CONDA: Adaptive Concept Bottleneck for Foundation Models Under Distribution Shifts

Jihye Choi, Jayaram Raghuram, Yixuan Li, Somesh Jha

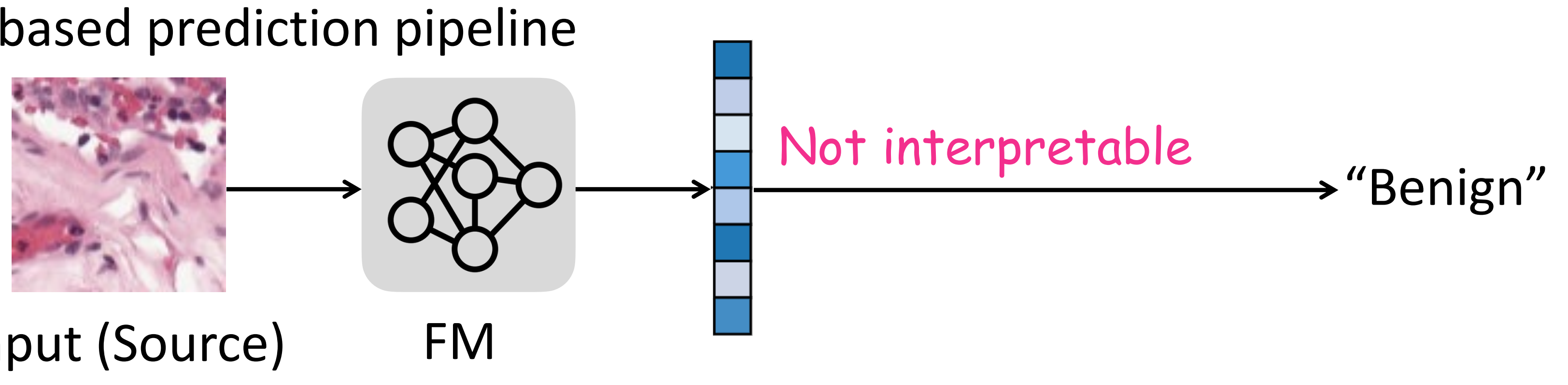
✉ jihye@cs.wisc.edu   ⌂ jihyechoi77@github.io

PAPER   CODE

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

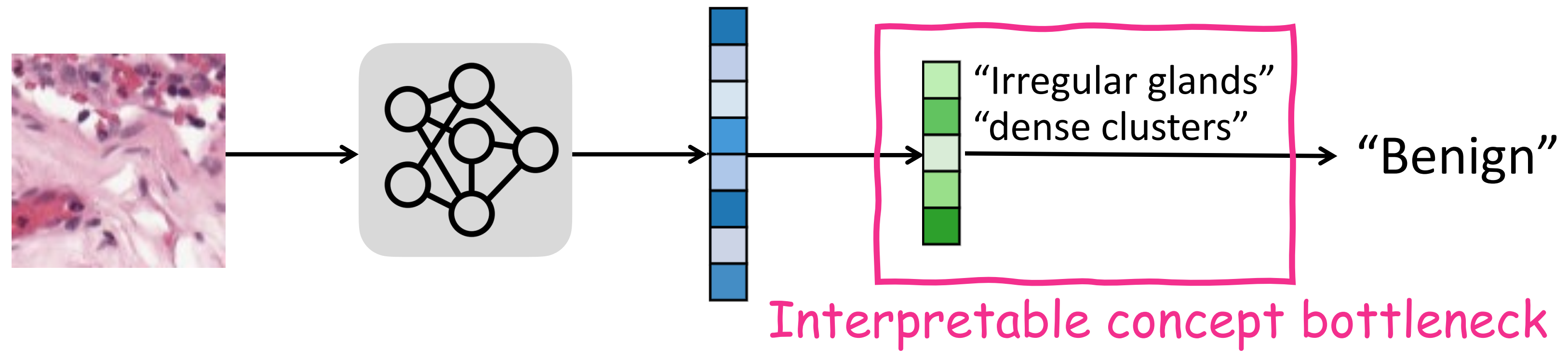
Background: Concept Bottleneck for FMs

Feature-based prediction pipeline



vs

Transformed into Concept-based prediction pipeline (PCBM: YWZ, ICLR'23, …)

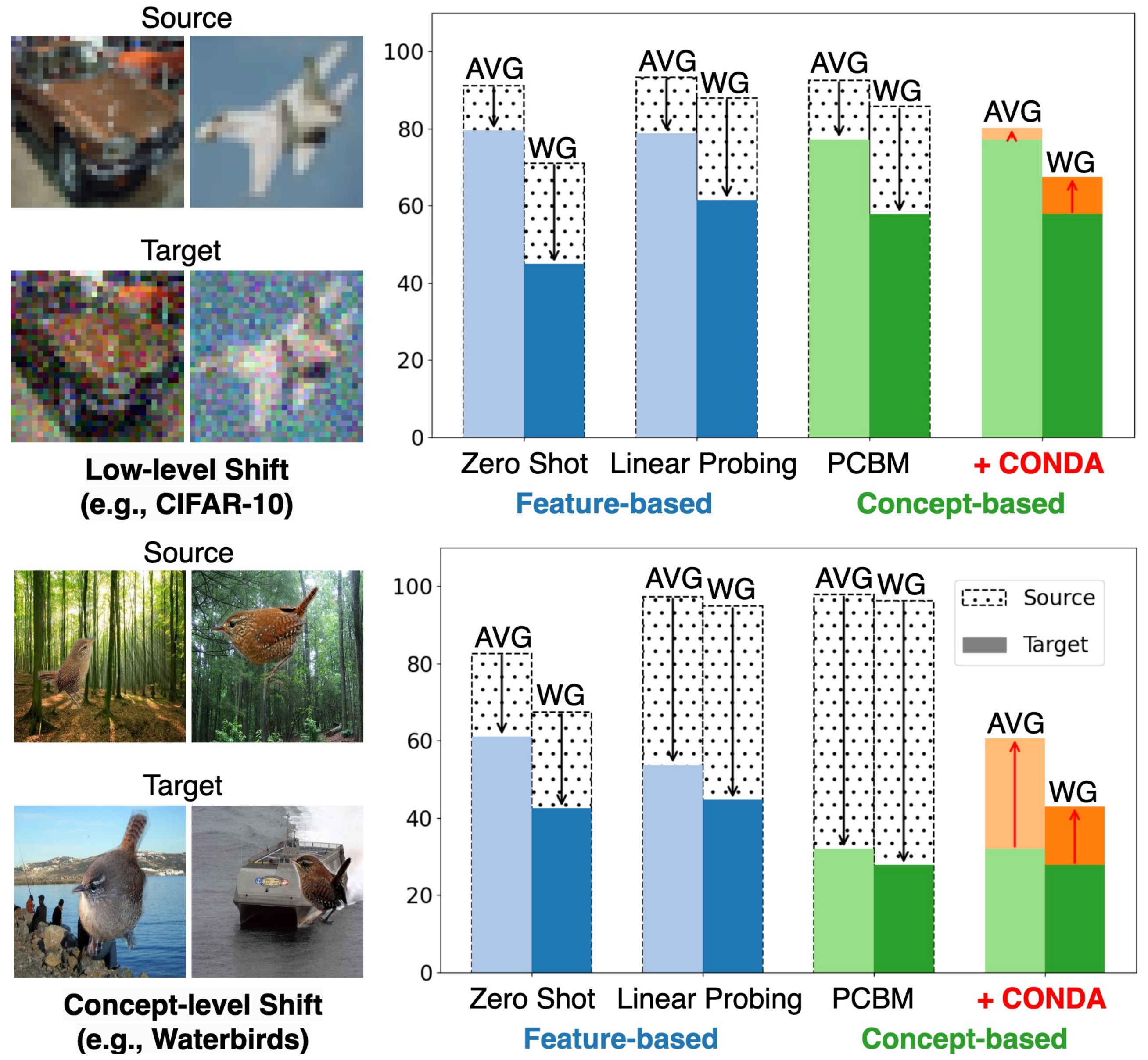


Various efforts to close the performance gap on in-distribution test set

How does it perform after deployment?

Motivation: When Deployed in the Wild

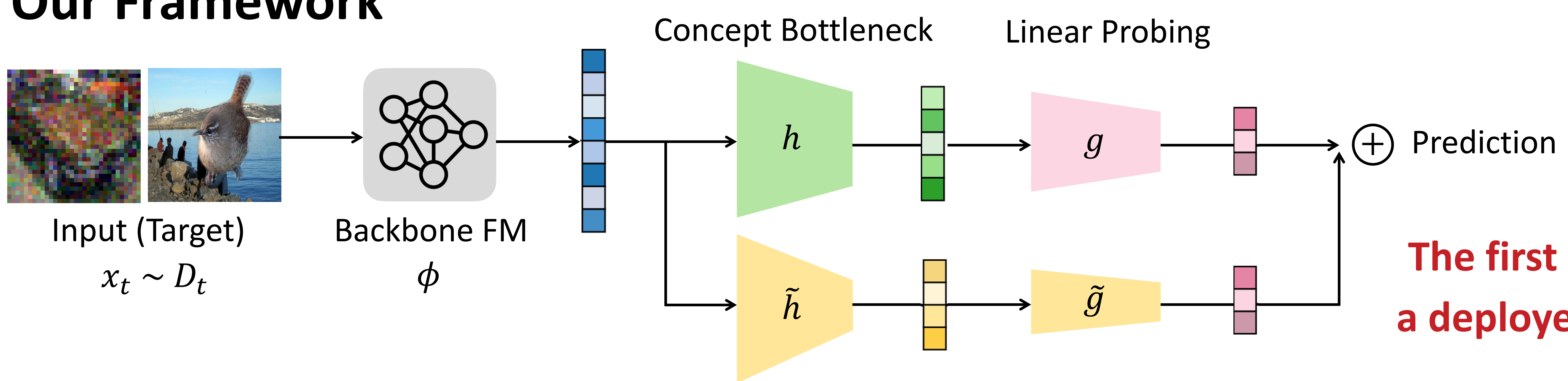
(AVG: average group acc, WG: worst group acc)



Low-level shift: Not necessarily more robust

Concept-level shift: Even more vulnerable

Our Framework



✓ No ground truth label at test time

✓ No access to source data

✓ Adapt on-the-fly with incoming batch

The first test-time adaptation framework for a deployed concept-based prediction pipeline

Failure Modes

1. Concept bottleneck is not robust

$$\mathbb{P}_{con}\,(D_t,\phi,h) \neq \mathbb{P}_{con}\,(D_s,\phi,h)$$

2. Concept reliance is not adapted

$$\mathbb{P}_{con}\,(D_t,\phi,h) \approx \mathbb{P}_{con}\,(D_s,\phi,h)$$

$$\mathbb{P}_{pred}\,(D_t,\phi,h,g) \neq \mathbb{P}_{pred}\,(D_s,\phi,h,g)$$

3. Concept set is not complete

There does not exist any g that satisfies

$$\mathbb{P}_{pred}\,(D_t,\phi,h,g) = \mathbb{P}_{pred}\,(D_s,\phi,h,g)$$

Corresponding Remedies

1. Concept Score Alignment (CSA)

Feature alignment of the concept scores of test inputs: their class-conditional distributions are close to that of the concept scores in the source dataset

2. Linear Probing Adaptation (LPA)

Label predictor is adapted, minimizing the cross-entropy loss with FM-based pseudo labels

3. Residual Concept Bottleneck (RCB)

Learning additional concept vectors and a linear predictor, minimizing test accuracy and overlap with existing concept vectors, while maximizing the concept coherency

Experiments

▨ Feature-based (Source)

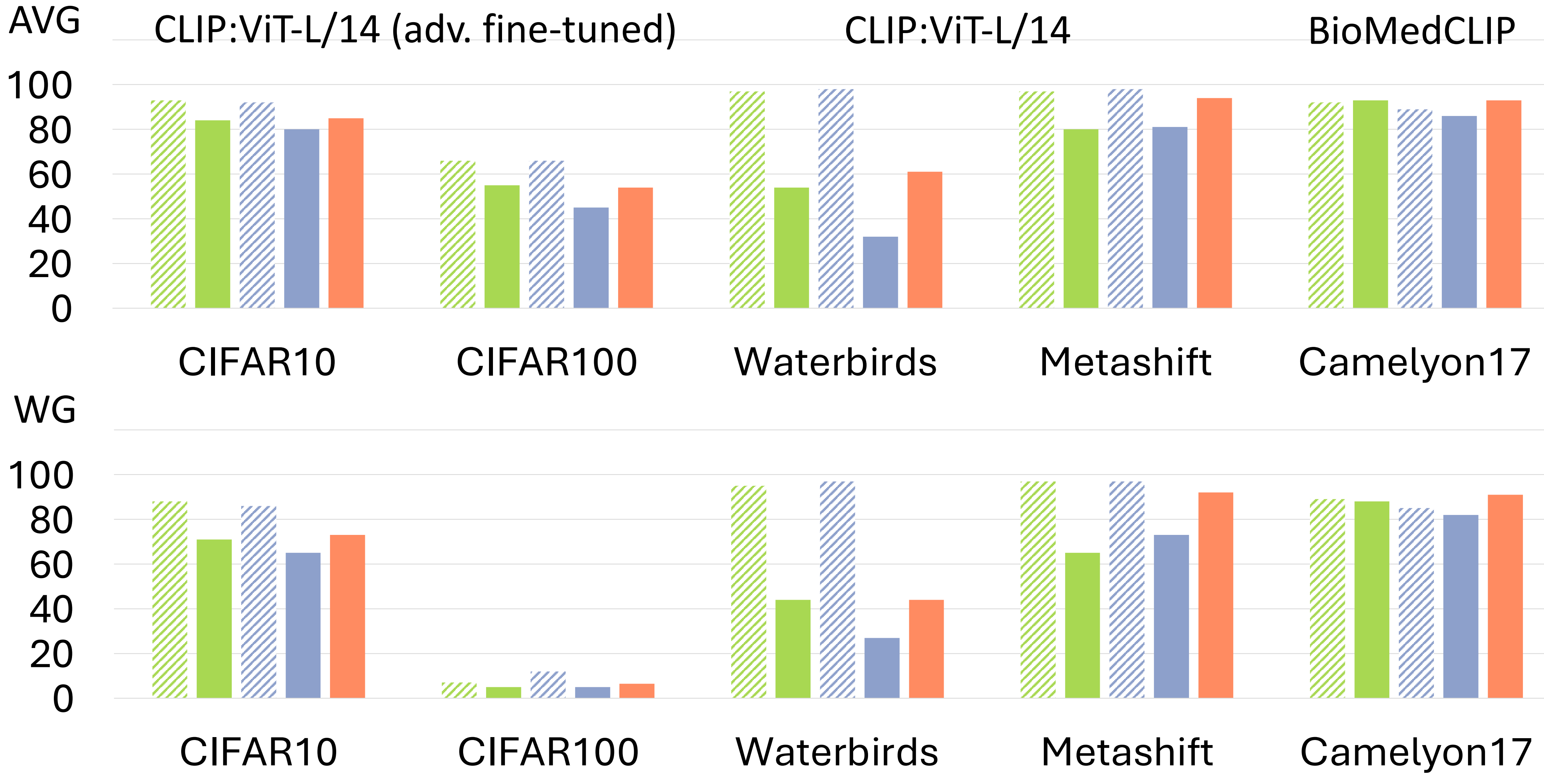
🟩 Feature-based (Target)

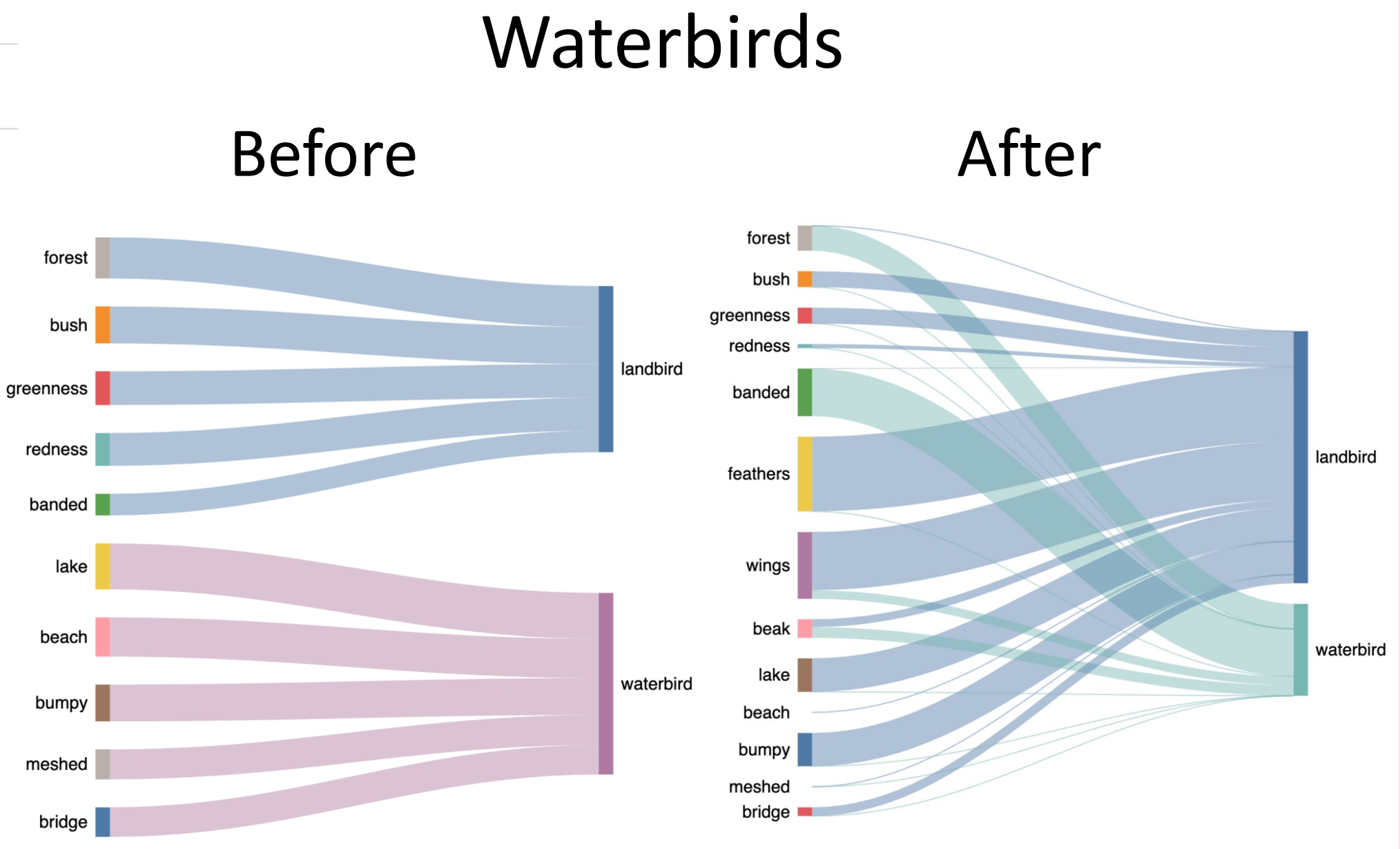
▨ Concept-based (Source)

■ Concept-based (Target)

🟧 (+adapted)

Boosted post-deployment performance + interpretability





More ablations in the paper:

Concept Bottleneck constructions

Backbone FMs

Individual contribution of each module

and so much more…