

Sparse Autoencoders Reveal Selective Remapping of Visual Concepts During Adaptation

ICLR 2025 (Paper 14240)

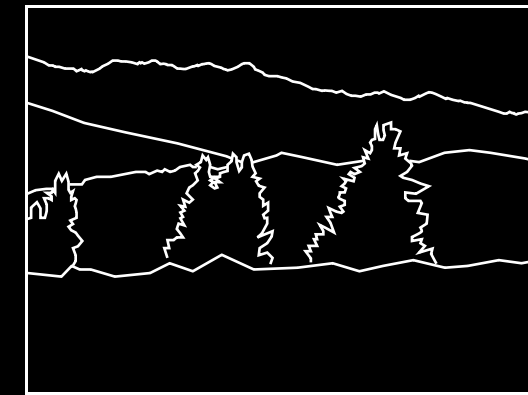
Hyesu Lim^{1,2} Jinho Choi² Jaegul Choo² Steffen Schneider^{1,3}



Motivation



color



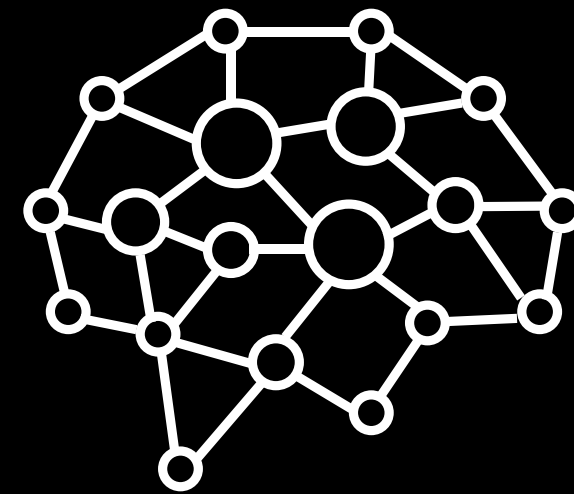
shape



object

Visual concepts

Motivation



CLIP*



Not human readable

* Learning Transferable Visual Models From Natural Language Supervision (Radford et al., ICML 2021)

Our method: PatchSAE



CLIP

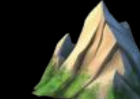
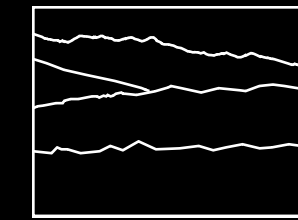
PatchSAE

Interpretability tool
for CLIP vision encoder

Our method: PatchSAE



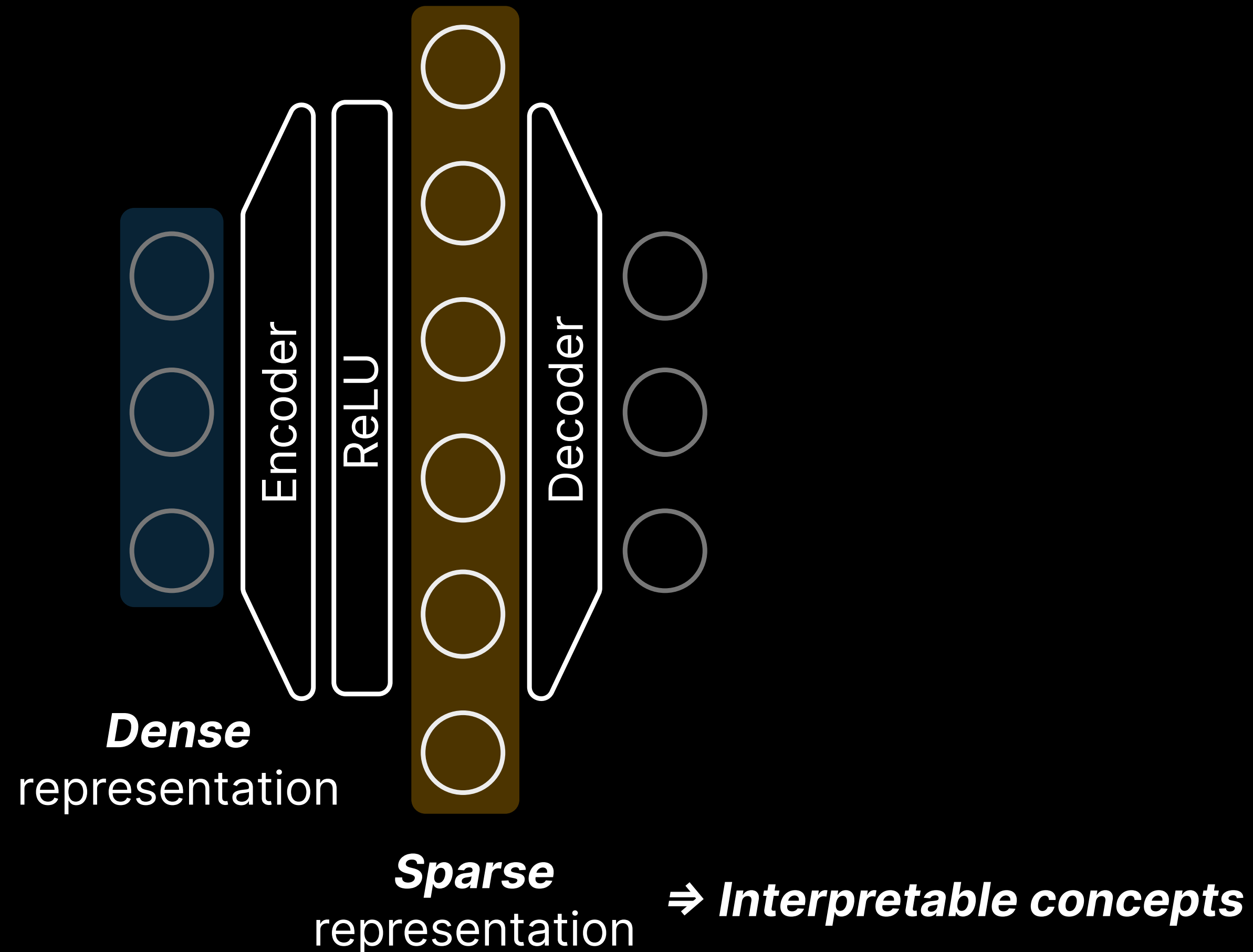
CLIP



Interpretable concepts

Sparse Autoencoder (SAE)*

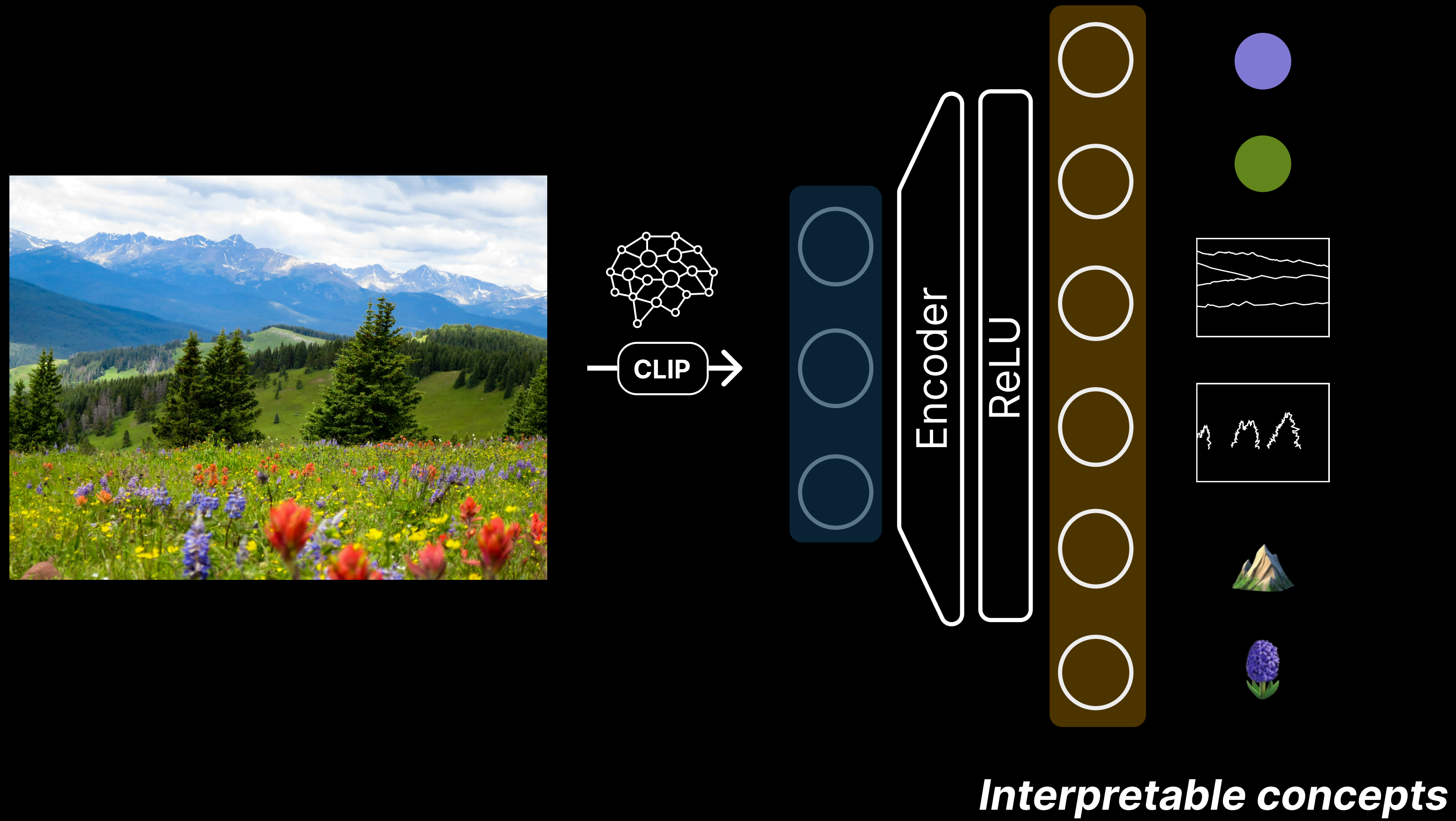
$$\text{SAE}(z) = (g \circ \phi \circ f)(z) = W_D^\top \phi(W_E^\top z)$$



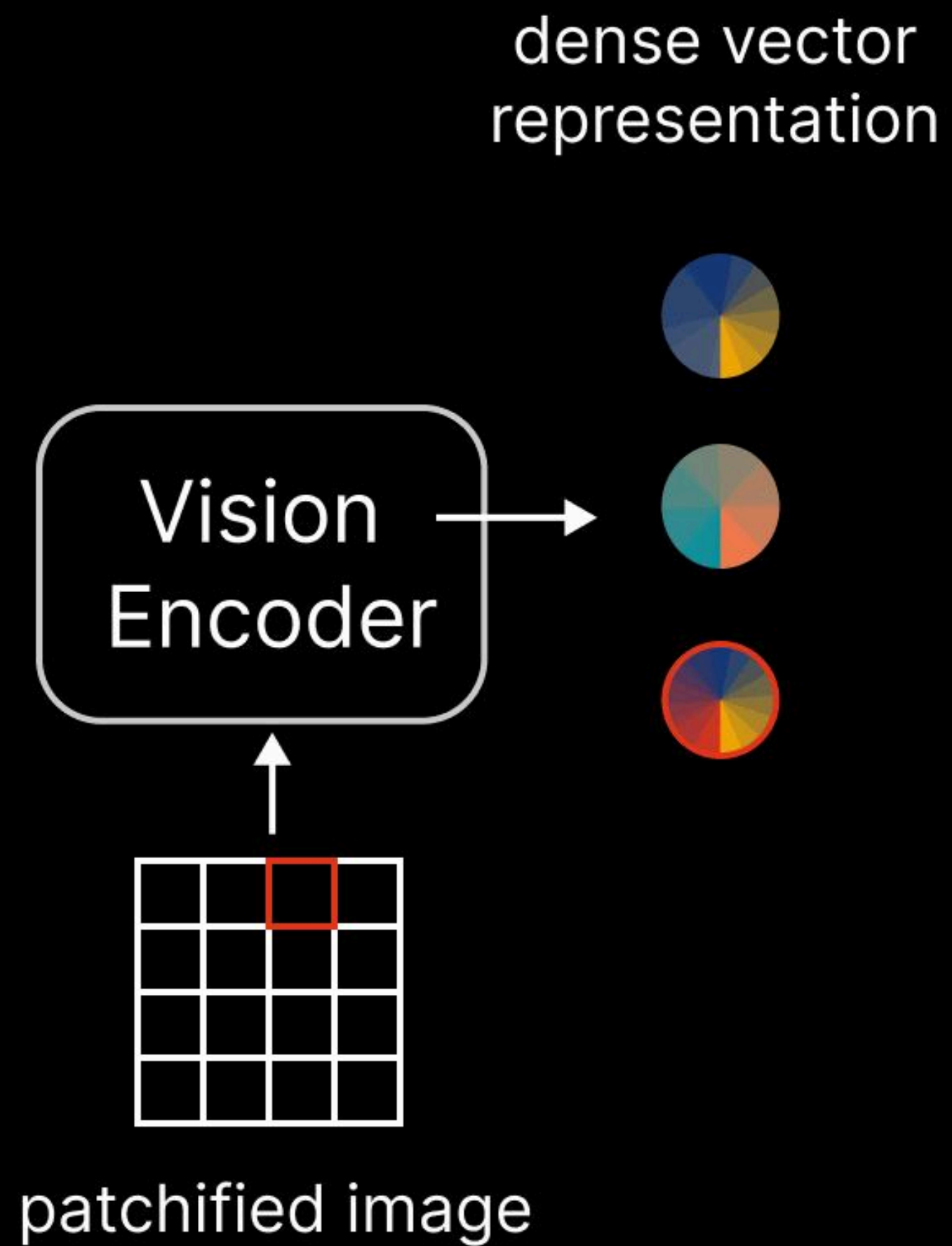
* Towards monosemanticity: Decomposing language models with dictionary learning (Bricken et al., Anthropic 2023)

Sparse Autoencoders Find Highly Interpretable Features in Language Models (Huben et al., ICLR 2024)

Our method: PatchSAE

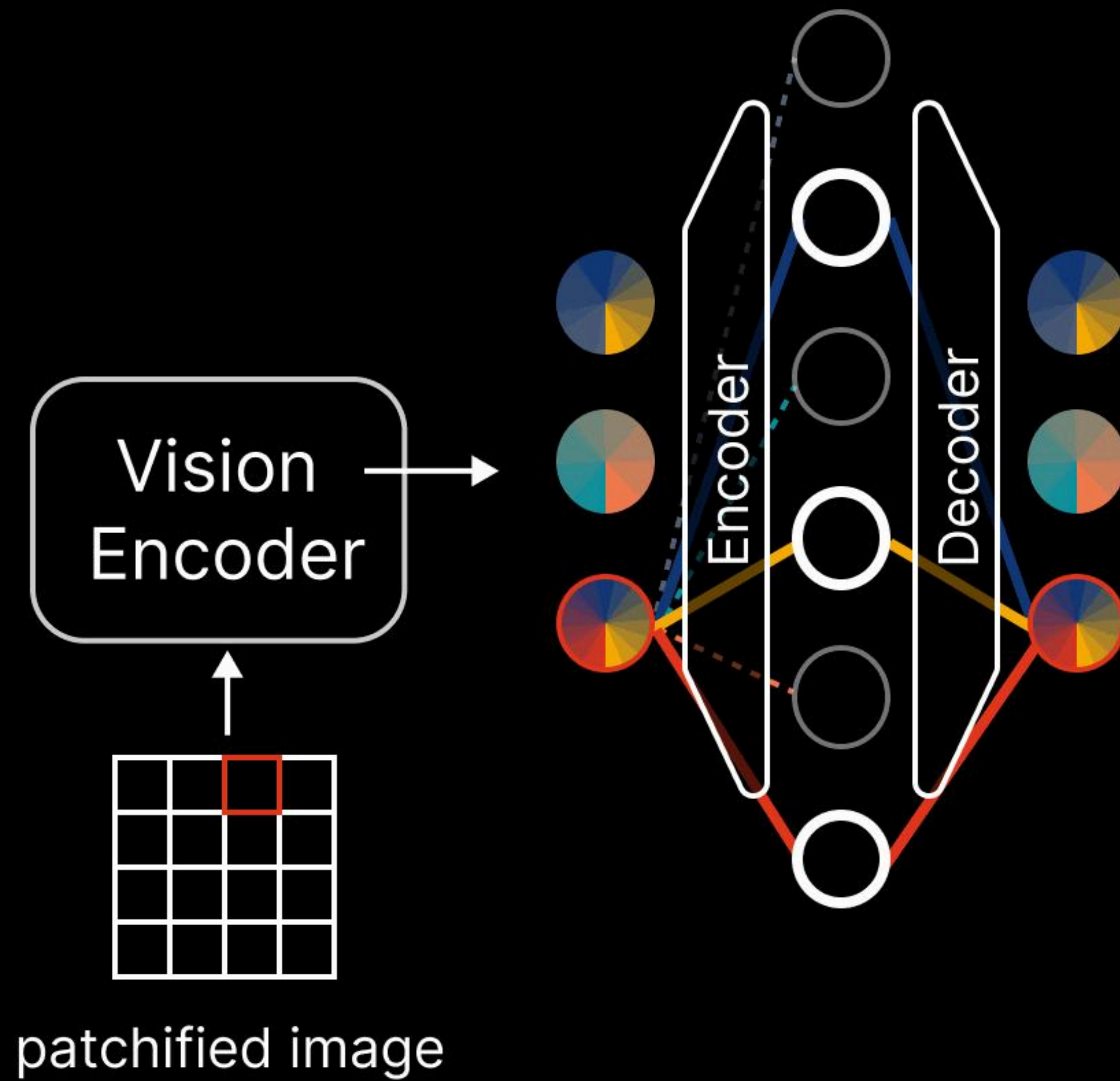


PatchSAE: Discovering concepts from patch-wise image representations



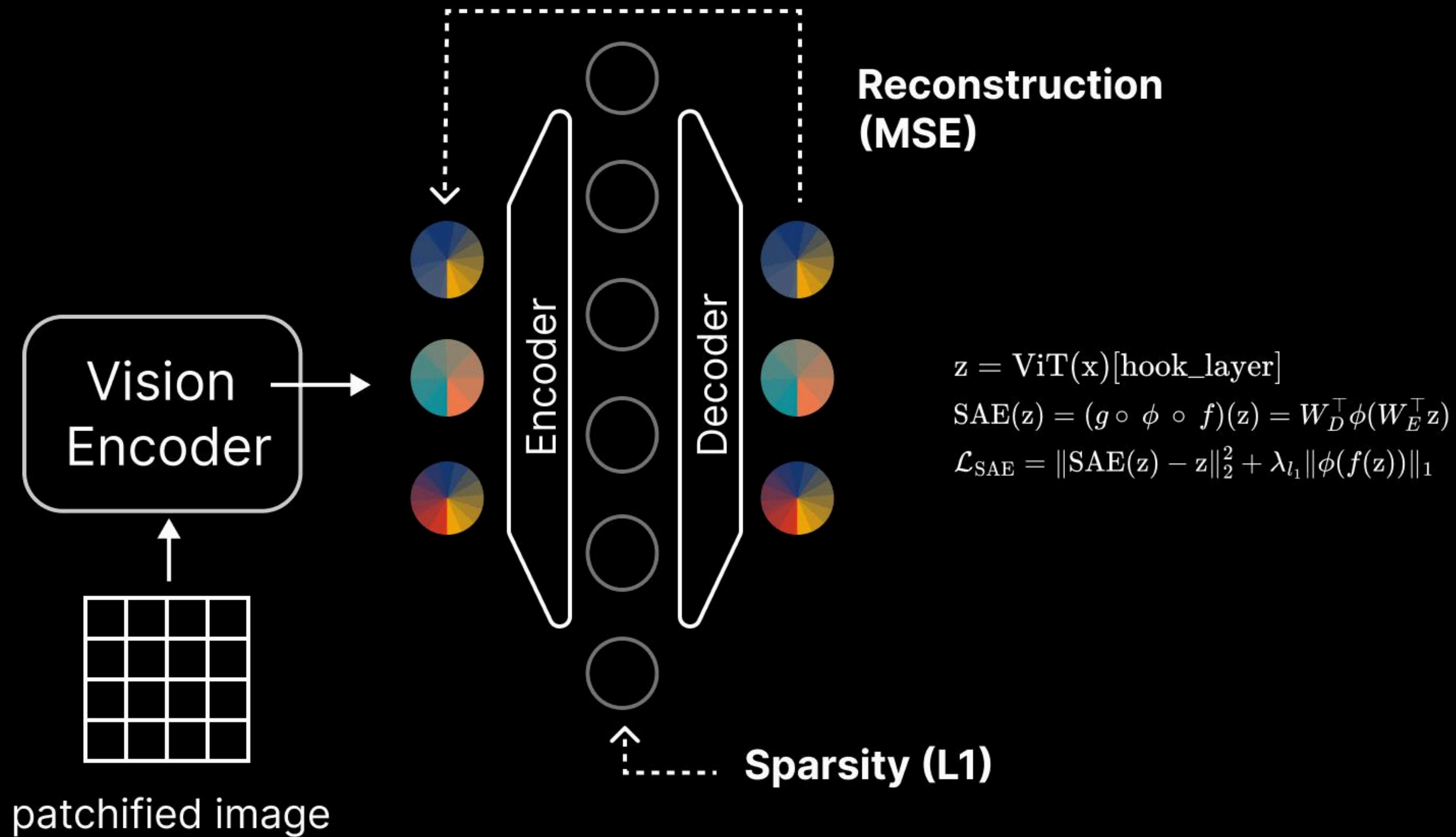
$$z = \text{ViT}(x)[\text{hook_layer}]$$

PatchSAE: Discovering concepts from patch-wise image representations

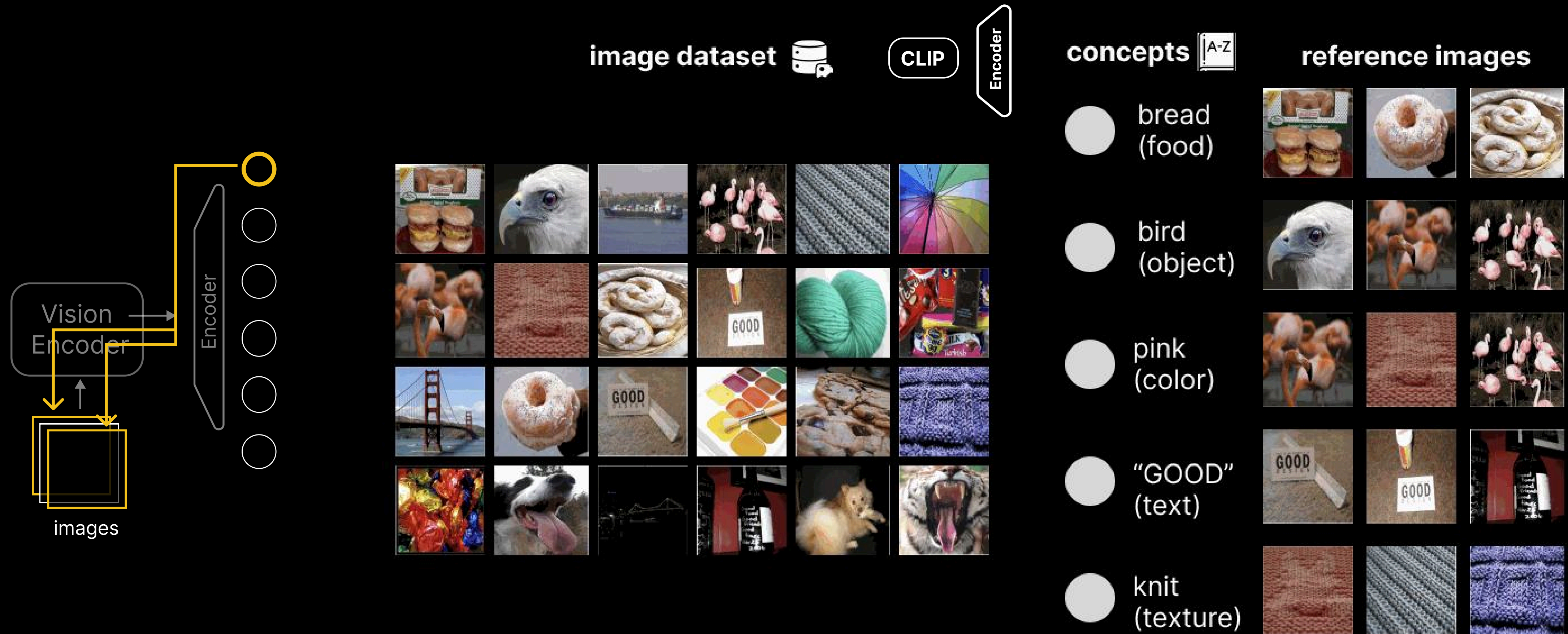


$$z = \text{ViT}(\mathbf{x})[\text{hook_layer}]$$
$$\text{SAE}(z) = (g \circ \phi \circ f)(z) = W_D^\top \phi(W_E^\top z)$$

PatchSAE: Discovering concepts from patch-wise image representations

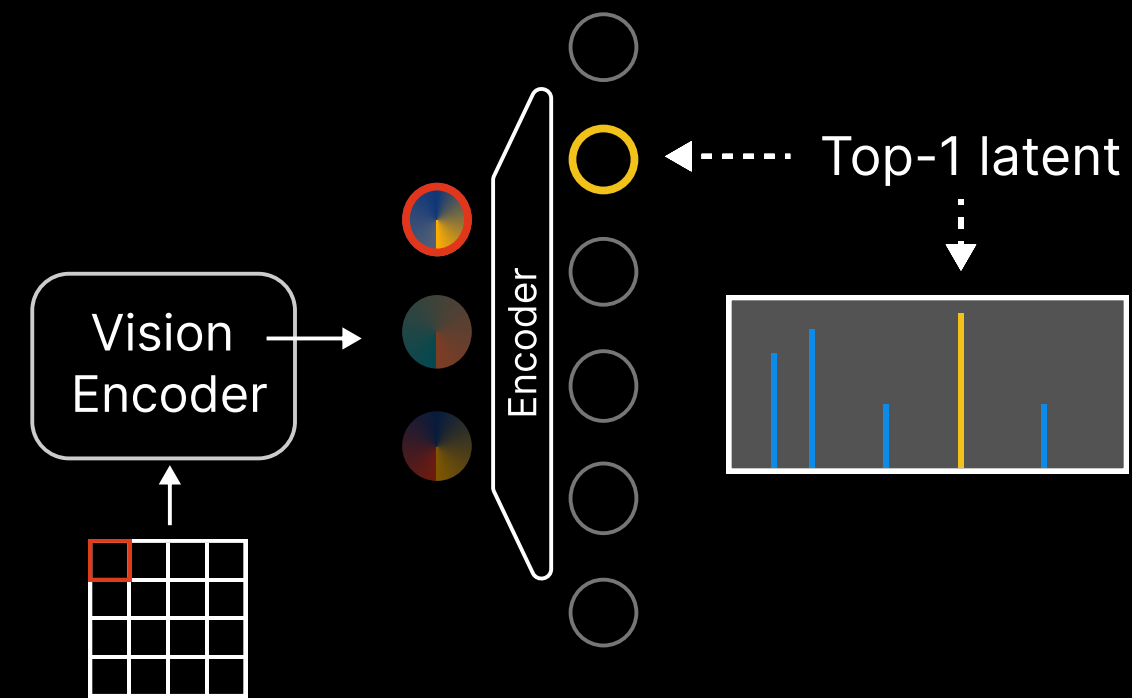


PatchSAE: Discovering concepts from patch-wise image representations



Learned concepts are identified by collecting **maximally activating images** of each latent.

PatchSAE: Discovering concepts from patch-wise image representations



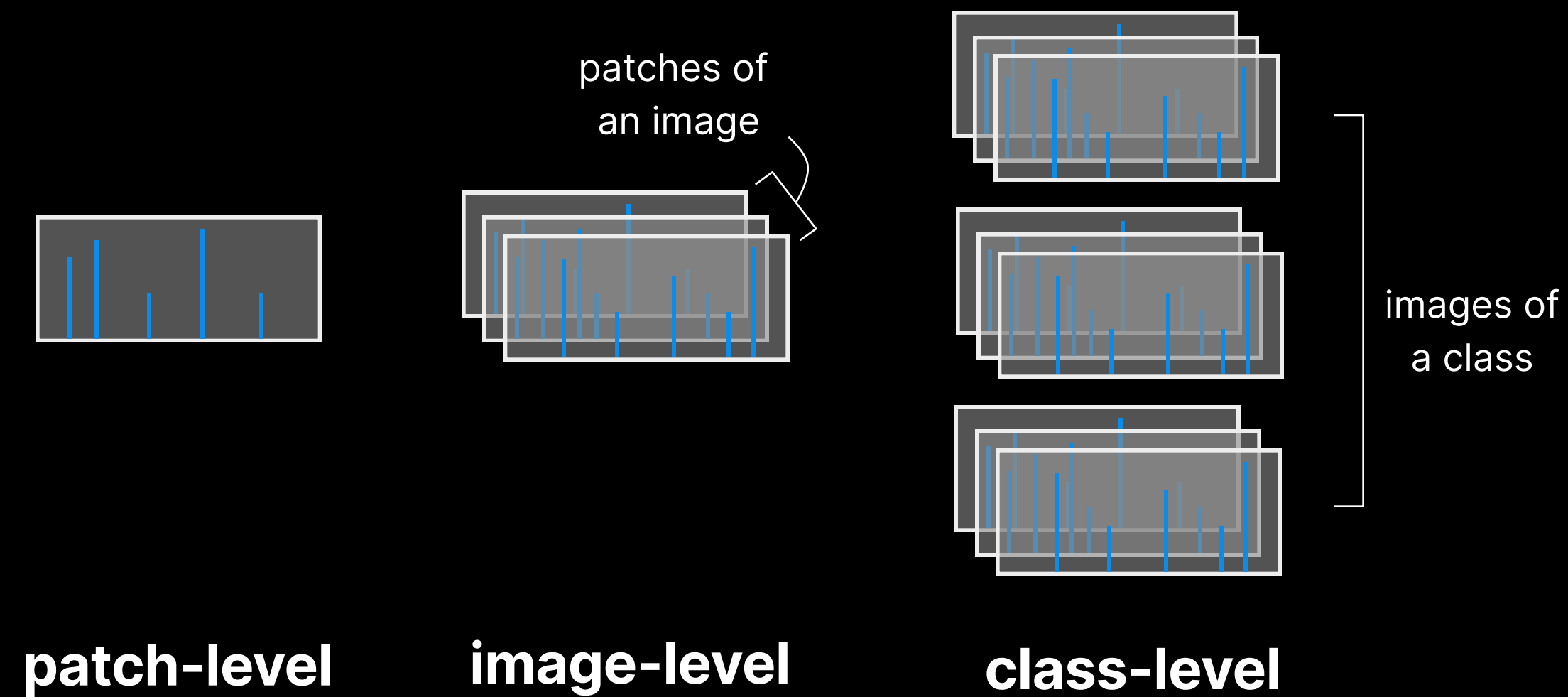
Input



Top-1 latent

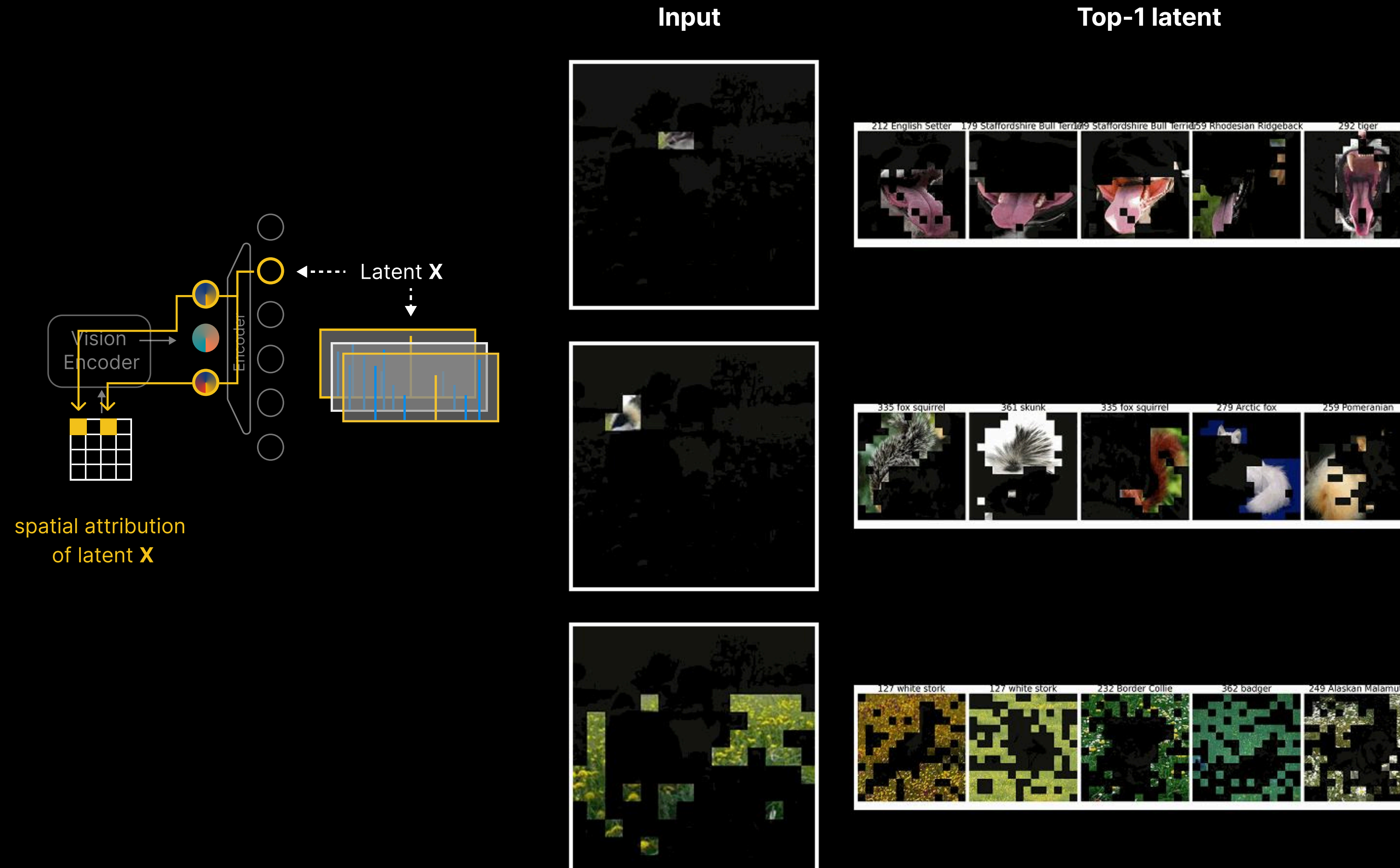


PatchSAE: Discovering concepts from patch-wise image representations



We can easily obtain image-, class-, and larger group-level concepts from patch-level activation through aggregation.

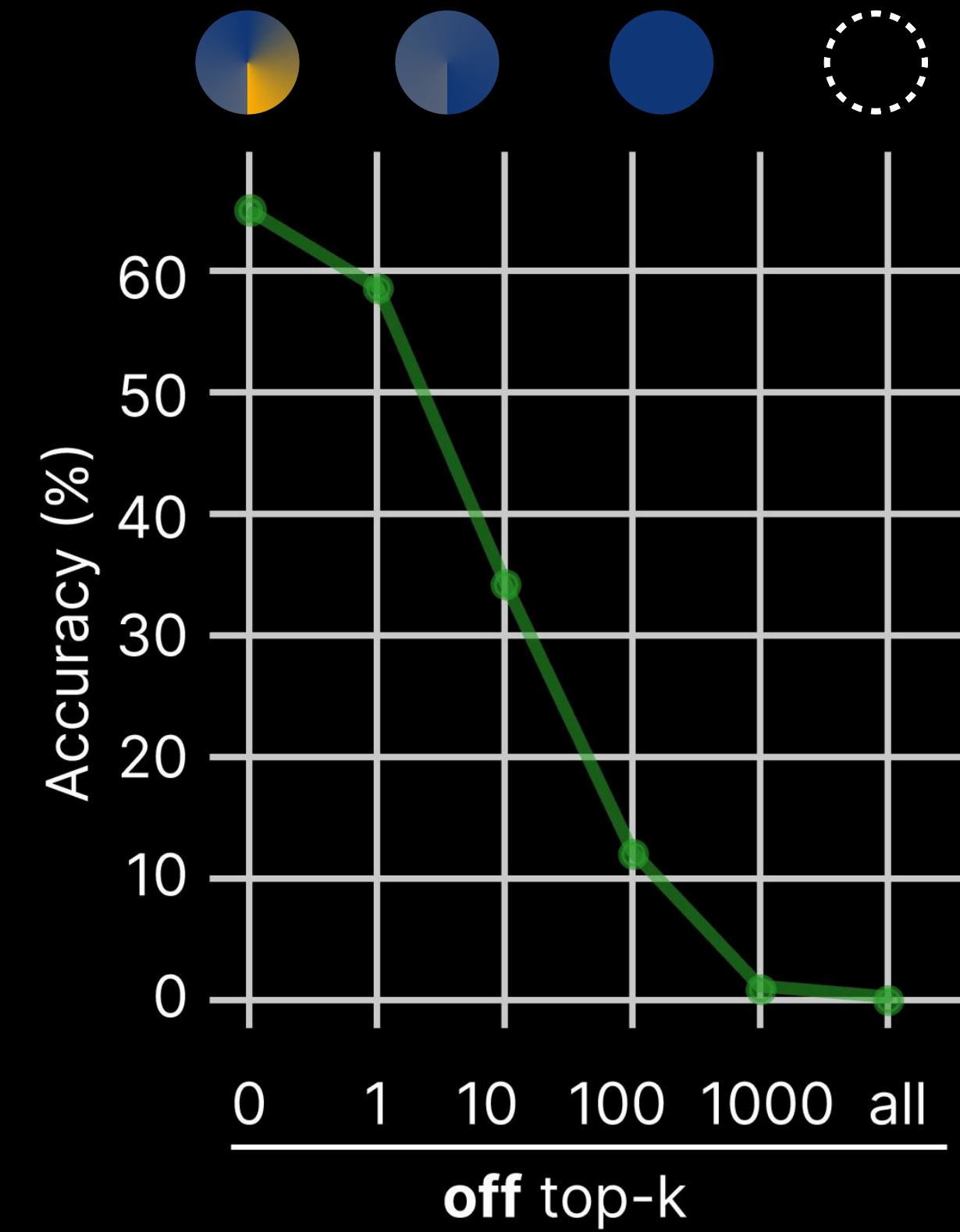
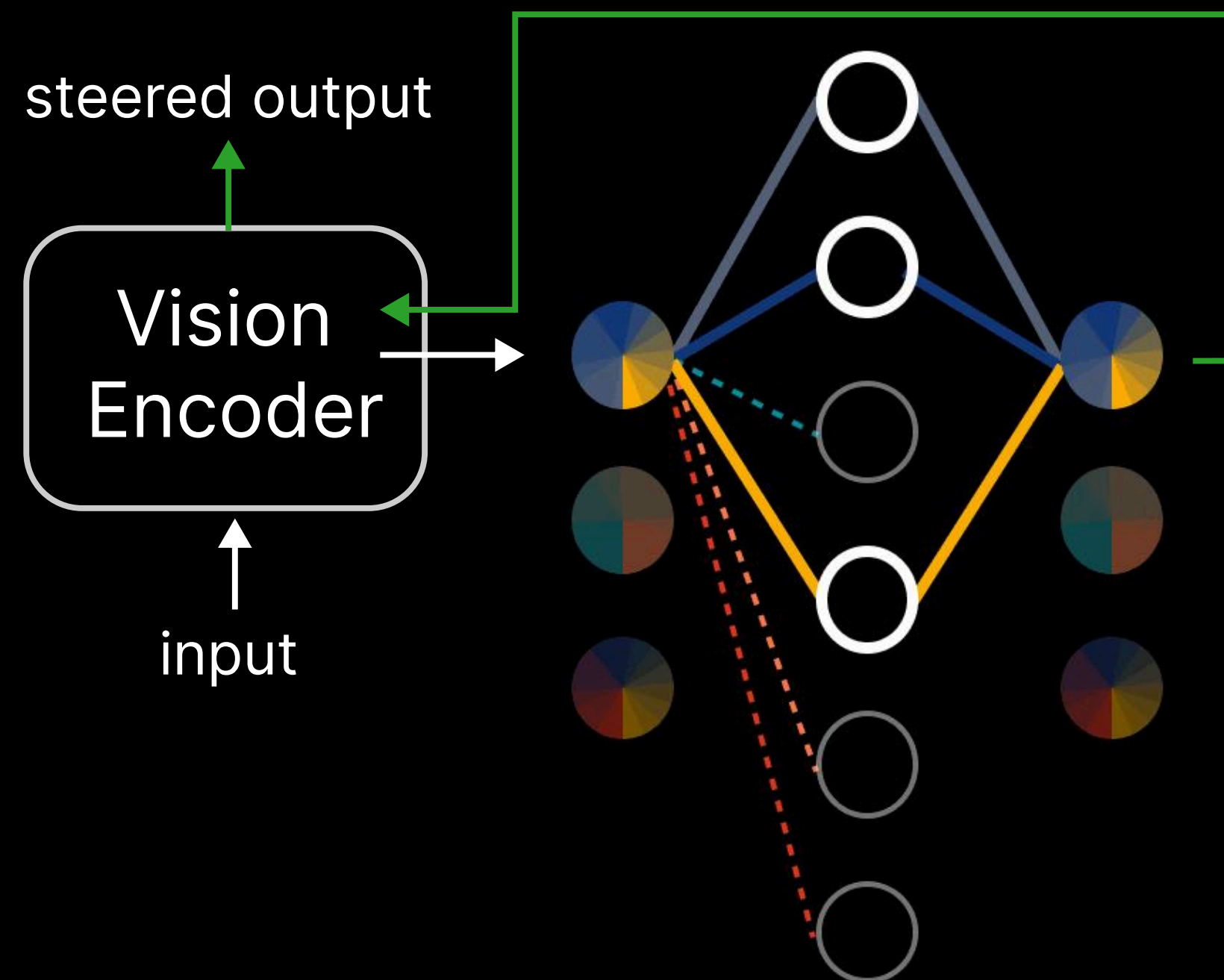
PatchSAE: Discovering concepts from patch-wise image representations



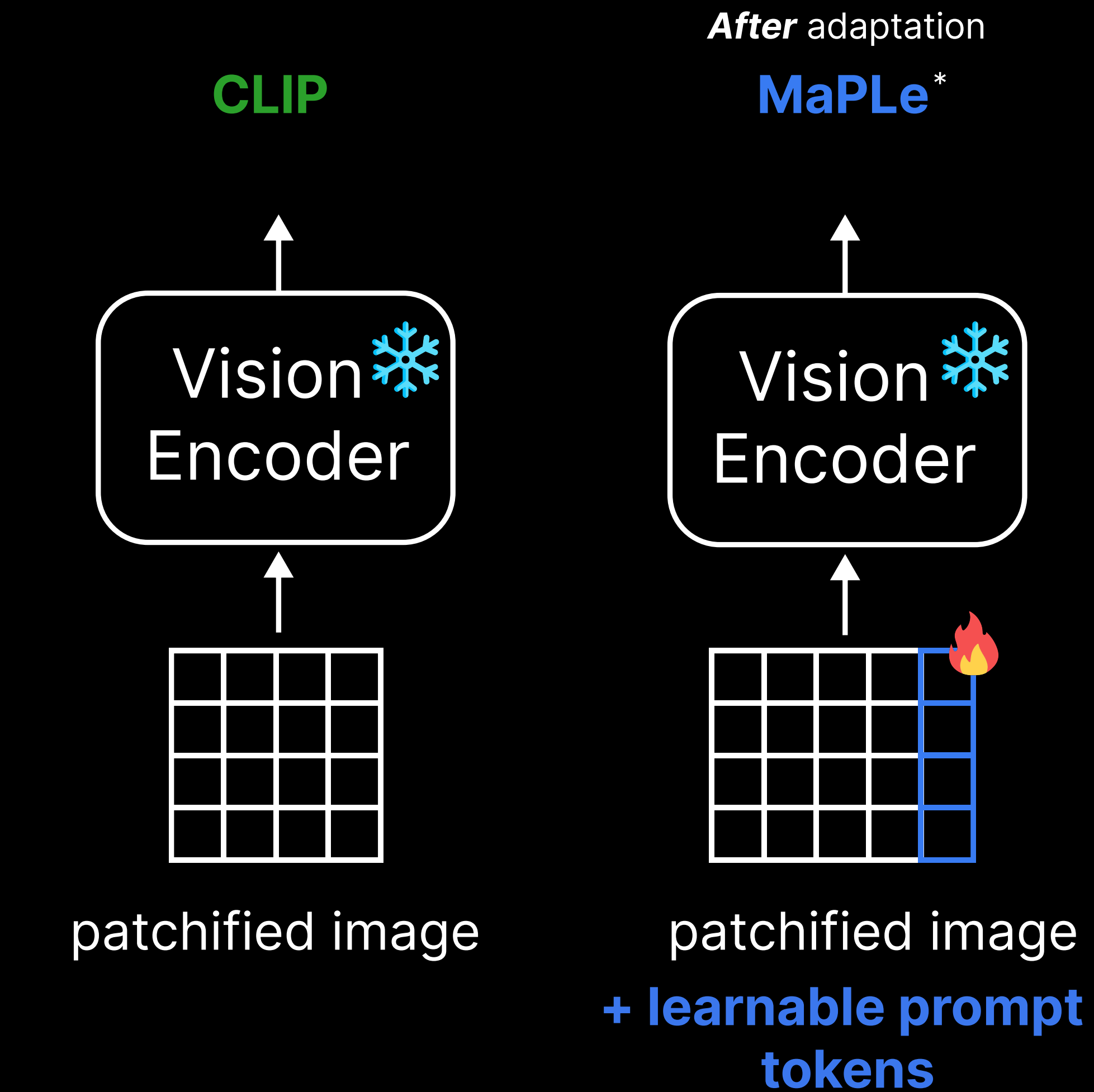
Understanding CLIP behavior via PatchSAE

Observing decision-making process &
Understanding **adaptation mechanisms**

Top-activating SAE concepts and model prediction

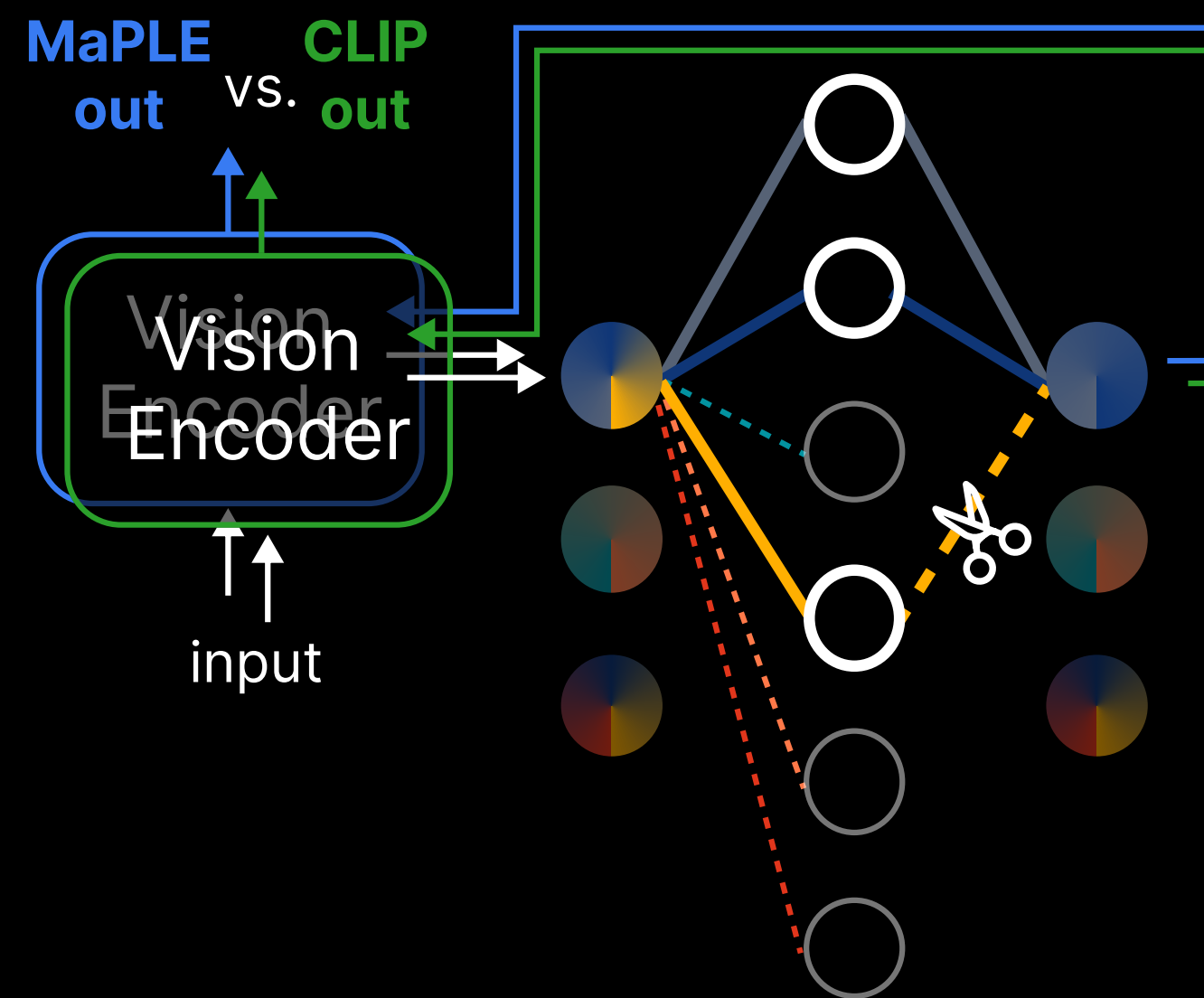


Understanding adaptation mechanisms

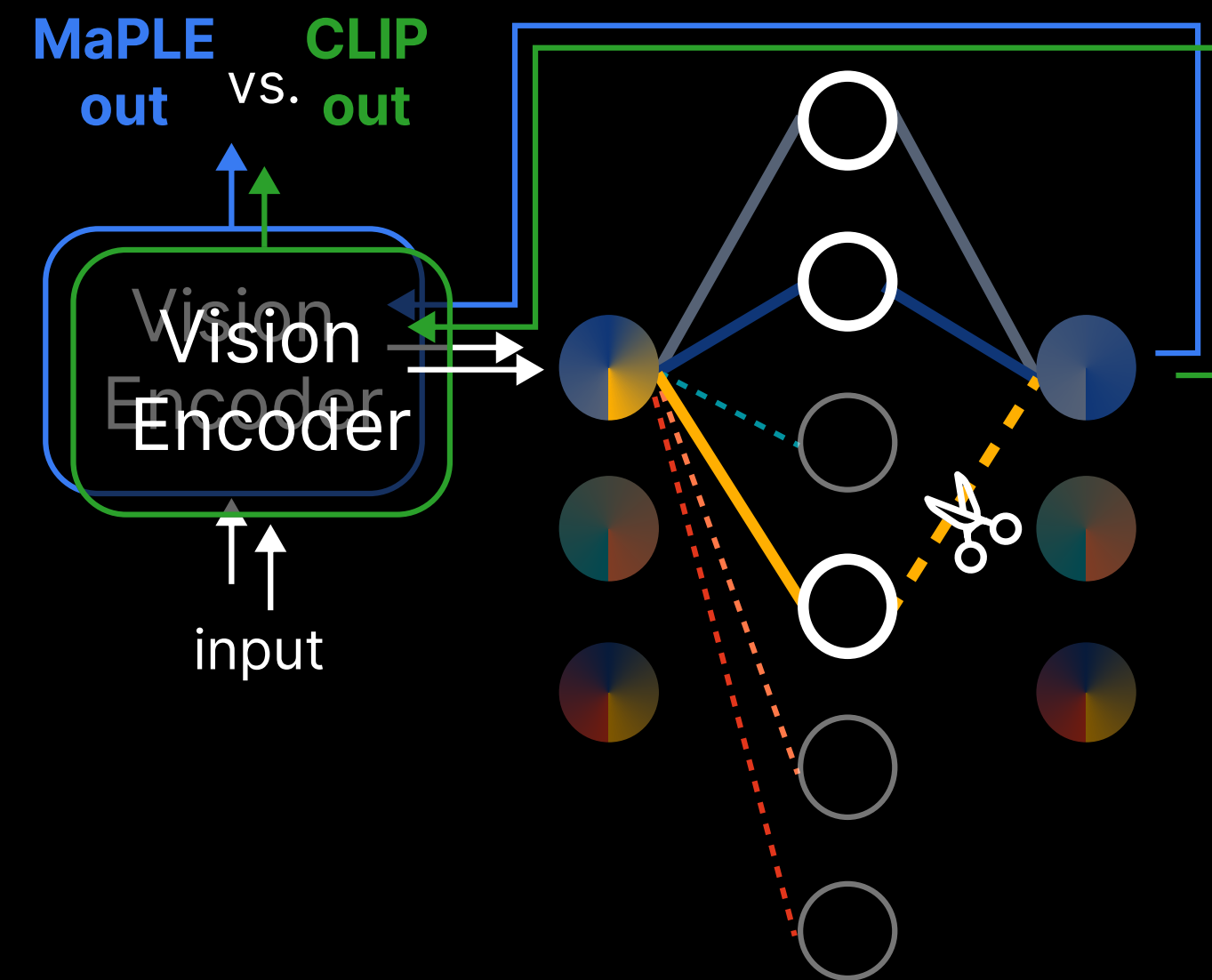


* MaPLe: Multi-modal Prompt Learning (Khattak et al., CVPR 2023)

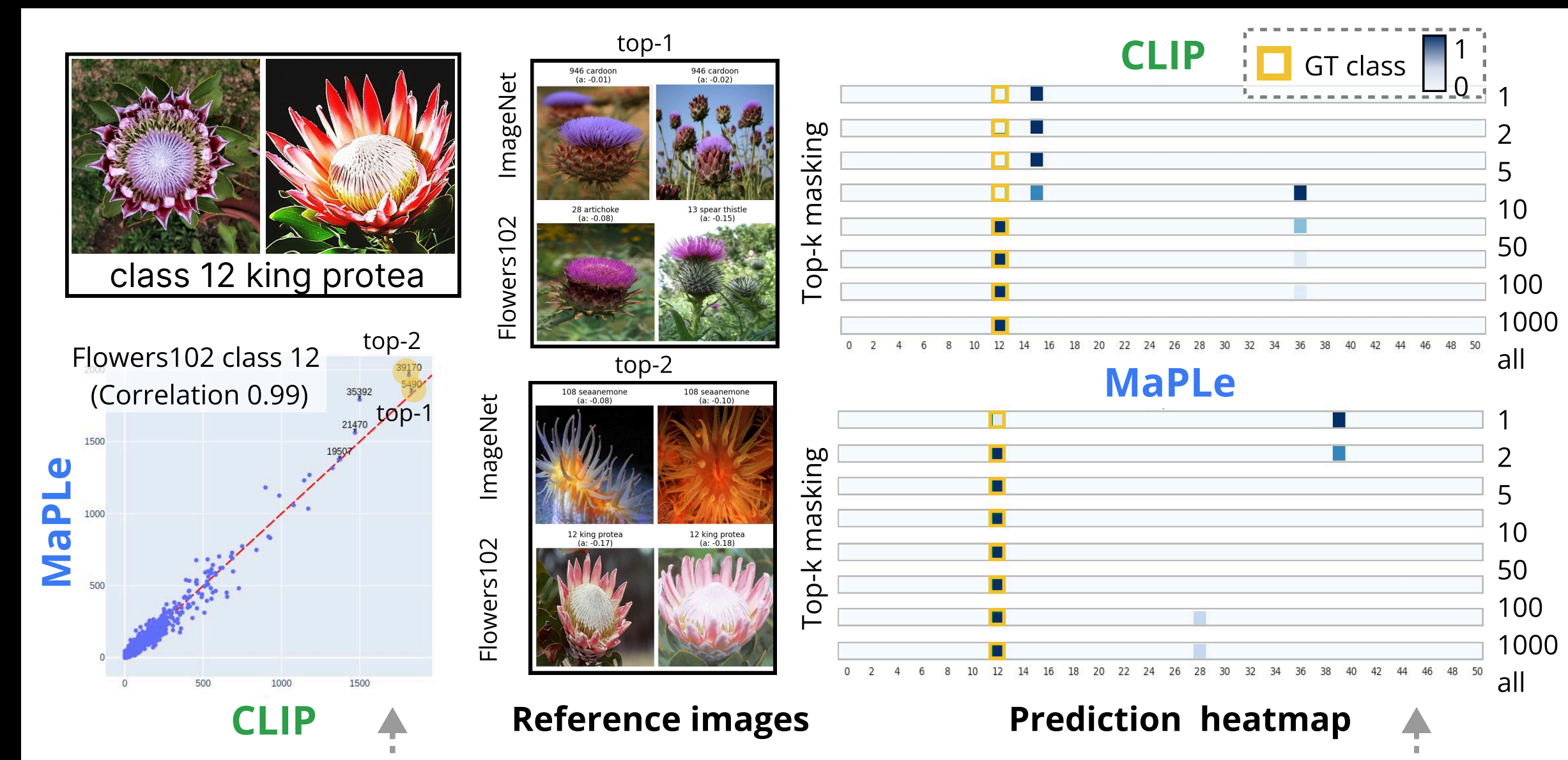
Understanding adaptation mechanisms



Understanding adaptation mechanisms



Before adaptation: **CLIP**
After adaptation: **MaPLE**

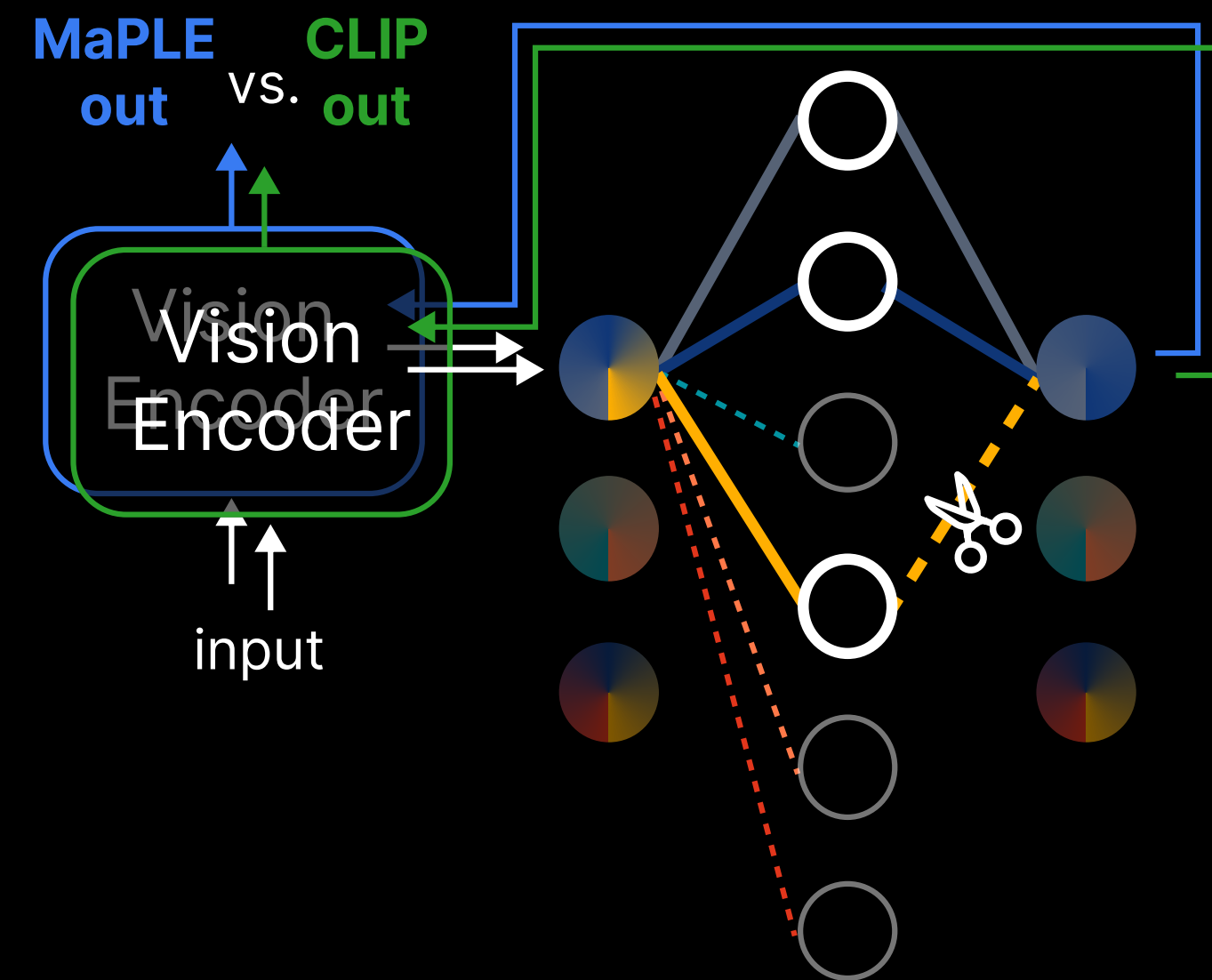


Adaptation **re-uses**
the concepts

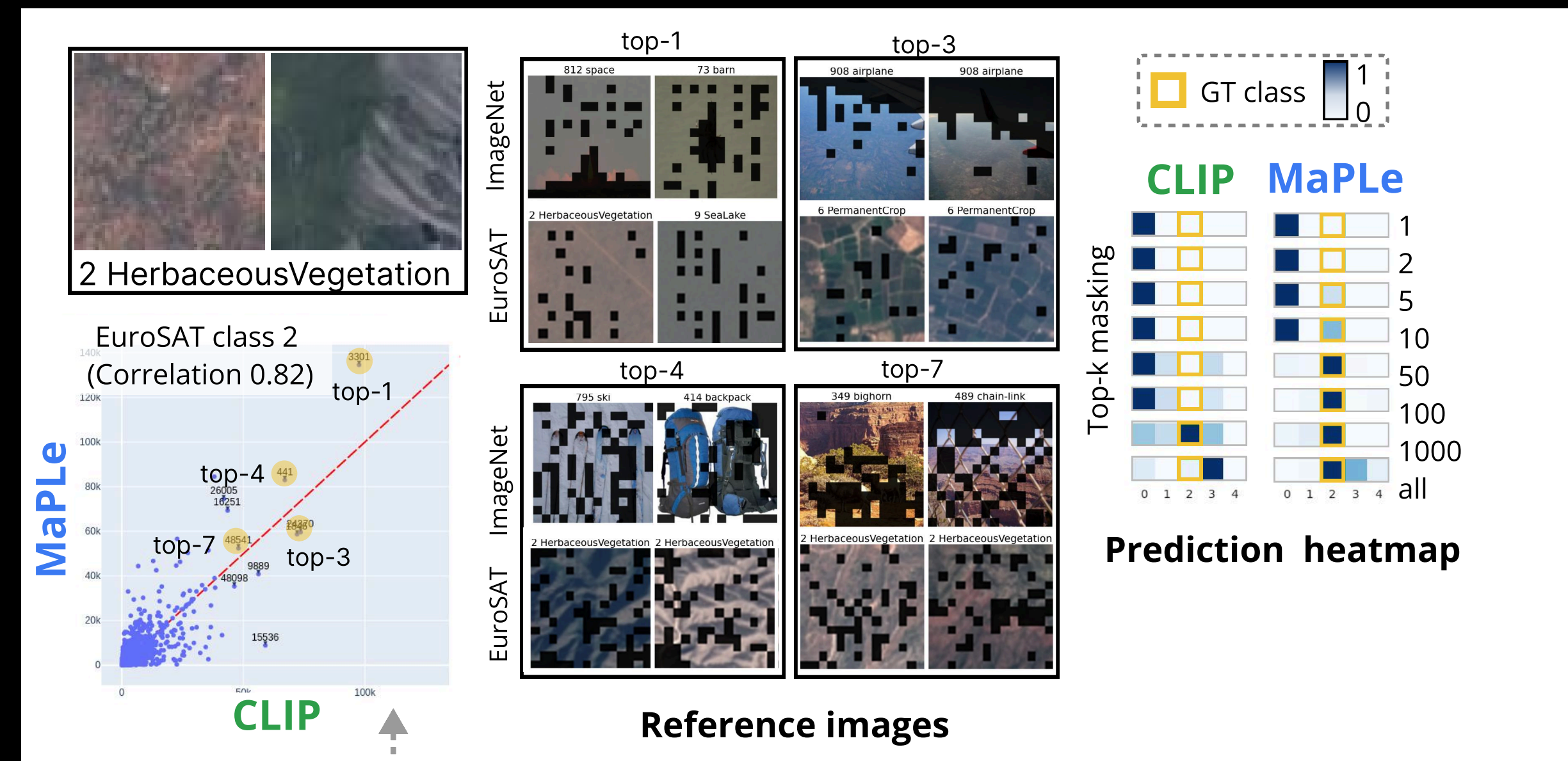
Adaptation **re-associates**
the concepts to the right classes

* Automated flower classification over a large number of classes (Nilsback & Zisserman, 2008)

Understanding adaptation mechanisms



Before adaptation: **CLIP**
After adaptation: **MaPLE**



Adaptation introduces
new concepts

* Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification (Helber et al., 2019)

PatchSAE

Interpreting CLIP representations

We introduce PatchSAE, an interpretability tool for vision models.

Observing decision-making process

PatchSAE shows that CLIP models recognize class-discriminative information and use it for downstream tasks

Understanding adaptation mechanism

Learnable prompts allow CLIP to link already recognized concepts towards the targeting task

Sparse Autoencoders Reveal Selective Remapping of Visual Concepts During Adaptation

ICLR 2025 (Paper 14240)

Hyesu Lim^{1,2} Jinho Choi² Jaegul Choo² Steffen Schneider^{1,3}

Project page

<https://dynamical-inference.ai/patchsae>

Code and Demo

<https://github.com/dynamical-inference/patchsae>

