Understanding and Improving Feature Learning for Out-of-Distribution Generalization

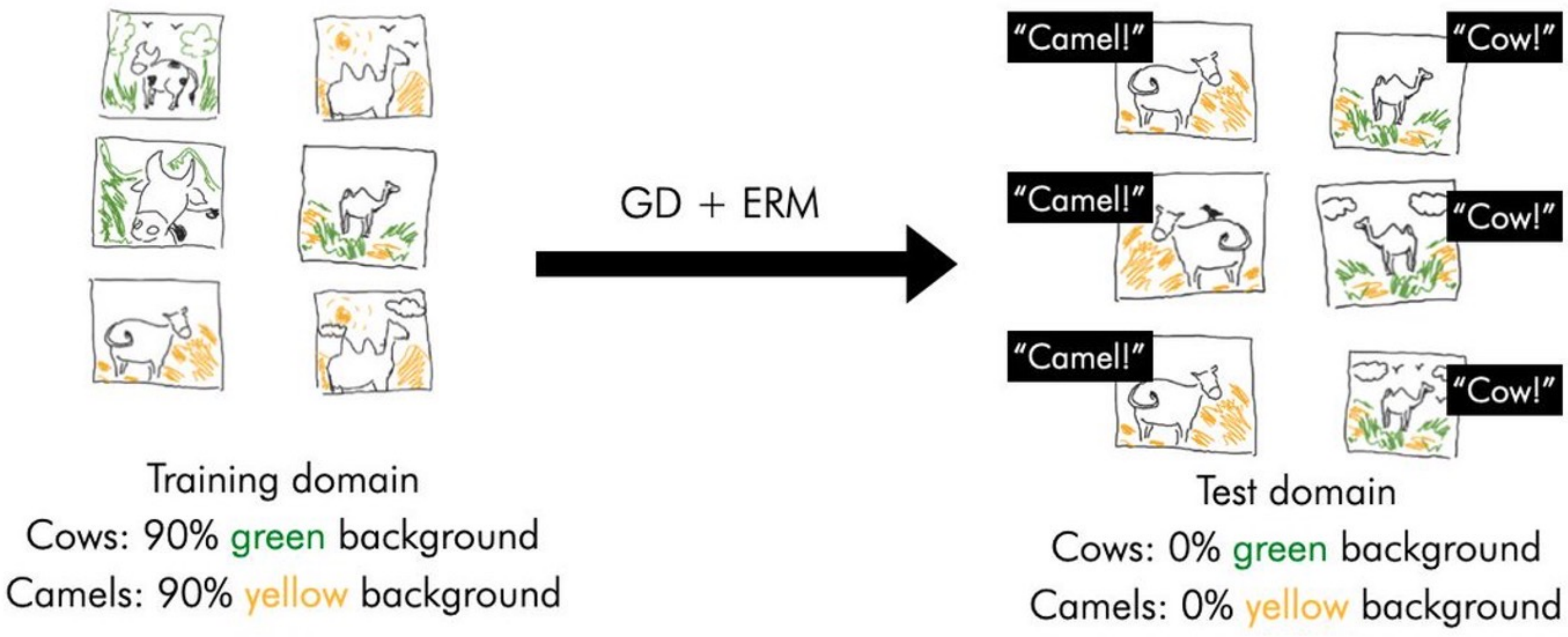
Yongqiang Chen*
CUHK, Tencent AI Lab

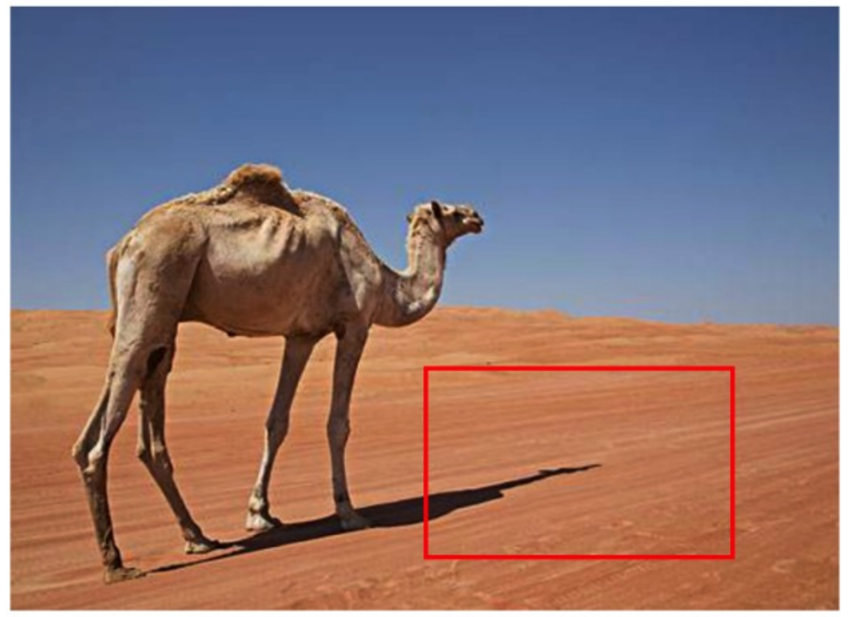
with Wei Huang, Kaiwen Zhou*, Yatao Bian, Bo Han, and James Cheng*

**equal contributions*

A Debate on ERM Feature Learning

ERM learns **predictive** but **spurious** features, that are **bad** for out-of-distribution (OOD) generalization.





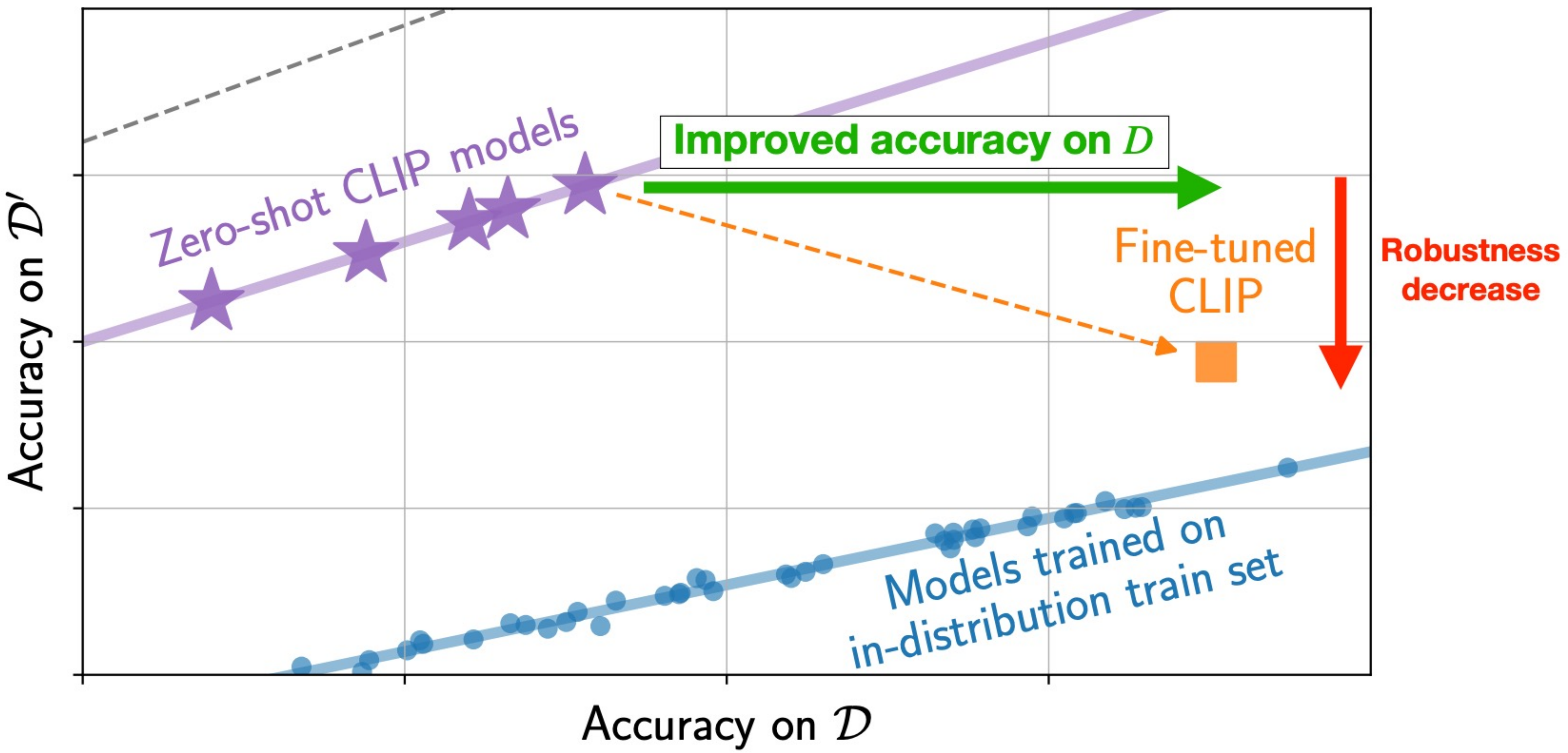
camel

cow

(Beery et al., 2018; Arjovsky et al., 2019; DeGrave et al. 2021; Ahuja et al., 2021)

A Debate on ERM Feature Learning

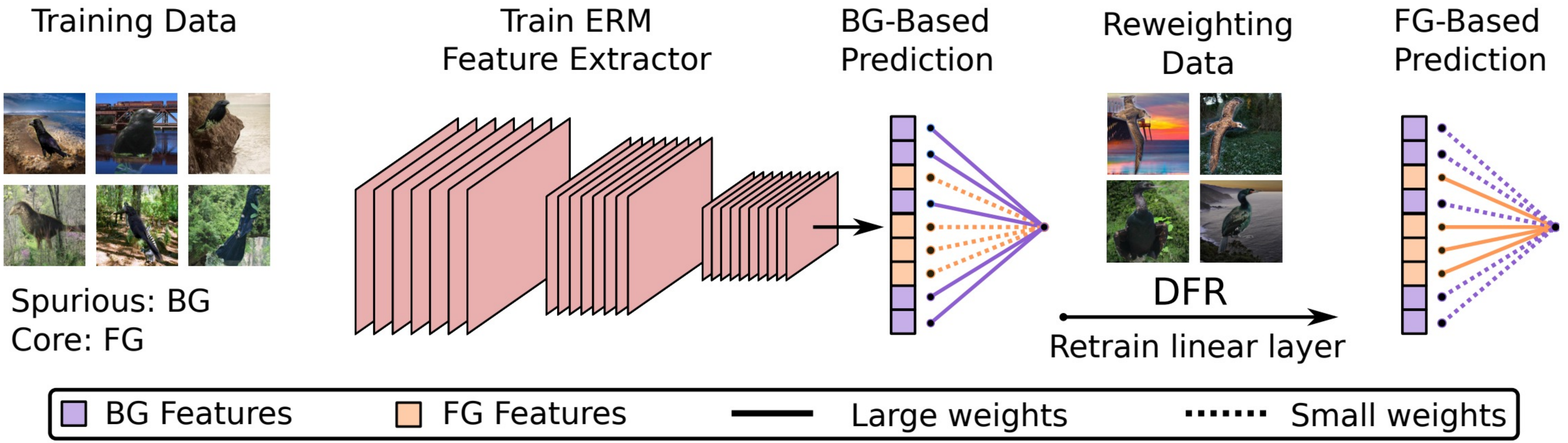
Fine-tuning generalist models with ERM can learns **predictive** but **spurious** features, that are **bad** for OOD generalization.



(Wortsman et al., 2021; Kumar et al., 2022)

A Debate on ERM Feature Learning

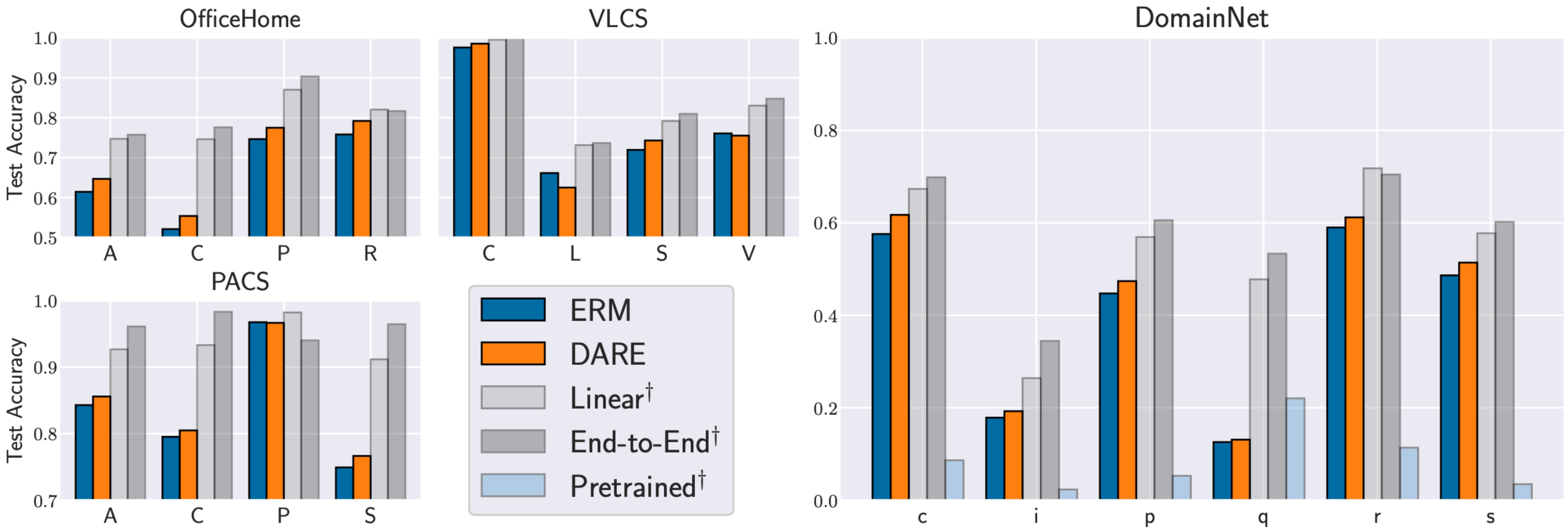
ERM already learns **invariant** features, that are **useful** for OOD generalization.



(Beery et al., 2018; Arjovsky et al., 2019; DeGrave et al. 2021; Ahuja et al., 2021)

A Debate on ERM Feature Learning

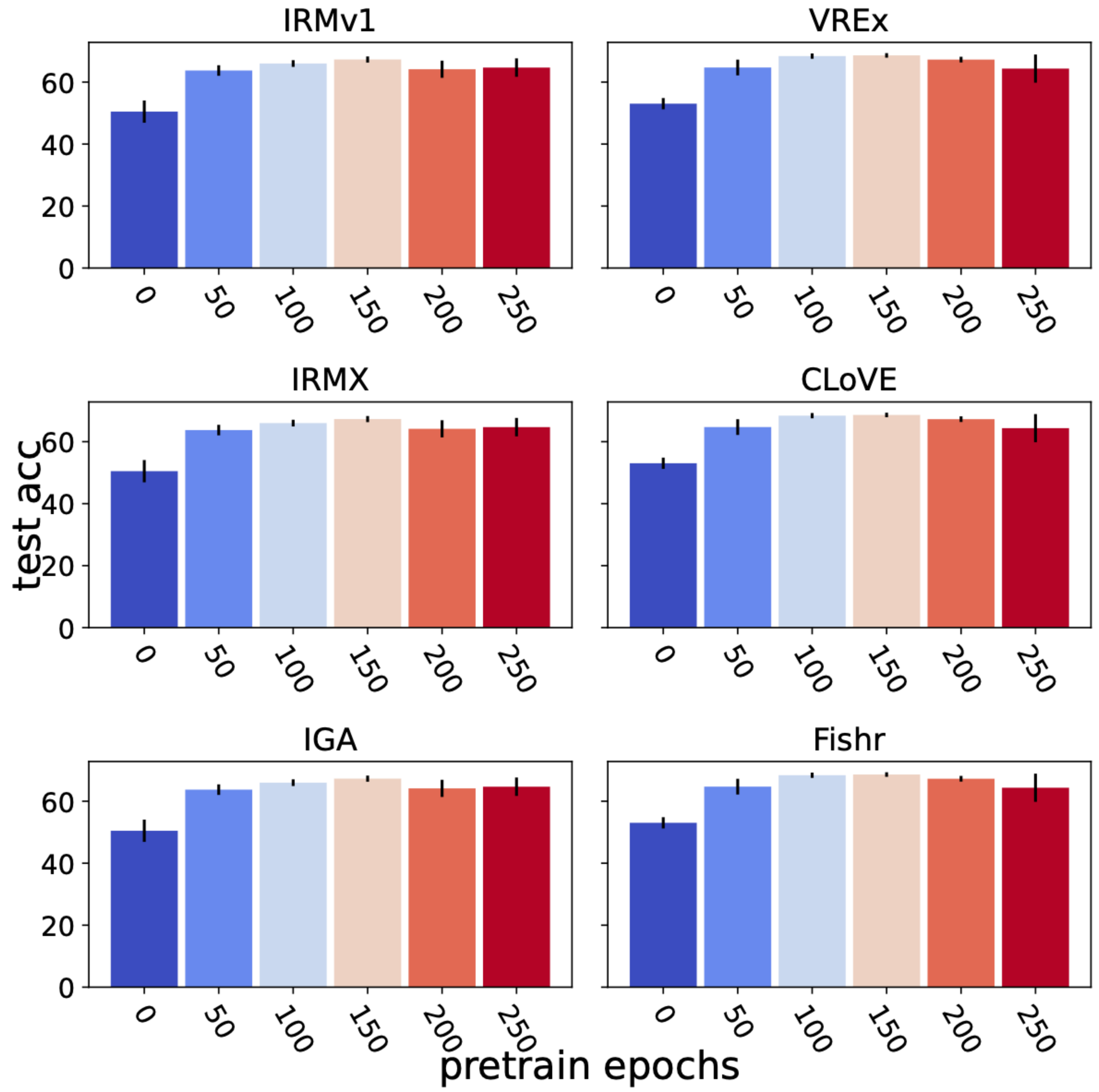
ERM already learns **invariant** features, that are **useful** for OOD generalization.



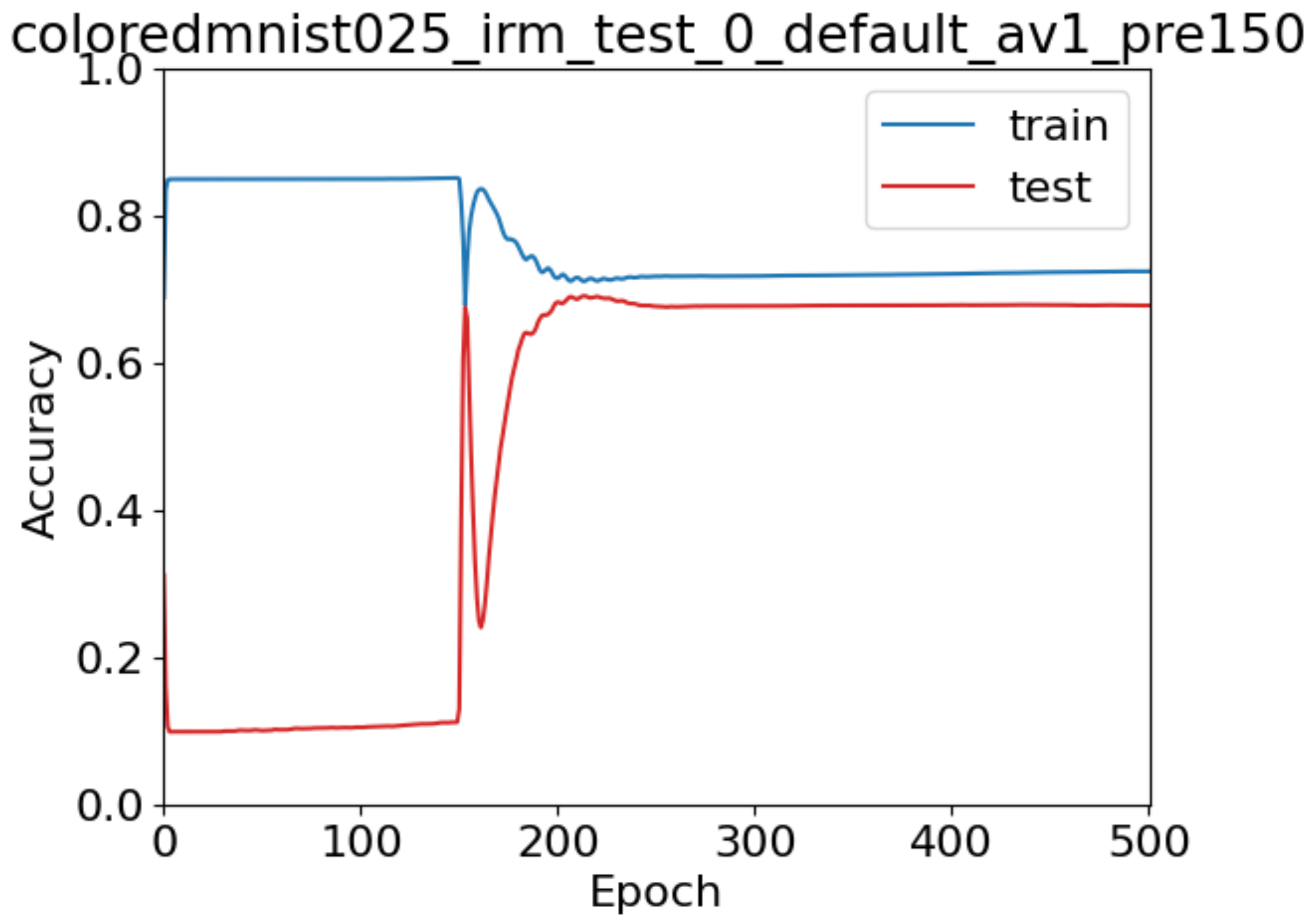
(*Beery et al., 2018; Arjovsky et al., 2019; DeGrave et al. 2021; Ahuja et al., 2021)*)

A Debate on ERM Feature Learning

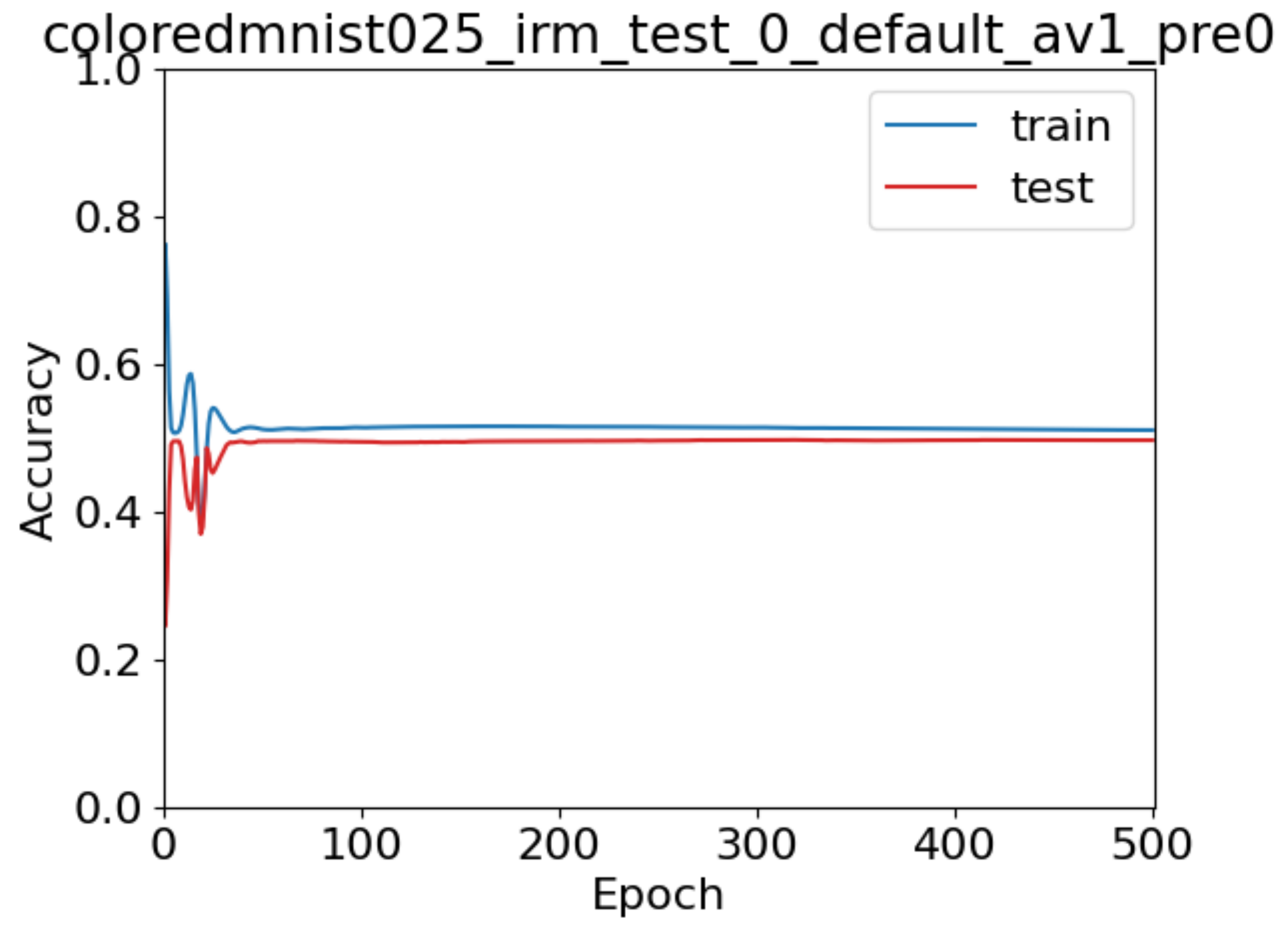
OOD generalization performance heavily **rely on** proper ERM pre-training.



OOD performance on ColoredMNIST



IRMv1 **with** ERM pretraining (150 epochs)



IRMv1 **w/o** ERM pretraining

(Zhang et al., 2022; Chen et al., 2022)

Is there a contradict?

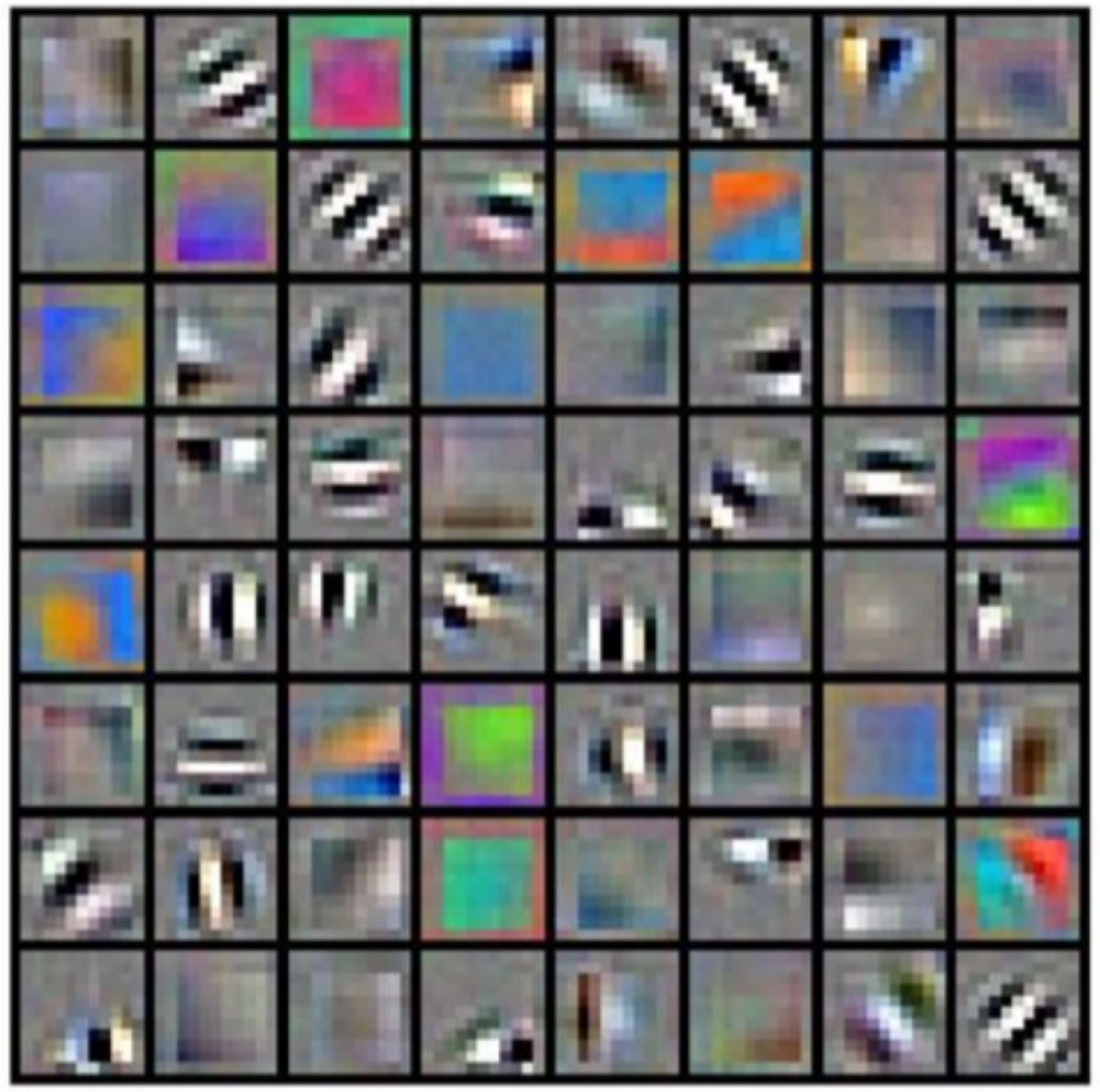
_or_ 🤔

A lack of understanding about feature learning in OOD generalization?

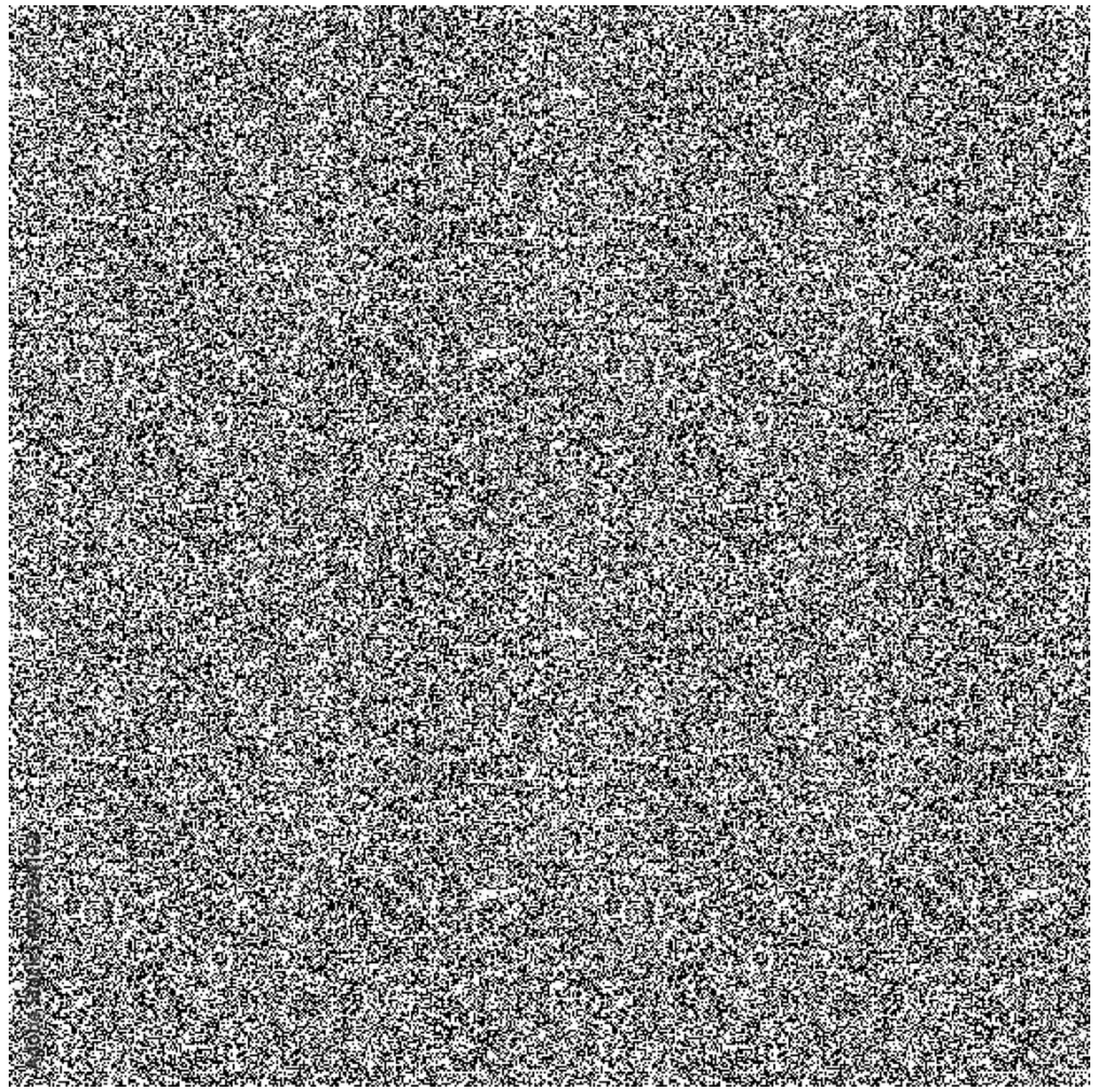
Data Model for OOD Generalization

- Two classes $y = \{-1, +1\}$
- The input $\mathbf{x} \in \mathbb{R}^{2d}$ is composed of

A feature patch $\mathbf{x}_1 \in \mathbb{R}^d$



A noise patch $\mathbf{x}_2 \in \mathbb{R}^d$



(Allen-Zhu & Li 2019)

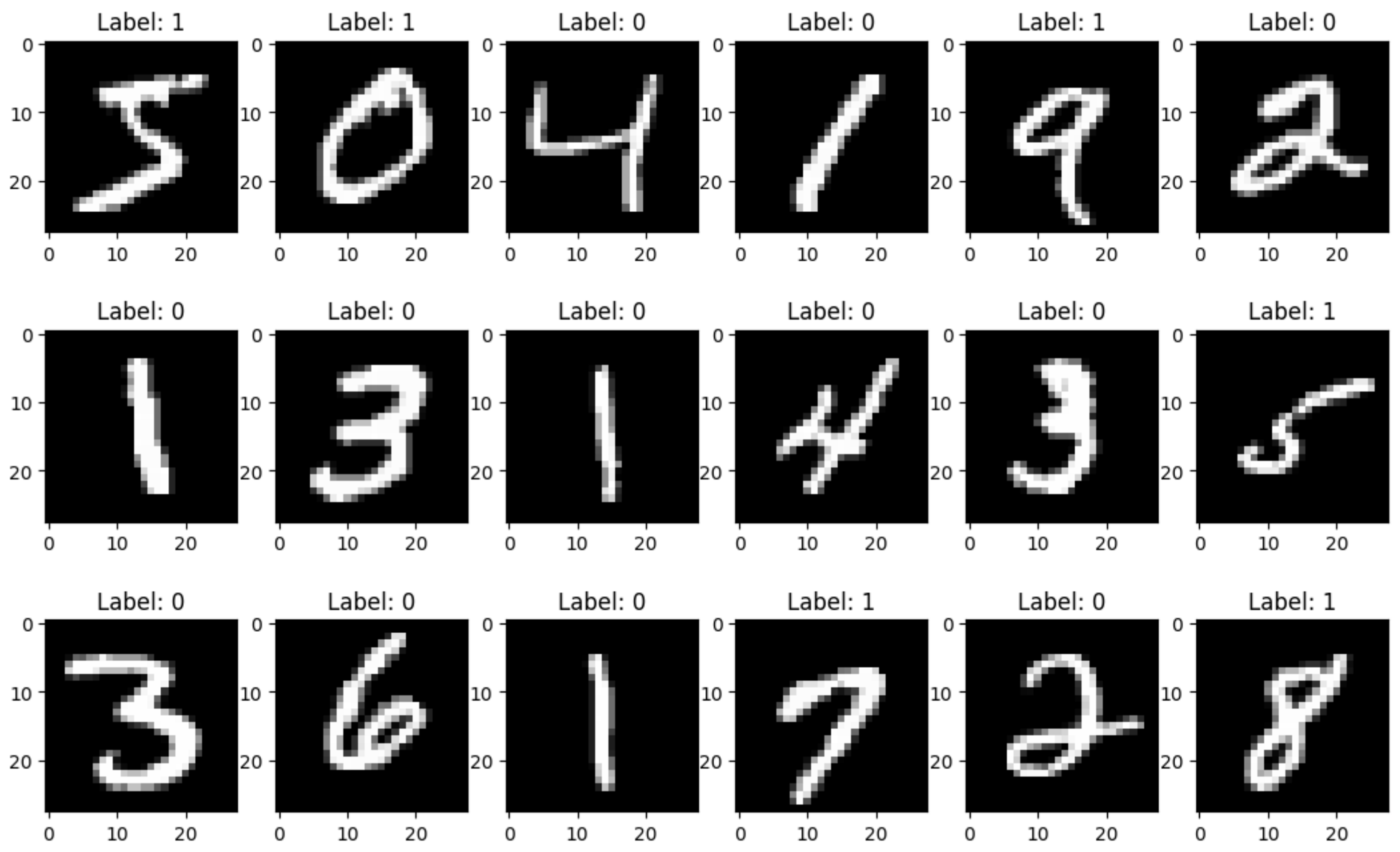
Data Model for OOD Generalization

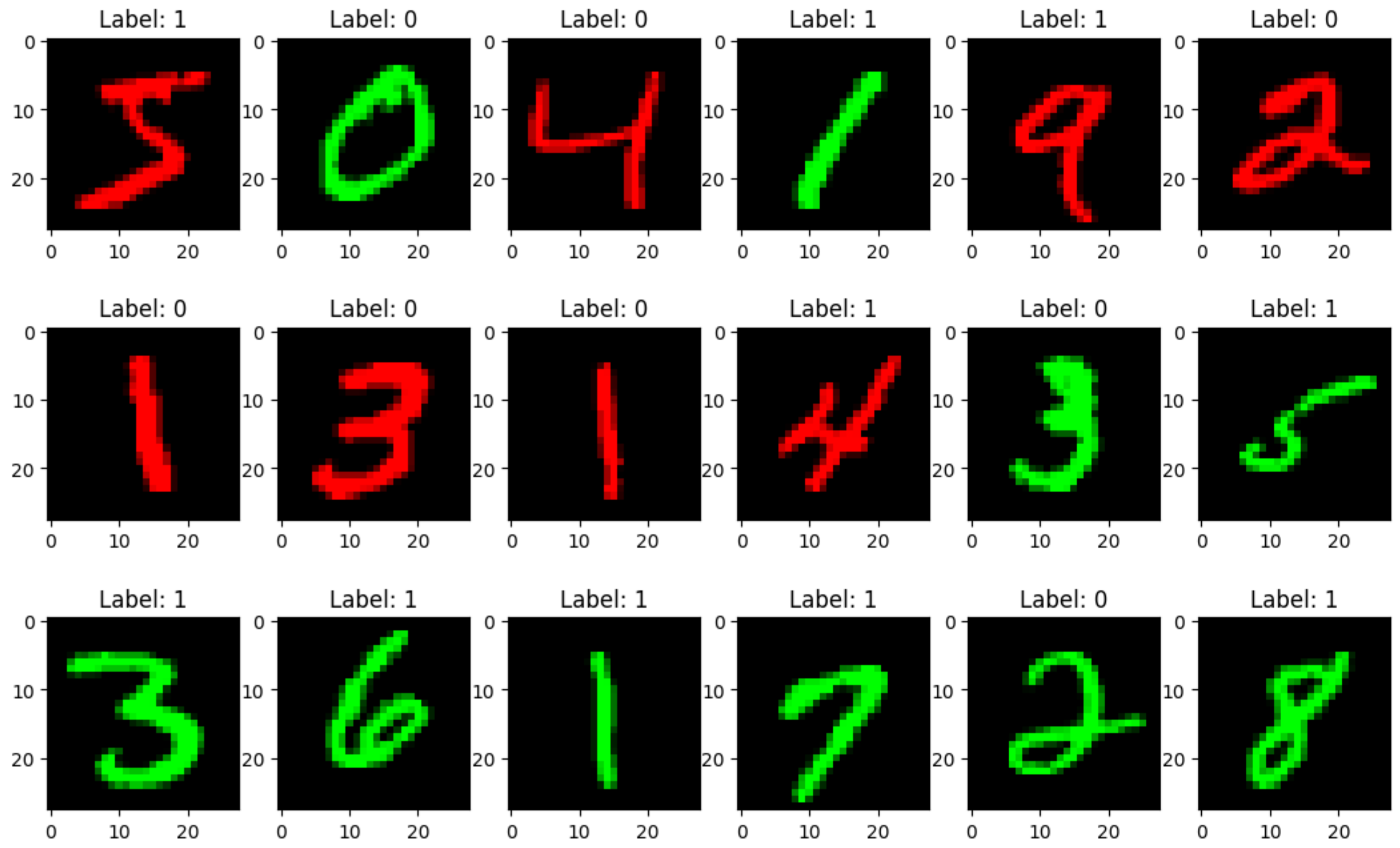
- Two classes $y = \{-1, +1\}$
- The input $\mathbf{x} \in \mathbb{R}^{2d}$ is composed of a feature patch $\mathbf{x}_1 \in \mathbb{R}^d$ and a noise patch $\mathbf{x}_2 \in \mathbb{R}^d$
- The feature patch $\mathbf{x}_1 \in \mathbb{R}^d$ is generated via:

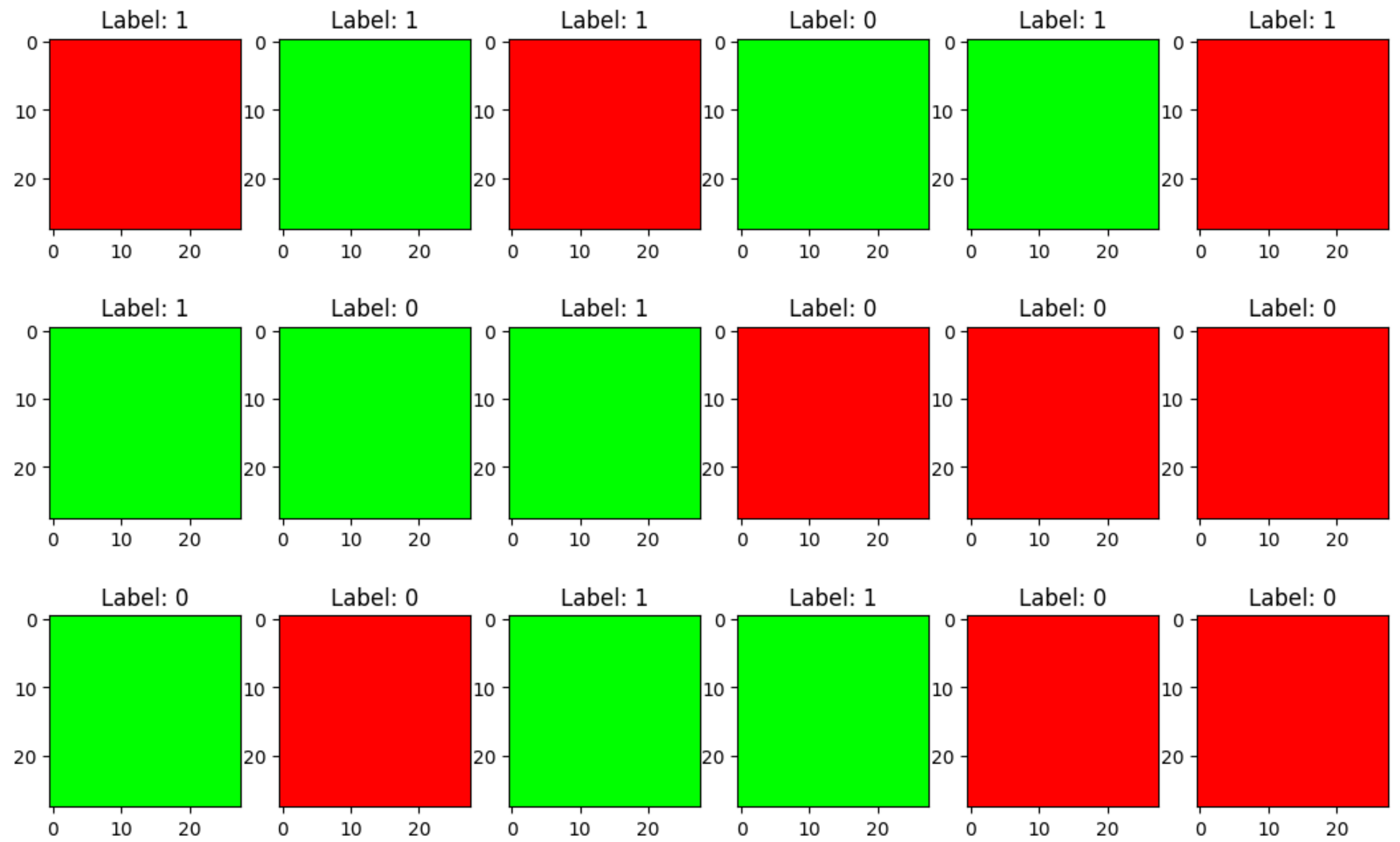
$$\mathbf{x}_1 = \boxed{y \cdot \mathrm{Rad}(\alpha) \cdot \mathbf{v}_1} + \boxed{y \cdot \mathrm{Rad}(\beta_e) \cdot \mathbf{v}_2}$$

Invariant signal

Spurious signal

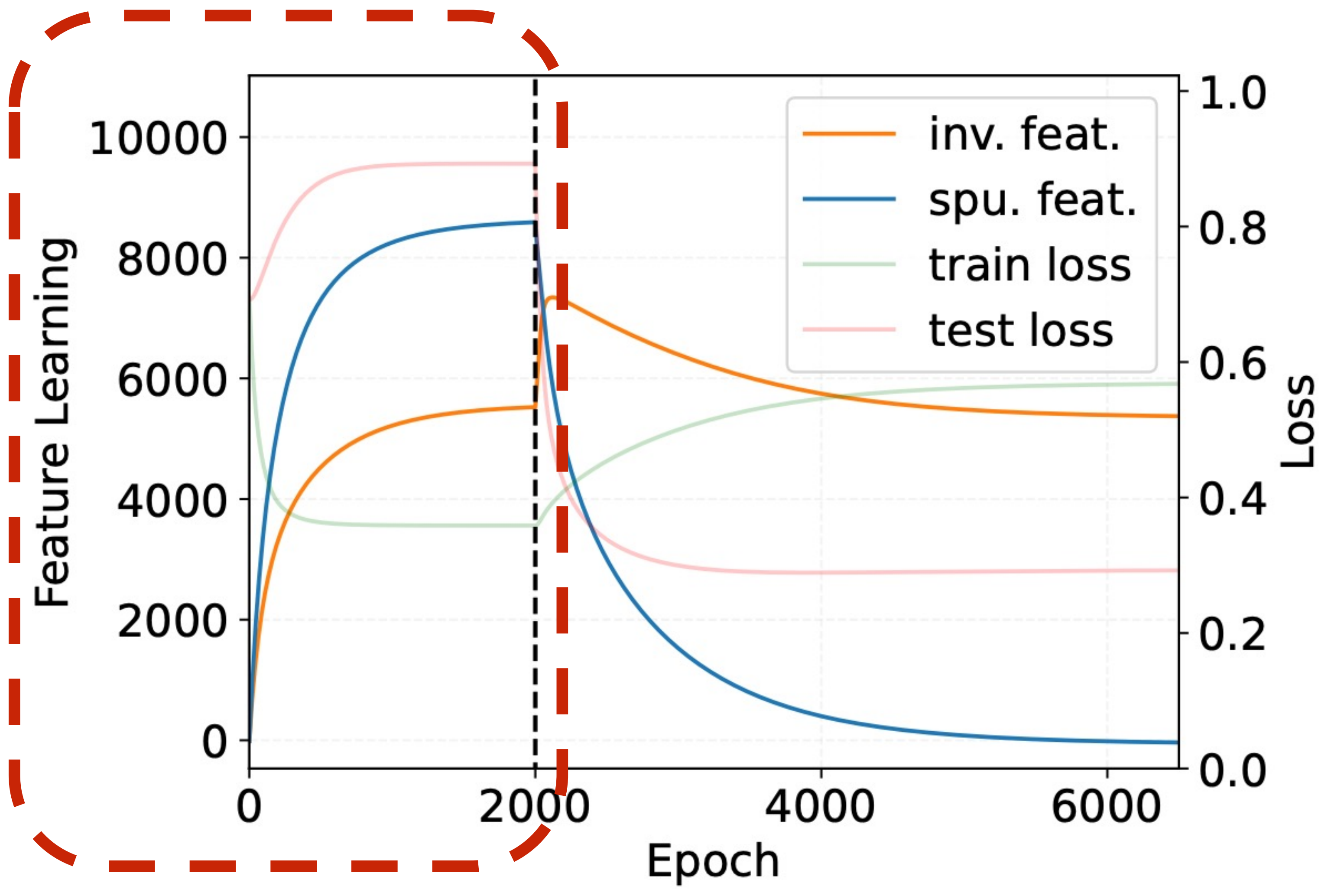






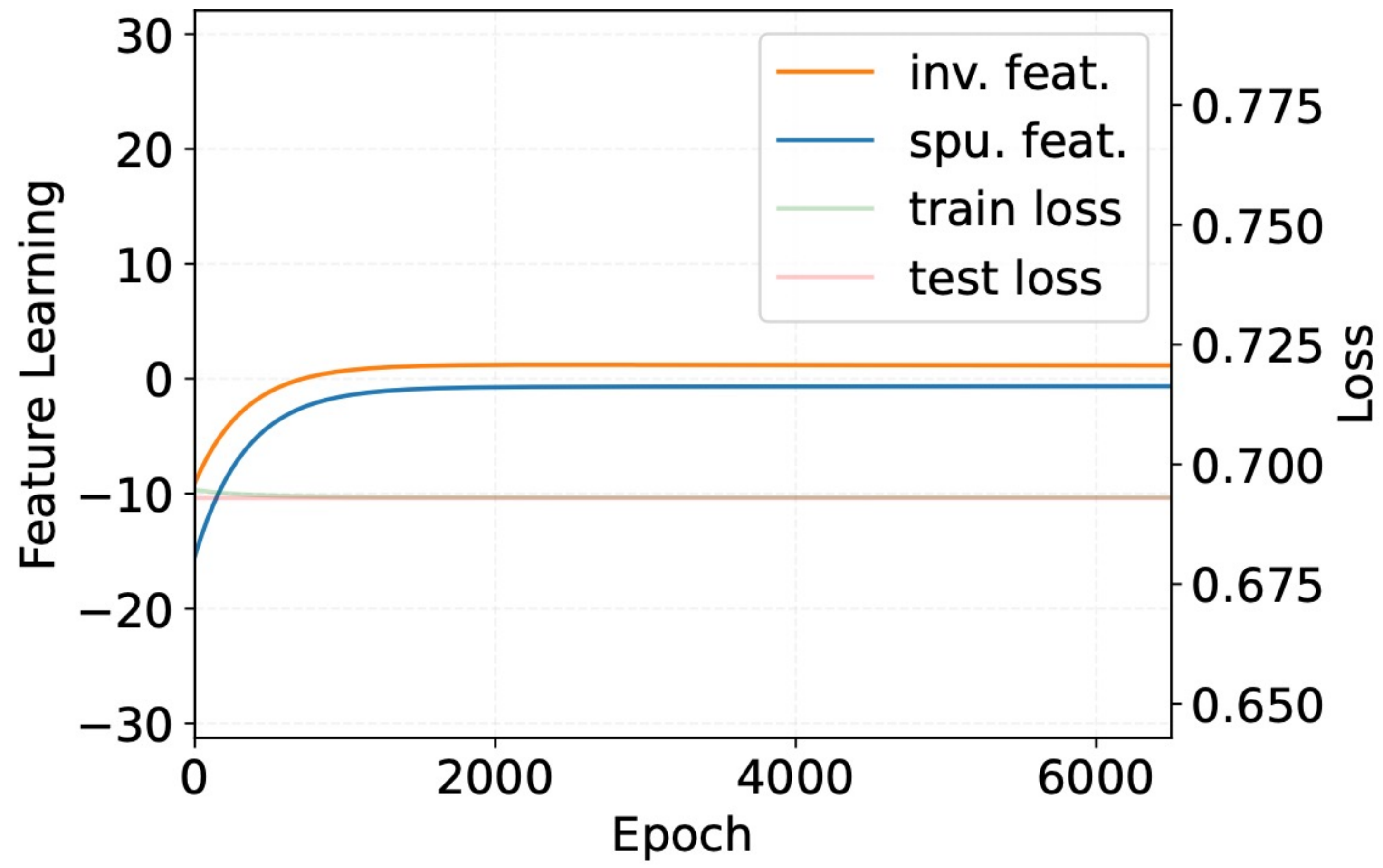
(Allen-Zhu & Li 2019)

ERM and IRM Feature Learning



ERM pre-training

FL w/ pre-training



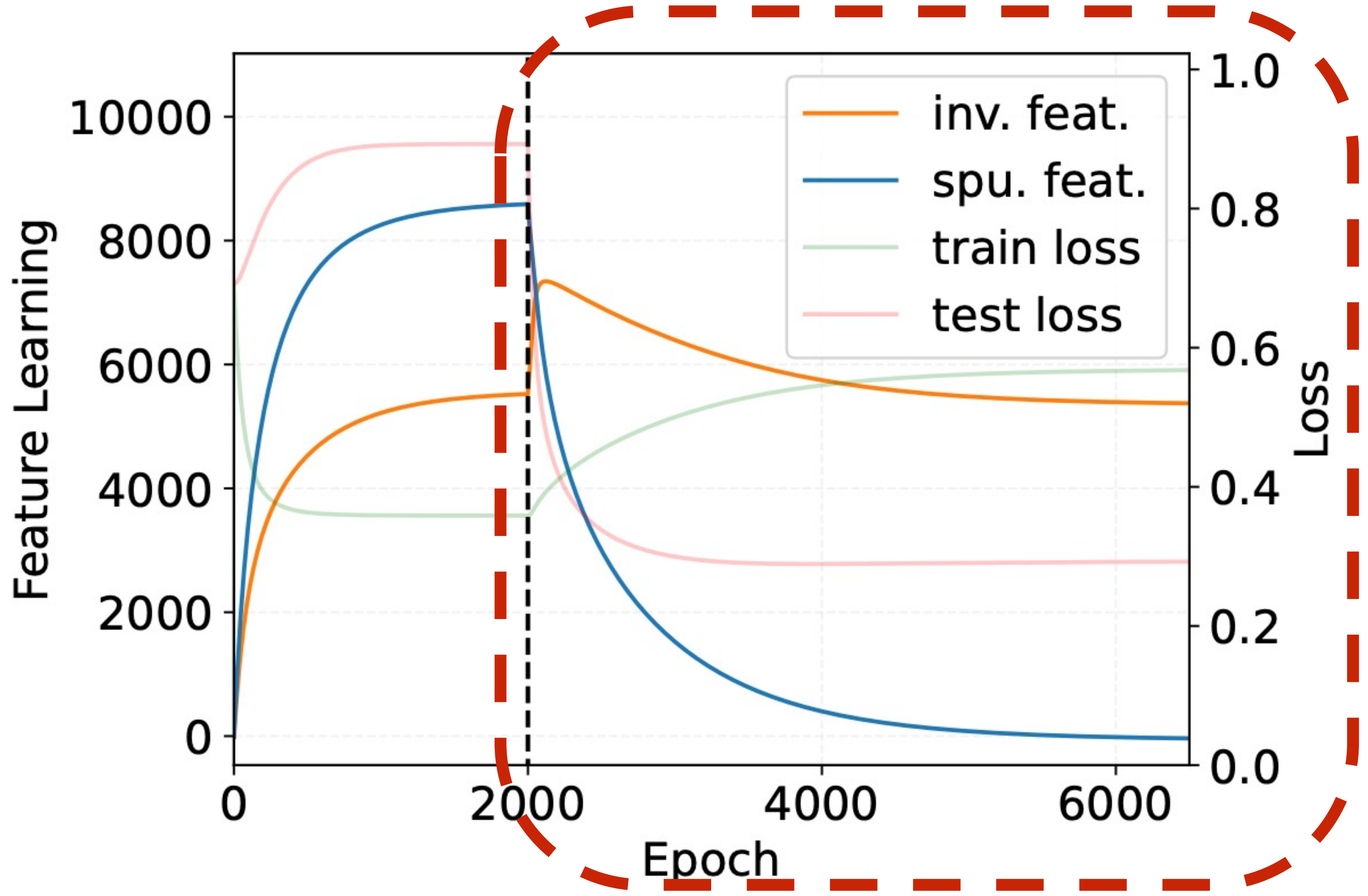
FL w/o pre-training

Theoretical Results (Informal):

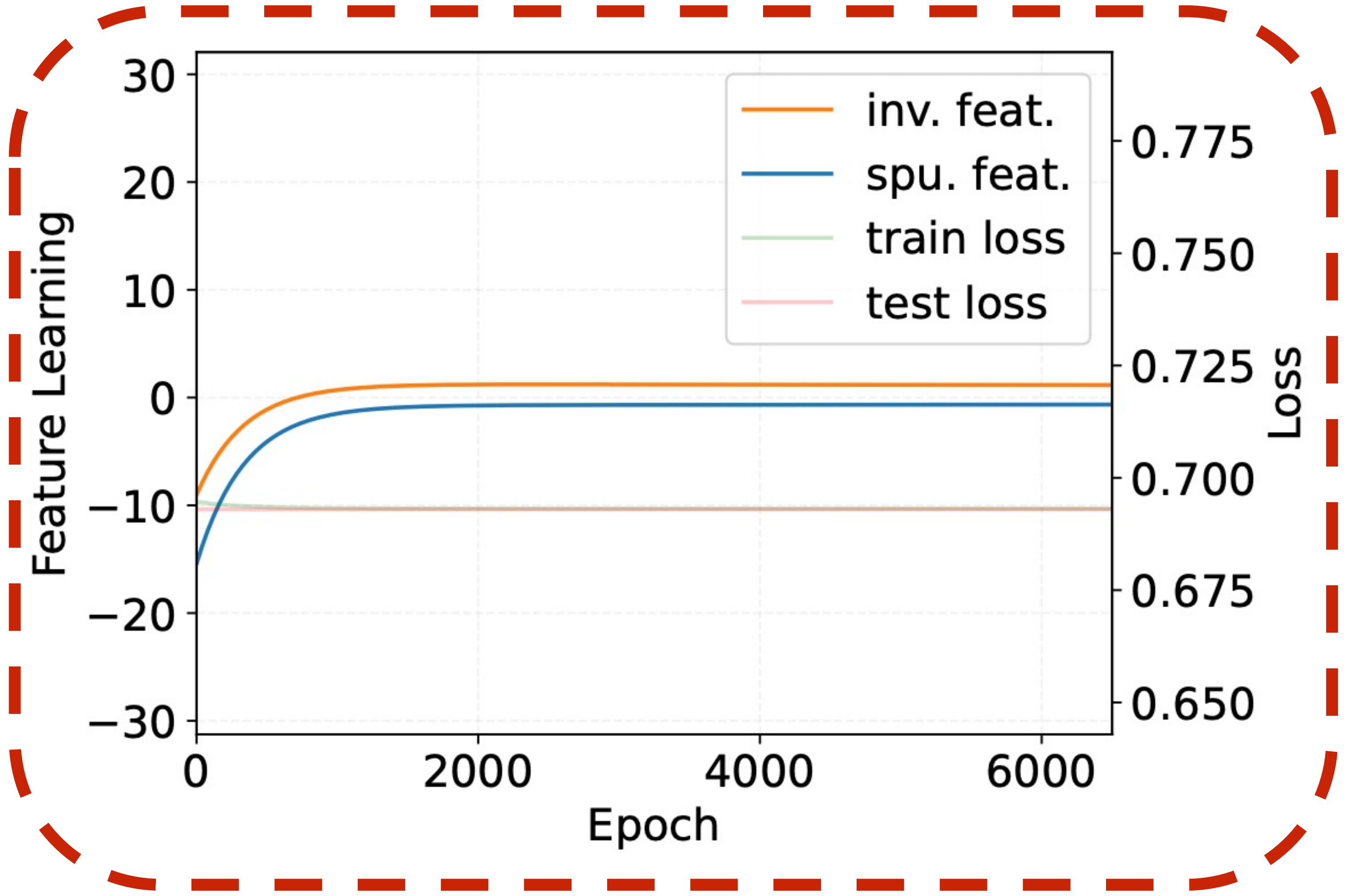
- ERM learns ***both*** invariant and spurious features.
- The invariant and spurious feature learning speed depends on the ***correlation strength*** with the labels.

ERM and IRM Feature Learning

OOD training with IRMv1



FL w/ pre-training

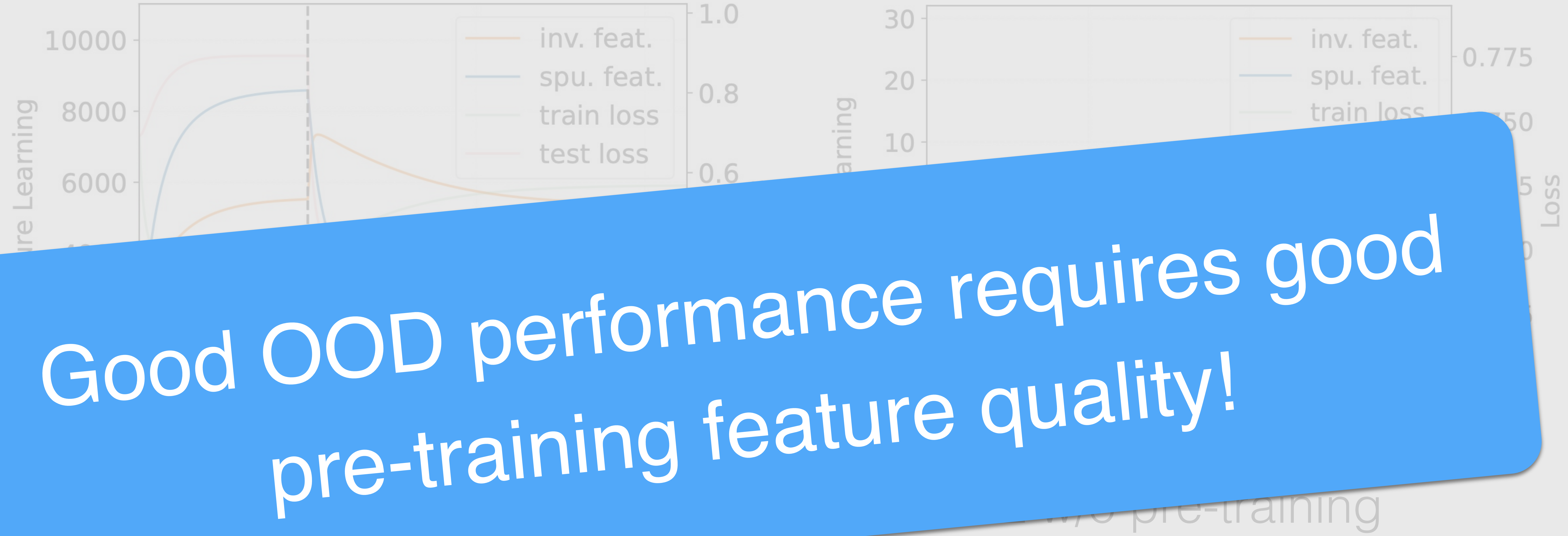


FL w/o pre-training

Theoretical Results (Informal):

- IRMv1 ***cannot*** learn any features even at the beginning of training;
- IRMv1 highly ***relies on*** ERM pre-training feature quality to extract invariant features.

ERM and IRM Feature Learning



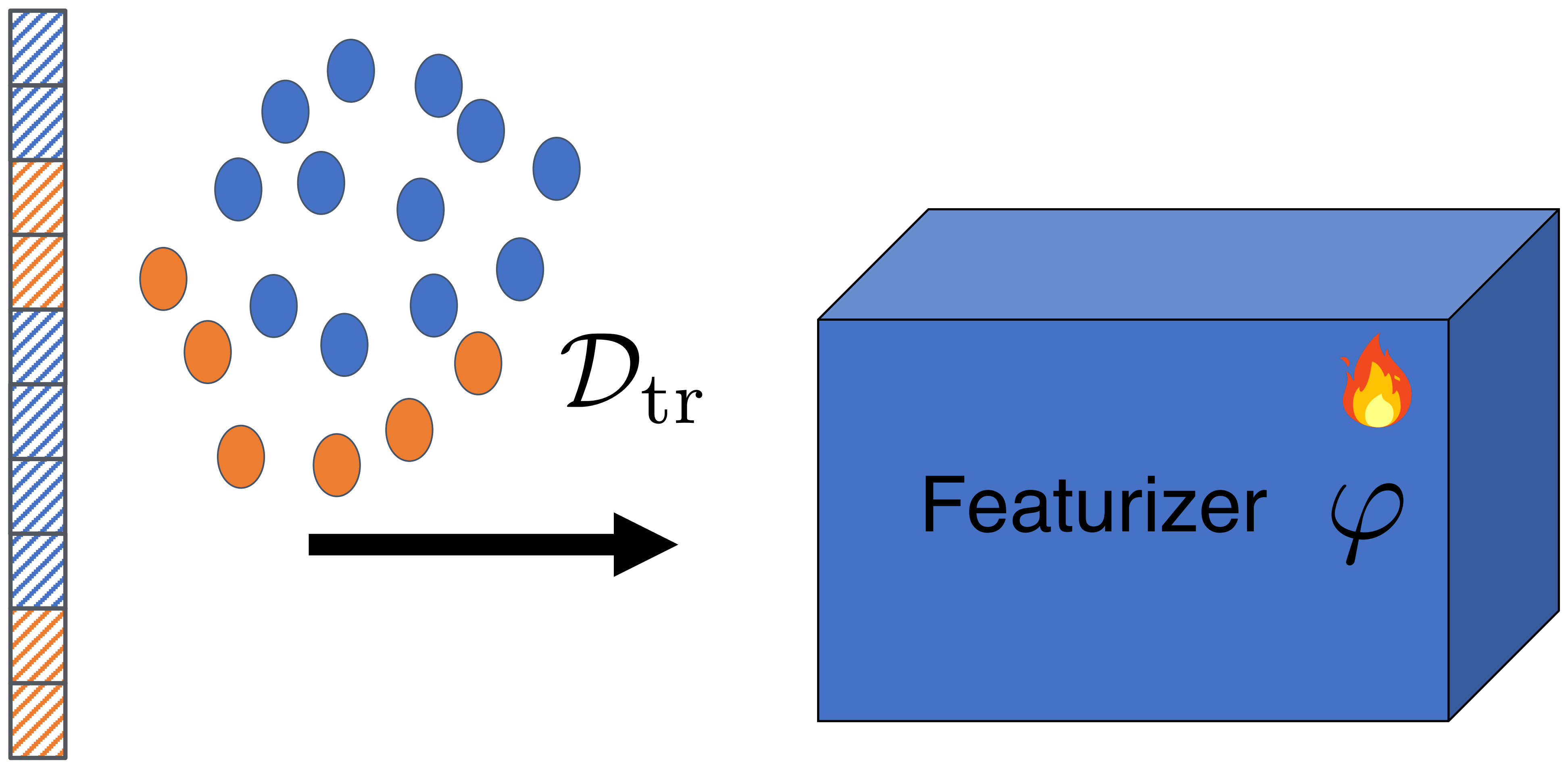
Good OOD performance requires good pre-training feature quality!

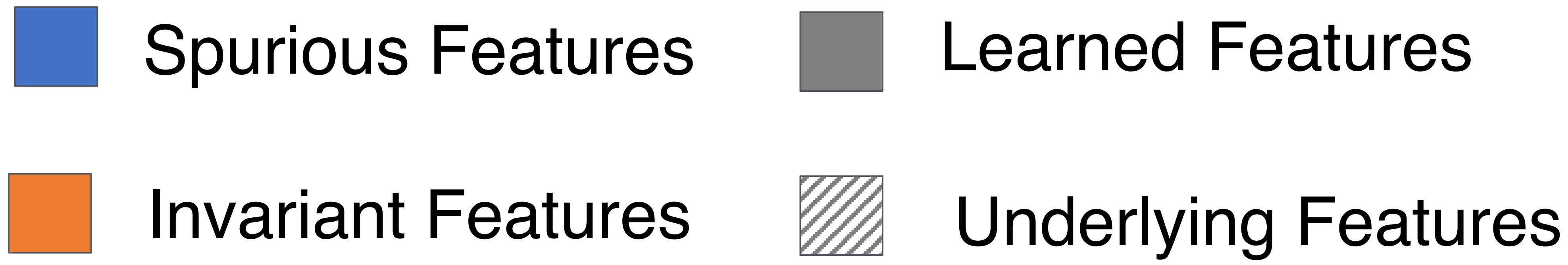
Theoretical Results (Informal):

- IRMv1 ***cannot*** learn any features even at the beginning of training;
- IRMv1 highly ***relies on*** ERM pre-training feature quality to extract invariant features.

Feature Learning with ERM

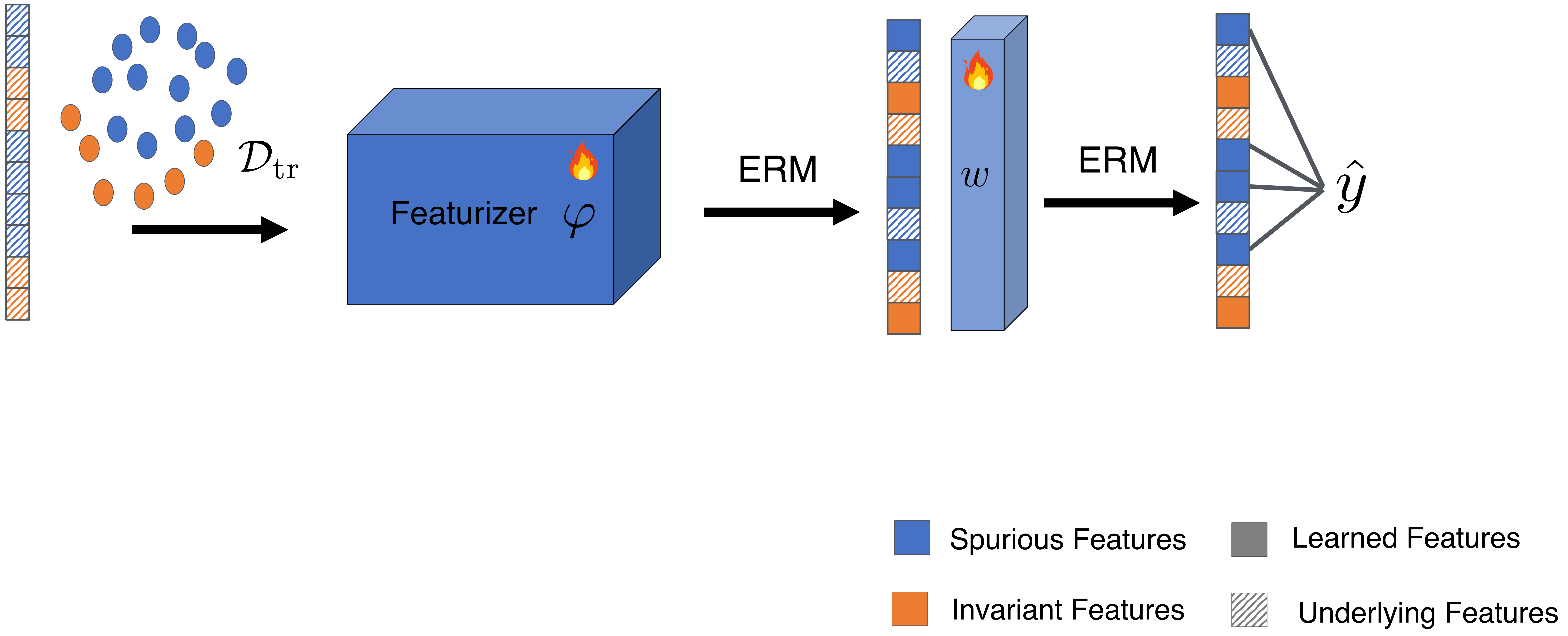
Consider the following dataset dominated by spurious features:





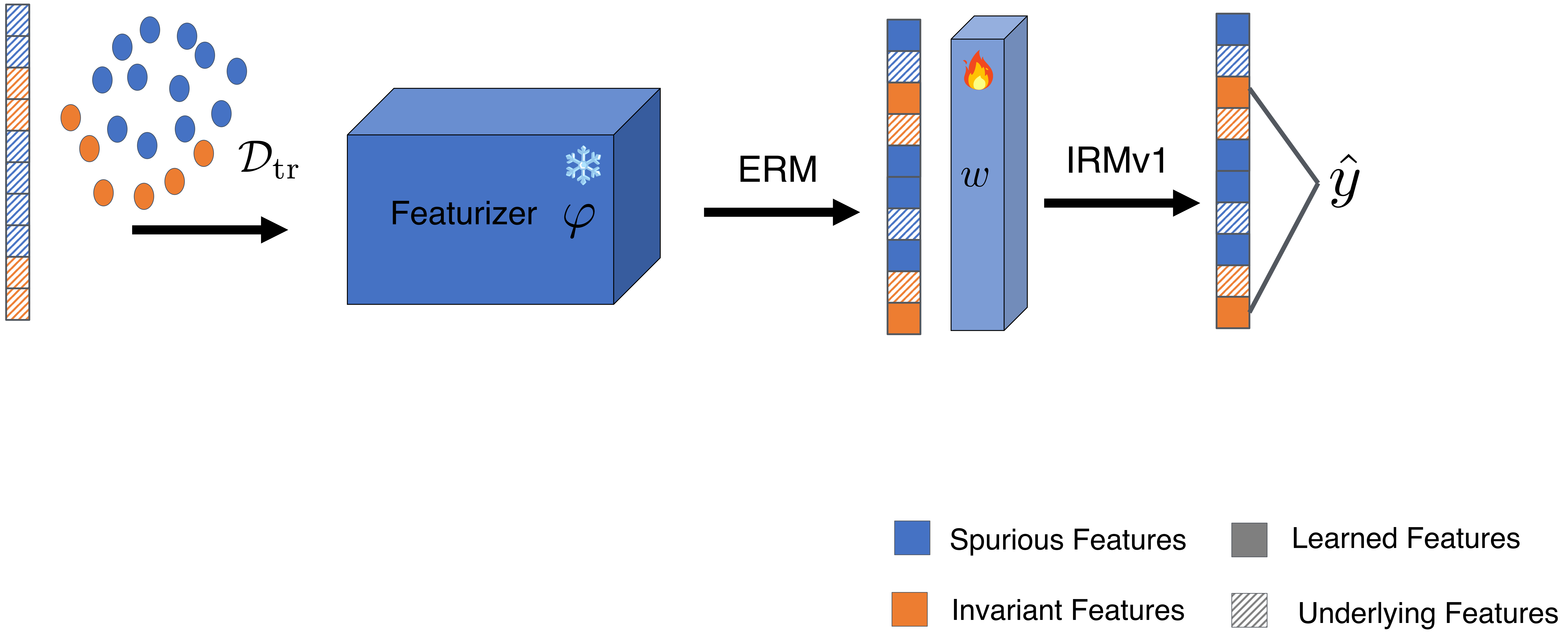
Feature Learning with ERM

ERM learns the spurious features ***more than*** the invariant features.



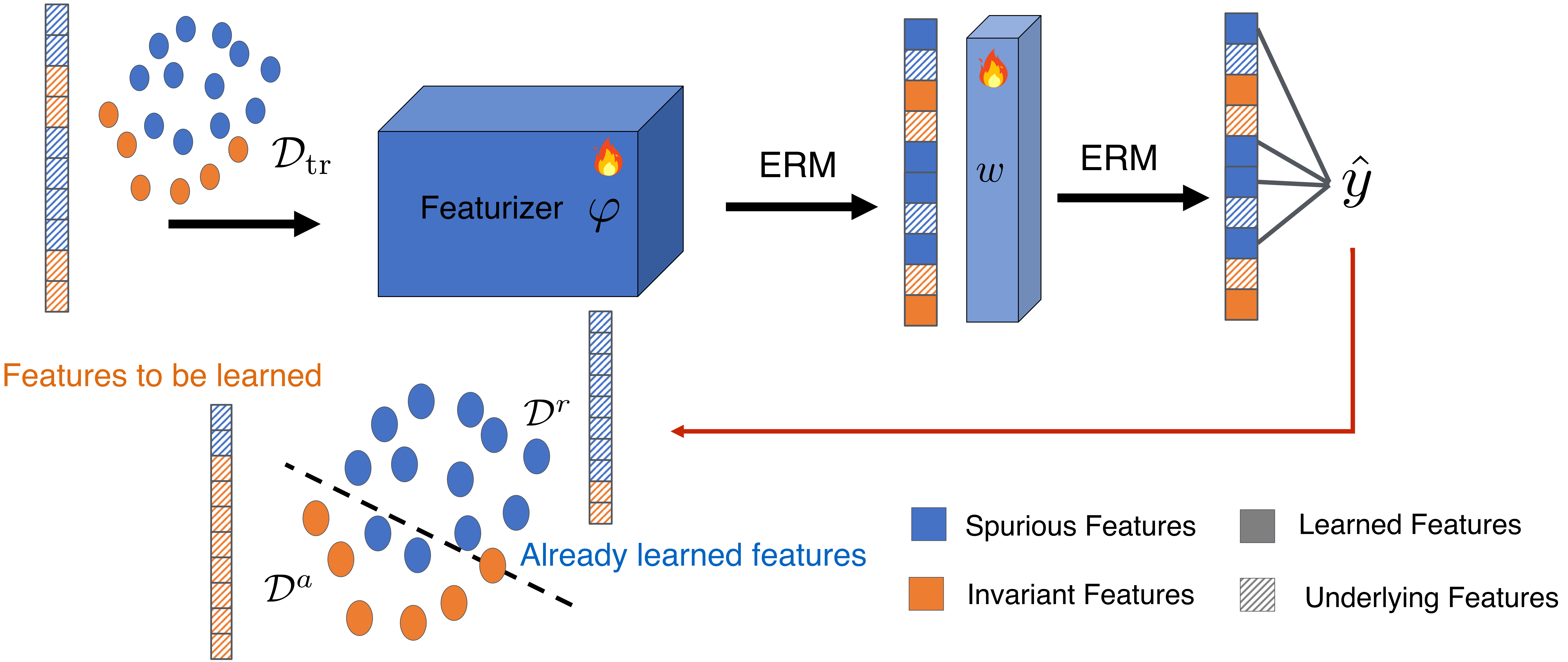
Feature Learning with ERM

OOD training can only leverage ***limited*** invariant features for prediction.



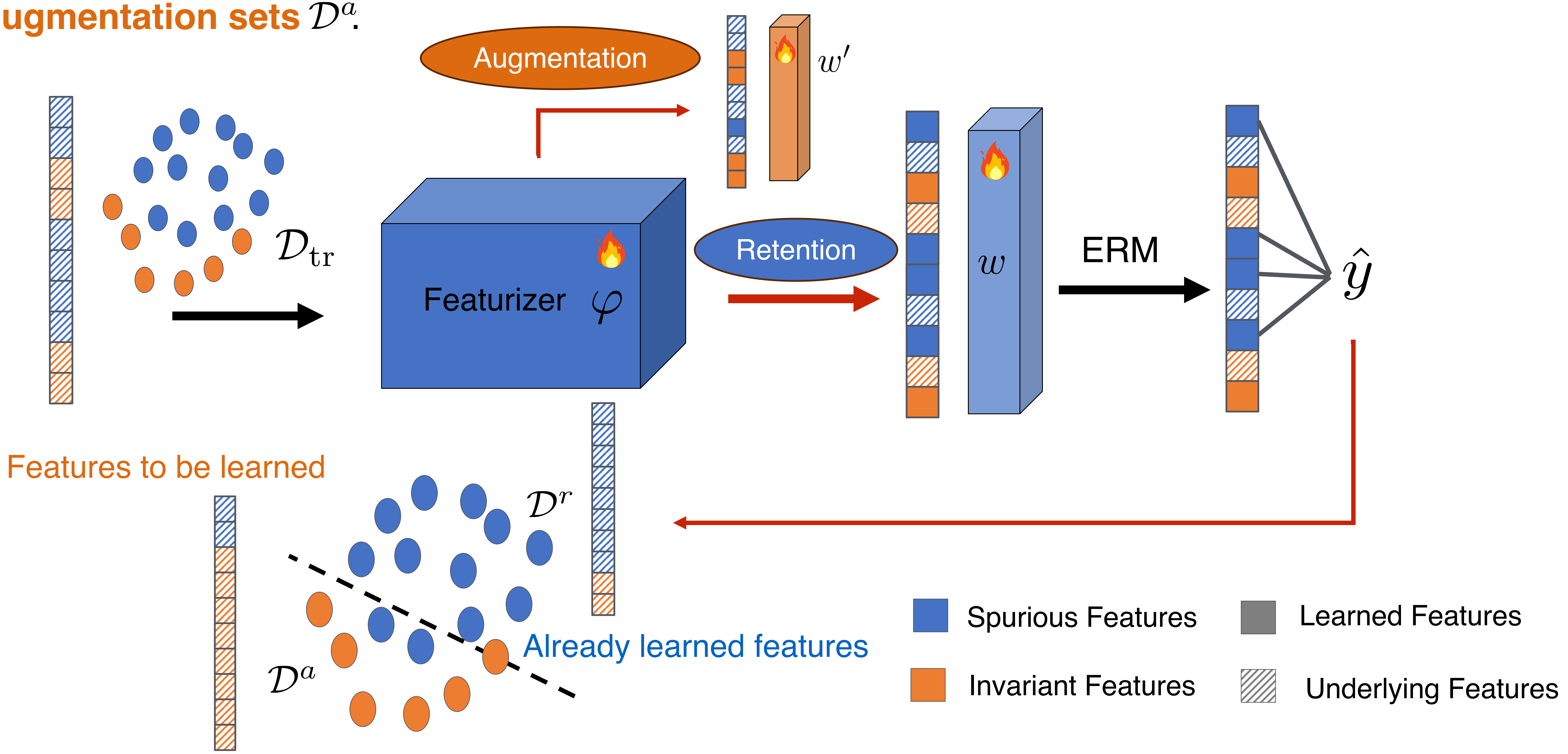
FeAT: Feature Augmented Training

Leveraging the feature learning information can partition the dataset into **retention sets** $\mathcal{D}^r$ and **augmentation sets** $\mathcal{D}^a$.



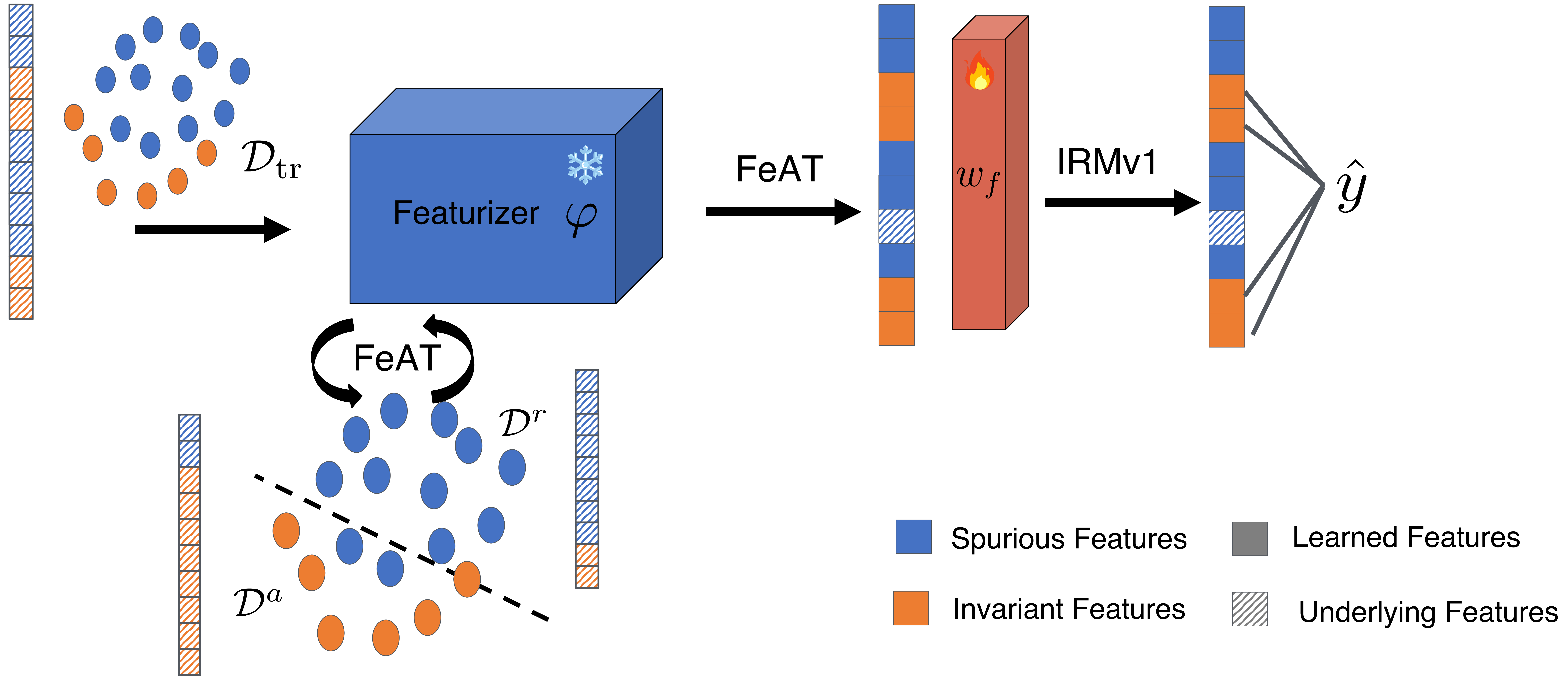
FeAT: Feature Augmented Training

Leveraging the feature learning information can partition the dataset into **retention sets** $\mathcal{D}^r$ and **augmentation sets** $\mathcal{D}^a$.



FeAT: Feature Augmented Training

Performing **feature augmentation** and **retention** several rounds, we can obtain richer feature representations that facilitate better OOD generalization.



Proof-of-Concept Experimental Results

FeAT boosts OOD performance of various objectives across various ColoredMNIST variant datasets.

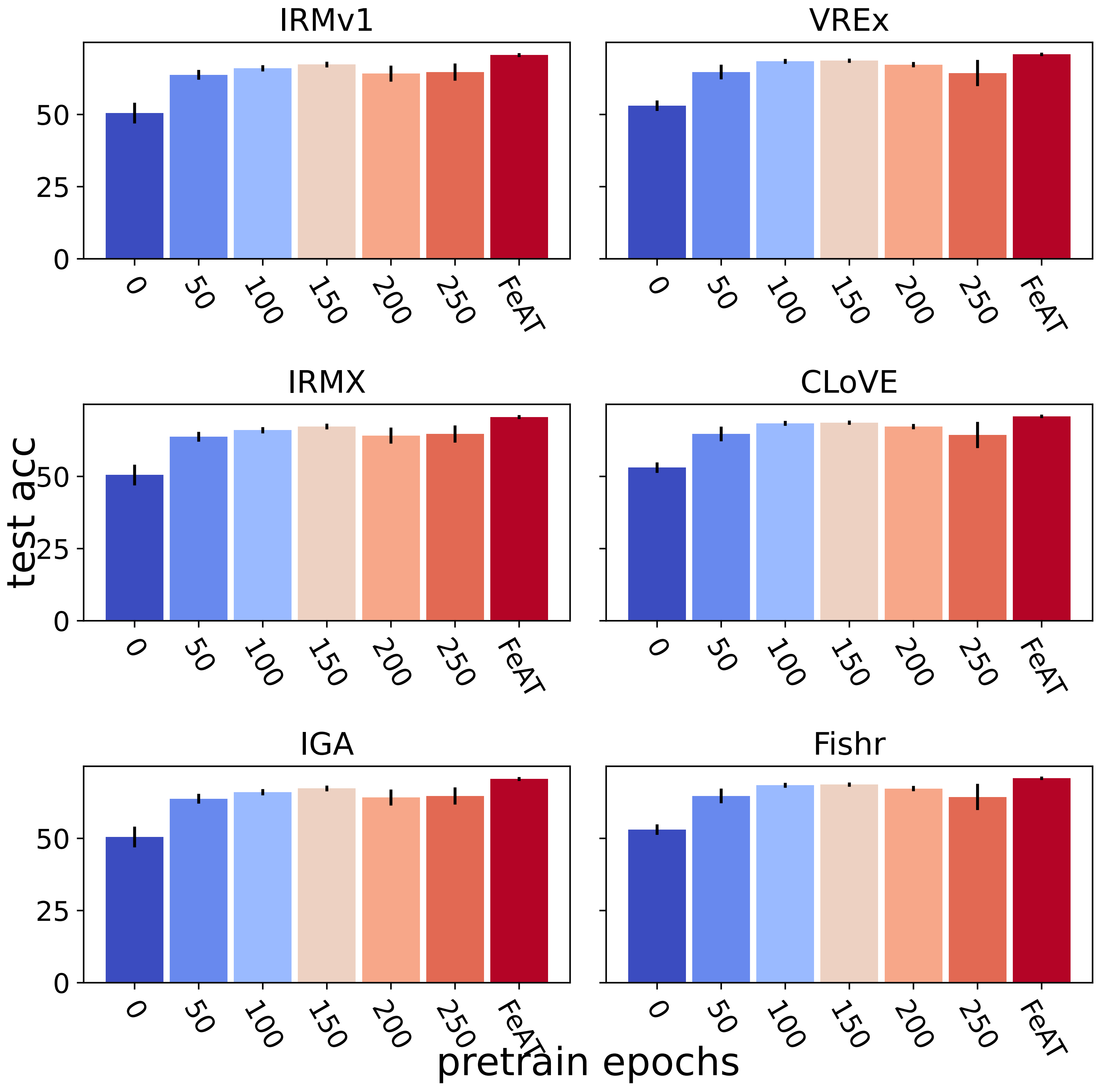


Table 1: OOD performance on ColoredMNIST datasets initialized with different representations.

	ColoredMNIST-025				ColoredMNIST-01			
	ERM-NF	ERM	Bonsai	FeAT	ERM-NF	ERM	Bonsai	FeAT
ERM	17.14 (±0.73)	12.40 (±0.32)	11.21 (±0.49)	**17.27** (±2.55)	73.06 (±0.71)	73.75 (±0.49)	70.95 (±0.93)	**76.05** (±1.45)
IRMv1	67.29 (±0.99)	59.81 (±4.46)	70.28 (±0.72)	**70.57** (±0.68)	76.89 (±3.25)	73.84 (±0.56)	76.71 (±4.10)	**82.33** (±1.77)
V-Rex	68.62 (±0.73)	65.96 (±1.29)	70.31 (±0.66)	**70.82** (±0.59)	83.52 (±2.52)	81.20 (±3.27)	82.61 (±1.76)	**84.70** (±0.69)
IRMX	67.00 (±1.95)	64.05 (±0.88)	70.46 (±0.42)	**70.78** (±0.61)	81.61 (±1.98)	75.97 (±0.88)	80.28 (±1.62)	**84.34** (±0.97)
IB-IRM	56.09 (±2.04)	59.81 (±4.46)	70.28 (±0.72)	**70.57** (±0.68)	75.81 (±0.63)	73.84 (±0.56)	76.71 (±4.10)	**82.33** (±1.77)
CLOvE	58.67 (±7.69)	65.78 (±0.00)	65.57 (±3.02)	**65.78** (±2.68)	75.66 (±10.6)	74.73 (±0.36)	72.73 (±1.18)	**75.12** (±1.08)
IGA	51.22 (±3.67)	62.43 (±3.06)	**70.17** (±0.89)	67.11 (±3.40)	74.20 (±2.45)	73.74 (±0.48)	74.72 (±3.60)	**83.46** (±2.17)
Fishr	69.38 (±0.39)	67.74 (±0.90)	68.75 (±1.10)	**70.56** (±0.97)	77.29 (±1.61)	82.23 (±1.35)	84.19 (±0.66)	**84.26** (±0.93)
Oracle	71.97 (±0.34)				86.55 (±0.27)			

Stronger spurious signal

Stronger invariant signal

Real-World Experimental Results

FeAT boosts OOD performance of various objectives across ***6*** challenging real-world OOD datasets.

Table 2: OOD generalization performances on WILDS benchmark.

INIT.	METHOD	CAMELYON17	CIVILCOMMENTS	FMOW	IWILDCAM	AMAZON	RXRX1
		Avg. acc. (%)	Worst acc. (%)	Worst acc. (%)	Macro F1	10-th per. acc. (%)	Avg. acc. (%)
ERM	DFR†	95.14 (±1.96)	**77.34** (±0.50)	41.96 (±1.90)	23.15 (±0.24)	48.00 (±0.00)	-
ERM	DFR-s†	-	82.24 (±0.13)	56.17 (±0.62)	52.44 (±0.34)	-	-
Bonsai	DFR†	95.17 (±0.18)	77.07 (±0.85)	43.26 (±0.82)	21.36 (±0.41)	46.67 (±0.00)	-
Bonsai	DFR-s†	-	81.26 (±1.86)	58.58 (±1.17)	50.85 (±0.18)	-	-
FAT	DFR†	**95.28** (±0.19)	**77.34** (±0.59)	**43.54** (±1.26)	**23.54** (±0.52)	**49.33** (±0.00)	-
FAT	DFR-s†	-	79.56 (±0.38)	57.69 (±0.78)	52.31 (±0.38)	-	-
ERM	ERM	74.30 (±5.96)	55.53 (±1.78)	33.58 (±1.02)	28.22 (±0.78)	51.11 (±0.63)	30.21 (±0.09)
ERM	GroupDRO	76.09 (±6.46)	69.50 (±0.15)	33.03 (±0.52)	28.51 (±0.58)	52.00 (±0.00)	29.99 (±0.13)
ERM	IRMv1	75.68 (±7.41)	68.84 (±0.95)	33.45 (±1.07)	28.76 (±0.45)	52.00 (±0.00)	30.10 (±0.05)
ERM	V-REx	71.60 (±7.88)	69.03 (±1.08)	33.06 (±0.46)	28.82 (±0.47)	52.44 (±0.63)	29.88 (±0.35)
ERM	IRMX	73.49 (±9.33)	68.91 (±1.19)	33.13 (±0.86)	28.82 (±0.47)	52.00 (±0.00)	30.10 (±0.05)
Bonsai	ERM	73.98 (±5.30)	63.34 (±3.49)	31.91 (±0.51)	28.27 (±1.05)	48.58 (±0.56)	24.22 (±0.44)
Bonsai	GroupDRO	72.82 (±5.37)	70.23 (±1.33)	33.12 (±1.20)	27.16 (±1.18)	42.67 (±1.09)	22.95 (±0.46)
Bonsai	IRMv1	73.59 (±6.16)	68.39 (±2.01)	32.51 (±1.23)	27.60 (±1.57)	47.11 (±0.63)	23.35 (±0.43)
Bonsai	V-REx	76.39 (±5.32)	68.67 (±1.29)	33.17 (±1.26)	25.81 (±0.42)	48.00 (±0.00)	23.34 (±0.42)
Bonsai	IRMX	64.77 (±10.1)	69.56 (±0.95)	32.63 (±0.75)	27.62 (±0.66)	46.67 (±0.00)	23.34 (±0.40)
FAT	ERM	77.80 (±2.48)	68.11 (±2.27)	33.13 (±0.78)	28.47 (±0.67)	**52.89** (±0.63)	**30.66** (±0.42)
FAT	GroupDRO	**80.41** (±3.30)	**71.29** (±0.46)	33.55 (±1.67)	28.38 (±1.32)	52.58 (±0.56)	29.99 (±0.11)
FAT	IRMv1	77.97 (±3.09)	70.33 (±1.14)	**34.04** (±0.70)	**29.66** (±1.52)	**52.89** (±0.63)	29.99 (±0.19)
FAT	V-REx	75.12 (±6.55)	70.97 (±1.06)	34.00 (±0.71)	29.48 (±1.94)	**52.89** (±0.63)	30.57 (±0.53)
FAT	IRMX	76.91 (±6.76)	71.18 (±1.10)	33.99 (±0.73)	29.04 (±2.96)	**52.89** (±0.63)	29.92 (±0.16)

†DFR/DFR-s use an additional OOD dataset to evaluate invariant and spurious feature learning, respectively.

Real-World Experimental Results

FeAT boosts OOD performance of various objectives across ***6*** challenging real-world OOD generalization datasets.

Table 2: OOD generalization performances on WILDS benchmark.

INIT.	METHOD	CAMELYON17 Avg. acc. (%)	CIVILCOMMENTS Worst acc. (%)	FMOW Worst acc. (%)	IWILDCAM Macro F1	AMAZON 10-th per. acc. (%)	RXRX1 Avg. acc. (%)
ERM	DFR†	95.14 (±1.96)	**77.34** (±0.50)	41.96 (±1.90)	23.15 (±0.24)	48.00 (±0.00)	-
ERM	DFR-s†	-	82.24 (±0.13)	56.17 (±0.62)	52.44 (±0.34)	-	-
Bonsai	DFR†	95.17 (±0.18)	77.07 (±0.85)	43.26 (±0.82)	21.36 (±0.41)	46.67 (±0.00)	-
Bonsai	DFR-s†	-	81.26 (±1.86)	58.58 (±1.17)	50.85 (±0.18)	-	-
FeAT	DFR†	**95.28** (±0.19)	**77.34** (±0.59)	**43.54** (±1.26)	**23.54** (±0.52)	**49.33** (±0.00)	-
FeAT	DFR-s†	-	79.56 (±0.38)	57.69 (±0.78)	52.31 (±0.38)	-	-
ERM	ERM	74.30 (±5.96)	55.53 (±1.78)	33.58 (±1.02)	28.22 (±0.78)	51.11 (±0.63)	30.21 (±0.09)
ERM	GroupDRO	76.09 (±6.46)	69.50 (±0.15)	33.03 (±0.52)	28.51 (±0.58)	52.00 (±0.00)	29.99 (±0.13)
ERM	IRMv1	75.68 (±7.41)	68.84 (±0.95)	33.45 (±1.07)	28.76 (±0.45)	52.00 (±0.00)	30.10 (±0.05)
ERM	V-REx	71.60 (±7.88)	69.03 (±1.08)	33.06 (±0.46)	28.82 (±0.47)	52.44 (±0.63)	29.88 (±0.35)
ERM	IRMX	73.49 (±9.33)	68.91 (±1.19)	33.13 (±0.86)	28.82 (±0.47)	52.00 (±0.00)	30.10 (±0.05)
Bonsai	ERM	73.98 (±5.30)	63.34 (±3.49)	31.91 (±0.51)	28.27 (±1.05)	48.58 (±0.56)	24.22 (±0.44)
Bonsai	GroupDRO	72.82 (±5.37)	70.23 (±1.33)	33.12 (±1.20)	27.16 (±1.18)	42.67 (±1.09)	22.95 (±0.46)
Bonsai	IRMv1	73.59 (±6.16)	68.39 (±2.01)	32.51 (±1.23)	27.60 (±1.57)	47.11 (±0.63)	23.35 (±0.43)
Bonsai	V-REx	76.39 (±5.32)	68.67 (±1.29)	33.17 (±1.26)	25.81 (±0.42)	48.00 (±0.00)	23.34 (±0.42)
Bonsai	IRMX	64.77 (±10.1)	69.56 (±0.95)	32.63 (±0.75)	27.62 (±0.66)	46.67 (±0.00)	23.34 (±0.40)
FeAT	ERM	77.80 (±2.48)	68.11 (±2.27)	33.13 (±0.78)	28.47 (±0.67)	**52.89** (±0.63)	**30.66** (±0.42)
FeAT	GroupDRO	**80.41** (±3.30)	**71.29** (±0.46)	33.55 (±1.67)	28.38 (±1.32)	52.58 (±0.56)	29.99 (±0.11)
FeAT	IRMv1	77.97 (±3.09)	70.33 (±1.14)	**34.04** (±0.70)	**29.66** (±1.52)	**52.89** (±0.63)	29.99 (±0.19)
FeAT	V-REx	75.12 (±6.55)	70.97 (±1.06)	34.00 (±0.71)	29.48 (±1.94)	**52.89** (±0.63)	30.57 (±0.53)
FeAT	IRMX	76.91 (±6.76)	71.18 (±1.10)	33.99 (±0.73)	29.04 (±2.96)	**52.89** (±0.63)	29.92 (±0.16)

†DFR/DFR-s use an additional OOD dataset to evaluate invariant and spurious feature learning, respectively.

FeAT Learns Richer Meaningful Features

(a) ERM 150 epochs (b) ERM 300 epochs (c) ERM 450 epochs (d) FeAT 2 rounds

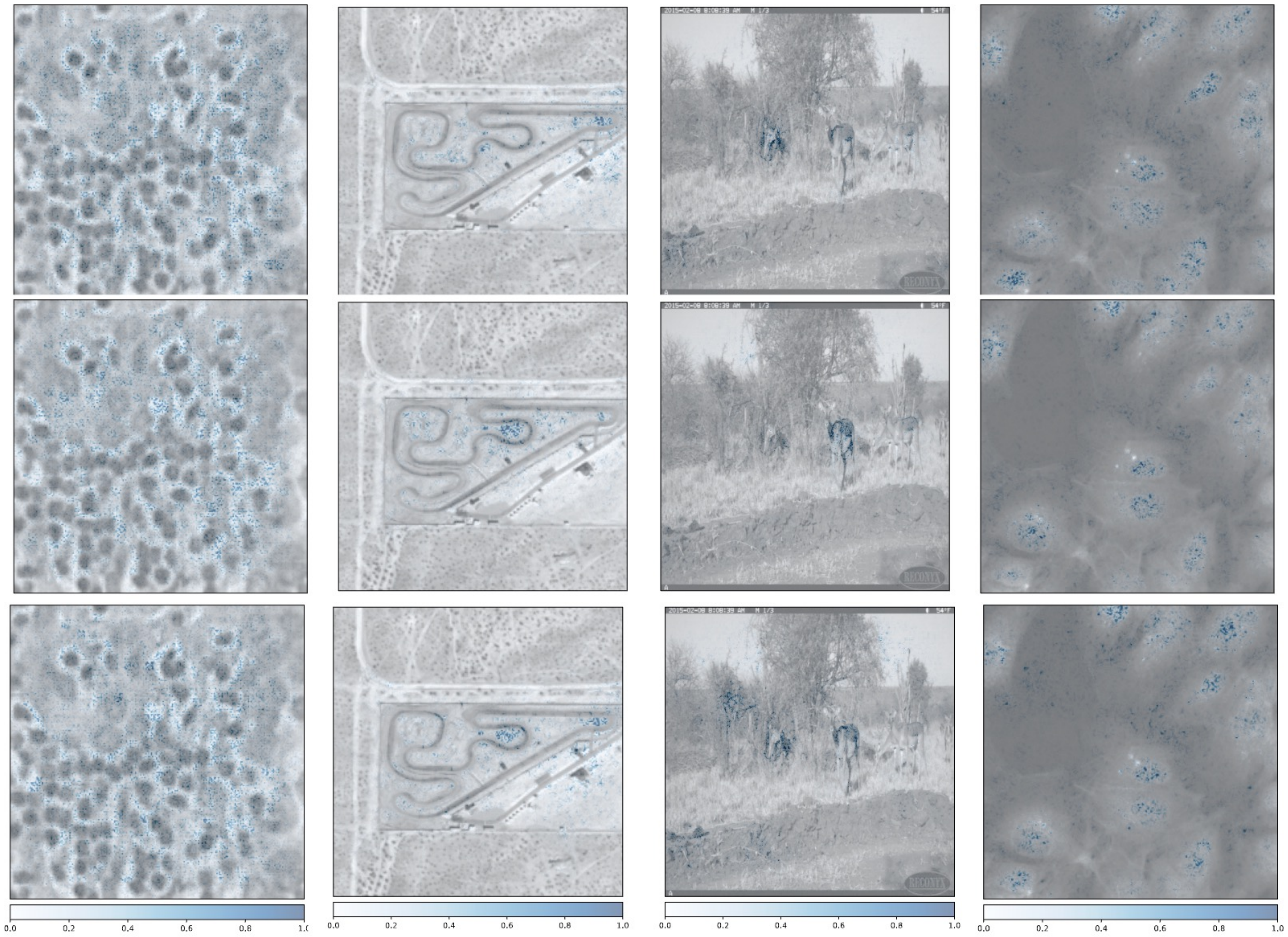
Figure 1: GradCAM visualization on COLOREDMNIST-025, where the shortcuts are now concentrated to a colored path at the up left. Three visualizations are drawn for each sample: the original figure, the gray-colored gradcam, and the gradcam. It can be found that ERM can not properly capture the desired features or even forget certain features with longer training epochs. FAT can stably capture the desired features.

FeAT Learns Richer Meaningful Features

ERM

Bonsai

FeAT



(i) CAMELYON17 (j) FMOW (k) IWILDCAM (l) RXRX1

Summary

We established a feature learning framework and theoretically revealed that ERM will learn both invariant and spurious features.

We also show that the performance of OOD objectives like IRM highly rely on the features quality, which motivates to learn richer features before OOD training.

We propose a novel rich feature learning algorithm FAT and conduct extensive experiments in challenging OOD benchmarks to verify the effectiveness of FAT.

Paper Code

Thank you!

Contact: yqchen@cse.cuhk.edu.hk