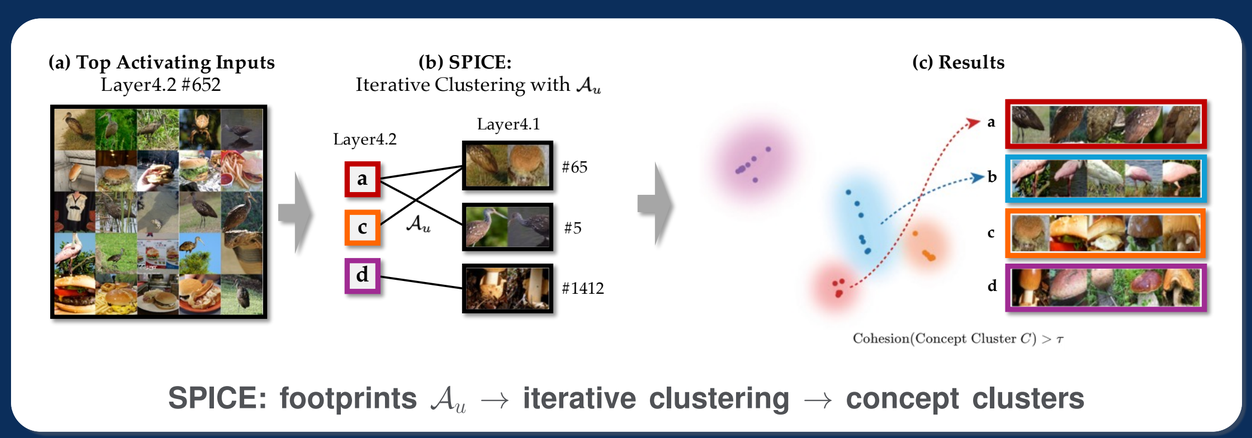
Sehyun Lee, Dahee Kwon, Damin Lee, Jaesik Choi
Mechanistic Interpretability Workshop, ICML 2026



SPICE: footprints $\mathcal{A}_u \rightarrow$ iterative clustering $\rightarrow$ concept clusters

One Neuron, Many Concepts.

SPICE disentangles them — automatically, on any architecture.

SPICE: Simple Polysemantic Feature Interpretation via Clustering-based Explanations

Sehyun Lee, Dahee Kwon, Damin Lee, Jaesik Choi — KAIST

1 TL;DR

MAIN 3 TAKEAWAYS

- **No architecture-specific rules are needed** to disentangle polysemantic concepts — clustering attribution footprints works on both **CNNs and Transformers**.
- **No manual K**: an adaptive cohesion threshold discovers the number of concepts per neuron automatically.
- **Distinct inductive biases**: CNNs keep locally cohesive concepts; ViTs diversify mid-network, then **reconsolidate** via global self-attention.

2 Background

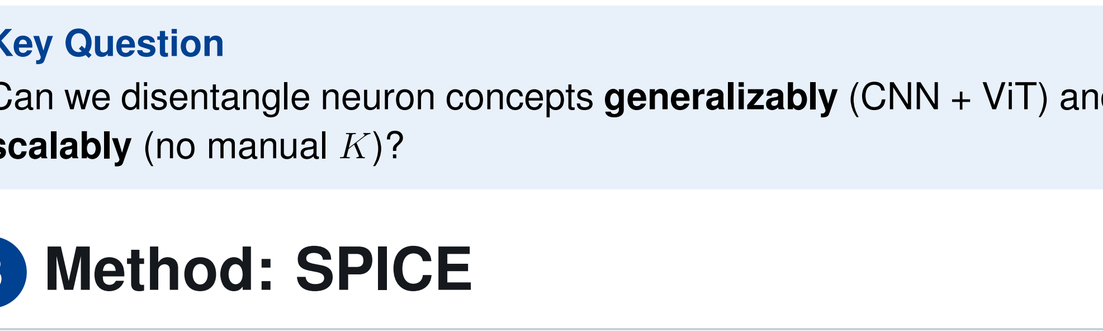
- Neurons are often **polysemantic**: one unit activates for multiple unrelated concepts, breaking single-neuron interpretation.
- SAEs learn a *dictionary* but do not trace the network's **intrinsic computational flow**; prior attribution clustering (PURE) is **CNN-only** with a **manually fixed** K.

Key Question

Can we disentangle neuron concepts **generalizably** (CNN + ViT) and **scalably** (no manual K)?

3 Method: SPICE

- **Attribution footprints.** Input $\times$ Gradient contribution of each preceding layer neuron $\Rightarrow$ attribution $\mathcal{A}_u \in \mathbb{R}^d$ (normalized against deep-layer anisotropy), capturing which upstream features the concept is built from.
- **Iterative clustering.** Start at $k{=}2$; keep clusters with $\mathrm{Cohesion}(C) > \tau$ (mean pairwise cosine sim.), remove them, adapt k, repeat — K **emerges from the data**.
- **Adaptive threshold** $\tau = \max(95\text{th pct},\ \text{KDE 2nd peak})$: handles unimodal early layers *and* bimodal deep layers.



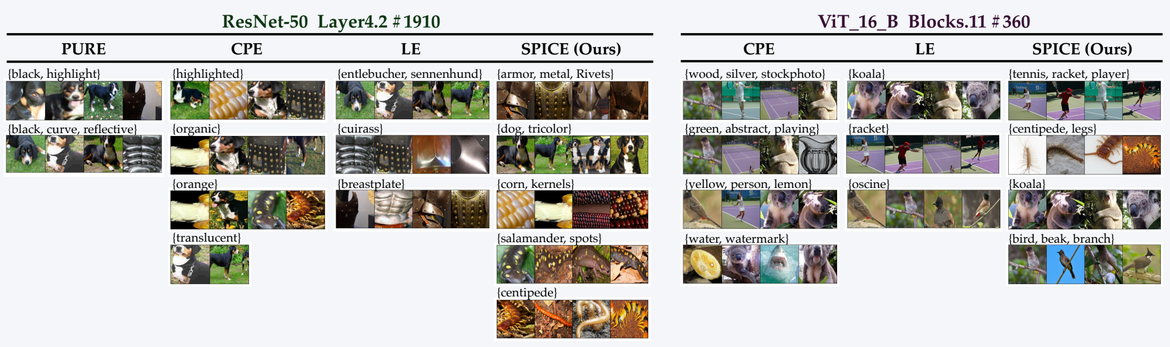
Concepts per neuron, intra-cluster similarity, and
the unimodal $\rightarrow$ bimodal shift behind the adaptive τ.

4 Results

- Best or near-best **separability** across four architectures — and the clear winner in the deepest, most polysemantic layers.
- Highest AUC/MAD in **generative (CoSy) eval**.

Deepest layer	PURE	LE	CPE	**Ours**
ResNet-50 (L4.2)	1.009	1.187	0.998	**1.291**
ViT-B (B11)	1.035	1.357	0.998	**1.577**
DenseNet (L4.16)	1.025	**1.191**	0.994	1.033
CLIP ViT-B (B11)	1.021	1.244	0.998	**1.410**

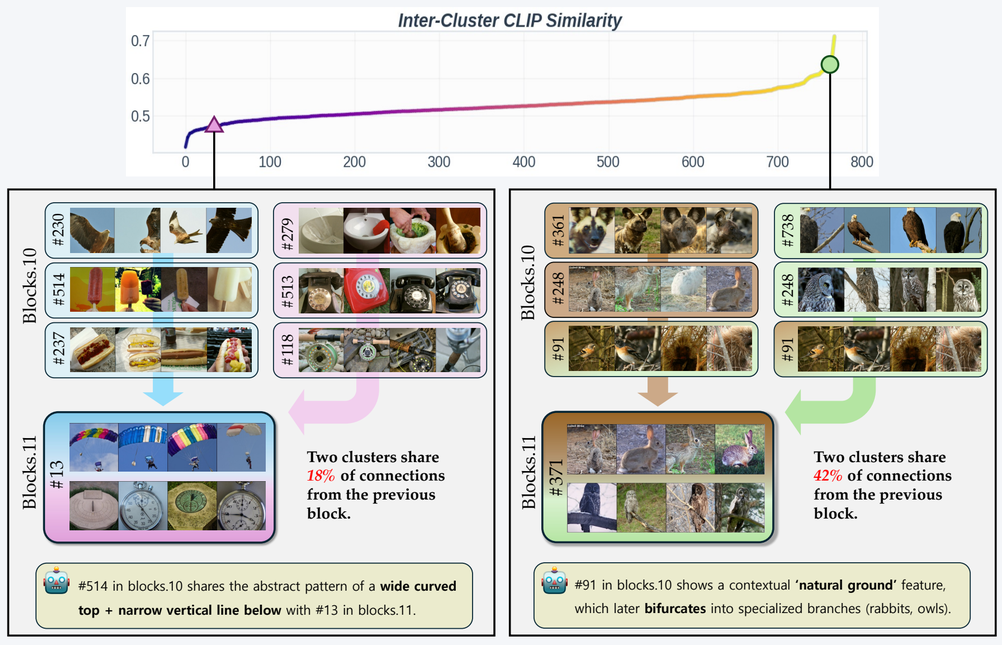
Separability, ratio of within-cluster to between-cluster concept margin



SPICE recovers **more, cleaner concepts** than baselines — PURE/CPE/LE.

Robust where prior work never looked.

- **Full-range evaluation** Standard protocols evaluate only the highly activation range; we extend to Top/Middle/Bottom ranges across depths.
- **Robust in weak ranges** SPICE stays strong in highly polysemantic middle/bottom ranges (e.g., 1.41 at ViT-B B10), whereas LE collapses (1.04) and PURE/CPE stay near 1.0.
- **Adaptive** τ $\approx$100% coverage vs. 6% for static. No over-segmentation.



How polysemanticity forms. 18% vs. 42% upstream overlap. Unrelated concepts arise from **disjoint circuits**; related ones share a **common pathway**

5 Limitations and Discussion

- Iterative clustering costs more than single-pass methods.
- CLIP verification penalizes low-level early-layer features.
- **Next** circuit-level disentanglement along pathways, beyond single neurons.

Accepted to ECCV 2026

See you in Malmö, Sweden

Come discuss with us there too!

PAPER

CODE

{sehyun.lee, daheekwon}@kaist.ac.kr