DeepREAL

ICML
International Conference
On Machine Learning

Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers

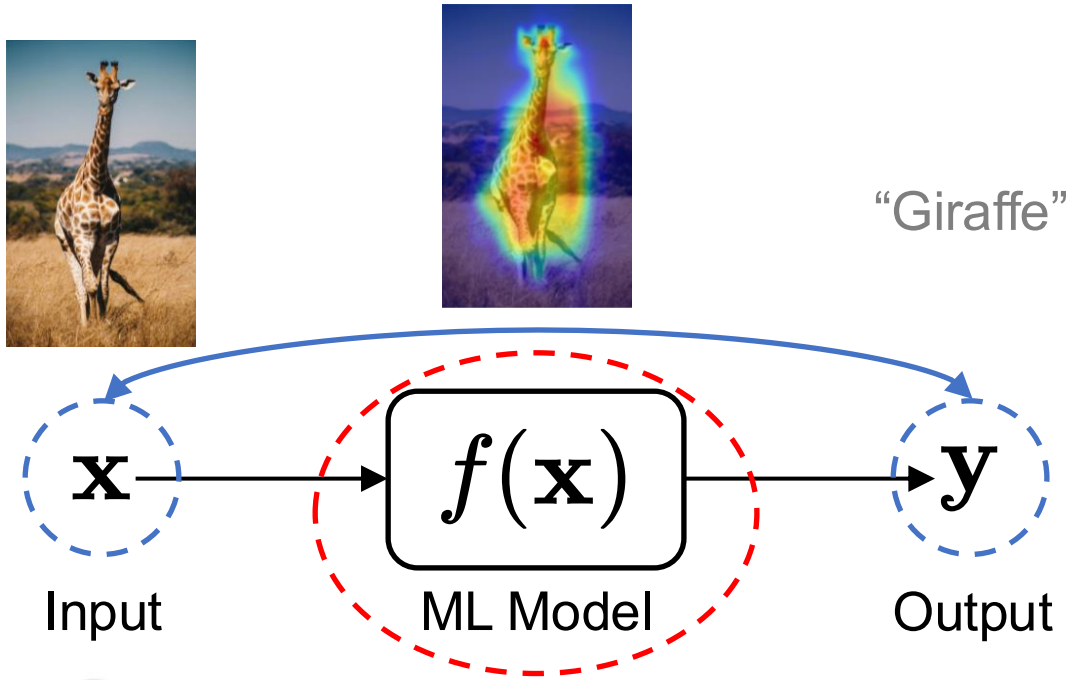
Tang Li Yanlin Chen Mengmeng Ma Xi Peng

Deep Robust & Explainable AI Lab (DeepREAL)

Tue, Jul 7, 2026 10:30 AM – 12:15 PM KST Coex: HALL A

From "correlation" to "cognition"

"What" lead to the prediction



but **"How"** does it think inside?

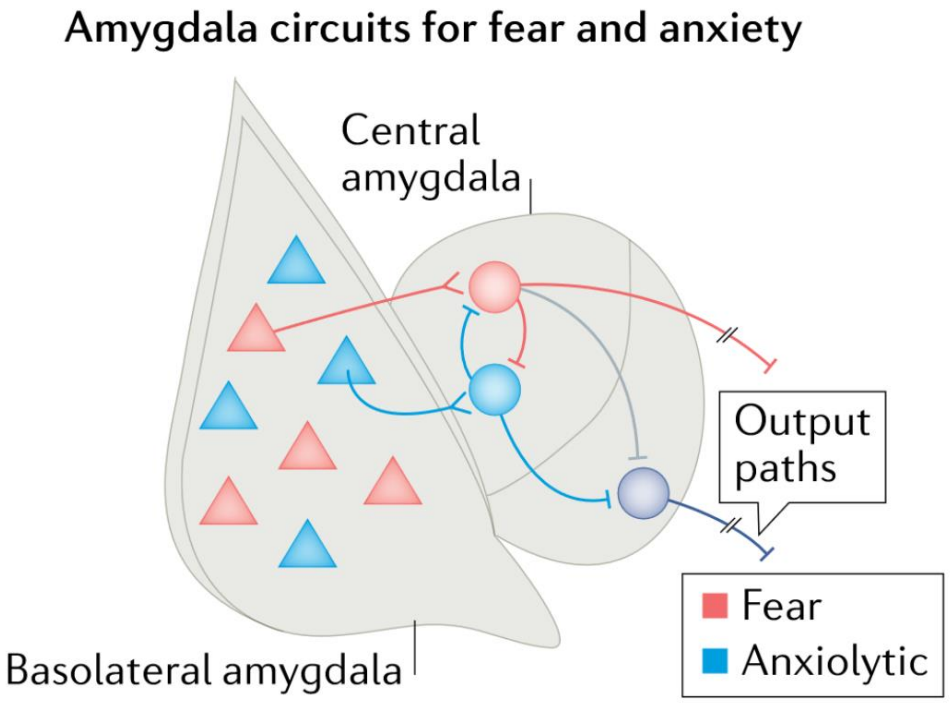
In other words, can we "read the mind" of AI?

[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Two views to understand biological brains: Cognitive Science of Human

Sherringtonian View

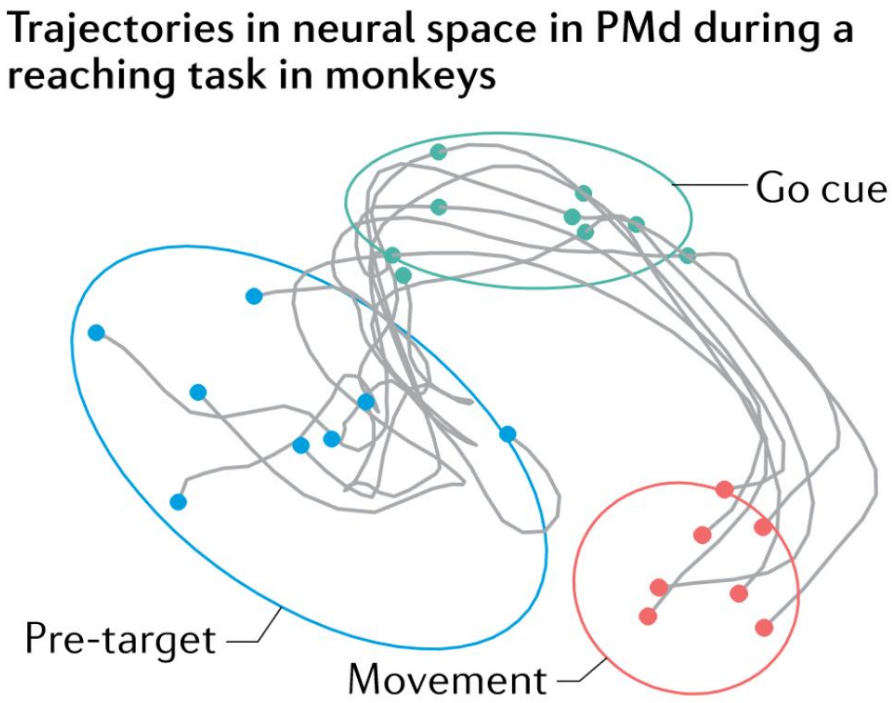
- Node-to-node connections
- Transfer between neuron computations



[Nature'15, Janak et al.]

Hopfieldian View

- Representational spaces
- Transformations between spaces



[Neurosci'10, Churchland et al.]

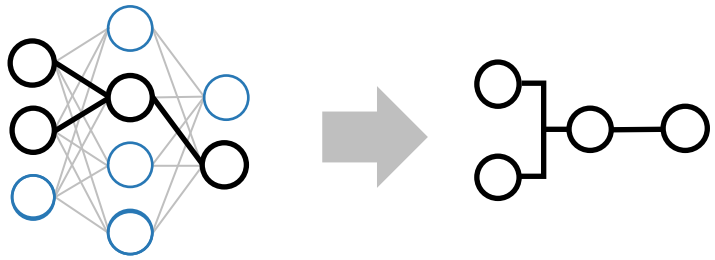
[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Two views to understand artificial brains (ML models):

Sherringtonian View

Bottom-up View

- Node-to-node connections
- Identify functional circuits



Sub-computation Graph → Circuit

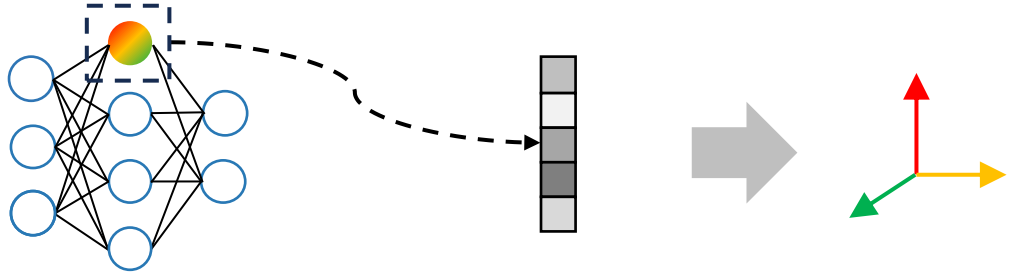
- *Activation patching* [Conmy'23]
- *Path patching* [Goldowsky-Dill'23]
- *Causal abstraction* [Geiger'23]

⊖ Local Semantics ⊕ Global Structure

Hopfieldian View

Top-down View

- Representation space transforms
- Disentangle features from representations



Polysemantic Neuron → Representations → Monosemantic features

- *Probing* [Burns'23]
- *Representation Engineering* [Zou'23]
- *Sparse dictionary learning* [Cunningham'24]

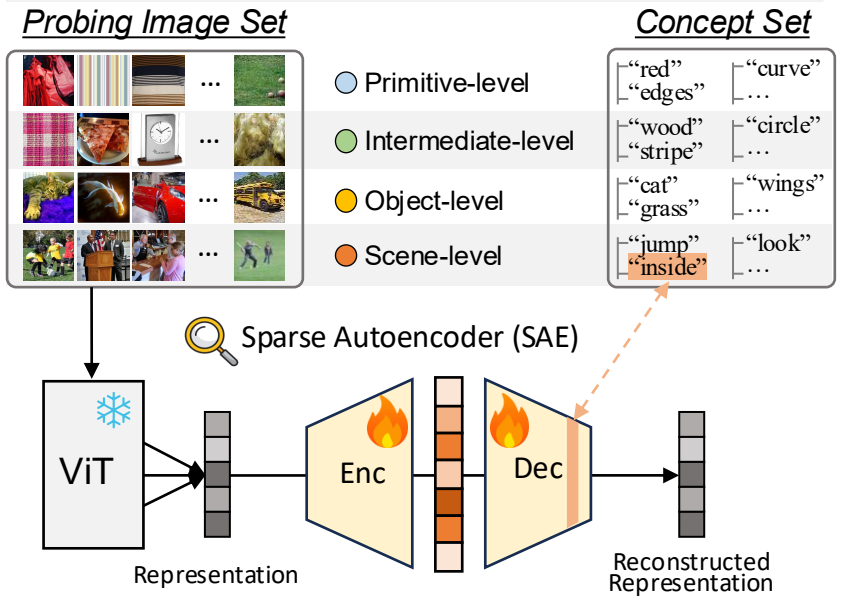
⊕ Local Semantics ⊖ Global Structure

Can we bridge Bottom-up and Top-down Interpretations to develop the "Cognitive Science of AI"?

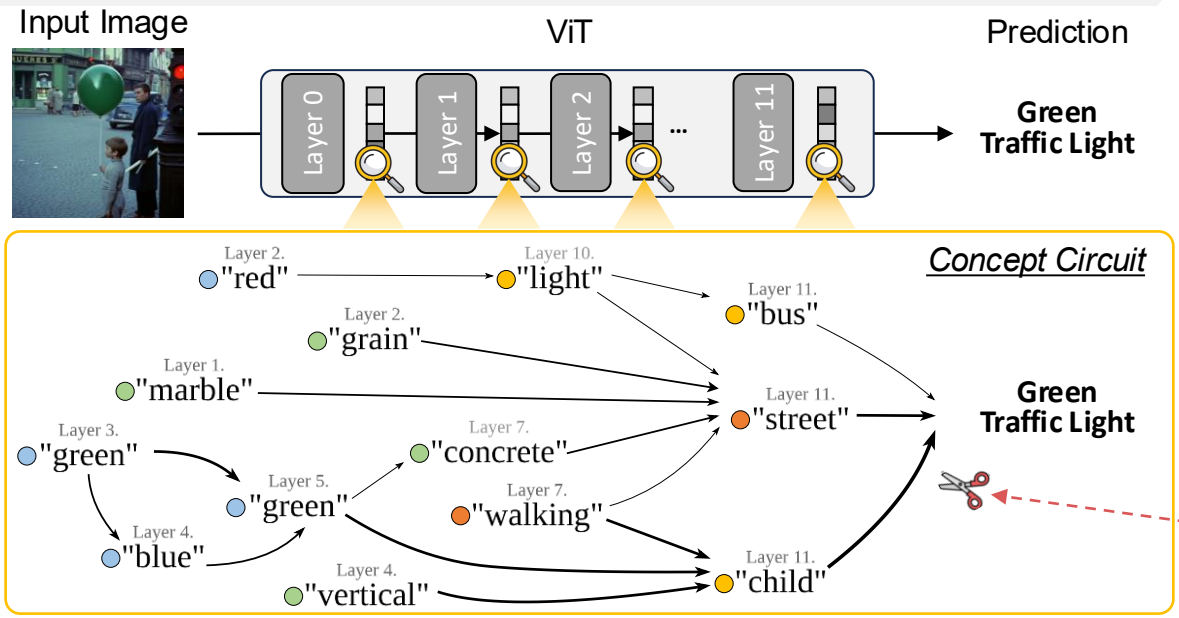
[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Our Solution: ViSAE Toolbox for Interpreting ViT Inner Workings

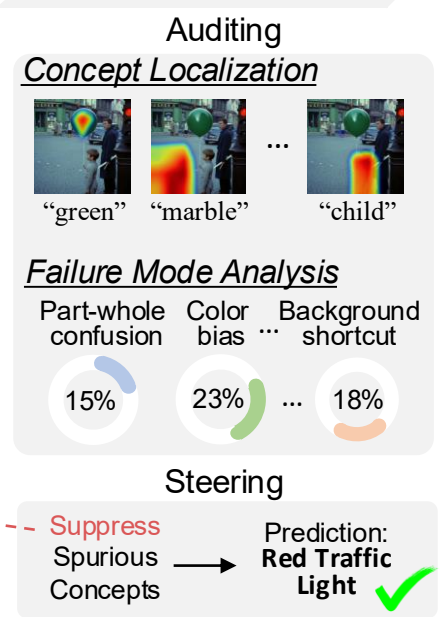
Neuroscience-motivated Probing Suite



Concept Circuits Tracing Algorithm

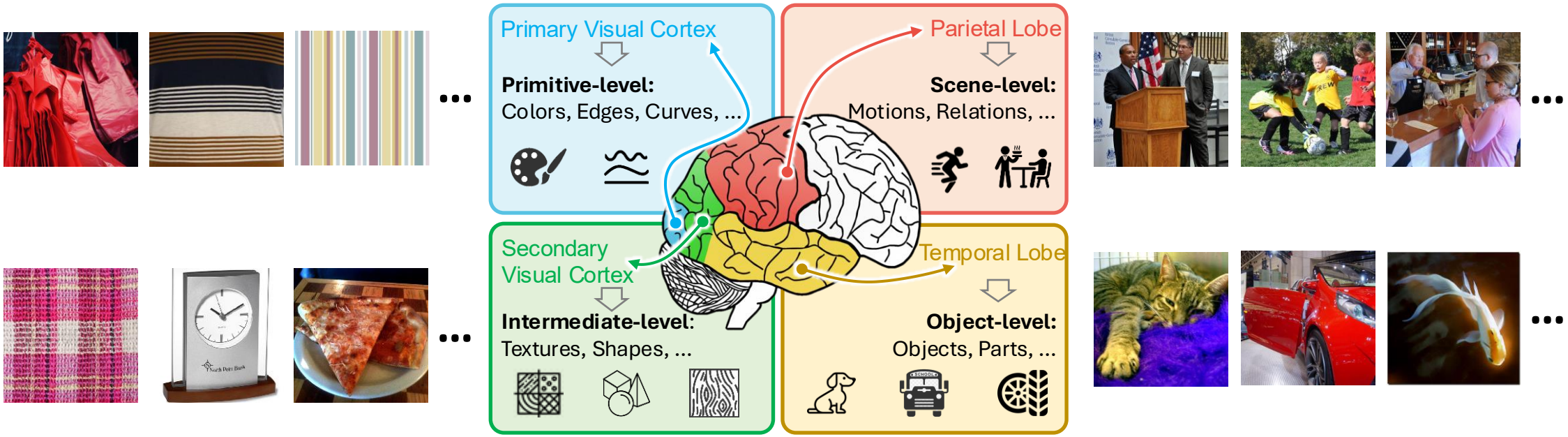


Applications



[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Neuroscience-Motivated Probing Suite



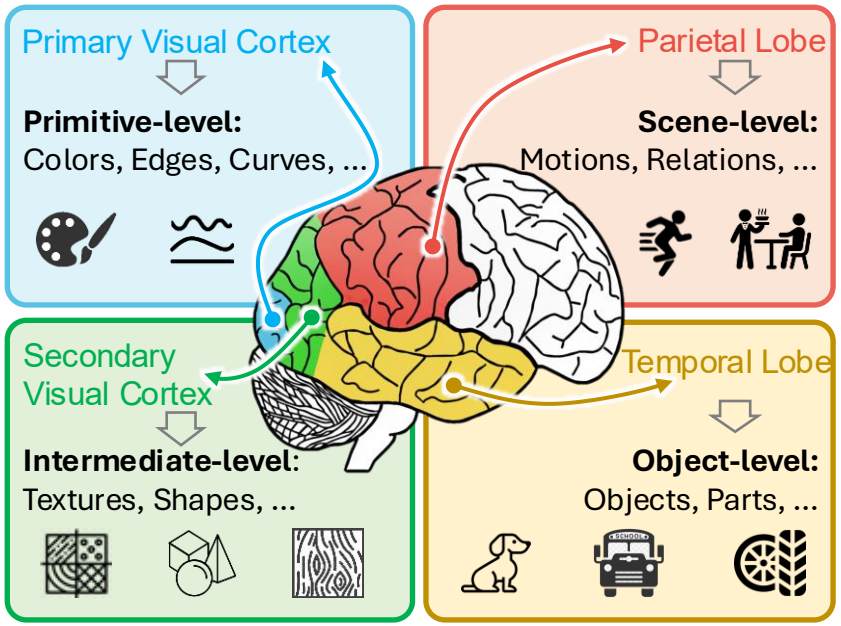
Probing Image Set	Data Source	# of Images	Concepts Covered by Images (%)					Coverage Efficiency ↑ (%/1K Images)
			Primitive	Intermediate	Object	Scene	Avg.	
ImageNet	ImageNet	1,281K	81.0	78.2	**97.7**	59.0	78.9	0.06
MSCOCO	MSCOCO	118K	69.6	65.4	80.4	**63.1**	69.6	0.59
Ours	Primitive Level: DTD, Broden; Intermediate Level: Broden, ShapeNet; Object Level: ImageNet, VisualGenome; Scene Level: Places365, MSCOCO;	64K	**87.1**	**80.6**	92.6	61.7	**80.5**	**1.26**

[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Neuroscience Motivated Probing Suite

"red", "blue", "lime", "lines", "edges", "dots", ...

"wooden", "striped", "triangle", "steel", "round", ...



"playing basketball", "leap", "above", ...

"dog", "airplane", "chair", bird, ...

Concept Statistics

Level	# of Concepts
Primitive	1,073
Intermediate	1,723
Object	10,534
Scene	2,720
Total	16,050

Human Evaluations

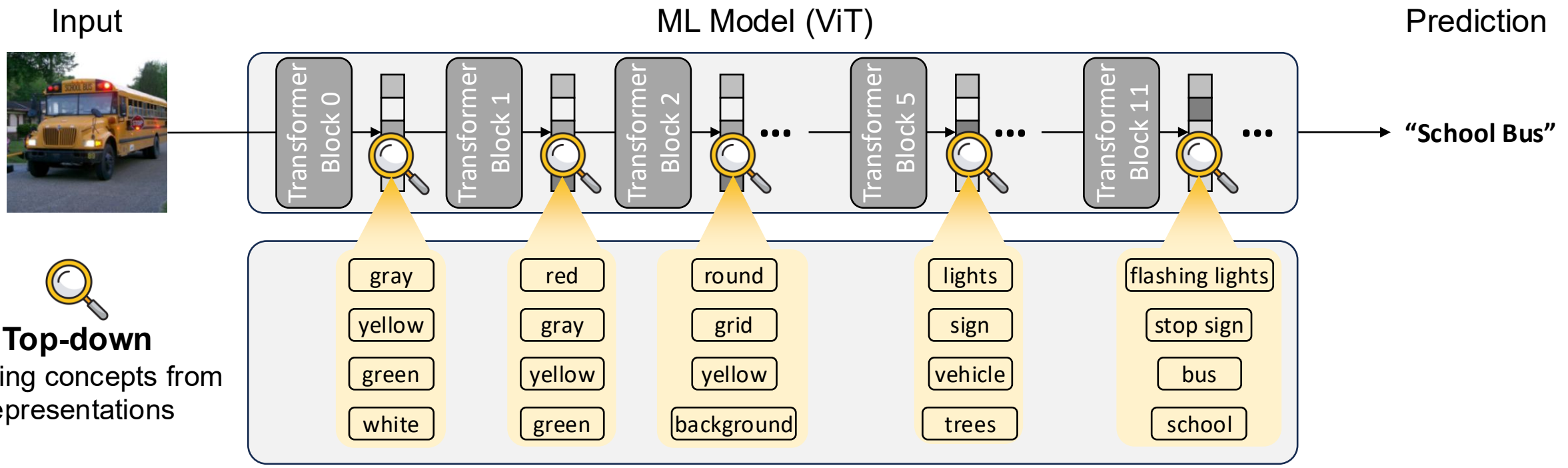
Evaluator	Faithfulness	Completeness
Human_A	4.65 ± 0.06	4.72 ± 0.04
Human_B	4.83 ± 0.20	4.73 ± 0.06
Human_C	4.90 ± 0.05	4.78 ± 0.18
Avg.	4.79	4.74

Comparison of Concept Quality

Concept Set	# of Concepts	Redundancy (↓)	Visually Grounded (↑)
LAION-freq	15K	0.419	0.602
Google-freq	20K	0.654	0.432
LaBo	10K	0.584	0.565
Ours	16K	**0.213**	**0.864**

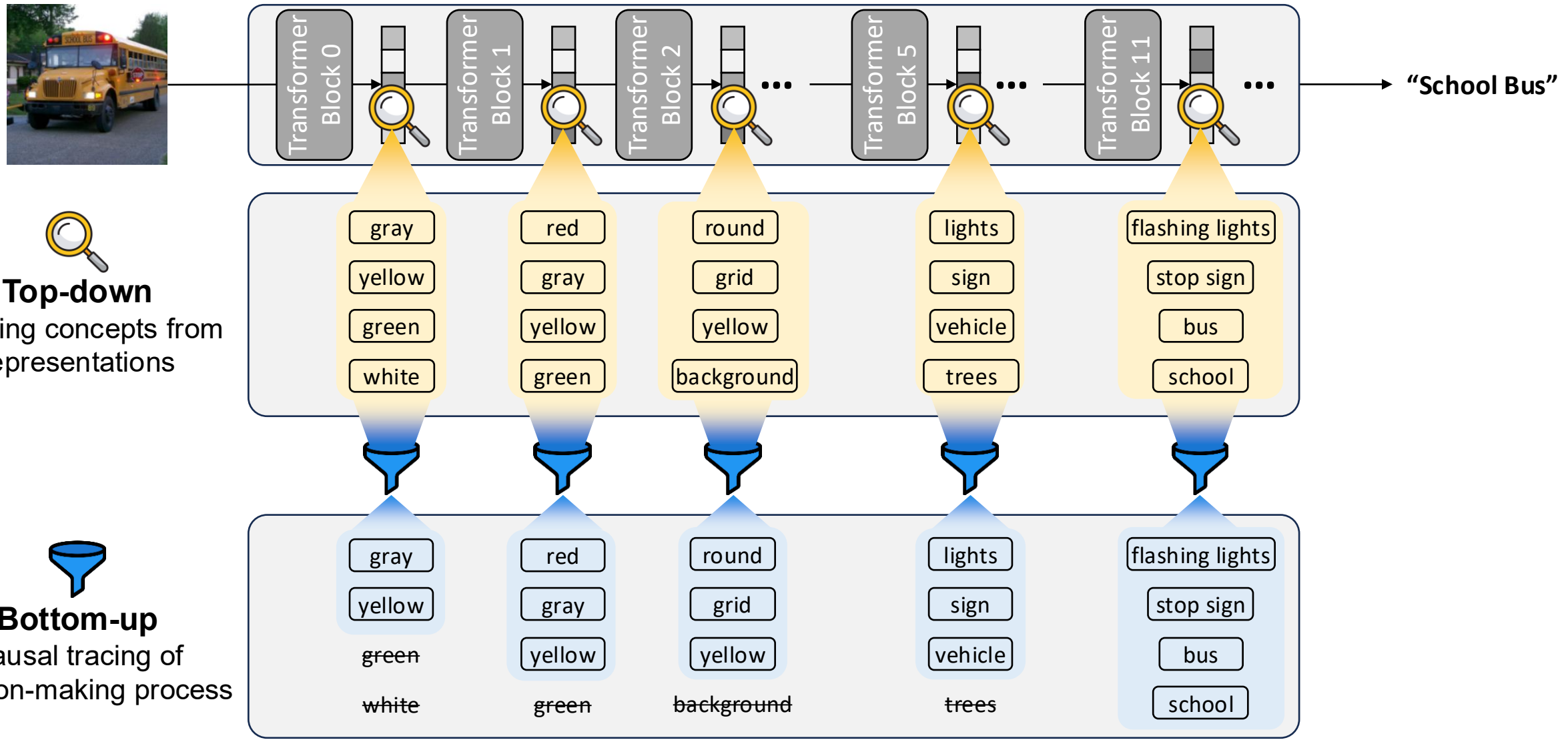
[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Concept Circuit Tracing Algorithm



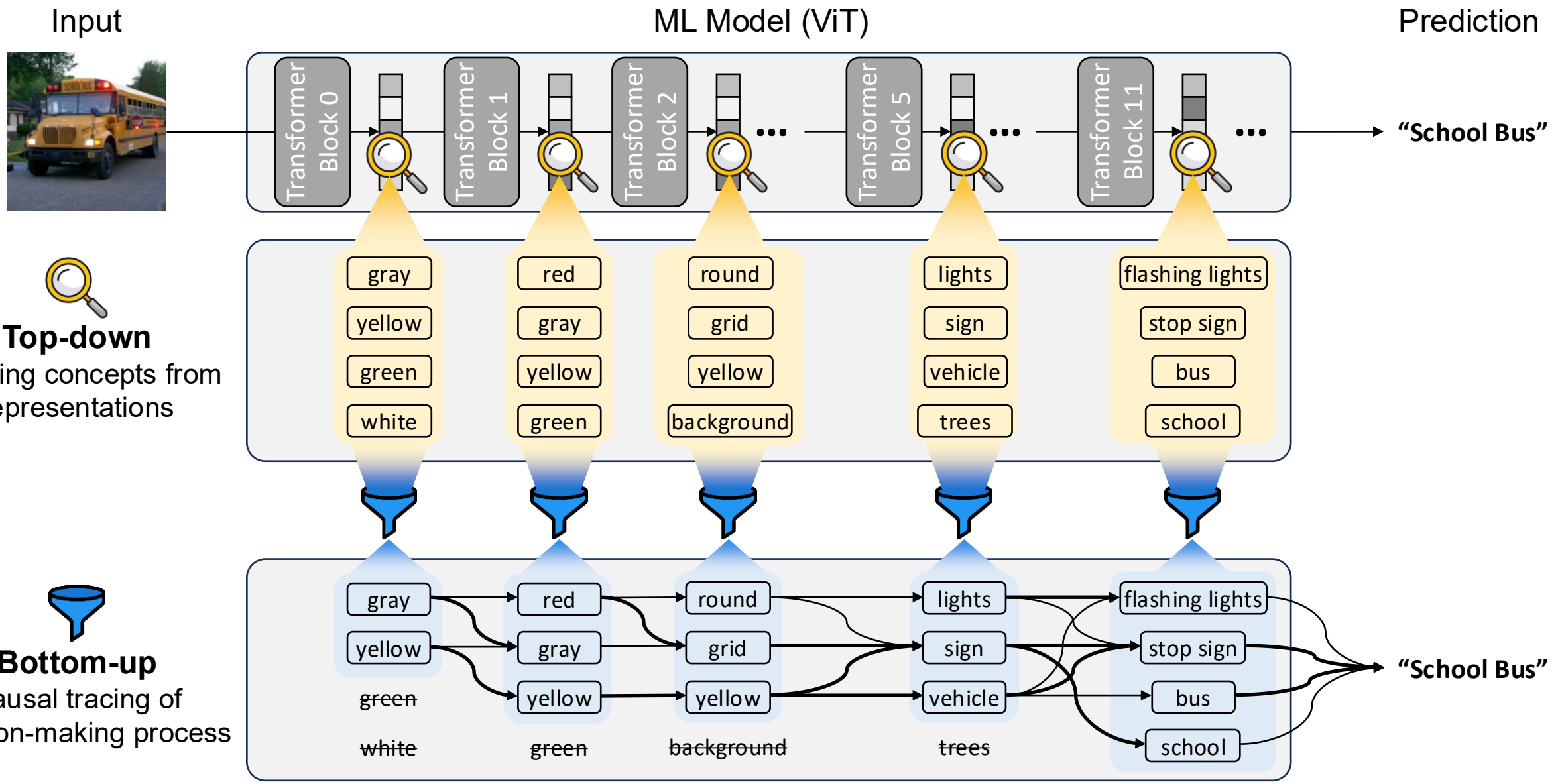
[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Concept Circuit Tracing Algorithm



[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Concept Circuit Tracing Algorithm

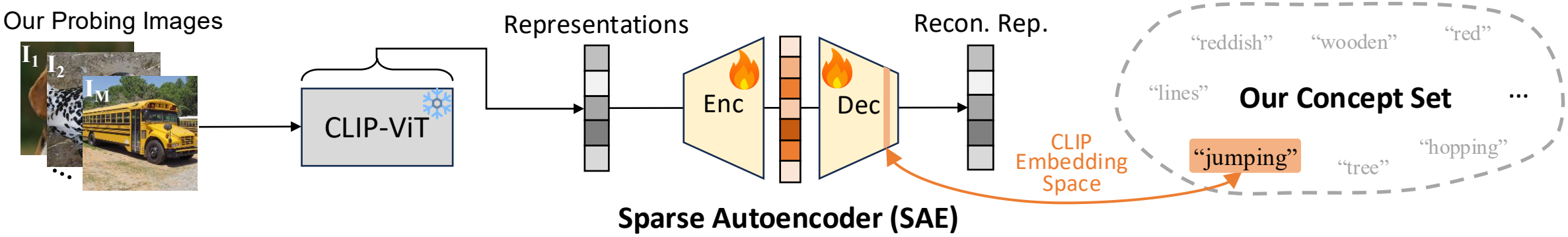


[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

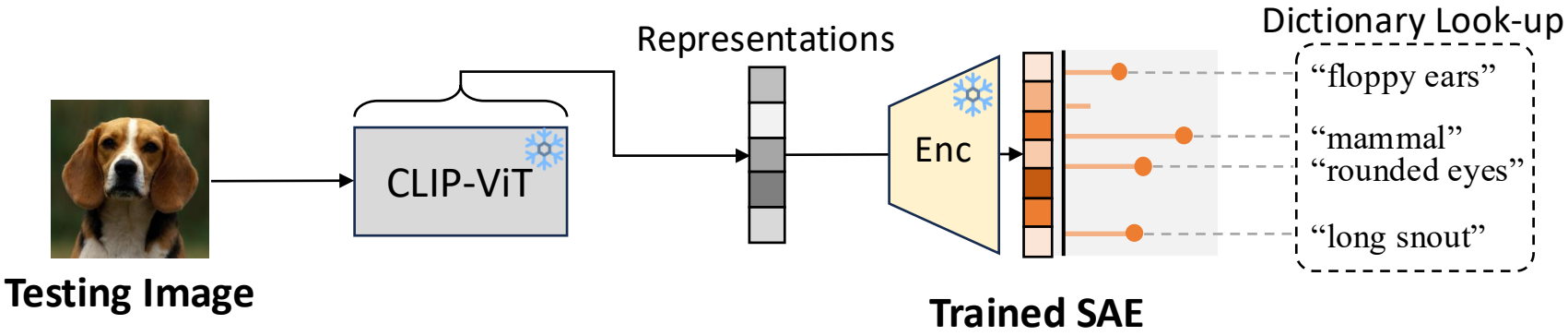
Top-down Concept Reading

➢ **Step 1:** Train SAE with the probing image set.

➢ **Step 2:** Map SAE features to concepts.



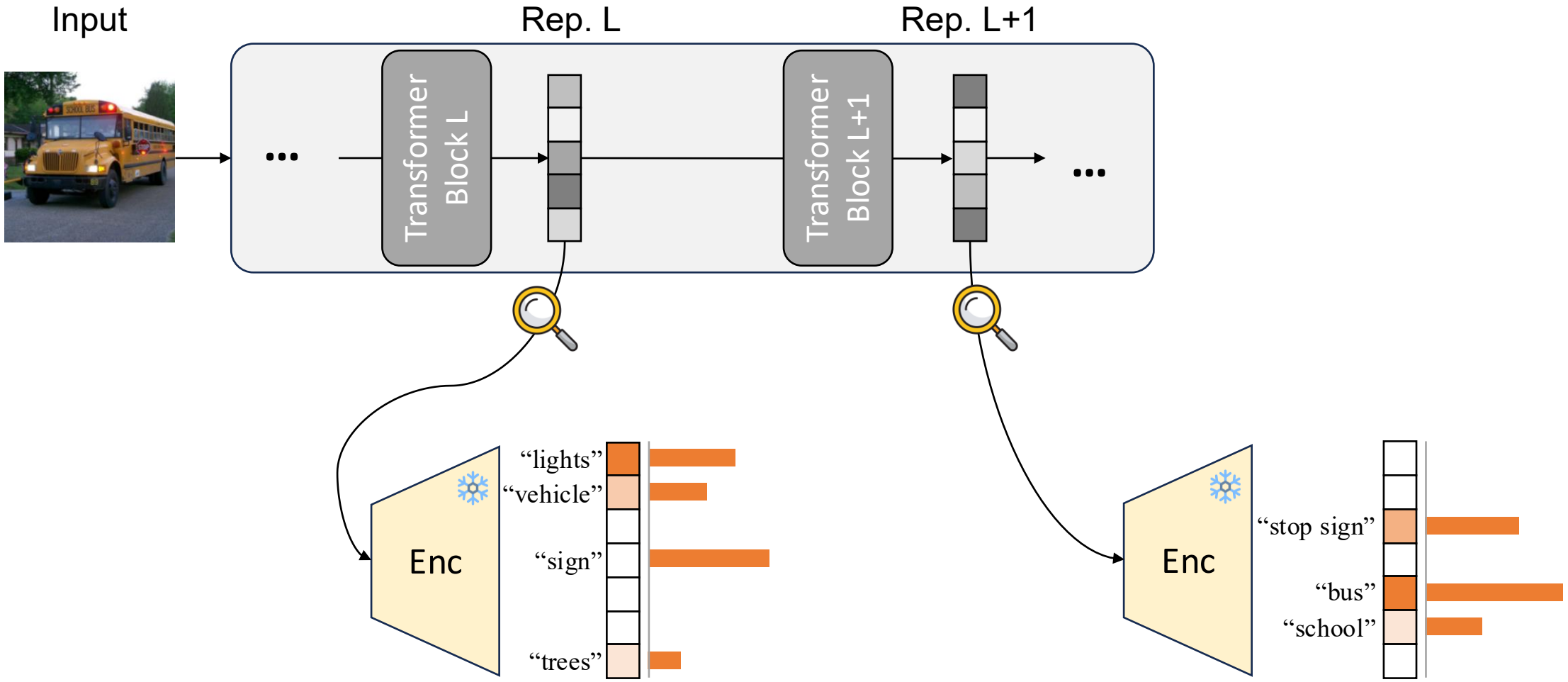
➢ **Step 3:** Read concepts from representations.



[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Bottom-up Circuit Tracing

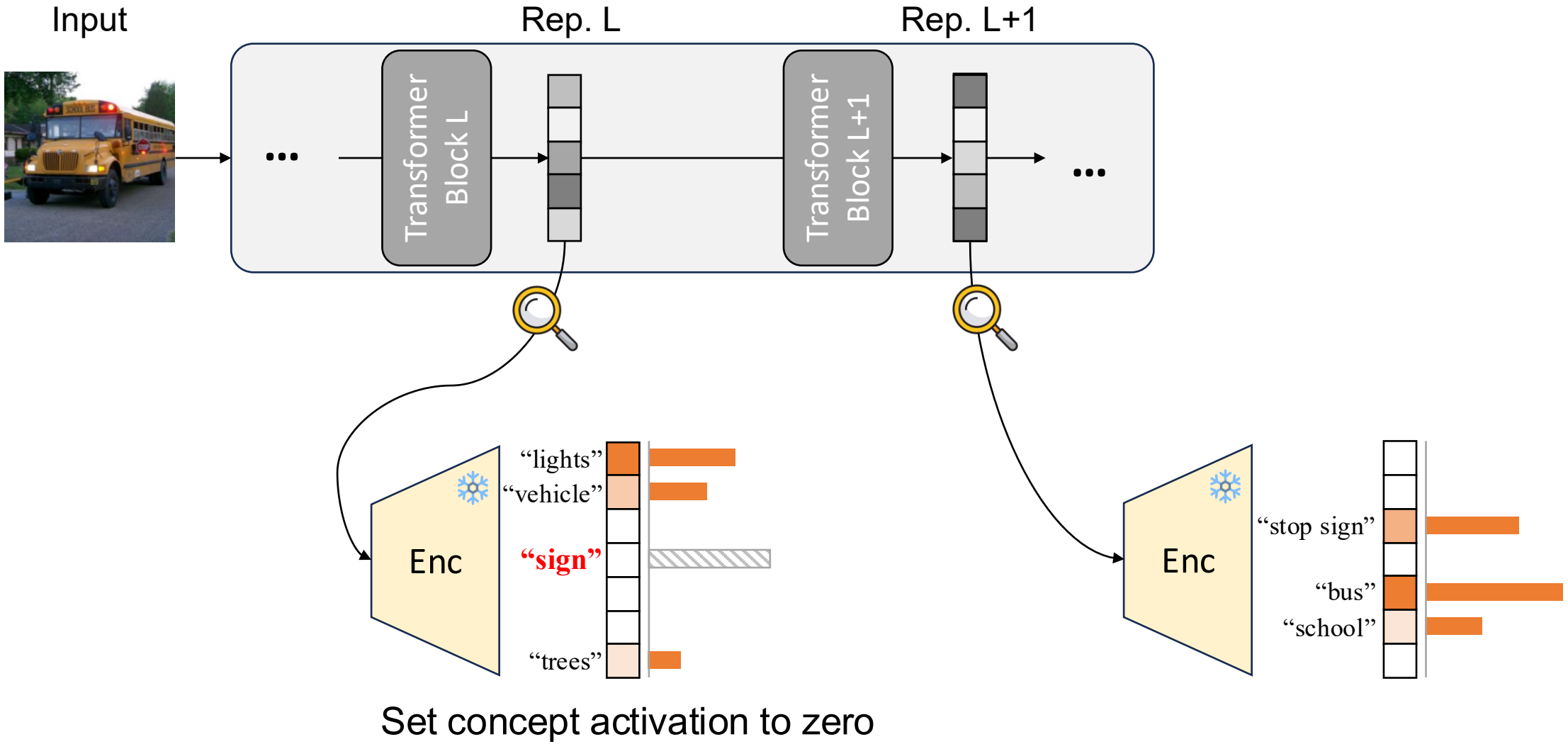
$$\mathrm{IE}_{i\to j}^{s\to t} = \mathrm{IE}\big(\alpha_j^t; c_i^s; r_{\text{clean}}, r_{\text{patch}}\big)$$
$$= \alpha_j^t(r_{\text{clean}}) - \alpha_j^t(r_{\text{clean}} \mid \mathrm{do}(\alpha_i^s = \alpha_i^s(r_{\text{patch}}))).$$



[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Bottom-up Circuit Tracing

$$\mathrm{IE}_{i\to j}^{s\to t} = \mathrm{IE}\big(\alpha_j^t; c_i^s; r_{\text{clean}}, r_{\text{patch}}\big)$$
$$= \alpha_j^t(r_{\text{clean}}) - \alpha_j^t(r_{\text{clean}} \mid \mathrm{do}(\alpha_i^s = \alpha_i^s(r_{\text{patch}}))).$$

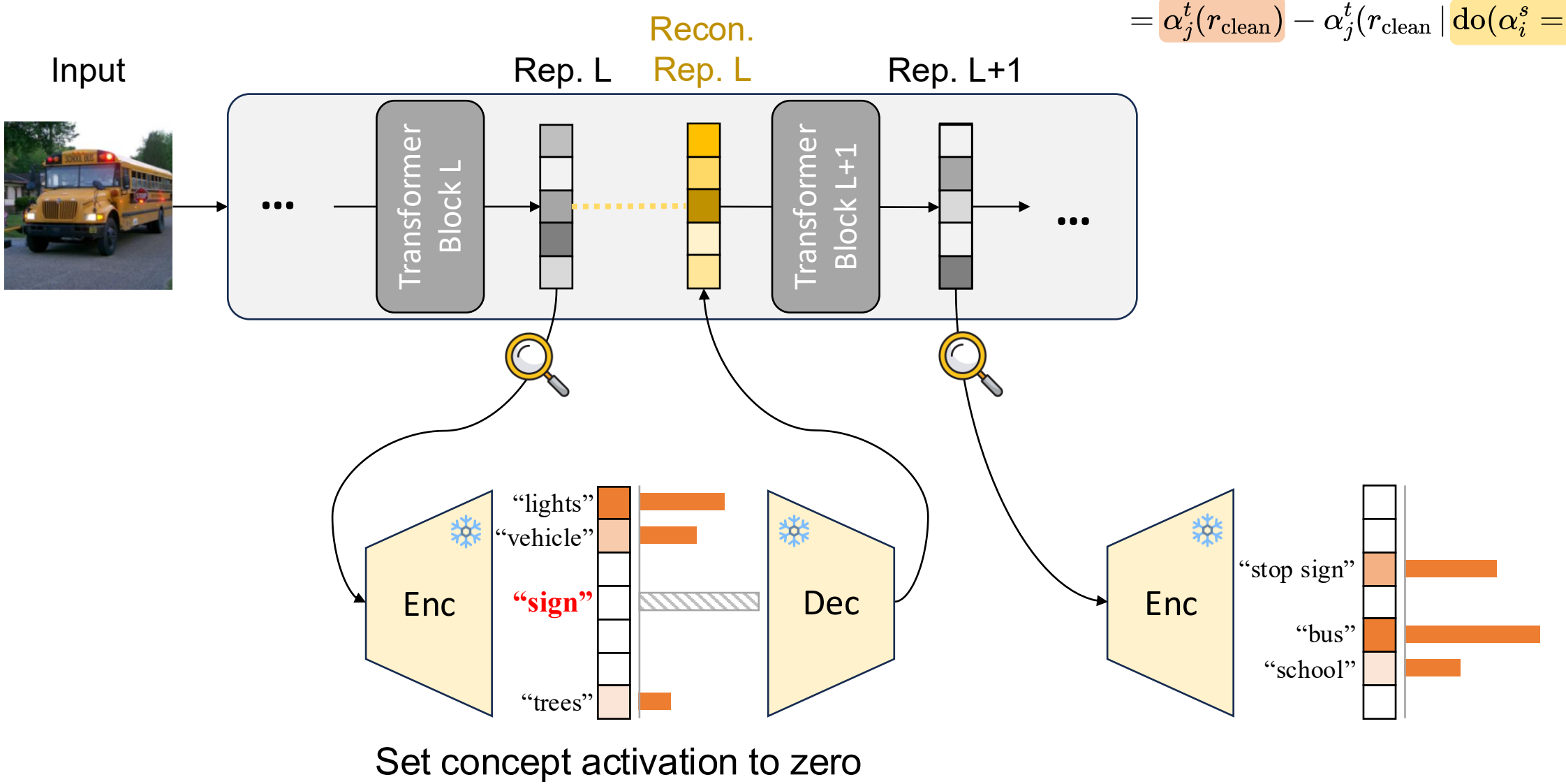


Set concept activation to zero

[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Bottom-up Circuit Tracing

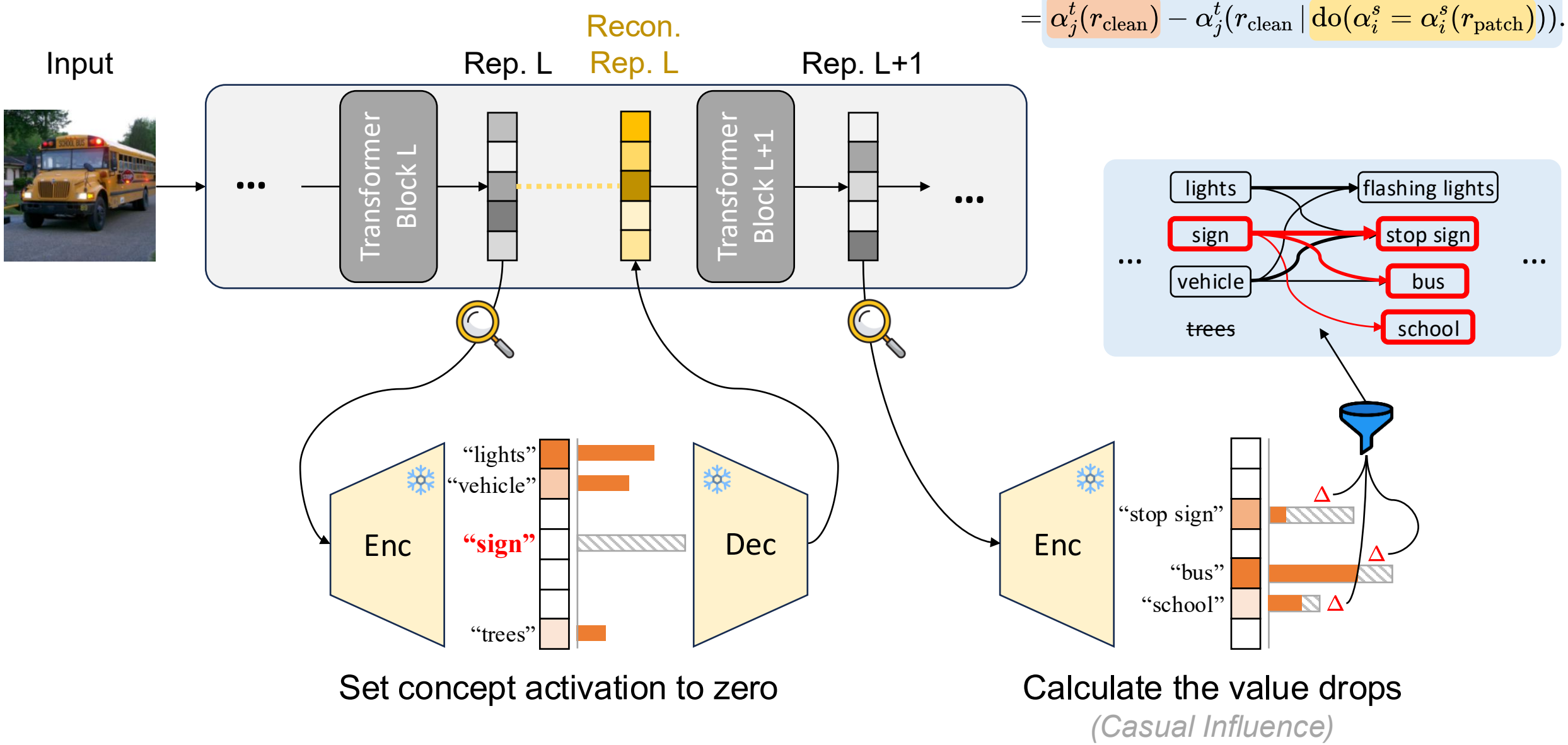
$$\mathrm{IE}_{i\to j}^{s\to t} = \mathrm{IE}\big(\alpha_j^t; c_i^s; r_{\text{clean}}, r_{\text{patch}}\big)$$
$$= \alpha_j^t(r_{\text{clean}}) - \alpha_j^t(r_{\text{clean}} \mid \mathrm{do}(\alpha_i^s = \alpha_i^s(r_{\text{patch}}))).$$



[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Bottom-up Circuit Tracing

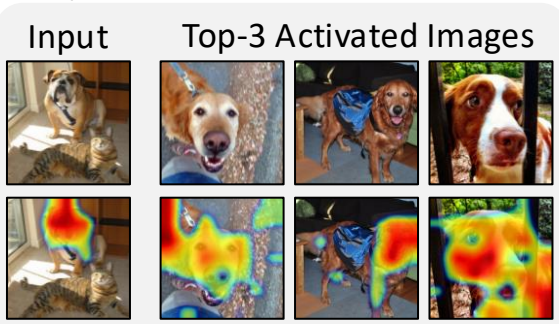
$$\mathrm{IE}_{i\to j}^{s\to t} = \mathrm{IE}\left(\alpha_j^t; c_i^s; r_{\text{clean}}, r_{\text{patch}}\right)$$
$$= \alpha_j^t(r_{\text{clean}}) - \alpha_j^t(r_{\text{clean}} \mid \mathrm{do}(\alpha_i^s = \alpha_i^s(r_{\text{patch}}))).$$



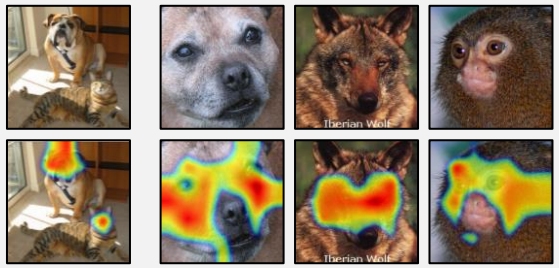
[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

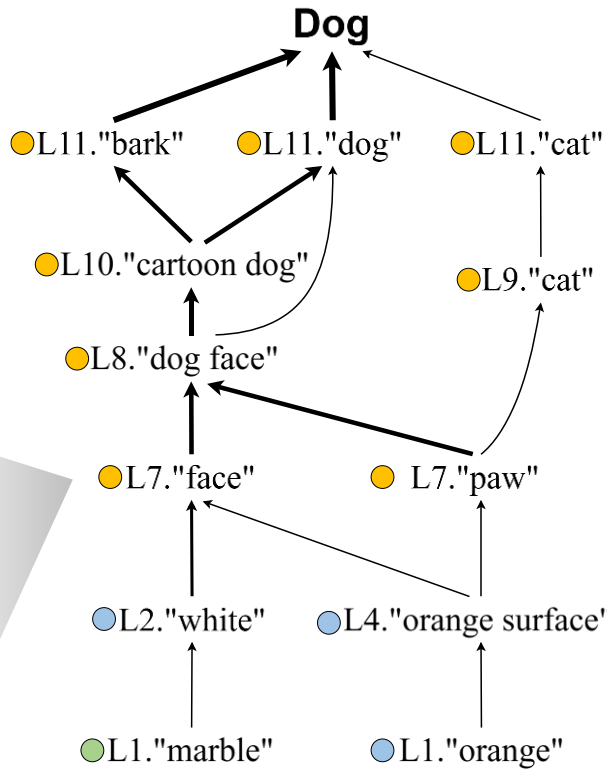
Audit inner workings

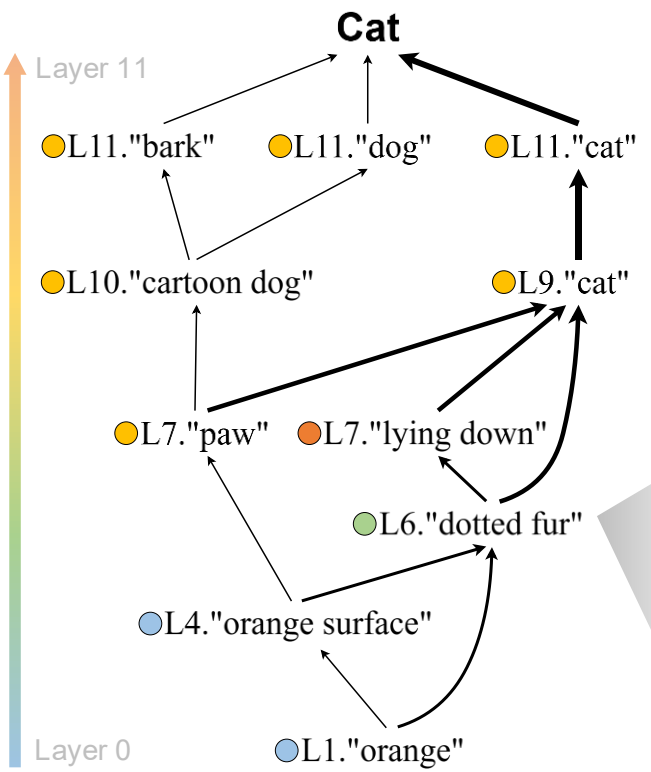
Layer 11, CLS token, #3167: **"bark"**



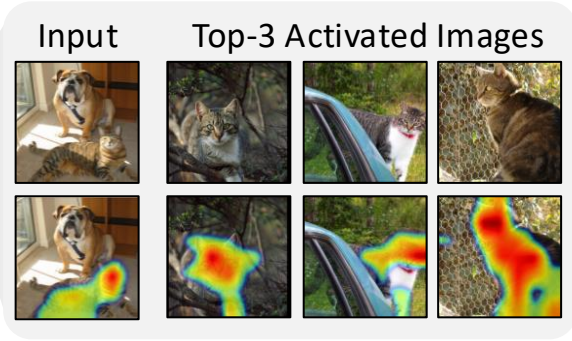
Layer 7, IMG token, #4114: **"face"**



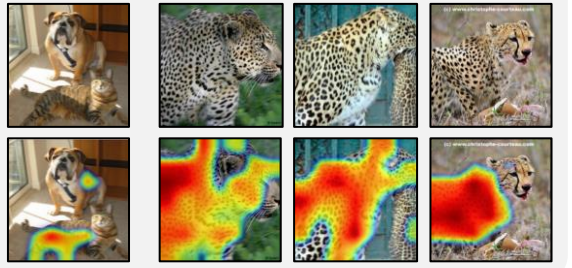




Layer 11, CLS token, #1354: **"cat"**



Layer 6, IMG token, #4375: **"dotted fur"**



Our method captures the nuanced difference of model in making predictions.

[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Locate concepts on pixels

Primitive-Level

Concept 2196 – “yellow”

Original

Localization

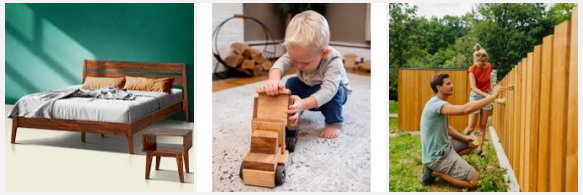
Concept 1747 – “lines”

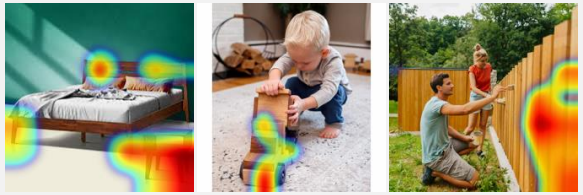
Original

Localization

Intermediate-Level

Concept 1067 – “wooden”



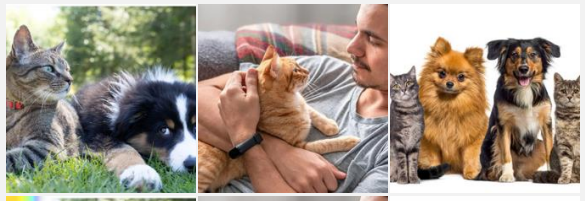


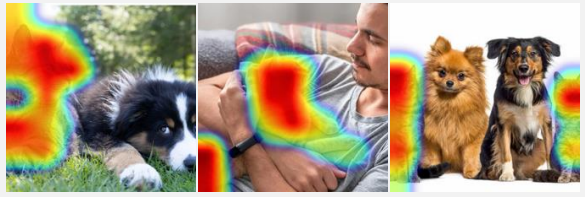
Concept 1875 – “circle”

Object-Level

Concept 2924 – “wings”

Concept 139 – “cat”





Scene-Level

Concept 1274 – “looking at”

Concept 2554 – “jumping”

Our method localize concepts on pixels, even for highly abstract ones.

[ICML 2026] *“Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers”* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Analyze failure modes

Failure Images
38 classes, 1.9K images

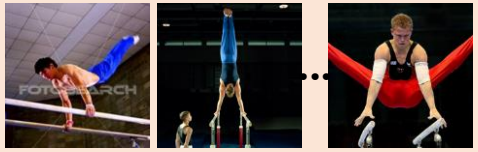
Right vs. Wrong
Difference in SAE Concepts

Failure Modes
Statistics (%)

Label: canoe
Pred: paddle

→ ["**holding**", "**people**", "mountain", "water", "shade", …]

Label: parallel bar
Pred: high bar (single)



→ ["gymnast", "**flying**", "bar", "**downward**", "indoor", …]

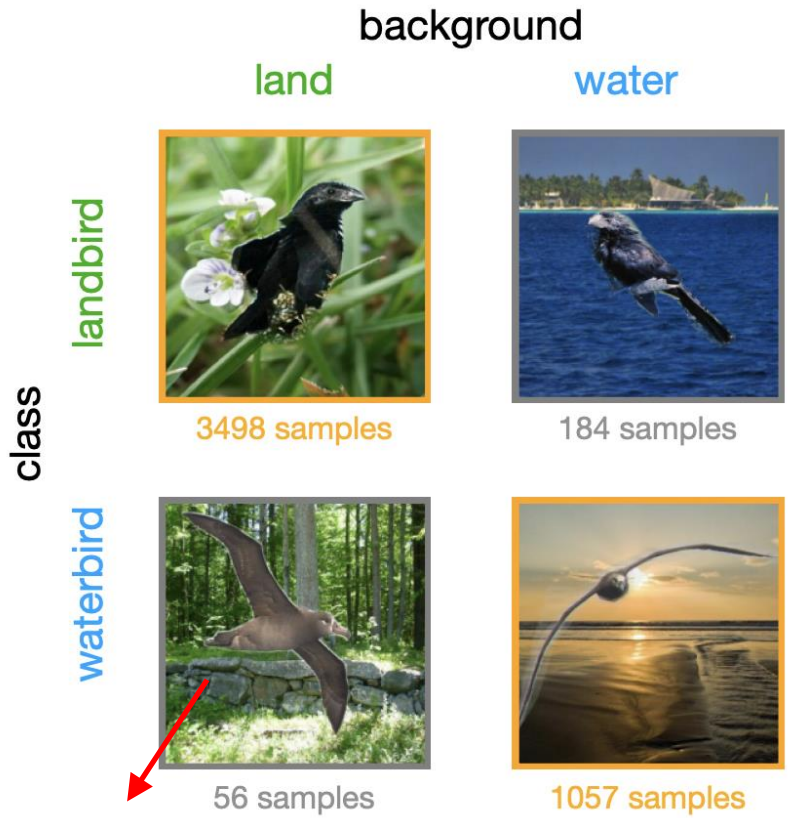
Our method can efficiently analyze the reasons behind failures.

[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Steer model prediction by editing concepts

WaterBird Dataset

Bird categories spuriously correlated to backgrounds.



[Sagawa'19]

Worst Group: waterbirds on land backgrounds.

Label: **waterbird**

SAE Top-4 Concepts →

"bird": 7.09
"water": 3.90
"tree": 2.53
"rock": 2.46

Pred: **landbird**
Confidence: **54%**

Remove Land Concepts

"bird": 7.09
"water": 3.90
"tree": **0.00**
"rock": **0.00**

Pred: **waterbird**
Confidence: **89%**

Enhance Land Concepts

"bird": 7.09
"water": 3.90
"tree": **5.00**
"rock": **5.00**

Pred: **landbird**
Confidence: **94%**

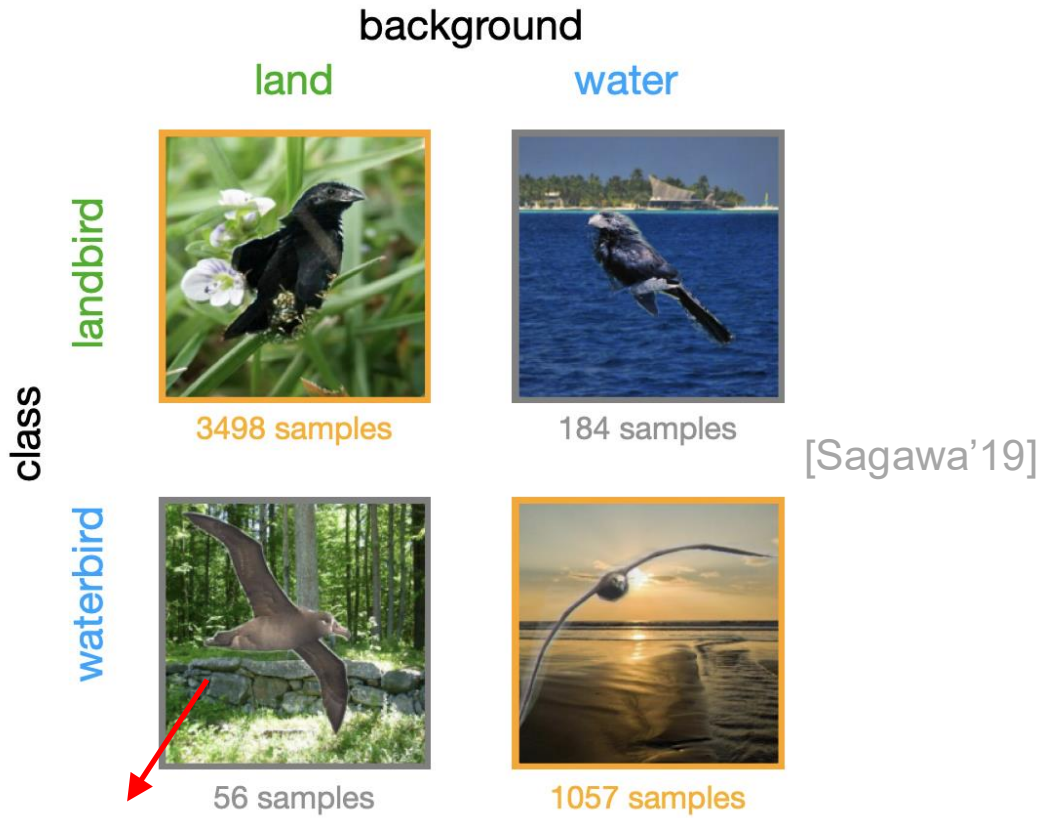
Our SAE works as precise control "knobs" of concepts.

[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

Steer model prediction by editing concepts

WaterBird Dataset

Bird categories spuriously correlated to backgrounds.



[Sagawa'19]

Worst Group: waterbirds on land backgrounds.

Comparison of Steering Power

Method	Steer Spuri. Corr.	Overall Acc. (%)	Worst Group Acc. (%)	Δ
CBM	None	-	37.3	-
	Remove	-	51.8	+ 14.5
SpLiCE	None	-	48.0	-
	Remove	-	60.0	+ 12.0
DN-CBM	None	-	57.5	-
	Remove	-	71.3	+ 13.8
PCBM	None	-	50.3	-
	Remove	-	74.7	+ 24.4
Ours	None	79.7	50.3	-
	Enhance	74.5	5.3	− 45.0
	Remove	85.2	**98.5**	+ **48.2**

Our SAE works as precise control "knobs" of concepts.

[ICML 2026] *"Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers"* Tang Li, Yanlin Chen, Mengmeng Ma, and Xi Peng.

DeepREAL

ICML
International Conference
On Machine Learning

Thank You!

Inside the Visual Mind: Neuroscience-Motivated Concept Circuits for Interpreting and Steering Vision Transformers

Tang Li Yanlin Chen Mengmeng Ma Xi Peng

Deep Robust & Explainable AI Lab (DeepREAL)

Tue, Jul 7, 2026 10:30 AM – 12:15 PM KST Coex: HALL A