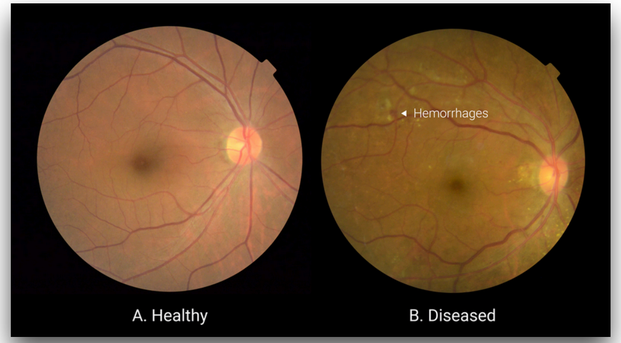
A data-centric view on robustness

Ludwig Schmidt

UNIVERSITY of WASHINGTON

Safety-critical applications of ML

Transportation



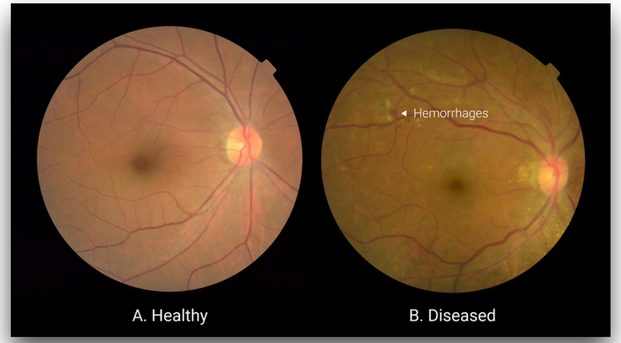
Health care

Robotics

Content moderation

Safety-critical applications of ML

Transportation



Health care

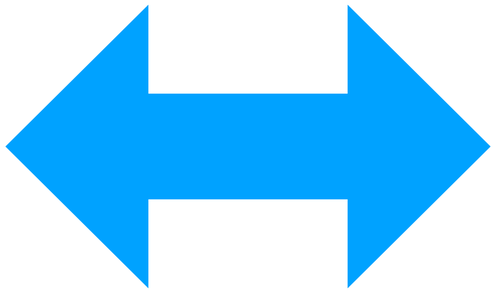
Robotics

Content moderation

Need **reliable** machine learning

How can we make ML reliable?

Better training **algorithms**, models, etc.



Better training **data**

Outline

1. Overview of the robustness landscape in computer vision

2. New image text-models (e.g., OpenAI's CLIP model) are (a lot) more robust

3. Where does CLIP's robustness come from? → Training data

Outline

1. Overview of the robustness landscape in computer vision

2. New image text-models (e.g., OpenAI's CLIP model) are (a lot) more robust

3. Where does CLIP's robustness come from? → Training data

[Deng, Dong, Socher, Li, Li, Fei-Fei'09]

[Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, Berg Fei-Fei'15]

Robustness on ImageNet

Lots of progress on ImageNet over the past 10 years, but models are still not robust.

Evaluation: **new test sets**

Robustness on ImageNet

Lots of progress on ImageNet over the past 10 years, but models are still not robust.

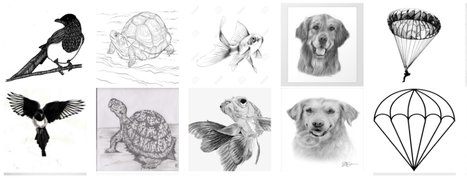
Evaluation: **new test sets**

ImageNetV2

[Recht, Roelofs, Schmidt, Shankar ’19]

ObjectNet

[Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz ’19]



ImageNet-Sketch

[Wang, Ge, Lipton, Xing ’19]

ImageNet-R

[Hendrycks, Basart, Mu, Kadavath, Wang, Dorundo, Desai, Zhu, Parajuli, Guo, Song, Steinhardt, Gilmer ’20]

Quantifying Robustness

Often in-distribution (“standard”) accuracy acts as a **confounder**.

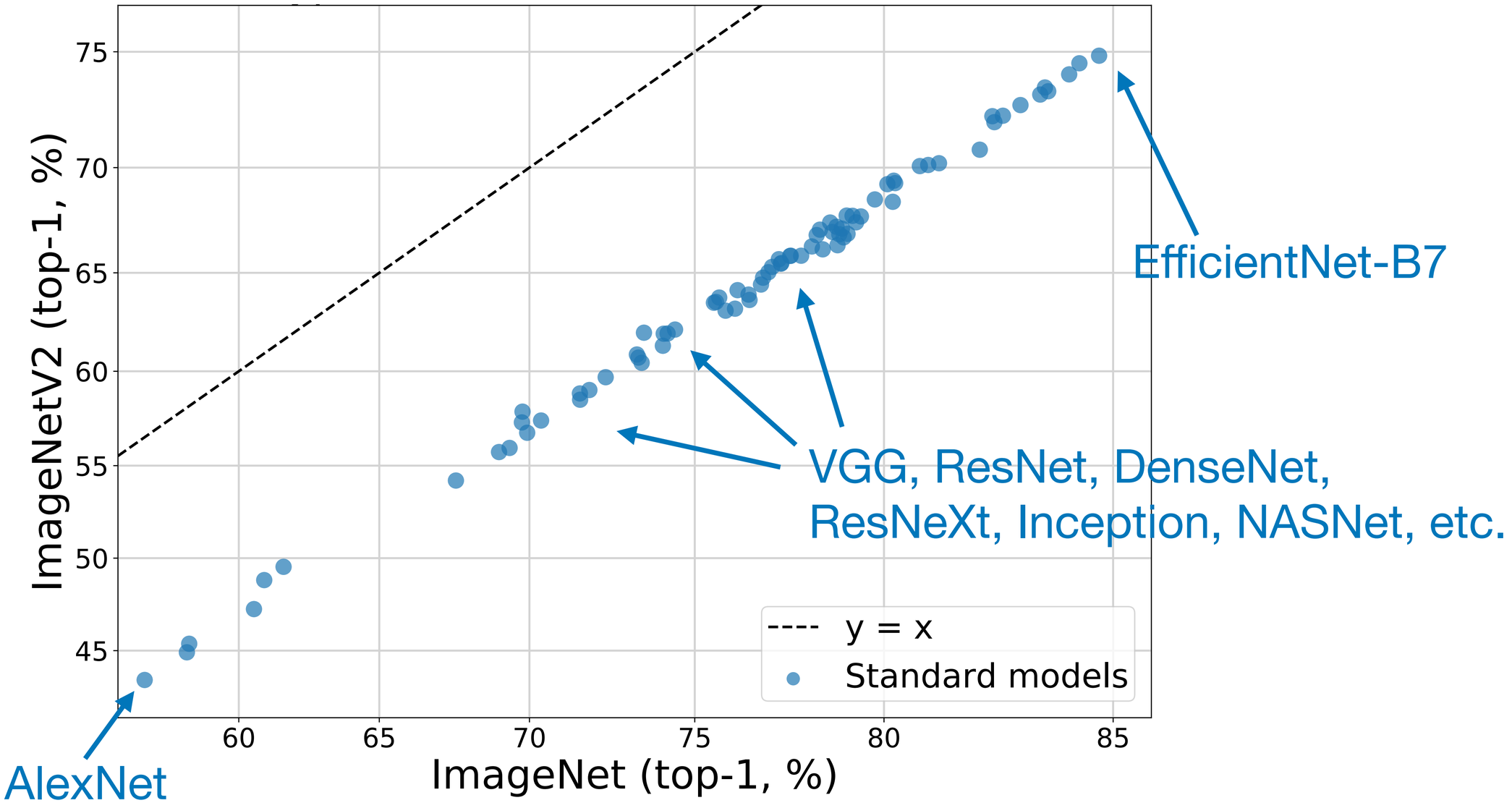
	In-distribution (Source) Accuracy	Out-of-distribution (Target) Accuracy
Model A	80%	75%
Model B	90%	77%

Quantifying Robustness

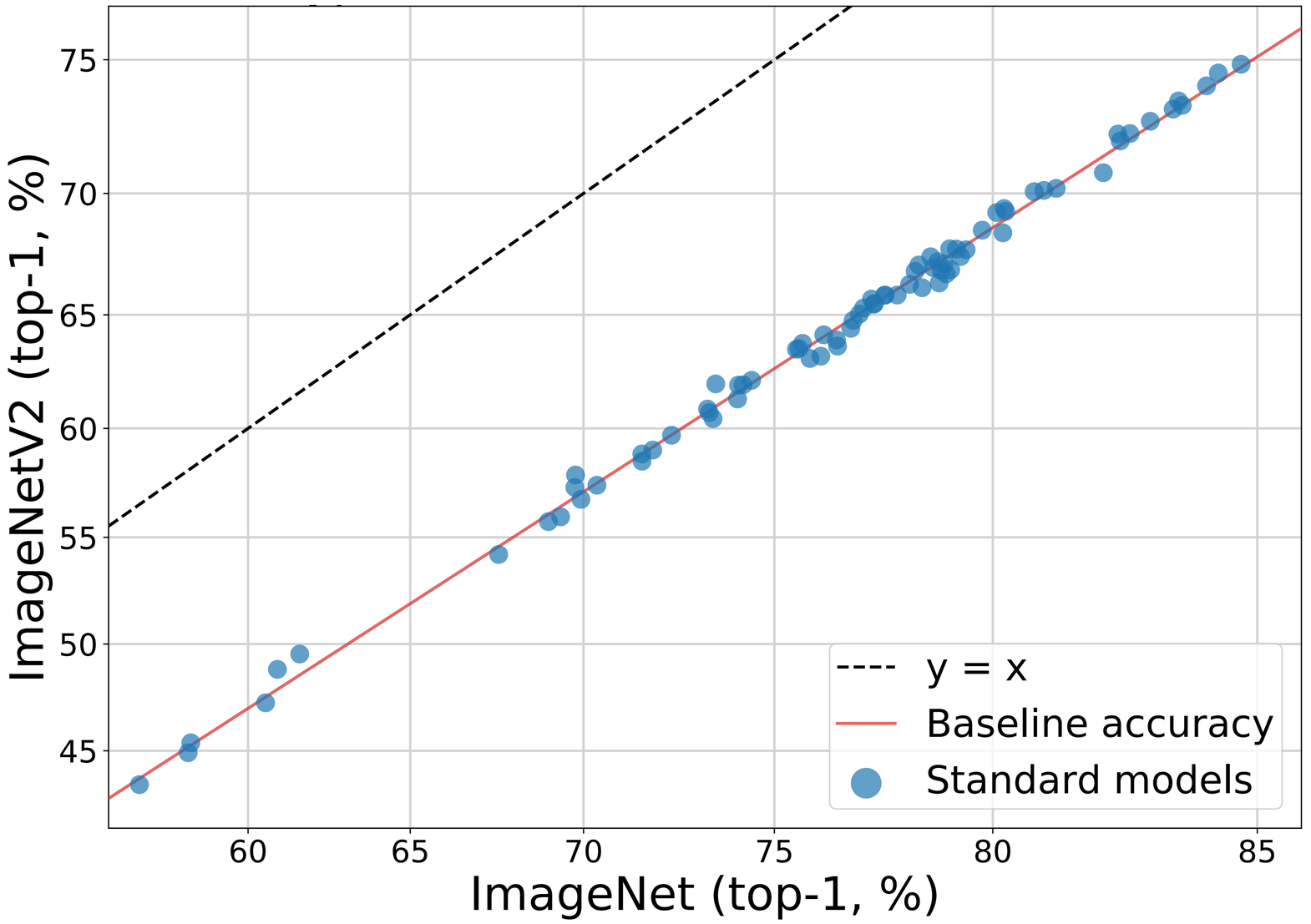
Often in-distribution (“standard”) accuracy acts as a **confounder**.

	In-distribution (Source) Accuracy	Out-of-distribution (Target) Accuracy	Accuracy Drop
Model A	80%	75%	5%
Model B	90%	77%	13%

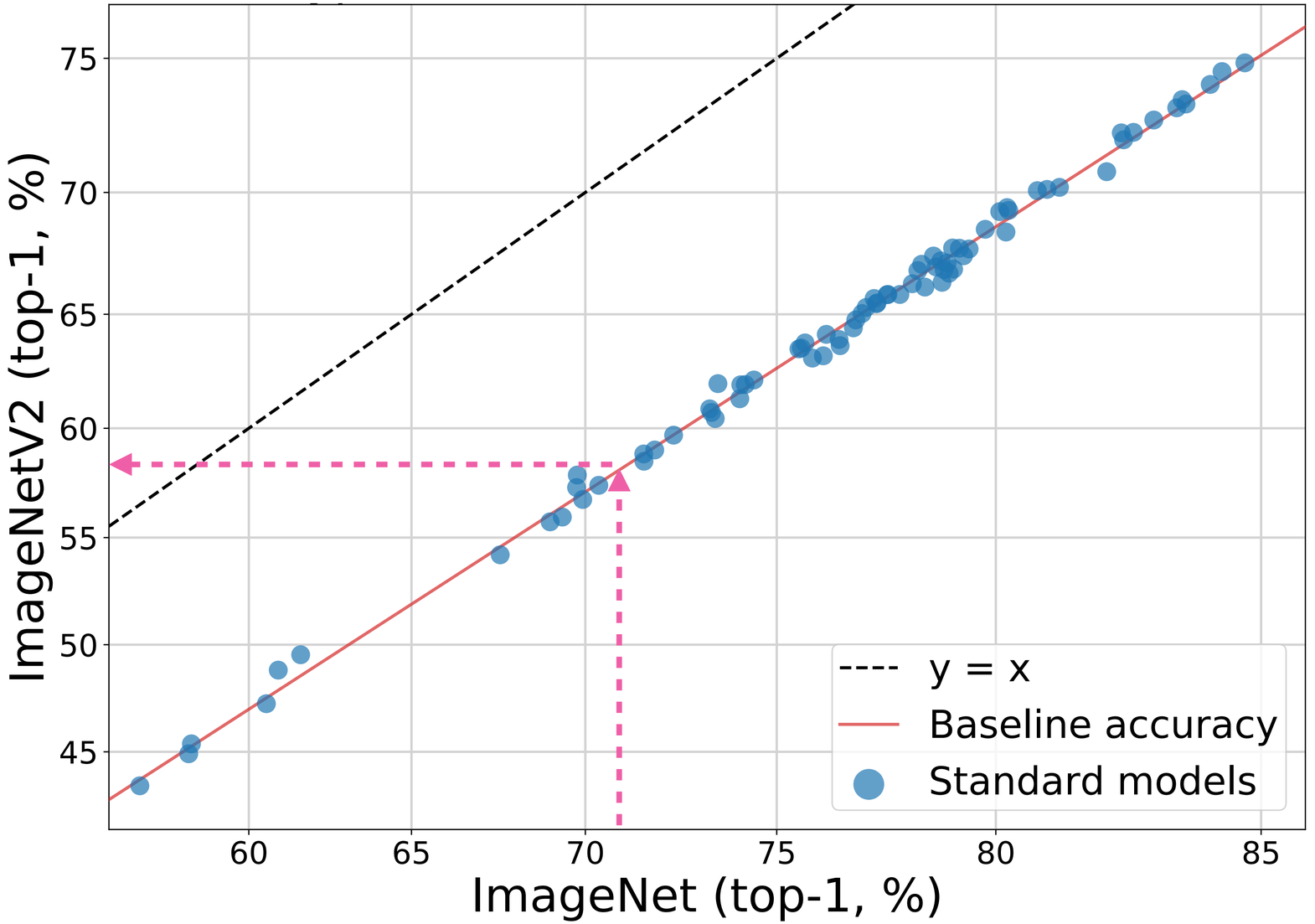
How do we compare models with different in-distribution accuracy?



[Taori, Dave, Shankar, Carlini, Recht, Schmidt ’20]

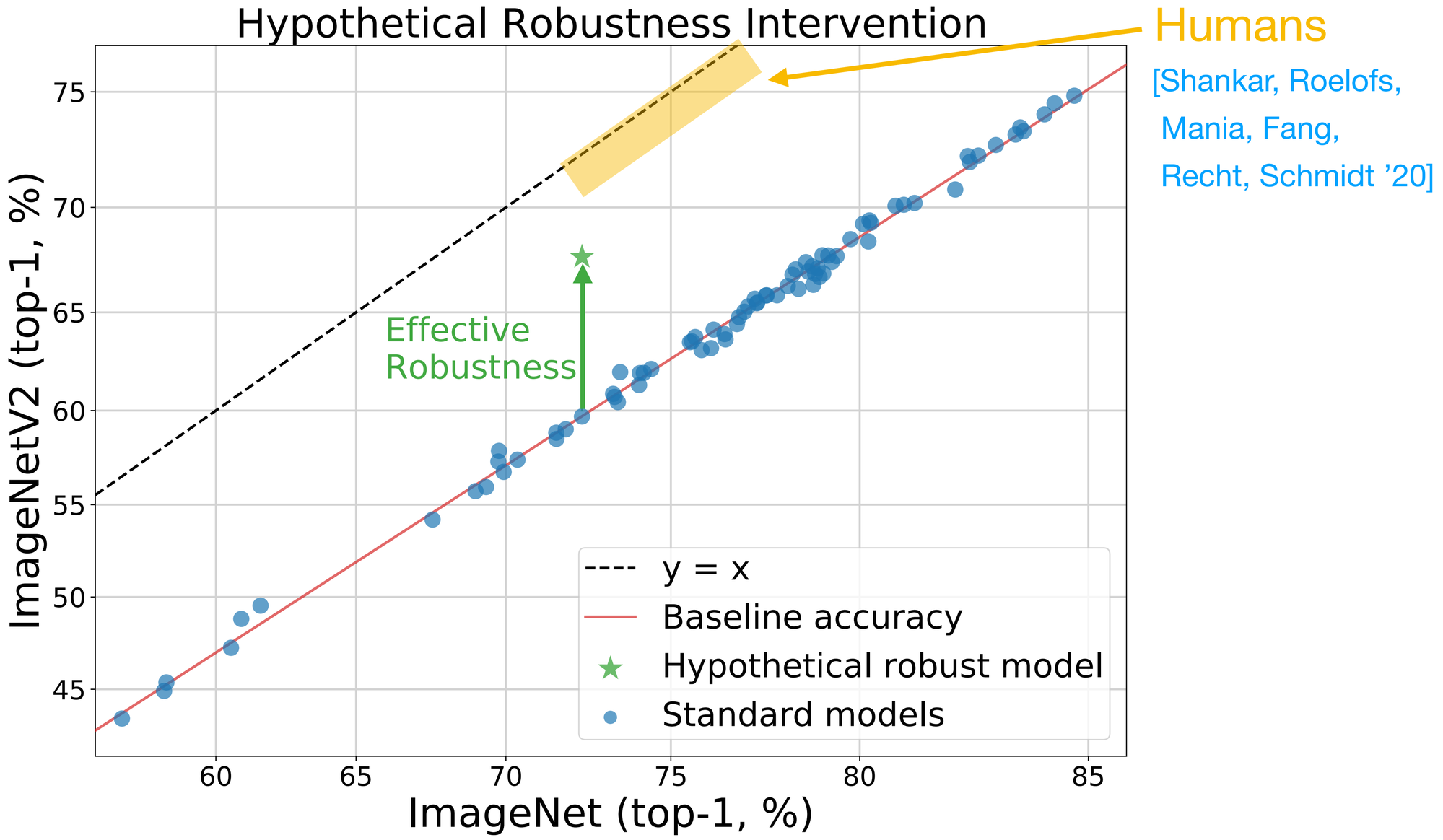

ImageNetV2 (top-1, %)
75
70
65
60
55
50
45
60
65
70
75
80
85
ImageNet (top-1, %)
y = x
Baseline accuracy
Standard models

Expected out-of-distribution accuracy



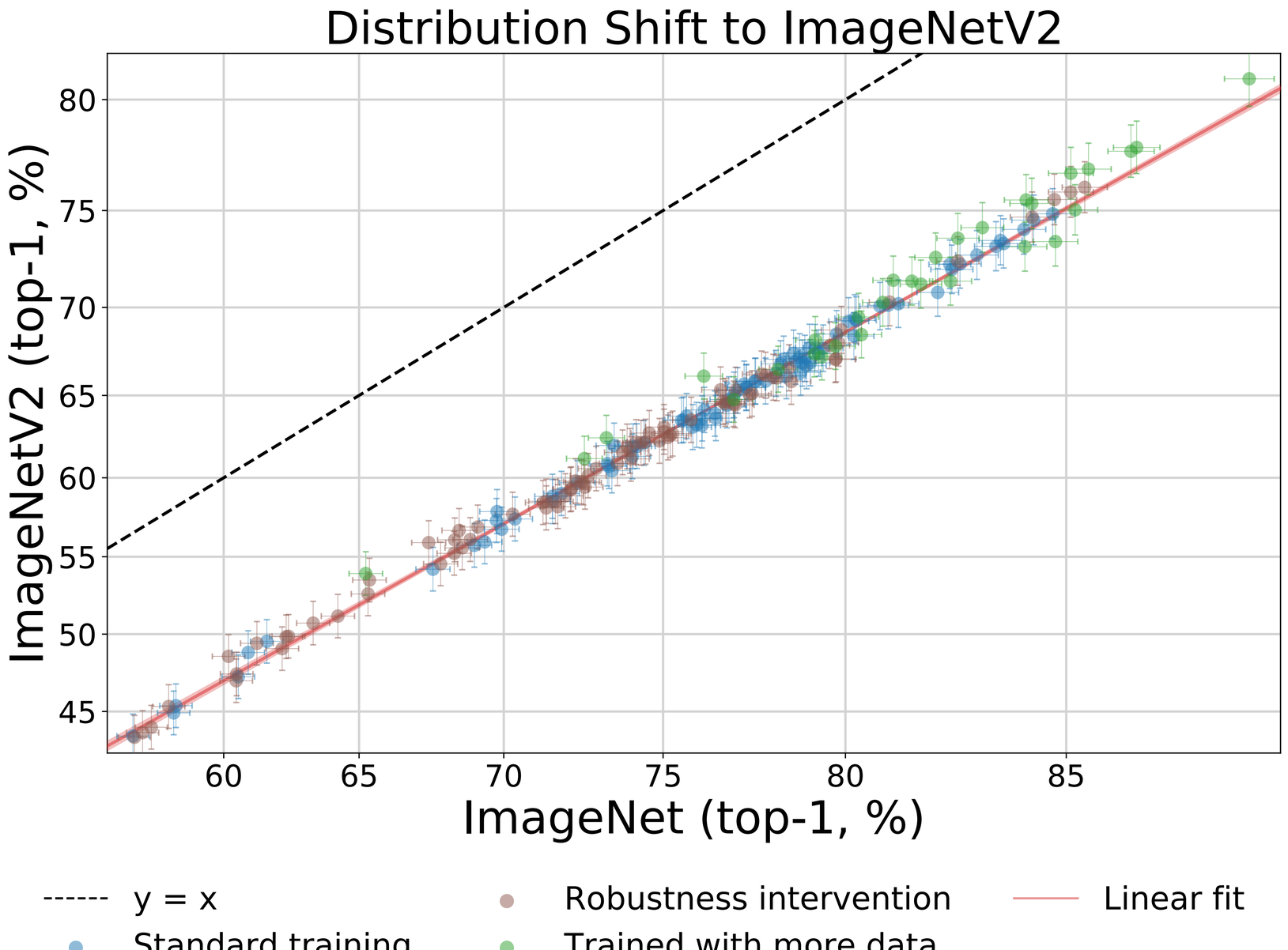
In-distribution accuracy

Baseline **out-of-distribution accuracy** from **in-distribution accuracy.**



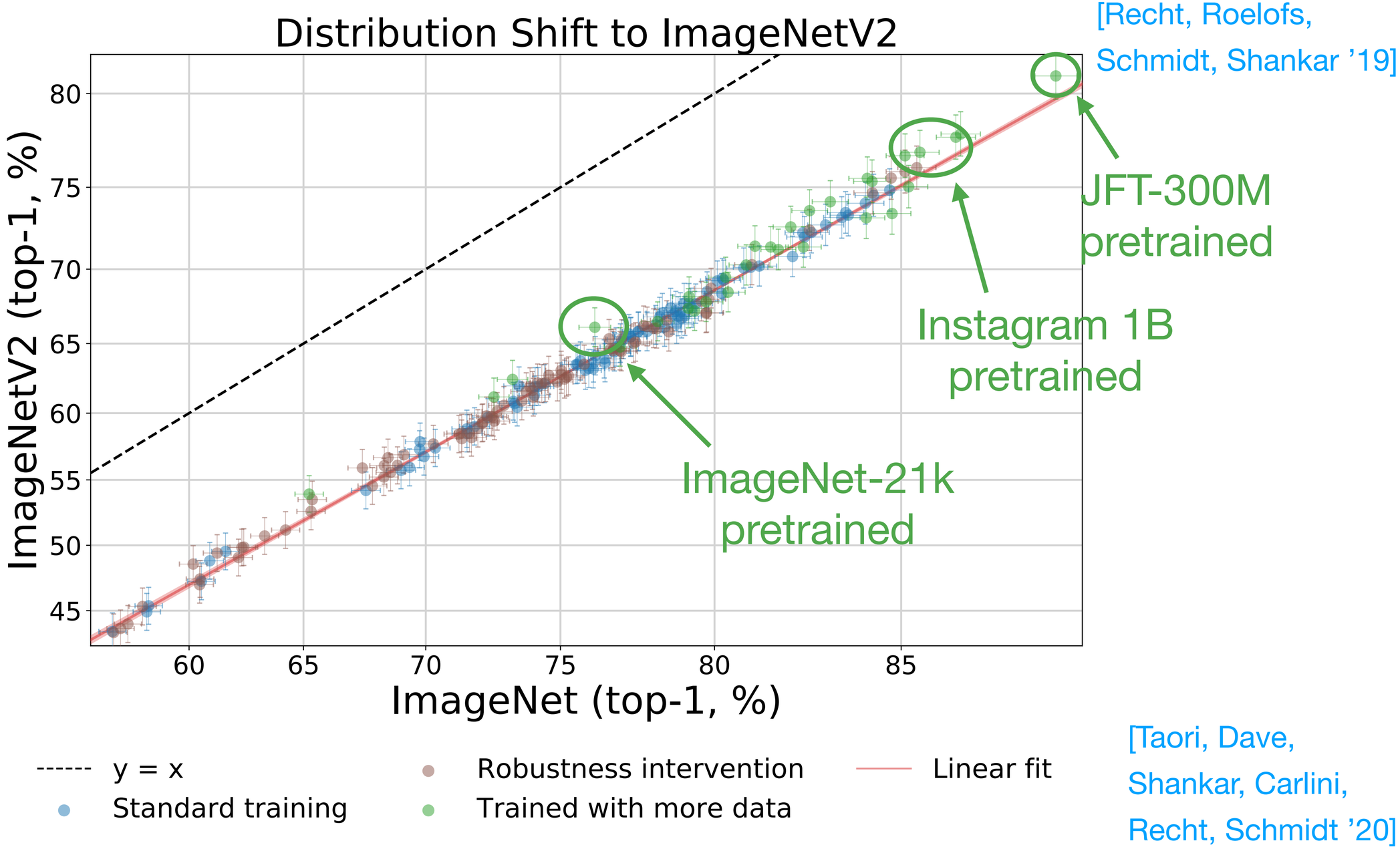
Do current robustness interventions achieve effective robustness?

[Recht, Roelofs, Schmidt, Shankar ’19]



[Taori, Dave, Shankar, Carlini, Recht, Schmidt ’20]

Only training on (a lot) **more data** gives a small amount of effective robustness.

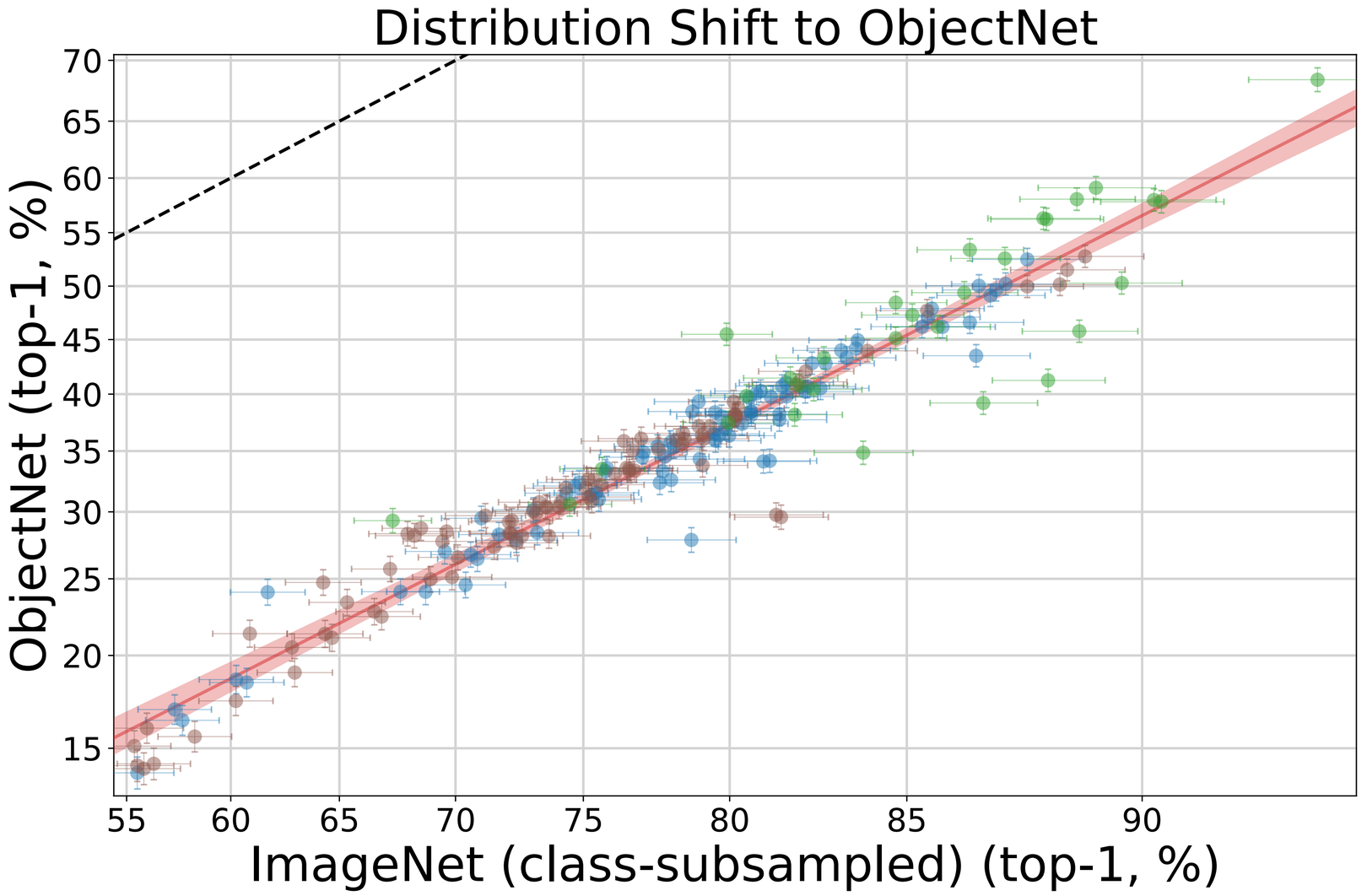


[Recht, Roelofs, Schmidt, Shankar '19]

[Taori, Dave, Shankar, Carlini, Recht, Schmidt '20]

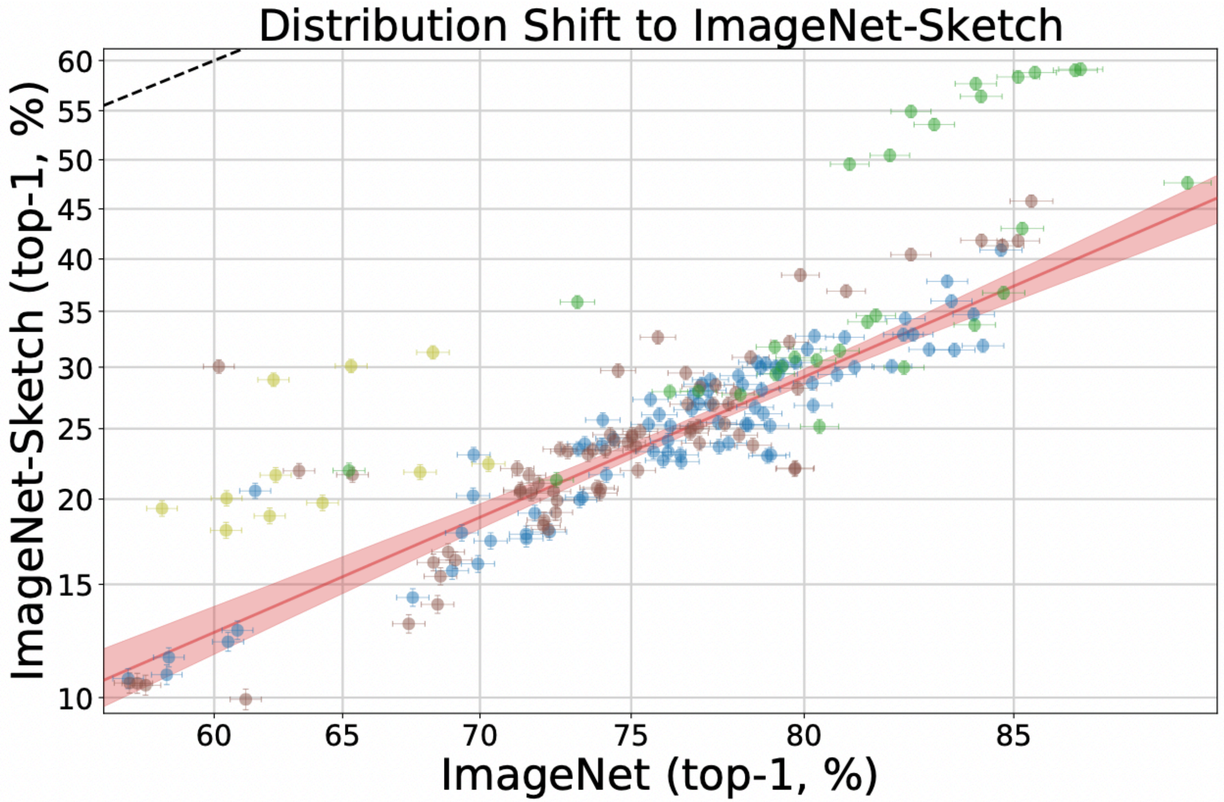
Only training on (a lot) **more data** gives a small amount of effective robustness.

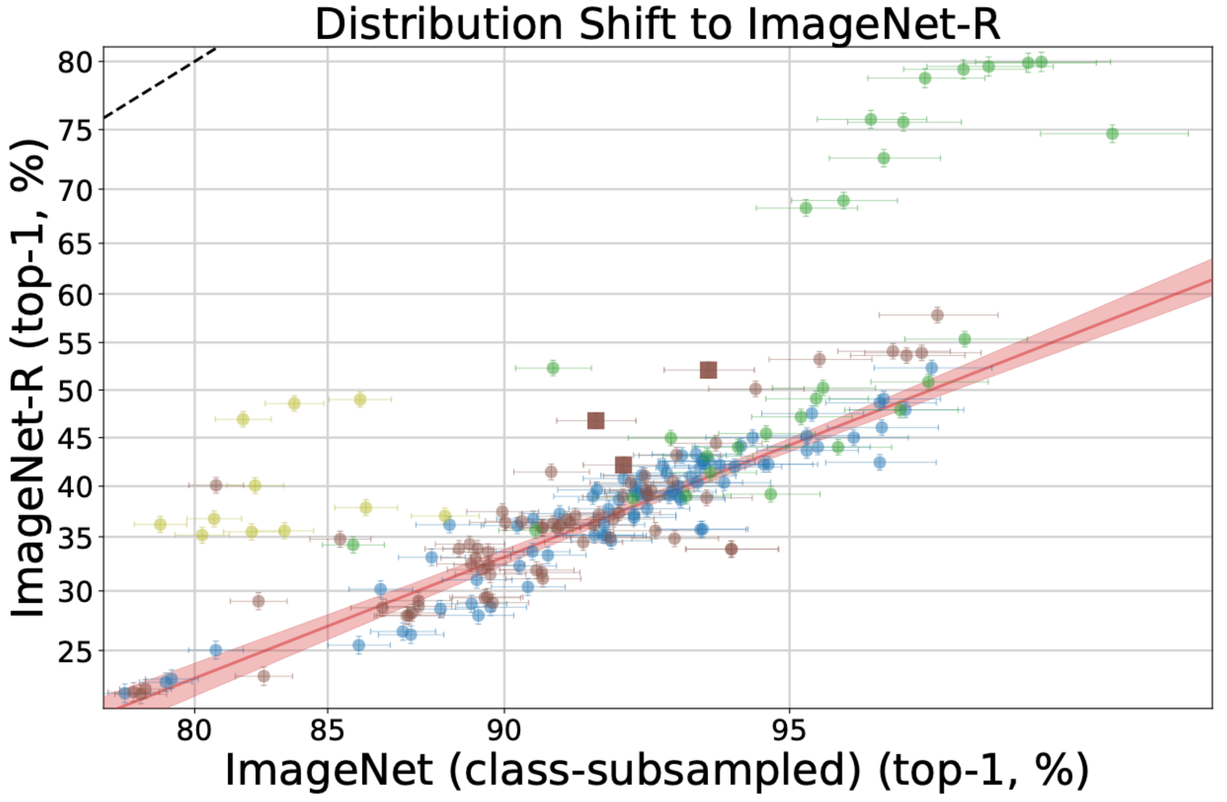
[Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz '19]

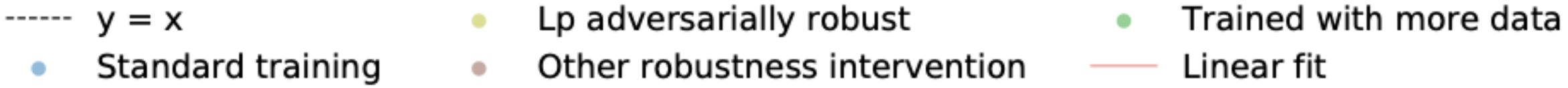


Same trend: only **more data** gives effective robustness.

[Taori, Dave, Shankar, Carlini, Recht, Schmidt '20]







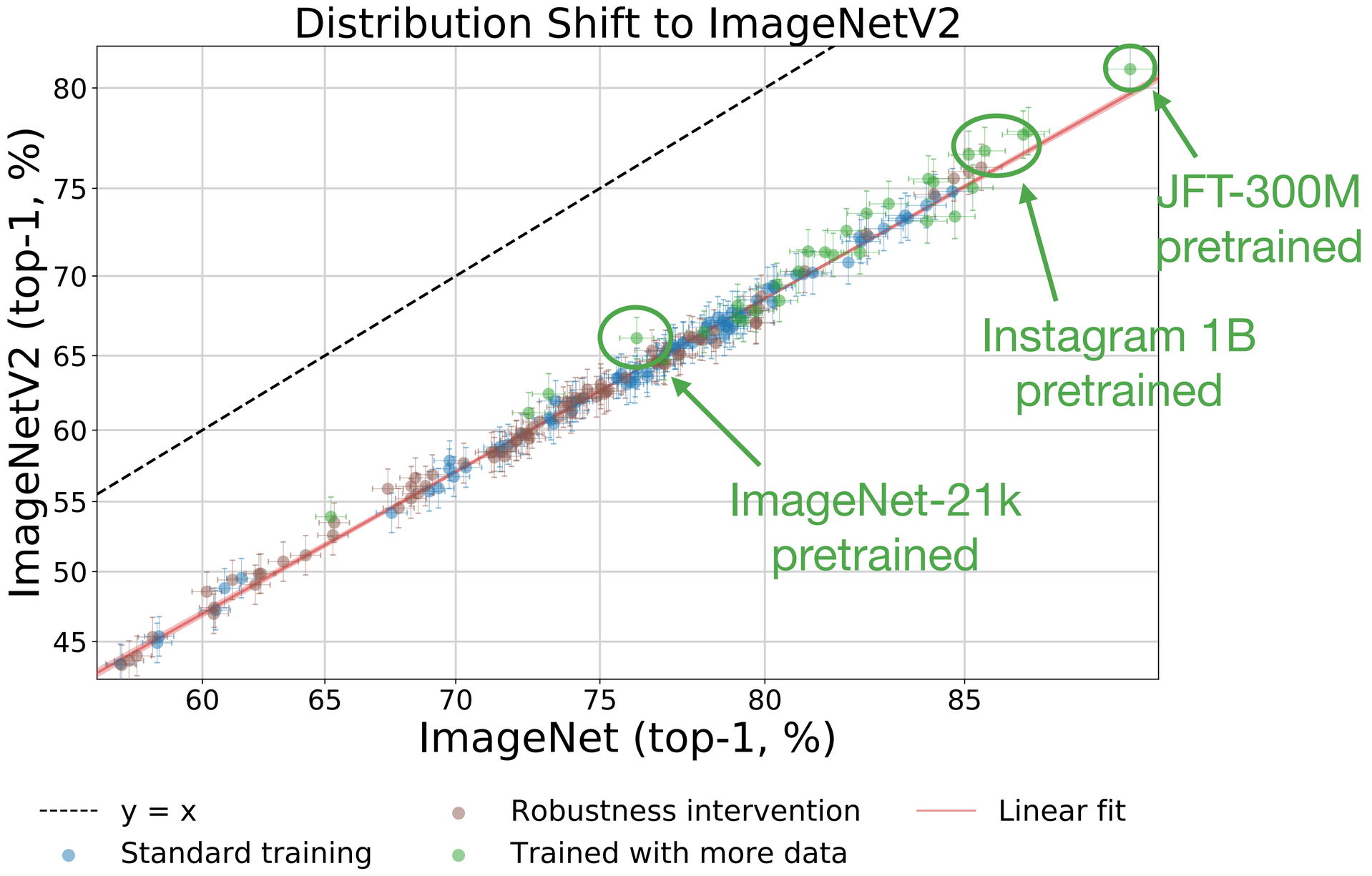
Some gains from **adv. training** and data augmentation. **More data** models still best.

Outline

1. Overview of the robustness landscape in computer vision

2. New image text-models (e.g., OpenAI's CLIP model) are (a lot) more robust

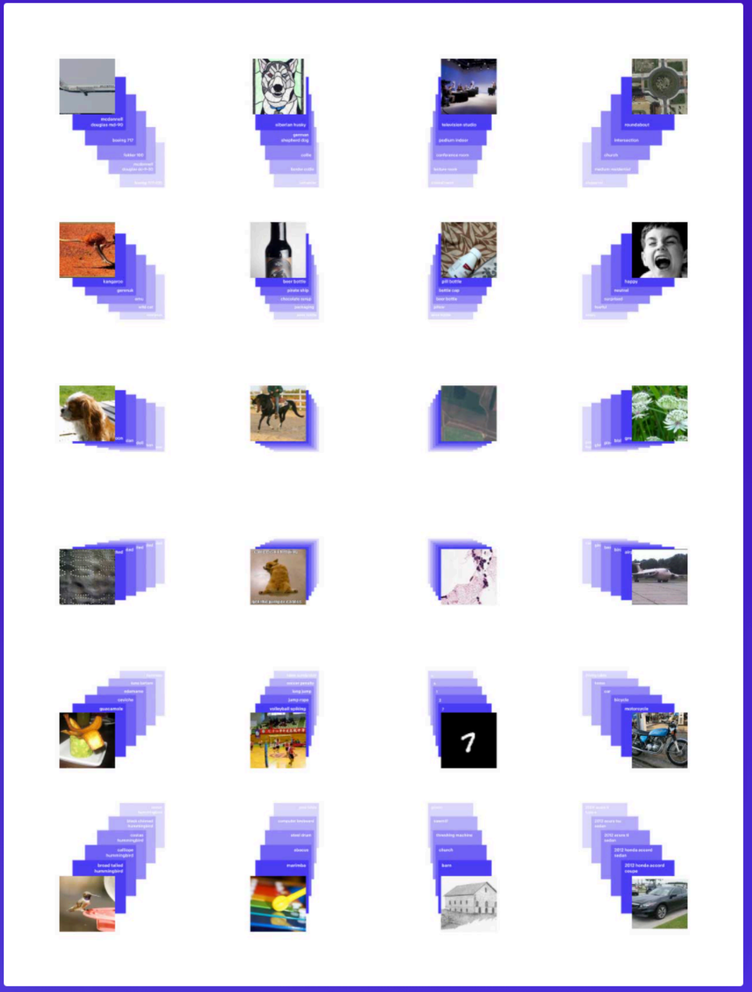
3. Where does CLIP's robustness come from? → Training data



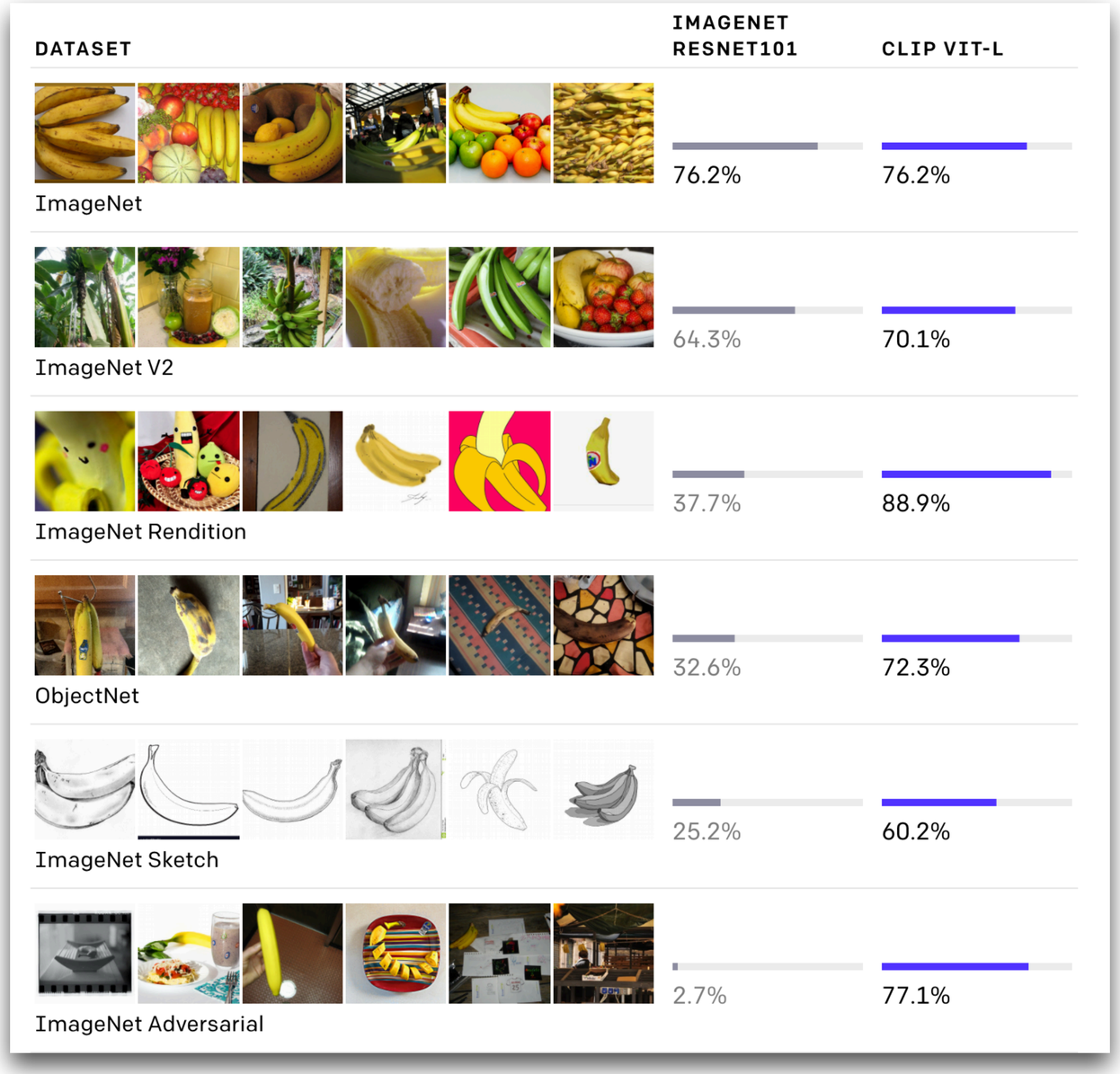
Training on (a lot) more data gives a **small** amount of effective robustness.

CLIP: Connecting Text and Images

We're introducing a neural network called CLIP which efficiently learns visual concepts from natural language supervision. CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the "zero-shot" capabilities of GPT-2 and GPT-3.



January 5, 2021
15 minute read



Effective robustness

+6%

+51%

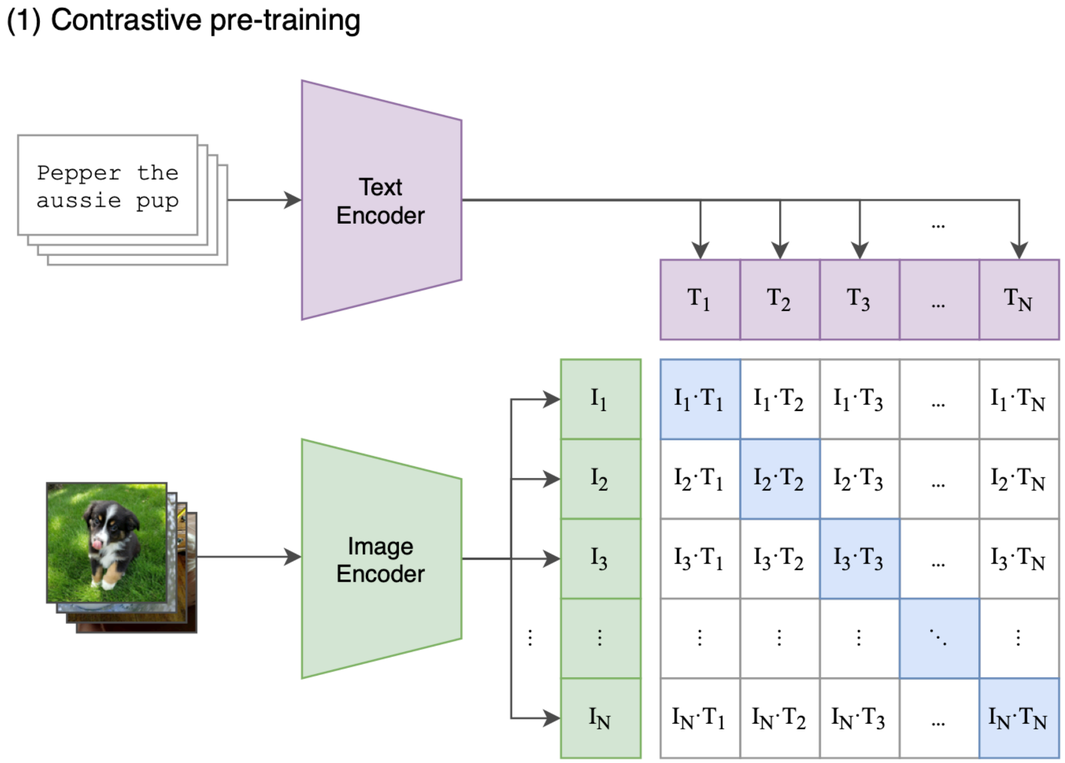
+40%

+35%

+74%

Very large improvements in out-of-distribution robustness.

CLIP is not (explicitly) designed for robustness

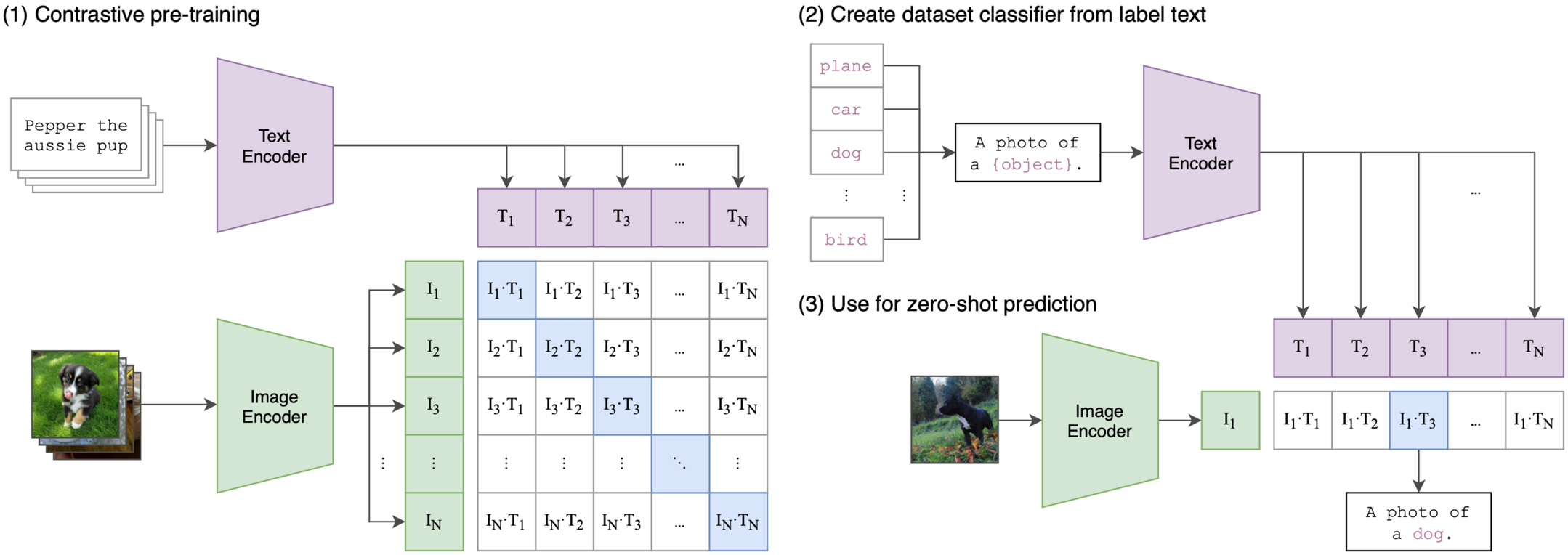


Training data: 400 million images collected from the web (dataset internal to OpenAI).

Compute: Trained on 250 - 600 GPUs for up to 18 days.

Model: ResNets and ViTs with up to 300M parameters.

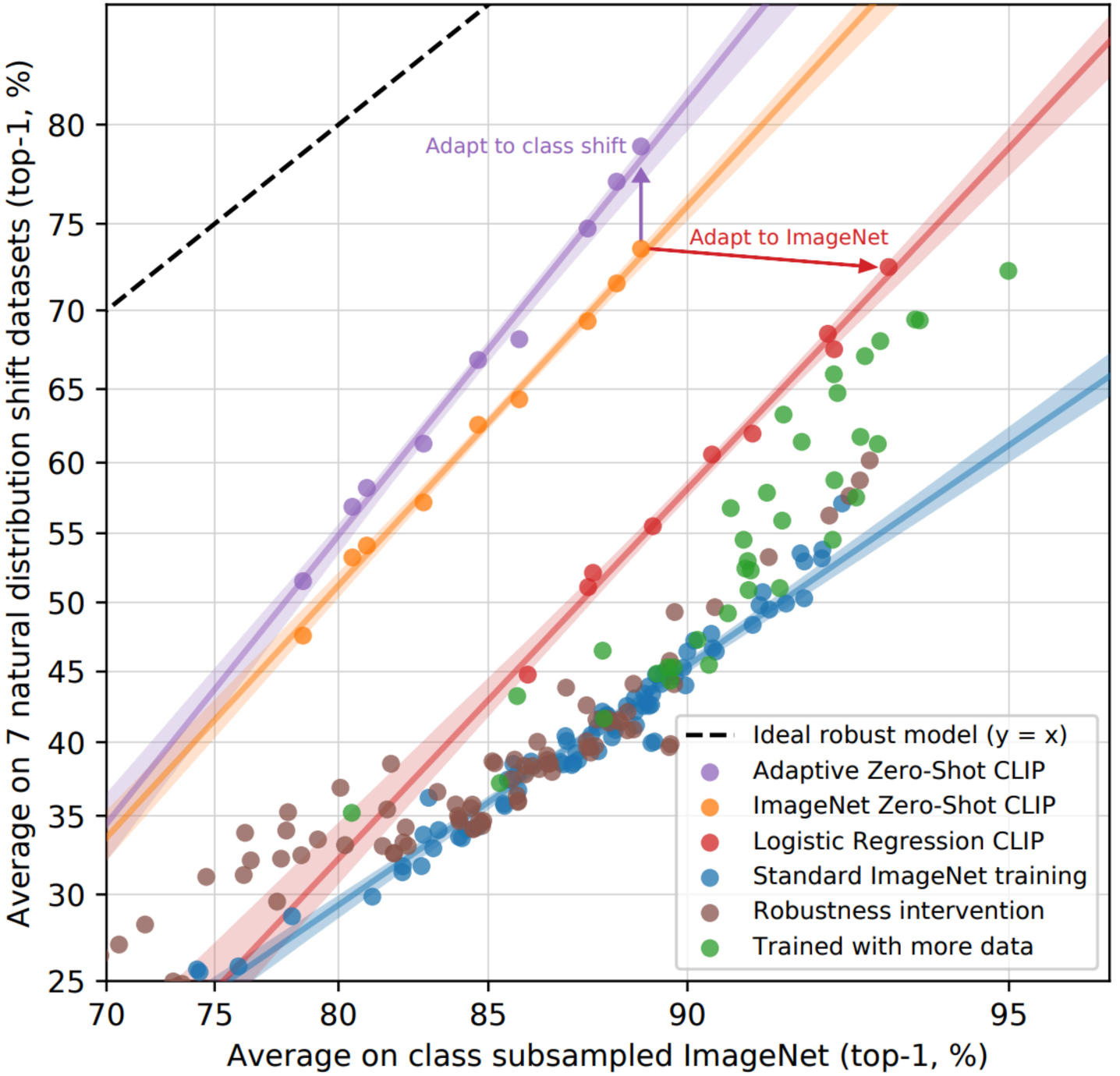
CLIP is not (explicitly) designed for robustness



Training data: 400 million images collected from the web (dataset internal to OpenAI).

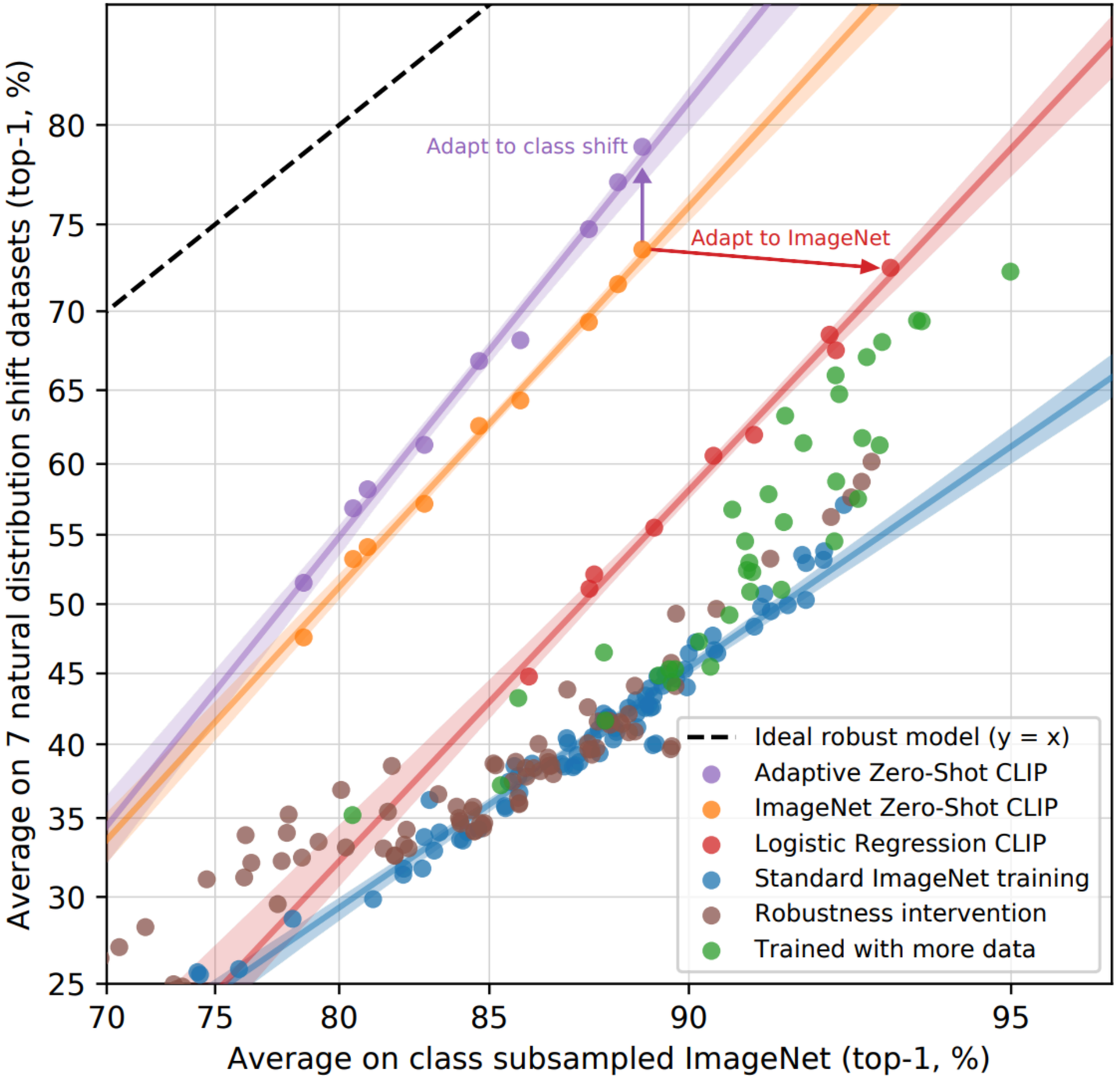
Compute: Trained on 250 - 600 GPUs for up to 18 days.

Model: ResNets and ViTs with up to 300M parameters.



[Radford, Kim, Hallacy, Ramesh, Goh, Agarwal, Sastry, Askell, Mishkin, Clark, Krueger, Sutskever '21]

Large robustness gains



[Radford, Kim, Hallacy, Ramesh, Goh, Agarwal, Sastry, Askell, Mishkin, Clark, Krueger, Sutskever '21]

Large robustness gains

→ Why makes CLIP robust?

But: fine-tuning reduces robustness

→ Can we get **both** high in-distribution **and** out-of-distribution accuracy?

Can we fine-tune CLIP without losing robustness?

Robust fine-tuning of zero-shot models

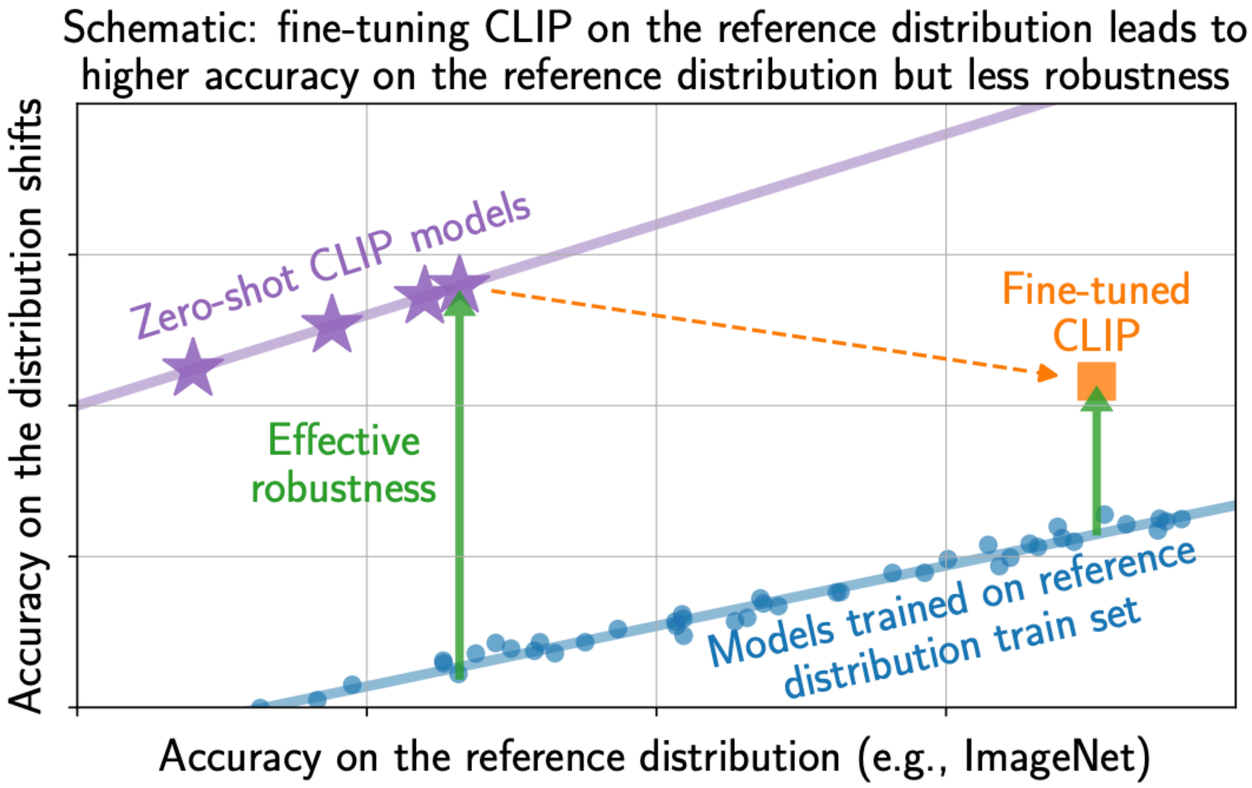
Mitchell Wortsman*† Gabriel Ilharco*† Jong Wook Kim§ Mike Li‡

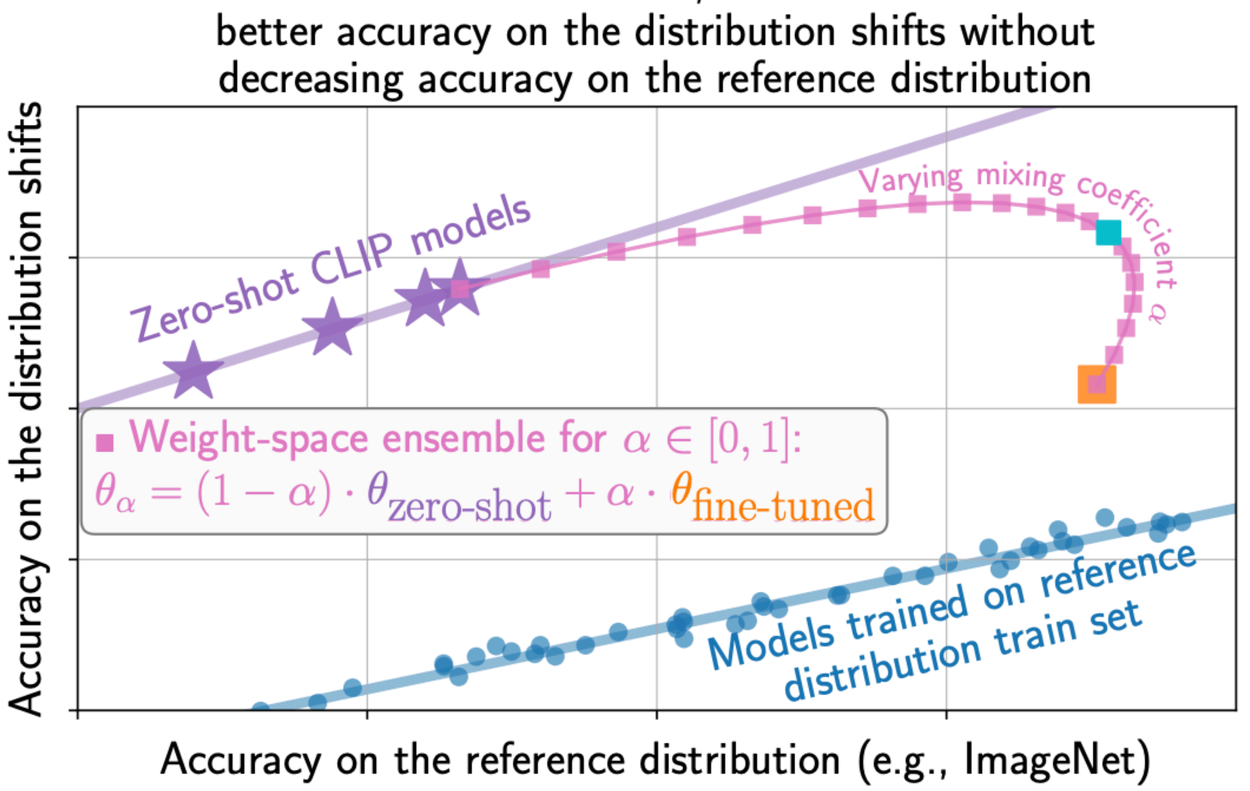
Simon Kornblith◇ Rebecca Roelofs◇ Raphael Gontijo-Lopes◇

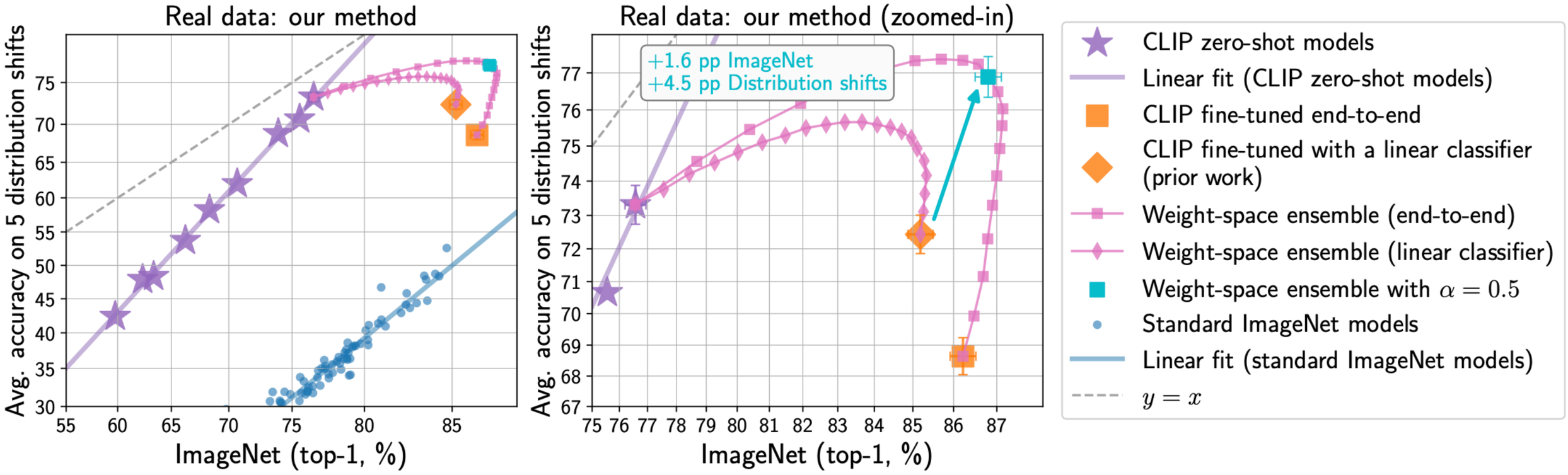
Hannaneh Hajishirzi†○ Ali Farhadi*† Hongseok Namkoong*‡ Ludwig Schmidt†△

Abstract

Large pre-trained models such as CLIP or ALIGN offer consistent accuracy across a range of data distributions when performing zero-shot inference (i.e., without fine-tuning on a specific dataset). Although existing fine-tuning methods substantially improve accuracy on a given target distribution, they often reduce robustness to distribution shifts. We address this tension by introducing a simple and effective method for improving robustness while fine-tuning: ensembling the weights of the zero-shot and fine-tuned models (WiSE-FT). Compared to standard fine-tuning, WiSE-FT provides large accuracy improvements under distribution shift, while preserving high accuracy on the target distribution. On ImageNet and five derived distribution shifts, WiSE-FT improves accuracy under distribution shift by 4 to 6 percentage points (pp) over prior work while increasing ImageNet accuracy by 1.6 pp. WiSE-FT achieves similarly large robustness gains (2 to 23 pp) on a diverse set of six further distribution shifts, and accuracy gains of 0.8 to 3.3 pp compared to standard fine-tuning on seven commonly used transfer learning datasets. These improvements come at no additional computational cost during fine-tuning or inference.


Schematic: fine-tuning CLIP on the reference distribution leads to higher accuracy on the reference distribution but less robustness
Accuracy on the distribution shifts
Zero-shot CLIP models
Fine-tuned CLIP
Effective robustness
Models trained on reference distribution train set
Accuracy on the reference distribution (e.g., ImageNet)


Schematic: our method, WiSE-FT leads to better accuracy on the distribution shifts without decreasing accuracy on the reference distribution
Accuracy on the distribution shifts
Zero-shot CLIP models
Varying mixing coefficient α
Weight-space ensemble for $\alpha \in [0, 1]$:
$\theta_\alpha = (1-\alpha) \cdot \theta_{\text{zero-shot}} + \alpha \cdot \theta_{\text{fine-tuned}}$
Models trained on reference distribution train set
Accuracy on the reference distribution (e.g., ImageNet)



5 - 9 percentage points improvements on ImageNet distribution shifts.

Gains on several more datasets, e.g., CIFAR-10, WILDS, etc. (also in-distribution)

Model interpolation also led to the state-of-the-art on ImageNet (“**model soups**”)

[Wortsman, Ilharco, Gadre, Roelofs, Gontijo-Lopes, Morcos, Namkoong, Farhadi, Carmon, Simon, Schmidt ’22]

Results on WILDS

iWildCam

Without unlabeled data

Rank	Algorithm	Model	Test ID Macro F1	Test ID Avg Acc	Test OOD Macro F1 ▼	Test OOD Avg Acc	Contact	References	Date
1	**Model Soups (CLIP ViT-L)**	ViT-L	57.6 (1.9) *	79.1 (0.4) *	43.3 (1.0) *	79.3 (0.3) *	Mitchell Wortsman	Paper / Code	March 12, 2022
2	**ERM (CLIP ViT-L)**	ViT-L	55.8 (1.9) *	77.0 (0.7) *	41.4 (0.5) *	78.3 (1.1) *	Mitchell Wortsman	Paper / Code	July 28, 2022
3	**ERM**	PNASNet-5-Large	52.8 (1.4) *	77.3 (0.7) *	38.5 (0.6) *	78.3 (1.4) *	John Miller	Paper / Code	July 20, 2021
4	CORAL	ResNet50	43.6 (3.3)	73.8 (0.3)	32.7 (0.2)	73.3 (4.3)	WILDS	Paper / Code	July 15, 2021
5	ERM w/ data aug	ResNet50	47.0 (1.4)	76.9 (0.6)	32.2 (1.2)	73.0 (0.4)	WILDS	Paper / Code	December 09, 20
6	ERM (more checkpoints)	ResNet50	47.9 (2.6)	76.2 (0.1)	32.0 (1.5)	69.0 (0.4)	Kazuki Irie	Paper / Code	February 10, 202
7	ERM (grid search)	ResNet50	47.1 (1.5)	75.7 (0.4)	30.8 (1.3)	71.5 (2.6)	WILDS	Paper / Code	July 15, 2021
8	ERM (rand search)	ResNet50	46.7 (0.6)	74.9 (1.2)	30.6 (1.1)	72.5 (3.2)	WILDS	Paper / Code	December 09, 20
9	**Group DRO**	ResNet50	37.5 (1.9)	71.6 (2.7)	23.8 (2.0)	72.7 (2.0)	WILDS	Paper / Code	July 15, 2021
10	**ARM-BN**	ResNet50	27.5 (5.4)	62.0 (4.0)	23.3 (2.8)	70.2 (2.4)	Marvin Zhang	Paper / Code	April 19, 2022
11	**Fish**	ResNet50	40.3 (0.6)	73.8 (0.1)	22.0 (1.8)	64.7 (2.6)	Yuge Shi	Paper / Code	December 14, 20
12	IRM	ResNet50	22.4 (7.7)	59.8 (8.2)	15.1 (4.9)	59.7 (3.8)	WILDS	Paper / Code	July 15, 2021
13	MixUp	ResNet50	31.2 (3.1)	66.1 (1.8)	13.8 (0.8)	48.6 (1.1)	Olivia Wiles	Paper / Code	June 16, 2022
14	Test-time BN adaptation	ResNet50	12.0 (0.3)	37.2 (0.7)	13.8 (0.6)	46.6 (0.9)	Marvin Zhang	Paper / Code	April 19, 2022
15	JTT	ResNet50	32.6 (4.4)	64.9 (2.8)	11.0 (2.5)	47.4 (2.2)	Olivia Wiles	Paper / Code	June 16, 2022
16	**ABSGD**	ResNet50	3.1 (0.9)	25.1 (1.1)	2.6 (0.2)	36.0 (1.4)	Qi Qi	Paper / Code	June 29, 2022

Results on WILDS

FMoW

Without unlabeled data

Rank	Algorithm	Model	Val Avg Acc	Test Avg Acc	Val Worst-region Acc	Test Worst-region Acc ▼	Contact	References	Date
1	**Model Soups (CLIP ViT-L)**	ViT-L	75.7 (0.07) *	69.5 (0.08) *	59.8 (0.43) *	47.6 (0.33) *	Mitchell Wortsman	Paper / Code	March 12, 202
2	**ERM (CLIP ViT-L)**	ViT-L	73.6 (0.23) *	66.9 (0.17) *	59.5 (1.31) *	46.1 (0.59) *	Mitchell Wortsman	Paper / Code	July 28, 2022
3	ERM w/ data aug	DenseNet121	62.1 (0.23)	55.5 (0.42)	53.2 (0.61)	35.7 (0.26)	WILDS	Paper / Code	December 09, 2
4	**LISA**	DenseNet121	58.7 (1.12)	52.8 (1.15)	48.7 (0.92)	35.5 (0.81)	Yu Wang	Paper / Code	March 13, 2022
5	**ERM**	se_resnext101_32x4d	62.1 (0.24) *	55.5 (0.14) *	51.3 (2.93) *	35.0 (0.78) *	John Miller	Paper / Code	July 15, 2021
6	ERM (more checkpoints)	DenseNet121	62.0 (0.06)	55.6 (0.23)	52.5 (1.25)	34.8 (1.9)	Kazuki Irie	Paper / Code	February 10, 20
7	**Fish**	DenseNet121	57.8 (0.15)	51.8 (0.32)	49.5 (2.34)	34.6 (0.18)	Yuge Shi	Paper / Code	December 14, 2
8	ERM (rand search)	DenseNet121	60.6 (0.57)	54.0 (0.4)	52.6 (0.25)	34.1 (1.42)	WILDS	Paper / Code	December 09, 2
9	IRM	DenseNet121	56.1 (0.61)	50.4 (0.75)	49.7 (0.97)	32.8 (2.09)	WILDS	Paper / Code	July 15, 2021
10	CORAL	DenseNet121	56.5 (0.15)	50.1 (0.07)	48.9 (1.31)	32.8 (0.66)	WILDS	Paper / Code	July 15, 2021
11	**CGD**	DenseNet121	57.0 (1.03)	50.6 (1.39)	49.8 (1.04)	32.0 (2.26)	Vihari Piratla	Paper / Code	April 10, 2022
12	ERM (grid search)	DenseNet121	59.2 (0.07)	52.7 (0.23)	49.8 (0.36)	31.3 (0.17)	WILDS	Paper / Code	July 15, 2021
13	**Group DRO**	DenseNet121	57.6 (0.7)	51.2 (0.38)	49.4 (0.45)	31.1 (1.66)	WILDS	Paper / Code	July 15, 2021
14	Test-time BN adaptation	DenseNet121	57.9 (0.36)	51.5 (0.25)	47.8 (0.52)	30.0 (0.23)	Marvin Zhang	Paper / Code	April 19, 2022
15	**ARM-BN**	DenseNet121	48.0 (0.65)	42.1 (0.26)	38.9 (2.17)	24.4 (0.54)	Marvin Zhang	Paper / Code	April 19, 2022

Outline

1. Overview of the robustness landscape in computer vision

2. New image text-models (e.g., OpenAI's CLIP model) are (a lot) more robust

3. Where does CLIP's robustness come from? → Training data

Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP)

Alex Fang[†] Gabriel Ilharco[†] Mitchell Wortsman[†] Yuhao Wan[†]

Vaishaal Shankar[◇] Achal Dave[◇] Ludwig Schmidt[†◦]

Abstract

Contrastively trained image-text models such as CLIP, ALIGN, and BASIC have demonstrated unprecedented robustness to multiple challenging natural distribution shifts. Since these image-text models differ from previous training approaches in several ways, an important question is what causes the large robustness gains. We answer this question via a systematic experimental investigation. Concretely, we study five different possible causes for the robustness gains: (i) the training set size, (ii) the training distribution, (iii) language supervision at training time, (iv) language supervision at test time, and (v) the contrastive loss function. Our experiments show that the more diverse training distribution is the main cause for the robustness gains, with the other factors contributing little to no robustness. Beyond our experimental results, we also introduce ImageNet-Captions, a version of ImageNet with original text annotations from Flickr, to enable further controlled experiments of language-image training.

1 Introduction

Large pre-trained language-image models such as CLIP [27], ALIGN [21], and BASIC [26] have recently demonstrated unprecedented robustness on a variety of natural distribution shifts. In contrast to prior models that are trained on images with class annotations, CLIP and relatives[1] are directly trained on images and their corresponding unstructured text from the web. The resulting models achieve large robustness even on challenging distribution shifts such as ImageNetV2 [28] and ObjectNet [2]. No prior algorithmic techniques had enhanced robustness on these datasets even after multiple years of intensive research in reliable machine learning [13, 35]. As CLIP also improves robustness on a wide range of other distribution shifts, an important question emerges: *What causes CLIP's unprecedented robustness?*

Hypotheses for CLIP's robustness

	CLIP	Standard ImageNet supervised learning
Language supervision	Yes	No

Hypotheses for CLIP's robustness

	CLIP	Standard ImageNet supervised learning
Language supervision	Yes	No
Training distribution	???	ImageNet

Hypotheses for CLIP's robustness

	CLIP	Standard ImageNet supervised learning
Language supervision	Yes	No
Training distribution	???	ImageNet
Training set size	400M	1.2M

Hypotheses for CLIP's robustness

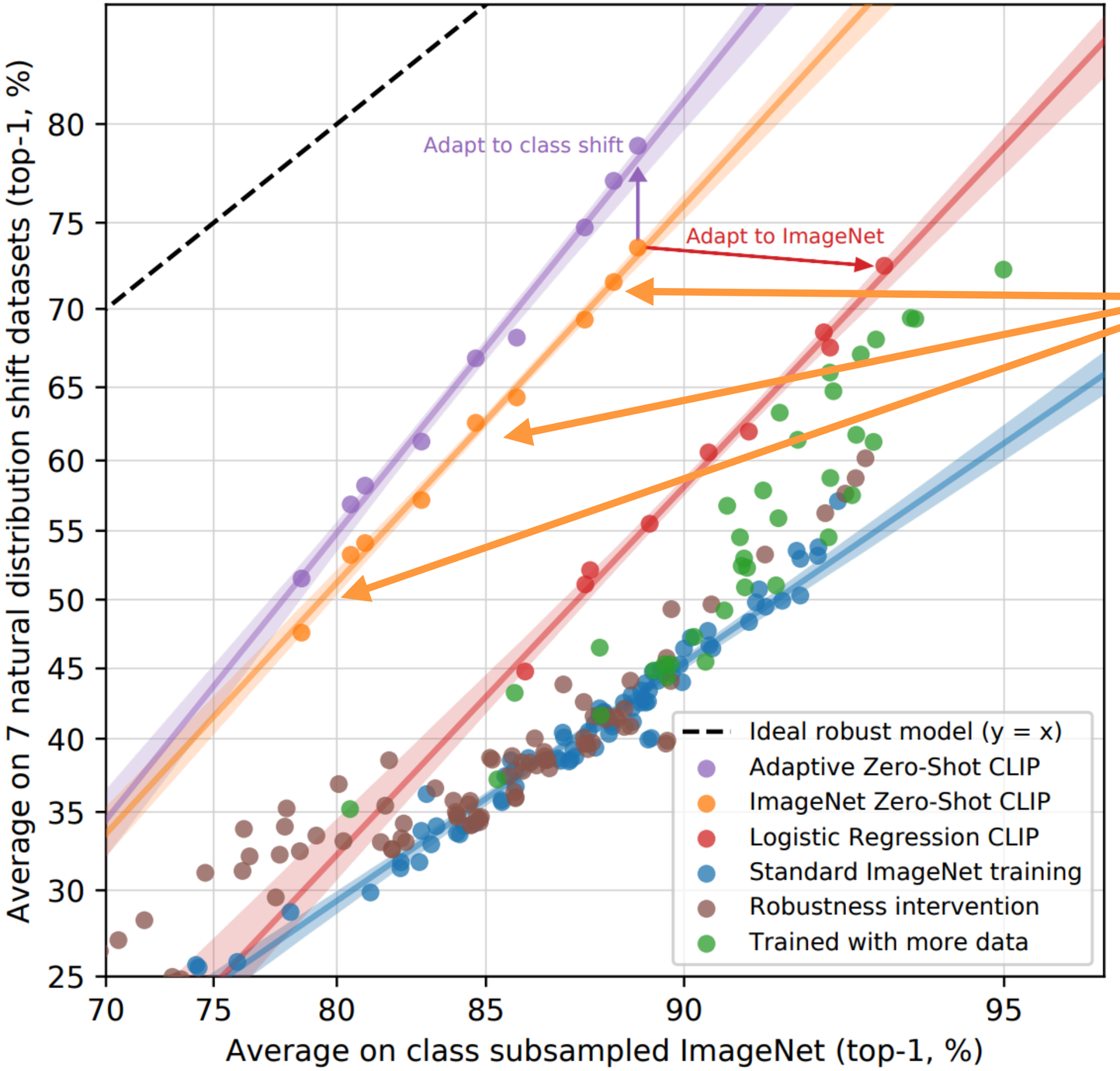
	CLIP	Standard ImageNet supervised learning
Language supervision	Yes	No
Training distribution	???	ImageNet
Training set size	400M	1.2M
Loss function	Contrastive	Supervised

Hypotheses for CLIP's robustness

	CLIP	Standard ImageNet supervised learning
Language supervision	Yes	No
Training distribution	???	ImageNet
Training set size	400M	1.2M
Loss function	Contrastive	Supervised
Test-time prompting	Yes	No

Hypotheses for CLIP's robustness

	CLIP	Standard ImageNet supervised learning
Language supervision	Yes	No
Training distribution	???	ImageNet
Training set size	400M	1.2M
Loss function	Contrastive	Supervised
Test-time prompting	Yes	No
Model architecture	ViTs	CNNs



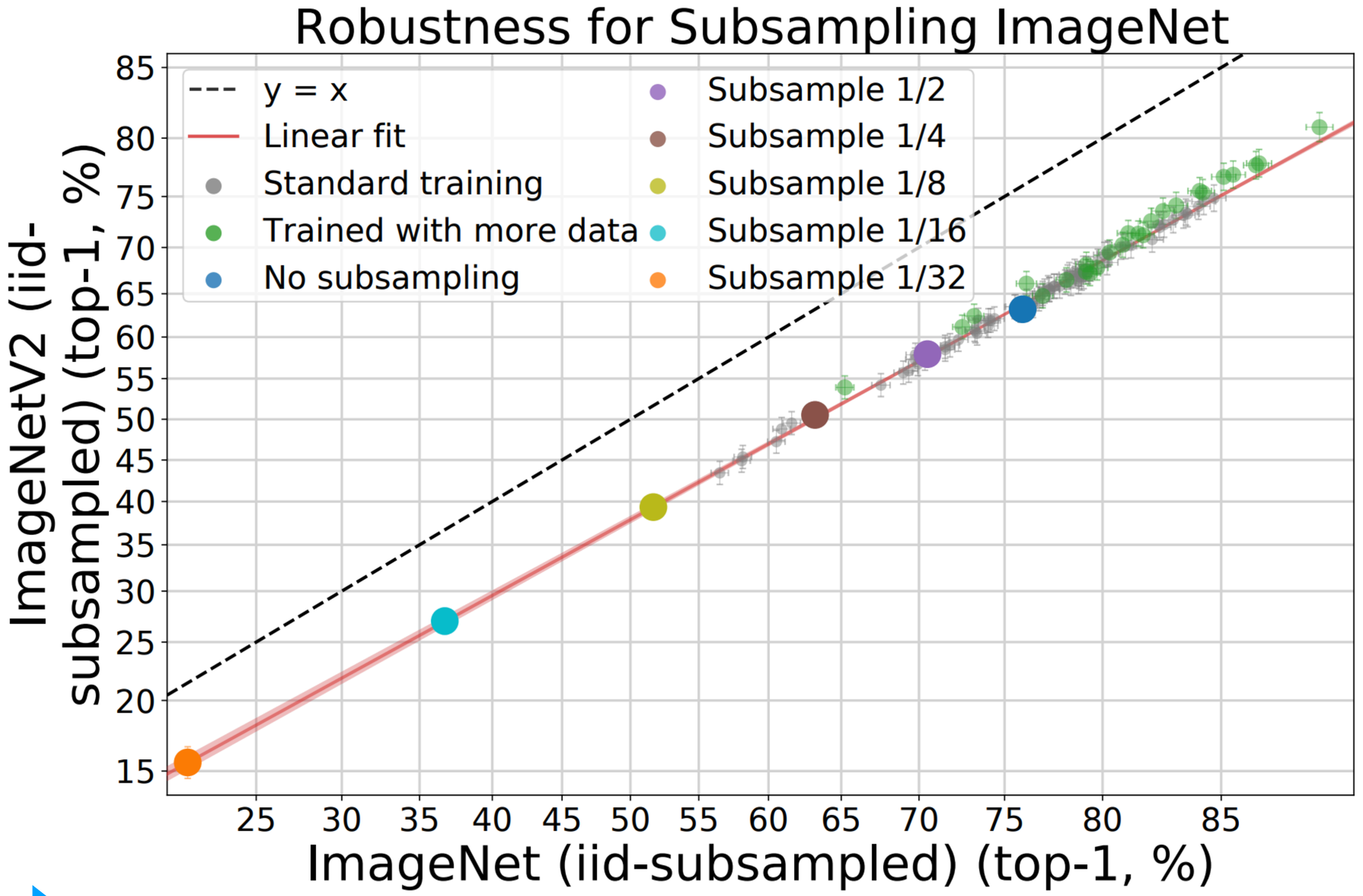
[Radford, Kim, Hallacy, Ramesh, Goh, Agarwal, Sastry, Askell, Mishkin, Clark, Krueger, Sutskever '21]

ViTs and CNNs are on the same linear trend.

→ Architecture does not change effective robustness.

Hypotheses for CLIP's robustness

	CLIP	Standard ImageNet supervised learning
Language supervision	Yes	No
Training distribution	???	ImageNet
Training set size	400M	1.2M
Loss function	Contrastive	Supervised
Test-time prompting	Yes	No
~~**Model architecture**~~	~~ViTs~~	~~CNNs~~



Training set size does not change effective robustness.

[Taori, Dave, Shankar, Carlini, Recht, Schmidt '20]

Hypotheses for CLIP's robustness

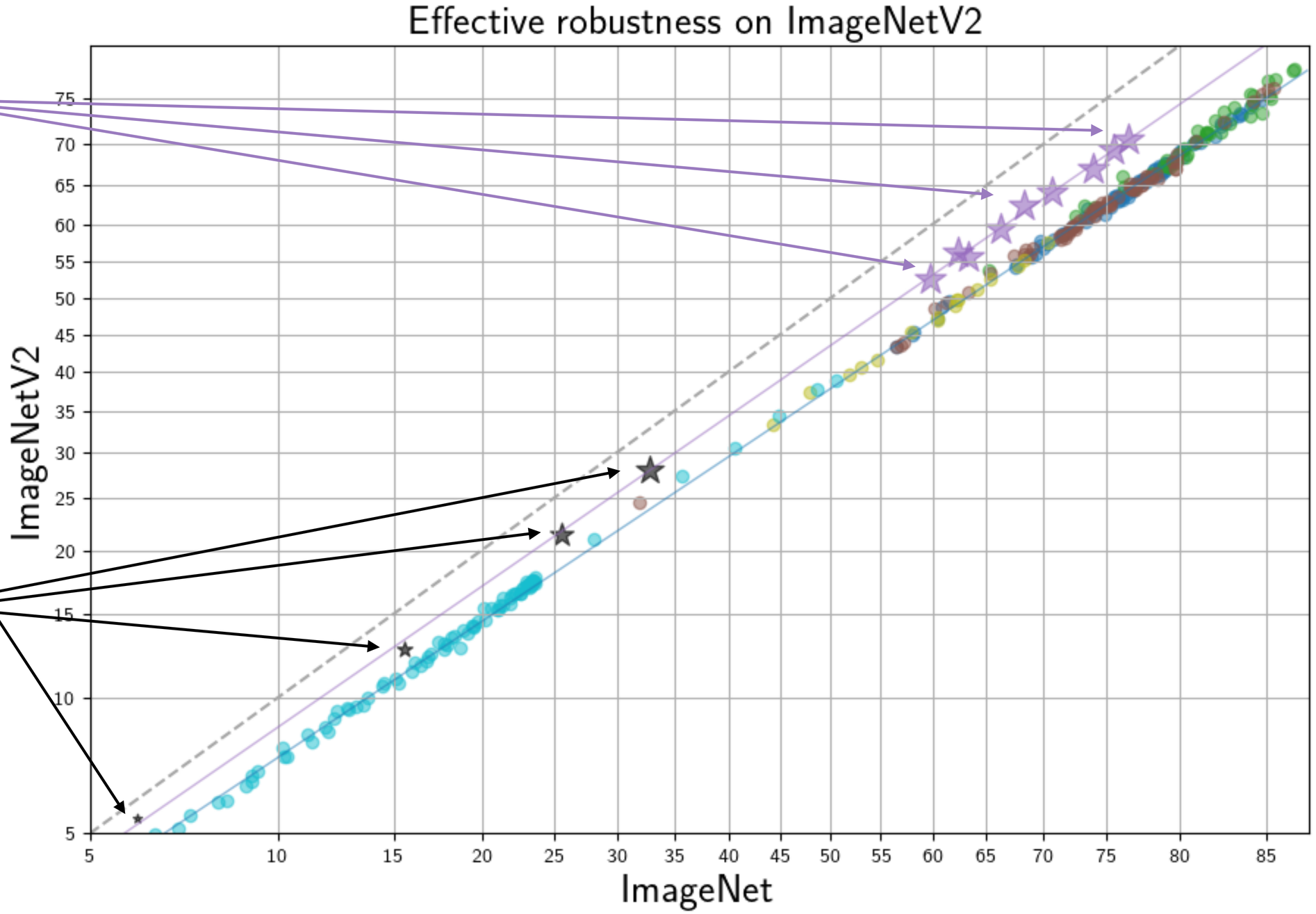
	CLIP	Standard ImageNet supervised learning
Language supervision	Yes	No
Training distribution	???	ImageNet
~~**Training set size**~~	~~400M~~	~~1.2M~~
Loss function	Contrastive	Supervised
Test-time prompting	Yes	No
~~**Model architecture**~~	~~ViTs~~	~~CNNs~~

Background: YFCC-15M

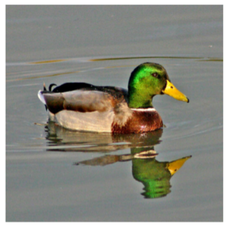
- A dataset of images with corresponding captions
- Subset of the 100M-size YFCC (Yahoo Flickr Creative Commons) dataset
- The 15M subset was released by OpenAI as a training set representative of their internal training set
- Part of OpenAI's 400M training set for CLIP

OpenAI's CLIP models (400M training data)

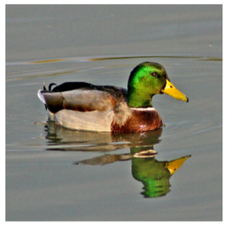
CLIP models trained on YFCC with OpenCLIP 1, 3, 7, 15M training data



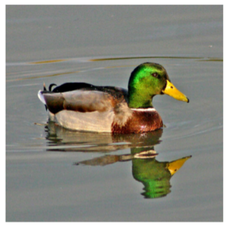
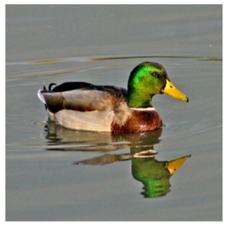
Experimental Setup

	Supervised (Without Language)	Contrastive (With Language)
ImageNet	ImageNet Standard (Baseline) 	
YFCC		YFCC-15M CLIP (Baseline) 

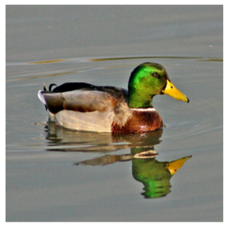
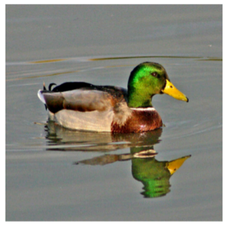
Experimental Setup

	Supervised (Without Language)	Contrastive (With Language)
ImageNet	ImageNet Standard (Baseline) 	ImageNet-Captions + CLIP (Us)  Title: `Reflected Duck` Description: Tags: `lake`, `water`, `bird` [6 tags omitted]
YFCC		YFCC-15M CLIP (Baseline)  one of the most dramatic mountain ranges I have seen

Experimental Setup

	Supervised (Without Language)	Contrastive (With Language)
ImageNet	ImageNet Standard (Baseline) 	ImageNet-Captions + CLIP (Us)  Title: `Reflected Duck` Description: Tags: `lake`, `water`, `bird` [6 tags omitted]
YFCC	YFCC Hardmatch + NoCLIP (Us) 	YFCC-15M CLIP (Baseline)  one of the most dramatic mountain ranges I have seen

Experimental Setup

	Supervised (Without Language)	Contrastive (With Language)
ImageNet	ImageNet Standard (Baseline) 	ImageNet-Captions + CLIP (Us)  Title: `Reflected Duck` Description: Tags: `lake`, `water`, `bird` [6 tags omitted]
YFCC	YFCC Hardmatch + NoCLIP (Us) 	YFCC-15M CLIP (Baseline)  one of the most dramatic mountain ranges I have seen

ImageNet-Captions

Ideal comparison: Train models on ImageNet with and without text annotations

Controlled experiment on a well-understood dataset

ImageNet-Captions

Ideal comparison: Train models on ImageNet with and without text annotations

➡ Controlled experiment on a well-understood dataset

ImageNet-Captions

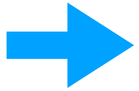
Ideal comparison: Train models on ImageNet with and without text annotations

➡ Controlled experiment on a well-understood dataset

BUT: ImageNet does not have any text annotations.

ImageNet-Captions

Ideal comparison: Train models on ImageNet with and without text annotations



Controlled experiment on a well-understood dataset

BUT: ImageNet does not have any text annotations.

Possible options:

- Templates like “A photo of a German shepherd”
- Image captioning models
- Collect new annotations from humans
- Get original text annotations - if they exist

ImageNet-Captions

Ideal comparison: Train models on ImageNet with and without text annotations

➡ Controlled experiment on a well-understood dataset

BUT: ImageNet does not have any text annotations.

Possible options:

- Templates like “A photo of a German shepherd”
- Image captioning models
- Collect new annotations from humans
- Get original text annotations - if they exist

We connected **460k** ImageNet images back to Flickr and retrieved the original captions.

➡ New image-text dataset “ImageNet-Captions”

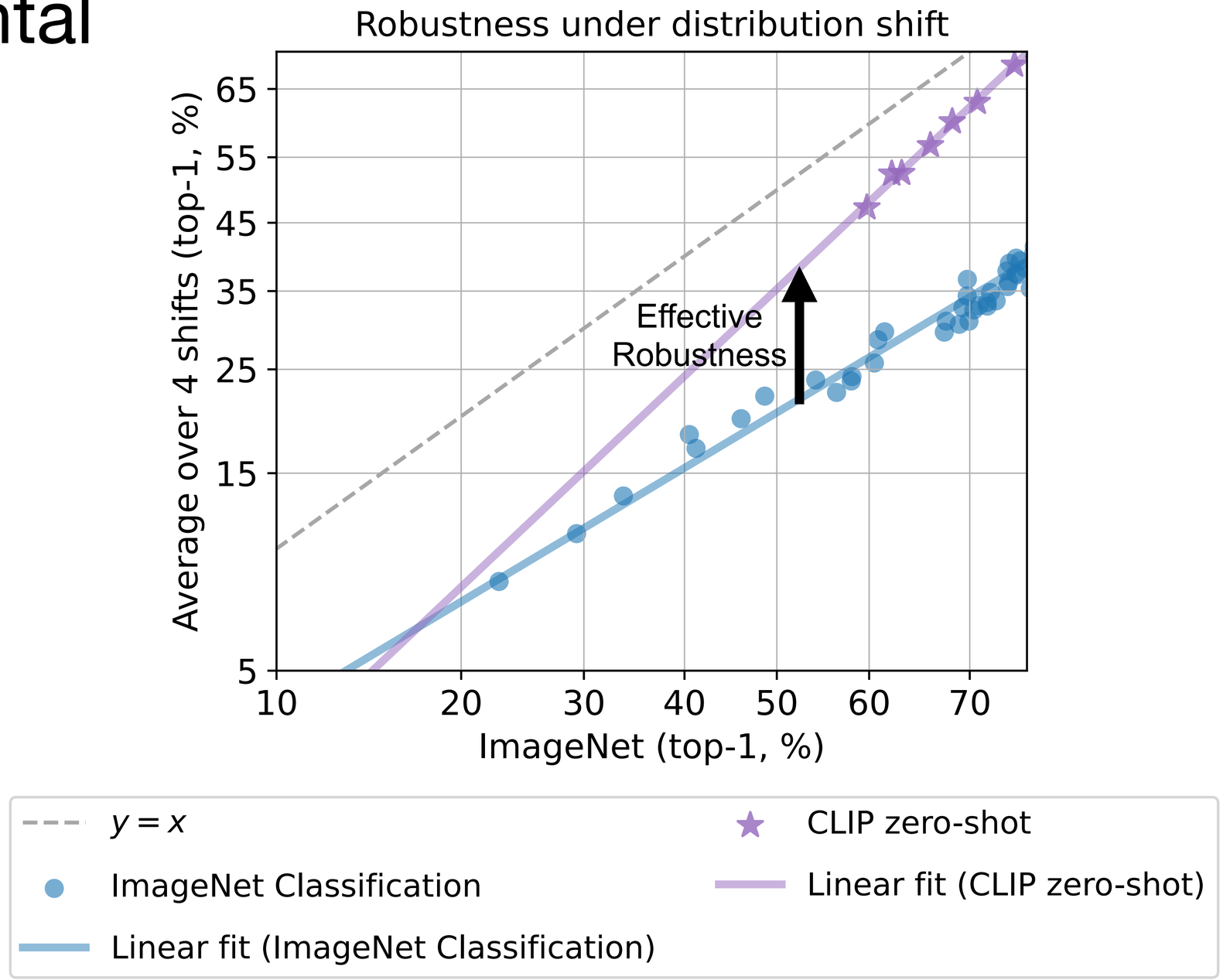
ImageNet-Captions examples

Title: Reflected Duck
Description:
Tags: lake, water, bird [6 tags omitted]

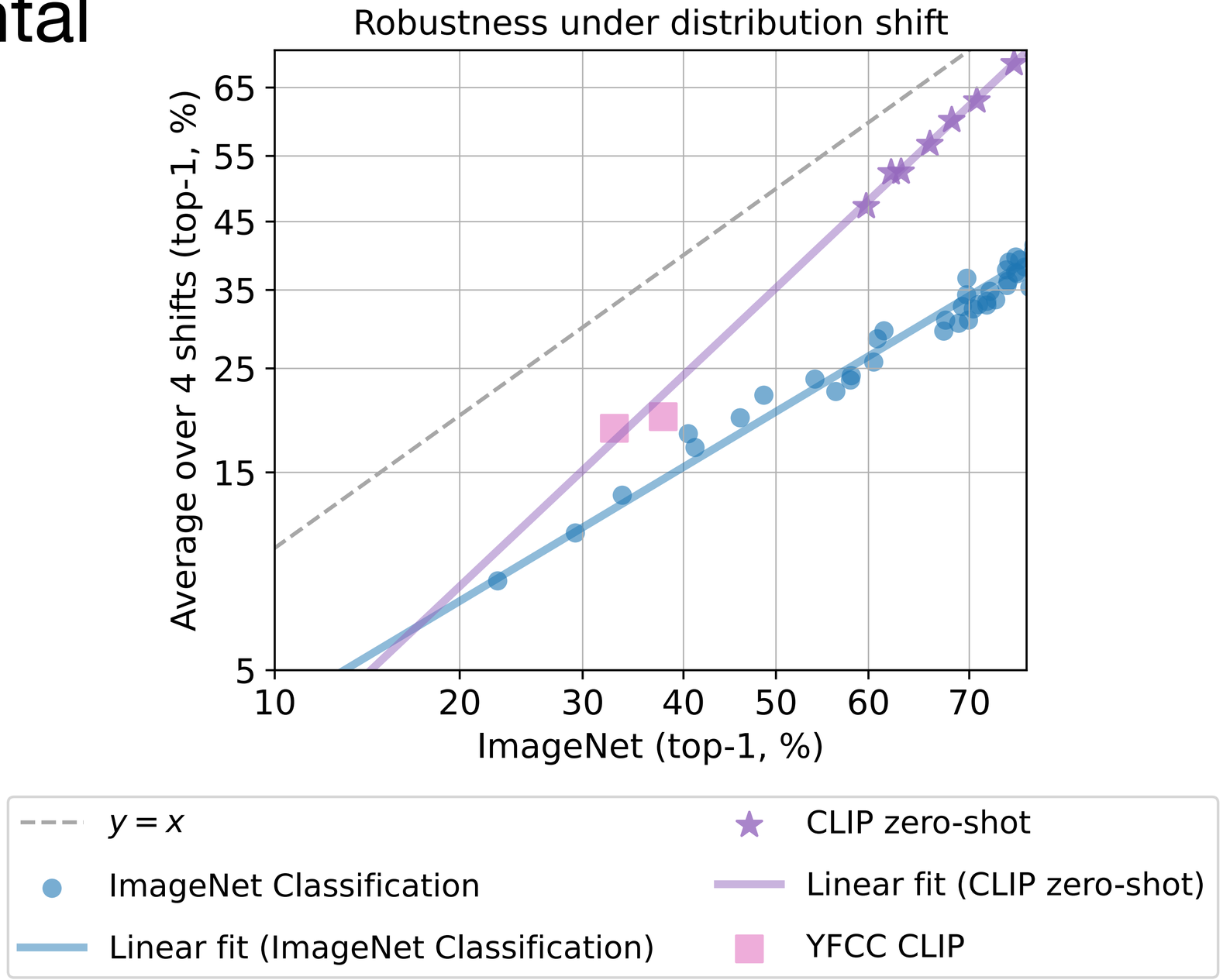
Title: SILENT ROCKER
Description: MOSE'S MOTHER HAS LEFT THE BuILDING [10 words omitted]
Tags: rockingchair, rock, chair [2 tags omitted]

Title: A Phone Call at Night
Description: I might have a thing with telephones [174 words omitted]
Tags: phone, telephone, blackandwhite [7 tags omitted]

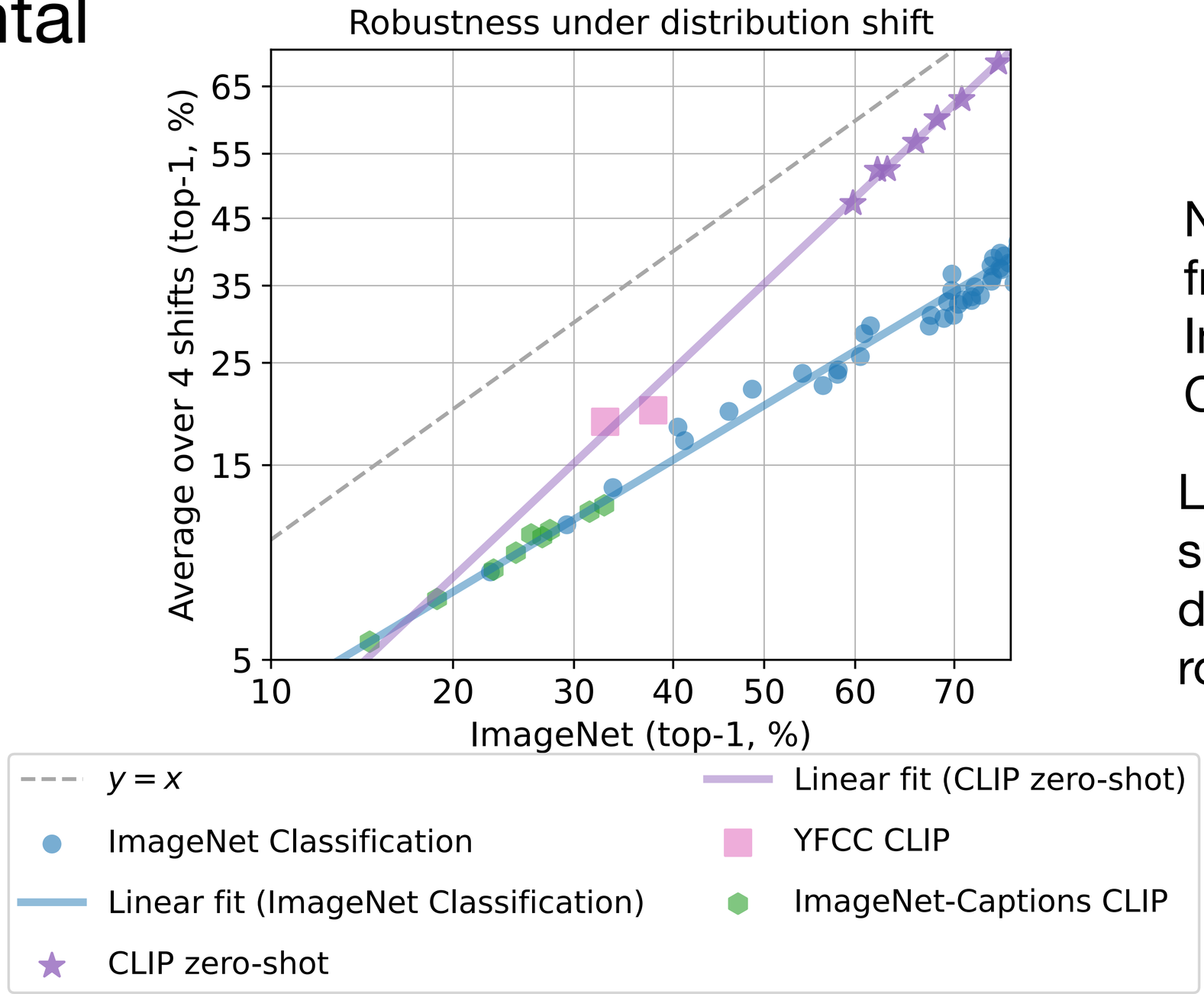
Experimental results



Experimental results



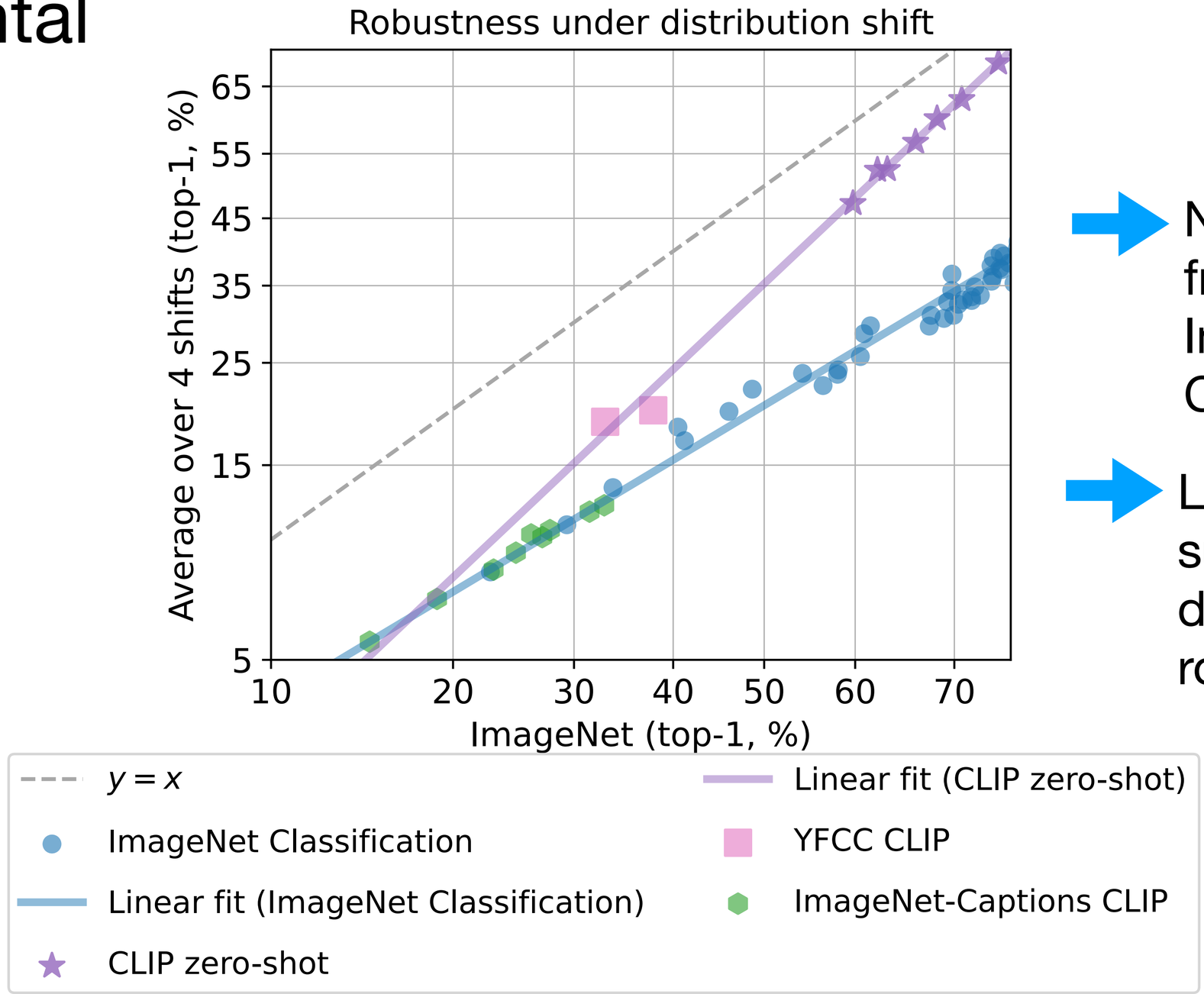
Experimental results



No robustness from CLIP on ImageNet-Captions.

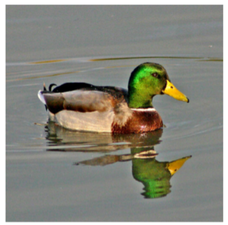
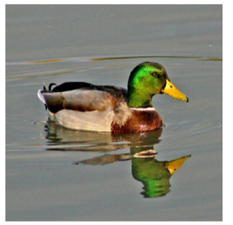
Language supervision alone does not promote robustness.

Experimental results



- No robustness from CLIP on ImageNet-Captions.
- Language supervision alone does not promote robustness.

Experimental Setup

	Supervised (Without Language)	Contrastive (With Language)
ImageNet	ImageNet Standard (Baseline) 	ImageNet-Captions + CLIP (Us)  Title: `Reflected Duck` Description: Tags: `lake`, `water`, `bird` [6 tags omitted]
YFCC	YFCC Hardmatch + NoCLIP (Us) 	YFCC-15M CLIP (Baseline)  one of the most dramatic mountain ranges I have seen

YFCC classification with minimal language

Recall: CLIP training on YFCC-15M resulted in a more robust model.

➡ Can we get the same robustness gains without language supervision?

YFCC classification with minimal language

Recall: CLIP training on YFCC-15M resulted in a more robust model.

➡ Can we get the same robustness gains without language supervision?

Yes, via the following training process **(“NoCLIP”)**:

YFCC classification with minimal language

Recall: CLIP training on YFCC-15M resulted in a more robust model.

➡ Can we get the same robustness gains without language supervision?

Yes, via the following training process **(“NoCLIP”)**:

1. Train a representation on **only the images** of YFCC-15M with SimCLR.

YFCC classification with minimal language

Recall: CLIP training on YFCC-15M resulted in a more robust model.

➡ Can we get the same robustness gains without language supervision?

Yes, via the following training process **(“NoCLIP”)**:

1. Train a representation on **only the images** of YFCC-15M with SimCLR.

   Problem: can’t use the representation as a classifier.

YFCC classification with minimal language

Recall: CLIP training on YFCC-15M resulted in a more robust model.

➡ Can we get the same robustness gains without language supervision?

Yes, via the following training process **(“NoCLIP”)**:

1. Train a representation on **only the images** of YFCC-15M with SimCLR.

   Problem: can’t use the representation as a classifier.

2. Create a subset of YFCC-15M via **substring matches** with the ImageNet class names.

   ➡ We call this subset YFCC-hardmatch.

   ➡ This is the only language-dependent step in the training process.

YFCC classification with minimal language

Recall: CLIP training on YFCC-15M resulted in a more robust model.

➡ Can we get the same robustness gains without language supervision?

Yes, via the following training process **(“NoCLIP”)**:

1. Train a representation on **only the images** of YFCC-15M with SimCLR.

   Problem: can’t use the representation as a classifier.

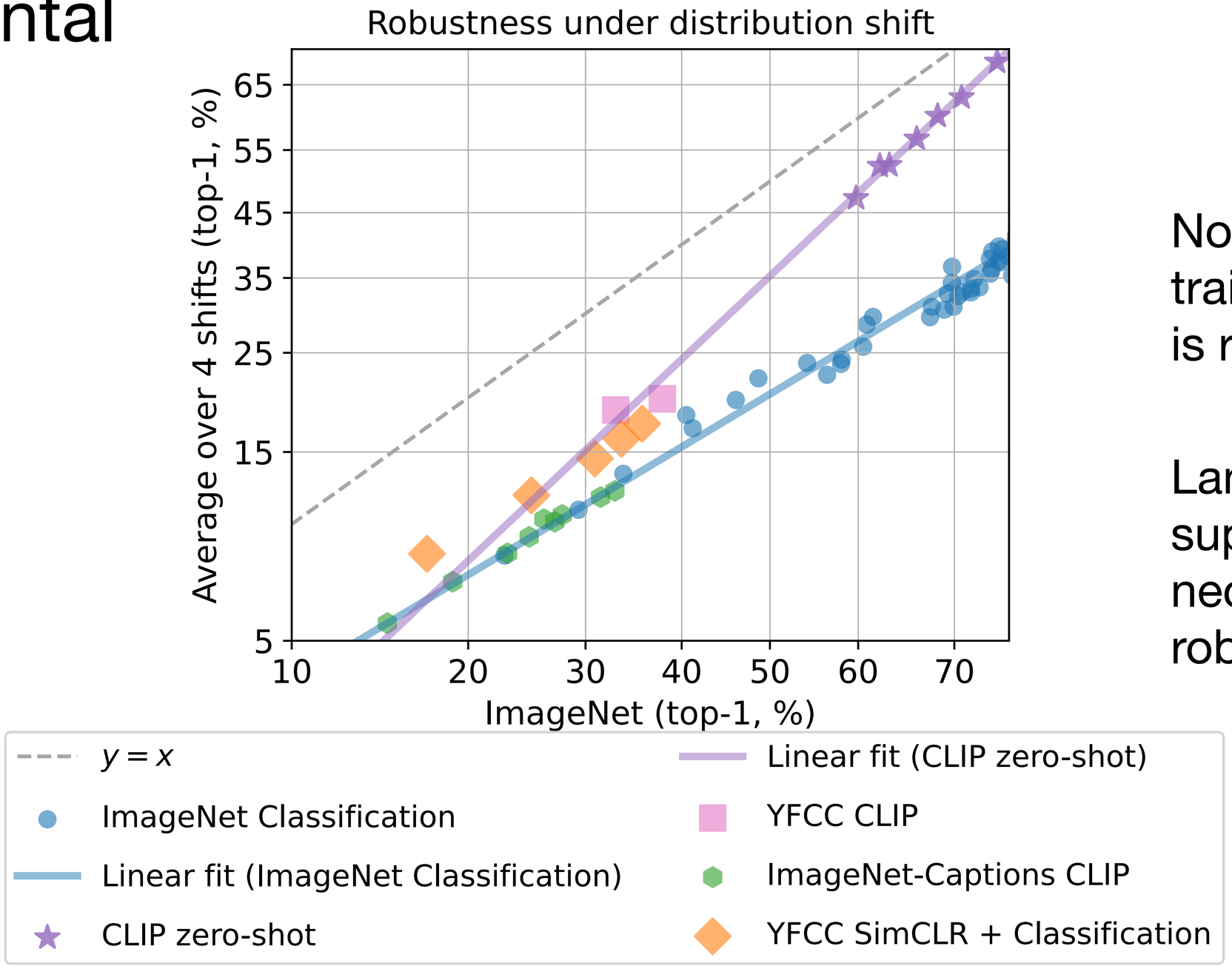
2. Create a subset of YFCC-15M via **substring matches** with the ImageNet class names.

   ➡ We call this subset YFCC-hardmatch.

   ➡ This is the only language-dependent step in the training process.

3. **Fine-tune** the SimCLR representation on YFCC-hardmatch as an ImageNet classifier.

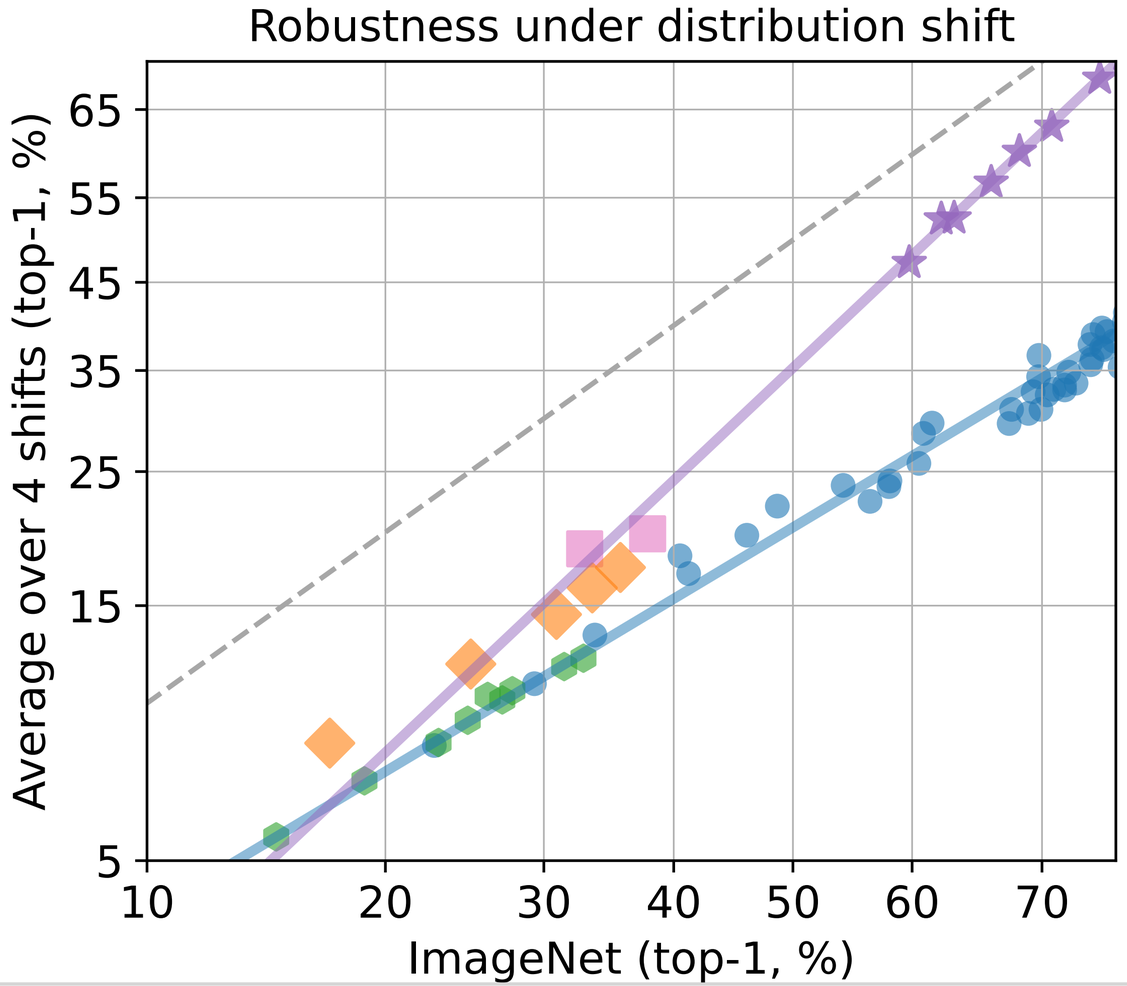
Experimental results



NoCLIP baseline trained on YFCC is robust.

Language supervision is not necessary for robustness.

Experimental results



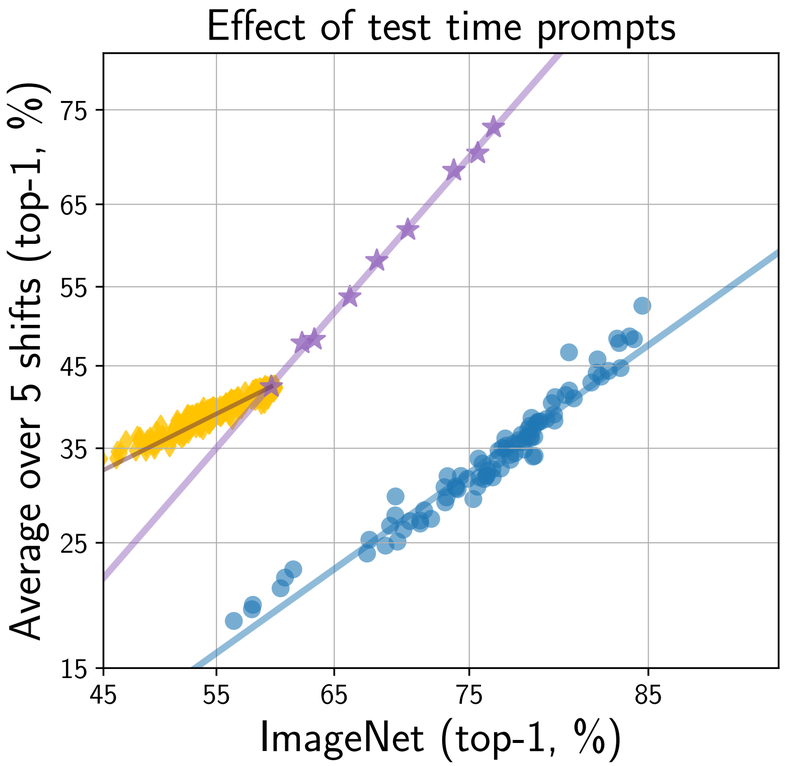
→ NoCLIP baseline trained on YFCC is robust.

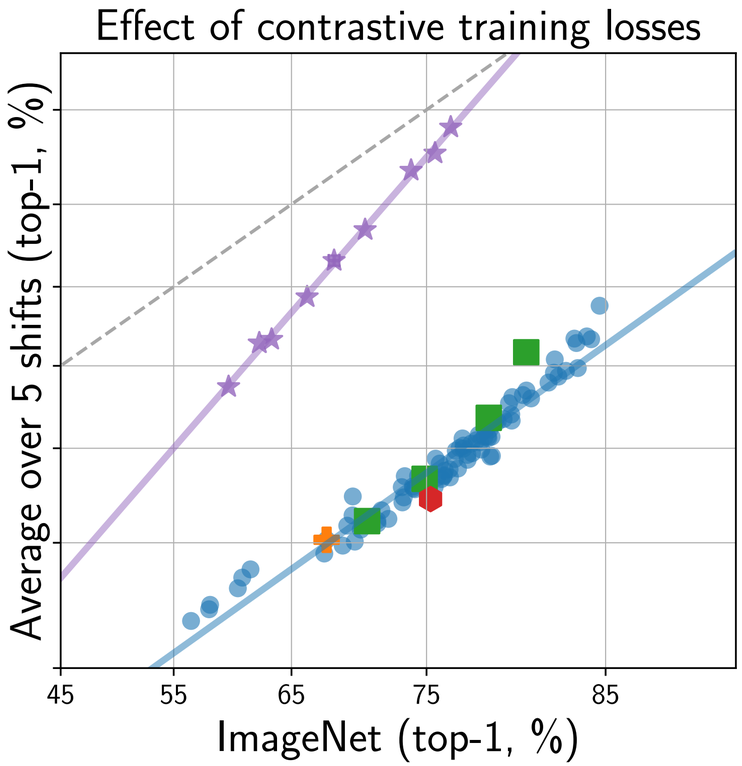
→ Language supervision is not necessary for robustness.

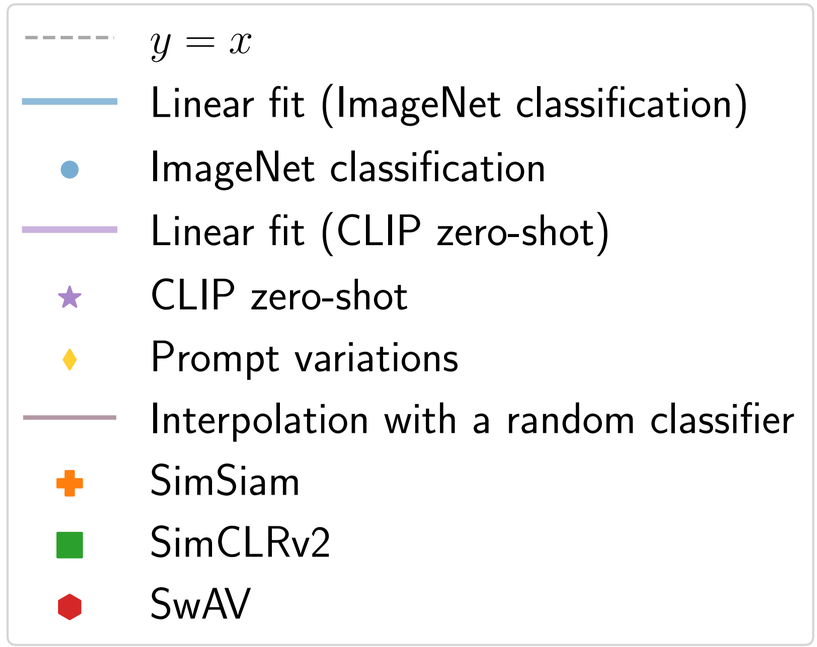
Hypotheses for CLIP's robustness

	CLIP	Standard ImageNet supervised learning
~~**Language supervision**~~	Yes	No
Training distribution	???	ImageNet
~~**Training set size**~~	400M	1.2M
Loss function	Contrastive	Supervised
Test-time prompting	Yes	No
~~**Model architecture**~~	ViTs	CNNs

Prompts and contrastive losses







No robustness gains from prompt variations or contrastive losses.

Hypotheses for CLIP's robustness

	CLIP	Standard ImageNet supervised learning
~~Language supervision~~	Yes	No
Training distribution	???	ImageNet
~~Training set size~~	400M	1.2M
~~Loss function~~	Contrastive	Supervised
~~Test-time prompting~~	Yes	No
~~Model architecture~~	ViTs	CNNs

Next step: can we design better pre-training datasets?

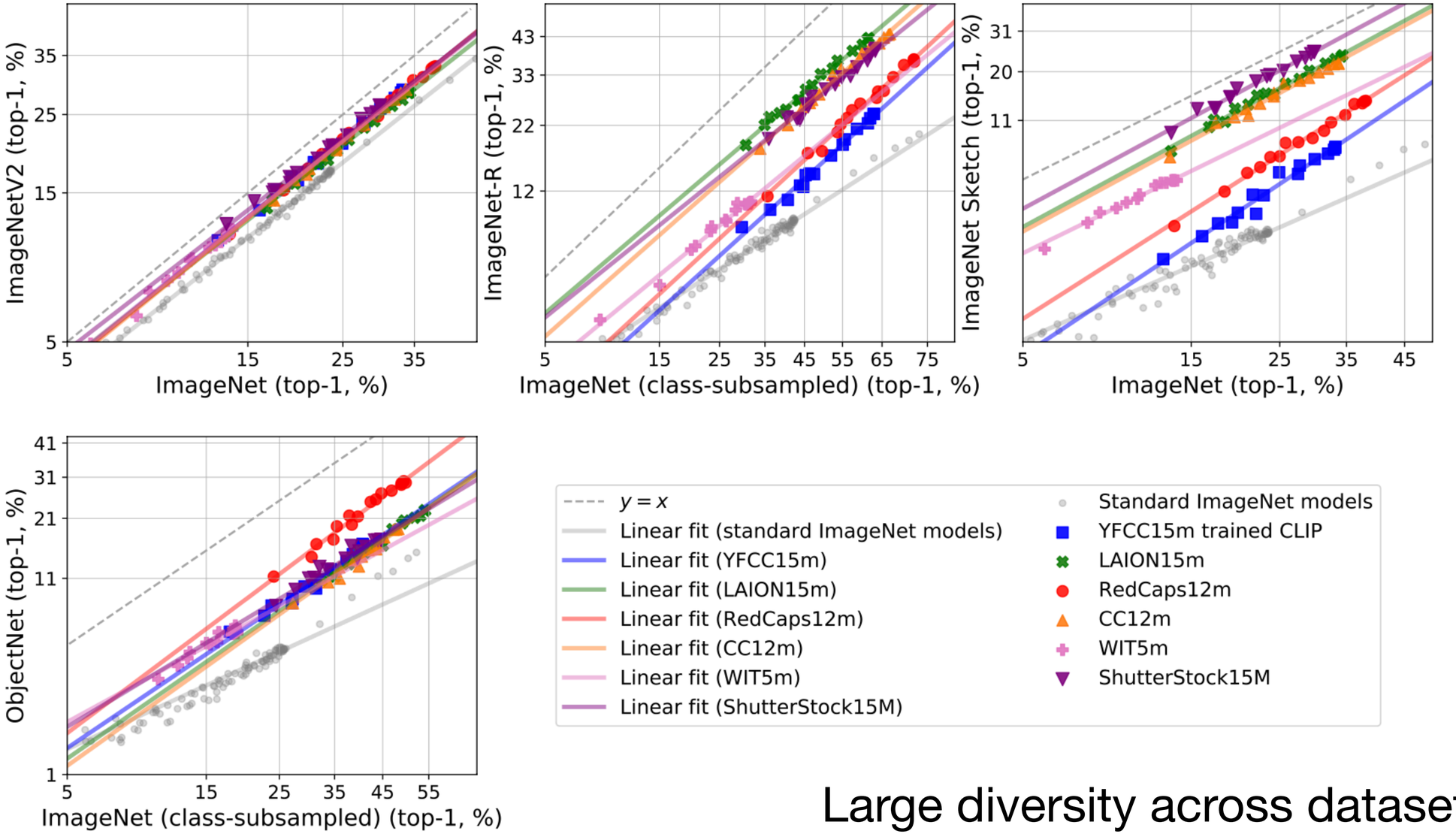
Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP

Thao Nguyen[1] Gabriel Ilharco[1] Mitchell Wortsman[1]
Sewoong Oh[1] Ludwig Schmidt[12]

Abstract

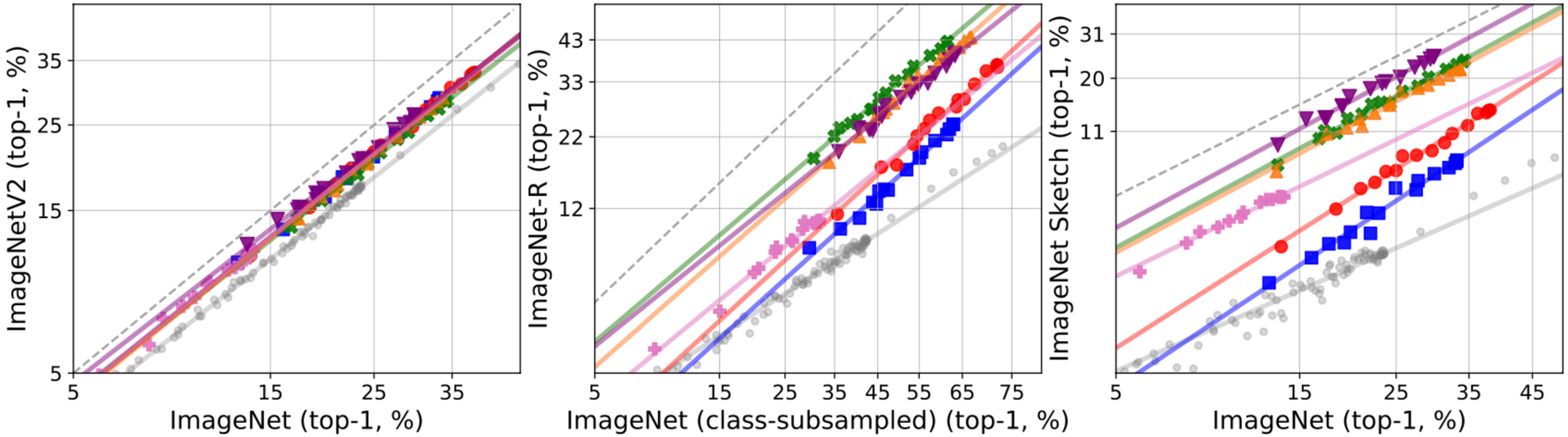
Web-crawled datasets have enabled remarkable generalization capabilities in recent image-text models such as CLIP (Contrastive Language-Image pre-training) or Flamingo, but little is known about the dataset creation processes. In this work, we introduce a testbed of six publicly available data sources—YFCC, LAION, Conceptual Captions, WIT, RedCaps, Shutterstock—to investigate how pre-training distributions induce robustness in CLIP. We find that the performance of the pre-training data varies substantially across distribution shifts, with no single data source dominating. Moreover, we systematically study the interactions between these data sources and find that combining multiple sources does not necessarily yield better models, but rather dilutes the robustness of the best individual data source. We complement our empirical findings with theoretical insights from a simple setting, where combining the training data also results in diluted robustness. In addition, our theoretical model provides a candidate explanation for the success of the CLIP-based data filtering technique recently employed in the LAION dataset. Overall our results demonstrate that simply gathering a large amount of data from the web is not the most effective way to build a pre-training dataset for robust generalization, necessitating further study into dataset design.

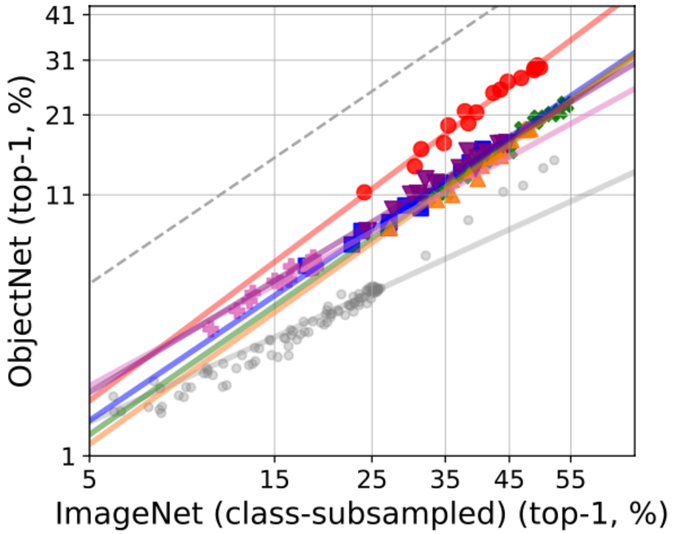
Experiments with six web data sources

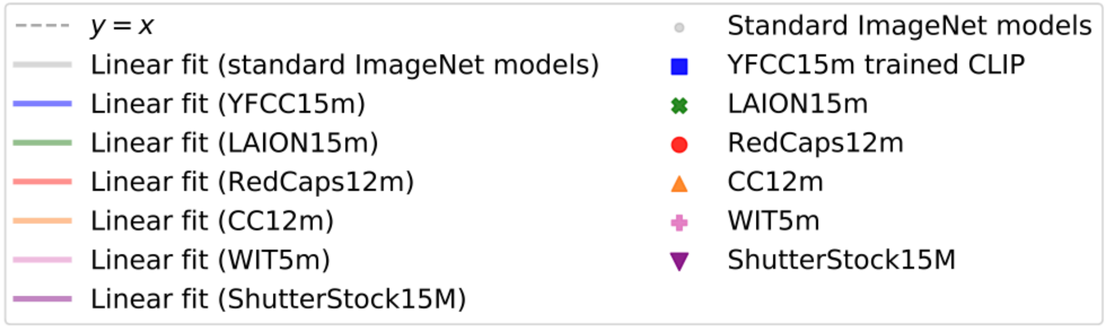


Large diversity across datasets

Experiments with six web data sources

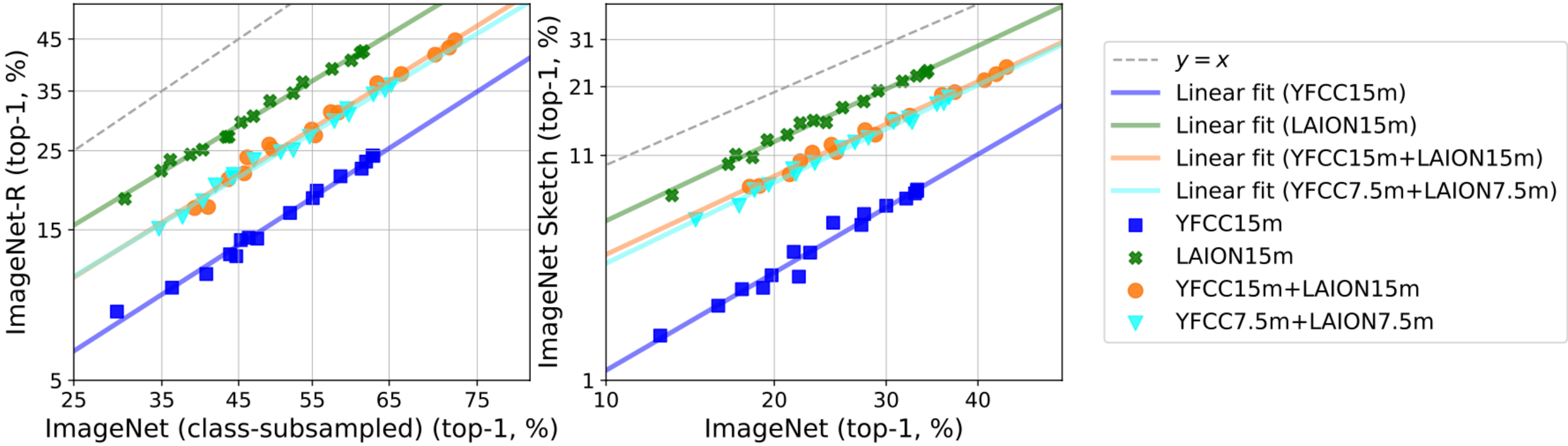






Large diversity across datasets

Combining datasets dilutes robustness

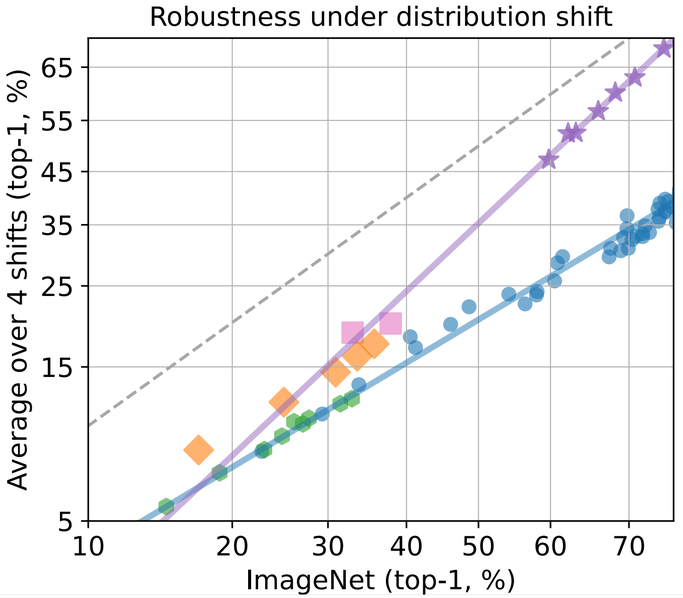


Conclusions

OpenAI's **CLIP** model led to large robustness gains in image classification.

Image distribution is the main reason for CLIP's robustness.

- Not only scale but also **"diversity".**
- Language supervision helps with robustness indirectly: makes it **easier to collect training data.**

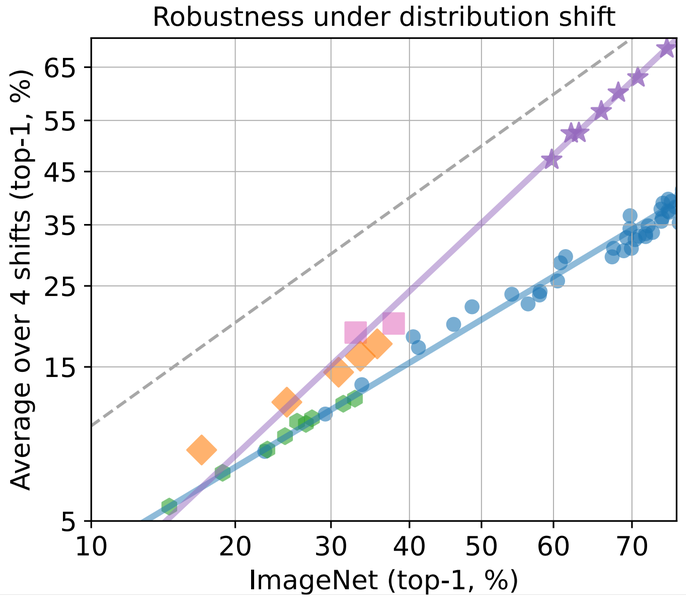


Conclusions

OpenAI's **CLIP** model led to large robustness gains in image classification.

Image distribution is the main reason for CLIP's robustness.

- Not only scale but also **"diversity".**
- Language supervision helps with robustness indirectly: makes it **easier to collect training data.**



CLIP's data sources differ substantially in their induced robustness.

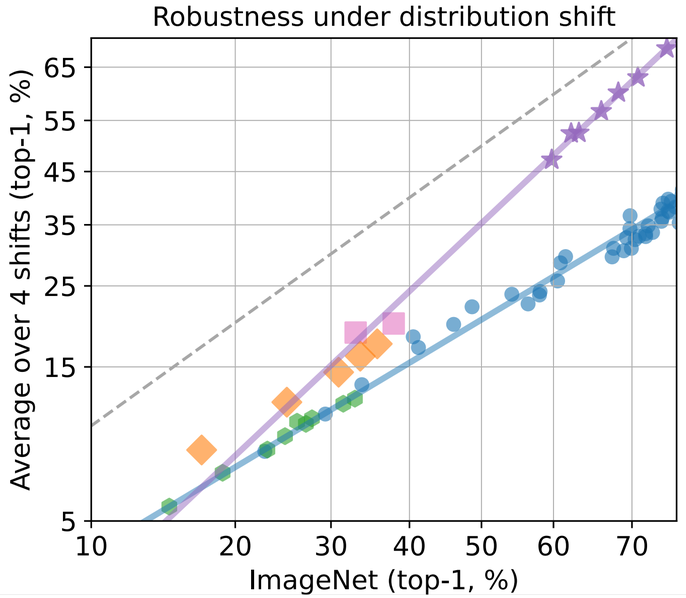
- How do we construct training sets that yield broadly reliable models?

Conclusions

OpenAI's **CLIP** model led to large robustness gains in image classification.

Image distribution is the main reason for CLIP's robustness.

- Not only scale but also **"diversity".**
- Language supervision helps with robustness indirectly: makes it **easier to collect training data.**



CLIP's data sources differ substantially in their induced robustness.

- How do we construct training sets that yield broadly reliable models?

github.com/mlfoundations/open_clip

robustness.imagenetv2.org