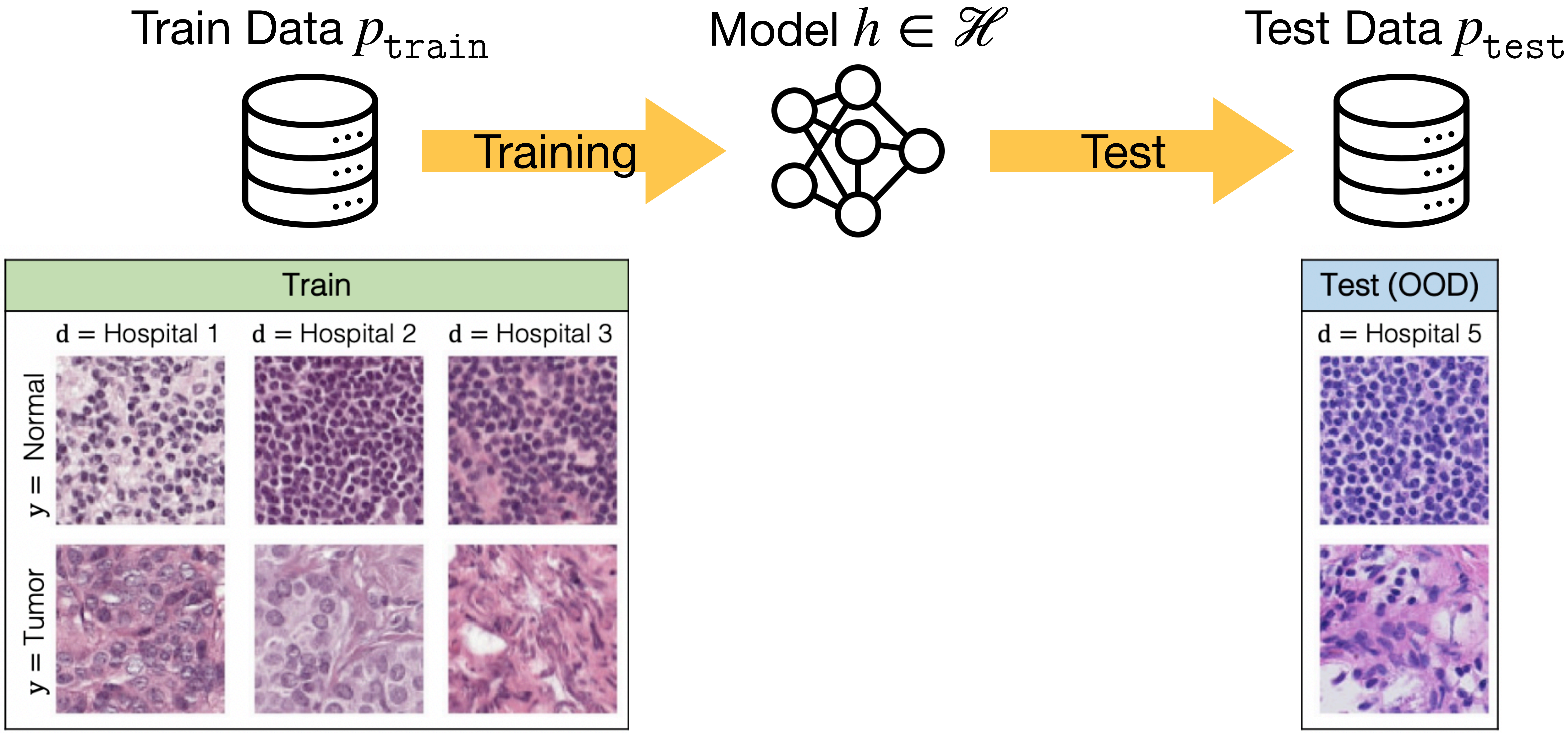
Test-Time Adaptation Induces Stronger Accuracy and Agreement-on-the-Line

Eungyeup Kim, Mingjie Sun, Christina Baek, Aditi Raghunathan, J. Zico Kolter

NeurIPS 2024

Motivation

Under distribution shifts, models often fail to generalize.

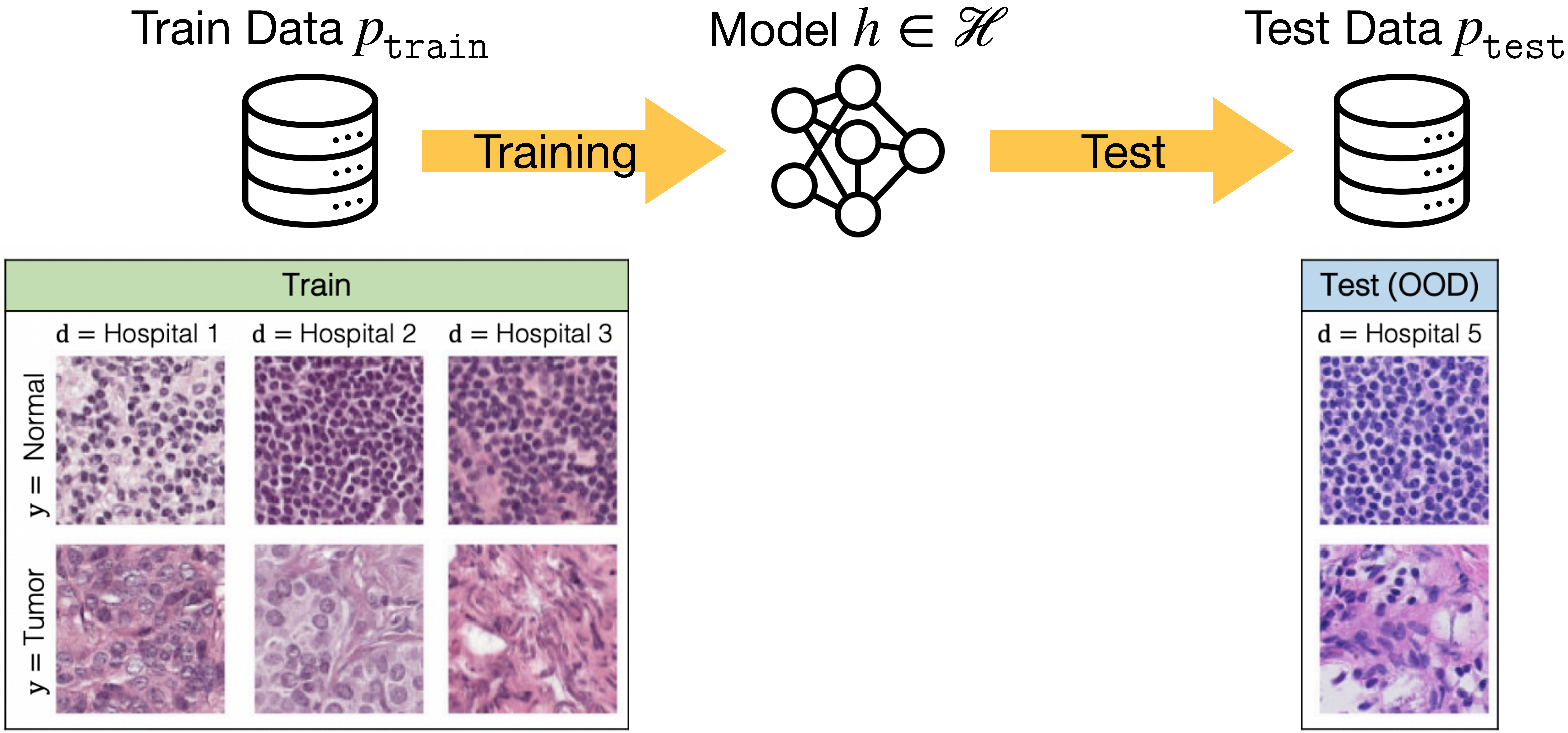


ID Accuracy: **85**%

OOD Accuracy $\ll$ **85**%

Motivation

Without labels, it is hard to predict models’ accuracy in OOD.



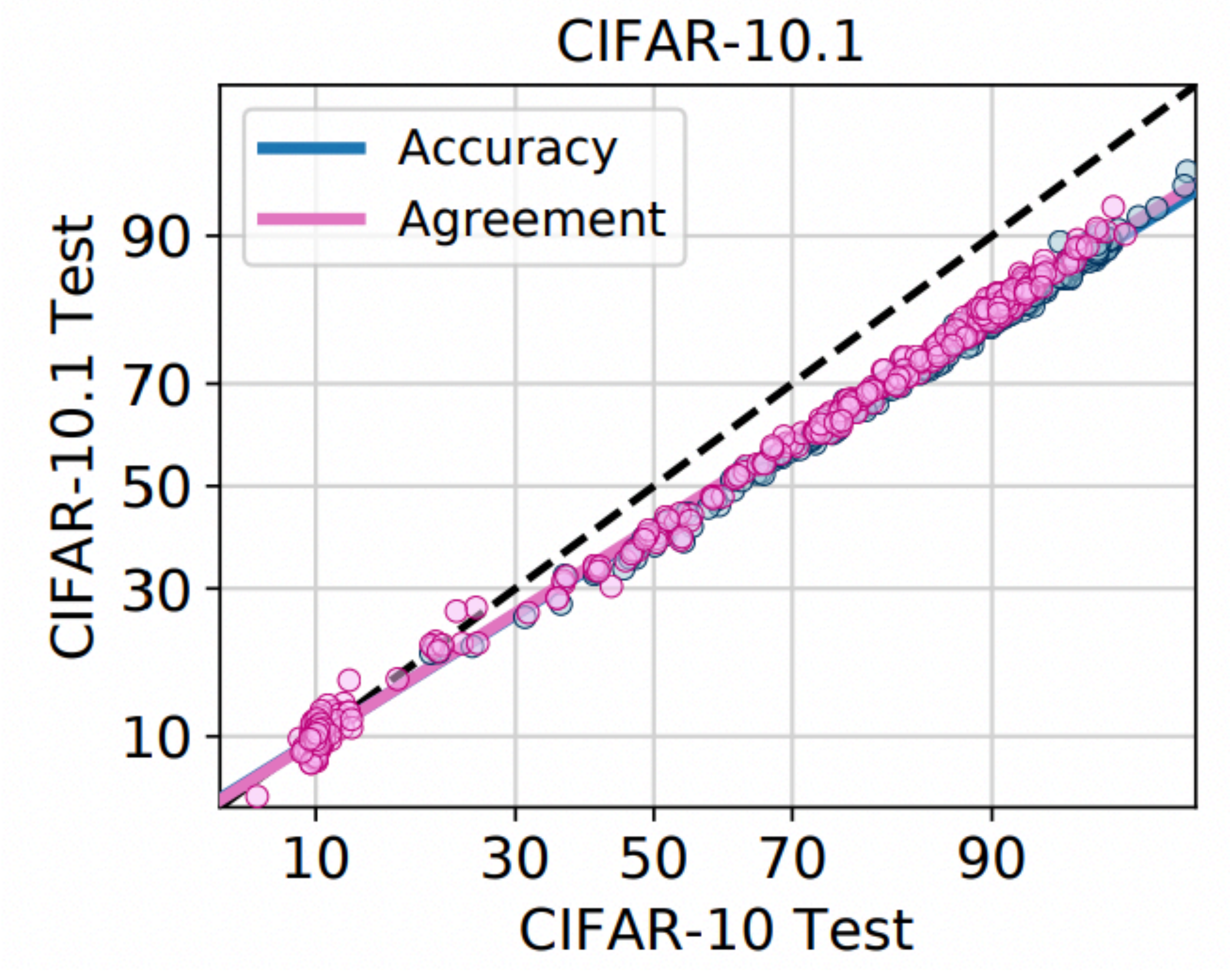
ID Accuracy: **85**%

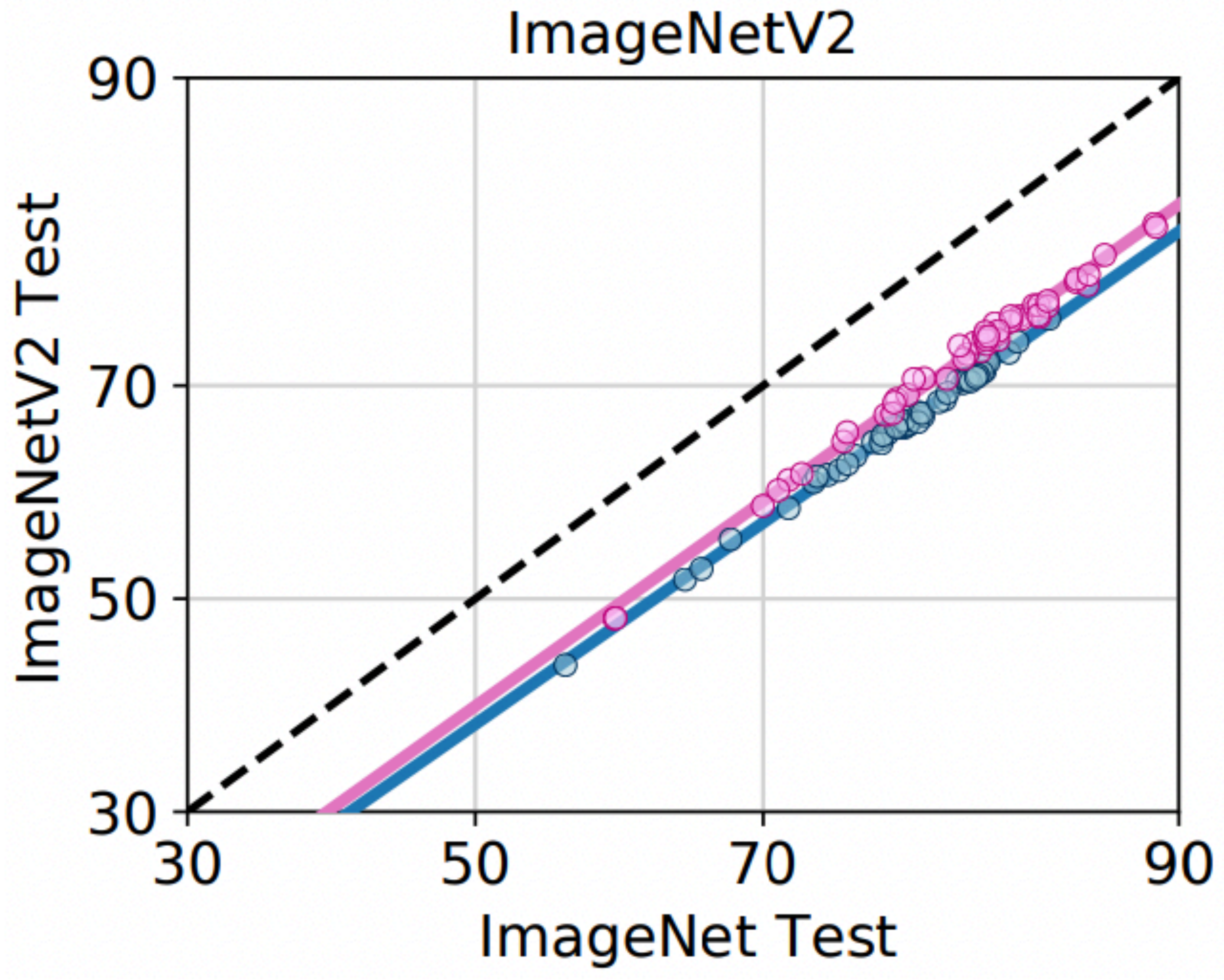
OOD Accuracy $\approx$ **?**%

Motivation

Recent studies [1,2] found simple empirical laws between ID and OOD.

- Models' ID vs. OOD accuracy are strongly correlated, termed as accuracy-on-the-line (**ACL**) [1].
- Additionally, when ACL, their agreements are correlated showing nearly identical linearity (**AGL**) [2].



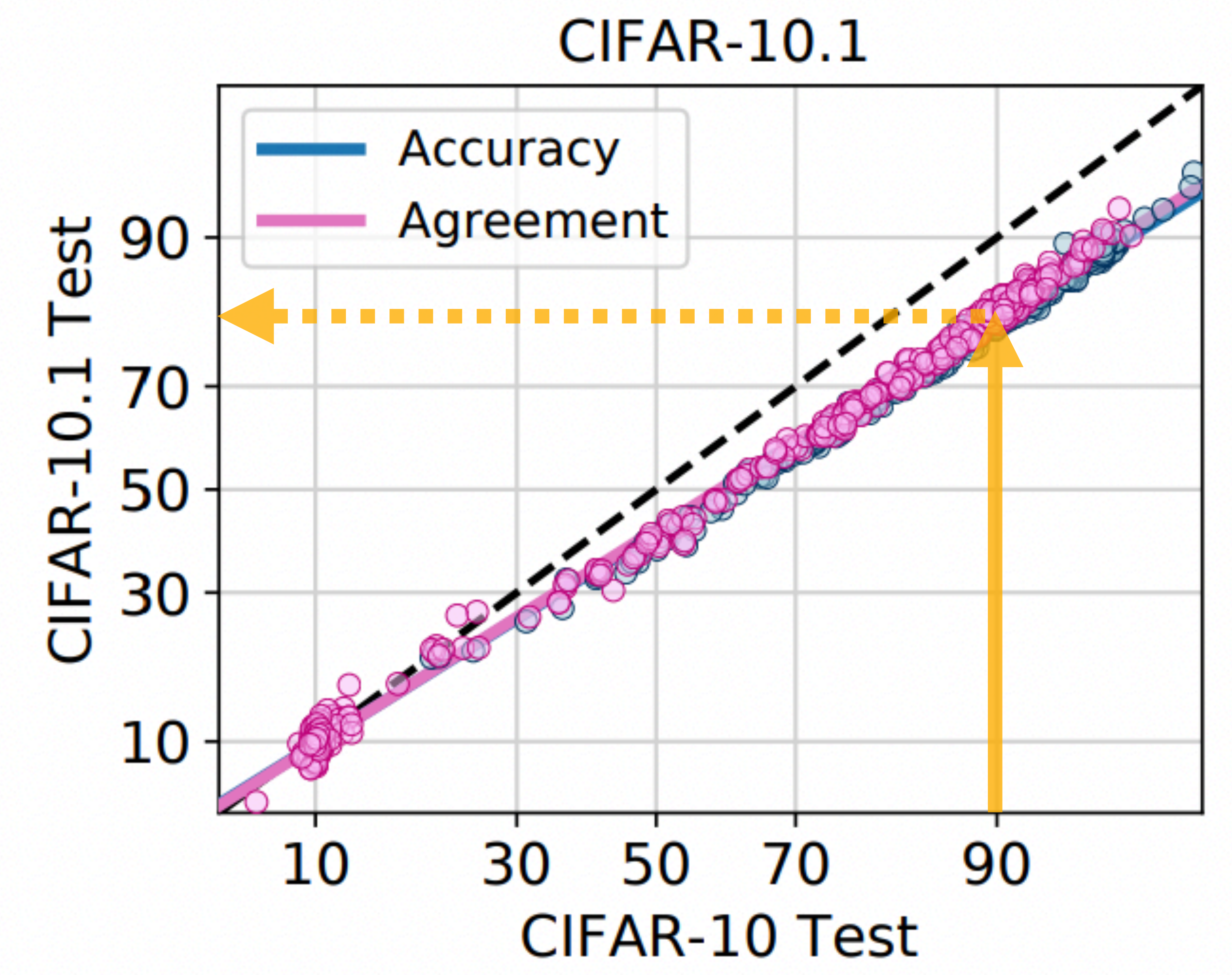


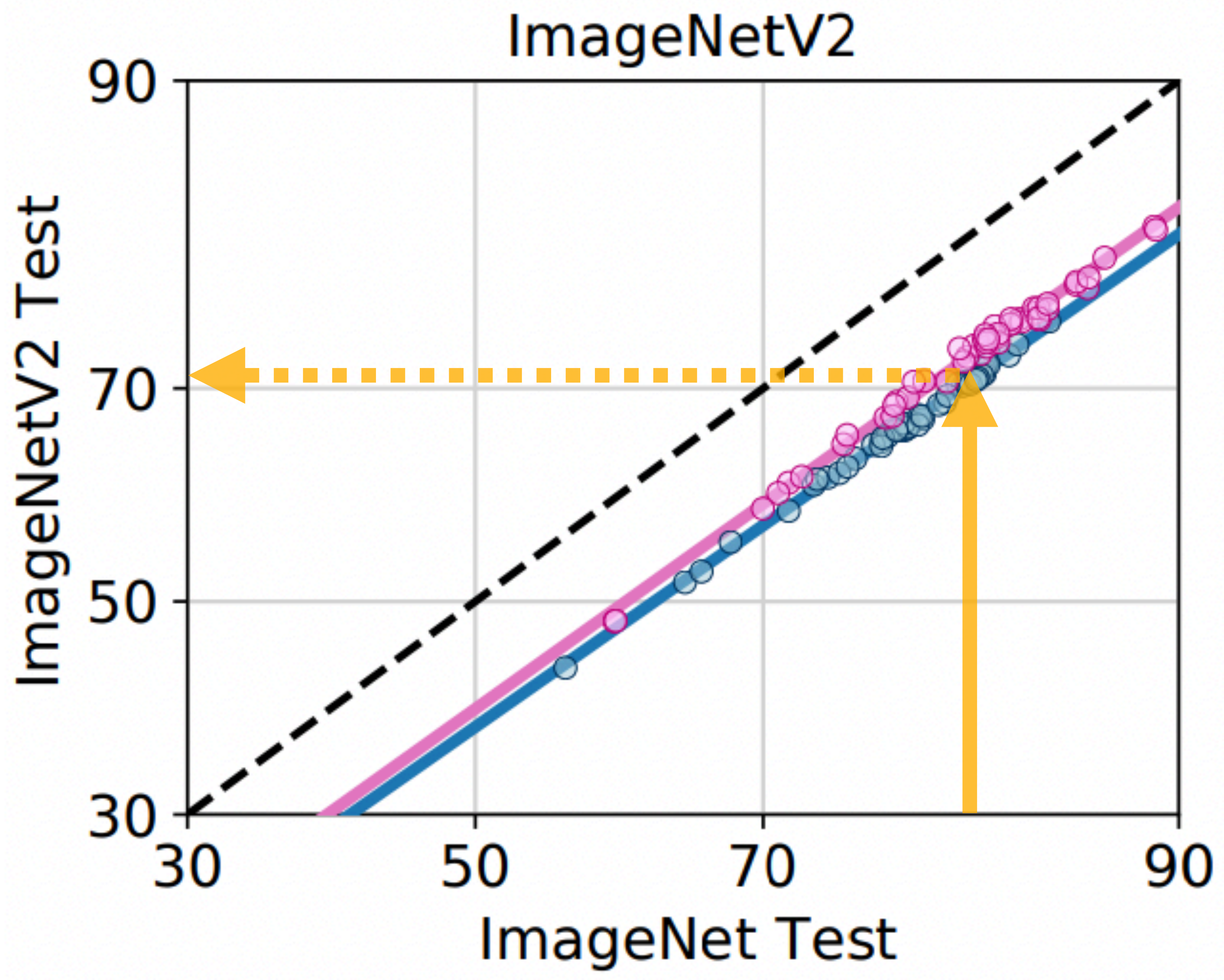
[1] Miller et al., Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization, ICML 2021
[2] Baek et al., Agreement-on-the-line: Predicting the performance of neural networks under distribution shift, NeurIPS 2022

Motivation

Recent studies [1,2] found simple empirical laws between ID and OOD.

- Models' ID vs. OOD accuracy are strongly correlated, termed as accuracy-on-the-line (**ACL**) [1].
- Additionally, when ACL, their agreements are correlated showing nearly identical linearity (**AGL**) [2].



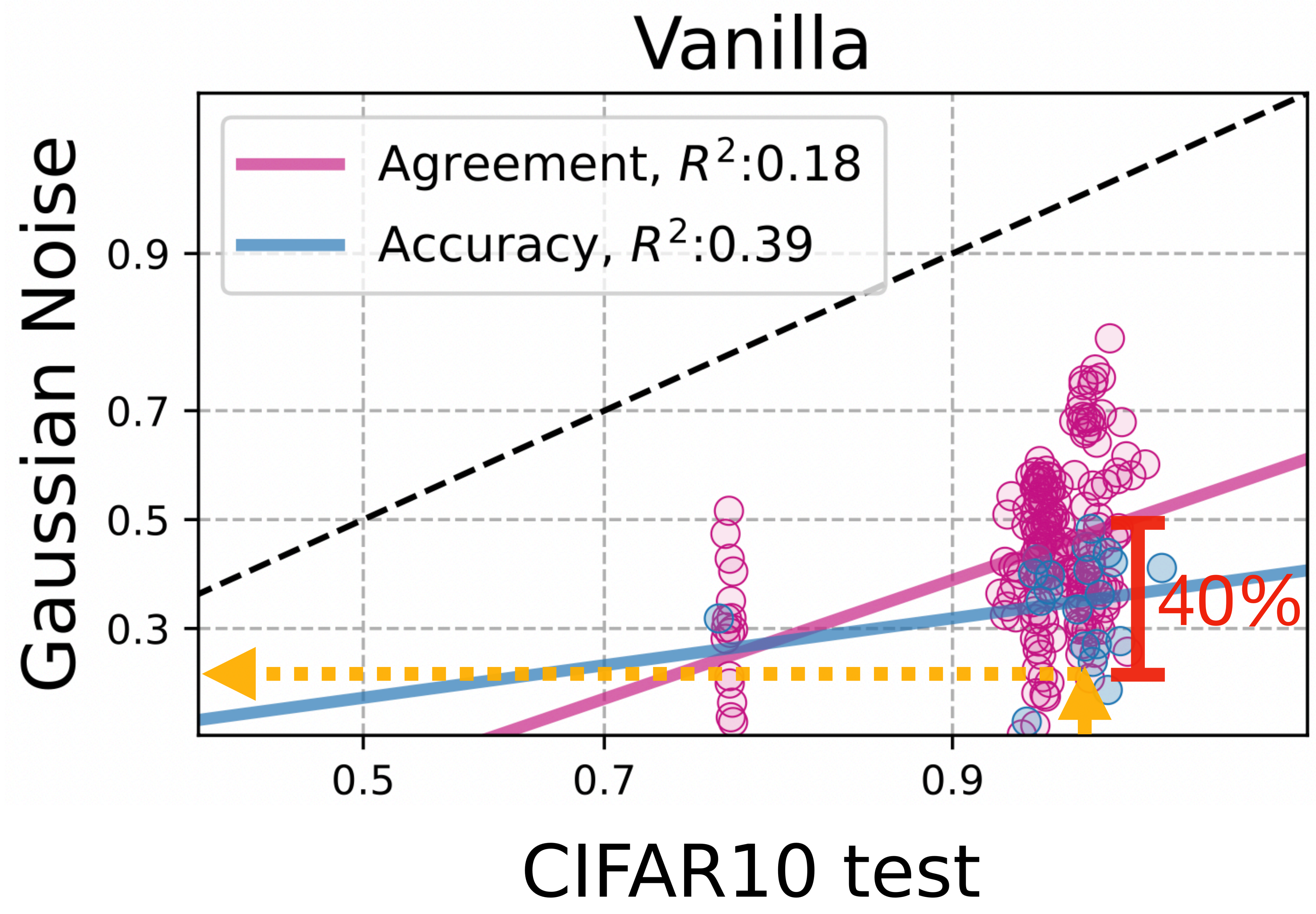


[1] Miller et al., Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization, ICML 2021
[2] Baek et al., Agreement-on-the-line: Predicting the performance of neural networks under distribution shift, NeurIPS 2022

Motivation

However, they often break in various distribution shifts.

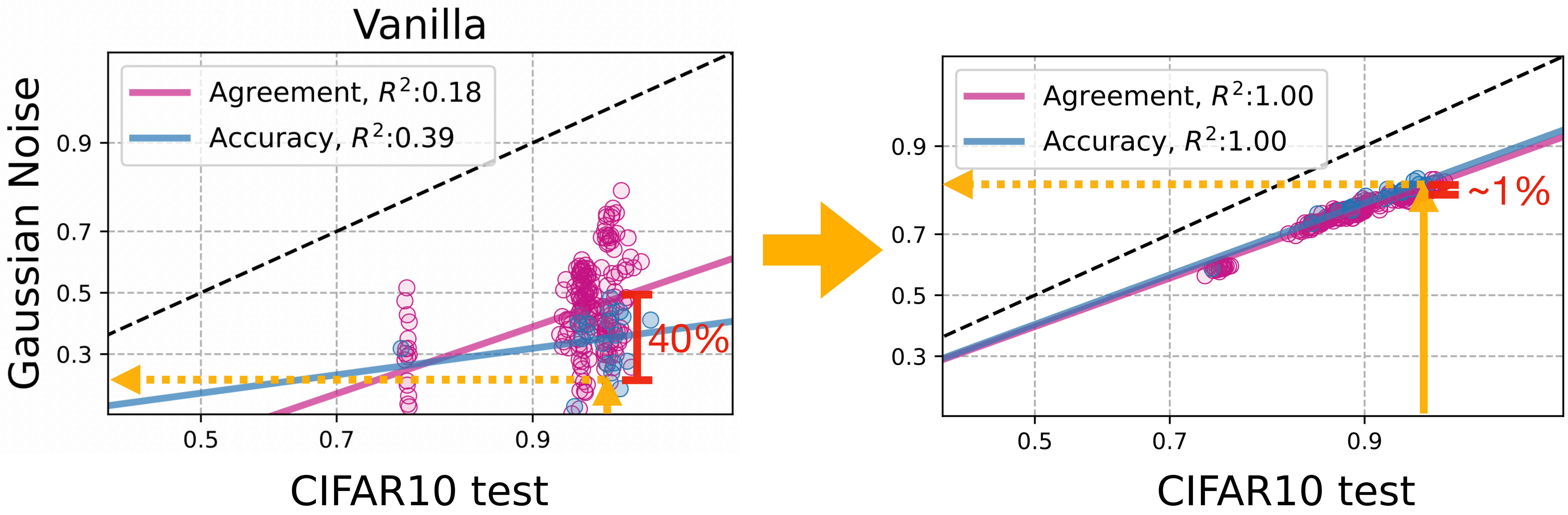
- Linearity breakdown leads to unreliable prediction on OOD performances.
- Any **intervention** to restore such linear trends?



Motivation

However, they often break in various distribution shifts.

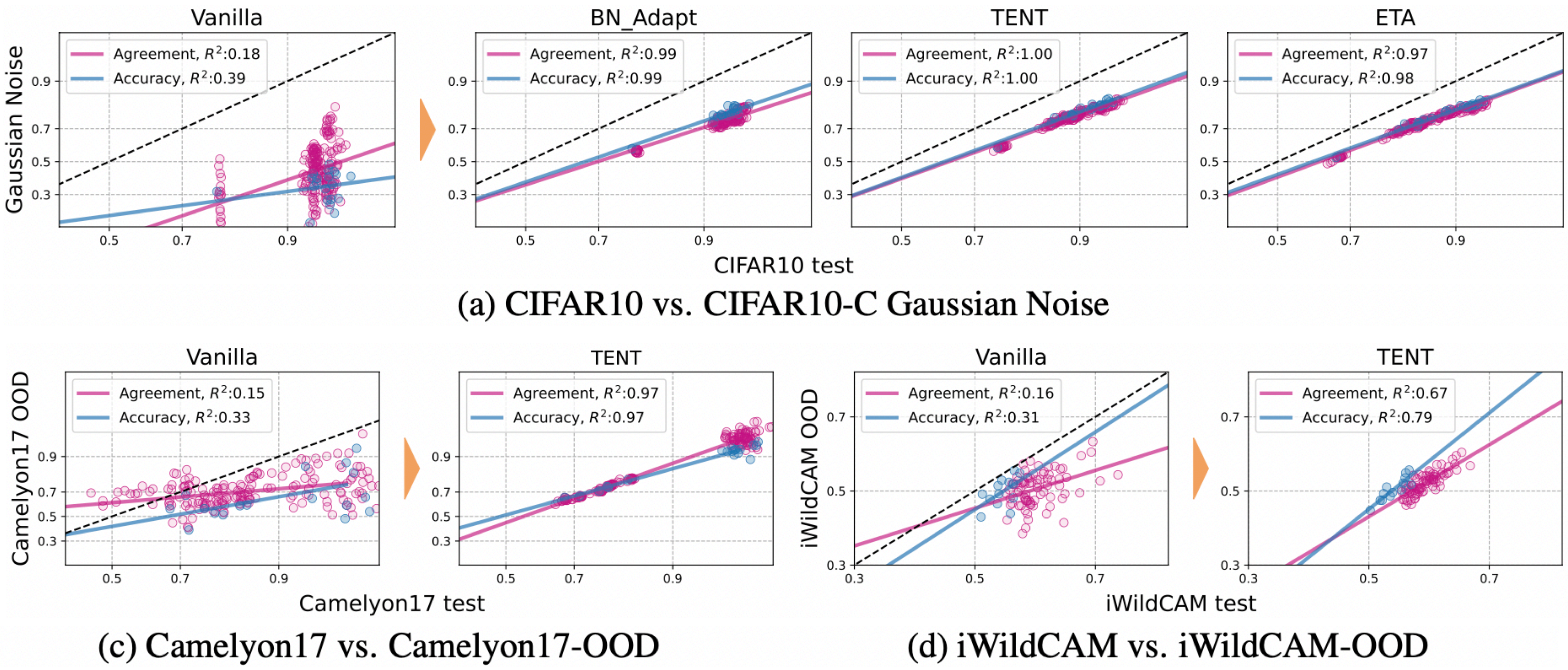
- Linearity breakdown leads to unreliable prediction on OOD performances.
- Any **intervention** to restore such linear trends?



🤔 Any intervention for restoring linear trends?

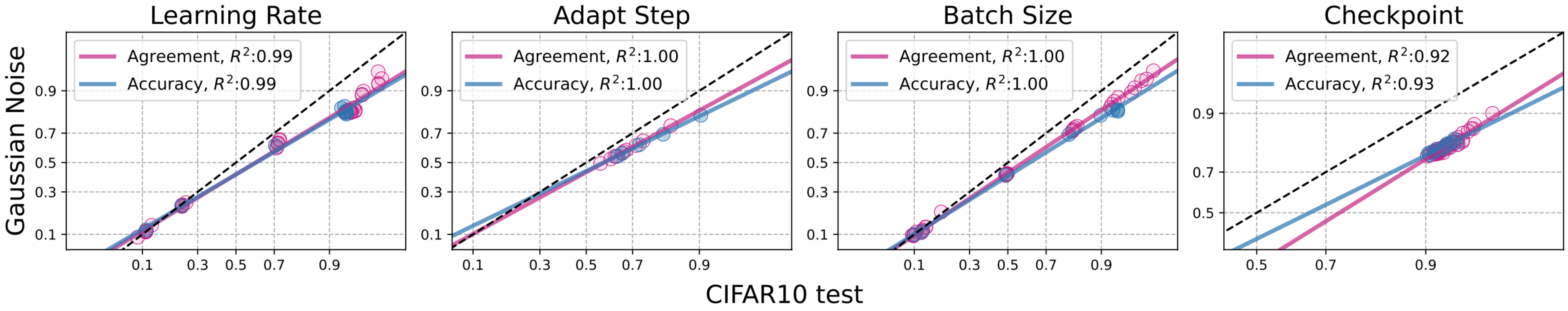
Observation

Test-Time Adaptation (TTA) empirically leads to stronger ACL / AGL.



Observation

Test-Time Adaptation (TTA) empirically leads to stronger ACL / AGL.



(a) TENT tested on CIFAR10 vs. CIFAR10-C Gaussian Noise

Explanation

Why TTA leads to stronger linear trends?

Explanation

Why TTA leads to stronger linear trends?

Theoretical Condition for Perfect ACL in Gaussian Toy Data

Out-of-distribution Q differs from in-distribution P by just some scaling constants $\alpha, \gamma > 0$,

$$P(x \mid y) = \mathcal{N}(y \cdot \mu; \Sigma), Q(x \mid y) = \mathcal{N}(y \cdot \alpha\mu; \gamma^2\Sigma).$$

[**Theorem 1**] Miller et al. (2021).

Under the Gaussian data setup, across all linear classifiers $f_\theta : x \mapsto \text{sign}(\theta^\top x)$, the profit-scaled accuracies over P and Q observes perfect linear correlation with a bias of zero and a slope of α/γ.

Explanation

Why TTA leads to stronger linear trends?

Theoretical Condition for Perfect ACL in Gaussian Toy Data

Out-of-distribution Q differs from in-distribution P by just some scaling constants $\alpha, \gamma > 0$,

$$P(x \mid y) = \mathcal{N}(y \cdot \mu; \Sigma), Q(x \mid y) = \mathcal{N}(y \cdot \alpha\mu; \gamma^2\Sigma).$$

[**Theorem 1**] Miller et al. (2021).

Under the Gaussian data setup, across all linear classifiers $f_\theta : x \mapsto \text{sign}(\theta^\top x)$, the profit-scaled accuracies over P and Q observes perfect linear correlation with a bias of zero and a slope of α/γ.

Explanation

Why TTA leads to stronger linear trends?

Theoretical Condition for Perfect ACL in Gaussian Toy Data

Out-of-distribution Q differs from in-distribution P by just some scaling constants $\alpha, \gamma > 0$,

$$P(x \mid y) = \mathcal{N}(y \cdot \mu; \Sigma), Q(x \mid y) = \mathcal{N}(y \cdot \alpha\mu; \gamma^2\Sigma).$$

[**Theorem 1**] Miller et al. (2021).

Under the Gaussian data setup, across all linear classifiers $f_\theta : x \mapsto \text{sign}(\theta^\top x)$, the profit-scaled accuracies over P and Q observes perfect linear correlation with a bias of zero and a slope of α/γ.

After TTA, in penultimate layer feature space, ID vs. OOD distributions have same mean direction and covariance shape (i.e., satisfying **Theorem 1**).

Explanation

Why TTA leads to stronger linear trends?

Setup	Cosine Similarity		Slope	
	Mean	Covariance	Theoretical	Empirical
Vanilla (Archs.)	0.691 ± 0.175	0.750 ± 0.109	–	–
BN_Adapt (Archs.)	0.988 ± 0.007	0.972 ± 0.011	0.751 ± 0.075	0.758
TENT (Archs.)	0.990 ± 0.005	0.974 ± 0.011	0.753 ± 0.072	0.778
Learning rates	0.993 ± 0.003	0.977 ± 0.006	0.759 ± 0.041	0.76
Batch Sizes	0.995 ± 0.003	0.982 ± 0.010	0.831 ± 0.101	0.809
Check Points	0.992 ±0.003	0.976 ± 0.008	0.782 ± 0.033	0.838

Table 1: Cosine similarity between mean direction and covariance shape of class-wise penultimate-layer features, followed by the comparison between theoretical and empirical slope. They are evaluated on CIFAR10 vs. CIFAR10-C Gaussian Noise, measured across architectures and hyperparameters. We report their means and standard deviations.

Experiments

Strong linear trends lead to OOD accuracy estimation.

Dataset	Method	Error	ATC	DOC-feat	AC	Agreement	ALine-S	ALine-D
CIFAR10-C	Vanilla	31.38	8.31	15.03	17.42	5.45	6.02	5.87
	SHOT	15.40	1.63	4.63	7.63	1.78	0.96	**0.77**
	BN_Adapt	16.87	3.69	4.79	7.53	1.93	1.12	**0.91**
	TENT	15.43	4.25	4.65	7.66	1.79	0.97	**0.77**
	ConjPL	16.62	1.80	6.16	11.46	2.02	1.18	**1.01**
	ETA	15.14	4.58	4.50	7.68	1.76	0.92	**0.72**
CIFAR100-C	Vanilla	59.04	5.05	12.82	18.34	6.96	7.49	7.22
	SHOT	40.79	2.21	5.44	14.36	2.52	1.64	**0.90**
	BN_Adapt	42.69	2.89	4.42	11.81	2.33	1.43	**1.13**
	TENT	41.11	6.60	5.59	14.85	2.65	1.64	**0.88**
	ConjPL	42.79	**1.09**	6.55	23.73	2.40	1.67	1.18
	ETA	44.27	7.15	4.92	16.49	4.96	1.44	**0.81**
ImageNet-C	Vanilla	80.41	3.95	13.72	17.34	9.06	6.00	5.95
	BN_Adapt	69.05	7.37	**2.63**	2.86	3.91	6.16	6.09
	TENT	56.58	5.98	6.54	12.70	7.48	4.62	**4.57**
	ETA	56.56	10.21	7.91	34.38	8.02	**3.66**	3.72
	SAR	43.30	5.39	8.61	13.68	5.51	5.19	**4.17**
Camelyon17 -WILDS	Vanilla	34.07	14.91	17.31	21.69	11.95	12.88	13.46
	TENT	14.37	3.00	3.43	6.94	6.49	2.29	**2.27**
	ETA	16.43	3.05	4.38	6.85	5.33	2.24	**1.42**
iWildCAM -WILDS	Vanilla	50.27	7.12	2.73	23.86	3.00	3.53	2.82
	TENT	47.39	5.44	3.20	28.03	3.55	**2.59**	2.96
	ETA	46.49	6.61	3.40	29.34	4.62	**2.14**	2.82

Experiments

Strong linear trends lead to unsupervised model validation.

HyperParameter	CIFAR10-C						ImageNet-C					
	MixVal	ENT	IM	Corr-C	SND	Ours	MixVal	ENT	IM	Corr-C	SND	Ours
Architecture	2.31	1.06	1.06	21.71	2.77	0.03	6.22	0.96	0.47	26.32	20.60	0.75
Learning Rate	6.97	8.88	2.24	11.56	1.87	0.72	12.75	20.49	1.49	20.18	12.61	9.70
Checkpoints	3.21	0.0	0.0	5.53	3.46	0.05	–	–	–	–	–	–
Batch Size	7.85	3.32	0.96	32.37	5.68	0.77	14.29	42.31	0.99	42.31	42.31	5.61
Adapt Step	0.85	0.0	0.0	1.02	0.0	0.23	1.85	1.94	1.25	3.09	2.17	0.30
Average	4.23	2.65	0.85	14.43	2.75	**0.36**	8.77	16.42	**1.05**	14.43	22.97	4.0

HyperParameter	ImageNet-R						Camelyon17-WILDS					
	MixVal	ENT	IM	Corr-C	SND	Ours	MixVal	ENT	IM	Corr-C	SND	Ours
Architecture	1.75	0.62	0.62	22.17	22.17	0.85	28.87	1.03	1.03	28.87	28.87	0.85
Learning Rate	3.12	10.16	4.73	19.16	19.16	2.8	0.91	48.37	46.41	48.37	48.37	1.14
Batch Size	1.83	35.88	0.08	35.88	35.88	1.74	0.0	46.67	46.67	40.45	40.45	1.37
Adapt Step	1.07	1.07	1.07	1.07	1.07	0.0	2.17	33.12	0.0	33.12	33.12	0.0
Average	1.94	14.18	1.62	19.57	19.57	**1.34**	7.98	32.29	23.52	37.70	37.70	**0.62**

Thank you