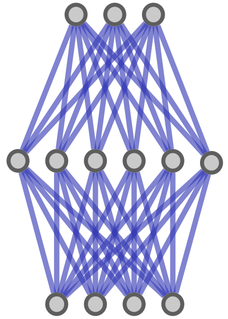
Robust fine-tuning of zero-shot models

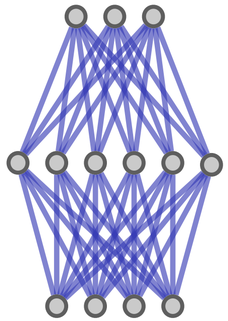
Mitchell Wortsman*, Gabriel Ilharco*, Jong Wook Kim, Mike Li,
Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, Ludwig Schmidt

Simplified machine learning workflow

Simplified machine learning workflow



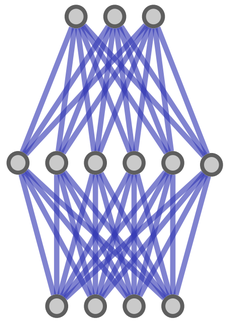
Simplified machine learning workflow



ImageNet (Deng et al.)

Distribution D

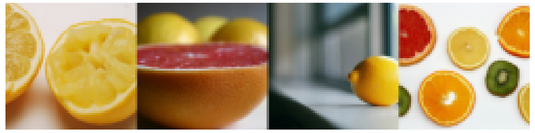
Simplified machine learning workflow



ImageNet (Deng et al.)

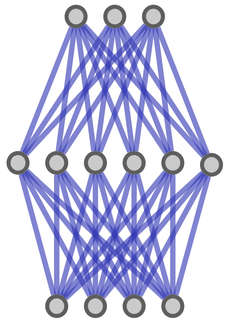
Distribution D

Train



Distribution D

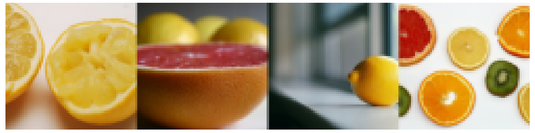
Simplified machine learning workflow



ImageNet (Deng et al.)

Distribution D

Train

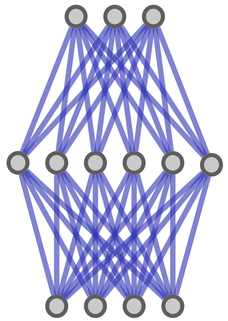


Distribution D

Test

Distribution D

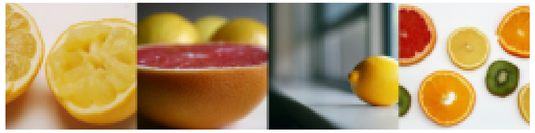
Simplified machine learning workflow



ImageNet (Deng et al.)

Distribution D

Train



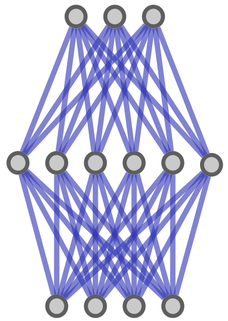
Distribution D

Test

Accuracy = 76.2%

Distribution D

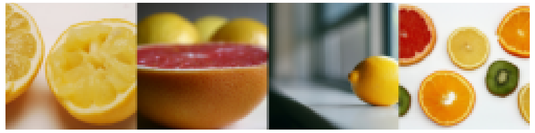
Simplified machine learning workflow



ImageNet (Deng et al.)

Distribution D

Train



Distribution D

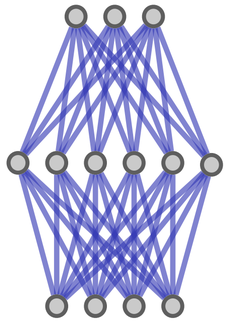
Test

Accuracy = 76.2%

Distribution D

Deploy

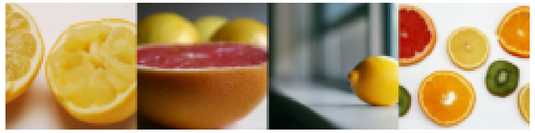
Simplified machine learning workflow



ImageNet (Deng et al.)

Distribution D

Train



Distribution D

Test

Accuracy = 76.2%

Distribution D

Deploy

???

Accuracy = 32.6%

Distribution D'

Evaluating models under natural distribution shift

Measuring robustness to natural distribution shift. Taori et al., 2020

Evaluating models under natural distribution shift

Distribution D

ImageNet (Deng et al.)

Evaluating models under natural distribution shift

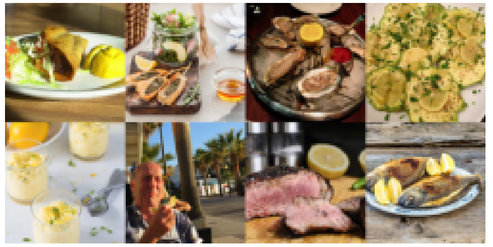
Distribution D

ImageNet (Deng et al.)

Distributions D'

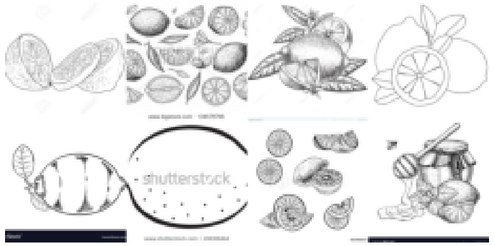
ObjectNet (Barbu et al.)

ImageNet-A (Hendrycks et al.)



ImageNetV2 (Recht et al.)

ImageNet Sketch (Wang et al.)



Measuring robustness to natural distribution shift. Taori et al., 2020

ImageNet (Deng et al.)

Images uploaded by humans to the internet

ImageNet (Deng et al.)

Images uploaded by humans to the internet

ObjectNet (Barbu et al.)

Centered objects from various viewpoints

DATASET	IMAGENET RESNET101
 ImageNet	76.2%

Source: https://openai.com/blog/clip/

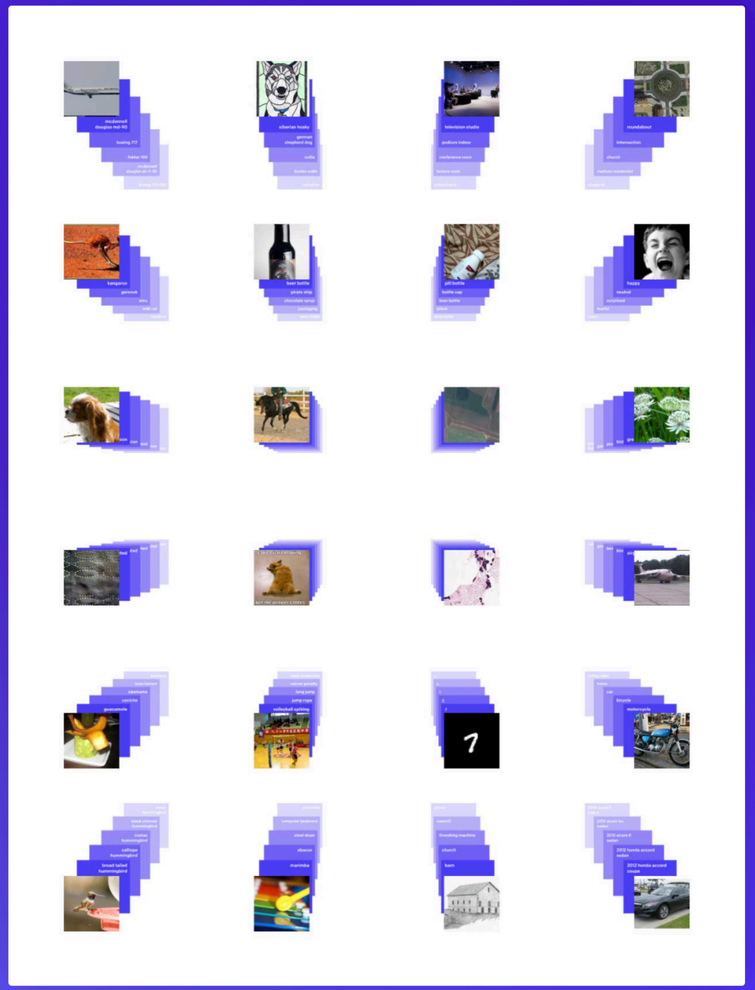
DATASET	IMAGENET RESNET101
 ImageNet	76.2%
 ObjectNet	32.6%

Source: https://openai.com/blog/clip/

CLIP: Connecting Text and Images

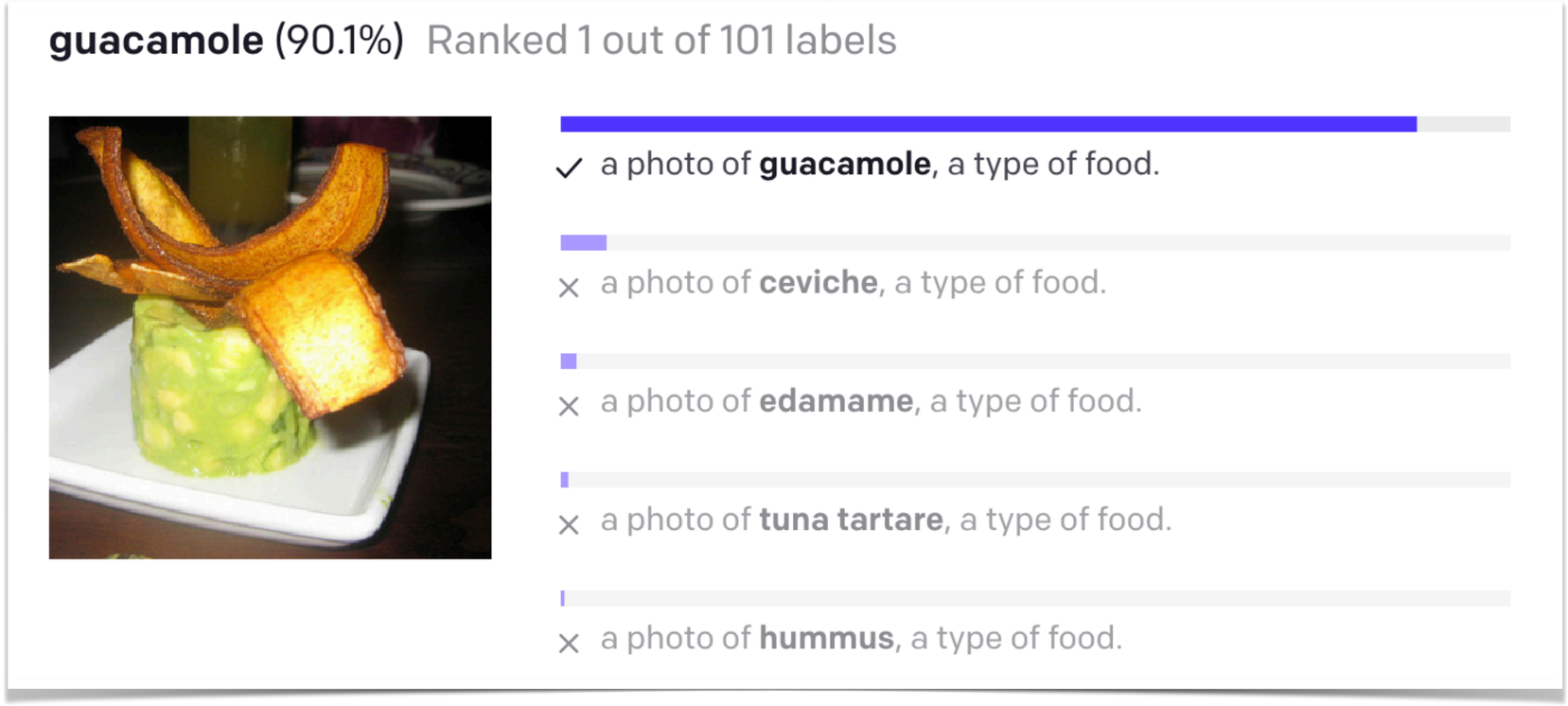
We're introducing a neural network called CLIP which efficiently learns visual concepts from natural language supervision. CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the "zero-shot" capabilities of GPT-2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

January 5, 2021
15 minute read



CLIP is trained to associate **image-caption pairs** on the internet, and can perform zero-shot inference.

CLIP is trained to associate **image-caption pairs** on the internet, and can perform zero-shot inference.



DATASET	IMAGENET RESNET101	CLIP VIT-L
 ImageNet	76.2%	76.2%

DATASET	IMAGENET RESNET101	CLIP VIT-L
 ImageNet	76.2%	76.2%
 ObjectNet	32.6%	72.3%

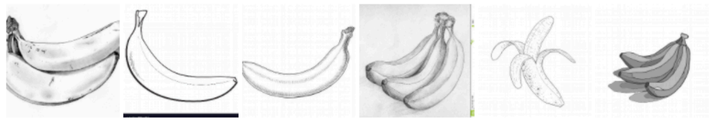
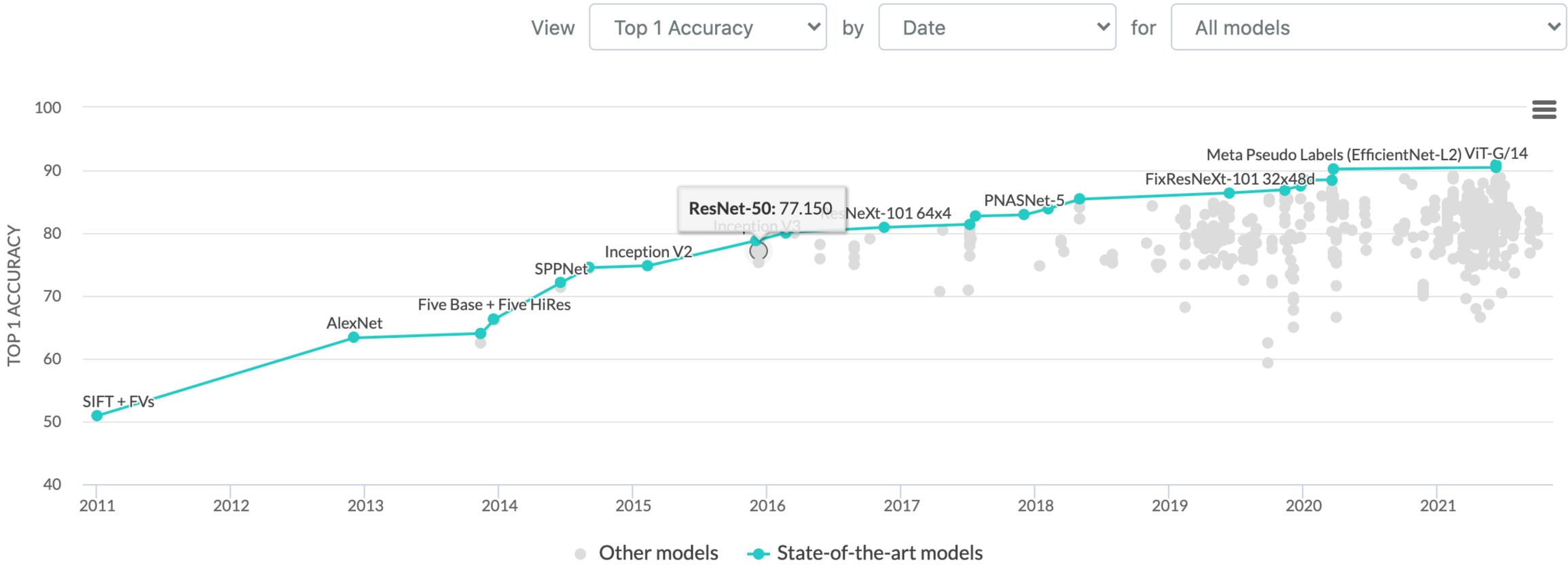
DATASET	IMAGENET RESNET101	CLIP VIT-L
 ImageNet	76.2%	76.2%
 ImageNet V2	64.3%	70.1%
 ImageNet Rendition	37.7%	88.9%
 ObjectNet	32.6%	72.3%
 ImageNet Sketch	25.2%	60.2%
 ImageNet Adversarial	2.7%	77.1%

Image Classification on ImageNet

Leaderboard Dataset



Source: https://paperswithcode.com/sota/image-classification-on-imagenet

To improve a models accuracy it is *fine-tuned* on data from the domain of interest.

To improve a models accuracy it is *fine-tuned* on data from the domain of interest.

But: despite accuracy improvements, ***fine-tuning* CLIP deteriorates robustness.**

To improve a models accuracy it is *fine-tuned* on data from the domain of interest.

But: despite accuracy improvements, ***fine-tuning* CLIP deteriorates robustness.**

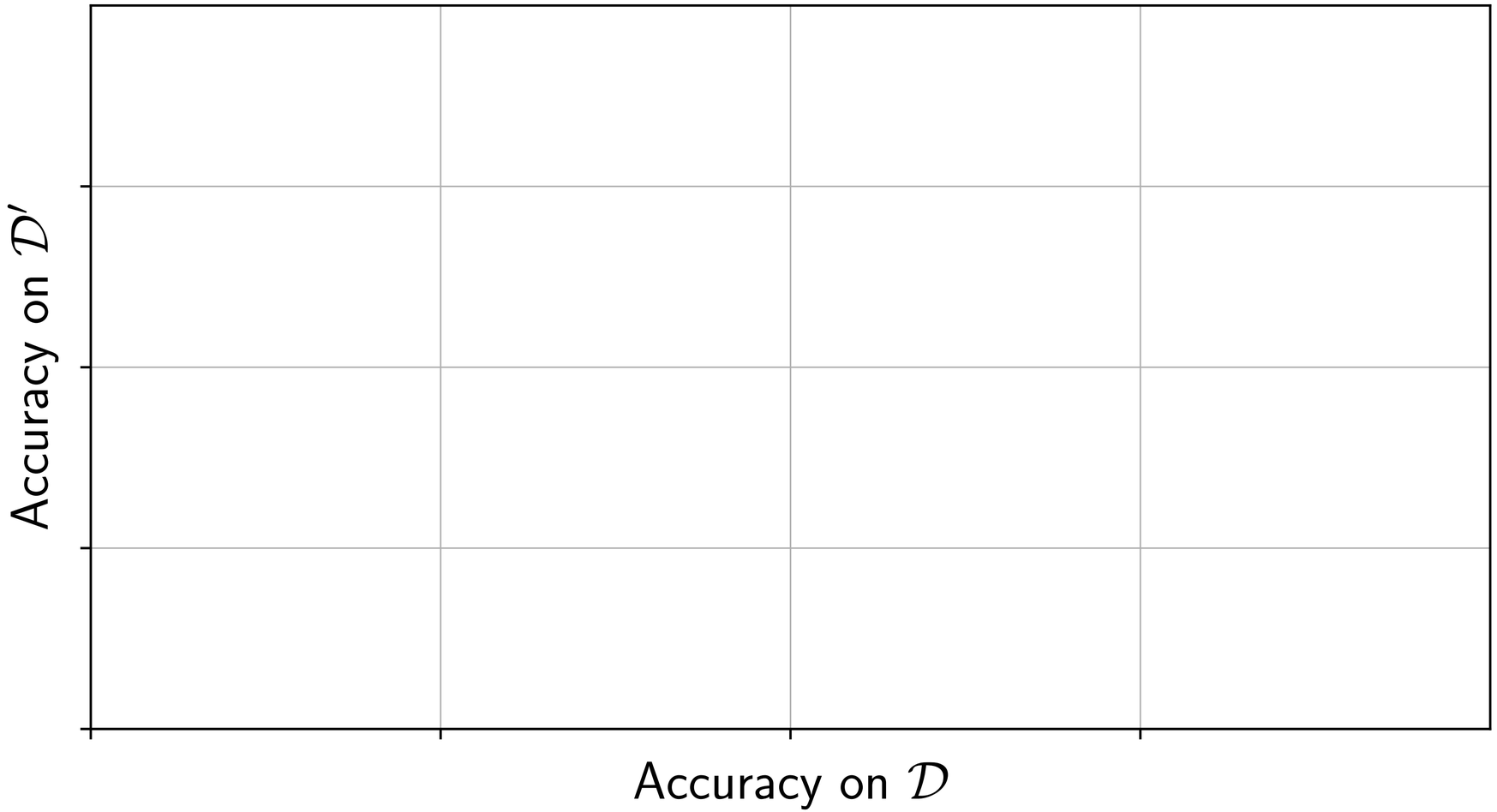
This work:

To improve a models accuracy it is *fine-tuned* on data from the domain of interest.

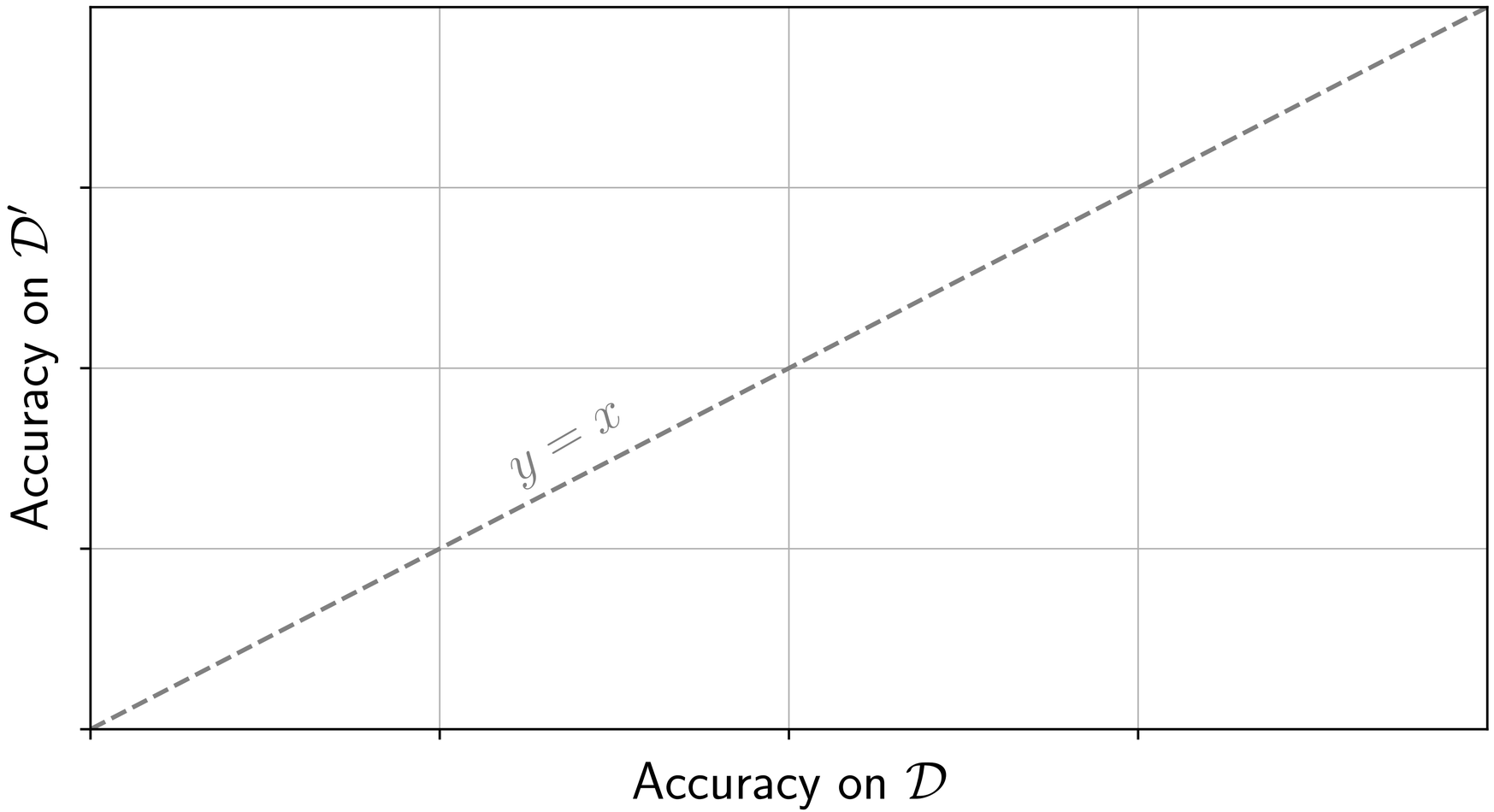
But: despite accuracy improvements, ***fine-tuning* CLIP deteriorates robustness.**

This work:

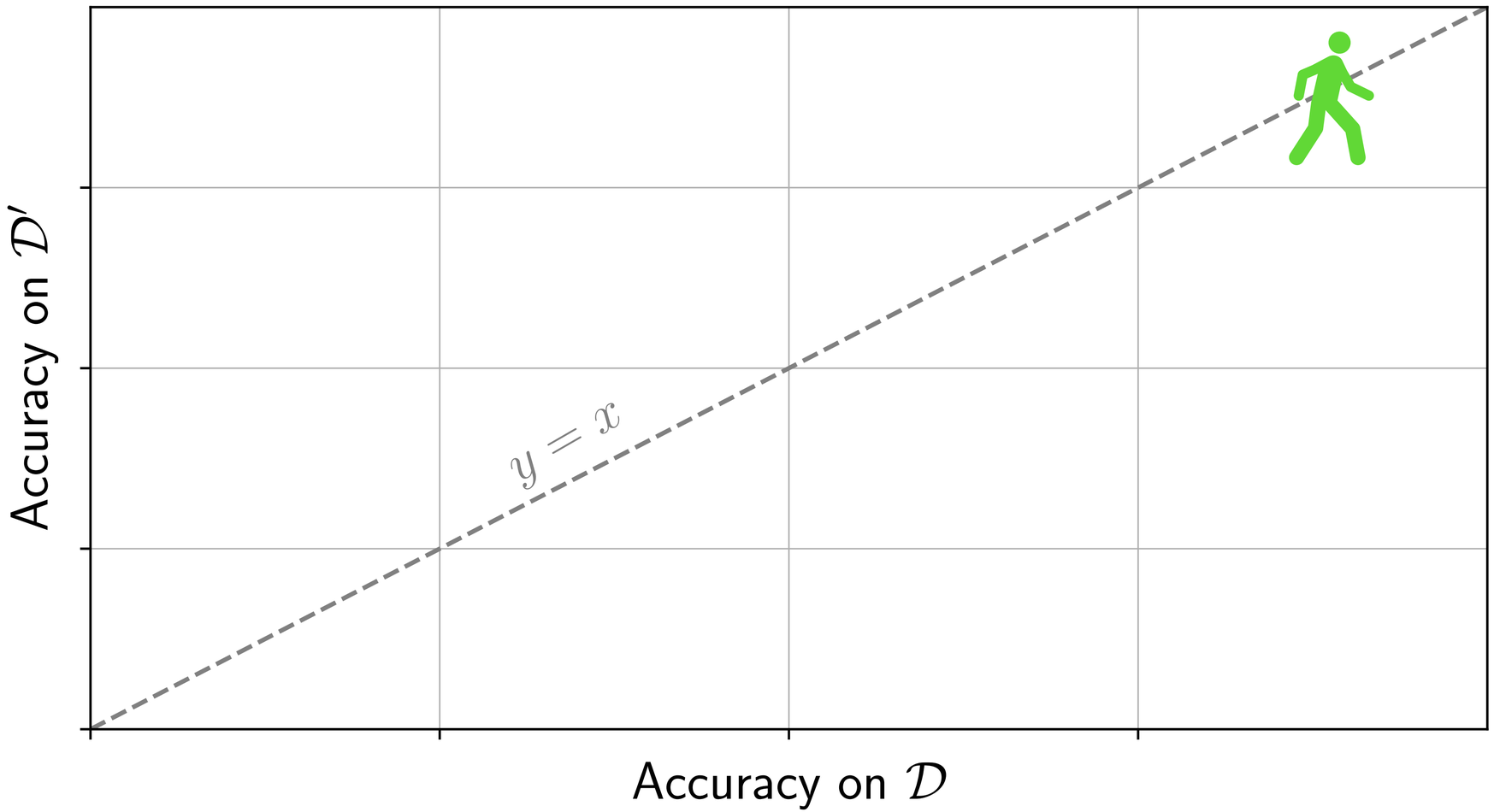
Can we fine-tune CLIP while maintaining robustness?



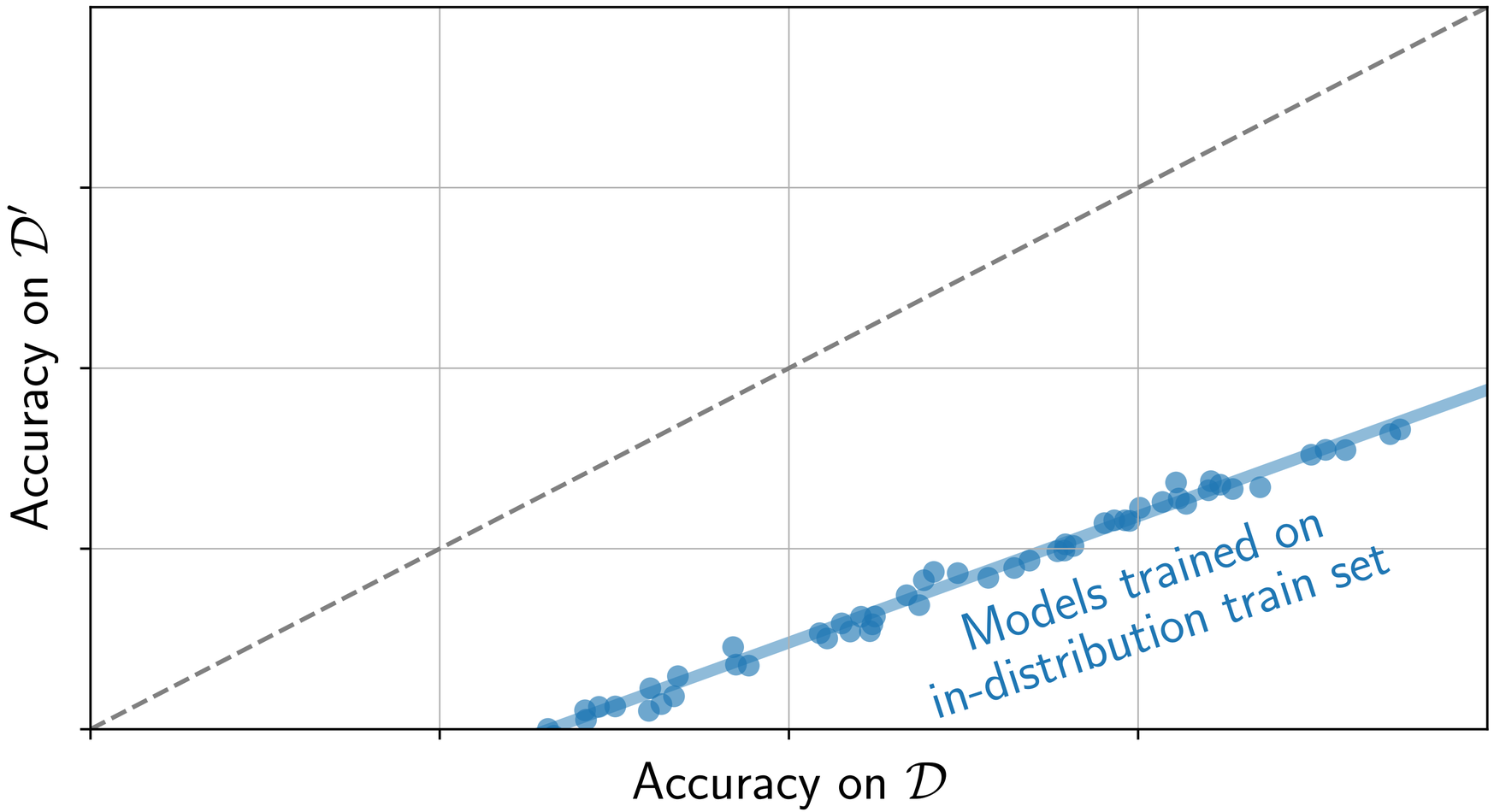
Measuring robustness to natural distribution shift. Taori et al., 2020



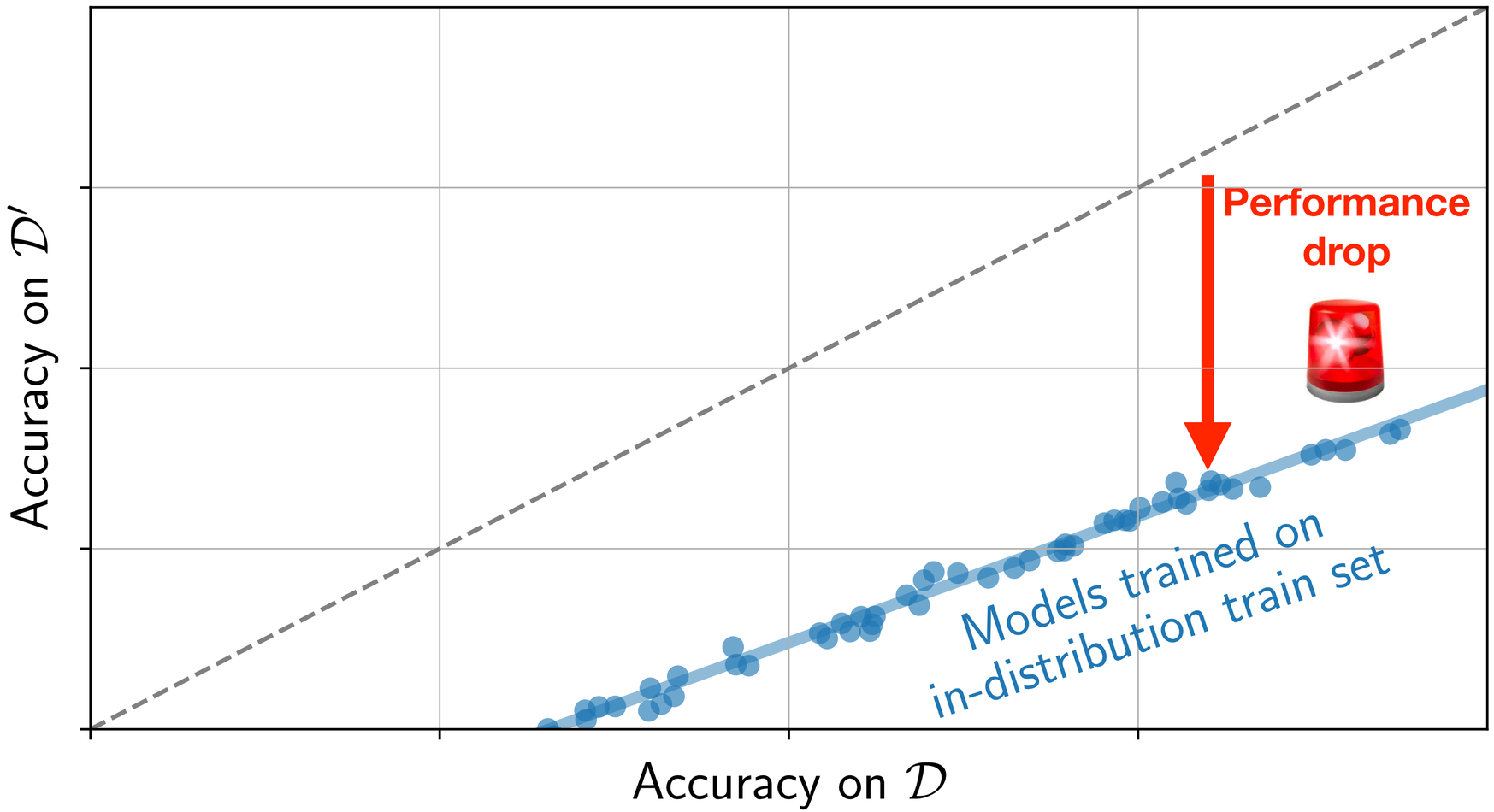
Measuring robustness to natural distribution shift. Taori et al., 2020



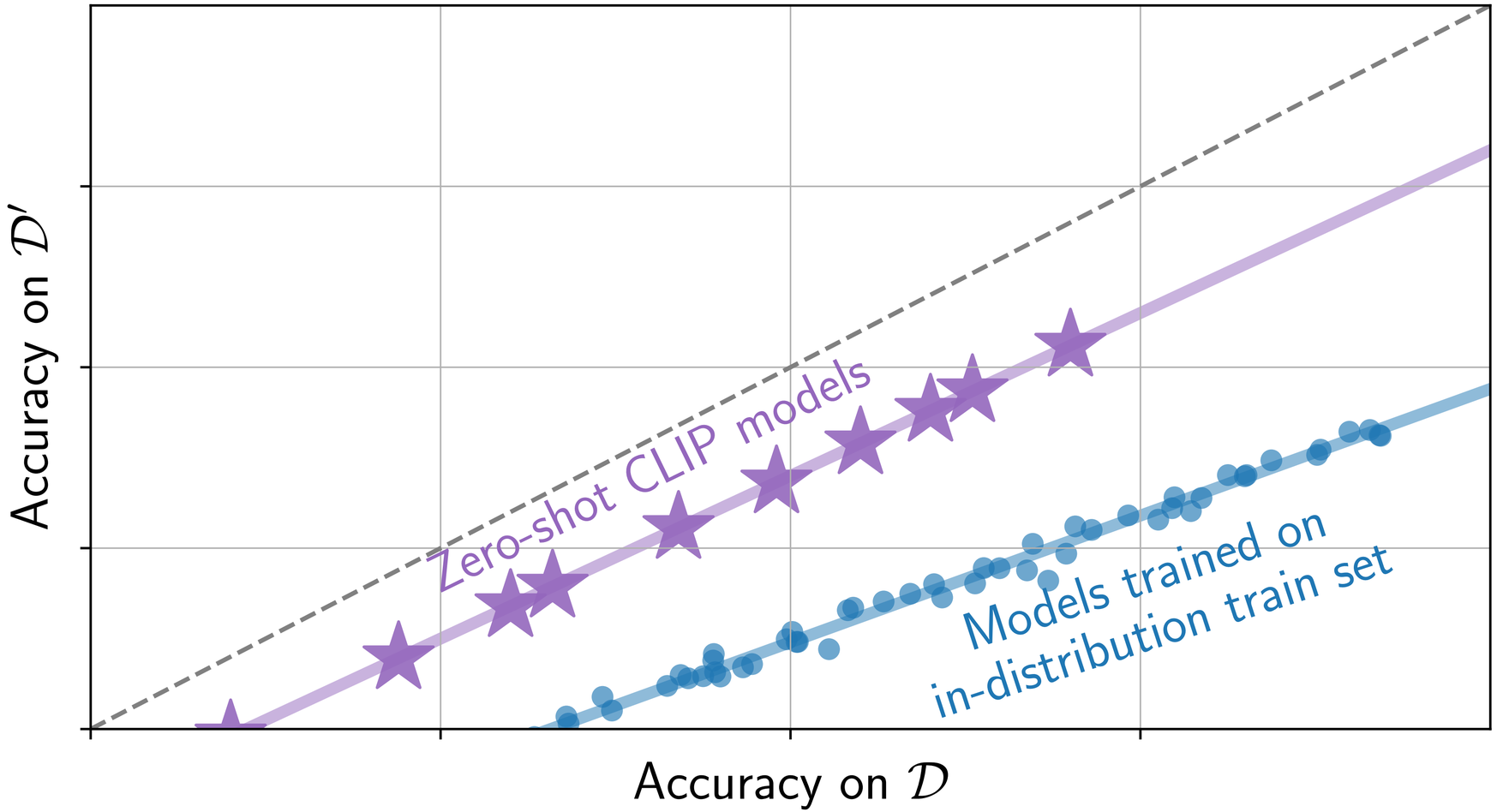
Evaluating Machine Accuracy on ImageNet. Shankar et al., 2020



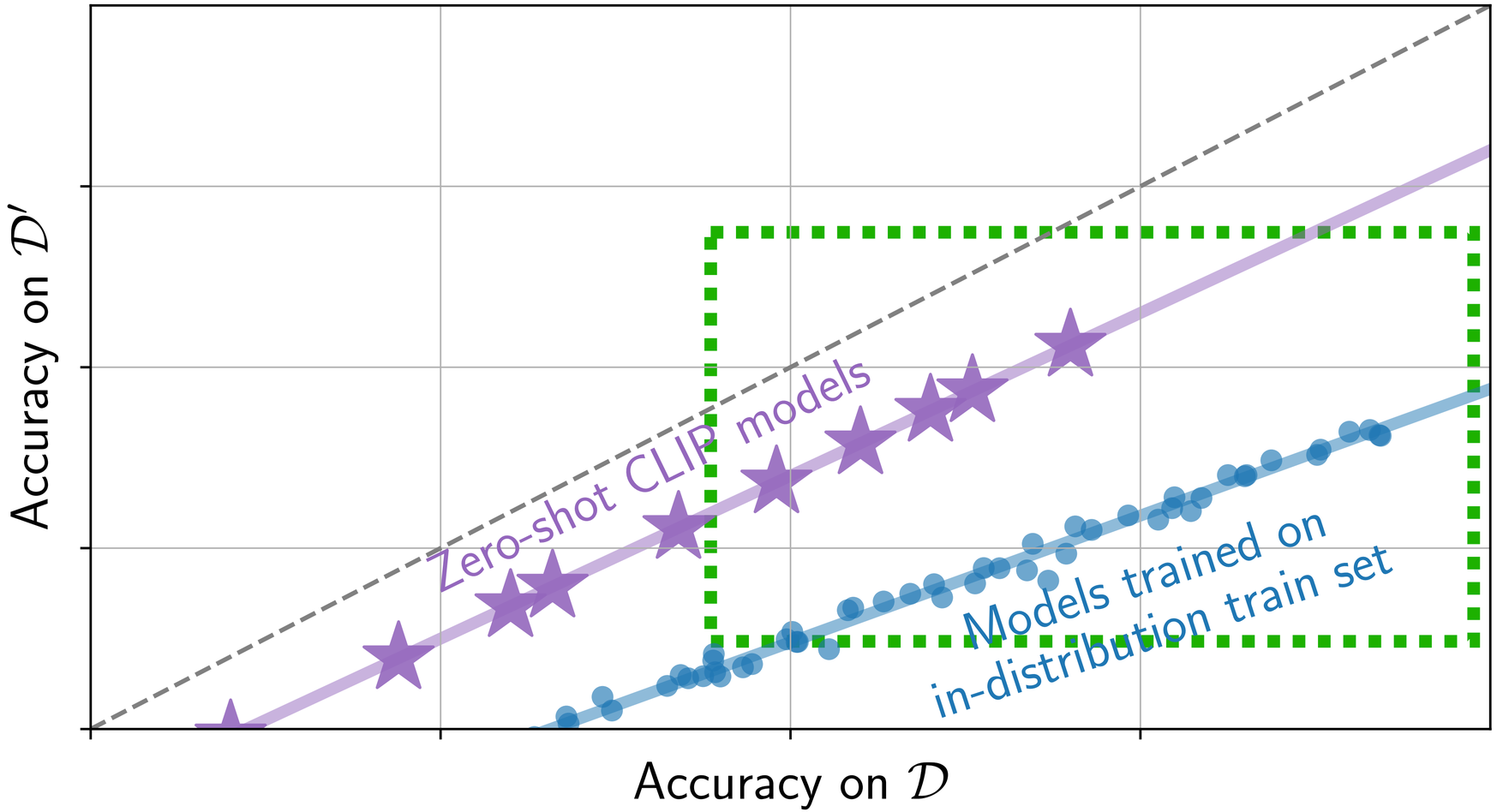
Measuring robustness to natural distribution shift. Taori et al., 2020



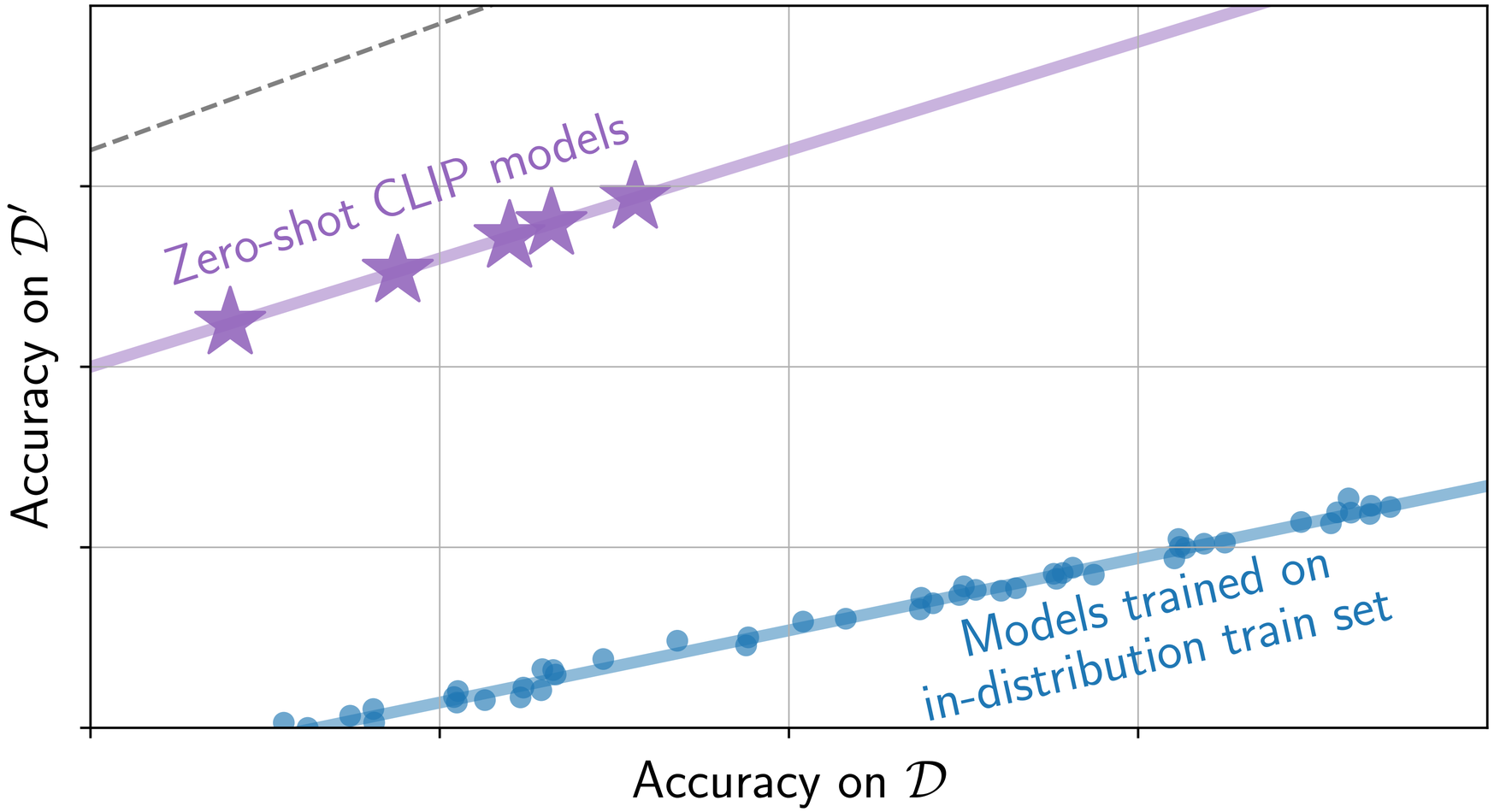
Measuring robustness to natural distribution shift. Taori et al., 2020



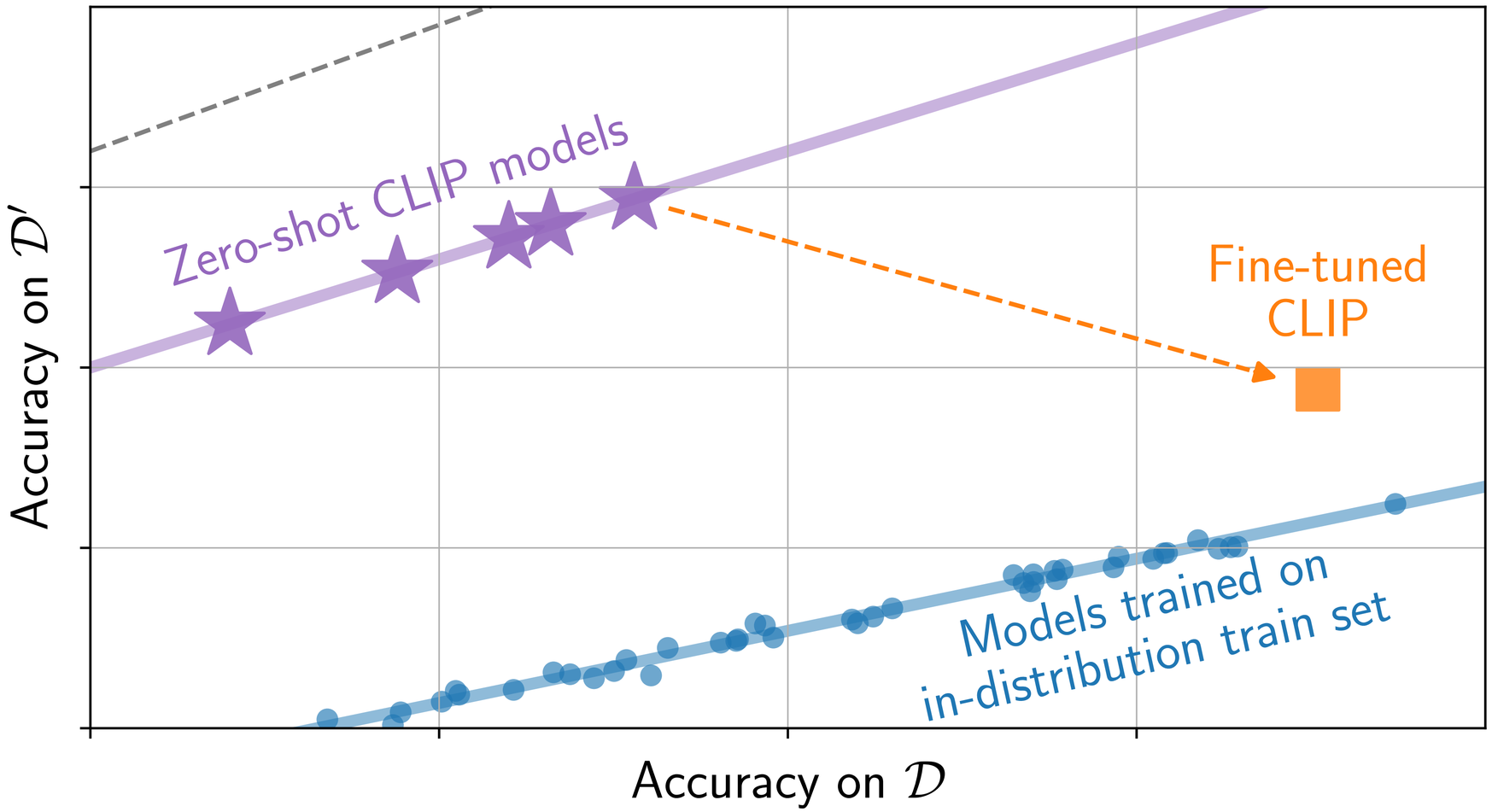
CLIP. Radford et al., 2020

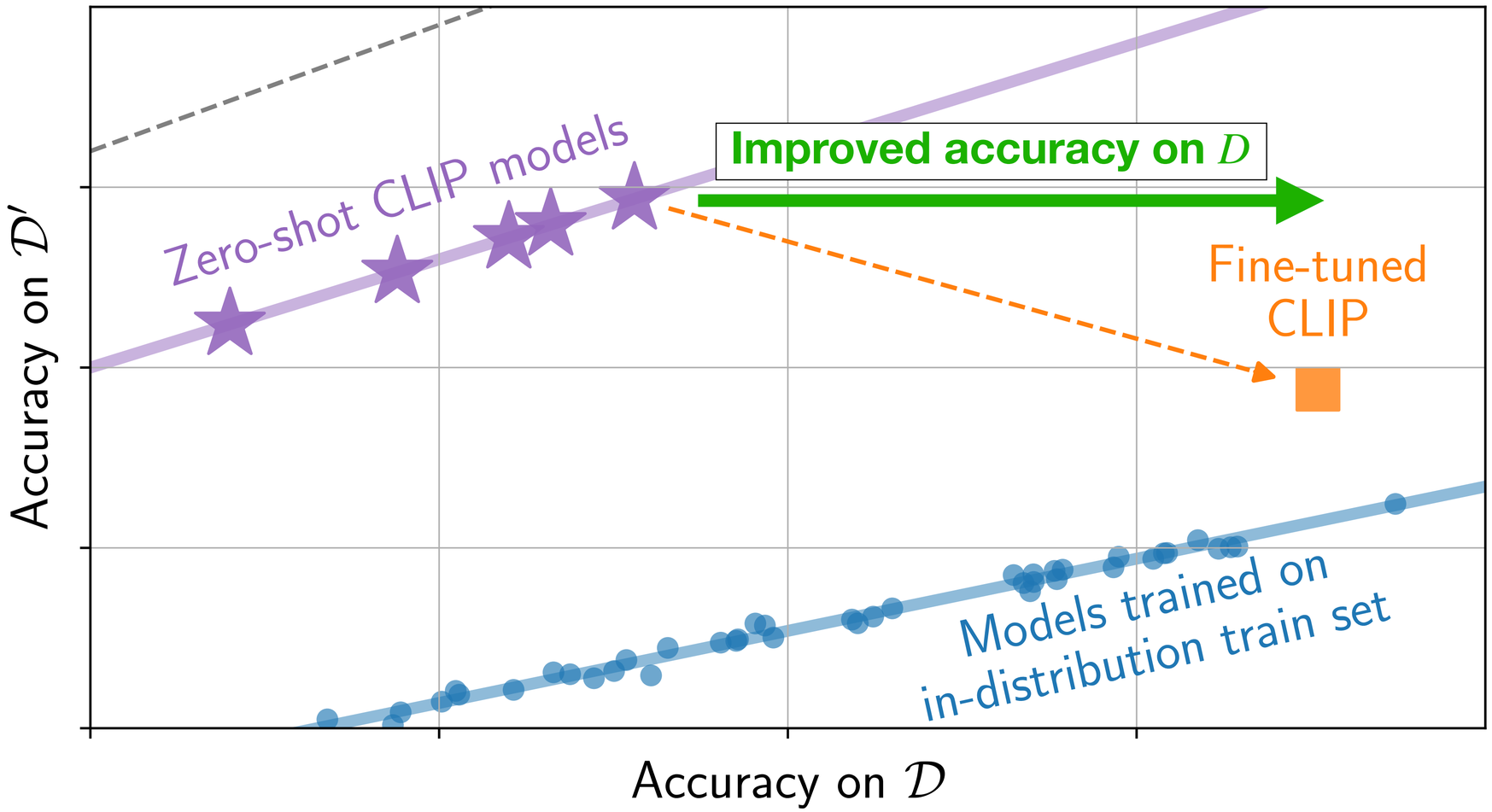


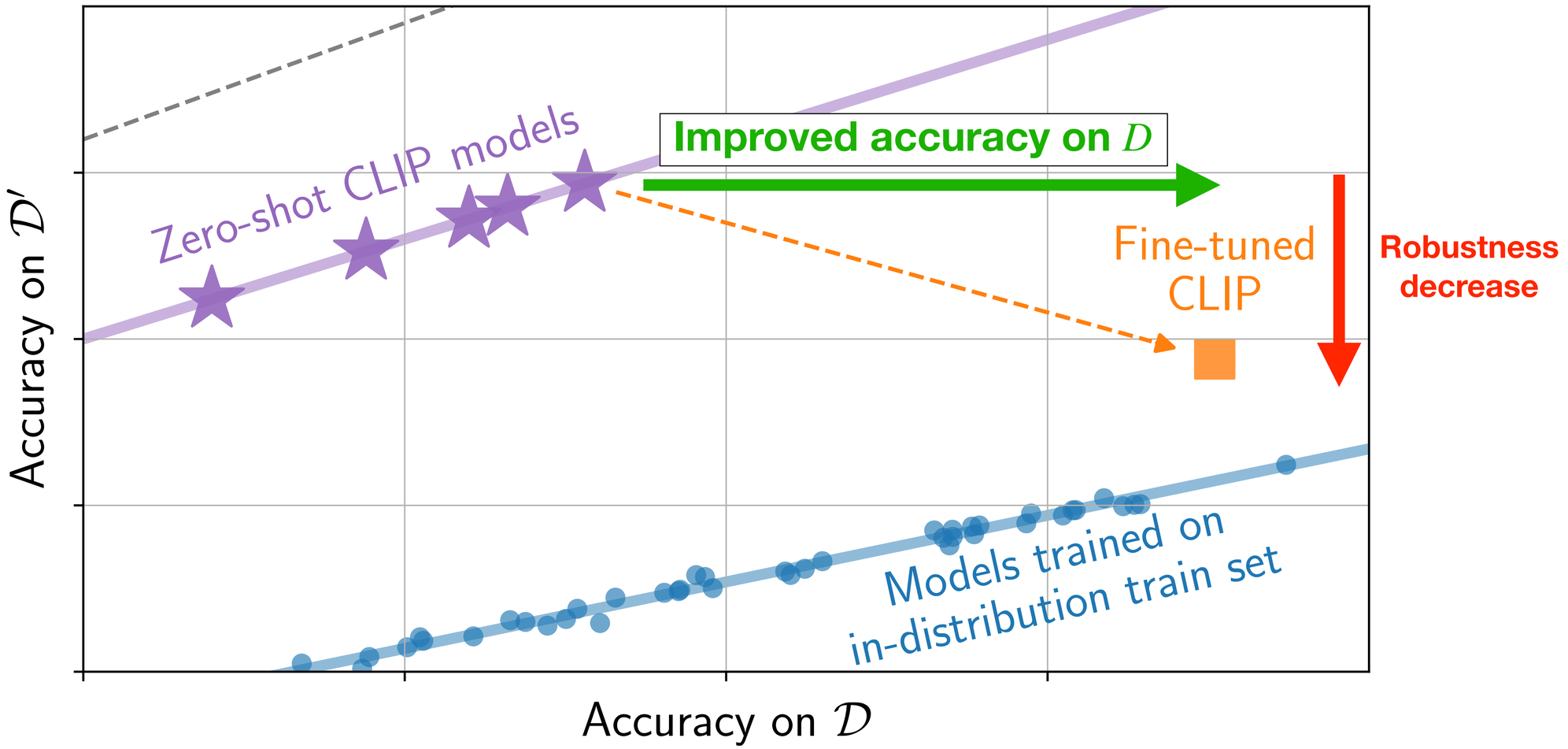
CLIP. Radford et al., 2020



CLIP. Radford et al., 2020

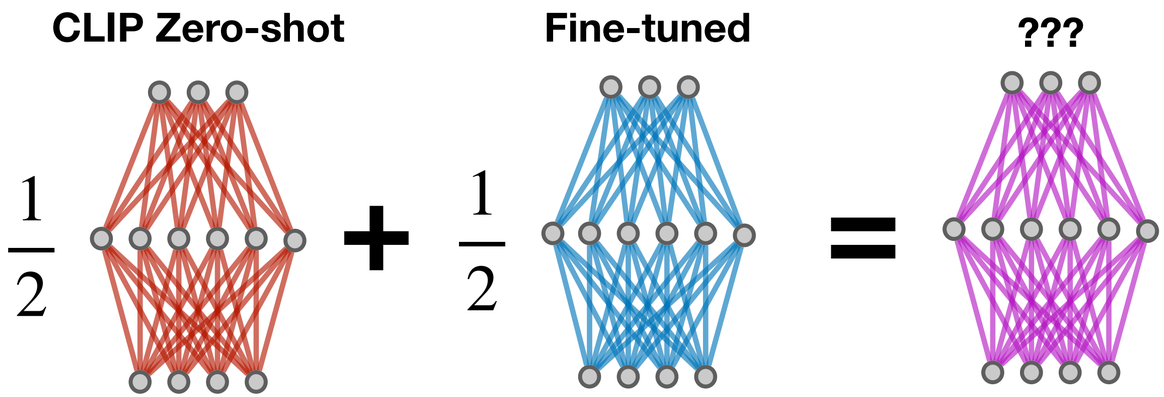

Zero-shot CLIP models
Fine-tuned
CLIP
Models trained on
in-distribution train set
Accuracy on $\mathcal{D}'$
Accuracy on $\mathcal{D}$


Zero-shot CLIP models
Improved accuracy on D
Fine-tuned CLIP
Models trained on in-distribution train set
Accuracy on $\mathcal{D}'$
Accuracy on $\mathcal{D}$


Zero-shot CLIP models
Improved accuracy on D
Fine-tuned CLIP
Robustness decrease
Models trained on in-distribution train set
Accuracy on $\mathcal{D}'$
Accuracy on $\mathcal{D}$

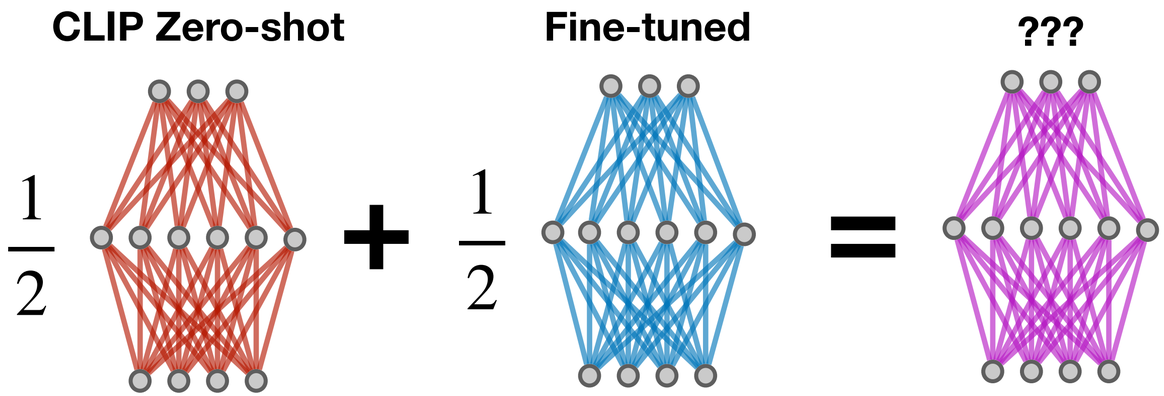
	CLIP Zero-shot	Fine-tuned
Accuracy on $\mathcal{D}$		
Accuracy on $\mathcal{D}'$		

	CLIP Zero-shot	Fine-tuned	???
Accuracy on $\mathcal{D}$			
Accuracy on $\mathcal{D}'$			



Accuracy on $\mathcal{D}$

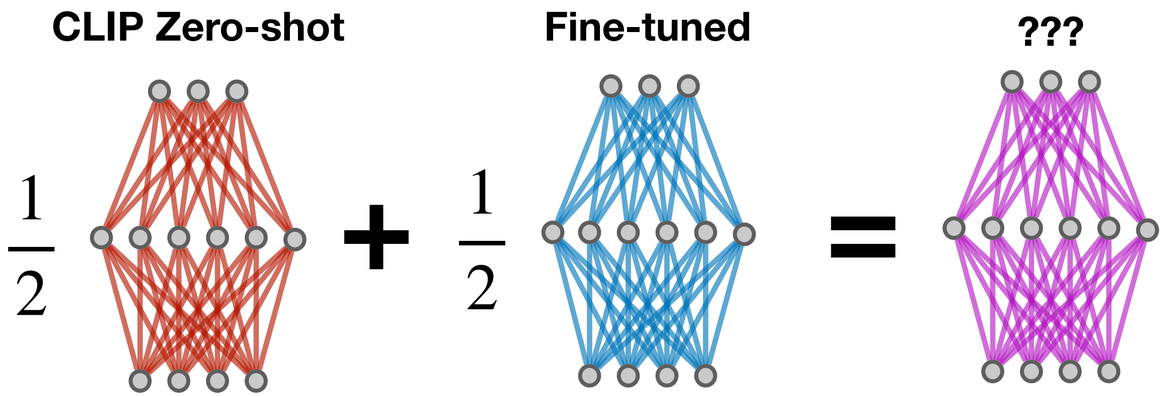
Accuracy on $\mathcal{D}'$



Accuracy on $\mathcal{D}$	➖	✅	✅
Accuracy on $\mathcal{D}'$	✅	➖	✅

Weight-space ensembling = linearly interpolating in weight space

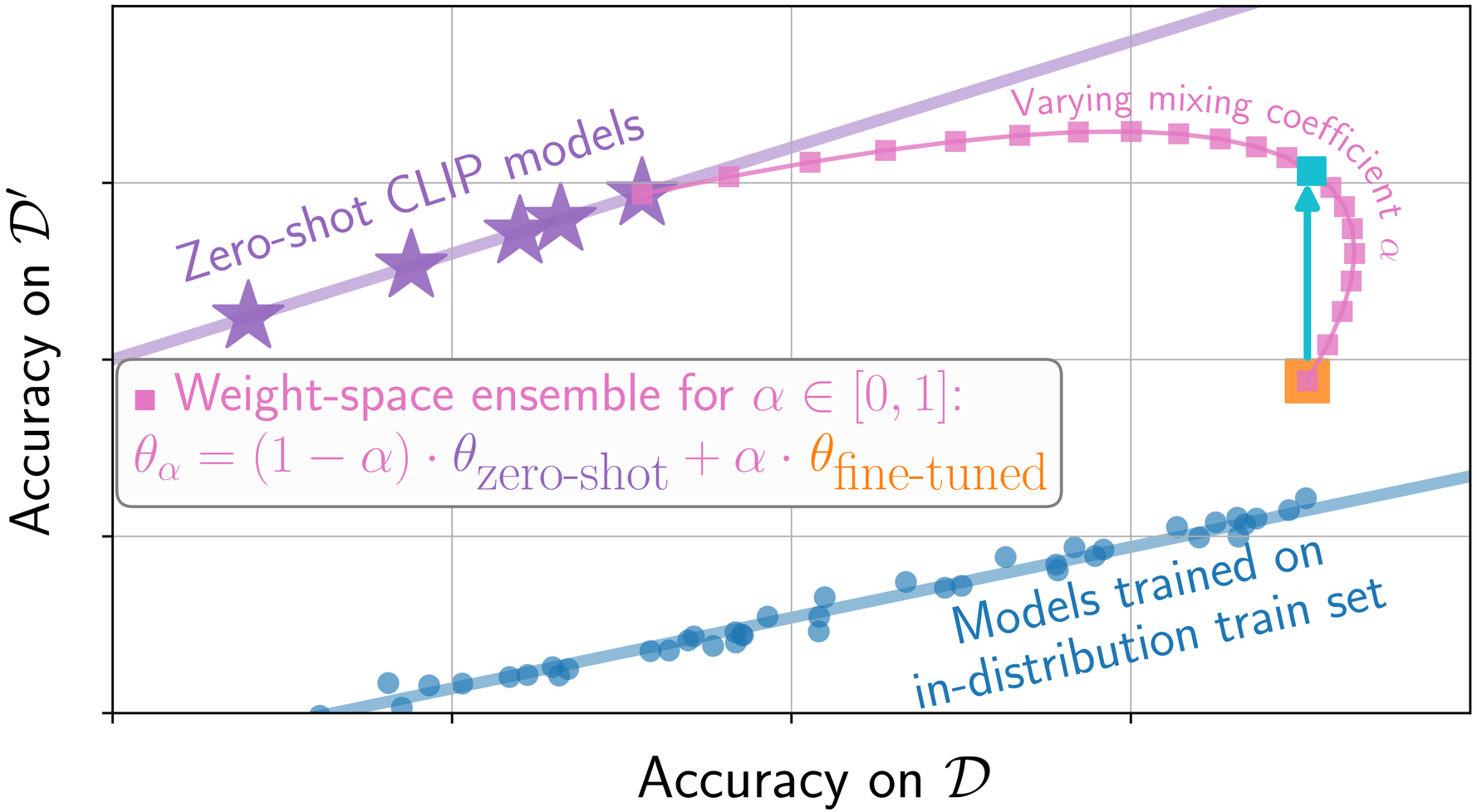
Weight-space ensembles for fine-tuning = WiSE-FT

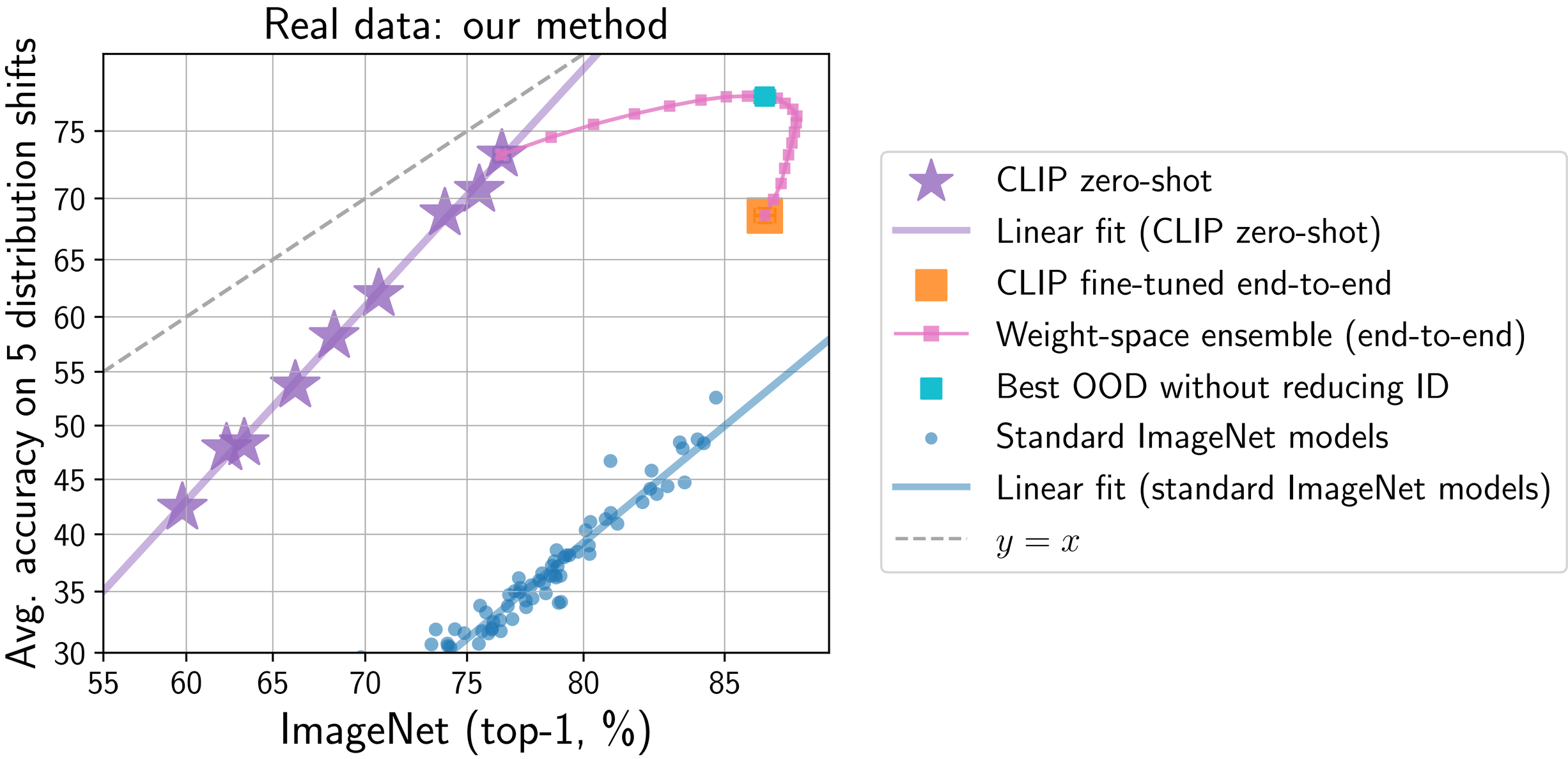


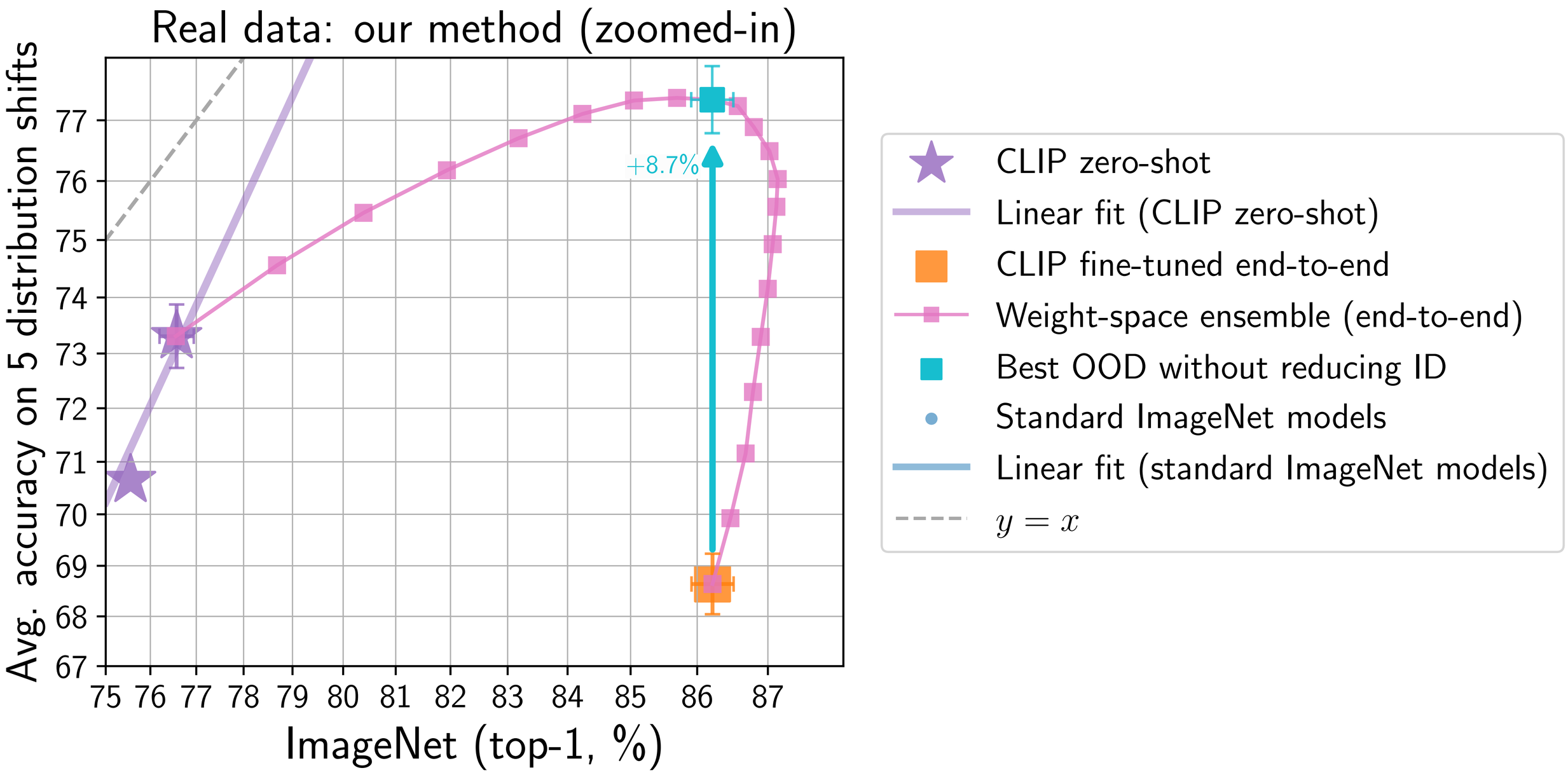
Accuracy on $\mathcal{D}$	➖	✅	✅
Accuracy on $\mathcal{D}'$	✅	➖	✅

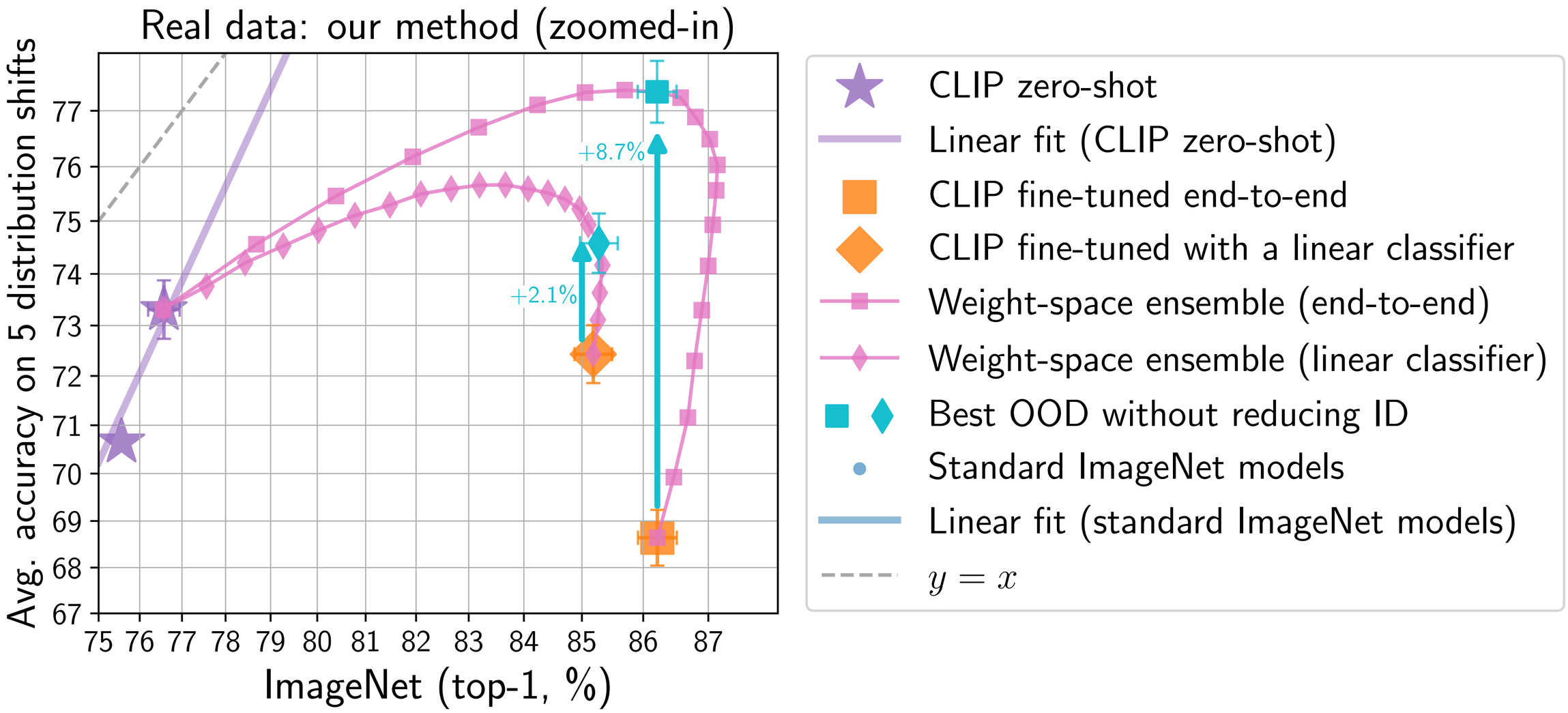
Weight-space ensembling = linearly interpolating in weight space

Schematic: our method, WiSE-FT




Real data: our method
Avg. accuracy on 5 distribution shifts
ImageNet (top-1, %)
30
35
40
45
50
55
60
65
70
75
55
60
65
70
75
80
85
CLIP zero-shot
Linear fit (CLIP zero-shot)
CLIP fine-tuned end-to-end
Weight-space ensemble (end-to-end)
Best OOD without reducing ID
Standard ImageNet models
Linear fit (standard ImageNet models)
$y = x$


Real data: our method (zoomed-in)
Avg. accuracy on 5 distribution shifts
ImageNet (top-1, %)
+8.7%
67
68
69
70
71
72
73
74
75
76
77
75 76 77 78 79 80 81 82 83 84 85 86 87
CLIP zero-shot
Linear fit (CLIP zero-shot)
CLIP fine-tuned end-to-end
Weight-space ensemble (end-to-end)
Best OOD without reducing ID
Standard ImageNet models
Linear fit (standard ImageNet models)
$y = x$

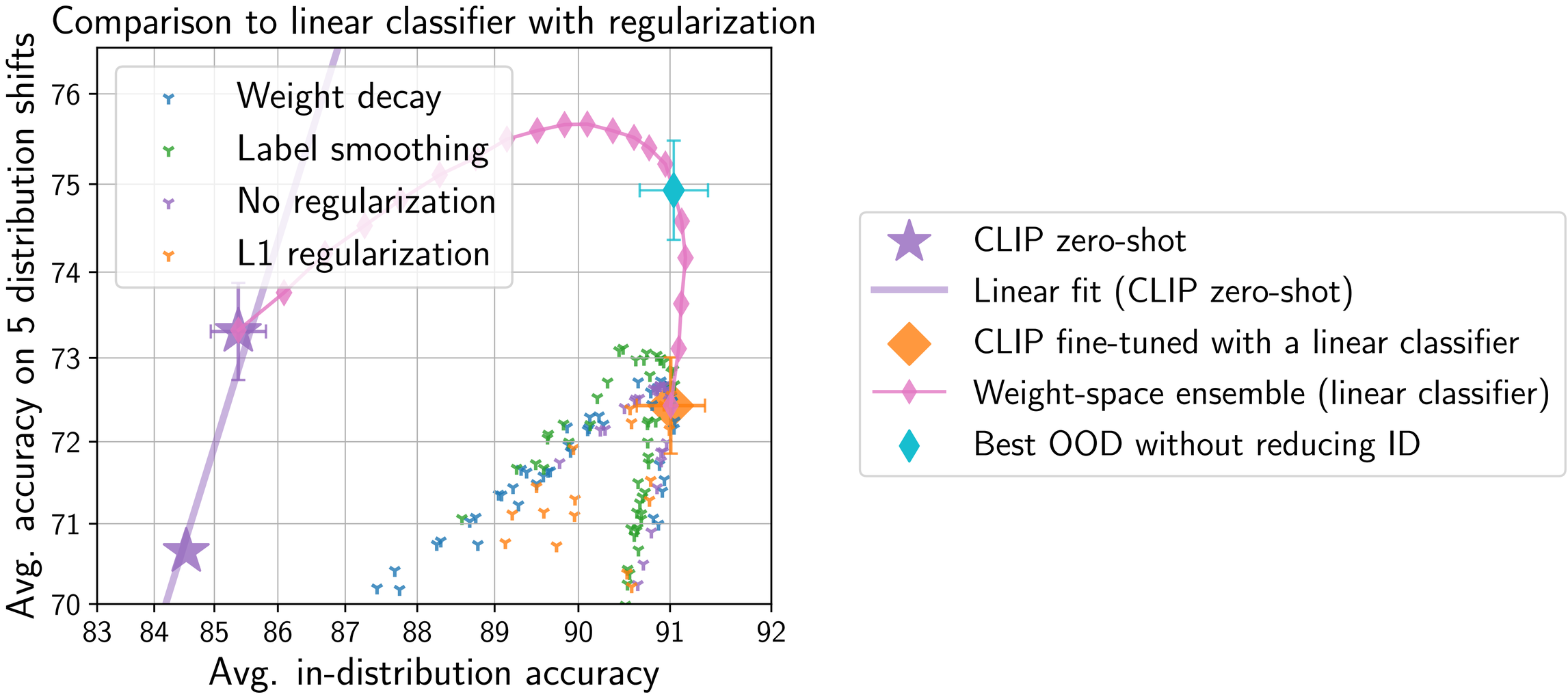

Real data: our method (zoomed-in)
Avg. accuracy on 5 distribution shifts
ImageNet (top-1, %)
+8.7%
+2.1%
CLIP zero-shot
Linear fit (CLIP zero-shot)
CLIP fine-tuned end-to-end
CLIP fine-tuned with a linear classifier
Weight-space ensemble (end-to-end)
Weight-space ensemble (linear classifier)
Best OOD without reducing ID
Standard ImageNet models
Linear fit (standard ImageNet models)
$y = x$

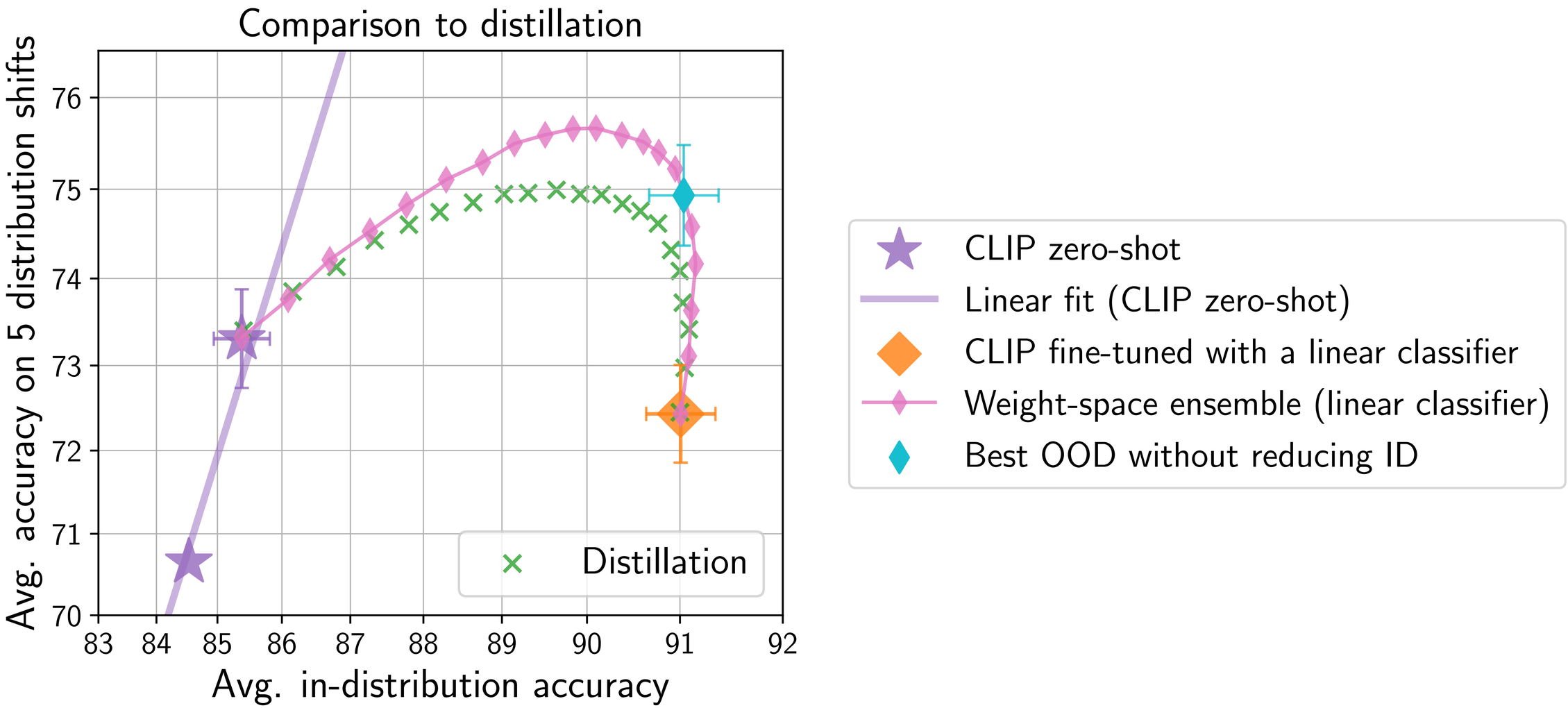
Outline

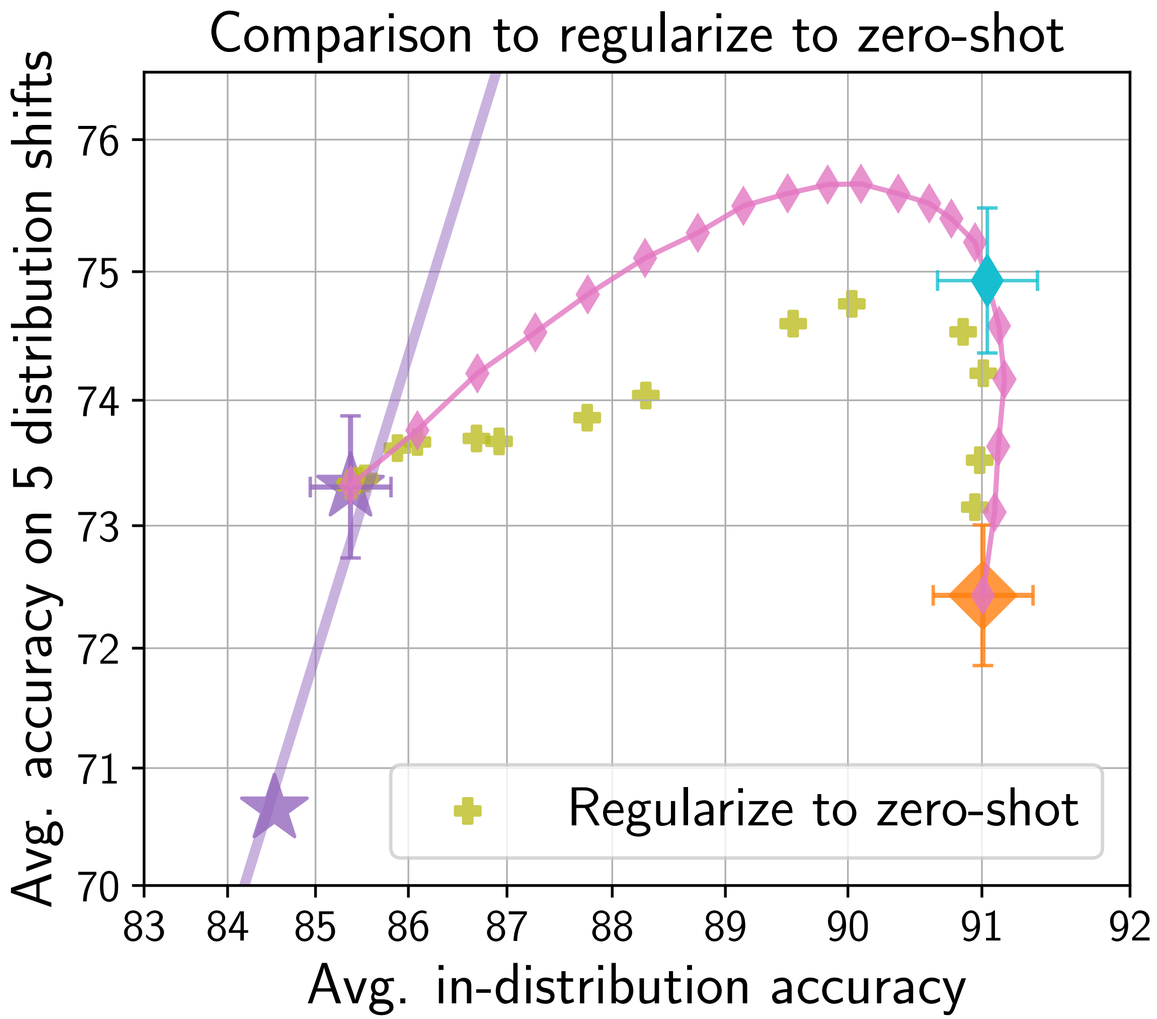
- Baselines
- Choosing the mixing coefficient
- More distribution shifts and datasets
- Why does this work

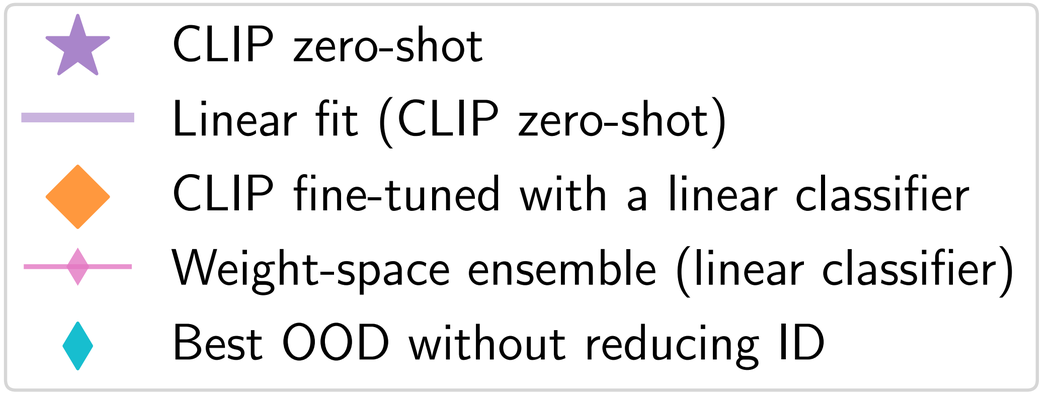
Outline

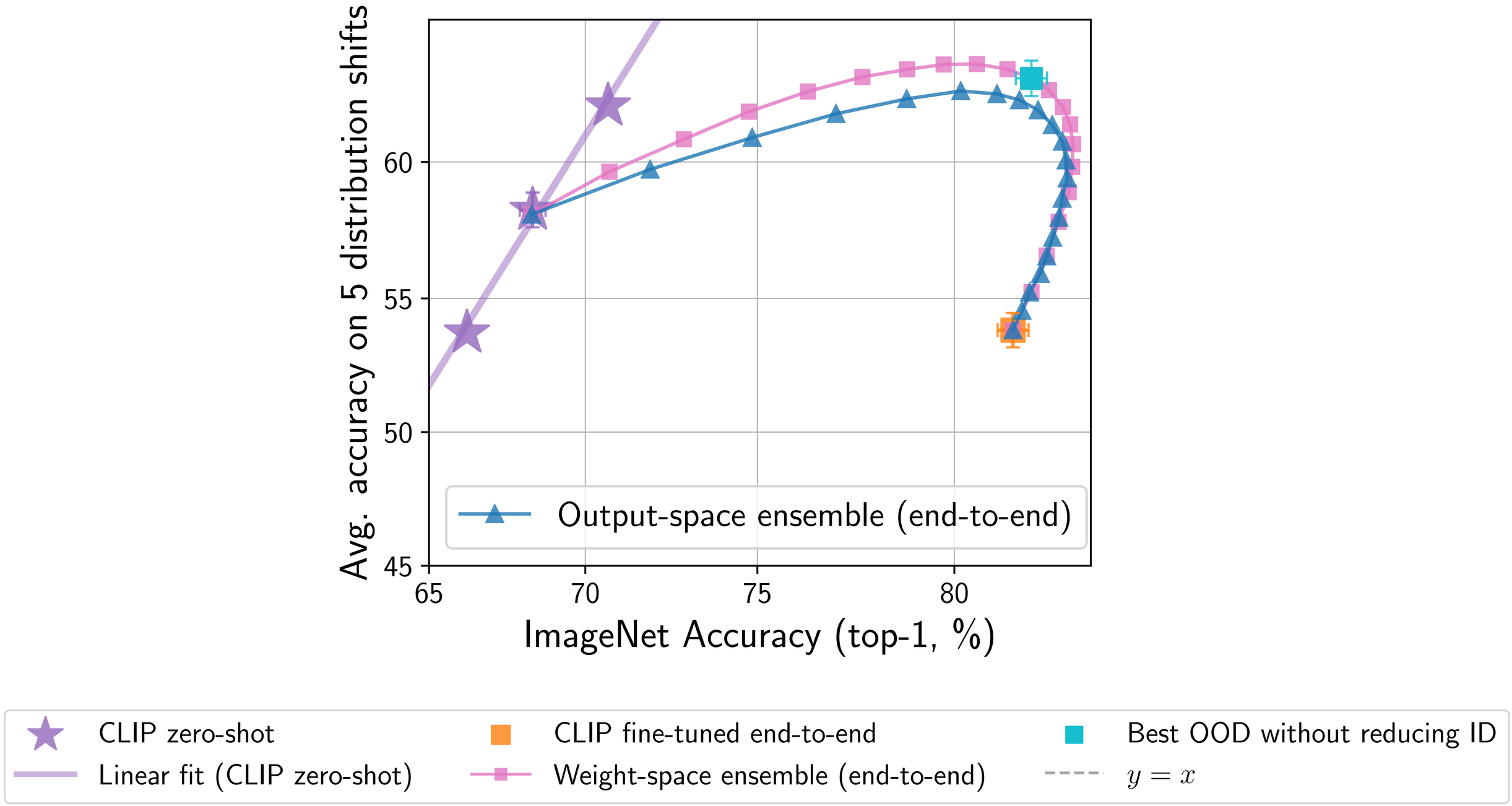
- Baselines
- Choosing the mixing coefficient
- More distribution shifts and datasets
- Why does this work

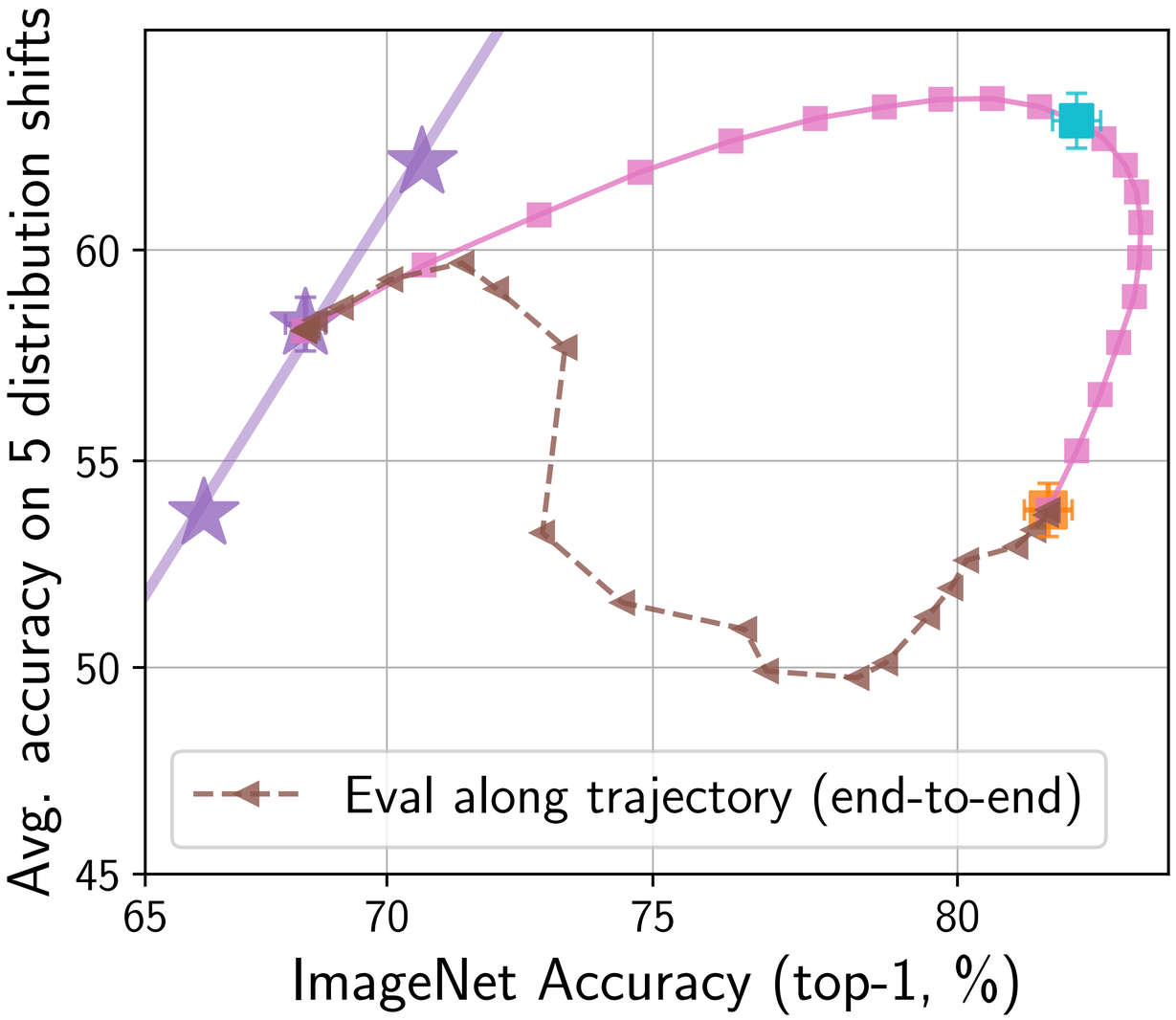

Comparison to linear classifier with regularization
Weight decay
Label smoothing
No regularization
L1 regularization
Avg. accuracy on 5 distribution shifts
Avg. in-distribution accuracy
70
71
72
73
74
75
76
83
84
85
86
87
88
89
90
91
92
CLIP zero-shot
Linear fit (CLIP zero-shot)
CLIP fine-tuned with a linear classifier
Weight-space ensemble (linear classifier)
Best OOD without reducing ID


Comparison to distillation
Avg. accuracy on 5 distribution shifts
Avg. in-distribution accuracy
70
71
72
73
74
75
76
83
84
85
86
87
88
89
90
91
92
Distillation
CLIP zero-shot
Linear fit (CLIP zero-shot)
CLIP fine-tuned with a linear classifier
Weight-space ensemble (linear classifier)
Best OOD without reducing ID


Comparison to regularize to zero-shot
Avg. accuracy on 5 distribution shifts
Avg. in-distribution accuracy
70
71
72
73
74
75
76
83
84
85
86
87
88
89
90
91
92
Regularize to zero-shot


CLIP zero-shot
Linear fit (CLIP zero-shot)
CLIP fine-tuned with a linear classifier
Weight-space ensemble (linear classifier)
Best OOD without reducing ID


Avg. accuracy on 5 distribution shifts
45
50
55
60
65
70
75
80
ImageNet Accuracy (top-1, %)
Output-space ensemble (end-to-end)
CLIP zero-shot
Linear fit (CLIP zero-shot)
CLIP fine-tuned end-to-end
Weight-space ensemble (end-to-end)
Best OOD without reducing ID
$y = x$



CLIP zero-shot
CLIP fine-tuned end-to-end
Best OOD without reducing ID
Linear fit (CLIP zero-shot)
Weight-space ensemble (end-to-end)
$y = x$

Outline

- Baselines
- Choosing the mixing coefficient
- More distribution shifts and datasets
- Why does this work

$\alpha = 0.4$

ImageNet (Deng et al.)

+0.0

ImageNetV2 (Recht et al.)

+2.4

ImageNet-R (Hendrycks et al.)

+10.1

ImageNet Sketch (Wang et al.)

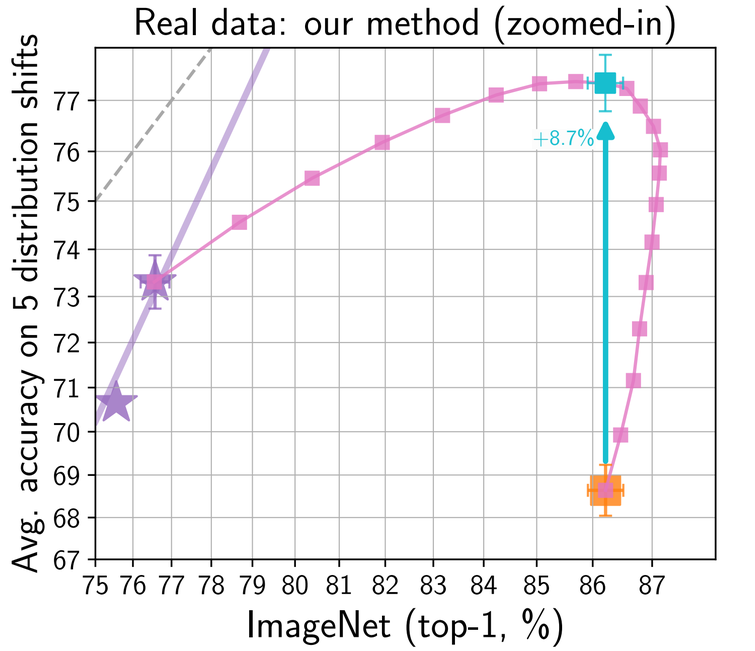
+7.1

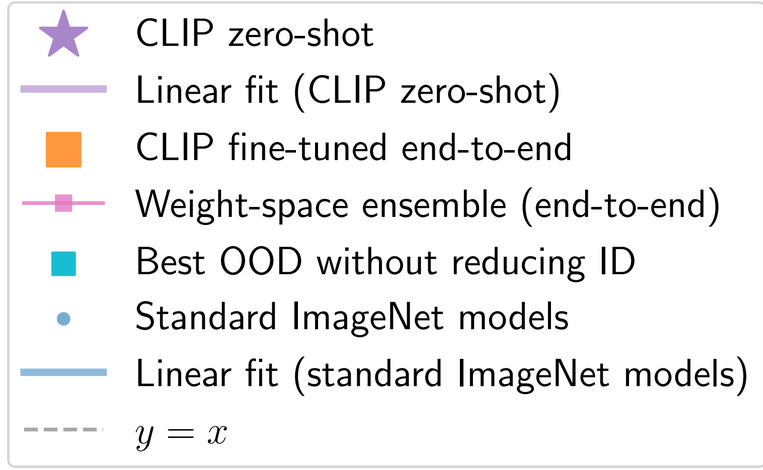
ObjectNet (Barbu et al.)

+8.6

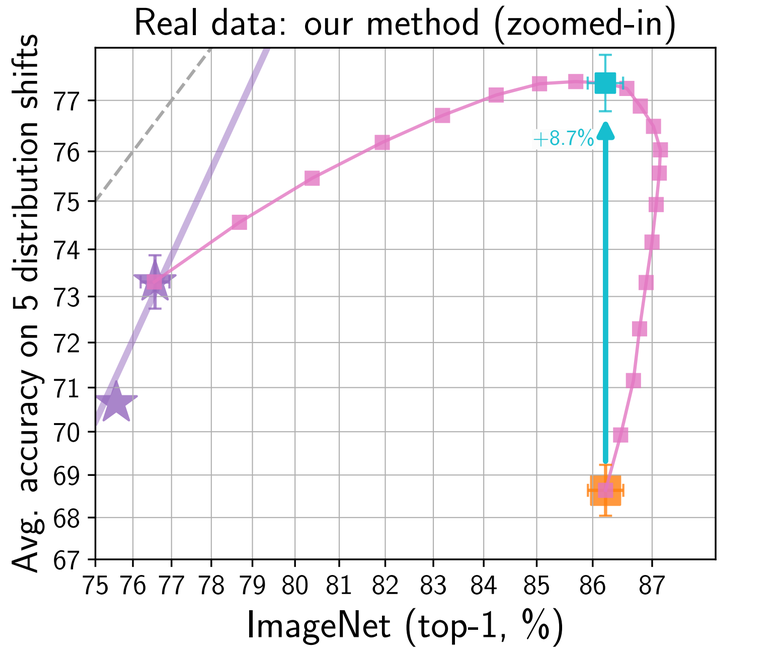
ImageNet-A (Hendrycks et al.)

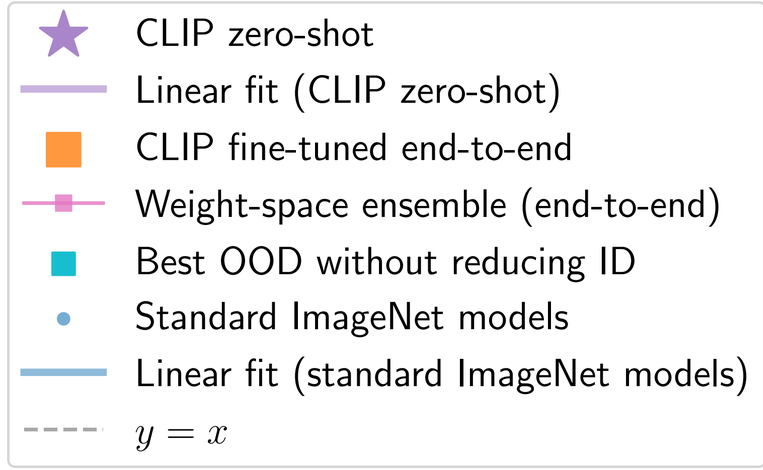
+15.3





	$\alpha = 0.4$	$\alpha = 0.75$
ImageNet (Deng et al.) 	+0.0	+0.8
ImageNetV2 (Recht et al.) 	+2.4	+2.0
ImageNet-R (Hendrycks et al.) 	+10.1	+6.3
ImageNet Sketch (Wang et al.) 	+7.1	+4.6
ObjectNet (Barbu et al.) 	+8.6	+4.8
ImageNet-A (Hendrycks et al.) 	+15.3	+9.8





Outline

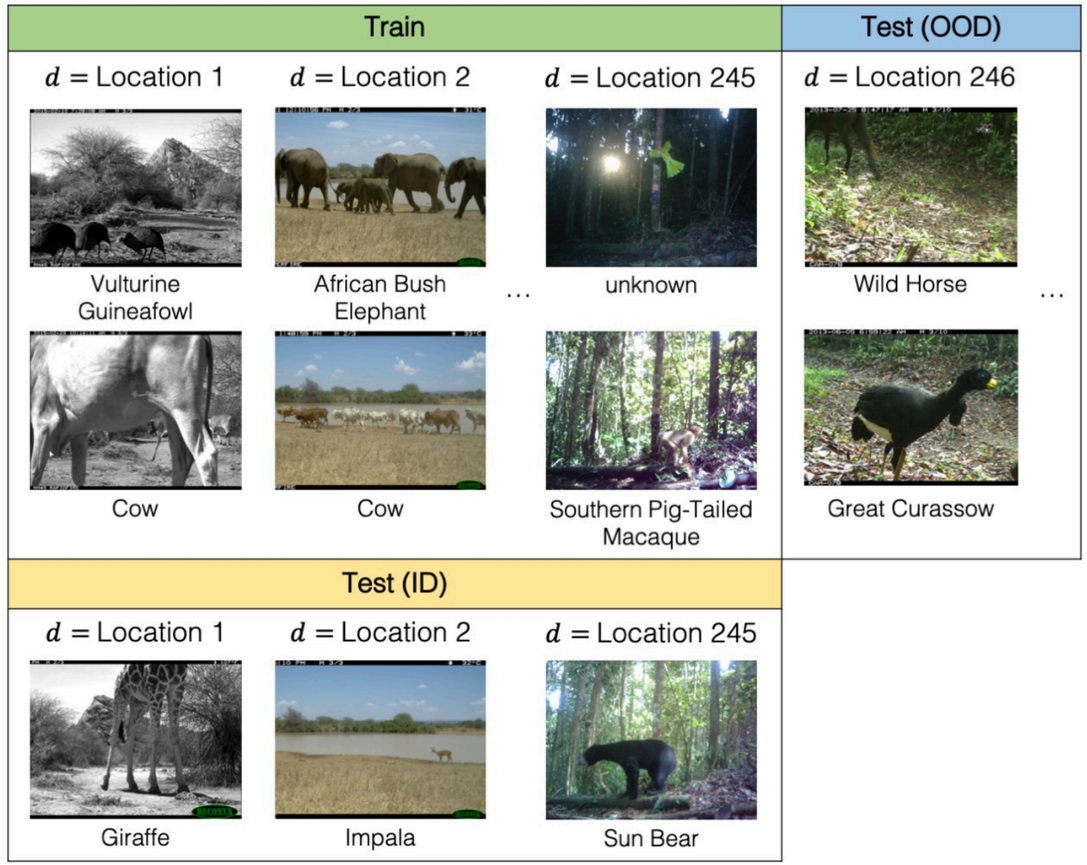
- Baselines
- Choosing the mixing coefficient
- More distribution shifts and datasets
- Why does this work

Koh et al., 2021

WILDS

Koh et al., 2021

iWildCam

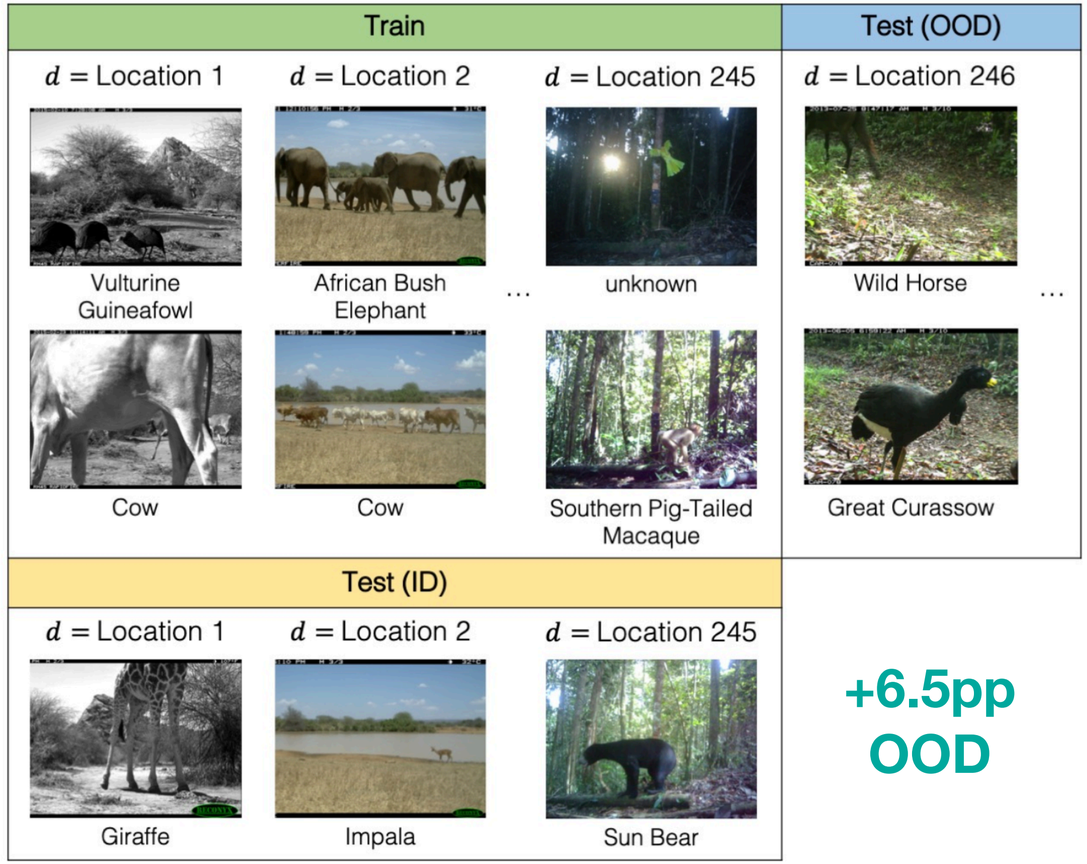


Beery et al., 2018

WILDS

Koh et al., 2021

iWildCam



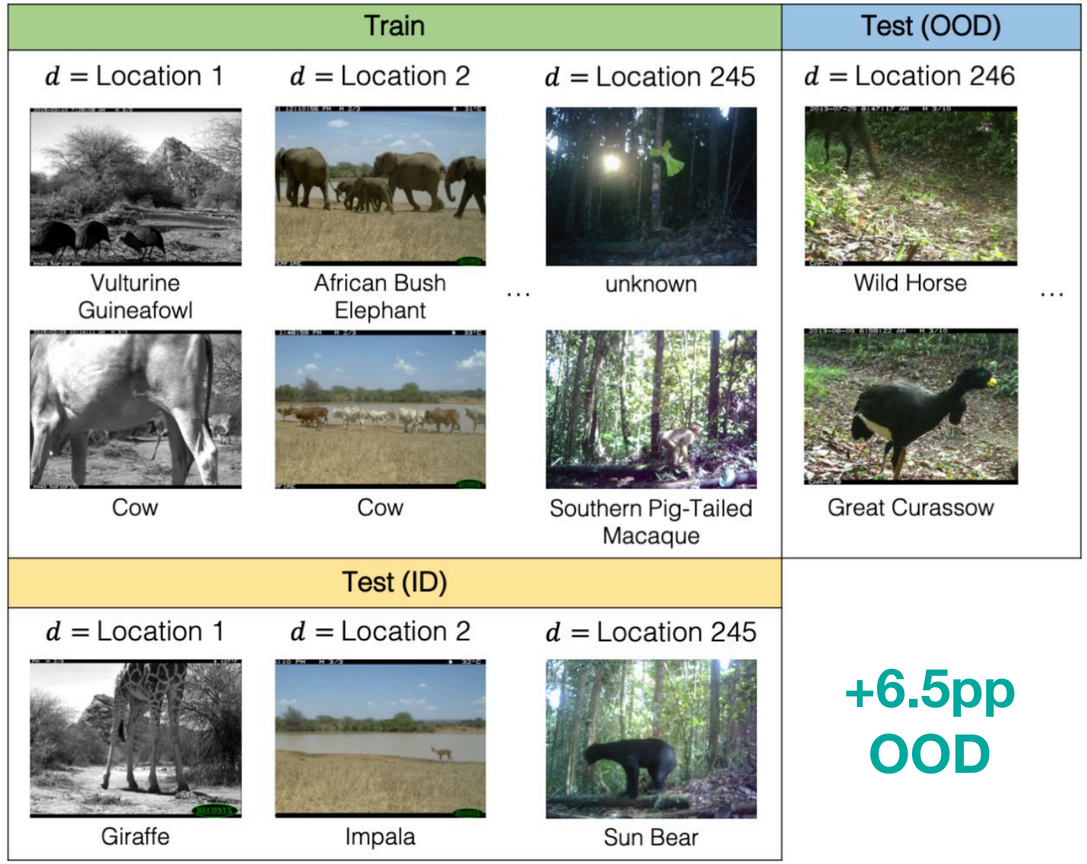
+6.5pp OOD

Beery et al., 2018

WILDS

Koh et al., 2021

iWildCam

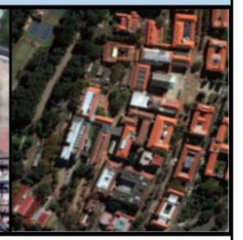


+6.5pp OOD

Beery et al., 2018

FMoW

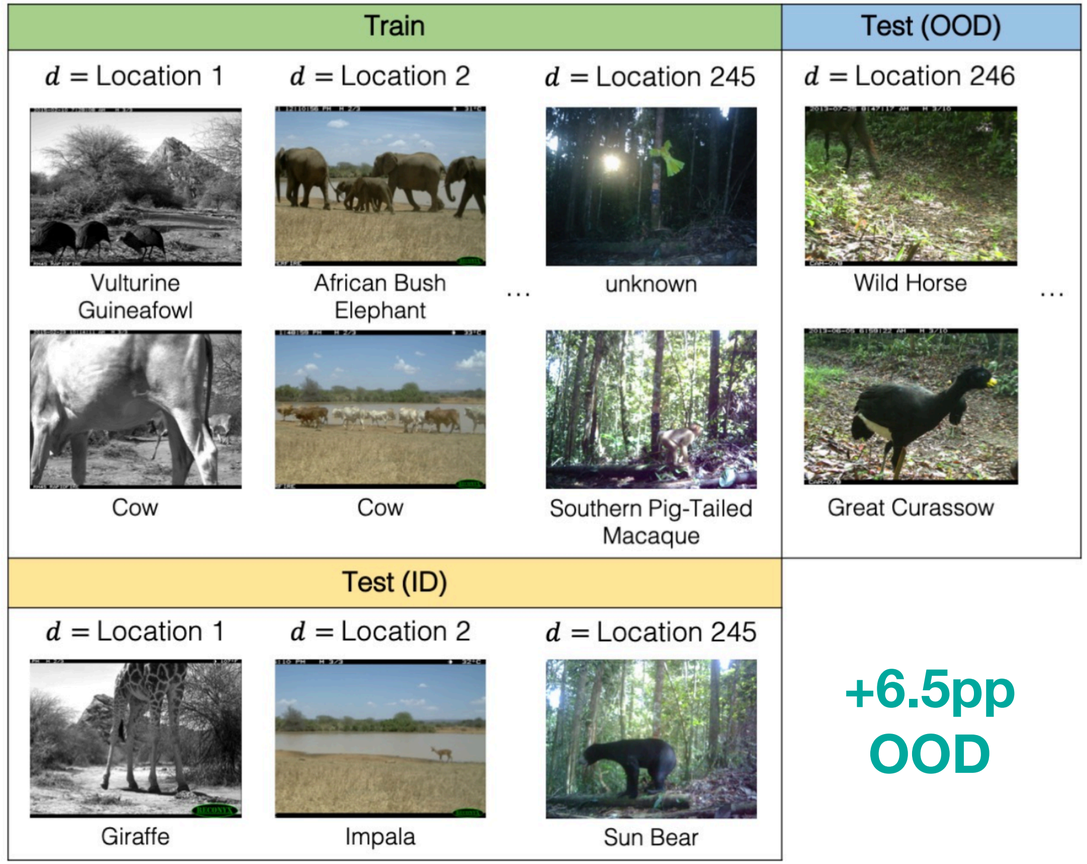
Christie et al., 2018

	Train			Test	
Satellite Image (x)					
Year / Region (d)	2002 / Americas	2009 / Africa	2012 / Europe	2016 / Americas	2017 / Africa
Building / Land Type (y)	shopping mall	multi-unit residential	road bridge	recreational facility	educational institution

WILDS

Koh et al., 2021

iWildCam



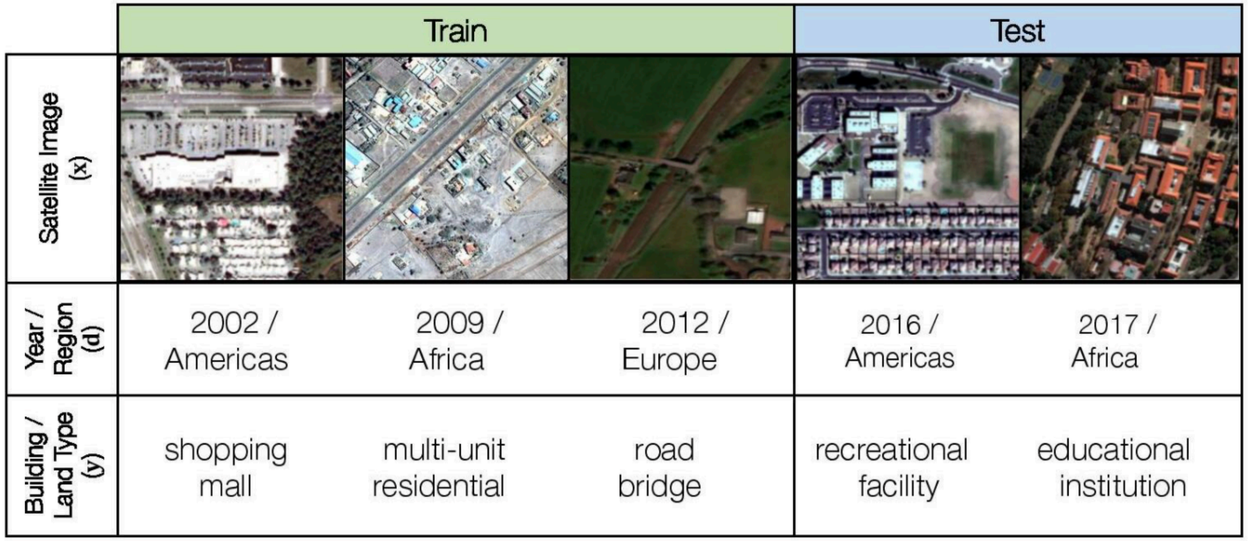
+6.5pp OOD

Beery et al., 2018

FMoW

+3.7pp OOD

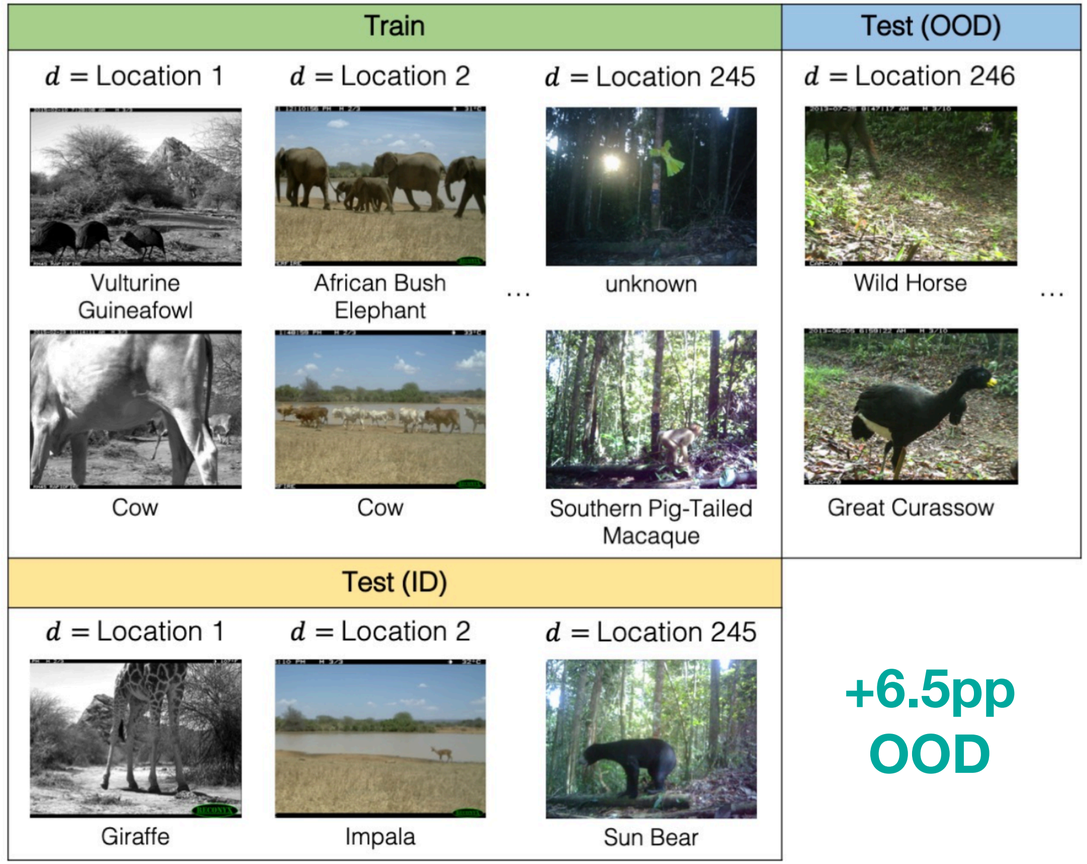
Christie et al., 2018



WILDS

Koh et al., 2021

iWildCam



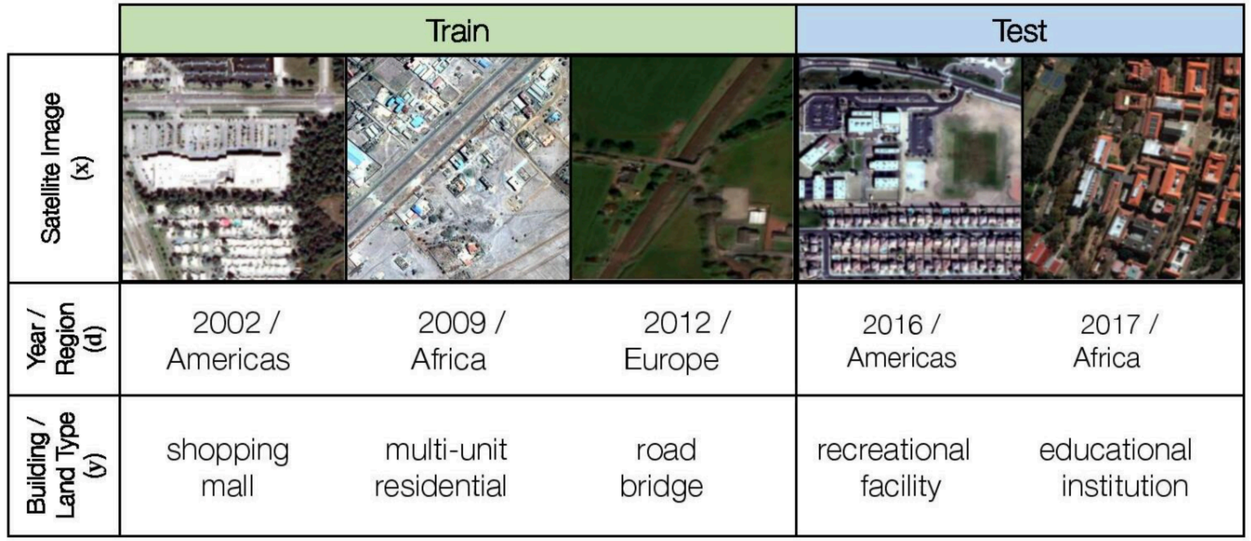
+6.5pp OOD

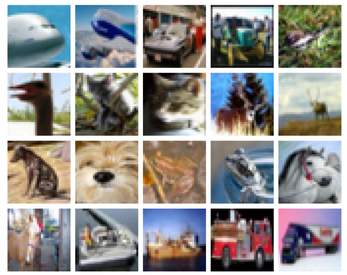
Beery et al., 2018

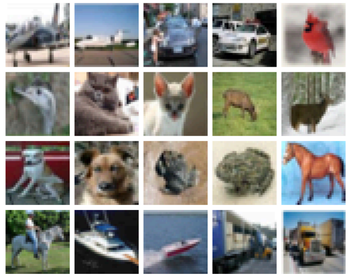
FMoW

+3.7pp OOD

Christie et al., 2018







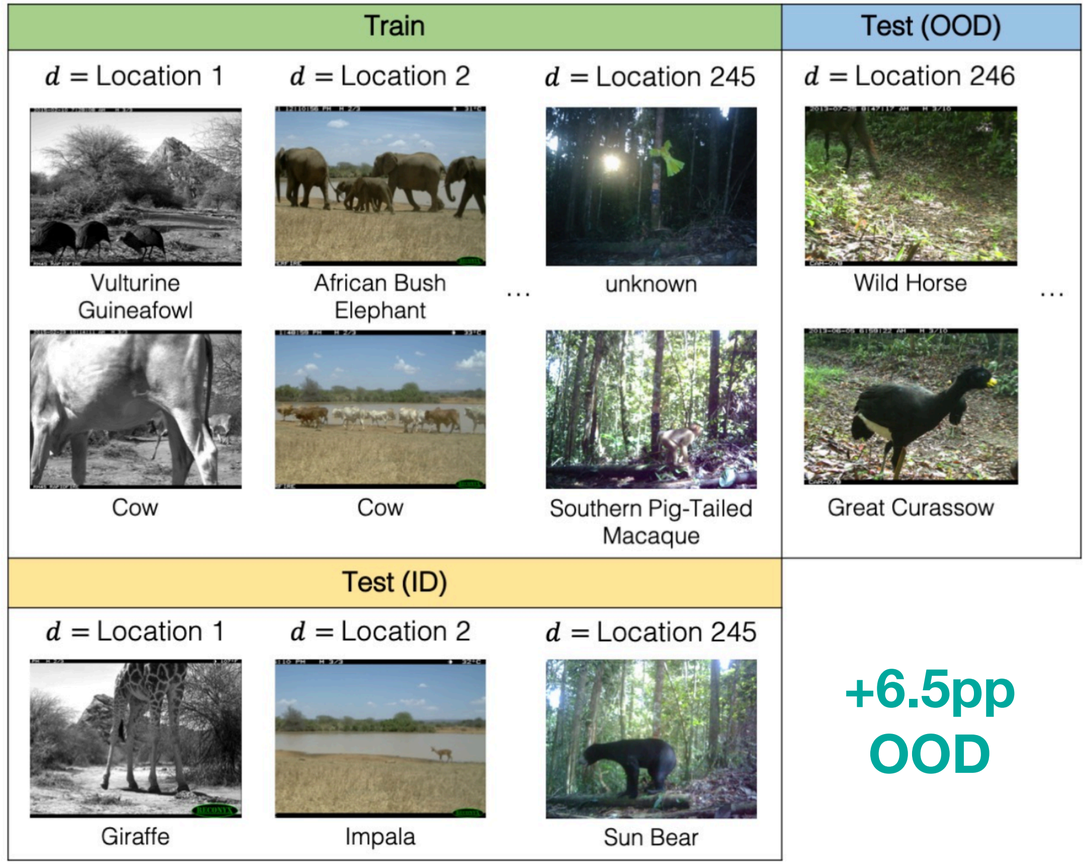
CIFAR-10.1.
Recht et al., 2019

CIFAR-10.2.
Lu et al., 2020

WILDS

Koh et al., 2021

iWildCam



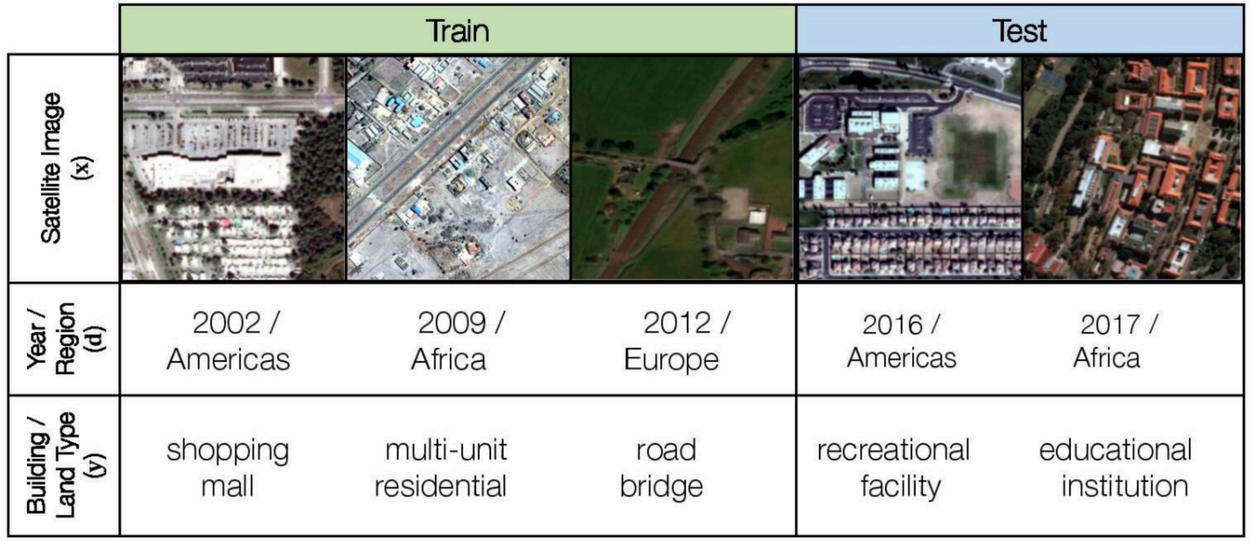
+6.5pp OOD

Beery et al., 2018

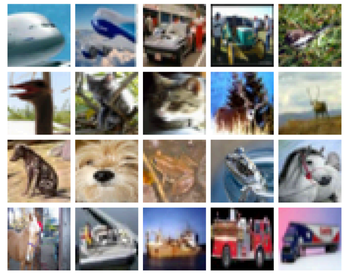
FMoW

+3.7pp OOD

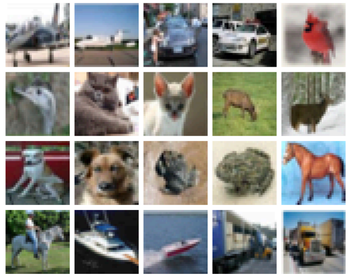
Christie et al., 2018



+2.2pp OOD



+3.0pp OOD



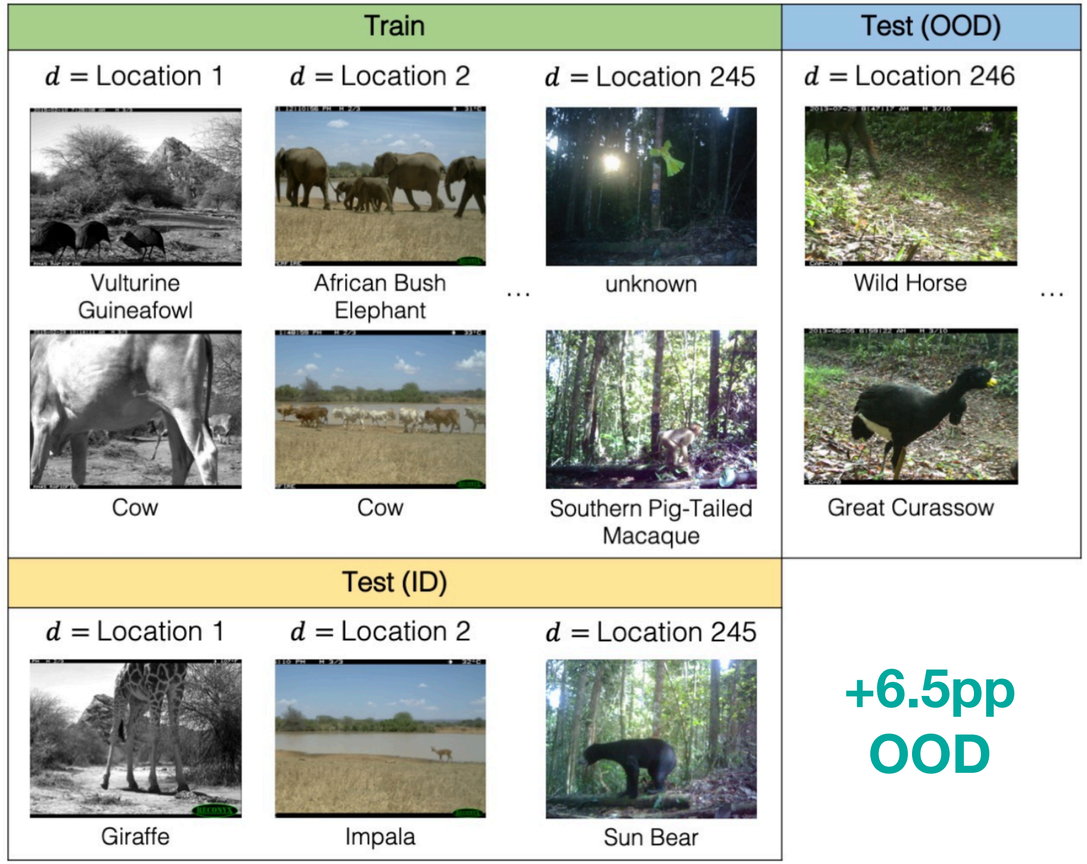
CIFAR-10.1.
Recht et al., 2019

CIFAR-10.2.
Lu et al., 2020

WILDS

Koh et al., 2021

iWildCam



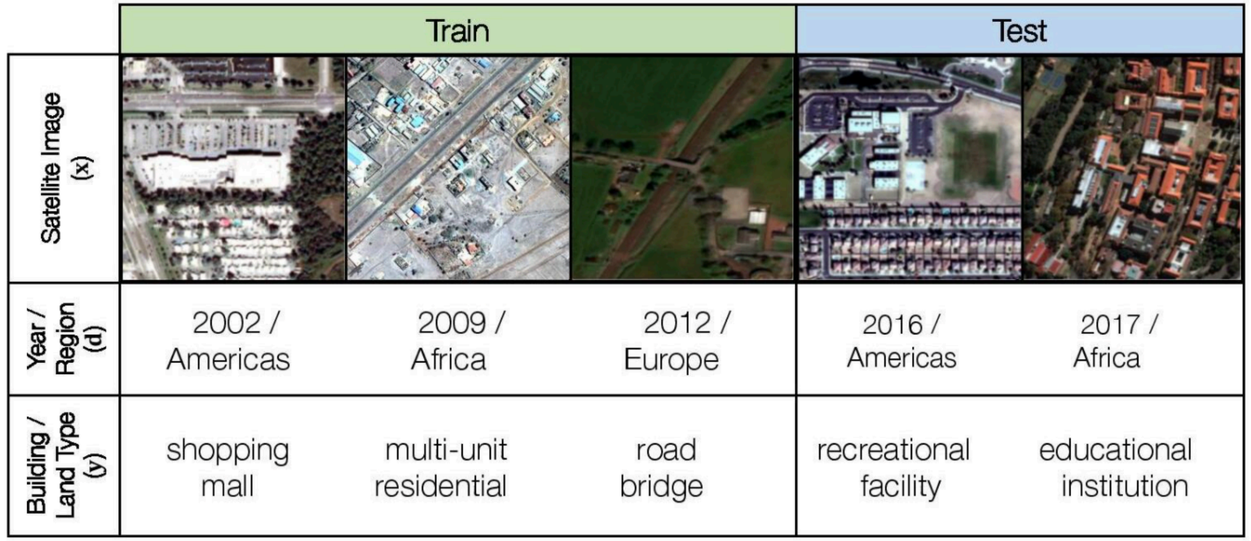
+6.5pp OOD

Beery et al., 2018

FMoW

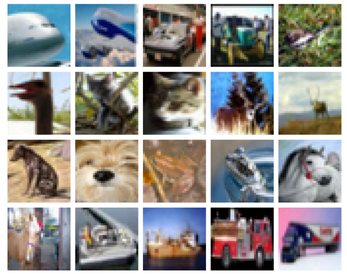
+3.7pp OOD

Christie et al., 2018



CIFAR-10.1.

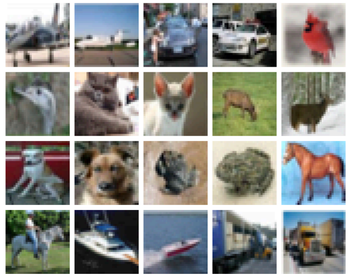
+2.2pp OOD



Recht et al., 2019

CIFAR-10.2.

+3.0pp OOD



Lu et al., 2020

ImageNet-Vid-Robust

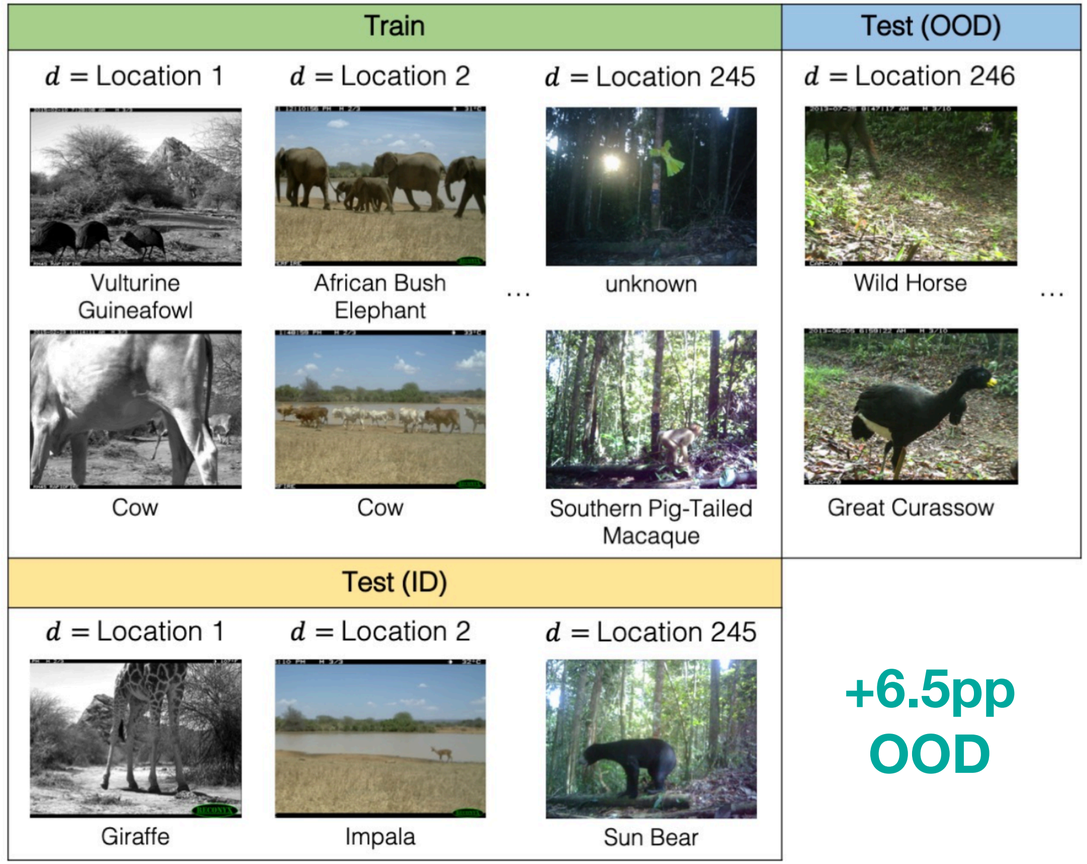
Shankar et al., 2019

YTBBRobust

WILDS

Koh et al., 2021

iWildCam



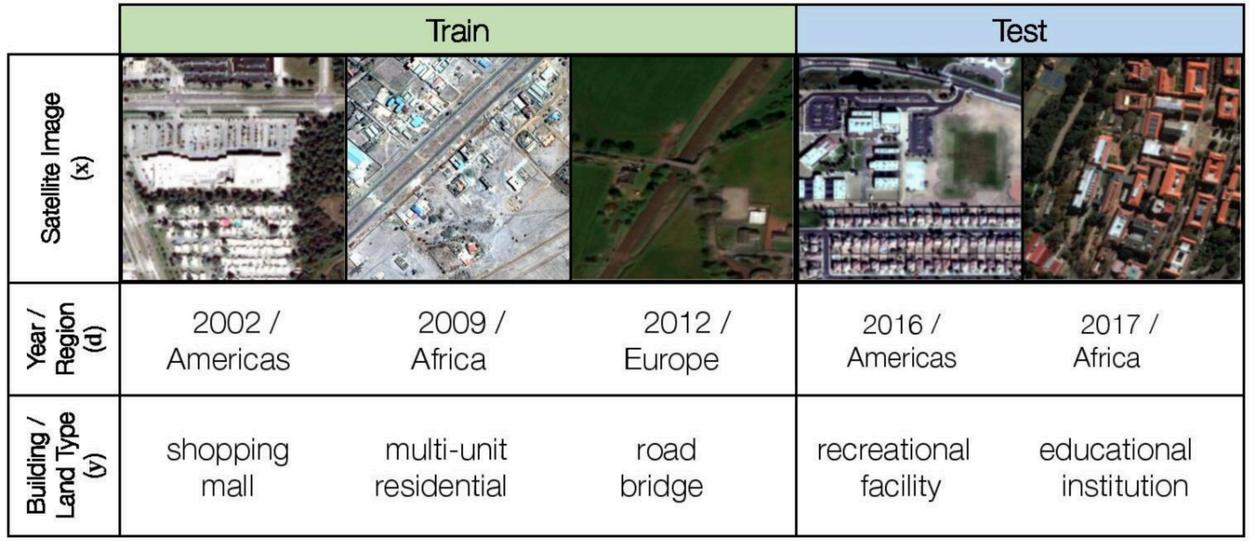
+6.5pp OOD

Beery et al., 2018

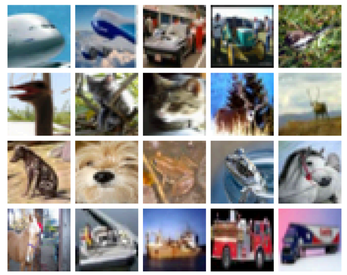
FMoW

+3.7pp OOD

Christie et al., 2018

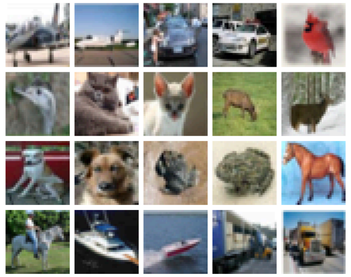


+2.2pp OOD



CIFAR-10.1.
Recht et al., 2019

+3.0pp OOD



CIFAR-10.2.
Lu et al., 2020

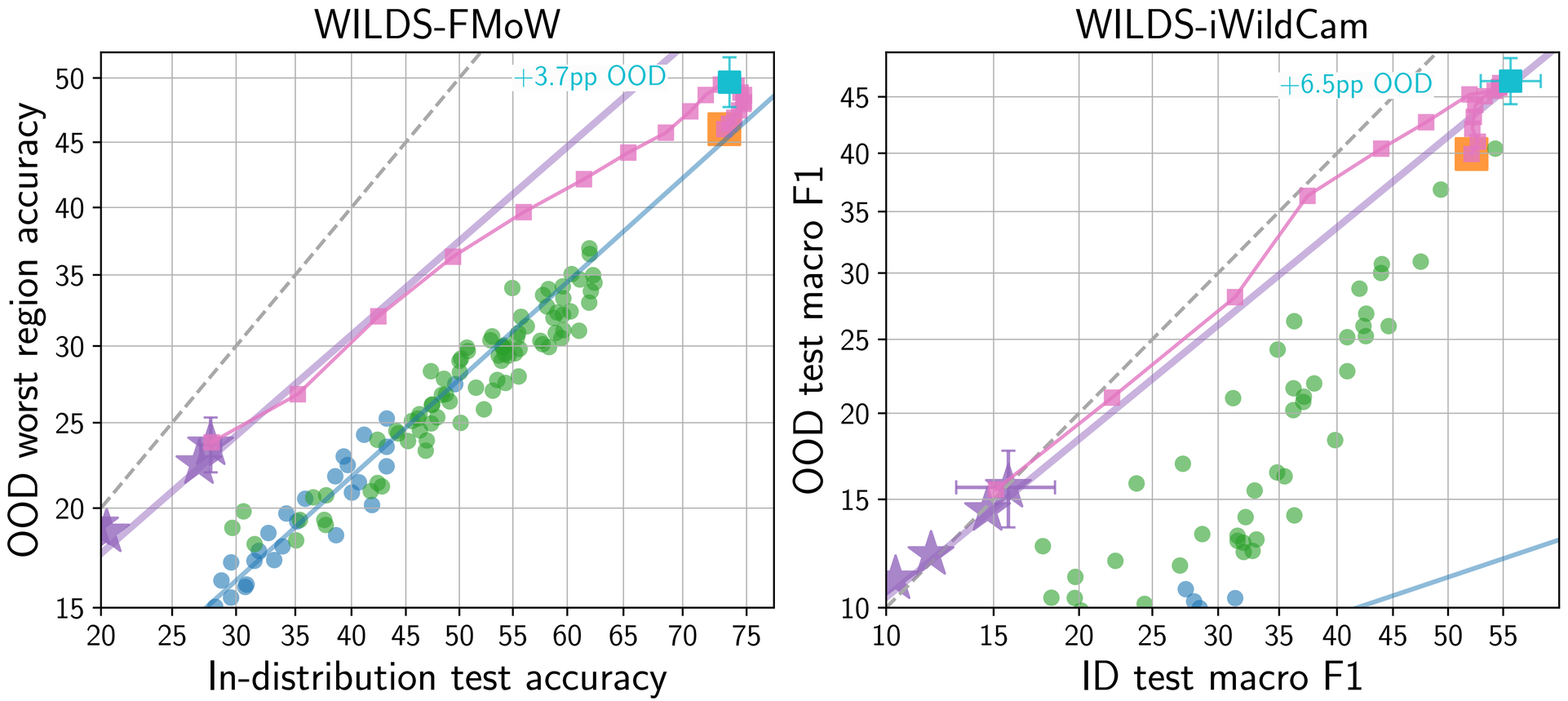
+8.3pp OOD

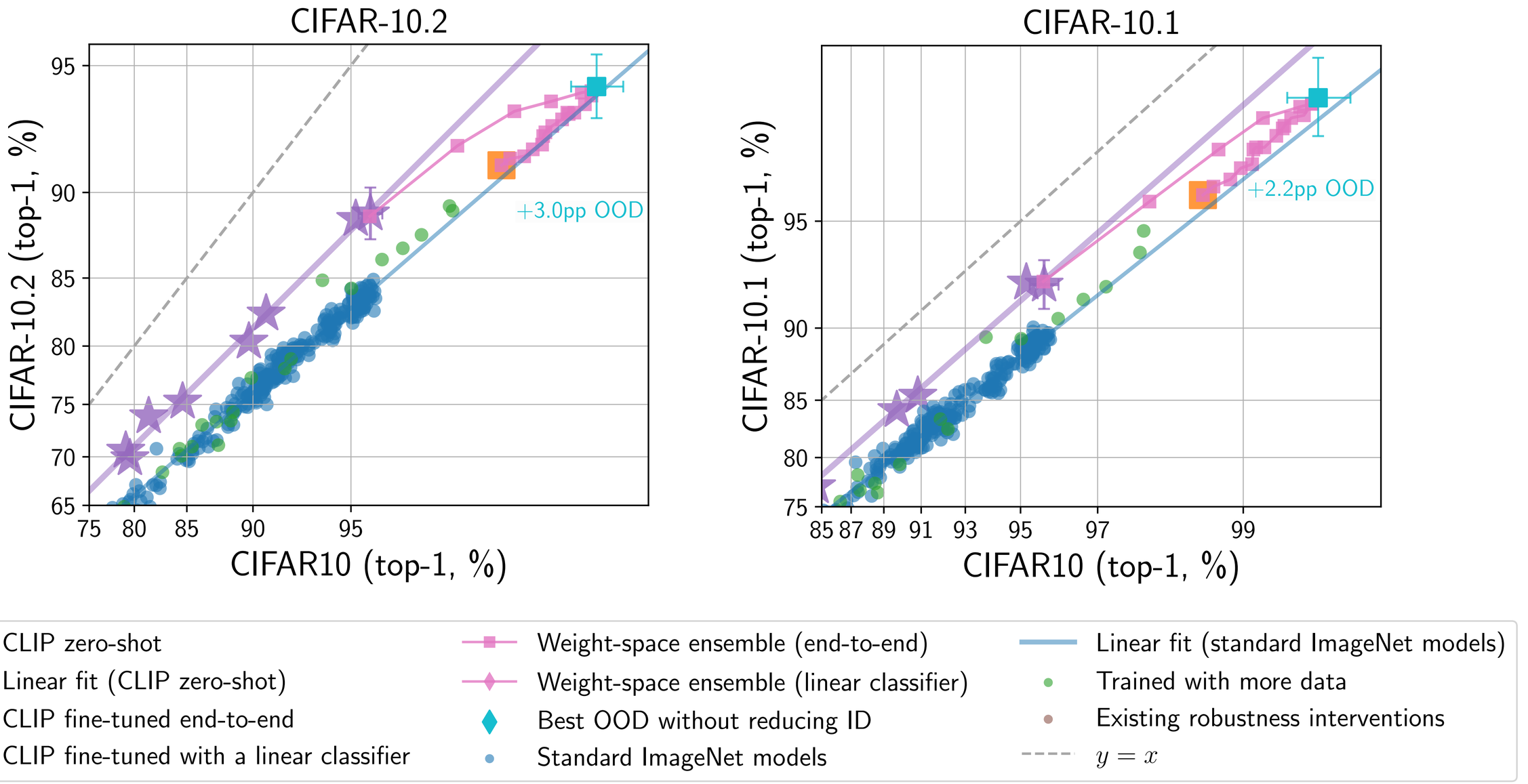
ImageNet-Vid-Robust

Shankar et al., 2019

YTBBRobust

+14.7pp OOD


WILDS-FMoW
+3.7pp OOD
OOD worst region accuracy
In-distribution test accuracy
WILDS-iWildCam
+6.5pp OOD
OOD test macro F1
ID test macro F1
CLIP zero-shot
Linear fit (CLIP zero-shot)
CLIP fine-tuned end-to-end
CLIP fine-tuned with a linear classifier
Weight-space ensemble (end-to-end)
Weight-space ensemble (linear classifier)
Best OOD without reducing ID
Standard ImageNet models
Linear fit (standard ImageNet models)
Trained with more data
Existing robustness interventions
$y = x$

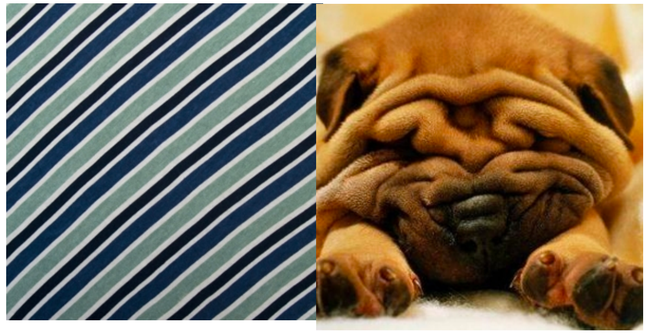

CIFAR-10.2
CIFAR-10.2 (top-1, %)
CIFAR10 (top-1, %)
+3.0pp OOD
CIFAR-10.1
CIFAR-10.1 (top-1, %)
CIFAR10 (top-1, %)
+2.2pp OOD
CLIP zero-shot
Linear fit (CLIP zero-shot)
CLIP fine-tuned end-to-end
CLIP fine-tuned with a linear classifier
Weight-space ensemble (end-to-end)
Weight-space ensemble (linear classifier)
Best OOD without reducing ID
Standard ImageNet models
Linear fit (standard ImageNet models)
Trained with more data
Existing robustness interventions
$y = x$

ImageNet (Deng et al., 2009)

CIFAR-10 & CIFAR-100 (Krizhevsky et al., 2009)

Stanford Cars (Krause et al., 2013)

Describable Textures (Cimpoi et al., 2014)



SUN397 (Xiao et al., 2016)

/s/subway_station/platform (337)

/c/control_tower/outdoor (110)

/b/bazaar/indoor (44)

/r/restaurant_kitchen (301)

/d/driveway (138)

/a/auto_factory (23)

Food101 (Bossard et al., 2014)

ImageNet (Deng et al., 2009)

+0.9pp ID

CIFAR-10 & CIFAR-100 (Krizhevsky et al., 2009)

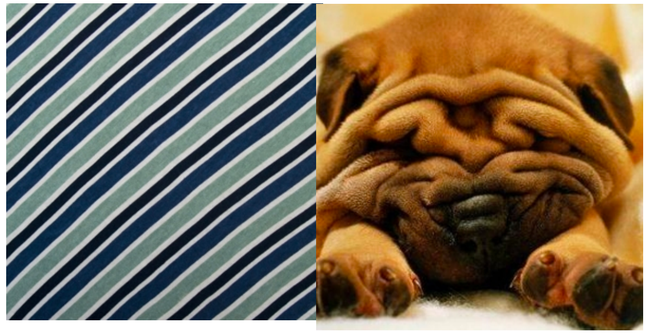
+0.8pp ID

+1.2pp ID

Stanford Cars (Krause et al., 2013)

+2.0pp ID

Describable Textures (Cimpoi et al., 2014)



+3.3pp ID

SUN397 (Xiao et al., 2016)

/s/subway_station/platform (337)

/c/control_tower/outdoor (110)

/b/bazaar/indoor (44)

/r/restaurant_kitchen (301)

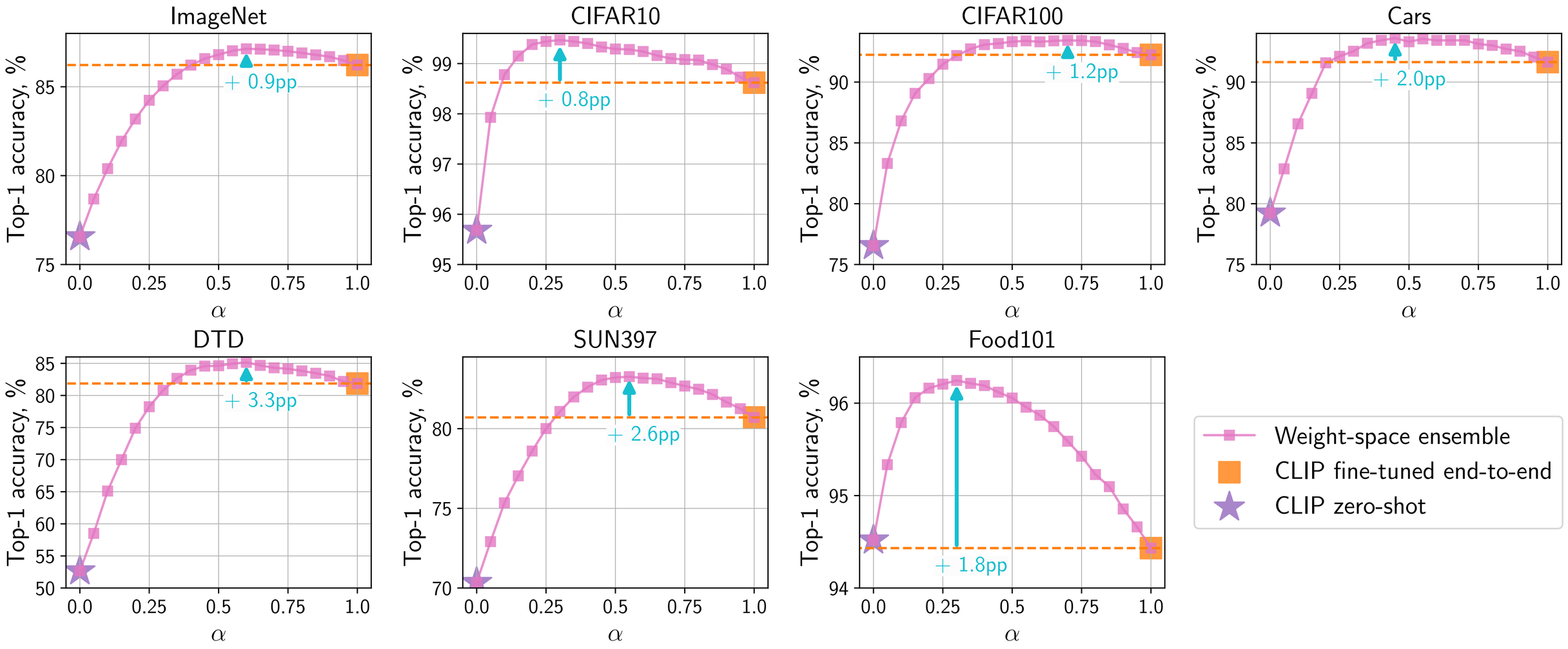
/d/driveway (138)

/a/auto_factory (23)

+2.6pp ID

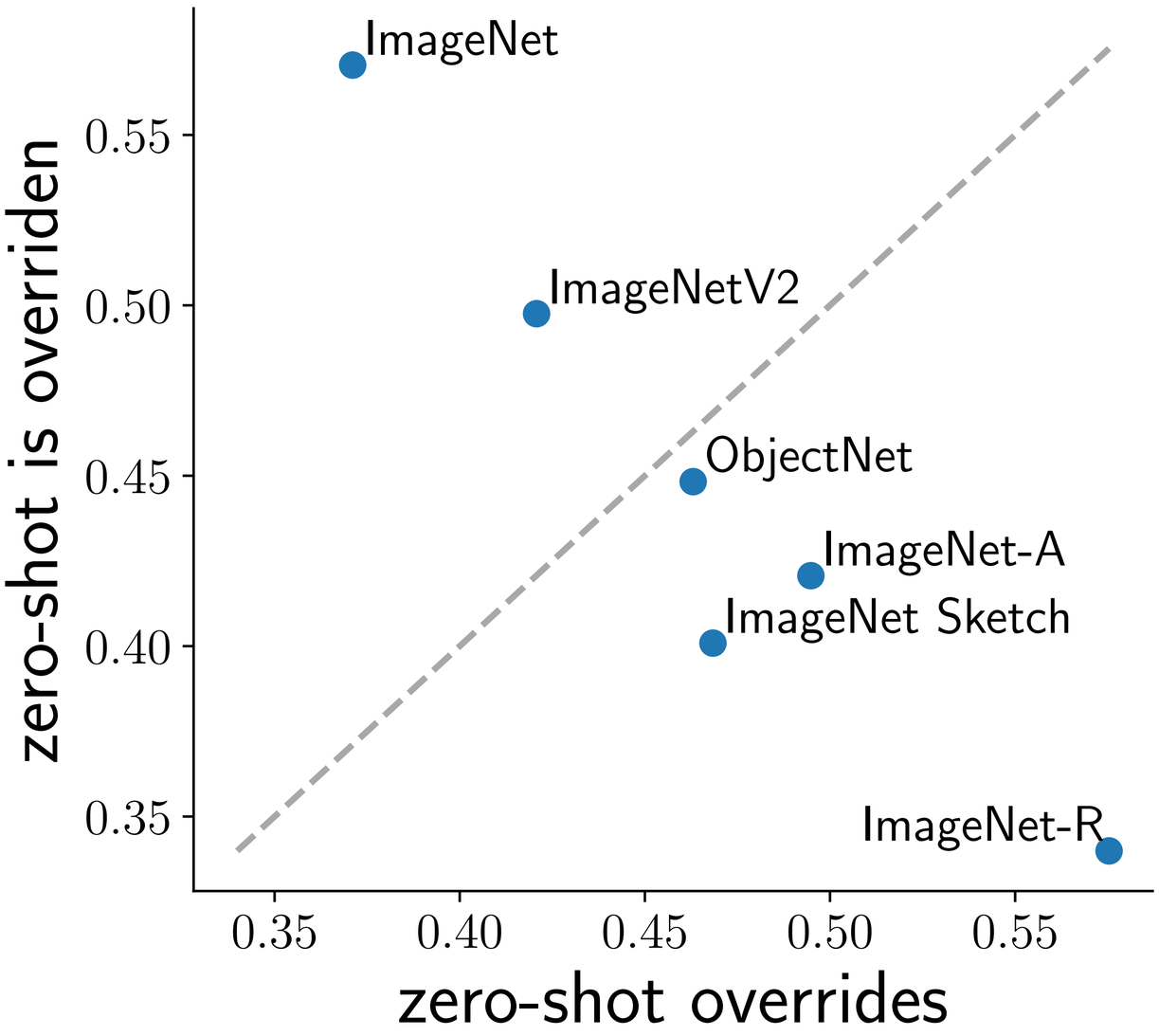
Food101 (Bossard et al., 2014)

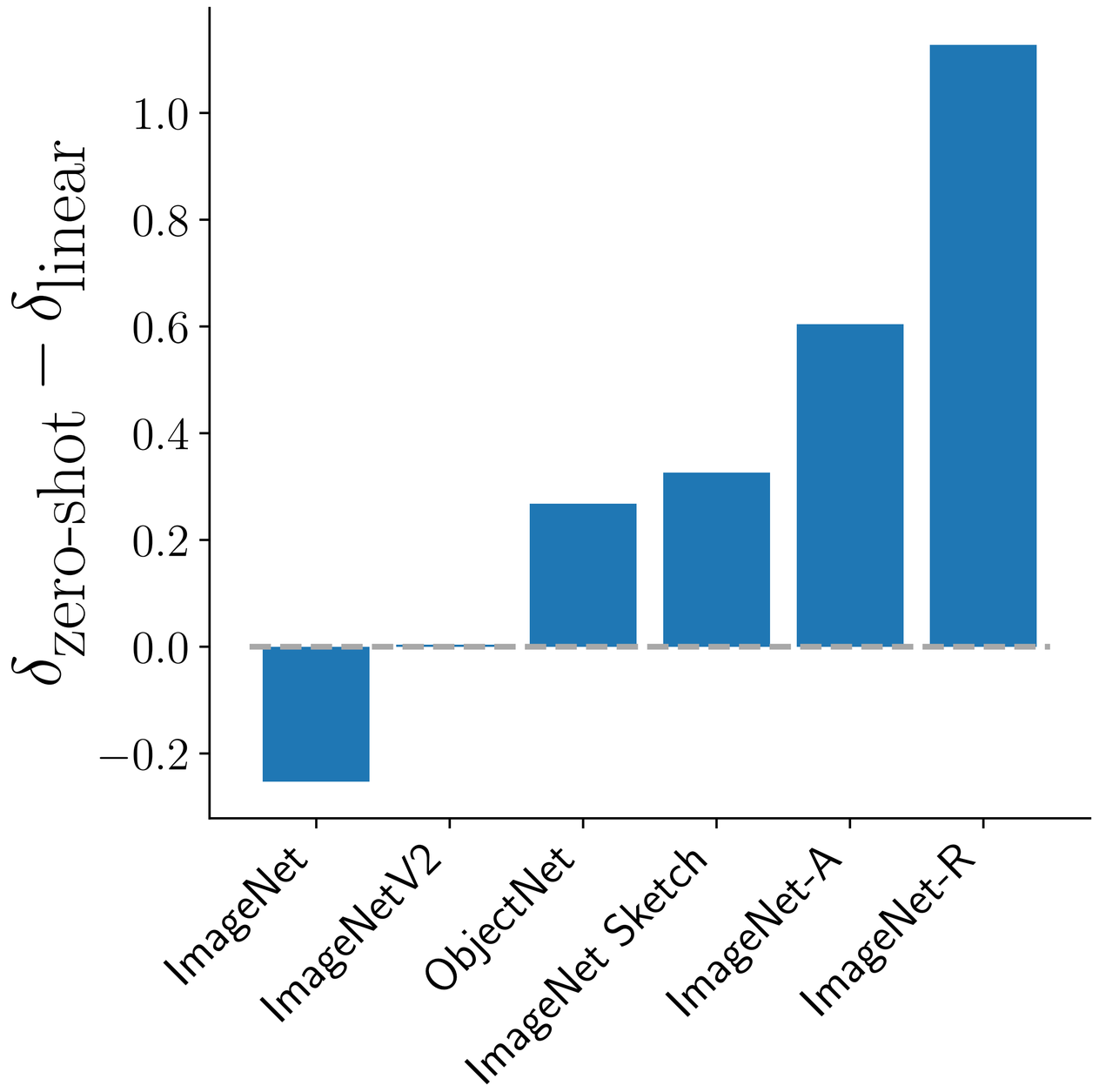
+1.8pp ID


ImageNet
CIFAR10
CIFAR100
Cars
DTD
SUN397
Food101
Top-1 accuracy, %
α
+ 0.9pp
+ 0.8pp
+ 1.2pp
+ 2.0pp
+ 3.3pp
+ 2.6pp
+ 1.8pp
Weight-space ensemble
CLIP fine-tuned end-to-end
CLIP zero-shot

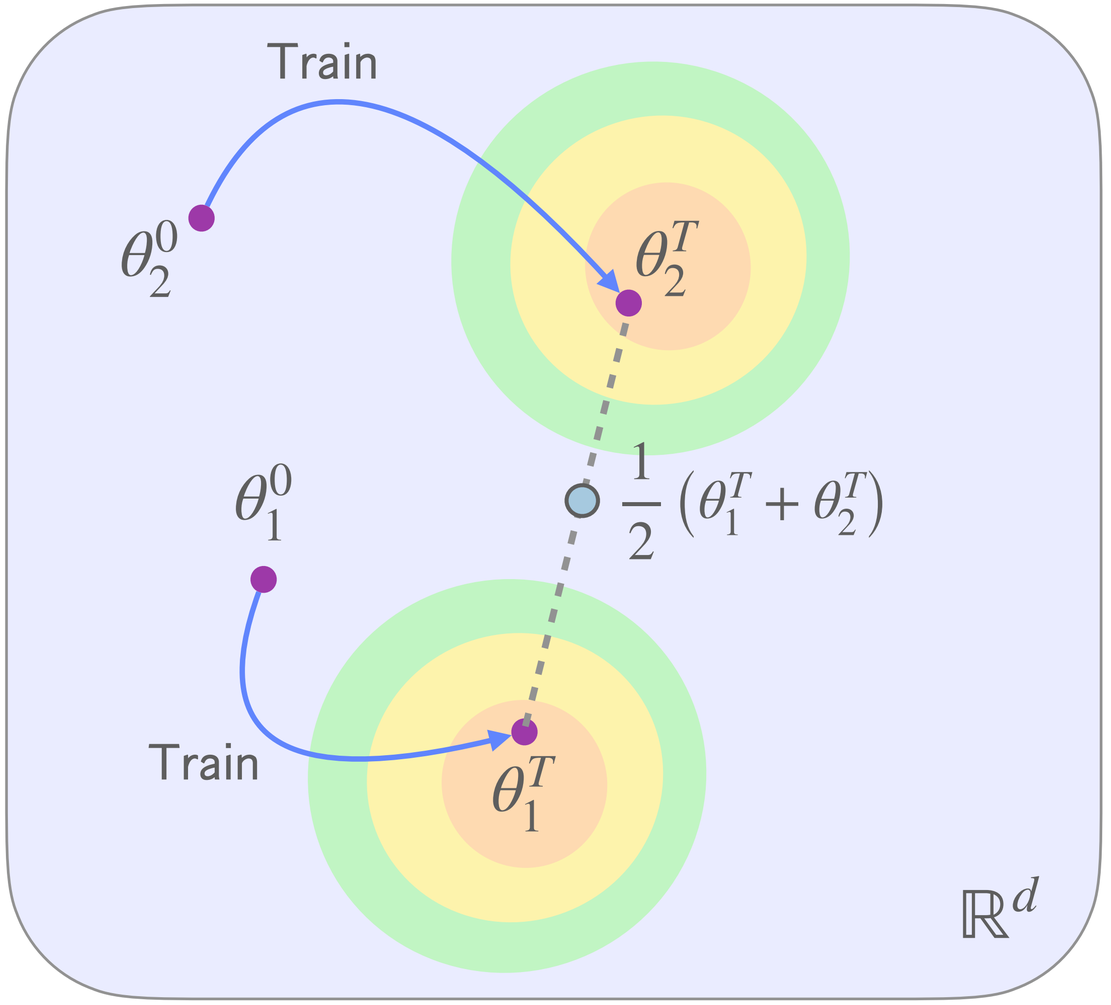
Outline

- Baselines
- Choosing the mixing coefficient
- More distribution shifts and datasets
- Why does this work


ImageNet
ImageNetV2
ObjectNet
ImageNet-A
ImageNet Sketch
ImageNet-R
0.55
0.50
0.45
0.40
0.35
0.35
0.40
0.45
0.50
0.55
zero-shot is overriden
zero-shot overrides


$\delta_{\text{zero-shot}} - \delta_{\text{linear}}$
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
ImageNet
ImageNetV2
ObjectNet
ImageNet Sketch
ImageNet-A
ImageNet-R

Observation 1: (Frankle et al., 2020)

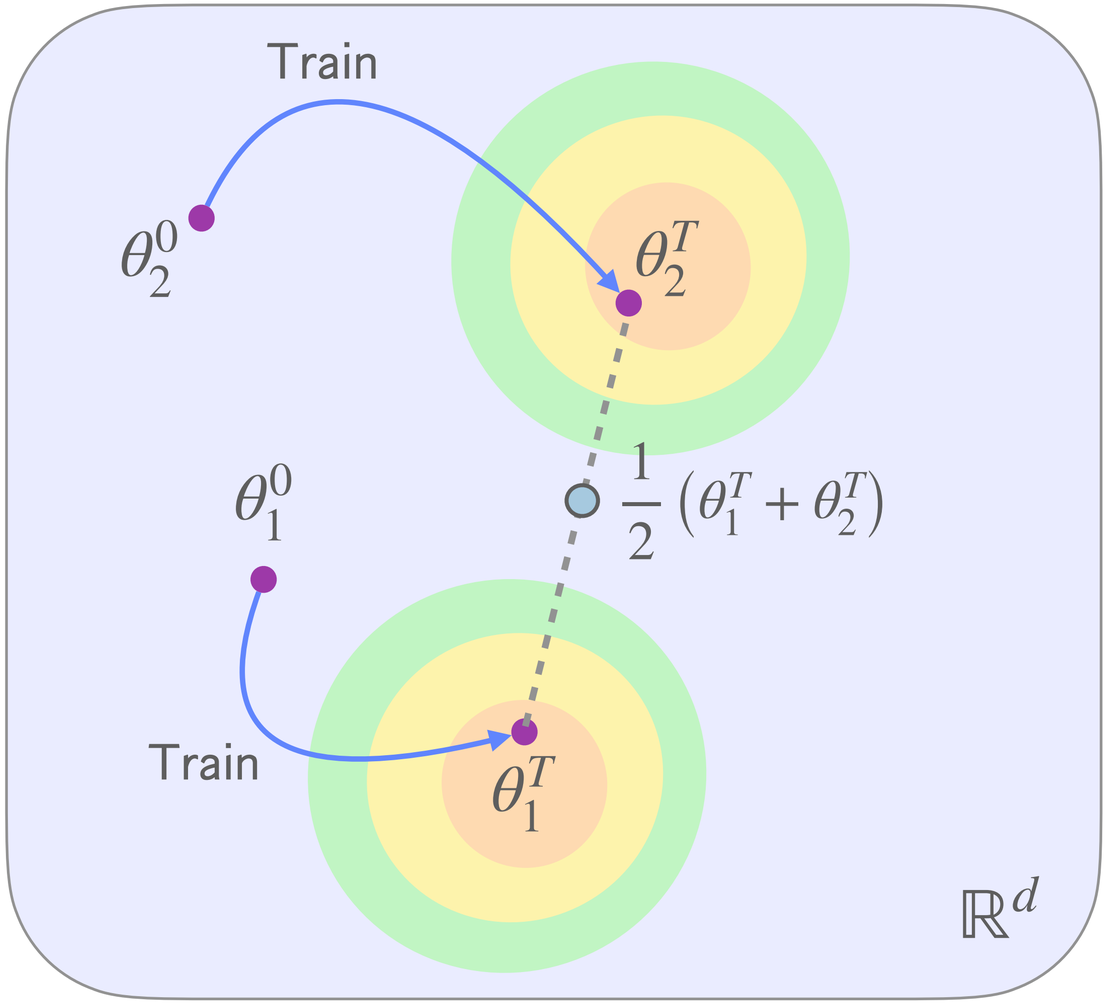


Observation 1: (Frankle et al., 2020)

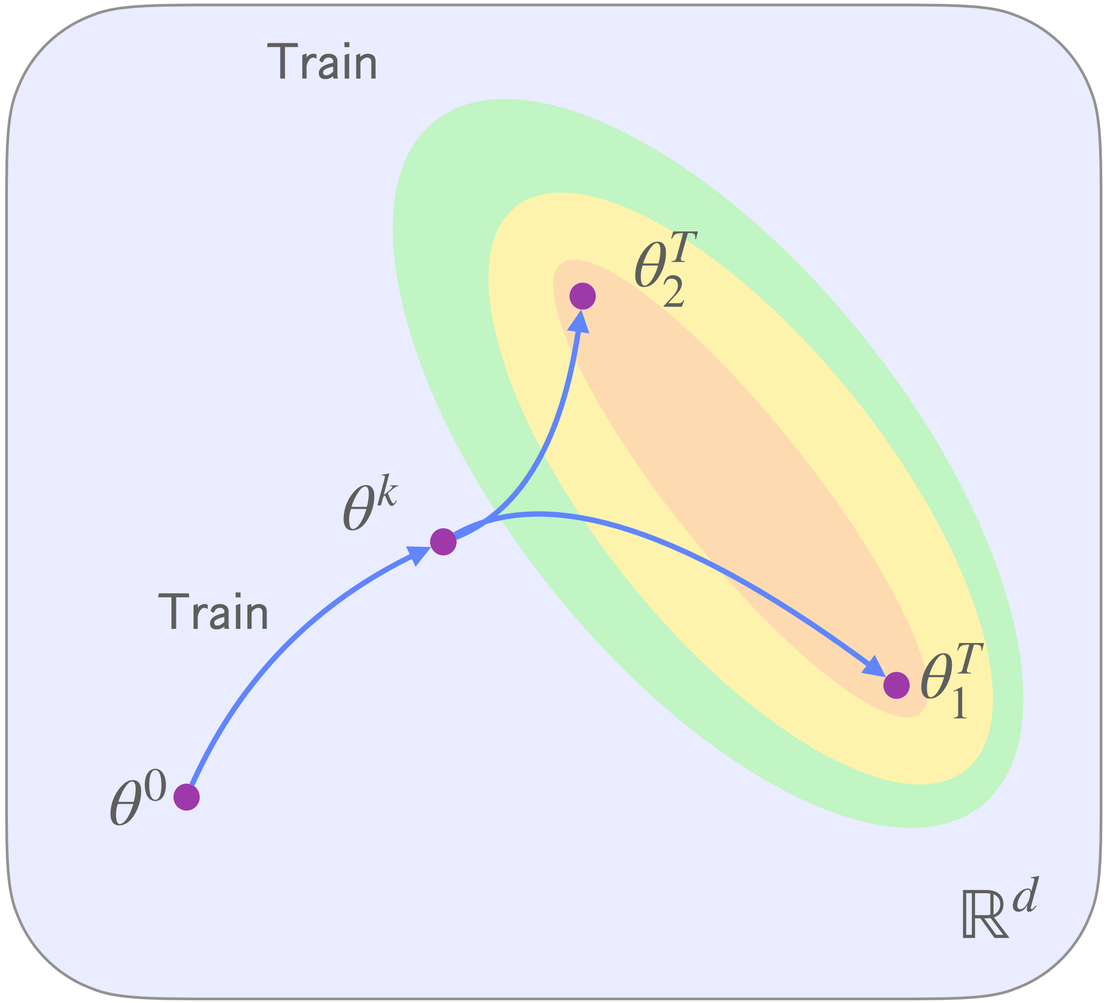
Ensembling in weight space — making predictions via

$$f\left(x, \frac{1}{2}\left(\theta_1^T + \theta_2^T\right)\right)$$

fails, performing no better than random chance.

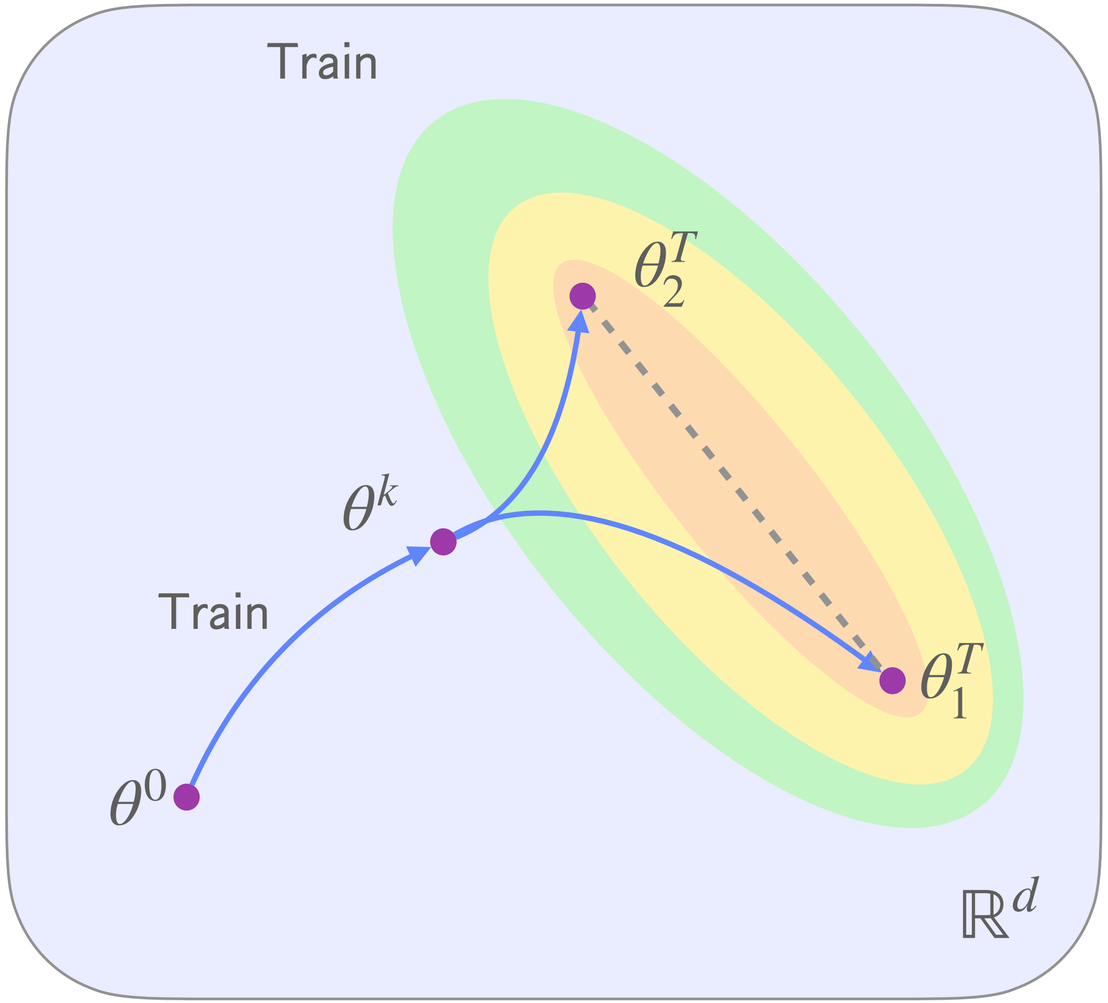


Observation 2: (Frankle et al., 2020)



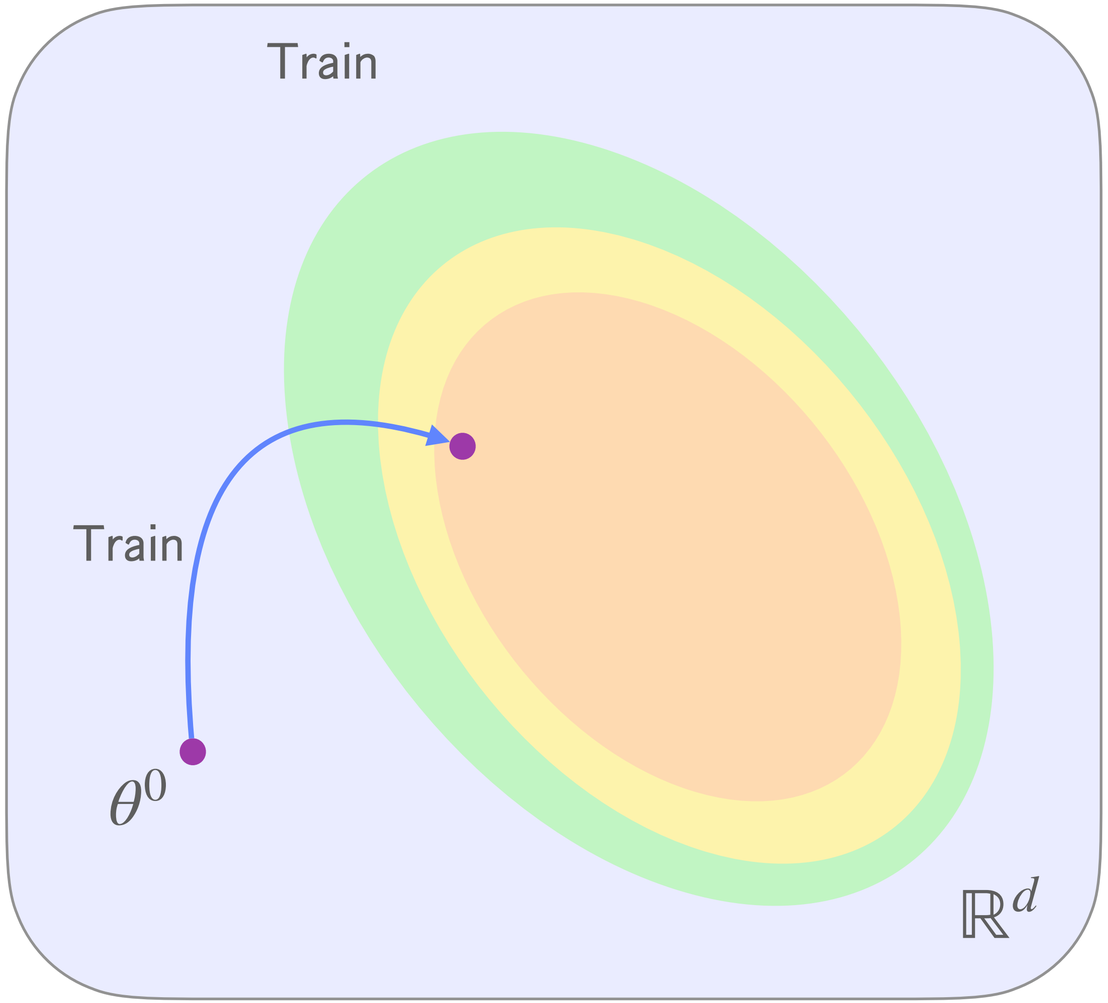
Observation 2 (Frankle et al., 2020)

When part of the training trajectory is shared, there exists a ***linear*** path in weight space between the two solutions along which loss remains low.



Observation 3 (Izmailov et al., 2018)

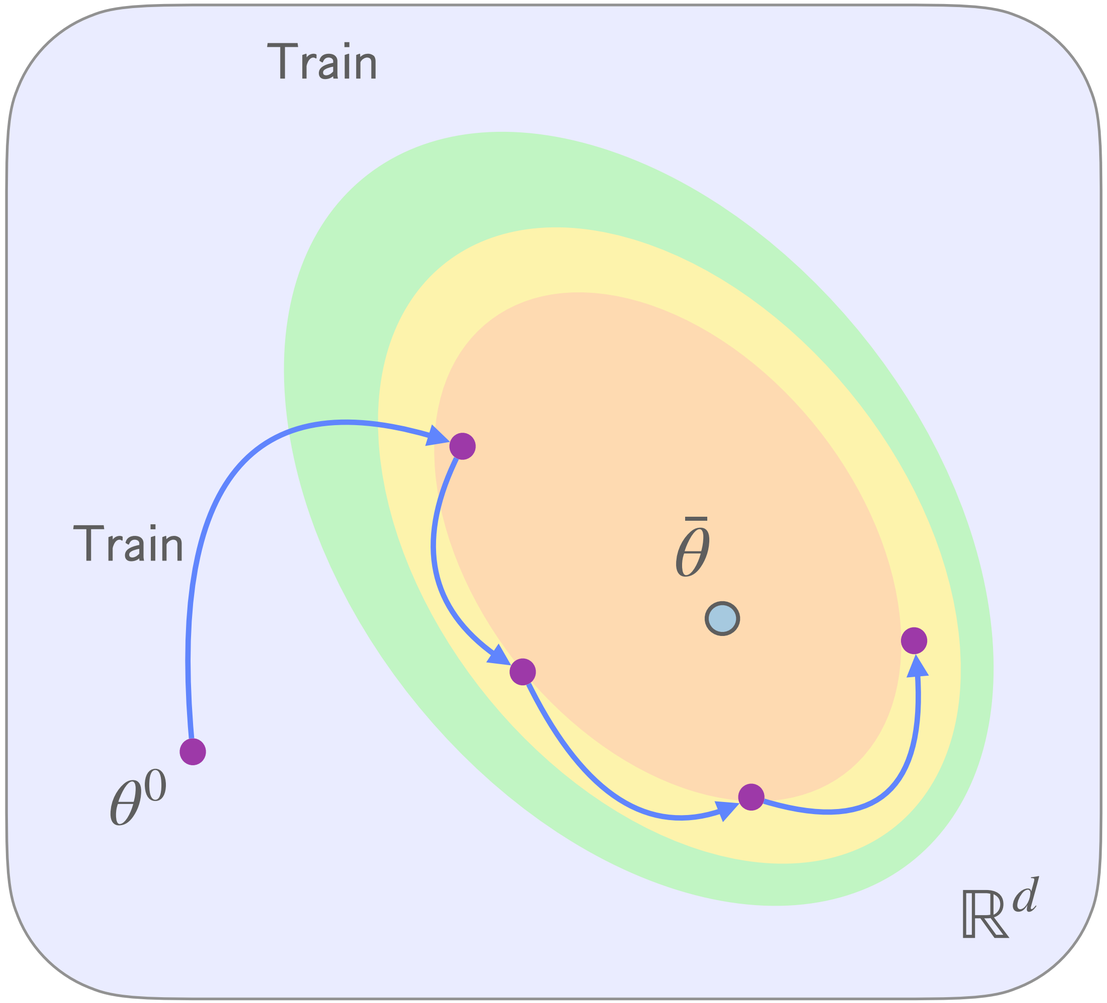
Weights found through SGD tend to lie at the periphery of a large, flat minimum.



Observation 3 (Izmailov et al., 2018)

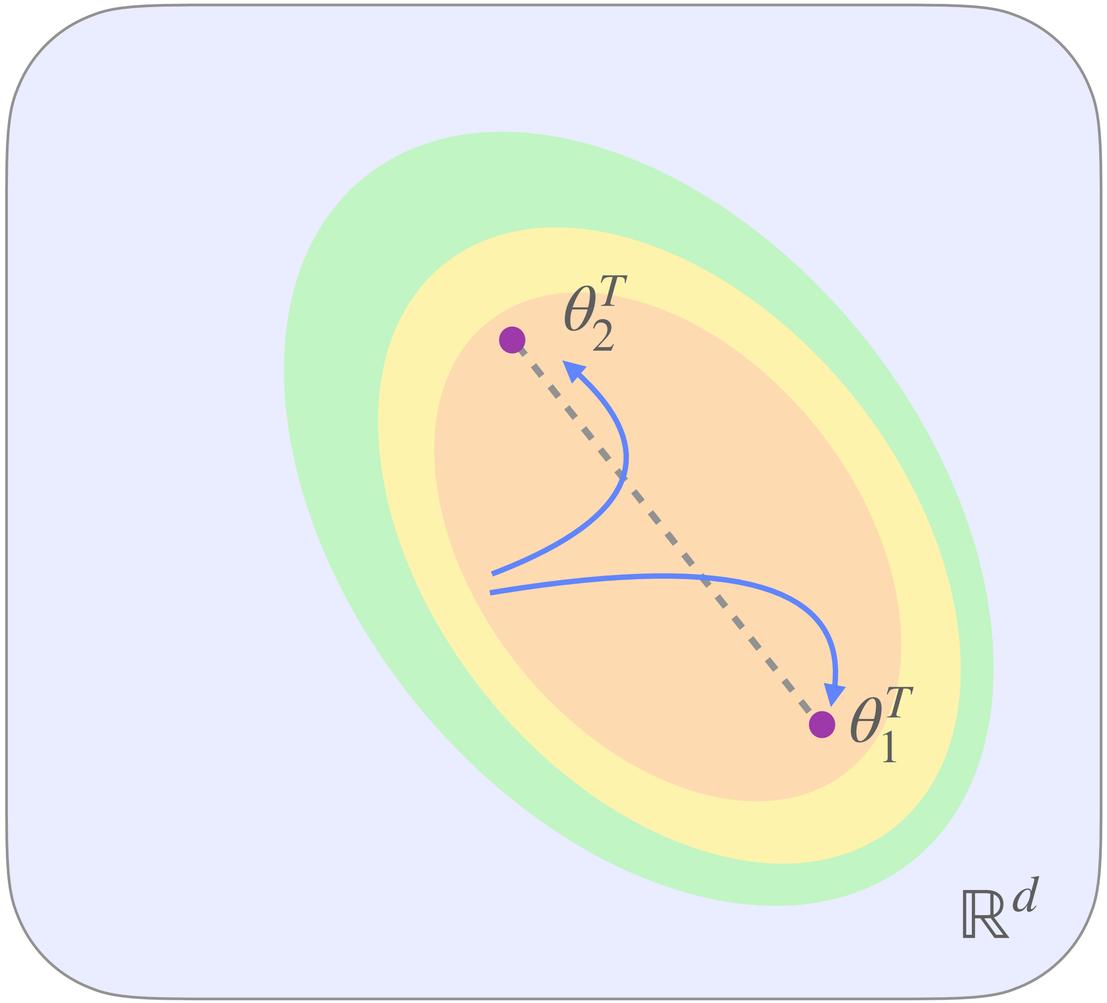
Weights found through SGD tend to lie at the periphery of a large, flat minimum.

To get closer to the center, Stochastic Weight Averaging (SWA) bounces around the minimum while saving checkpoints and returns the average.



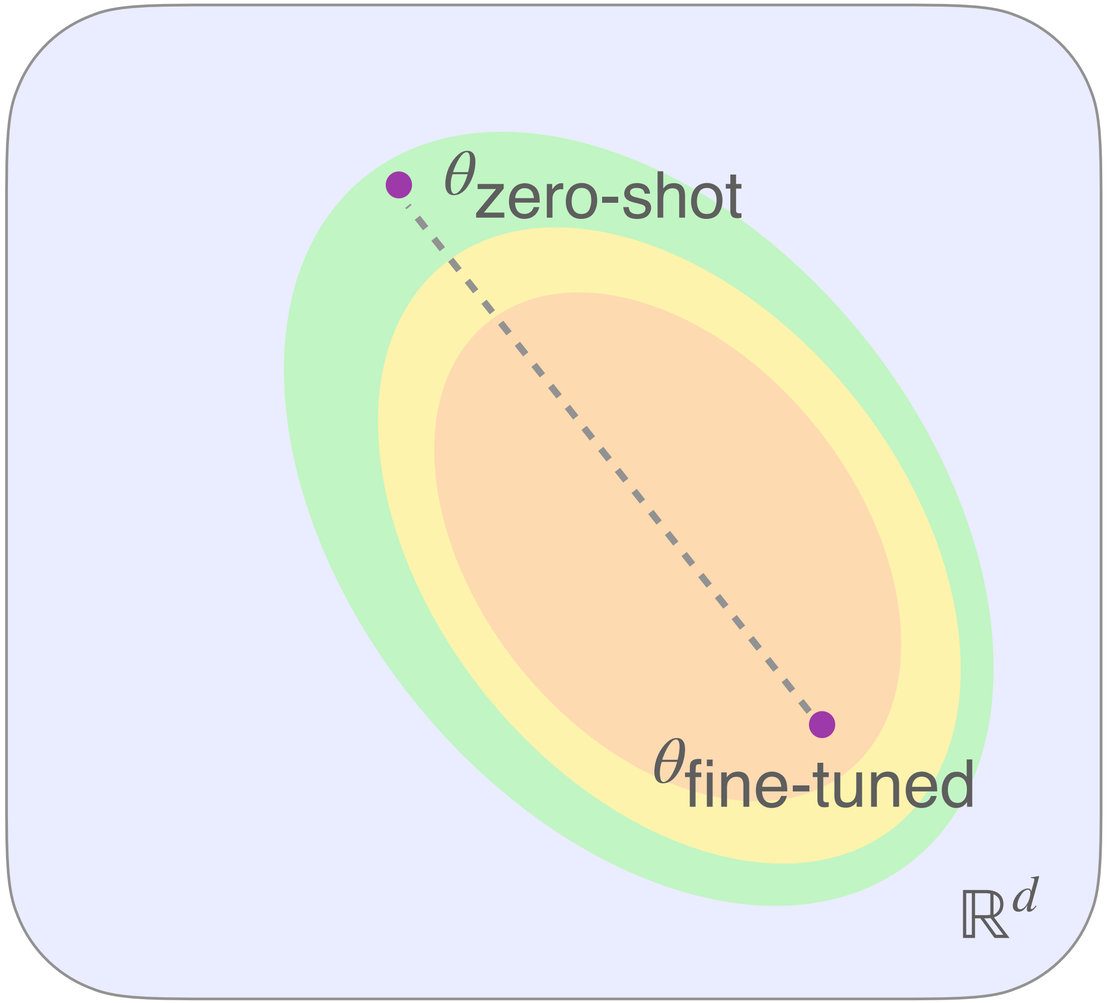
Observation 4 (Neyshabur, Sedghi, Zhang, 2020)

Linear mode connectivity between fine-tuned solutions.



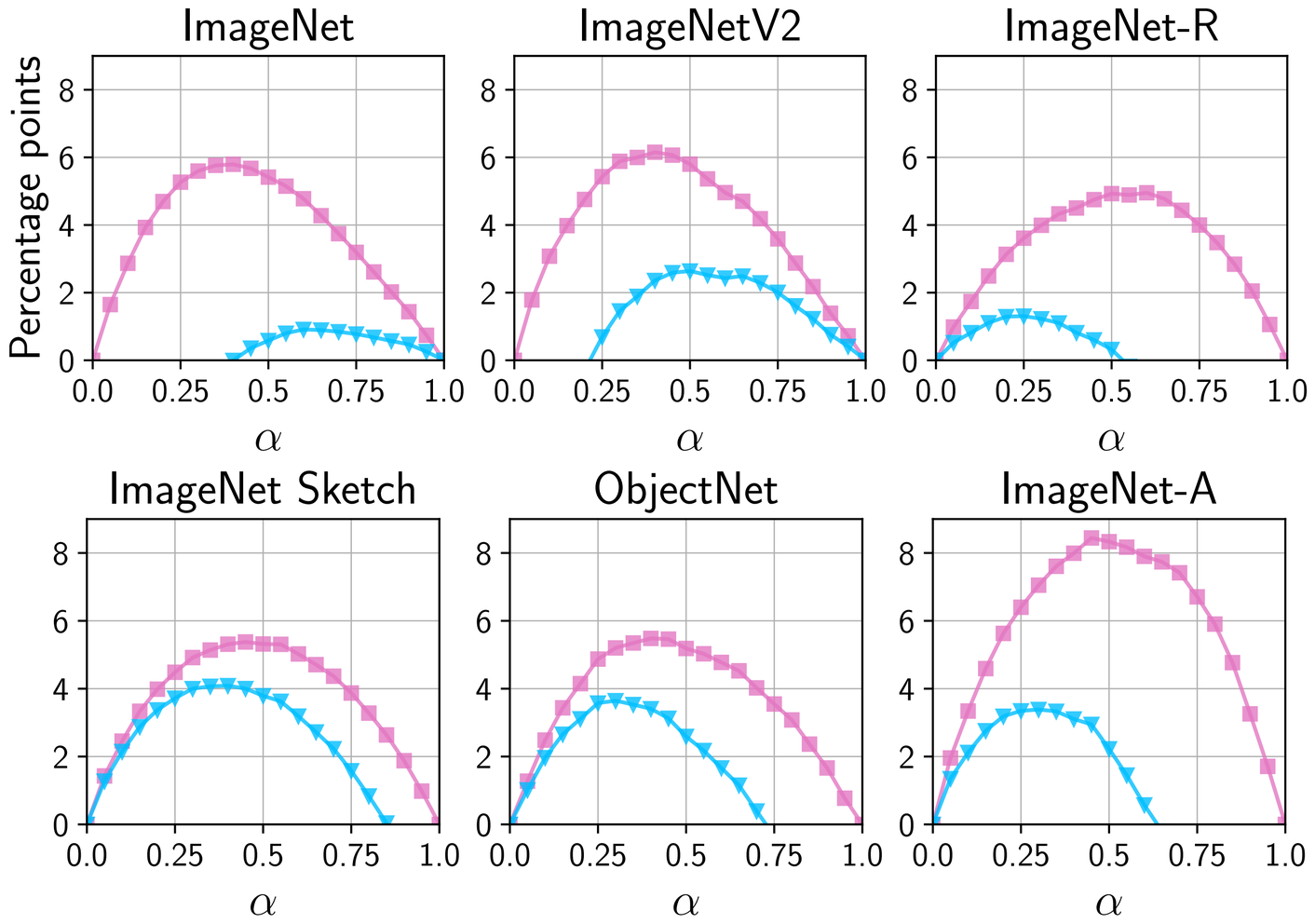
Observation 5

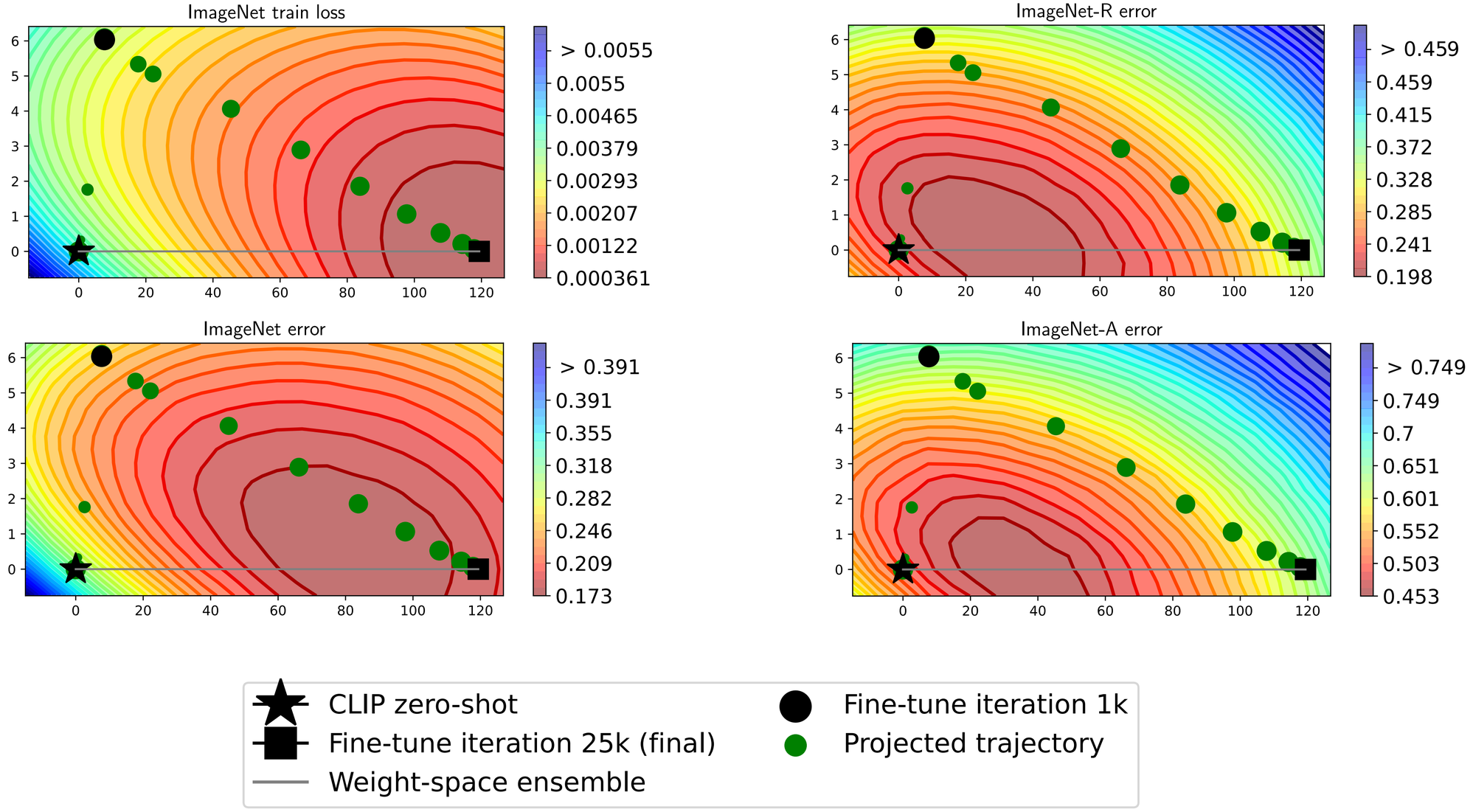
Linear mode connectivity between the zero-shot and fine-tuned solution on both the OOD and ID dataset.

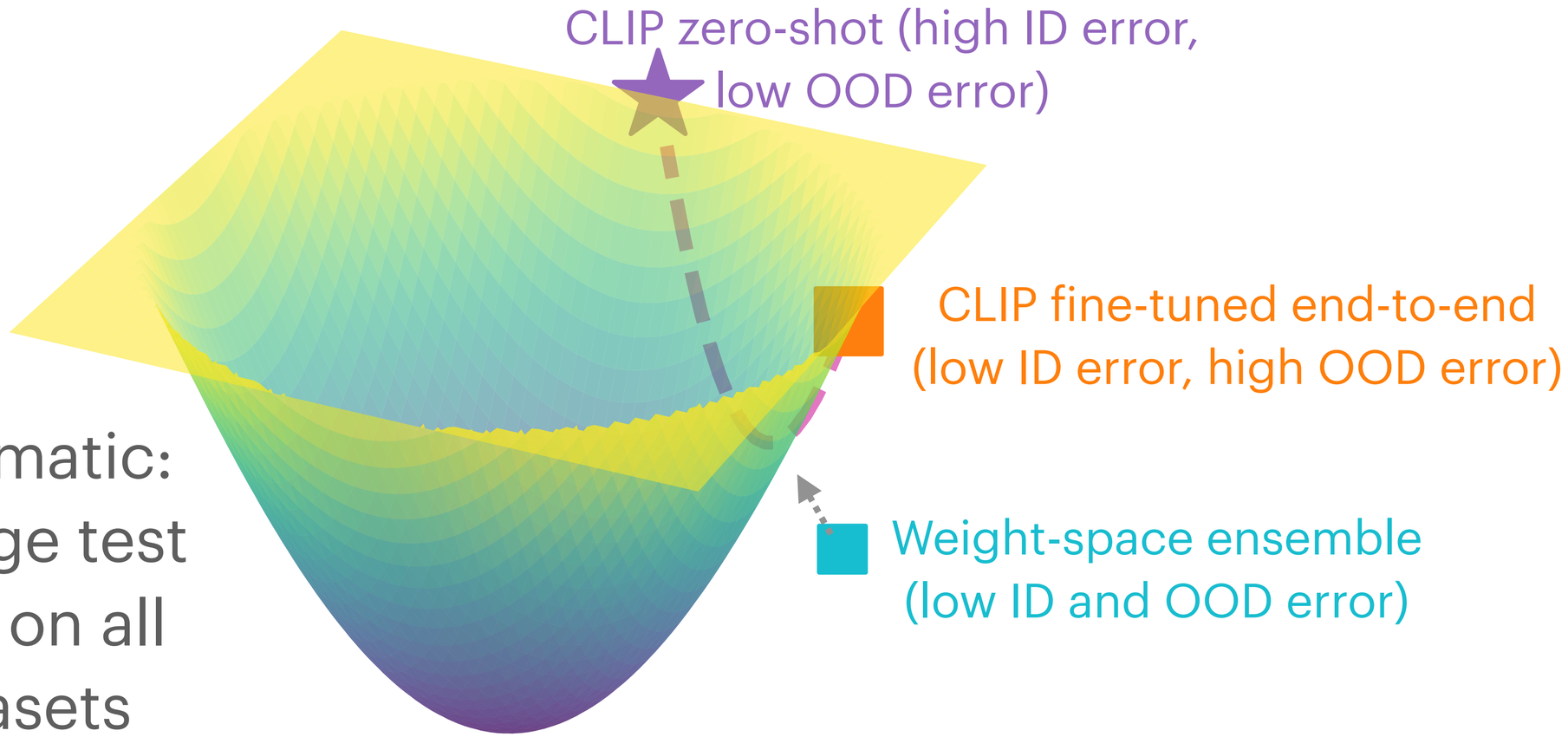


$$\mathrm{Acc}((1-\alpha)\cdot\theta_0+\alpha\cdot\theta_1)\geq(1-\alpha)\cdot\mathrm{Acc}(\theta_0)+\alpha\cdot\mathrm{Acc}(\theta_1)$$

$$\text{Acc}((1-\alpha)\cdot\theta_0+\alpha\cdot\theta_1)\geq(1-\alpha)\cdot\text{Acc}(\theta_0)+\alpha\cdot\text{Acc}(\theta_1)$$




ImageNet train loss
> 0.0055
0.0055
0.00465
0.00379
0.00293
0.00207
0.00122
0.000361
ImageNet-R error
> 0.459
0.459
0.415
0.372
0.328
0.285
0.241
0.198
ImageNet error
> 0.391
0.391
0.355
0.318
0.282
0.246
0.209
0.173
ImageNet-A error
> 0.749
0.749
0.7
0.651
0.601
0.552
0.503
0.453
CLIP zero-shot
Fine-tune iteration 25k (final)
Weight-space ensemble
Fine-tune iteration 1k
Projected trajectory


CLIP zero-shot (high ID error,
low OOD error)
CLIP fine-tuned end-to-end
(low ID error, high OOD error)
Weight-space ensemble
(low ID and OOD error)
Schematic:
average test
error on all
datasets

Takeaways

Takeaways

Robustness to distribution shift is an important step in ML.

Takeaways

Robustness to distribution shift is an important step in ML.

Large-scale pre-trained models such as CLIP are a promising direction.

Takeaways

Robustness to distribution shift is an important step in ML.

Large-scale pre-trained models such as CLIP are a promising direction.

Standard fine-tuning improves accuracy in-distribution, but deteriorates robustness.

Takeaways

Robustness to distribution shift is an important step in ML.

Large-scale pre-trained models such as CLIP are a promising direction.

Standard fine-tuning improves accuracy in-distribution, but deteriorates robustness.

WiSE-FT mitigates the compromise between high accuracy and robustness with no extra compute during fine-tuning or inference.