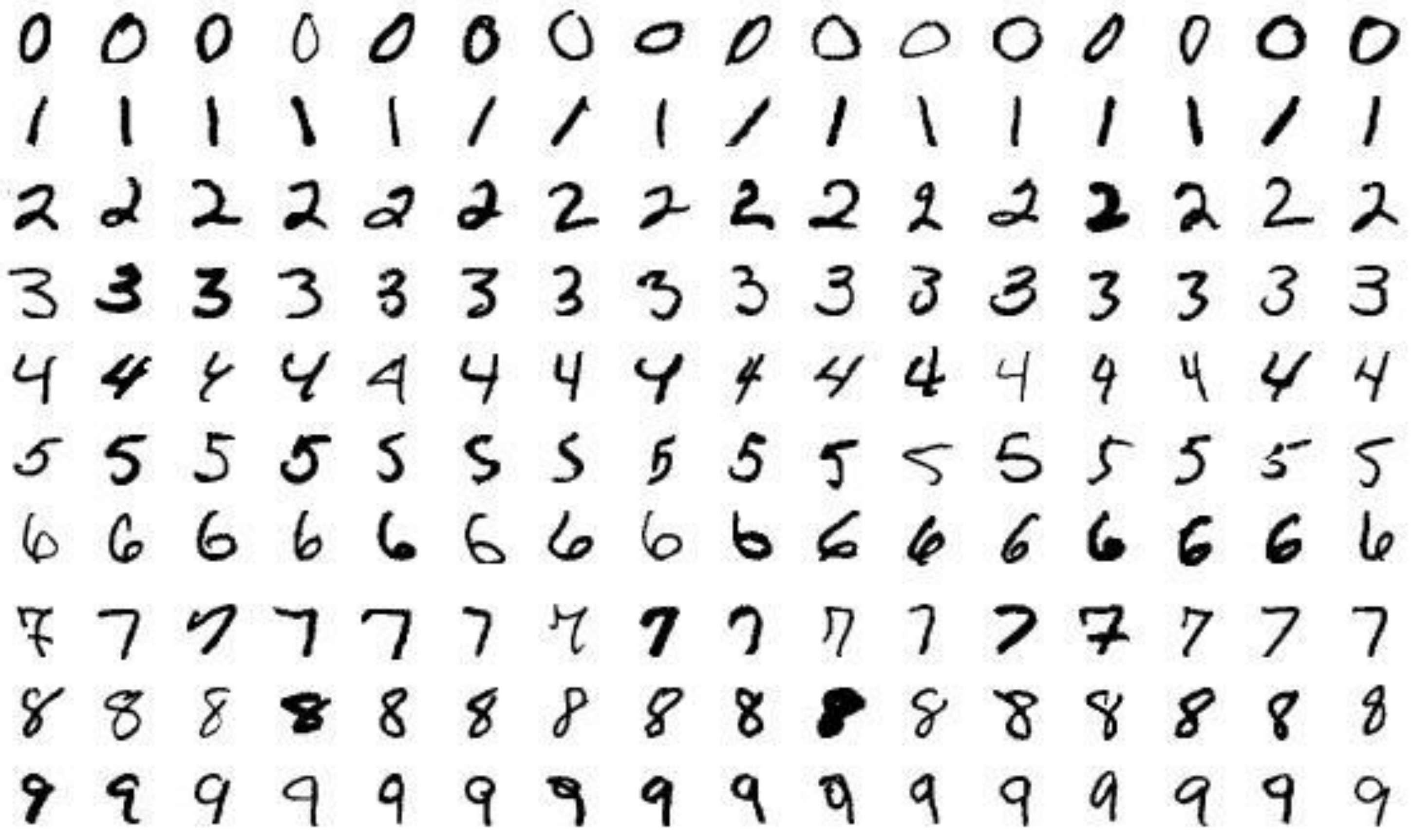
Scaling Training Dataset and Model Size Hardly Helps for Reasoning and Relations



Digit Recognition and Counting are Notable Limitations for VLMs

Digit Recognition and Counting are Notable Limitations for VLMs

MNIST

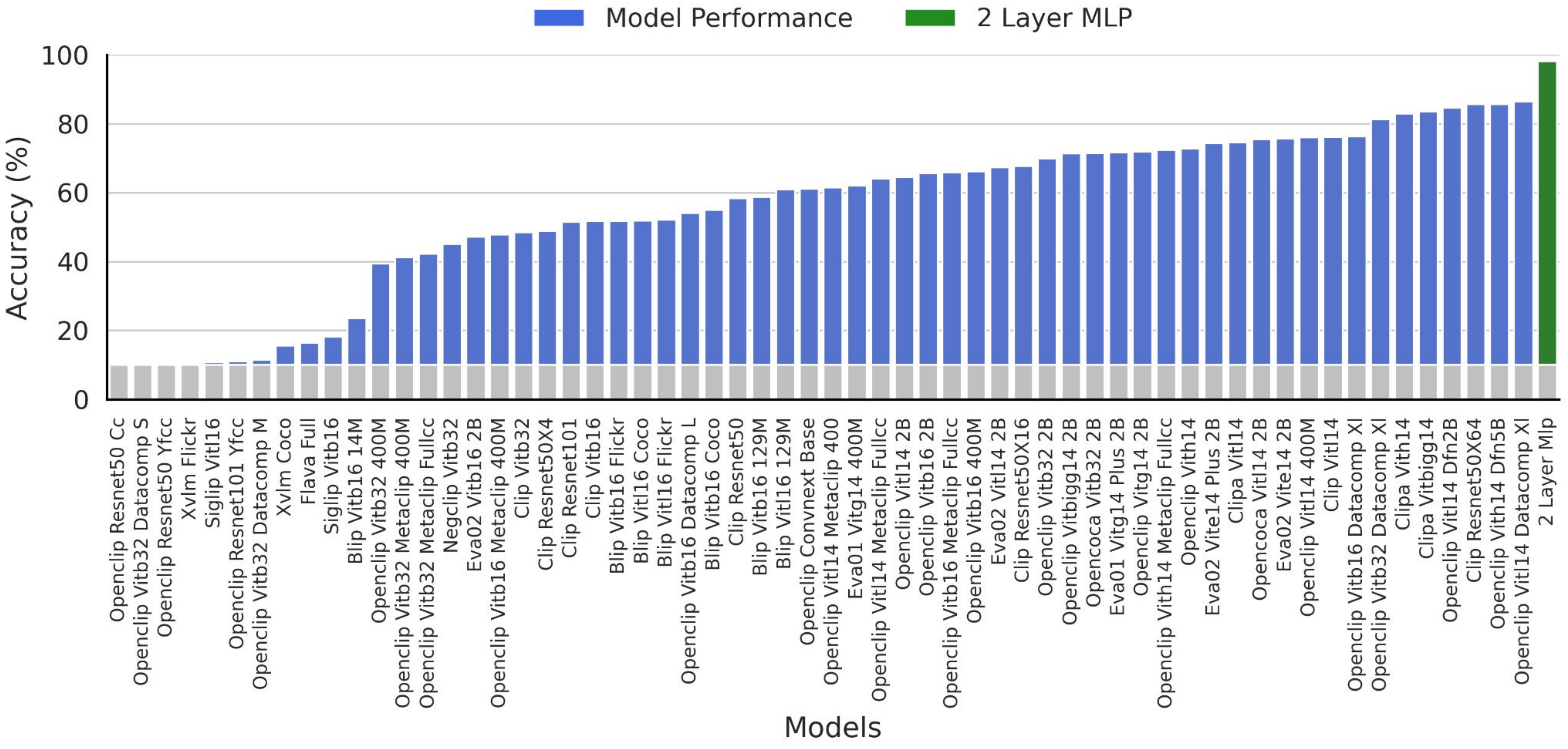


(LeCun, Yann et al., 1998)

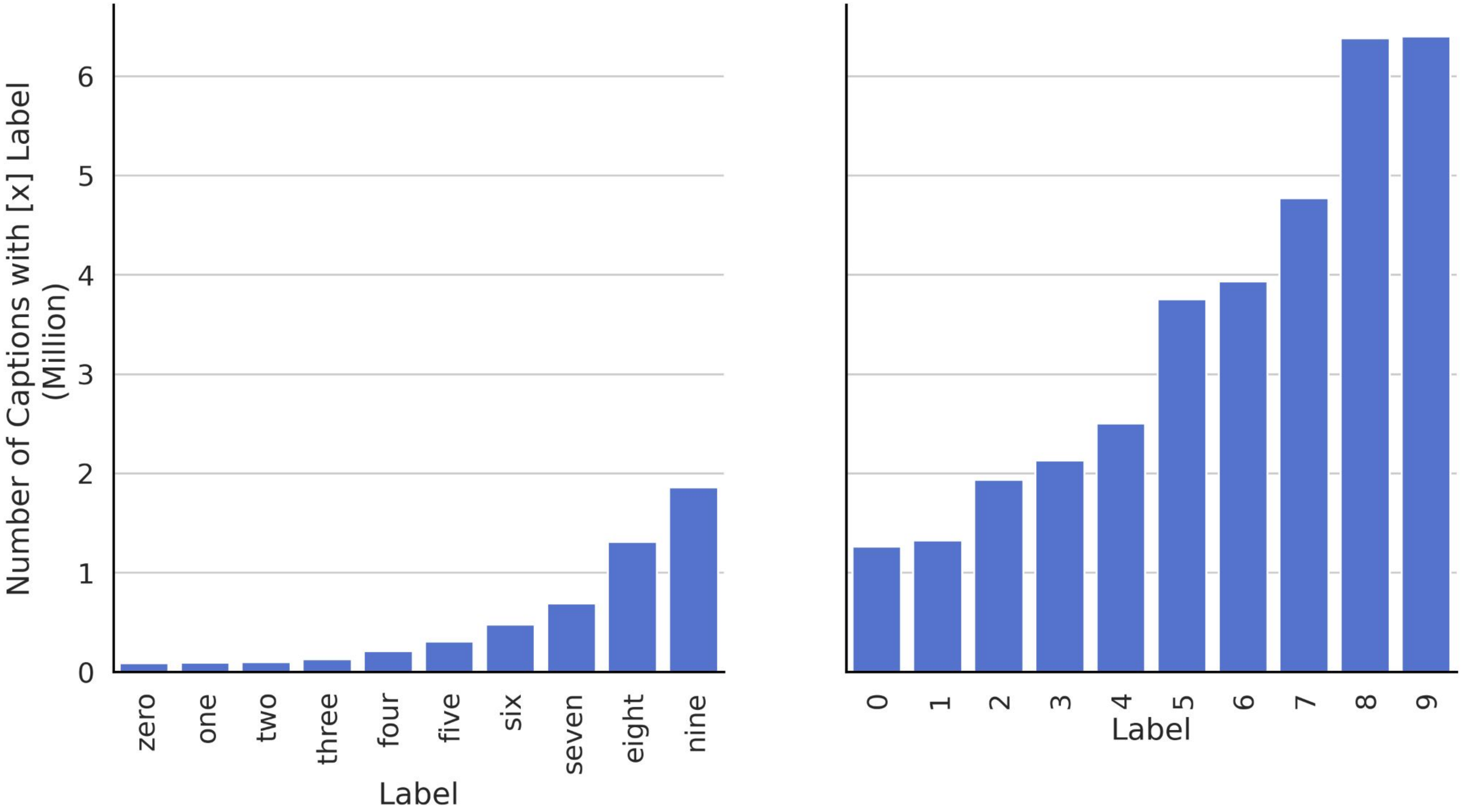
SVHN

(Netzer, Yuval, et al., 2011)

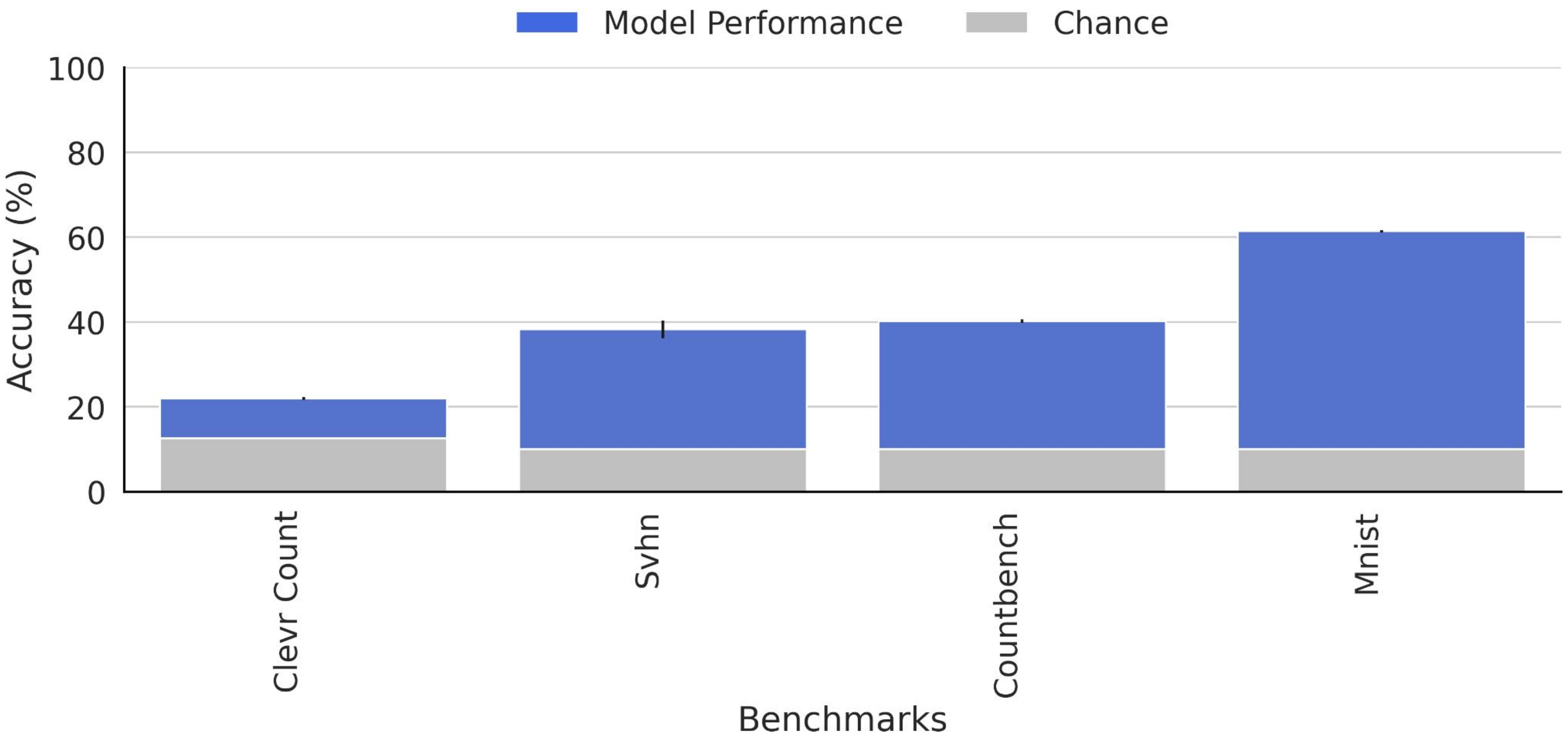
Digit Recognition are Notable Limitations for VLMs



Digit Recognition Limitation not Due to Lack of Data



Digit Recognition and Counting are Notable Limitations for VLMs



Outline

- Does Scaling **Models** helps in performance?

Not for relational understanding and reasoning tasks!

- How do I select a **Model** for my task?
- Do we need to evaluate on all these **Benchmarks**?

Tailored Learning Objectives can Help where Scale does not

Benchmark Type	Mean Performance	Top		Top vs Worst Scale		Worst	
		Model	Performance	Training Dataset Size	Model Size	Performance	Model
Corruption	46.2	EVA02 ViT E 14	74.3	153×	50×	2.4	DataComp ViT B 32
Non-Natural Images	54.1	EVA02 ViT E 14	74.6	153×	50×	16.1	DataComp ViT B 32
Object Recognition	55.0	CLIPA ViT G 14	71.1	98×	21×	12.1	DataComp ViT B 32
Reasoning	14.9	OpenCLIP ViT g 14	19.0	133×	18×	10.6	OpenCLIP ResNet101
Relation	46.7	NegCLIP ViT B 32	66.8	30×	1×	33.2	DataComp ViT B 32
Robustness	52.1	EVA02 ViT E 14	72.8	153×	50×	3.8	DataComp ViT B 32
Texture	53.5	MetaCLIP ViT H 14	72.5	192×	7×	5.4	DataComp ViT B 32
Overall	46.1	EVA02 ViT E 14	61.2	153×	50×	12.1	DataComp ViT B 32

Tailored Learning Objectives can Help where Scale does not

Benchmark Type	Mean Performance	Top		Top vs Worst Scale		Worst	
		Model	Performance	Training Dataset Size	Model Size	Performance	Model
Corruption	46.2	EVA02 ViT E 14	74.3	153×	50×	2.4	DataComp ViT B 32
Non-Natural Images	54.1	EVA02 ViT E 14	74.6	153×	50×	16.1	DataComp ViT B 32
Object Recognition	55.0	CLIPA ViT G 14	71.1	98×	21×	12.1	DataComp ViT B 32
Reasoning	14.9	OpenCLIP ViT g 14	19.0	133×	18×	10.6	OpenCLIP ResNet101
Relation	46.7	NegCLIP ViT B 32	66.8	30×	1×	33.2	DataComp ViT B 32
Robustness	52.1	EVA02 ViT E 14	72.8	153×	50×	3.8	DataComp ViT B 32
Texture	53.5	MetaCLIP ViT H 14	72.5	192×	7×	5.4	DataComp ViT B 32
Overall	46.1	EVA02 ViT E 14	61.2	153×	50×	12.1	DataComp ViT B 32

Tailored Learning Objectives can Help where Scale does not

Benchmark Type	Mean Performance	Top		Top vs Worst Scale		Worst	
		Model	Performance	Training Dataset Size	Model Size	Performance	Model
Corruption	46.2	EVA02 ViT E 14	74.3	153×	50×	2.4	DataComp ViT B 32
Non-Natural Images	54.1	EVA02 ViT E 14	74.6	153×	50×	16.1	DataComp ViT B 32
Object Recognition	55.0	CLIPA ViT G 14	71.1	98×	21×	12.1	DataComp ViT B 32
Reasoning	14.9	OpenCLIP ViT g 14	19.0	133×	18×	10.6	OpenCLIP ResNet101
Relation	46.7	NegCLIP ViT B 32	66.8	30×	1×	33.2	DataComp ViT B 32
Robustness	52.1	EVA02 ViT E 14	72.8	153×	50×	3.8	DataComp ViT B 32
Texture	53.5	MetaCLIP ViT H 14	72.5	192×	7×	5.4	DataComp ViT B 32
Overall	46.1	EVA02 ViT E 14	61.2	153×	50×	12.1	DataComp ViT B 32

Tailored Learning Objectives can Help where Scale does not

Benchmark Type	Mean Performance	Top		Top vs Worst Scale		Worst	
		Model	Performance	Training Dataset Size	Model Size	Performance	Model
Corruption	46.2	EVA02 ViT E 14	74.3	153×	50×	2.4	DataComp ViT B 32
Non-Natural Images	54.1	EVA02 ViT E 14	74.6	153×	50×	16.1	DataComp ViT B 32
Object Recognition	55.0	CLIPA ViT G 14	71.1	98×	21×	12.1	DataComp ViT B 32
Reasoning	14.9	OpenCLIP ViT g 14	19.0	133×	18×	10.6	OpenCLIP ResNet101
Relation	46.7	NegCLIP ViT B 32	66.8	30×	1×	33.2	DataComp ViT B 32
Robustness	52.1	EVA02 ViT E 14	72.8	153×	50×	3.8	DataComp ViT B 32
Texture	53.5	MetaCLIP ViT H 14	72.5	192×	7×	5.4	DataComp ViT B 32
Overall	46.1	EVA02 ViT E 14	61.2	153×	50×	12.1	DataComp ViT B 32

Tailored Learning Objectives can Help where Scale does not

Benchmark Type	Mean Performance	Top		Top vs Worst Scale		Worst	
		Model	Performance	Training Dataset Size	Model Size	Performance	Model
Corruption	46.2	EVA02 ViT E 14	74.3	153×	50×	2.4	DataComp ViT B 32
Non-Natural Images	54.1	EVA02 ViT E 14	74.6	153×	50×	16.1	DataComp ViT B 32
Object Recognition	55.0	CLIPA ViT G 14	71.1	98×	21×	12.1	DataComp ViT B 32
Reasoning	14.9	OpenCLIP ViT g 14	19.0	133×	18×	10.6	OpenCLIP ResNet101
Relation	46.7	NegCLIP ViT B 32	66.8	30×	1×	33.2	DataComp ViT B 32
Robustness	52.1	EVA02 ViT E 14	72.8	153×	50×	3.8	DataComp ViT B 32
Texture	53.5	MetaCLIP ViT H 14	72.5	192×	7×	5.4	DataComp ViT B 32
Overall	46.1	EVA02 ViT E 14	61.2	153×	50×	12.1	DataComp ViT B 32

Which model should you use?

Benchmark Type	Mean Performance	Top		Top vs Worst Scale		Worst	
		Model	Performance	Training Dataset Size	Model Size	Performance	Model
Corruption	46.2	EVA02 ViT E 14	74.3	153×	50×	2.4	DataComp ViT B 32
Non-Natural Images	54.1	EVA02 ViT E 14	74.6	153×	50×	16.1	DataComp ViT B 32
Object Recognition	55.0	CLIPA ViT G 14	71.1	98×	21×	12.1	DataComp ViT B 32
Reasoning	14.9	OpenCLIP ViT g 14	19.0	133×	18×	10.6	OpenCLIP ResNet101
Relation	46.7	NegCLIP ViT B 32	66.8	30×	1×	33.2	DataComp ViT B 32
Robustness	52.1	EVA02 ViT E 14	72.8	153×	50×	3.8	DataComp ViT B 32
Texture	53.5	MetaCLIP ViT H 14	72.5	192×	7×	5.4	DataComp ViT B 32
Overall	46.1	EVA02 ViT E 14	61.2	153×	50×	12.1	DataComp ViT B 32

Outline

- Does Scaling **Models** helps in performance?

Not for relational understanding and reasoning tasks!

- How do I select a **Model** for my task?

Based on the task, or EVA02…

- Do we need to evaluate on all these **Benchmarks**?

No need to evaluate on all 53 benchmarks, only 8 is enough

Benchmark Type	Most Correlated Benchmark	Correlation Value
Object recognition	ImageNet-1k	0.82
Reasoning (Counting)	CountBench	0.76
Reasoning (Spatial)	DSPR Position	0.29
Relation	VG Attribution	0.57
Texture	DTD	1
Non-Natural Images	Resisc45	0.72
Robustness	ImageNet-v2	0.81
Corruption	ImageNet-c	1

No need to evaluate on all 53 benchmarks, only 8 is enough

Benchmark Type	Most Correlated Benchmark	Correlation Value
Object recognition	ImageNet-1k	0.82
Reasoning (Counting)	CountBench	0.76
Reasoning (Spatial)	DSPR Position	0.29
Relation	VG Attribution	0.57
Texture	DTD	1
Non-Natural Images	Resisc45	0.72
Robustness	ImageNet-v2	0.81
Corruption	ImageNet-c	1

Takes **5 minutes** on a single GPU to evaluate on the 8 set of Benchmarks

Outline

- Does Scaling **Models** helps in performance?

Not for relational understanding and reasoning tasks!

- How do I select a **Model** for my task?

Based on the task, or EVA02…

- Do we need to evaluate on all these **Benchmarks**?

No, 8 is all you need!

UniBench Ease of Use

```python
import uni_bench
from torchvision.datasets import MNIST

evaluator = uni_bench.Evaluator()

# add a new model
evaluator.add_model(vision, text, tokenizer, model_name)

# add a new benchmark, accepts any torch.utils.data dataset
evaluator.add_benchmark(MNIST)

evaluator.evaluate()
```

UniBench

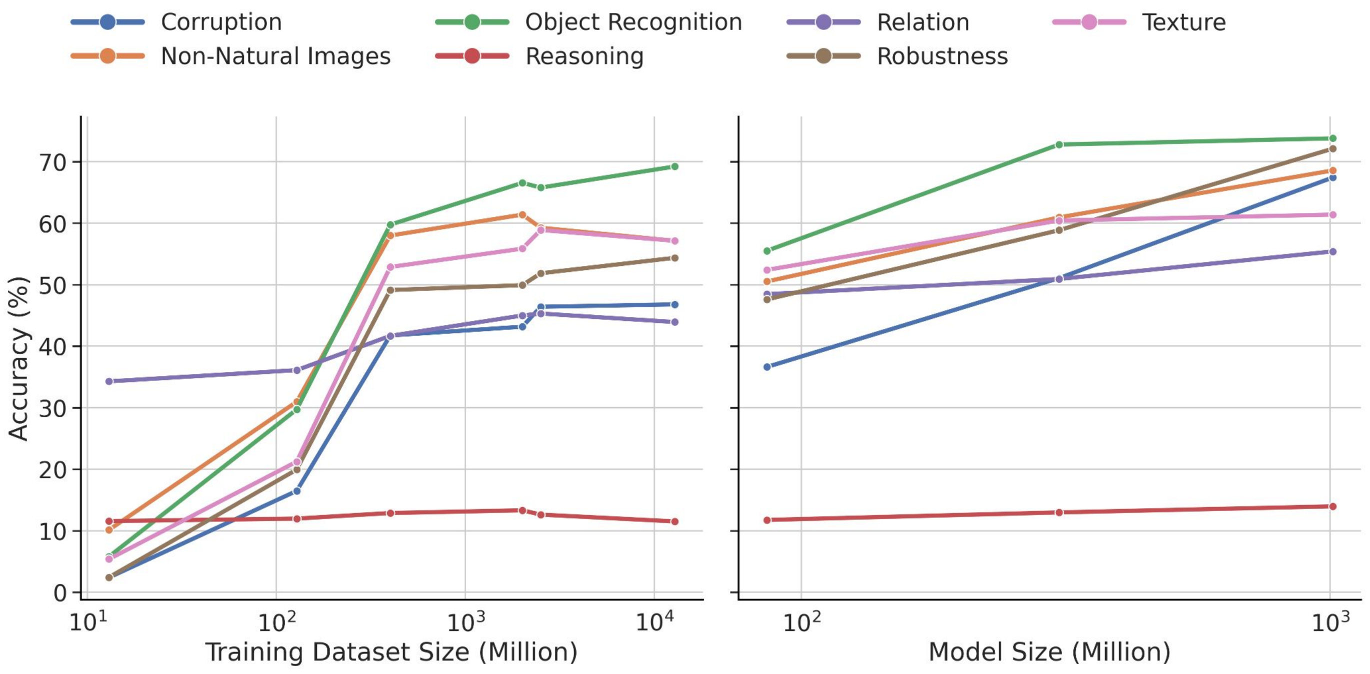
Repo: github.com/facebookresearch/unibench

50+

vision-language model **benchmarks** in a [unified codebase]()

reveals the **limits of scaling**

for visual reasoning & relationals



runs **5 minutes** on a single GPU

for representative capabilities

```python
import uni_bench

evaluator = uni_bench.Evaluator()
evaluator.evaluate()
```