FaCT: Faithful Concept Traces for Explaining Neural Network Decisions

Amin Parchami-Araghi, Sukrut Rao, Jonas Fischer, Bernt Schiele

MAX PLANCK INSTITUTE FOR INFORMATICS

SIC Saarland Informatics Campus

Presented by Siddhartha Gairola

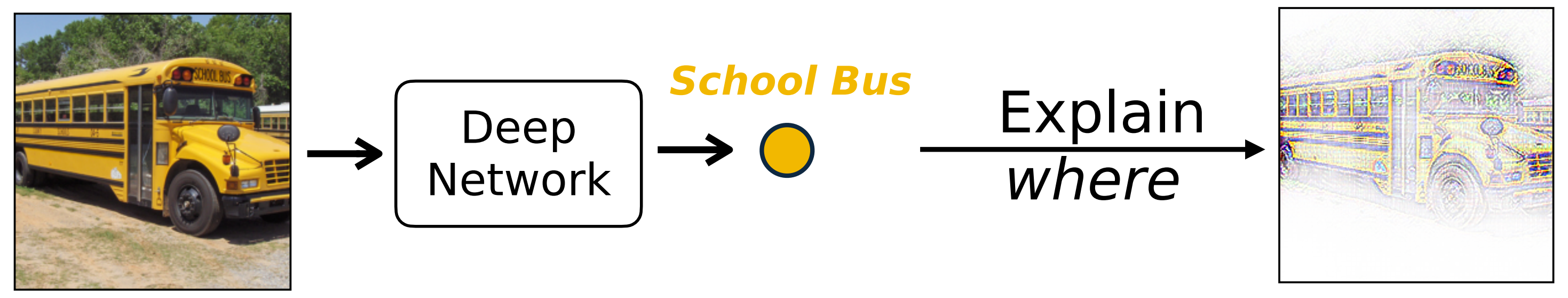
Code & Models

NEURAL INFORMATION PROCESSING SYSTEMS

Overview

Deep Networks are performant, yet opaque!

Attribution explanations → ***where*** the model looks



Deep Network

School Bus

Explain *where*

Concept explanations → ***what concepts*** the model uses.

In comparison to prior work:

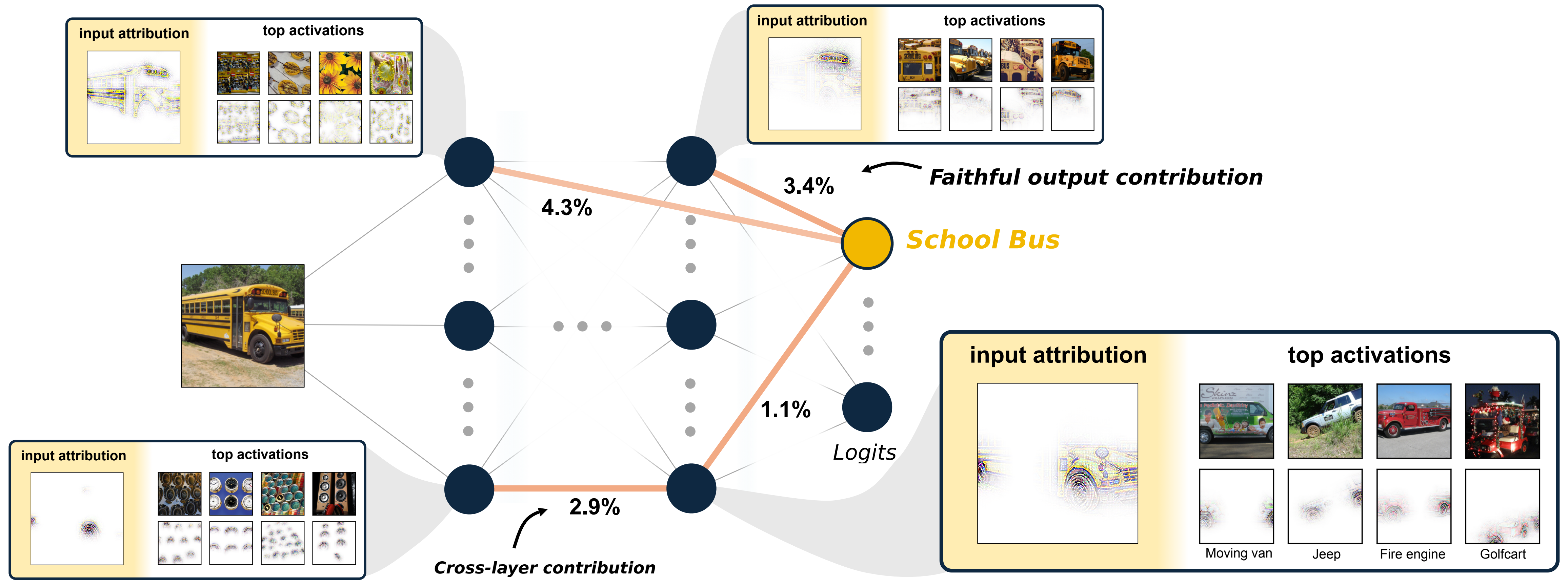
- ❌ Concepts are pre-defined, class-specific, or part-based.
- ✅ We have non-restricted diverse concepts across layers.
- ❌ Concepts are not part of the decision making.
- ✅ Our model *only uses concepts & reports their contributions.*
- ❌ Approximate concept visualizations are used.
- ✅ Our model provides faithful pixel-level concept visualization.
- ❌ Evaluation is limited to human annotations
- ✅ We propose a robust evaluation framework with DINOv2.

Faithful Concept Traces

- *Every logit is faithfully decomposed to concept contributions.*
- *Every concept is faithfully visualized at pixel-level.*

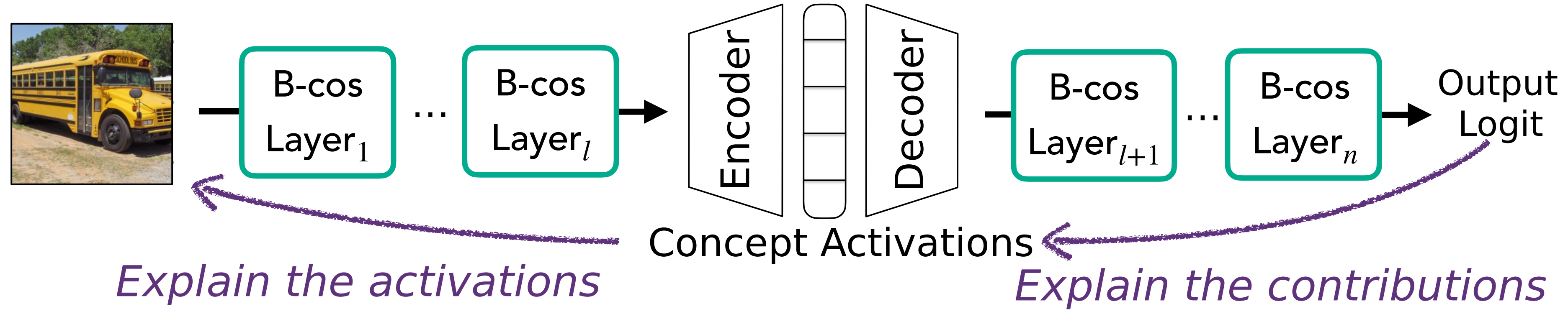
Concept Activation = $\sum$ Pixel Contribution

Output Logit = $\sum$ Concept Contribution


input attribution
top activations
input attribution
top activations
3.4%
Faithful output contribution
4.3%
School Bus
1.1%
Logits
2.9%
Cross-layer contribution
input attribution
top activations
input attribution
top activations
Moving van
Jeep
Fire engine
Golfcart

How we do it

Sparse Auto-Encoder



B-cos Layer$_1$ … B-cos Layer$_l$ → Encoder Decoder → B-cos Layer$_{l+1}$ … B-cos Layer$_n$ → Output Logit

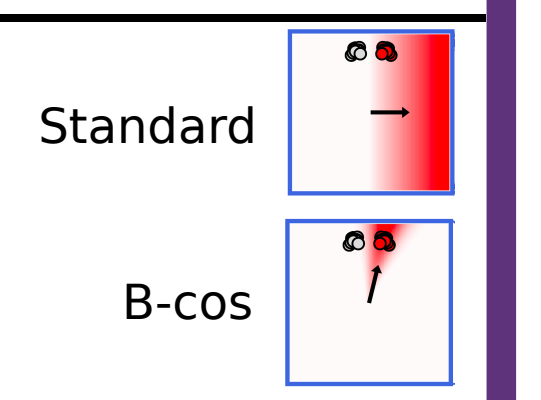
Concept Activations

Explain the activations

Explain the contributions

Standard non-linear layer $f(\mathbf{x}) := \text{ReLU}(\mathbf{W}\mathbf{x}+\mathbf{b})$

B-cos non-linear layer $f(\mathbf{x}) := \mathbf{W}\mathbf{x}\,|\cos(\mathbf{W},\mathbf{x})|^{\mathbf{B}} \underset{Interpret}{=} \tilde{\mathbf{W}}(\mathbf{x})\mathbf{x}$


Standard
B-cos

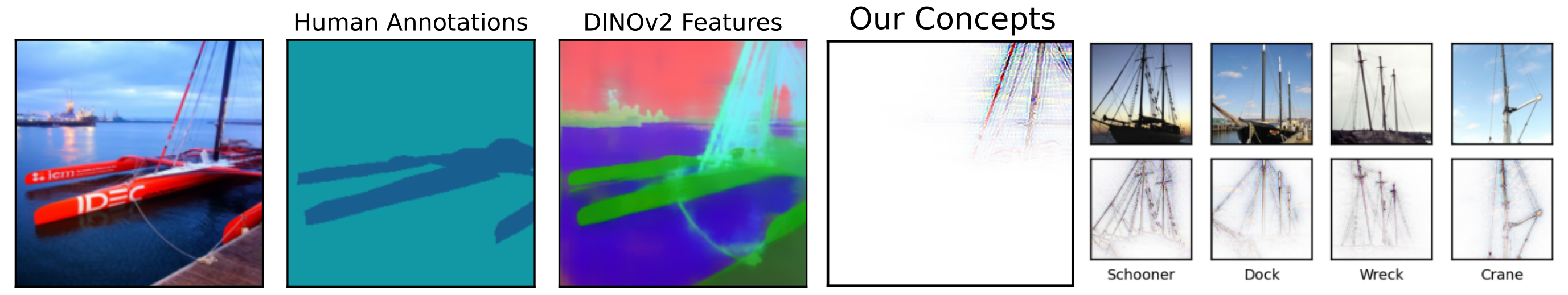
input — B-cos Layer$_1$ … B-cos Layer$_n$ → **output** $\underset{Interpret}{=} \tilde{\mathbf{W}}_{1\to n}(\textbf{input}) \times \textbf{input}$

$\tilde{\mathbf{W}}_1(\mathbf{x}_1)$ $\quad$ $\tilde{\mathbf{W}}_n(\mathbf{x}_n)$

(B-cos; Böhle et al, CVPR 2022)

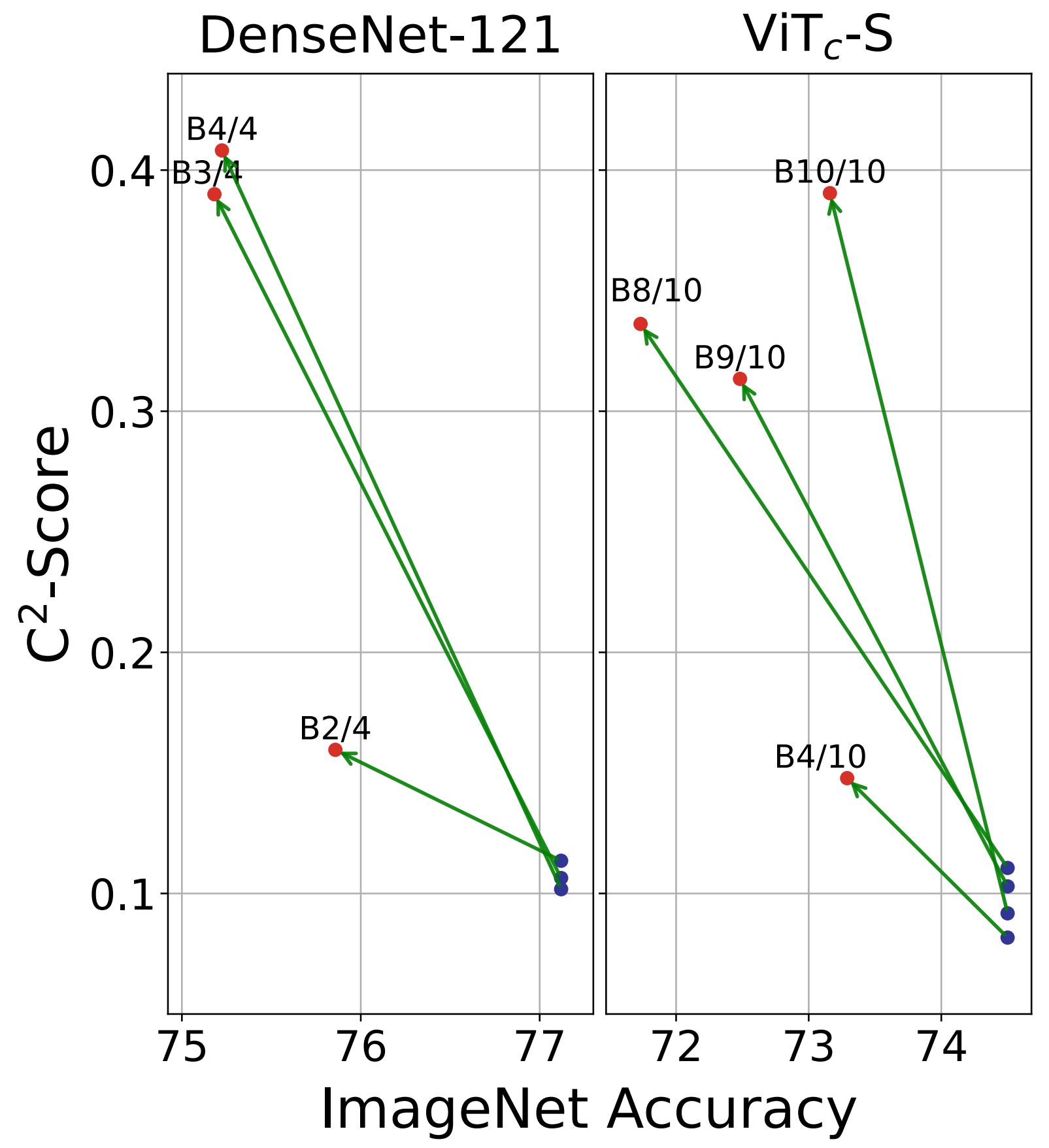
Measuring Consistency

Annotations are not suitable for evaluating concepts


Human Annotations
DINOv2 Features
Our Concepts
Schooner
Dock
Wreck
Crane

We propose using DINOv2 Features

Consistent & Competitive on ImageNet

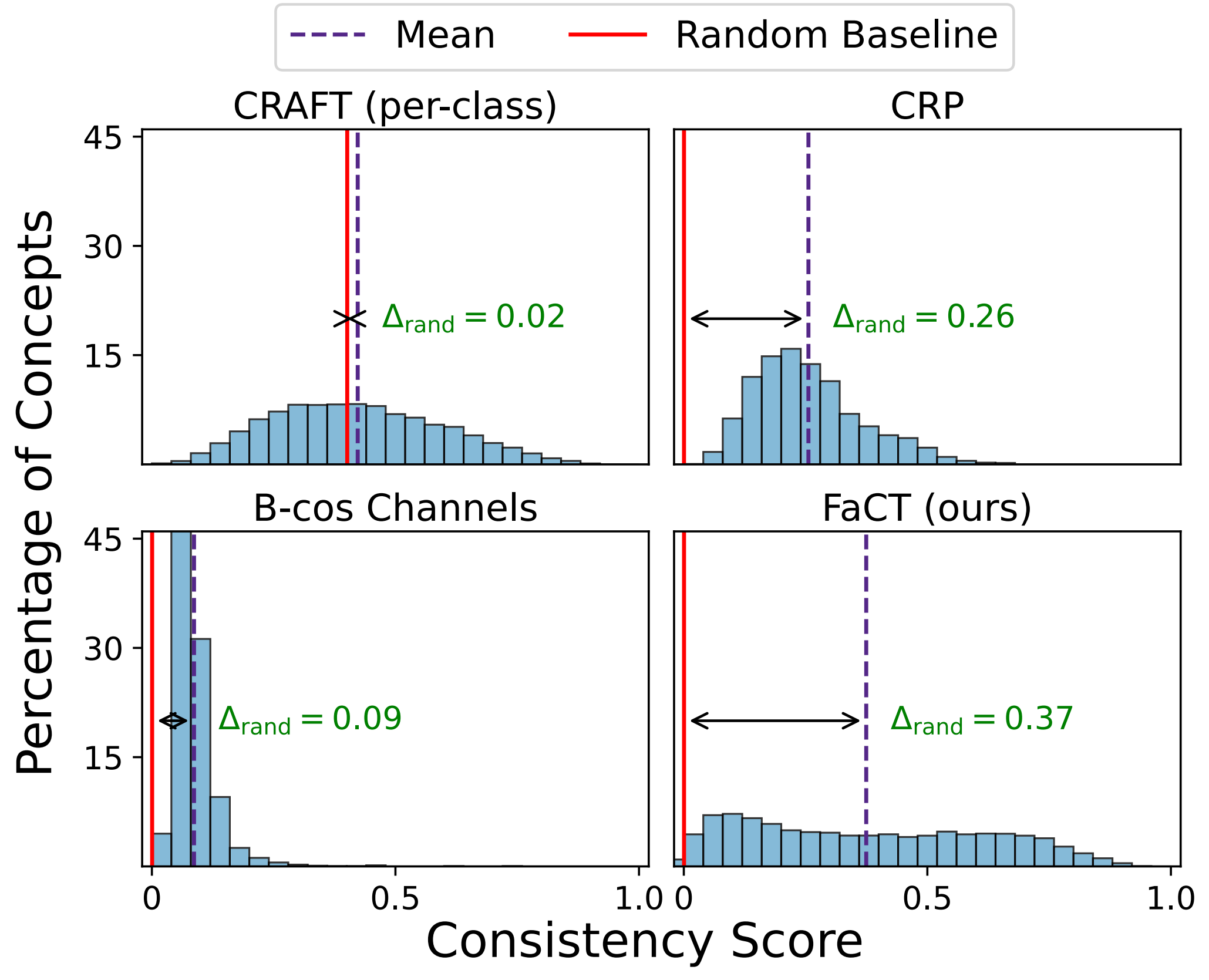


Unicycle
Maillot
Mountain bike
Bicycle-built-f
Bookshop
Library
Bookshop
Rocking chair

Competitive & far more consistent on ImageNet

Generalizing across layers and architectures

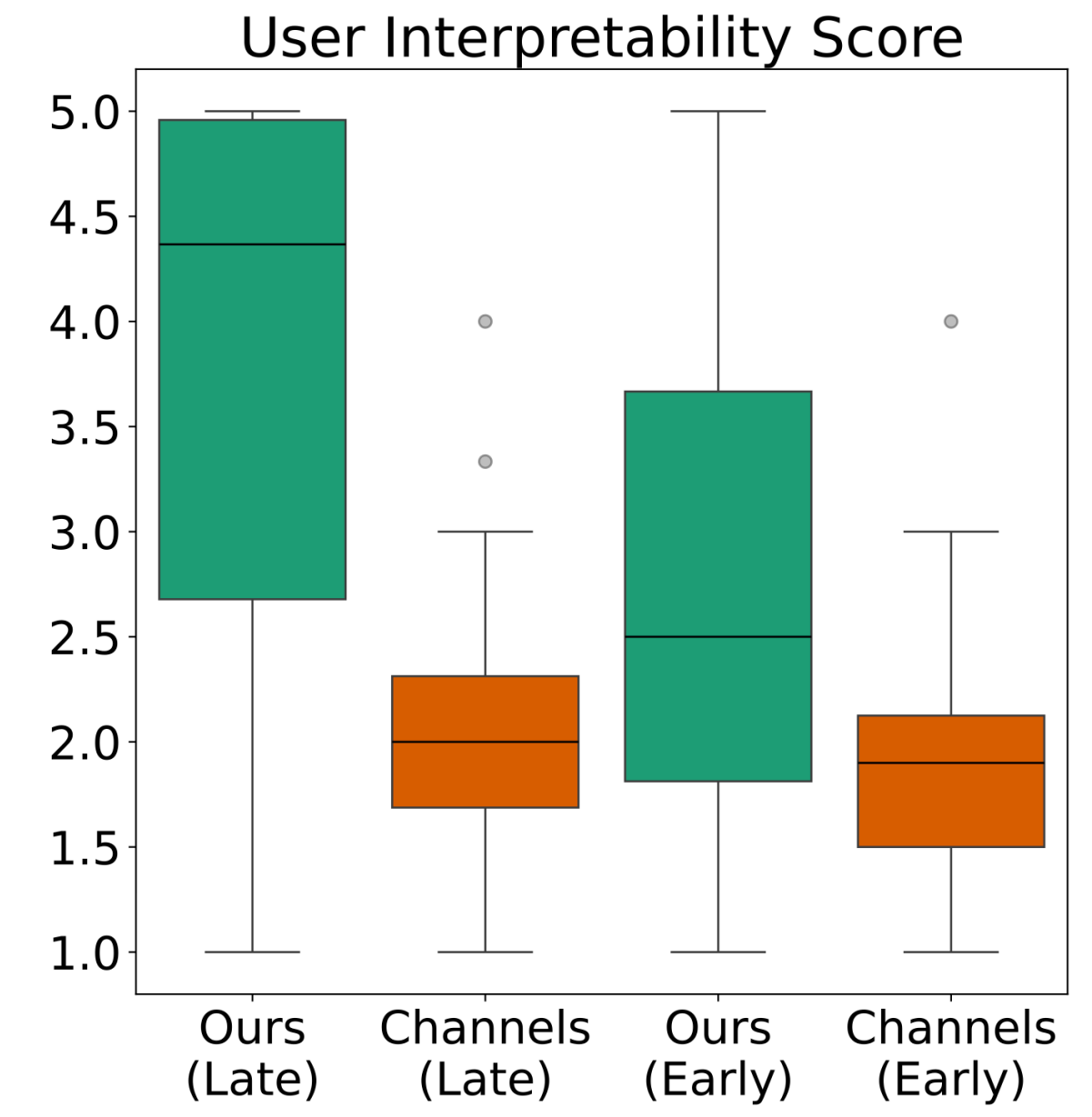
More Consistent Concepts



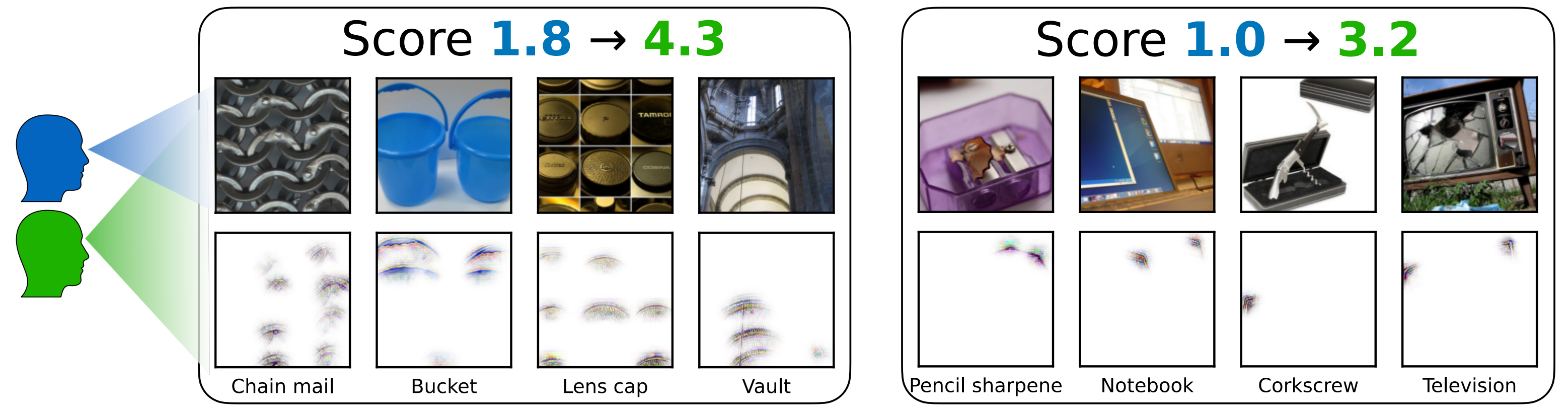
Our concepts are more consistent compared to other concept-based explanation methods.

(CRAFT; Fel et al, CVPR 2023) (CRP; Achtibat et al, Nature M. Intel. 2023)

More Interpretable to Users

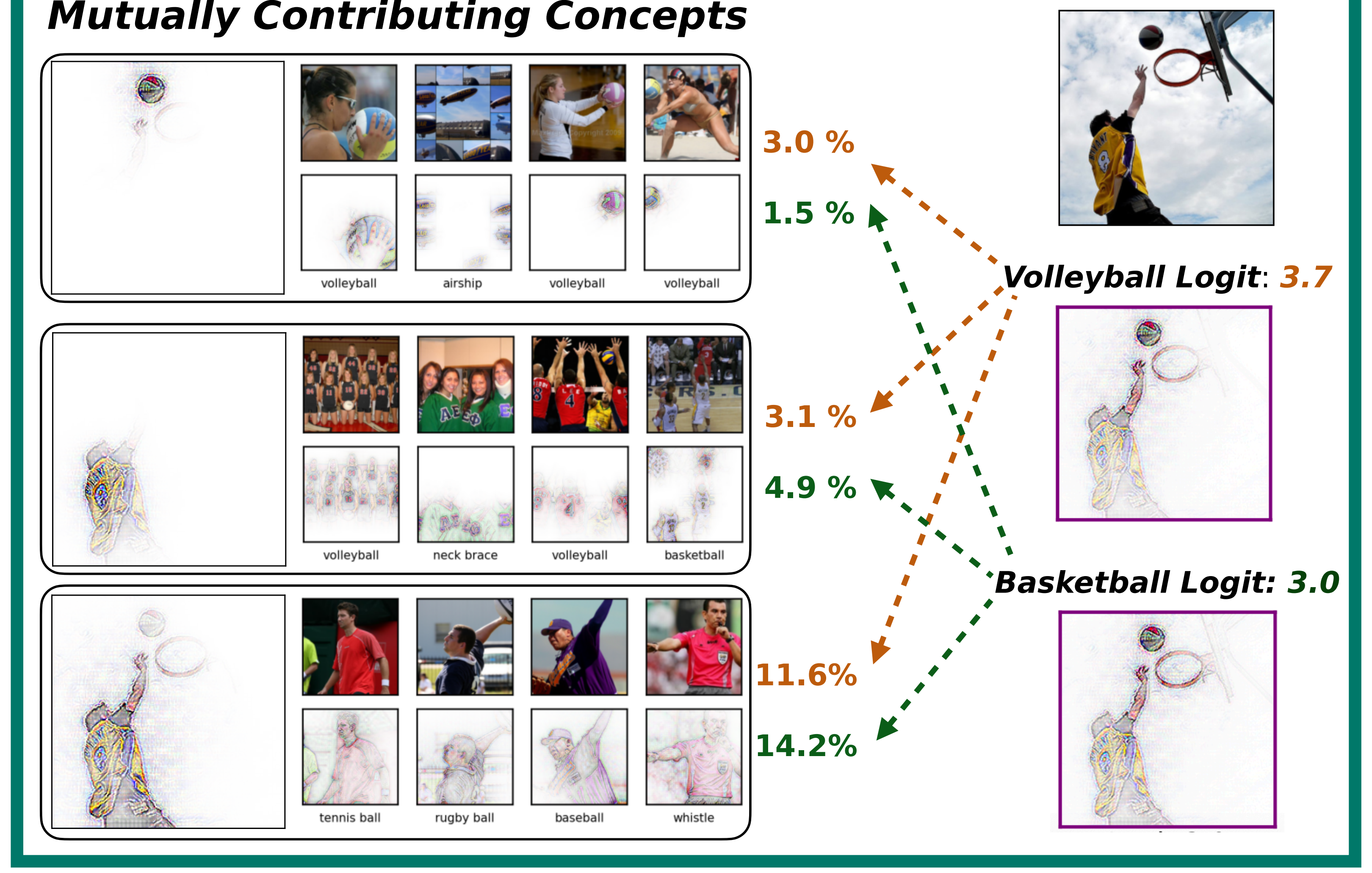


- Users found our concepts more Interpretable than B-cos channels.
- The pixel-level visualizations help the interpretability!


Score 1.8 → 4.3
Chain mail
Bucket
Lens cap
Vault
Score 1.0 → 3.2
Pencil sharpene
Notebook
Corkscrew
Television

Concept-level Analysis of Misclassification

Mutually Contributing Concepts


volleyball
airship
volleyball
volleyball
volleyball
neck brace
volleyball
basketball
tennis ball
rugby ball
baseball
whistle
3.0 %
1.5 %
3.1 %
4.9 %
11.6%
14.2%
Volleyball Logit: 3.7
Basketball Logit: 3.0