ICML 2026
Forty-Third International Conference
on Machine Learning

SAEs-BrainMap: Unveiling the Emergence of Specialized Concepts in Deep Models via Brain Alignment

Ziming Mao

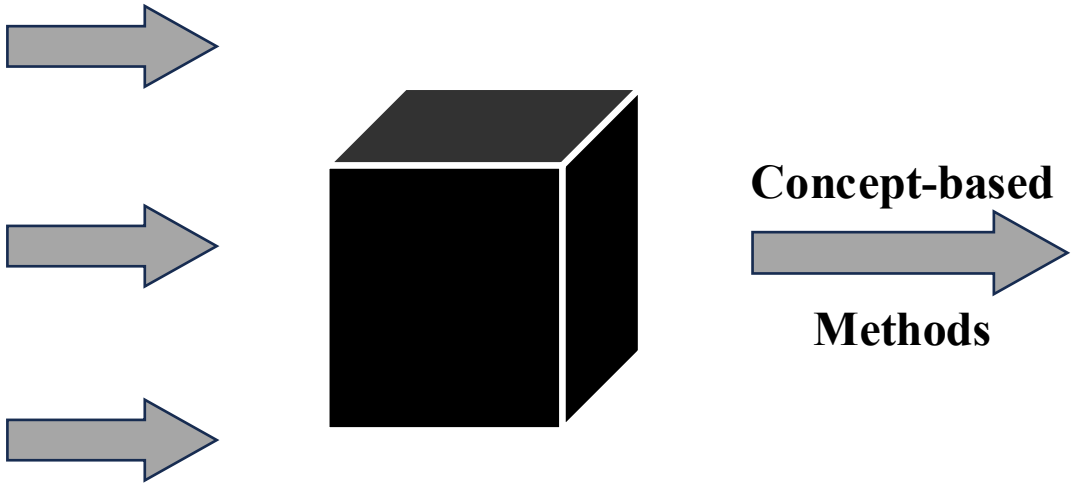
Beijing Institute of Technology & Westlake University

maoziming@westlake.edu.cn

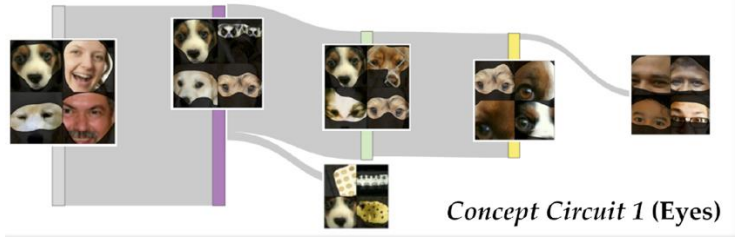
01

Introduction & Method

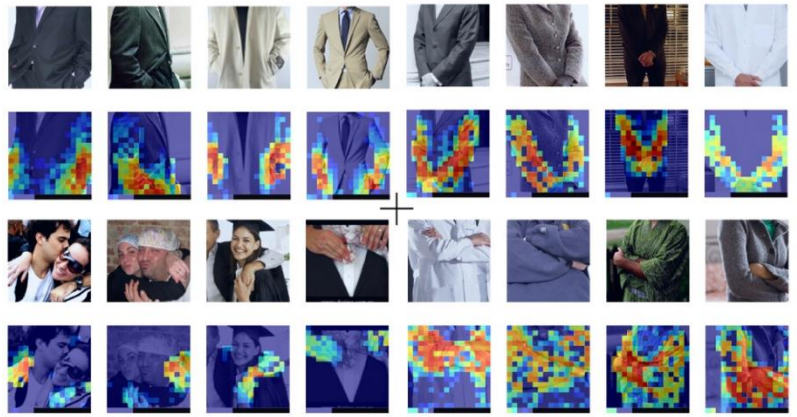
Concept-based Interpretability Methods



Network dissection



Granular Concept Circuits



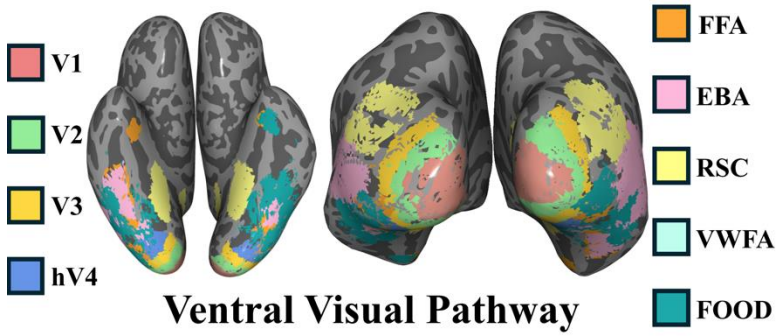
Sparse Autoencoder

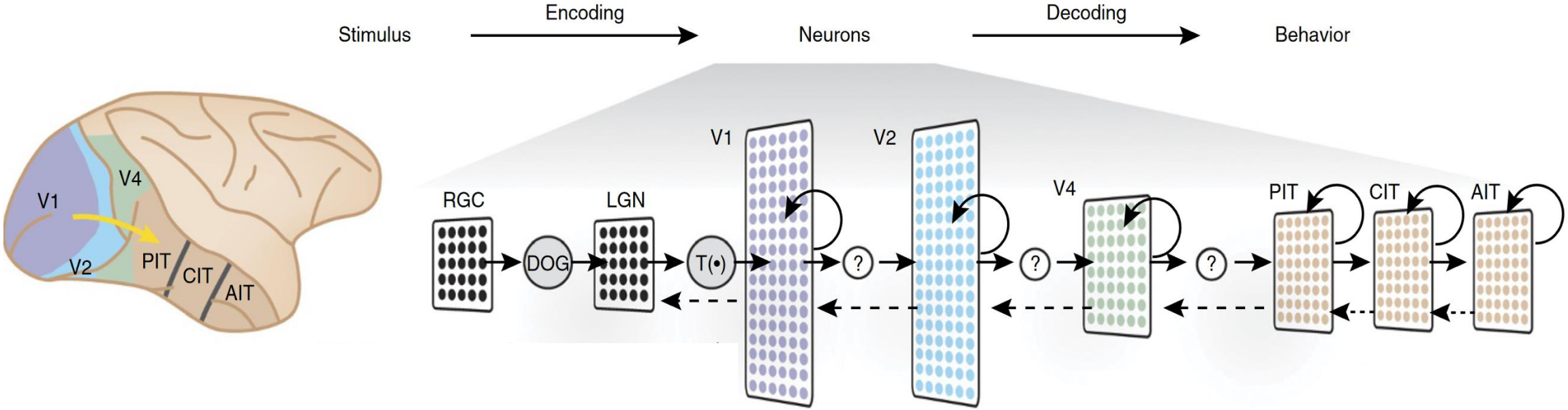
Granular Concept Circuits: Toward a Fine-Grained Circuit Discovery for Concept Representations. Dahee Kwon, et.al.

Network Dissection: Quantifying Interpretability of Deep Visual Representations. David Bau, et.al.

Archetypal SAE: Adaptive and Stable Dictionary Learning for Concept Extraction in Large Vision Models. Thomas Fel, et.al.

The ventral visual pathway

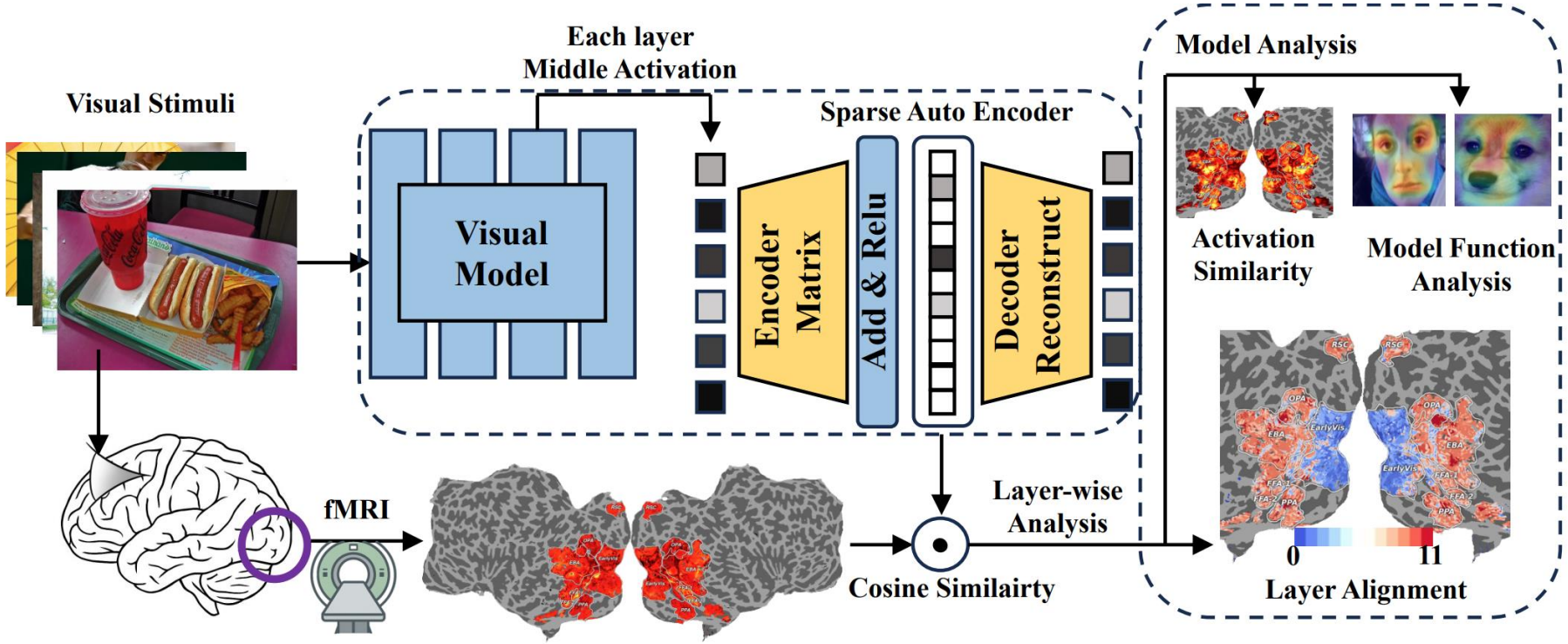




The transmission of visual information in the ventral visual pathway

Using goal-driven deep learning models to understand sensory cortex. Daniel L K Yamins, et.al.

SAEs-BrainMap Structure



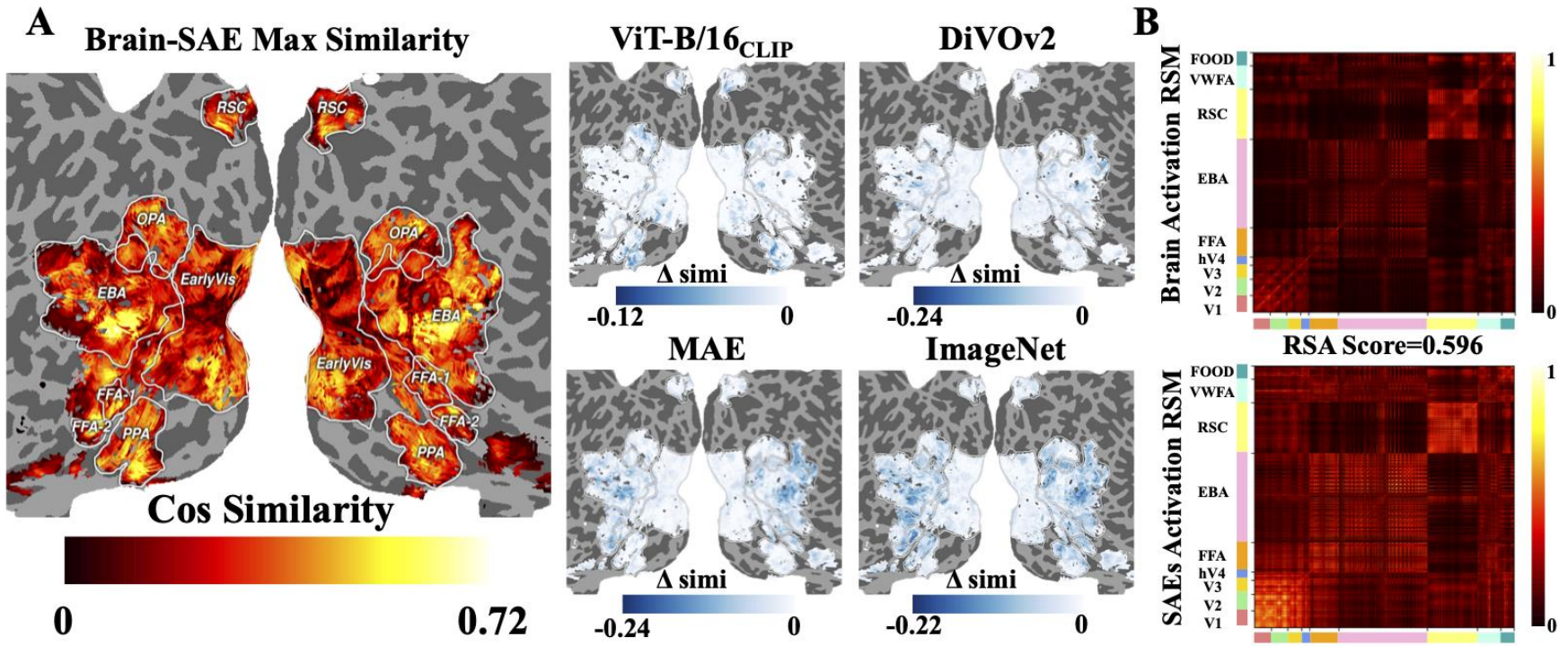
Procedure:

1. Layer-wise SAEs Training
2. Brain & SAE Feature Alignment
3. Concept Emergence Analysis

02

Brain & SAEs Feature Alignment

Activation & Structure

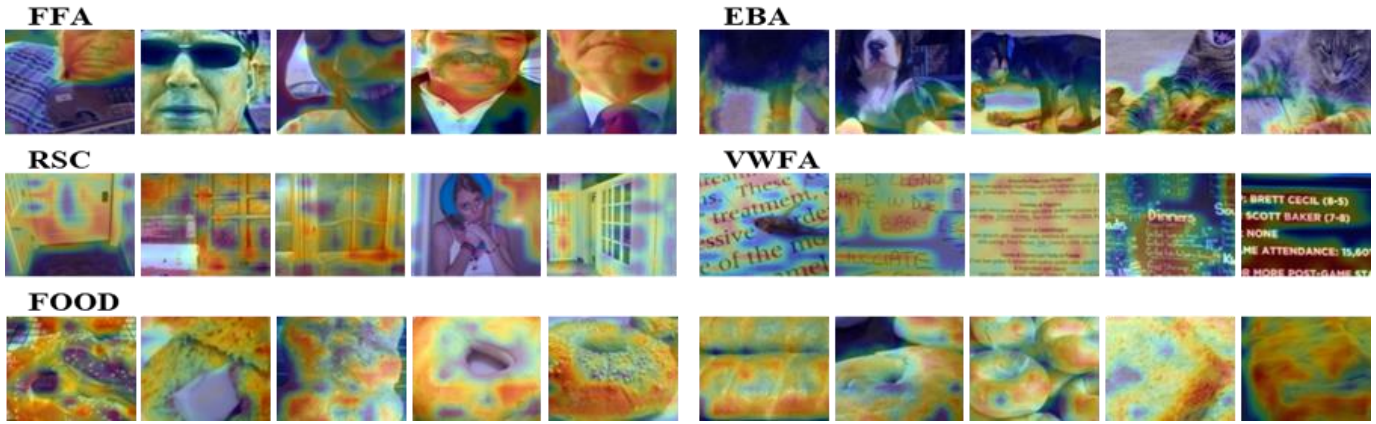


Model	Max Simi	Mean Simi	RSA Score
CLS SAEs			
CLIP	0.7247	0.2811	0.580
DINOV2	0.7172	0.2705	0.686
MAE	0.6198	0.2582	0.646
IMAGENET	0.6431	0.2495	0.630
Patch SAEs			
CLIP	0.6858	0.2708	0.596
DINOV2	0.6517	0.2670	0.592
MAE	0.6376	0.2327	0.586
IMAGENET	0.6019	0.2392	0.656
Raw Neurons			
CLIP	0.5875	0.2354	0.511
DINOV2	0.6961	0.2421	0.579
MAE	0.5108	0.2087	0.551
IMAGENET	0.5319	0.2245	0.596

A. Activation Similarity.
Evaluate with cosine similarity.

B. Structural Correlation.
Evaluate with Representation Similarity Analysis (RSA)

Functional Alignment (ROI Level)



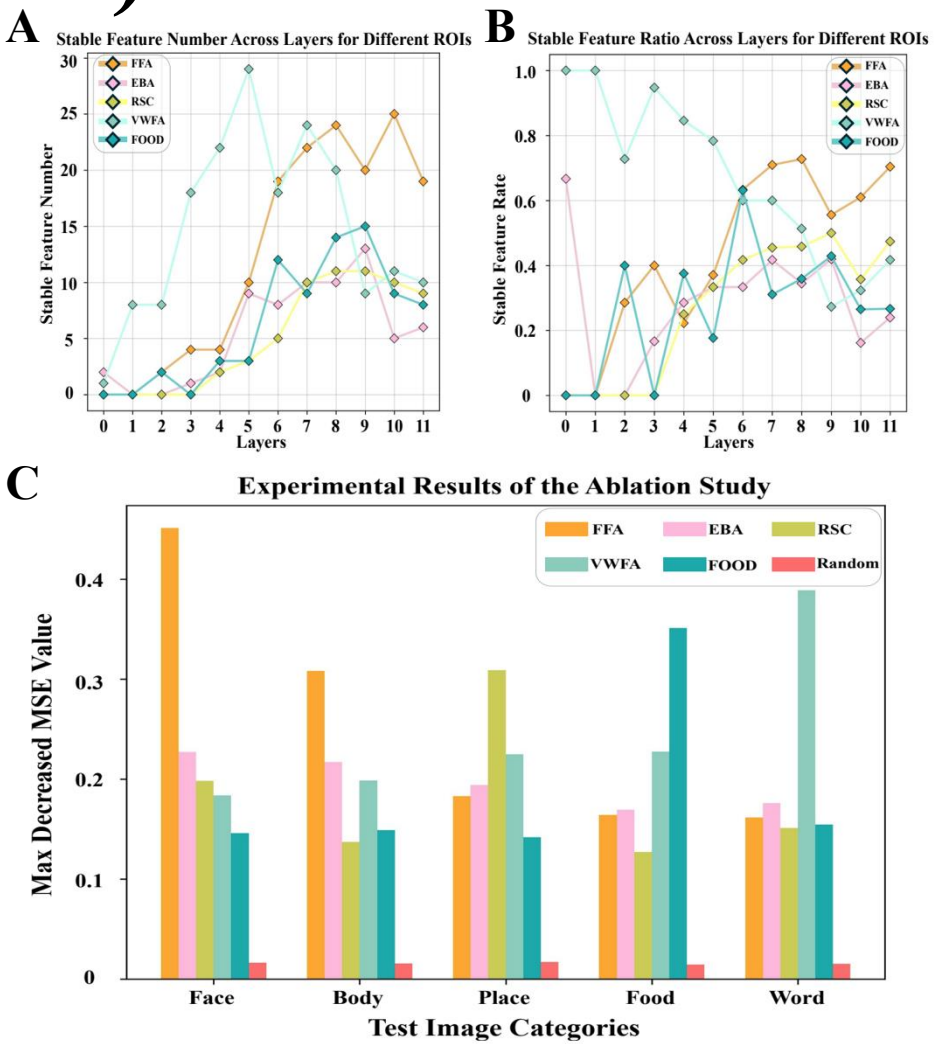
Features selected from the layer that have highest functional-aligned number.

Functional Alignment Evaluation.

1. Select Top 100 features that most correlate with a certain ROI

2. Use CLIP model and text prompt to evaluate the stability and functional alignment with the brain of the features' highest activated images.

3. Visualize Features Selectivity.



Stability Evaluation & Ablation Study

02

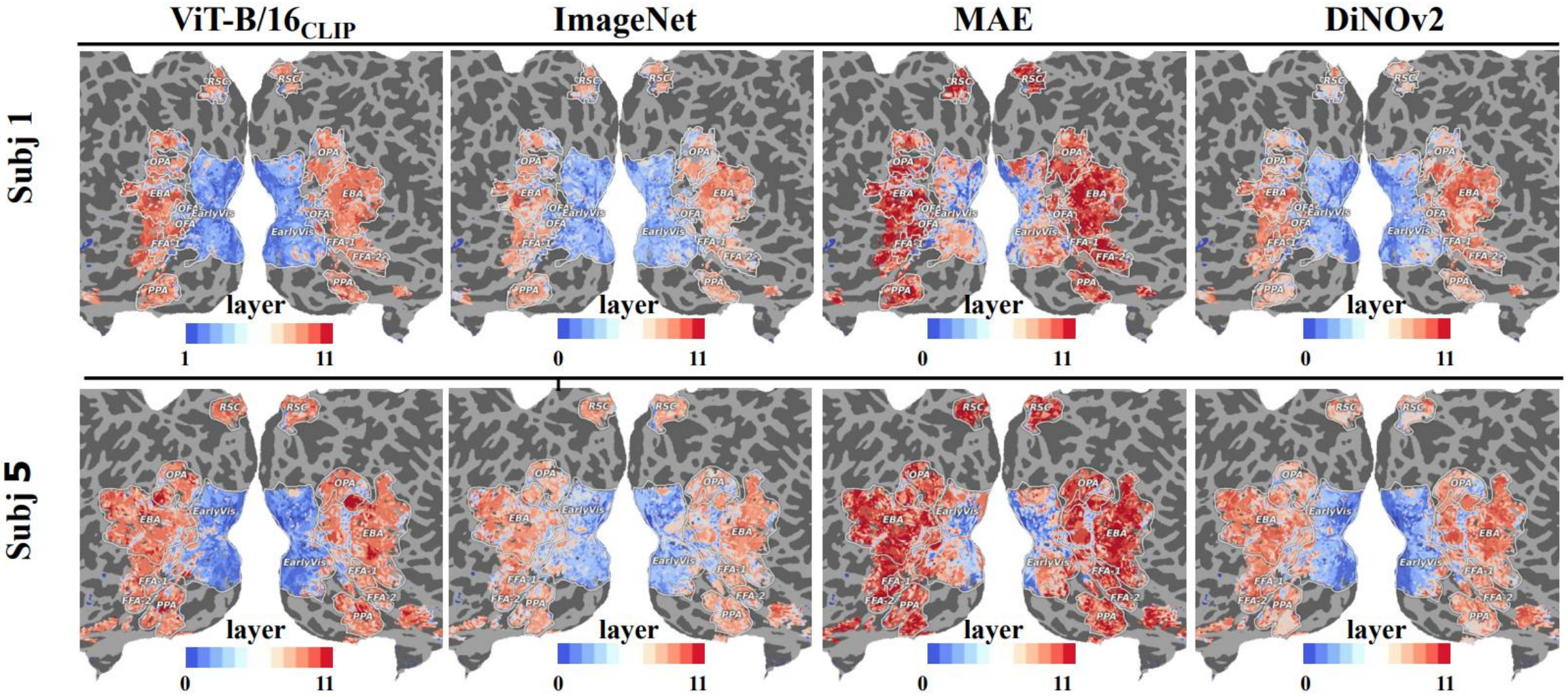
Concepts Emergence Analysis

Concept Emergence In CLIP ViT-B/16

Layer-Wise Visualization of features that selected by FFA

Partial Results for Five High-Level ROIs

Layer Mapping On Cortex



Main Contribution

1. We explore a brain-guided framework, SAEs-BrainMap, for analyzing layer-wise SAE features in deep visual models.

2. We provide empirical evidence that SAE features can show measurable activation-level and structure-level alignment with fMRI responses from the ventral visual pathway.

3. Based on ROI-level brain signals, we identify a subset of SAE features whose visual selectivity appears to be consistent with known functional preferences of high-level visual regions.

4. Our analysis offers a possible way to trace the layer-wise emergence of several generic visual concepts, including faces, bodies, places, words, and food, within the studied vision models.

ICML 2026
Forty-Third International Conference
on Machine Learning

Thank you for your listening!

Email:maoziming@westlake.edu.cn