MAX PLANCK INSTITUTE FOR INFORMATICS

SIC Saarland Informatics Campus

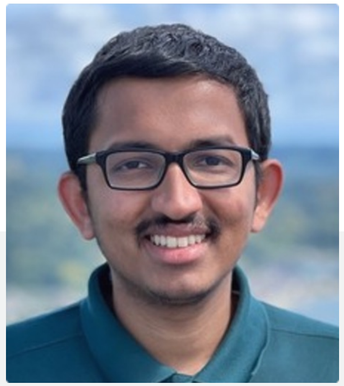
NEURAL INFORMATION PROCESSING SYSTEMS

FaCT: Faithful Concept Traces for Explaining Neural Network Decisions

Accepted to NeurIPS 2025
Presented at Theory of Explainable Machine Learning Workshop @ ELLIS UnConference



Amin Parchami-Araghi



Sukrut Rao



Jonas Fischer‡

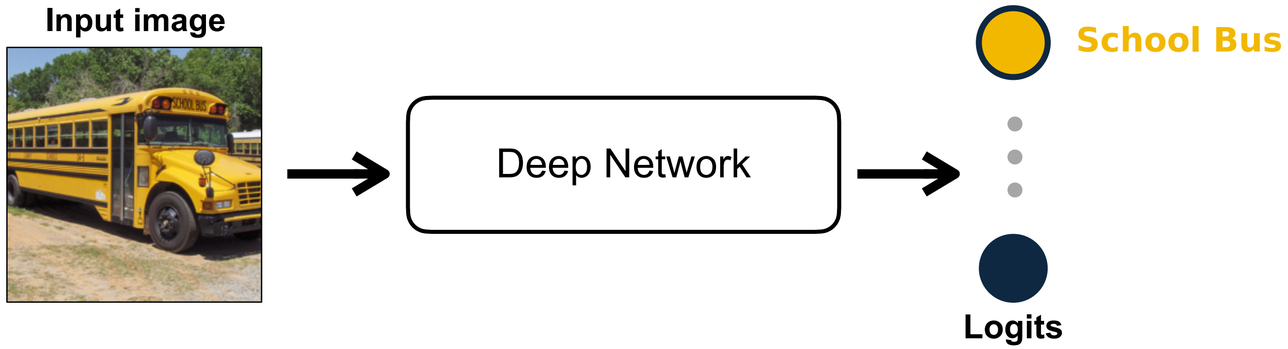


Bernt Schiele‡

‡ equal contribution as advisors

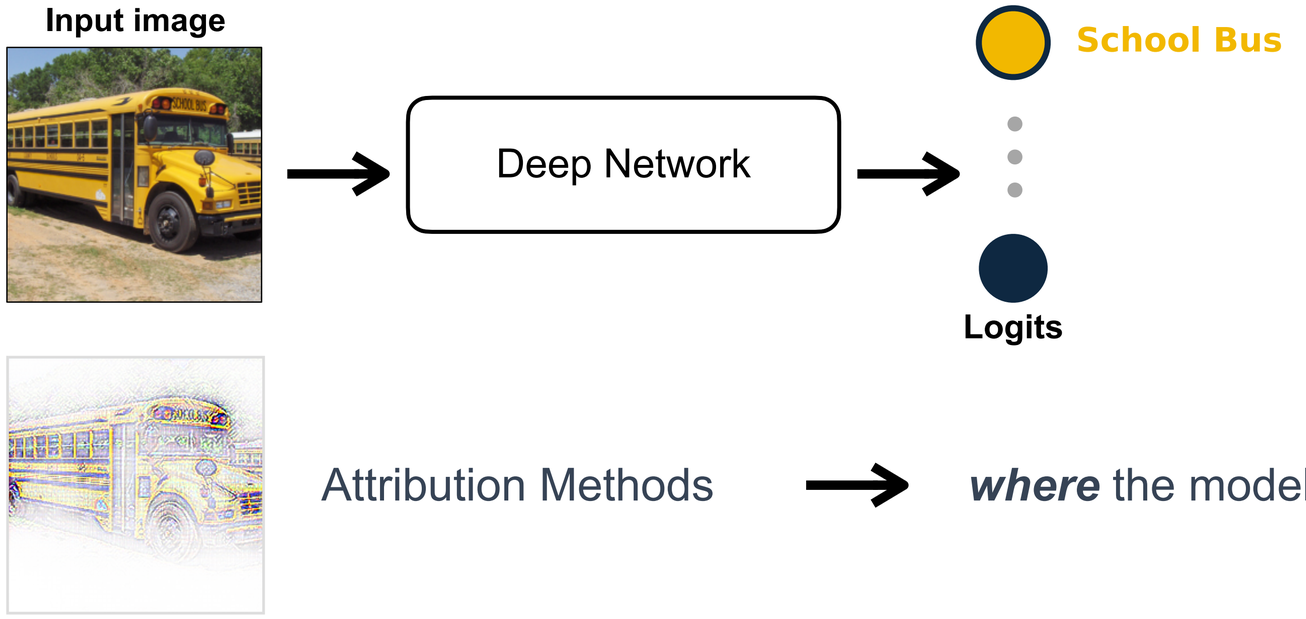
Background: Explanation Methods

How did the model arrive at this decision?



Background: Attribution Methods

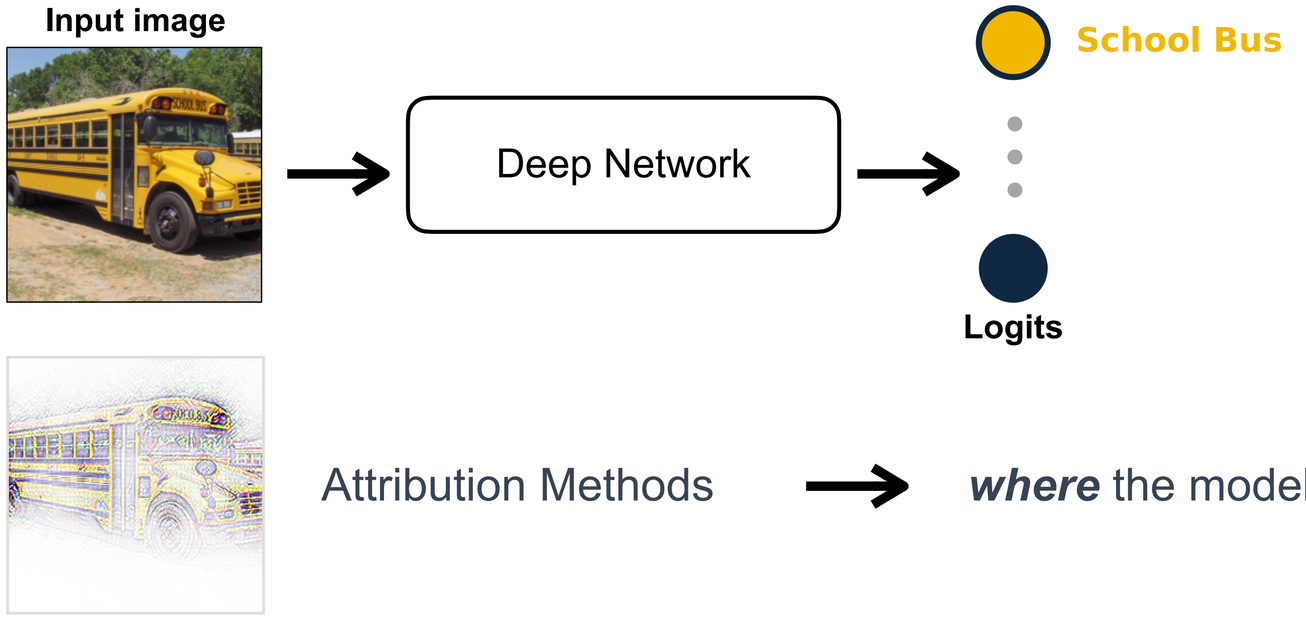
How did the model arrive at this decision?



Attribution method from B-cos Networks (Böhle et al., CVPR 2022, TPAMI 2024)

Background: Attribution Methods

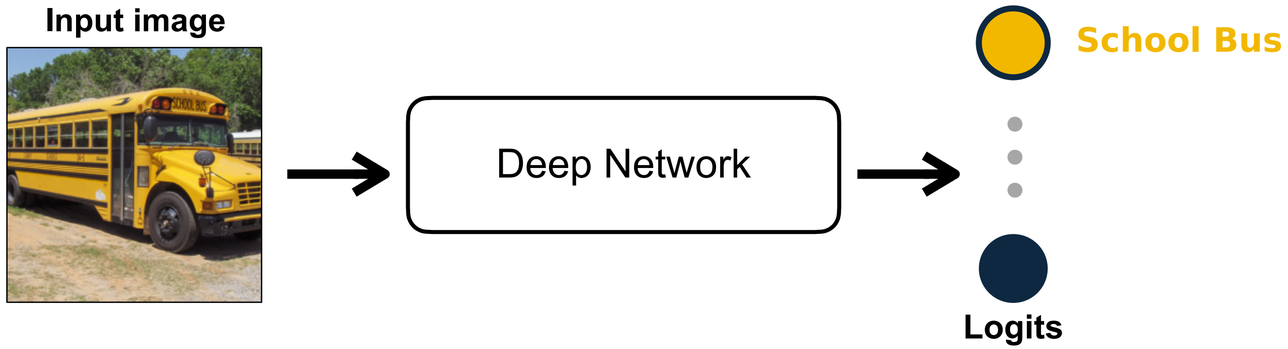
How did the model arrive at this decision?



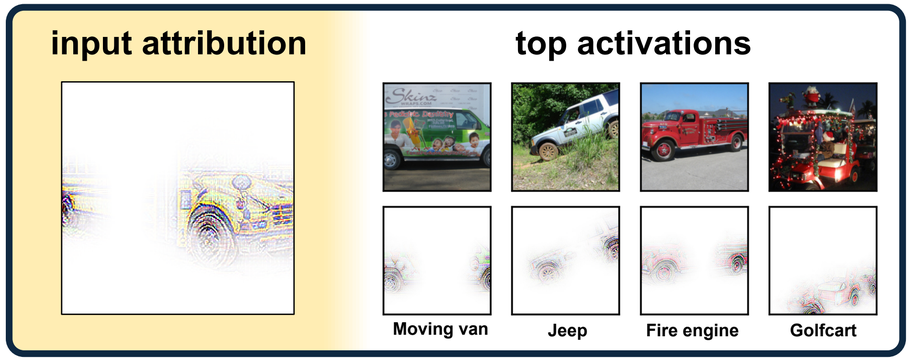
Question: **what semantic concepts** that the model is using?

Attribution method from B-cos Networks (Böhle et al., CVPR 2022, TPAMI 2024)

Explanation Methods: Concept-based Explanations

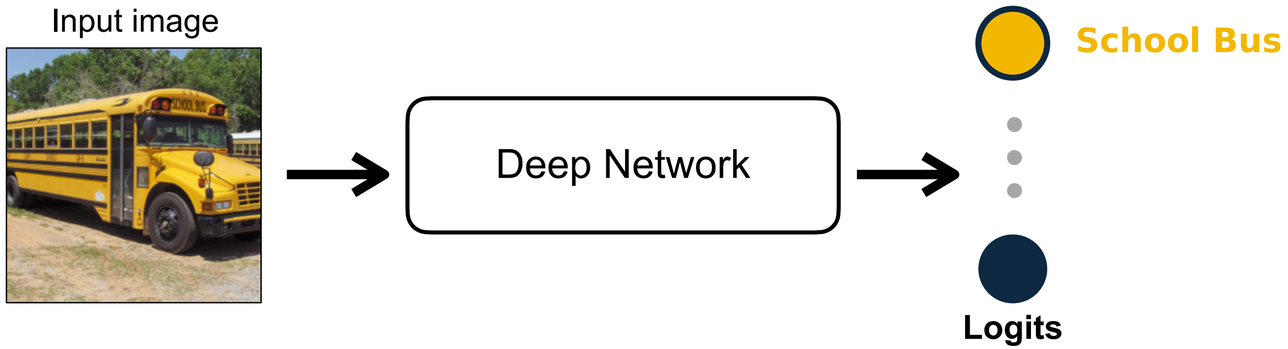


The **[wheel]** concept



How important is **[wheel]** for ***School Bus***?

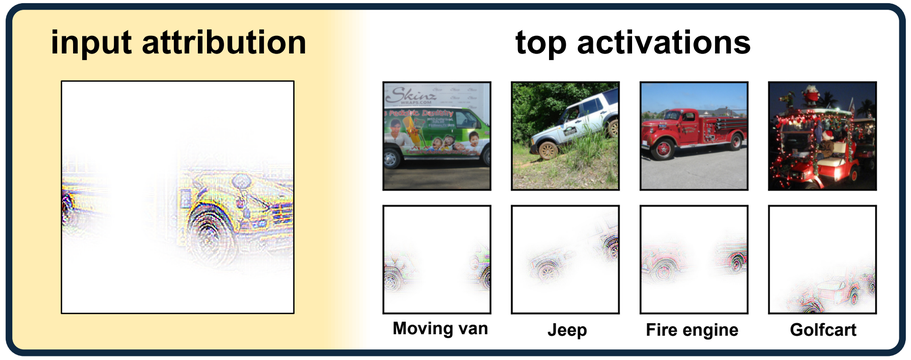
Explanation Methods: Concept-based Explanations



Concept Definition

The **[wheel]** concept

Concept Visualization



Concept Importance

How important is **[wheel]** for ***School Bus***?

Explanation Methods: Concept-based Explanations

Concept Definition

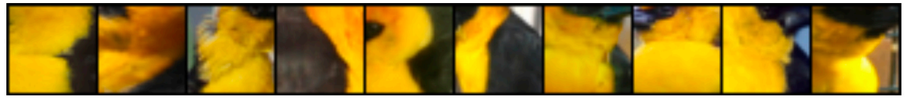
Restrictive →

- class-specific concepts [1,2,3,7]
- part-based or predefined [3,4,5]
- fixed spatial size or crops [1,4]

Class A, Concept 1



Class B, Concept 1



Concept Visualization

Concept Importance

[1]: CRAFT; Fel et al. [2]: ACE; Ghorbani et al. [3] ProtoPNet; Chen et al. [4] Pip-Net; Nauta et al.
[5] CBM; Koh et al. [6] Benchmark for Prototypical Parts, Sacha et al. [7] VCC, Kowal et al.; Figure partially from [4]

Explanation Methods: Concept-based Explanations

Concept Definition

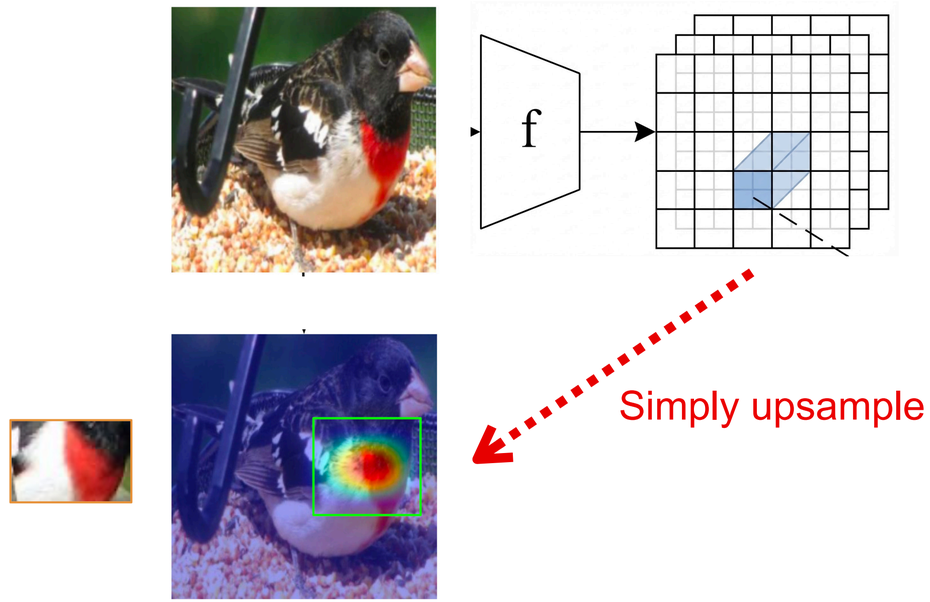
Restrictive →

- class-specific concepts [1,2,3,7]
- part-based or predefined [3,4,5]
- fixed spatial size or crops [1,4]

Concept Visualization

Approximate →

- up-sampled heatmaps (worse for ViTs) [1,3,6]
- top-activating crops [1,2,3,4]



Concept Importance

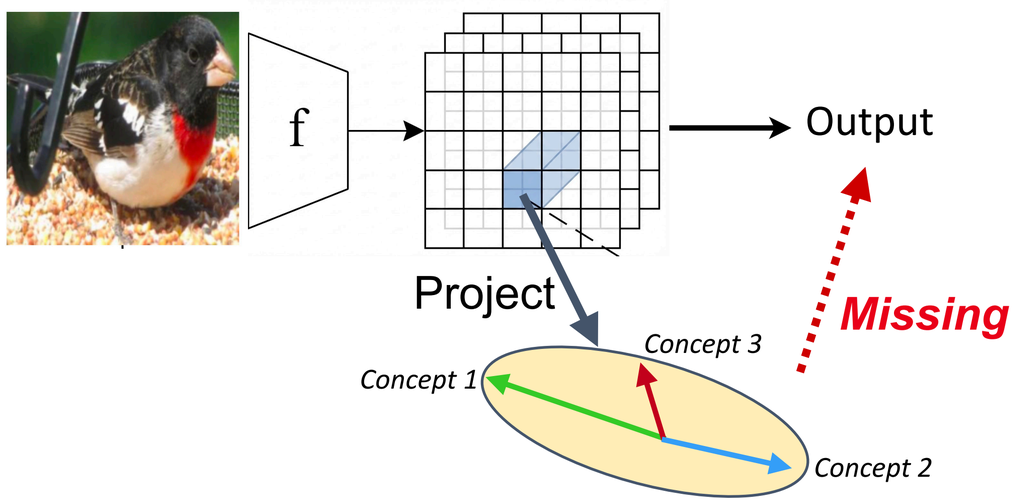
[1]: CRAFT; Fel et al. [2]: ACE; Ghorbani et al. [3] ProtoPNet; Chen et al. [4] Pip-Net; Nauta et al.
[5] CBM; Koh et al. [6] Benchmark for Prototypical Parts, Sacha et al. [7] VCC, Kowal et al.; Figure partially from [6]

Explanation Methods: Concept-based Explanations

Concept Definition —*Restrictive*→
- class-specific concepts [1,2,3,7]
- part-based or predefined [3,4,5]
- fixed spatial size or crops [1,4]

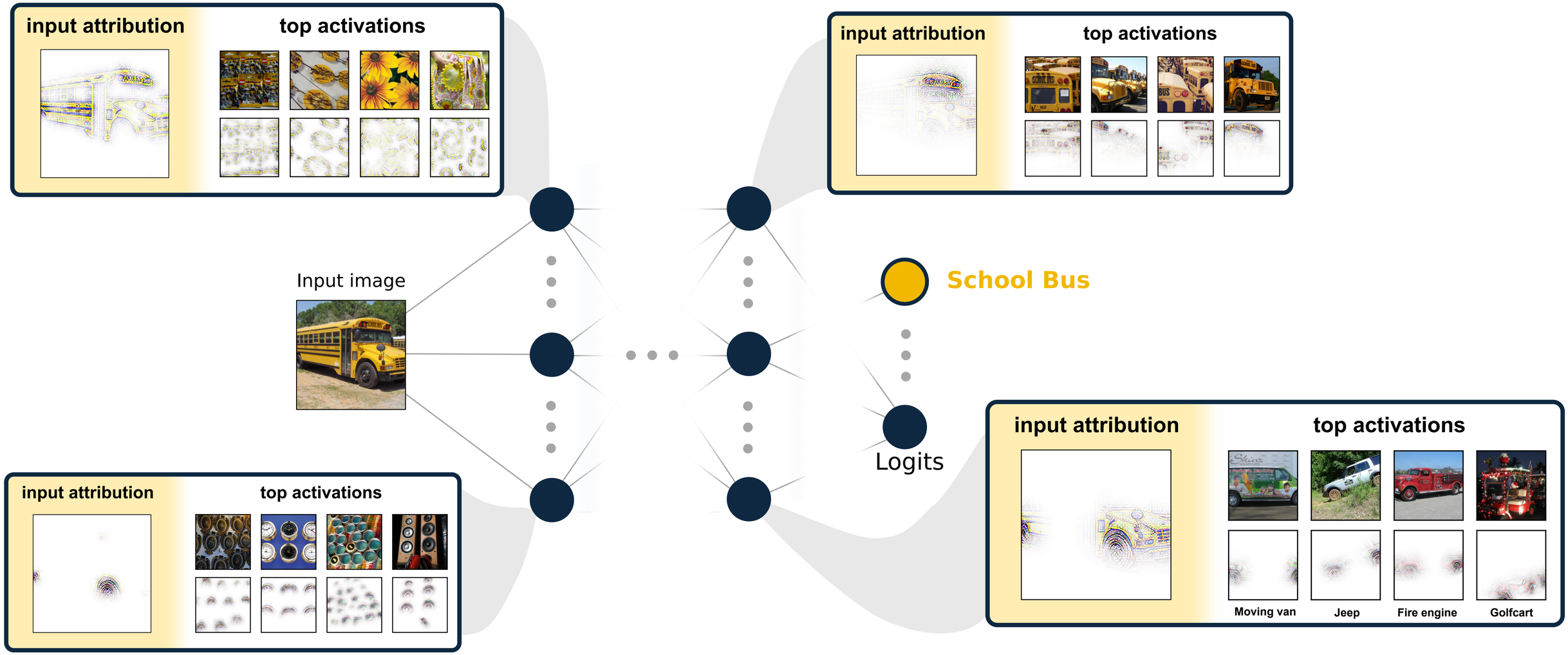
Concept Visualization —*Approximate*→
- up-sampled heatmaps (worse for ViTs) [1,3,6]
- top-activating crops [1,2,3,4]

Concept Importance —*Post-hoc*→
- concepts not part of the model [1,2,7]
- rely on approximate importance measures [1,2,7]

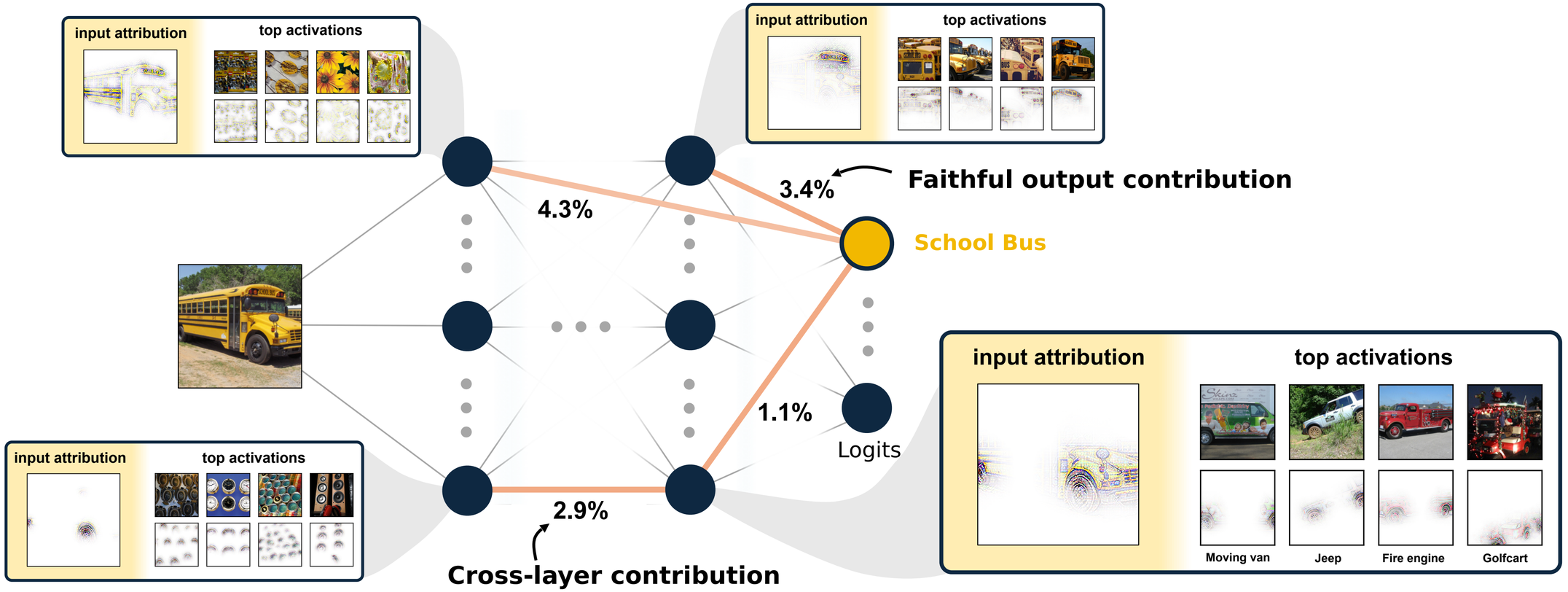


[1]: CRAFT; Fel et al. [2]: ACE; Ghorbani et al. [3] ProtoPNet; Chen et al. [4] Pip-Net; Nauta et al.
[5] CBM; Koh et al. [6] Benchmark for Prototypical Parts, Sacha et al. [7] VCC, Kowal et al.

FaCT: Shared Concepts Across the Layers



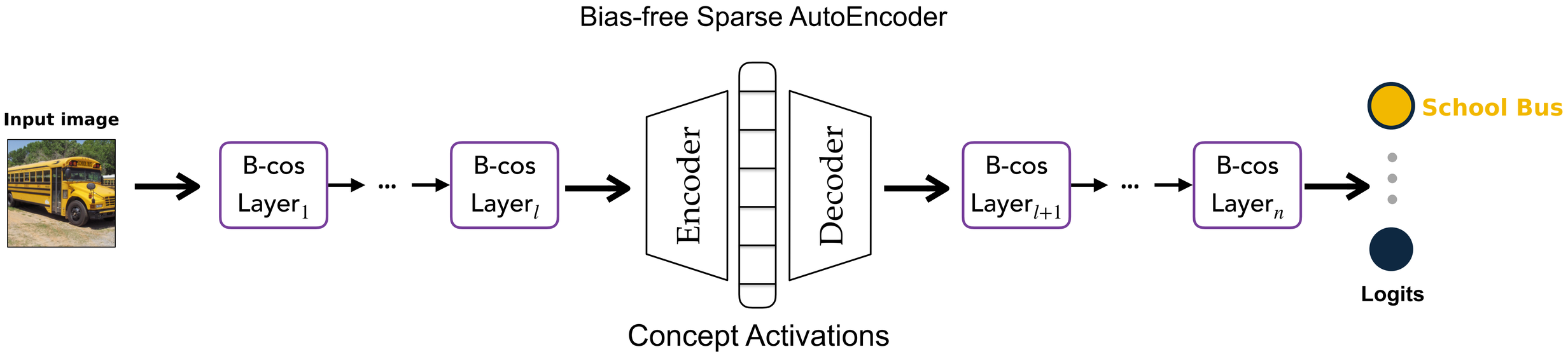
Our Work: Faithful Tracing



$$\text{Output Logit} = \sum \text{Concept Contribution}$$

$$\text{Concept Activation} = \sum \text{Pixel Contribution}$$

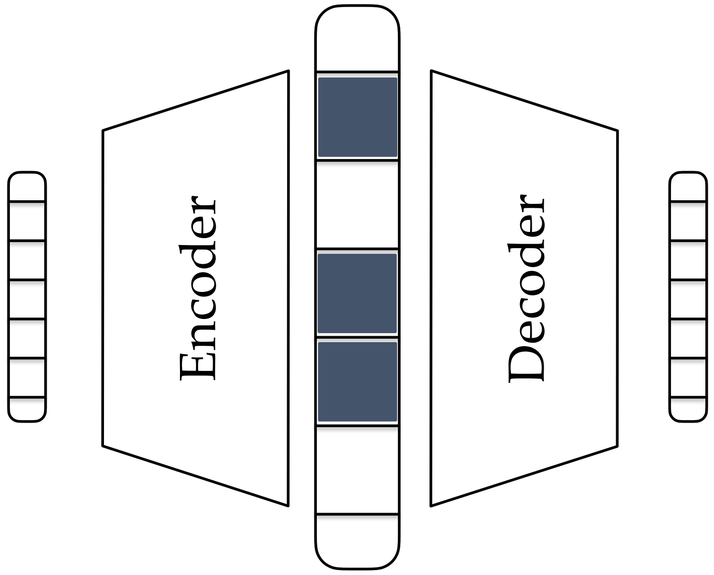
FaCT Architecture



B-cos: (Böhle et al. CVPR 2022, TPAMI 2024)

(Top-K) Sparse Autoencoder (Makhzani et al. ICLR 2014, Cunningham et al. ICLR 2024, Gao et al. 2024)

Detour: Sparse AutoEncoder (SAE)



SAE Architecture

$$\text{Encoder} = \text{ReLU}\left(\mathbf{W}^{\text{Encoder}}\mathbf{x} + \mathbf{b}^{\text{Encoder}}\right)$$

$$\text{Decoder} = \mathbf{W}^{\text{Decoder}}\mathbf{x} + \mathbf{b}^{\text{Decoder}}$$

Sparsity regularization e.g. (ℓ_1 loss, K-sparsity)

We use a bias-free TopK-SAE

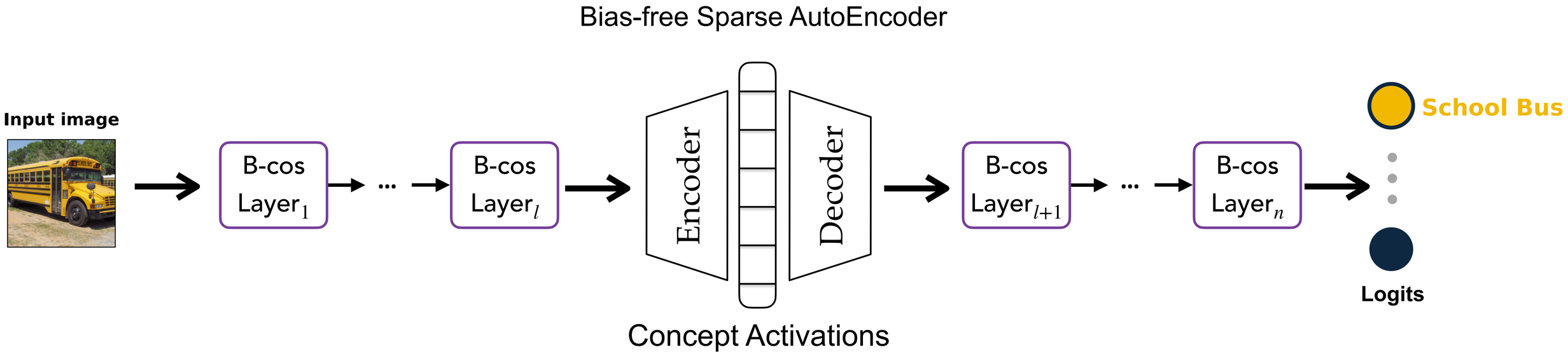
$$\text{Encoder} = \text{ReLU}\left(\mathbf{W}^{\text{Encoder}}\mathbf{x}\right)$$

$$\text{Decoder} = \mathbf{W}^{\text{Decoder}}\mathbf{x}$$

Sparsity regularization: K-sparsity

(Top-K) Sparse Autoencoder (Makhzani et al. ICLR 2014, Cunningham et al. ICLR 2024, Gao et al. 2024)

FaCT Architecture



B-cos: (Böhle et al. CVPR 2022, TPAMI 2024)

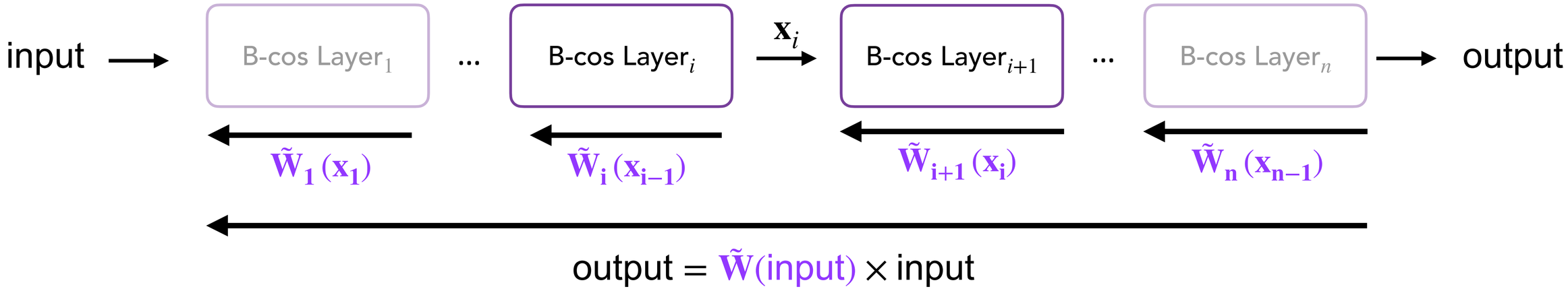
(Top-K) Sparse Autoencoder (Makhzani et al. ICLR 2014, Cunningham et al. ICLR 2024, Gao et al. 2024)

Detour: B-cos Transforms

Standard Layer $\text{ReLU}(\mathbf{W}\mathbf{x} + \mathbf{b})$

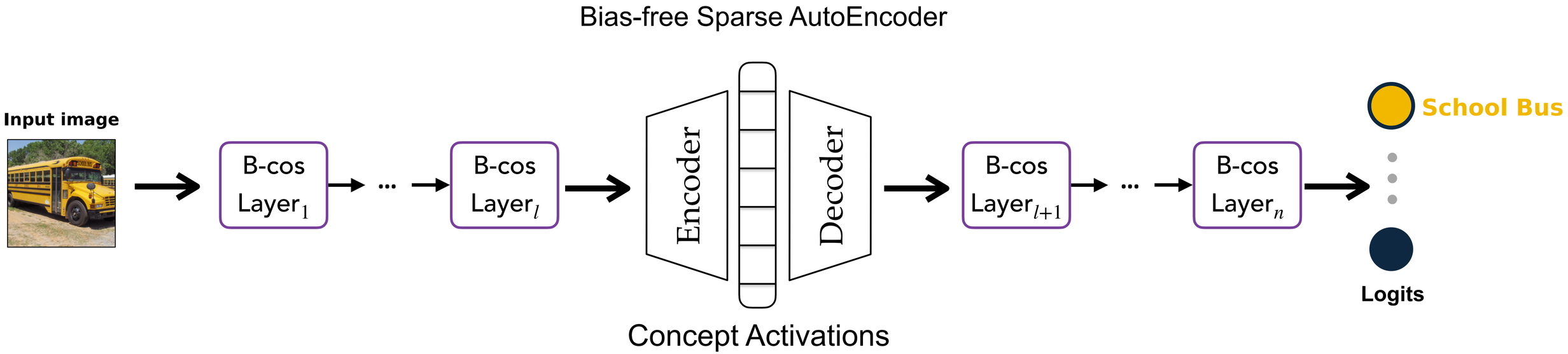
B-cos Layer $\mathbf{W}\mathbf{x}\,|\cos(\mathbf{W}, \mathbf{x})|^{\mathbf{B}}$ $\underset{\textit{Interpret}}{=}$ $\tilde{\mathbf{W}}(\mathbf{x})\mathbf{x}$

Non-linear transform, can be interpreted as a dynamic-linear transform



B-cos: Böhle et al. (CVPR 2022, TPAMI 2024)

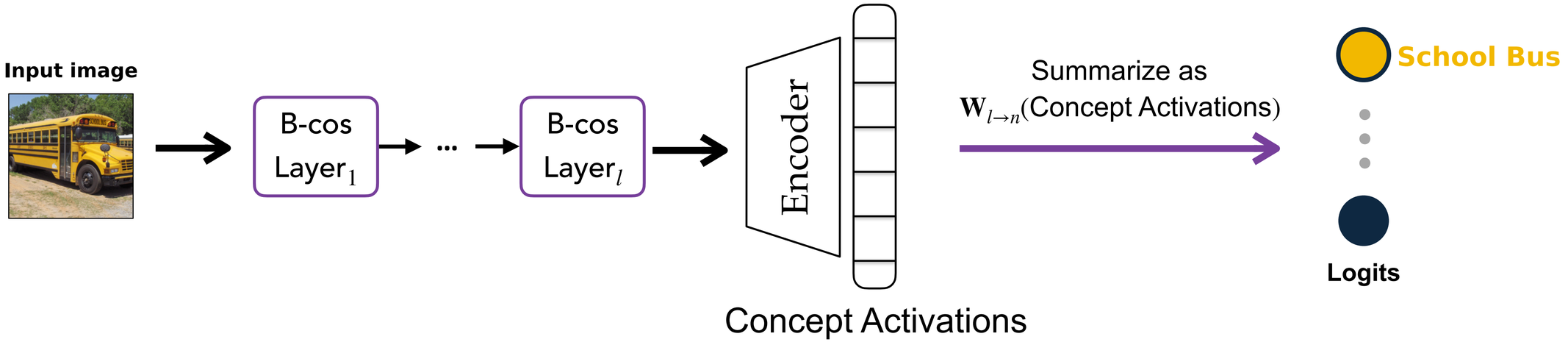
FaCT Architecture



B-cos (Böhle et al. CVPR 2022, TPAMI 2024)

(Top-K) Sparse Autoencoder (Makhzani et al. ICLR 2014, Cunningham et al. ICLR 2024, Gao et al. 2024)

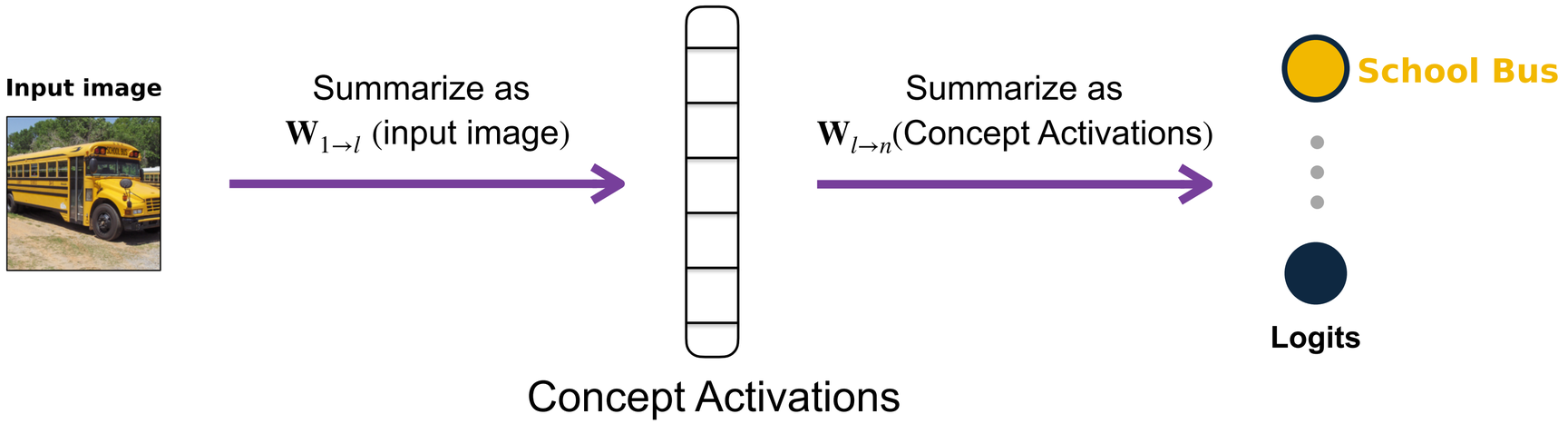
FaCT: Competitive and Diverse



$$\text{Output Logit} = \sum W_{l\to n}(\text{activations}) \times \text{activations}$$

$$= \sum \text{Concept Contribution}$$

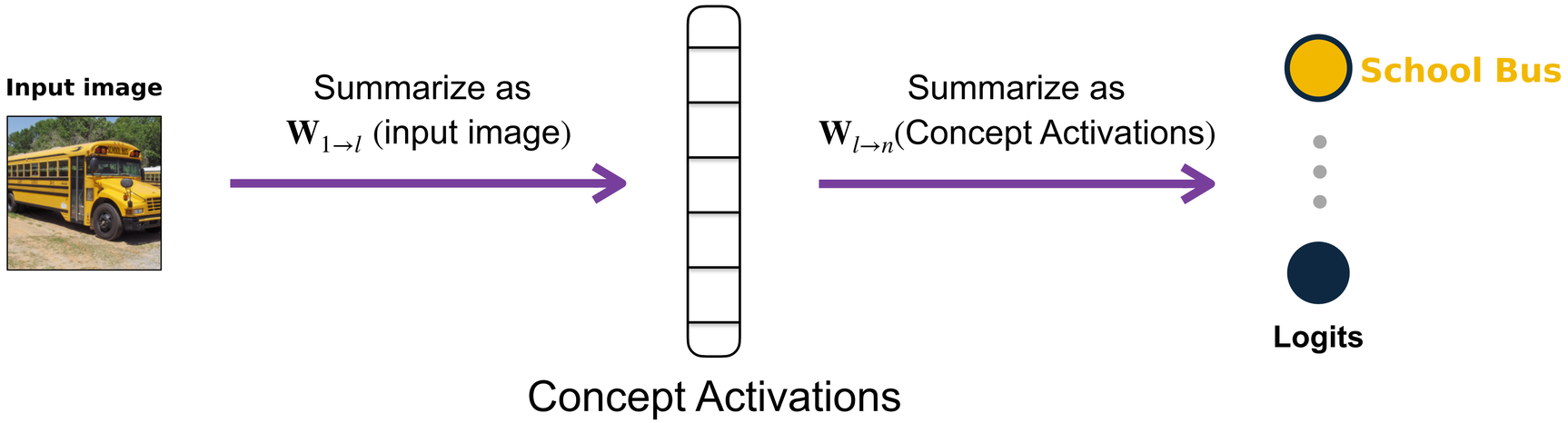
FaCT: Competitive and Diverse



$$\text{Concept Activation} = \sum W_{l\to n}(\text{pixels}) \times \text{pixels}$$
$$= \sum \text{Pixel Contribution}$$

$$\text{Output Logit} = \sum W_{l\to n}(\text{activations}) \times \text{activations}$$
$$= \sum \text{Concept Contribution}$$

FaCT: Competitive and Diverse



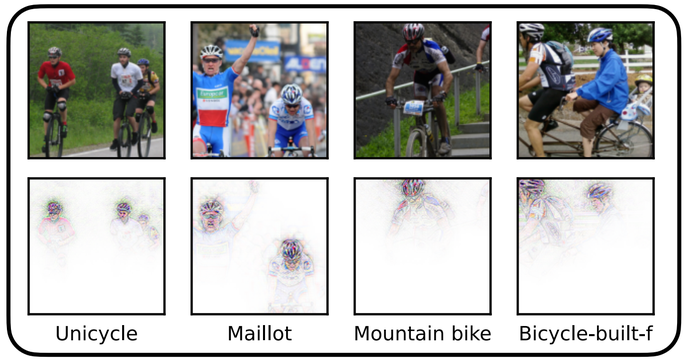
- ✓ The model ***only*** uses the concepts
- ✓ Concepts are shared across classes
- ✓ No (spatial) constraints on concepts
- ✓ Contribution of concepts to the output can be faithfully measured
- ✓ Concepts can be faithfully visualized at input level

FaCT Yields Diverse Concepts

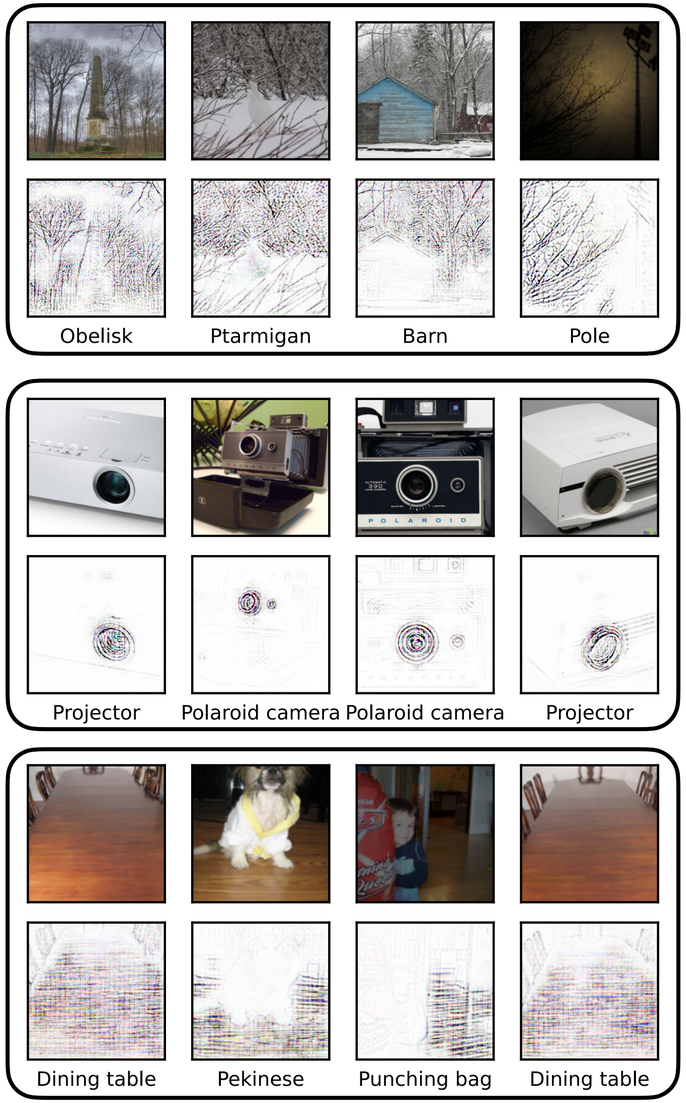
Late Layer

Early Layer

DenseNet

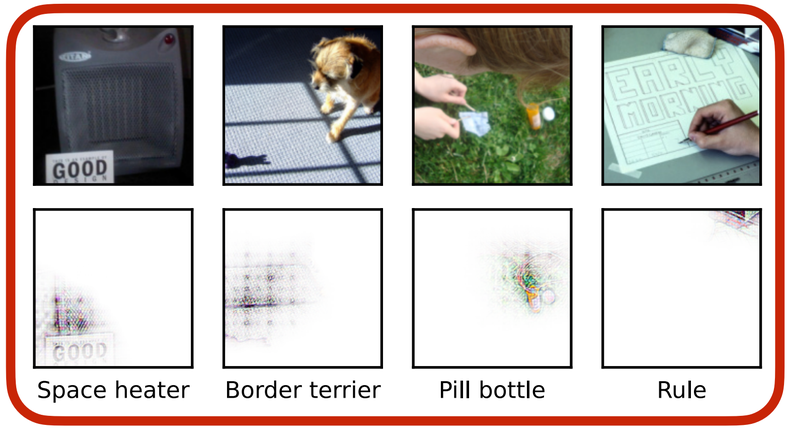


ViT

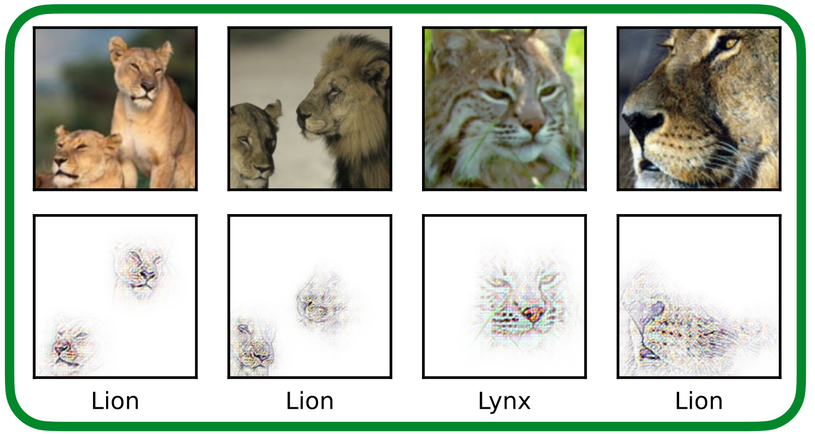


Concept Consistency

How to evaluate the consistency of concepts?

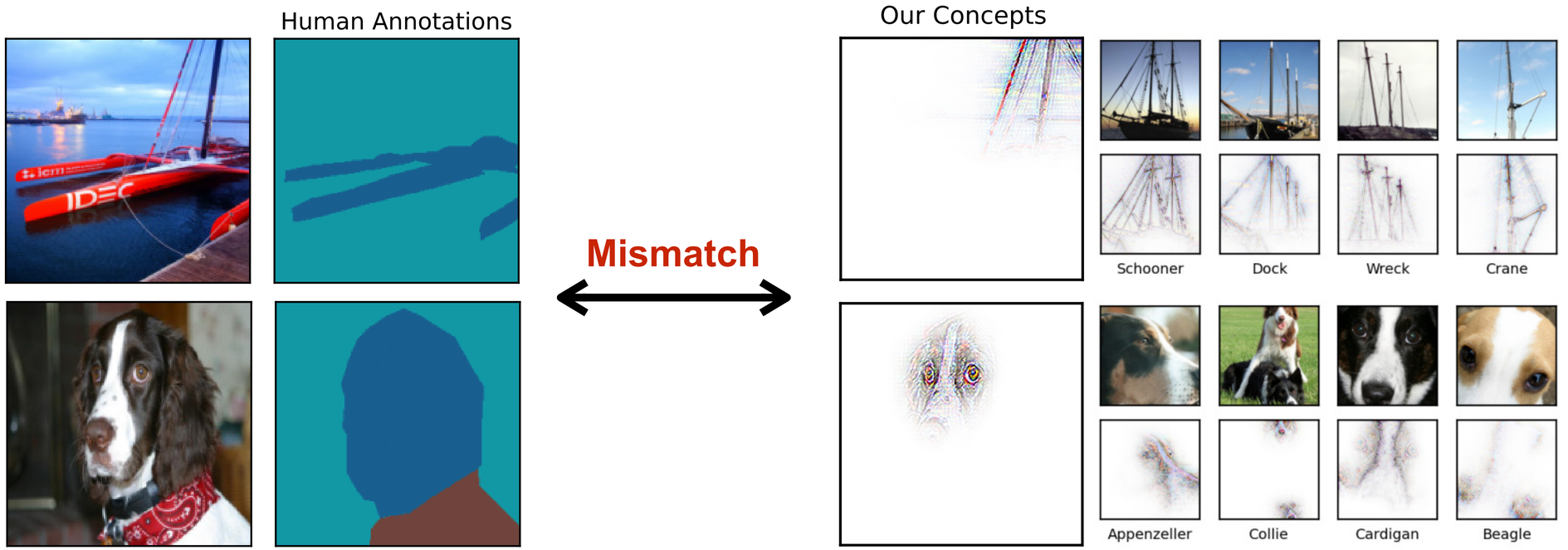


Low Consistency



High Consistency

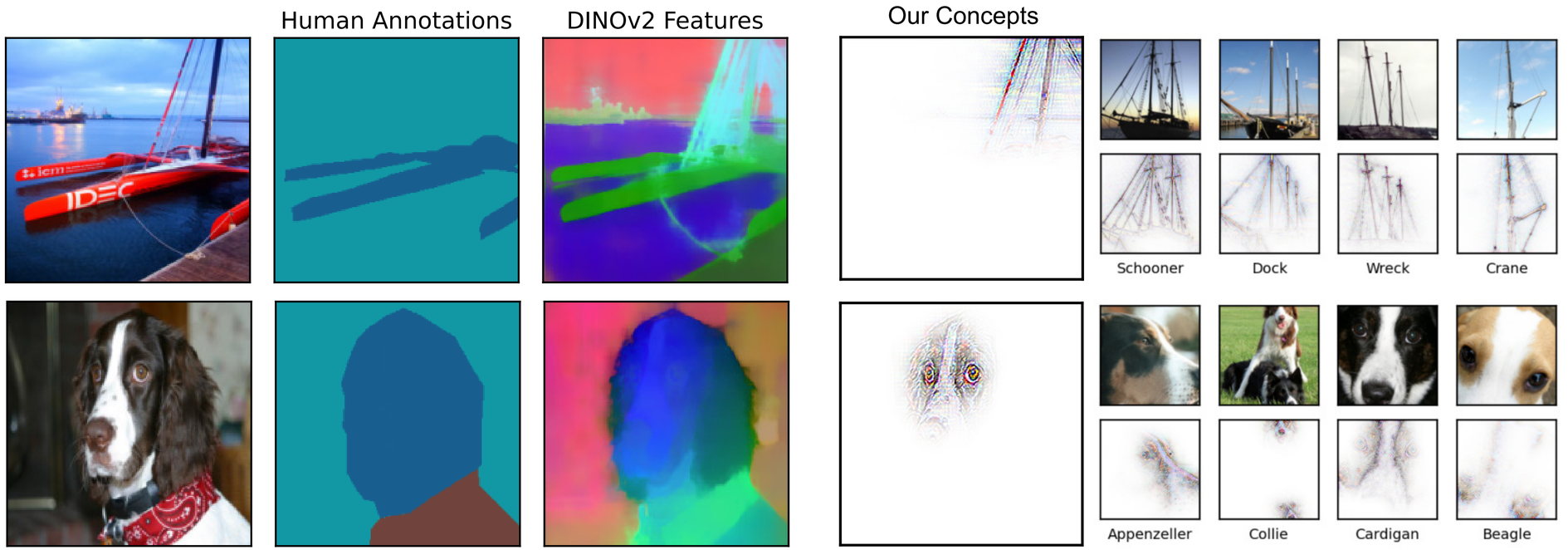
Evaluating Concept Consistency beyond Annotations



Not easy to anticipate the concepts during annotation!

Annotations from PartImageNet (He et al. ECCV 2022); Measuring consistency with annotations from ProtoEval (Huang et al. ICCV 2023)

Evaluating Concept Consistency beyond Annotations



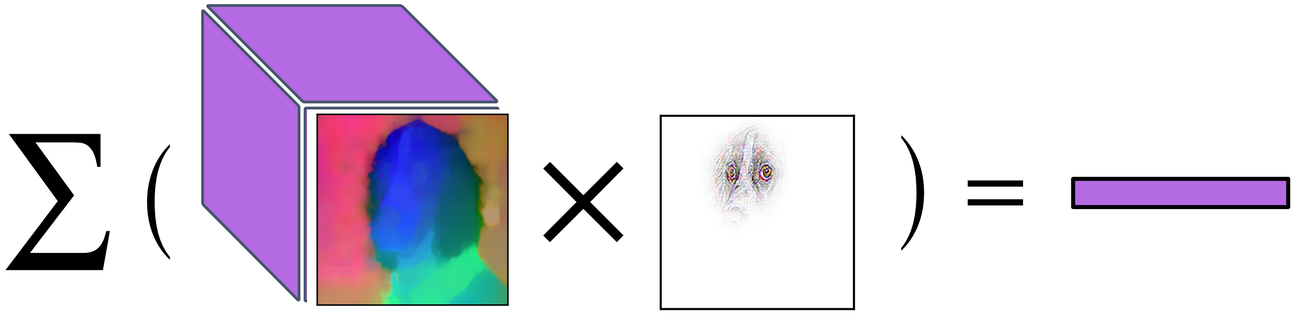


C^2-score: Measure concept consistency in DINOv2 space

PCA visualization from DINOv2 (Oquab et a. TMLR 2024) upsampled with LoftUP (Huang et al. ICCV 2025); bottom figure from (Zhang et al. NeurIPS 2023)

Evaluating Concept Consistency

C^2-score: Measure concept consistency in DINOv2 space

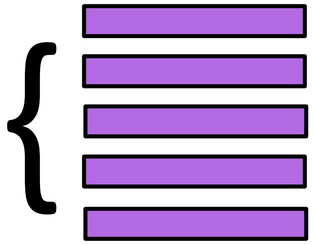


DINOv2 features (upsampled with LoftUP)

Concept attribution

Measure consistency across top-activating images

(average pair-wise cosine similarity)



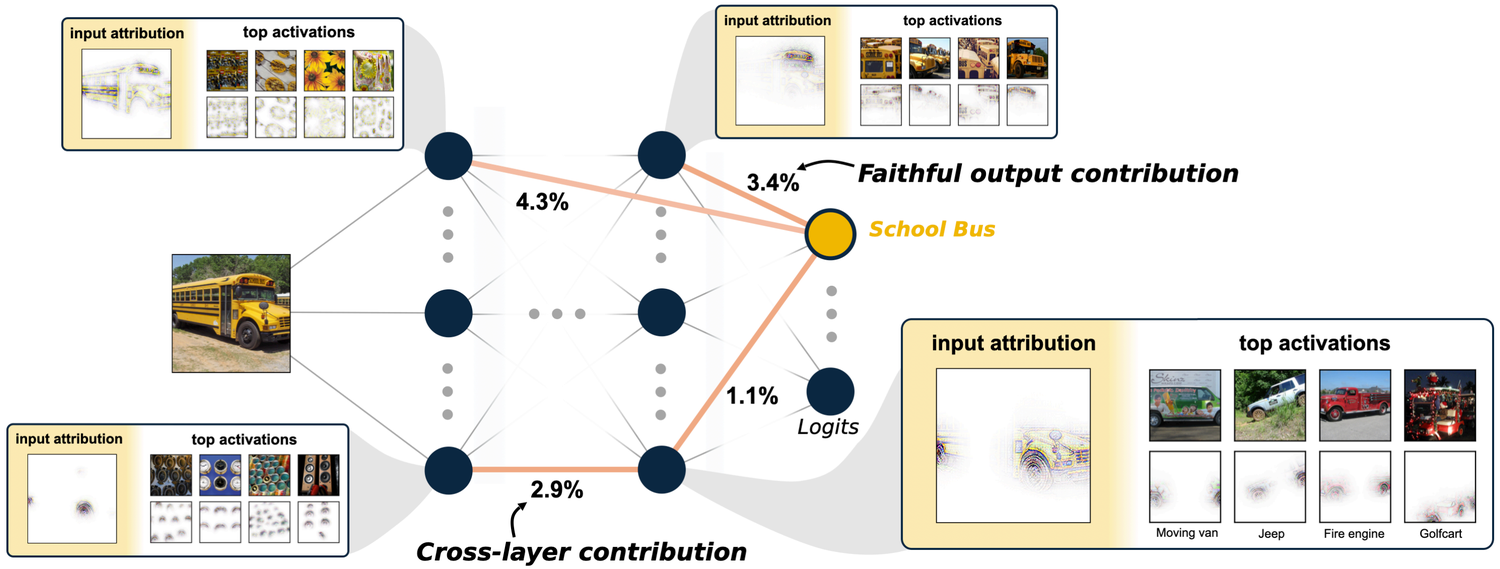
★ Rich and dense representation

★ Independent of annotations (part- and image-level)

PCA visualization from DINOv2 (Oquab et a. TMLR 2024) upsampled with LoftUP (Huang et al. ICCV 2025)

Our Setup

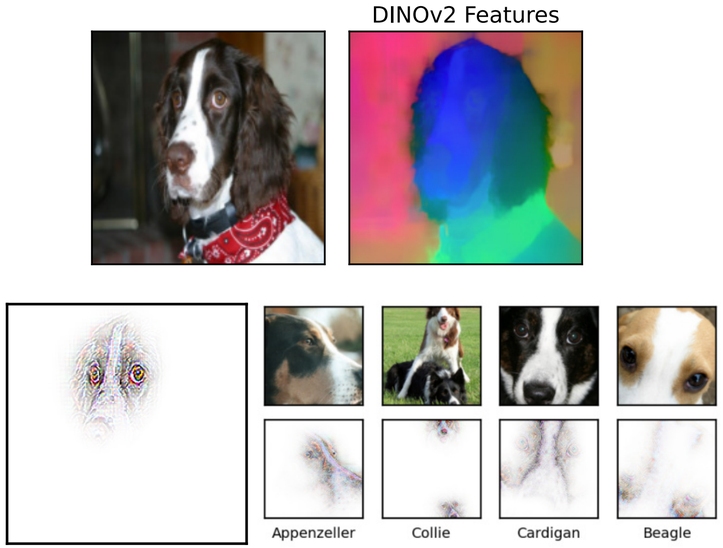
FaCT: Faithful Concept Tracing



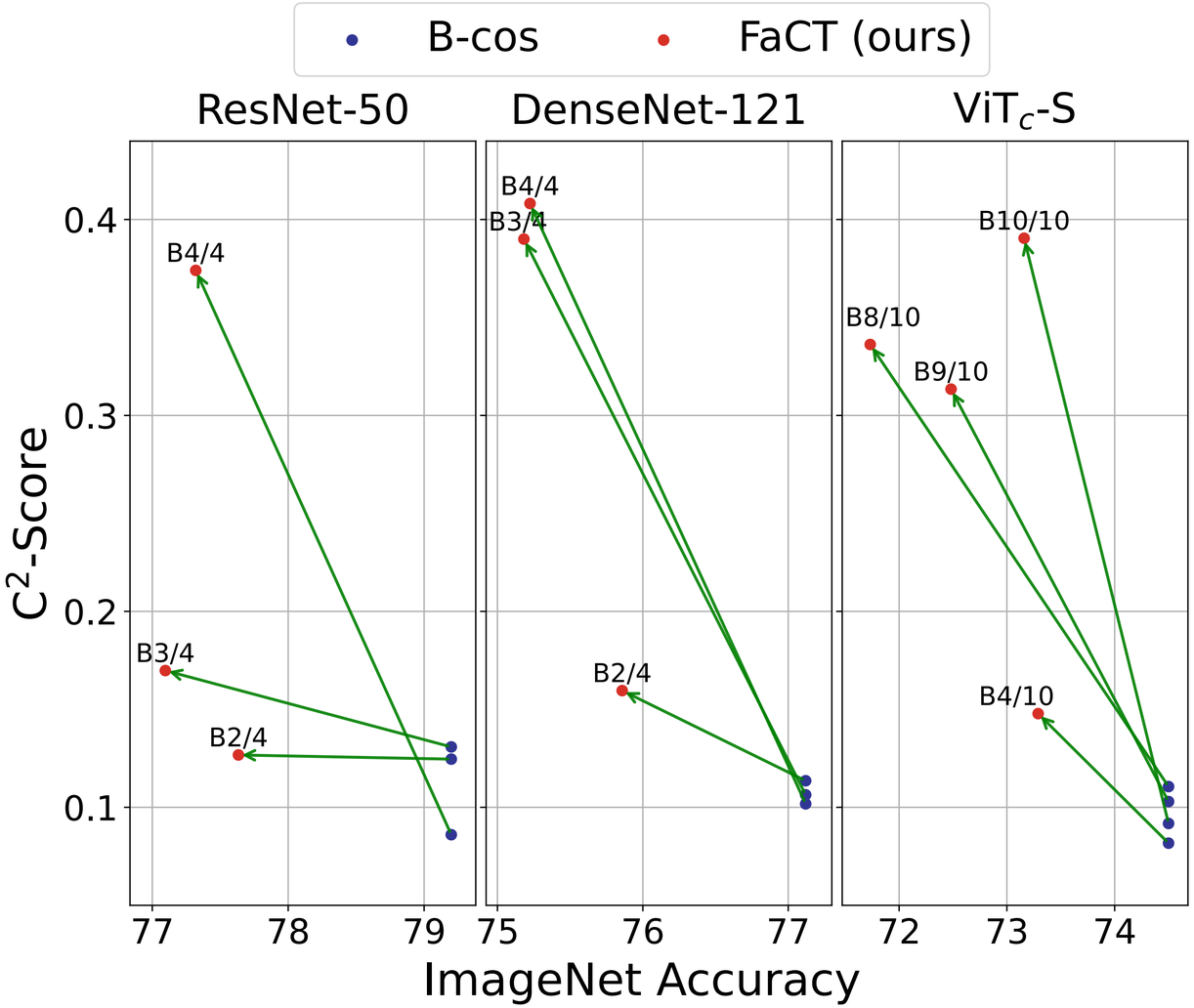
$$\text{Concept Activation} = \sum \text{Pixel Contribution}$$

$$\text{Output Logit} = \sum \text{Concept Contribution}$$

C2-Score: Consistency Evaluation



FaCT: Competitive and Consistent

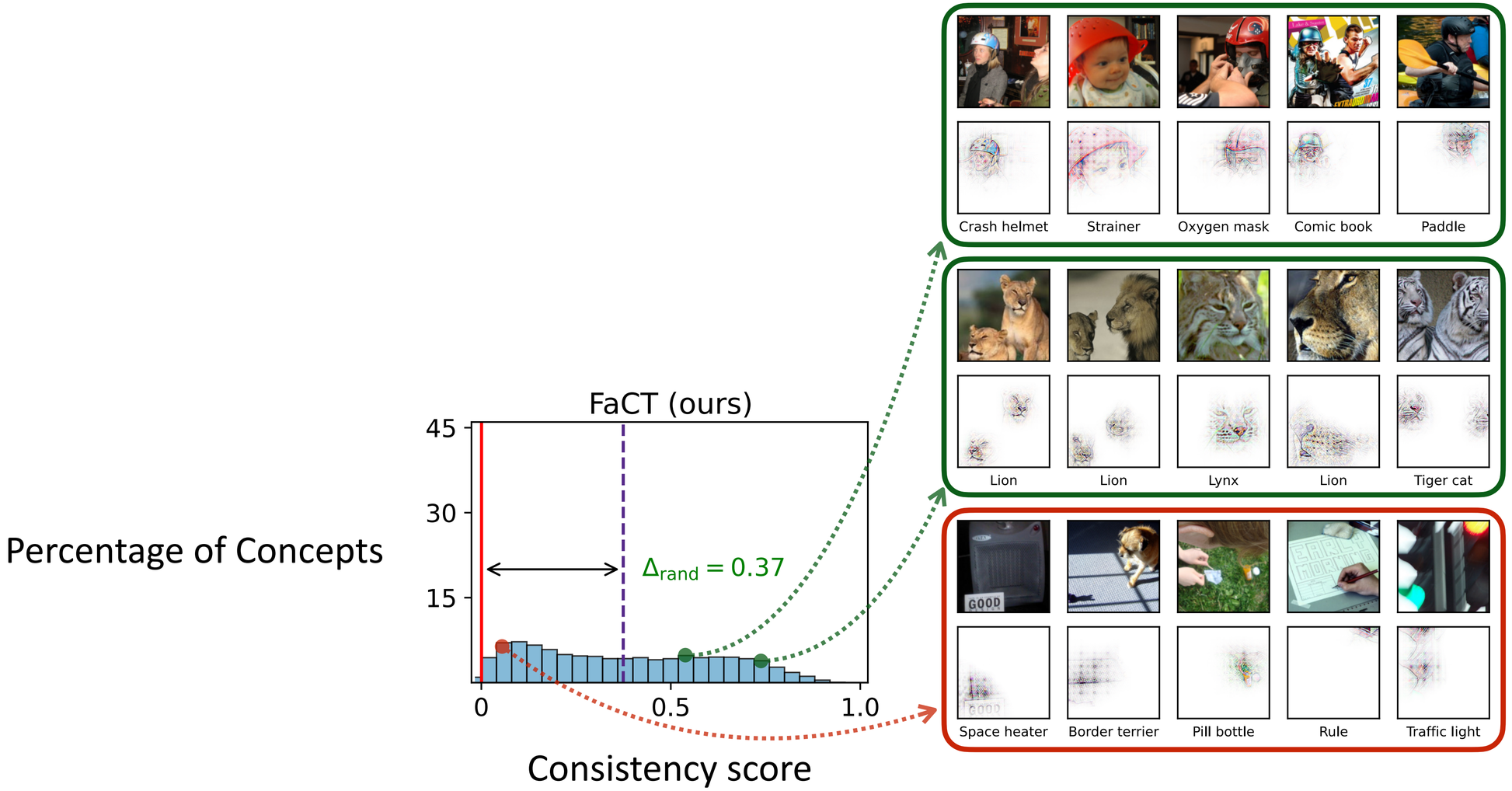


Competitive ImageNet performance

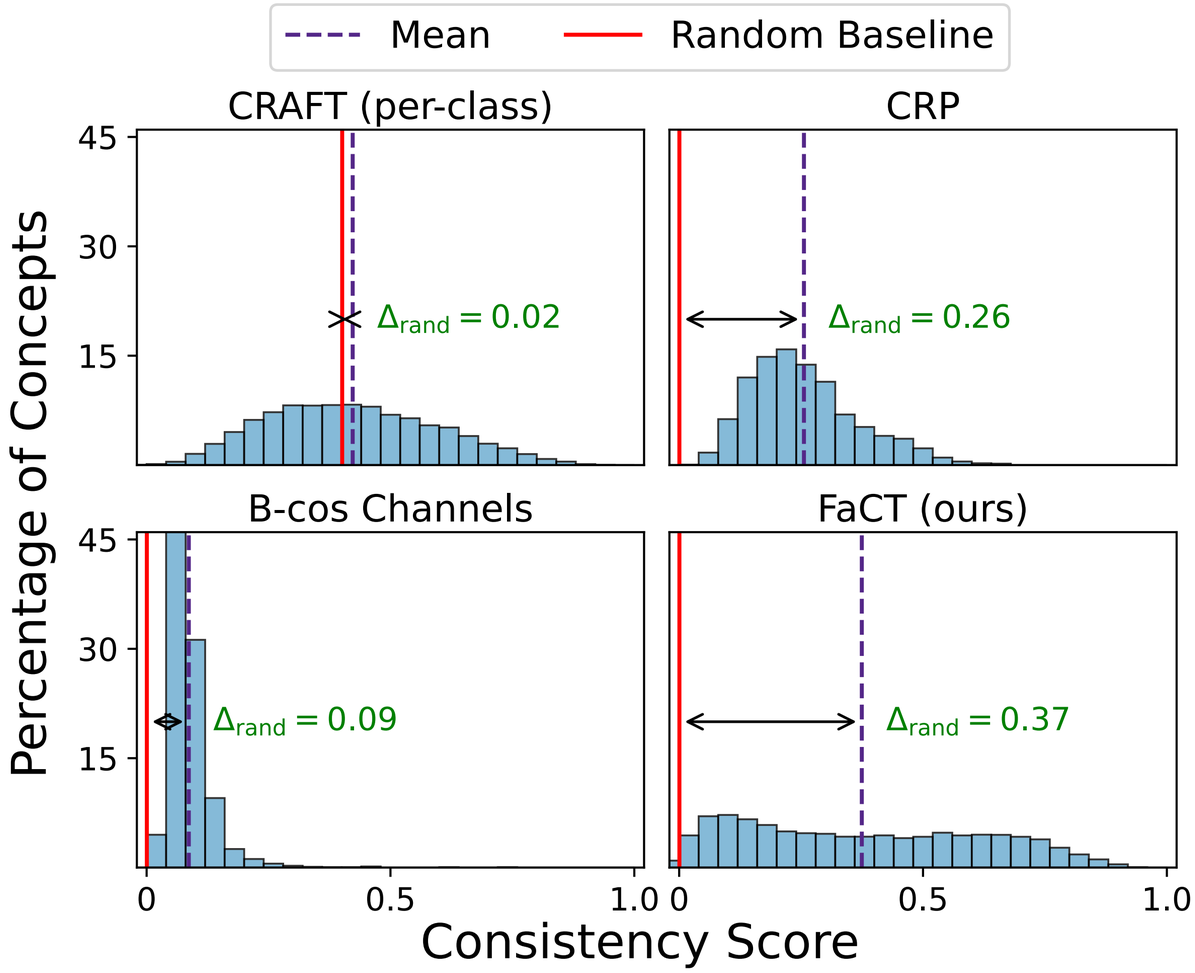
Generalization across layers and architectures

Significantly more consistent concepts

Quantitatively Evaluating the Concepts



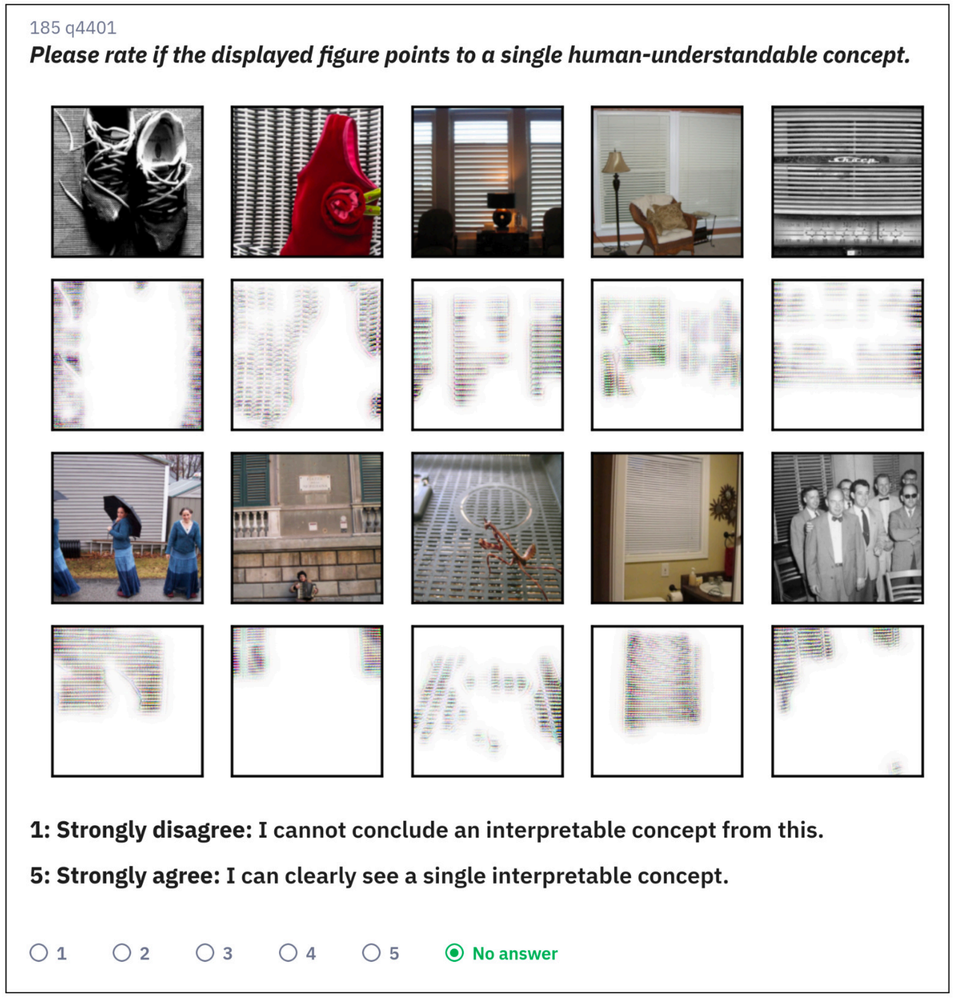
Comparing to Prior Work



Quantitatively more consistent concepts!

CRAFT; Fel et al. CVPR 2023; CRP; Achtibat et al. Nature M. Intel. 2023; B-cos Böhle et al.; CVPR 2022 TPAMI 2024

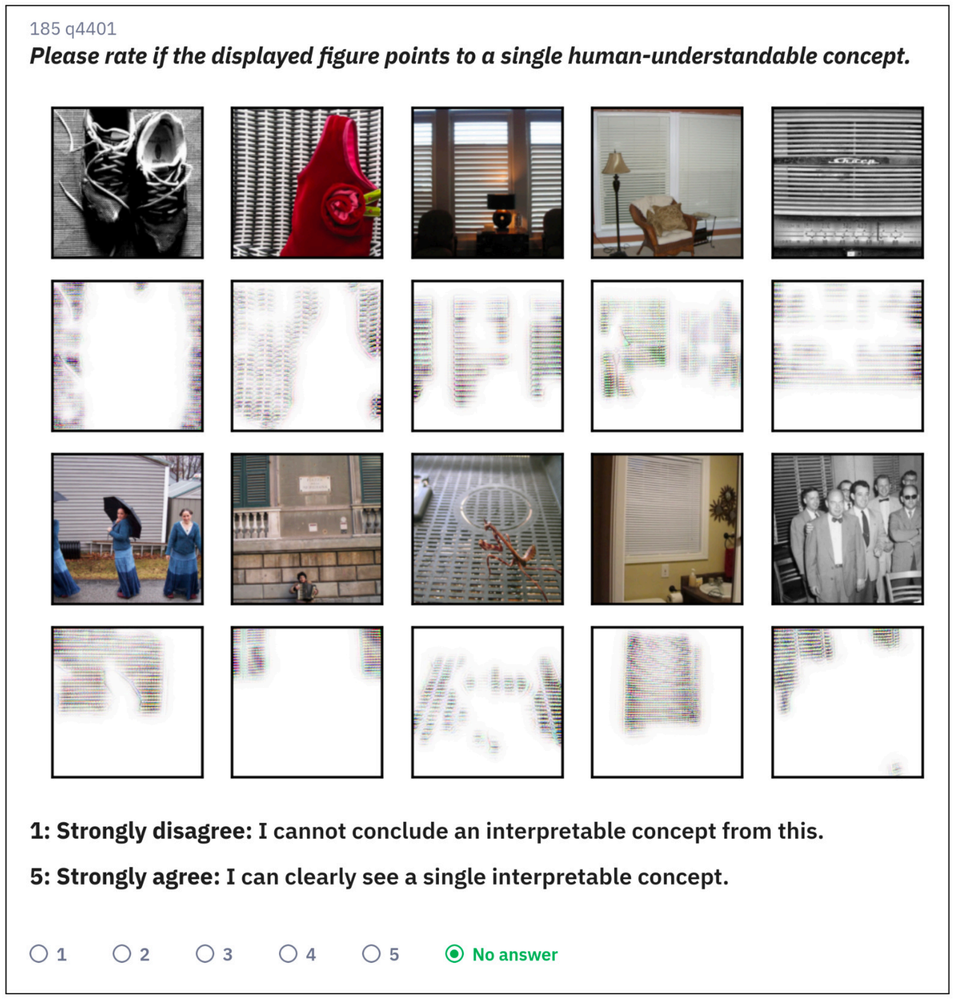
Evaluating Concept Interpretability (User Study)

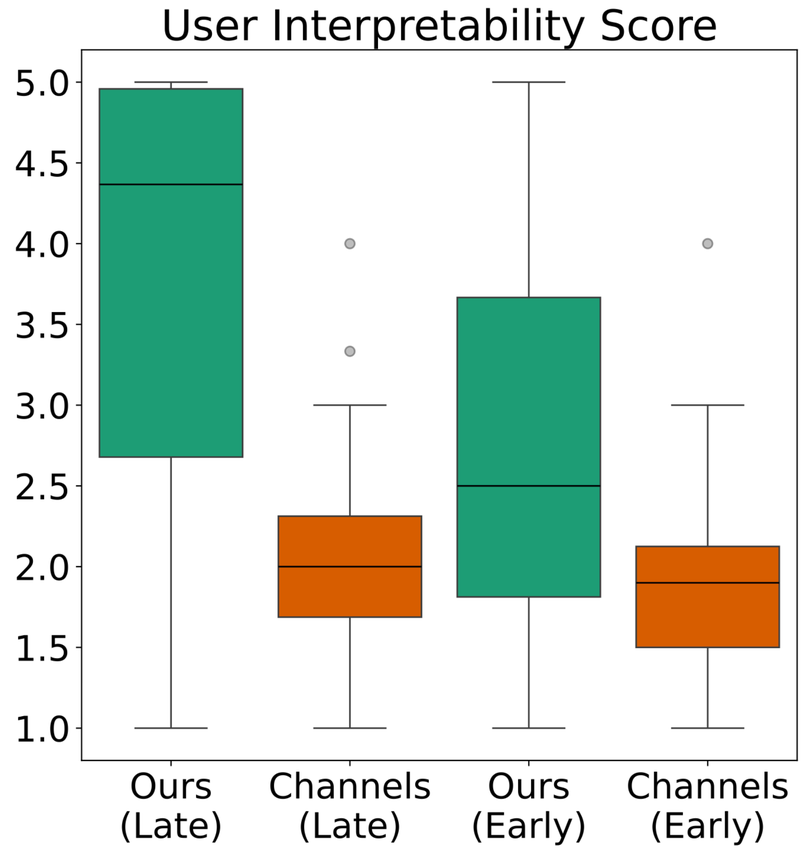




Score from 1 to 5

Evaluating Concept Interpretability (User Study)

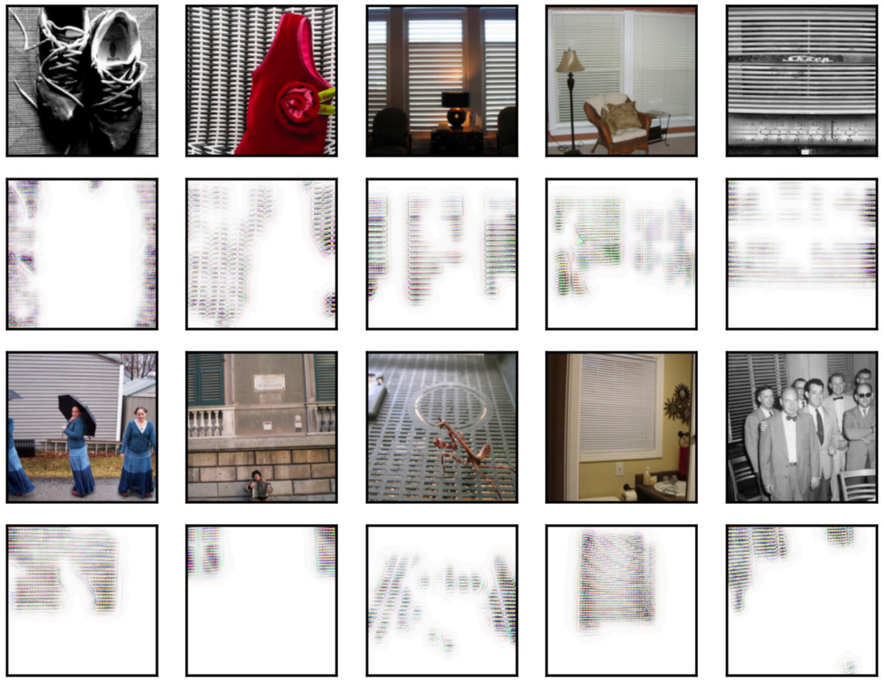




Evaluating Concept Interpretability (User Study)

185 q4401

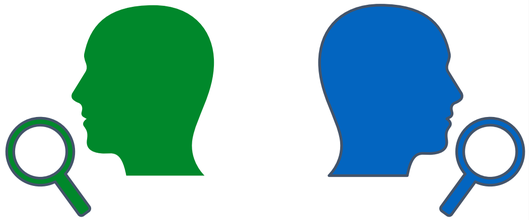
Please rate if the displayed figure points to a single human-understandable concept.



1: Strongly disagree: I cannot conclude an interpretable concept from this.

5: Strongly agree: I can clearly see a single interpretable concept.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ◉ No answer



300 q4721

Please rate if the displayed figure points to a single human-understandable concept.



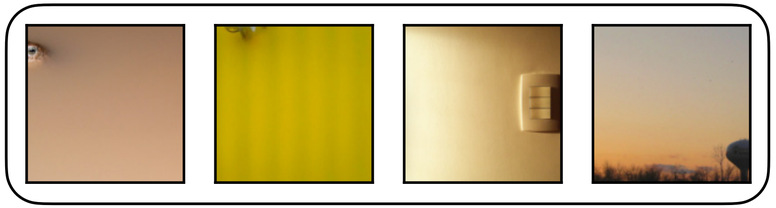
1: Strongly disagree: I cannot conclude an interpretable concept from this.

5: Strongly agree: I can clearly see a single interpretable concept.

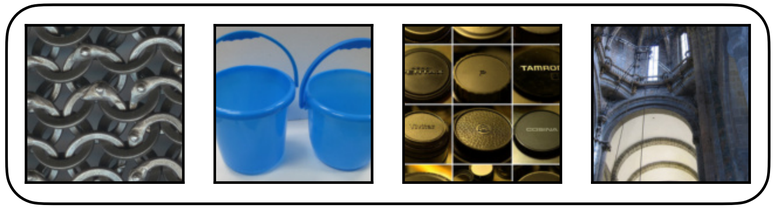
○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ◉ No answer

Visualizations Matter!

Score **2.0**



Score **1.8**



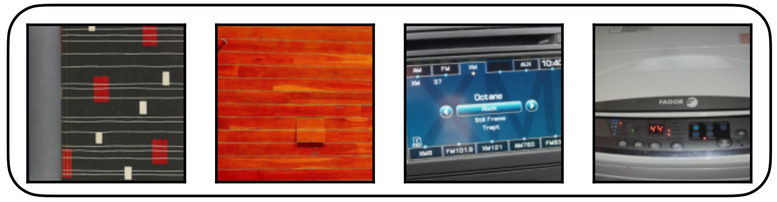
Range (1 - 5)



Score **1.0**

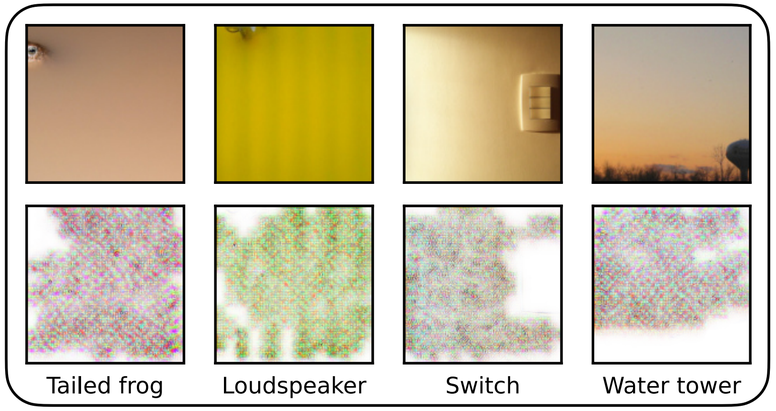


Score **2.0**

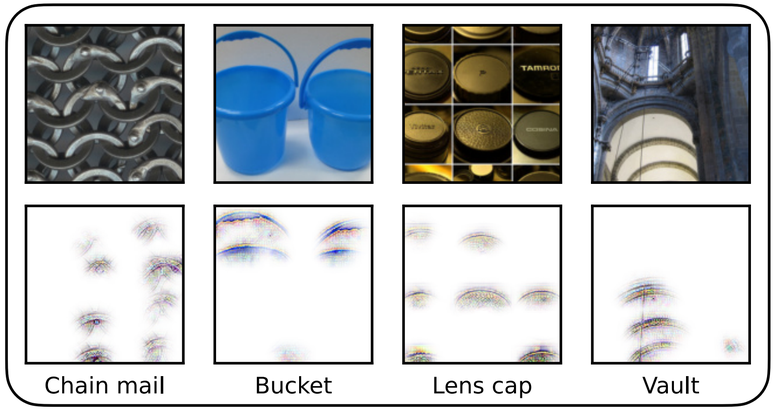


Visualizations Matter!

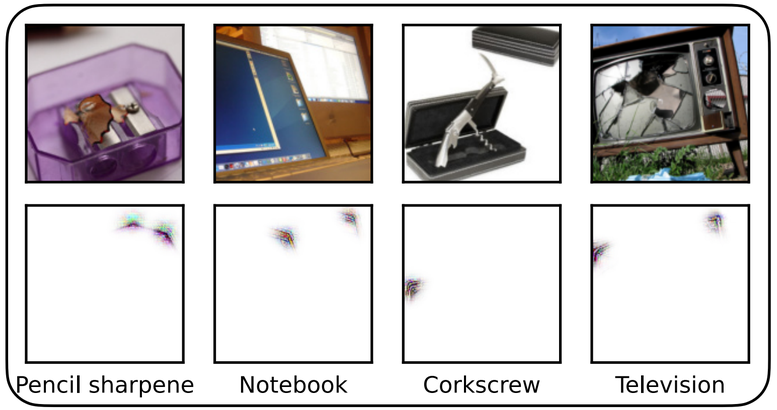
Score 2.0 → 5.0



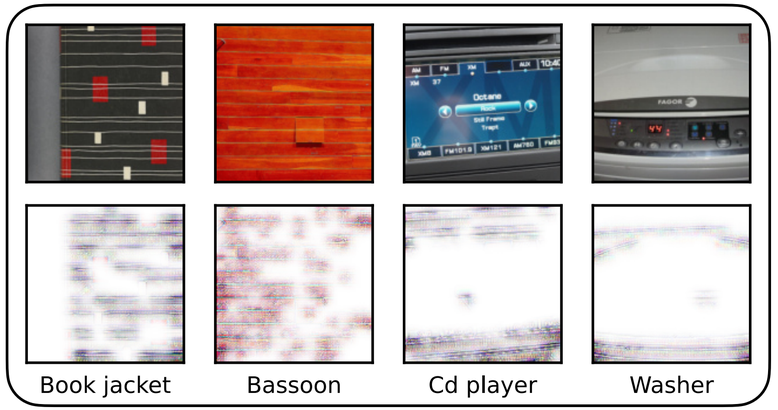
Score 1.8 → 4.3



Score 1.0 → 3.2



Score 2.0 → 3.8



Range (1 - 5)



Input attributions aid the interpretability!

Leveraging the Shared Concepts



Volleyball Logit: 3.7

Basketball Logit: 3.0

Why is the model confused between the two classes?

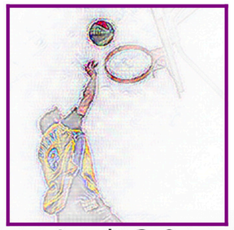
Leveraging the Shared Concepts



Volleyball Logit: ***3.7***



Basketball Logit: 3.0

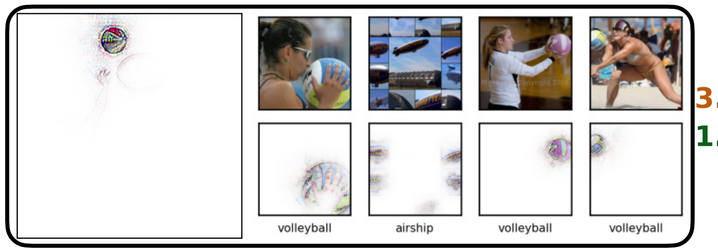


Why is the model confused between the two classes?

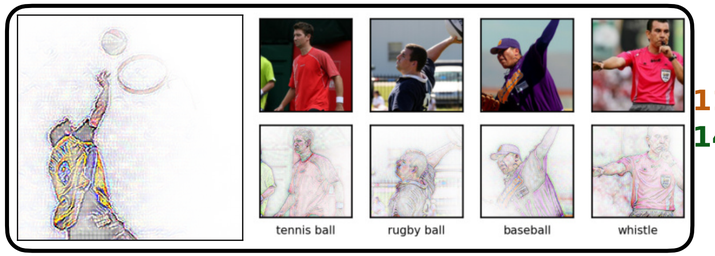
Attribution method from B-cos Networks, Böhle et al., CVPR 2022, TPAMI 2024

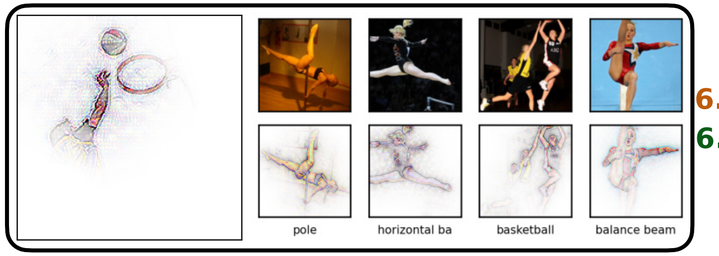
Leveraging the Shared Concepts

Mutually Contributing Concepts









3.0 %
1.5 %

3.1 %
4.9 %

11.6%
14.2%

6.3 %
6.1 %



***Volleyball Logit**: 3.7*



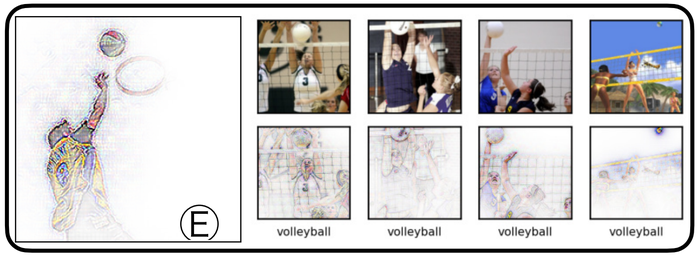
Basketball Logit: 3.0

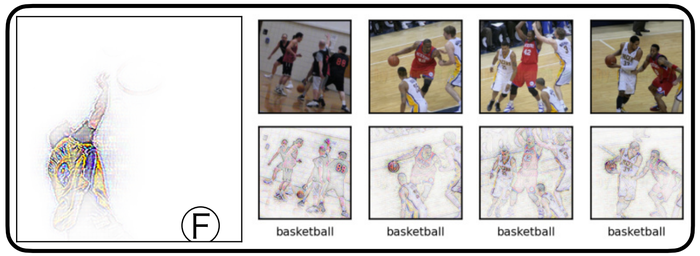


14.7%
0.0 %

0.7 %
4.0 %

Exclusively Contributing Concepts

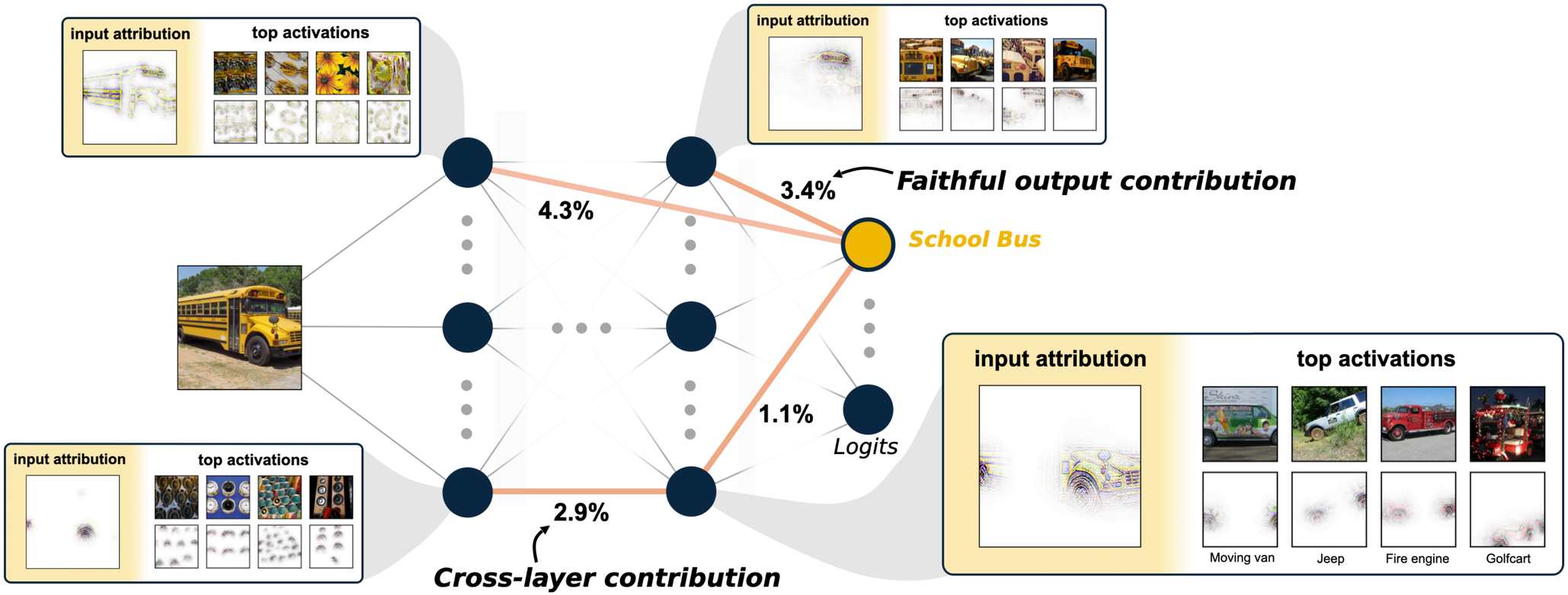




Leveraging the Shared Concepts



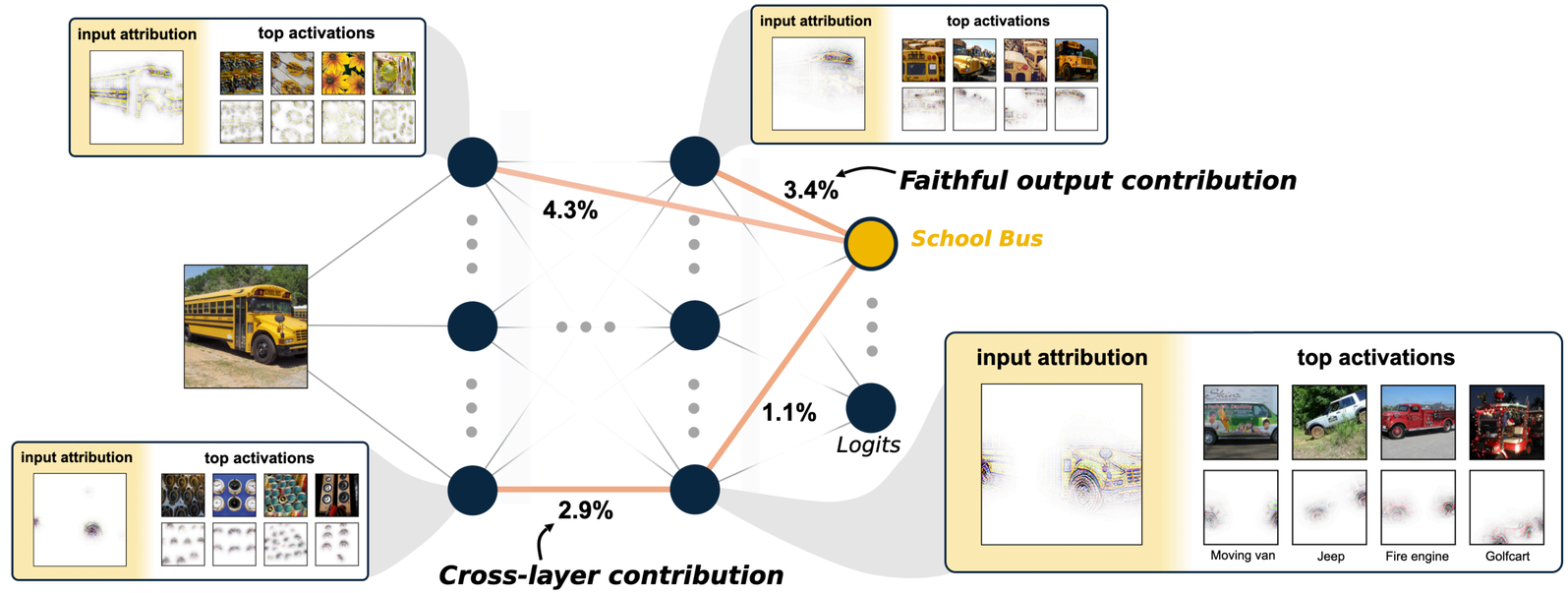
Faithful Concept Traces



$$\text{Concept Activation} = \sum \text{Pixel Contribution}$$

$$\text{Output Logit} = \sum \text{Concept Contribution}$$

Happy to hear your thoughts!



Paper

arxiv.org/abs/2510.25512

Code & Models

github.com/m-parchami/FaCT

Poster Sessions

San Diego (NeurIPS)

Thu 4 Dec 4:30-7:30 p.m. PST

Exhibit Hall C/D/E #1000

Copenhagen (EurIPS)

Thu 4 Dec 10:30 AM -12:30 PM. CET

Hall D3 #21