VISION,

Possible way to Understand visual comprehension and More via AI

Unidirectional Brain-Computer Interface: Artificial Neural Network Encoding Natural Images to FMRI Response in the Visual Cortex

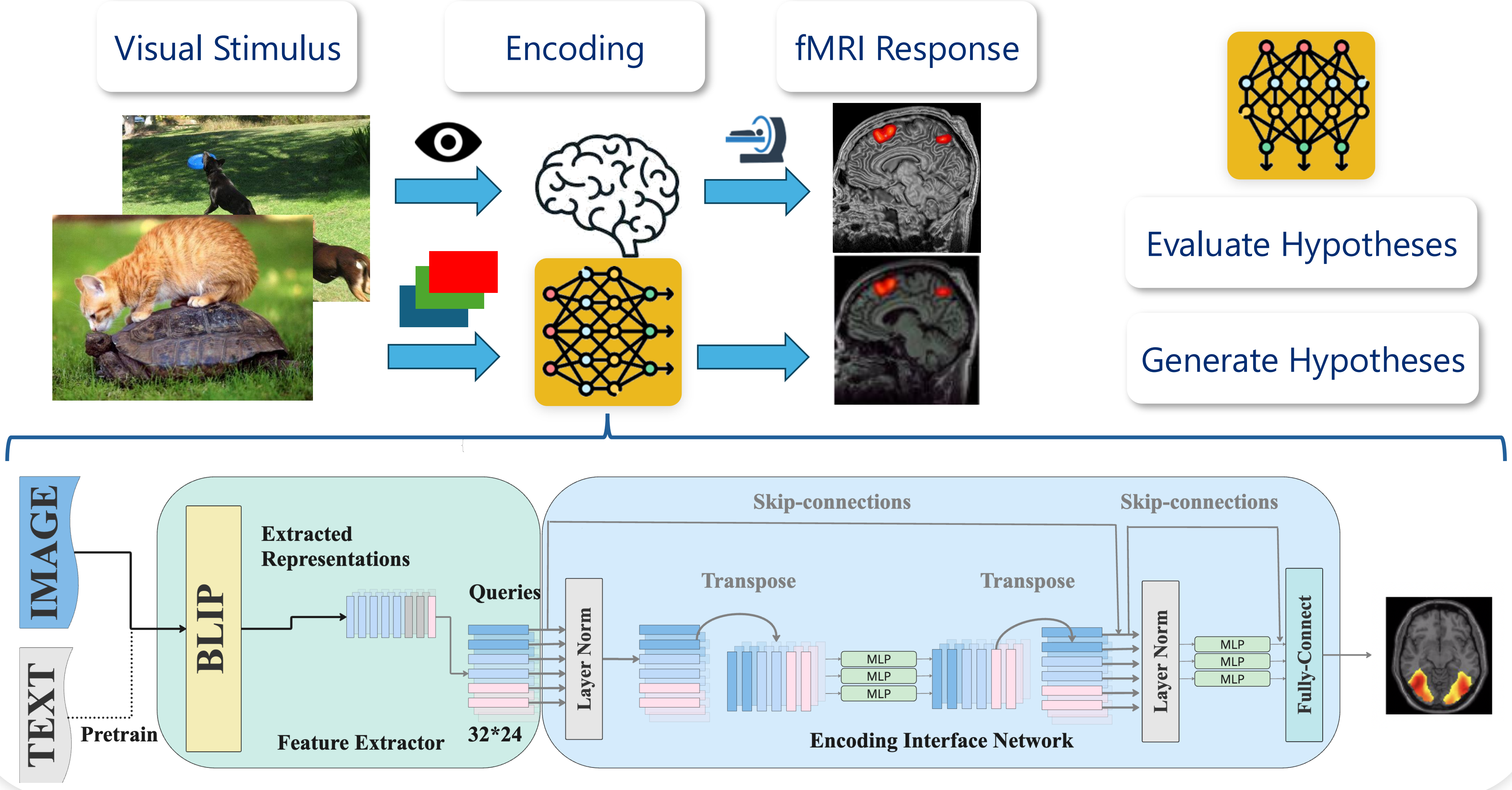
Abstract

While significant advancements in artificial intelligence (AI) have catalyzed progress across various domains, its full potential in understanding visual perception remains underexplored. We propose an artificial neural network dubbed VISION, an acronym for "Visual Interface System for Imaging Output of Neural activity," to mimic the human brain and show how it can foster neuroscientific inquiries. Using visual and contextual inputs, this multimodal model predicts the brain's functional magnetic resonance imaging (fMRI) scan response to natural images. VISION successfully predicts human hemodynamic responses as fMRI voxel values to visual inputs with an accuracy exceeding state-of-the-art performance by 45%. We further probe the trained networks to reveal representational biases in different visual areas, generate experimentally testable hypotheses, and formulate an interpretable metric to associate these hypotheses with cortical functions.

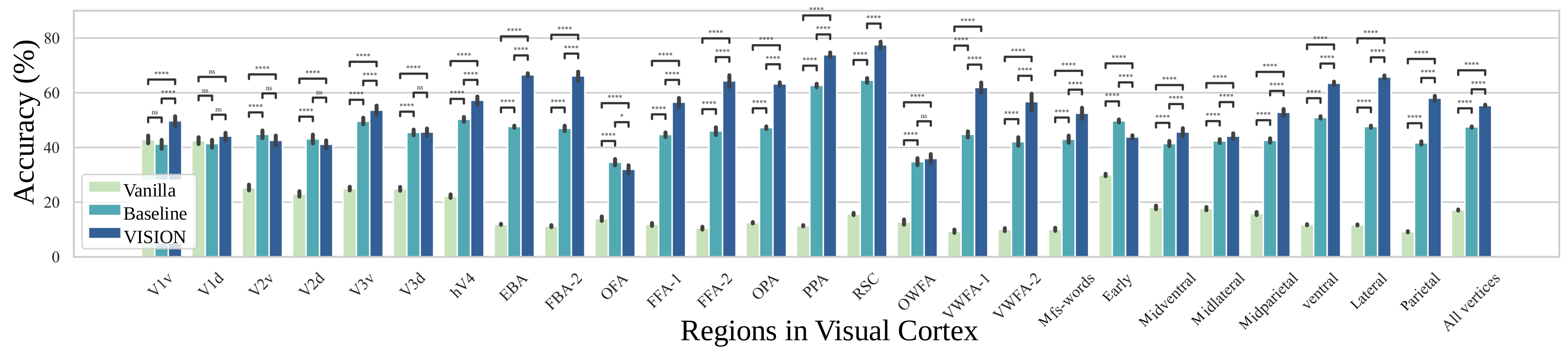
With both a model and evaluation metric, the cost and time burdens associated with designing and implementing functional analysis on the visual cortex could be reduced. Our work suggests that the evolution of computational models may shed light on our fundamental understanding of the visual cortex and provide a viable approach toward reliable brain-machine interfaces.

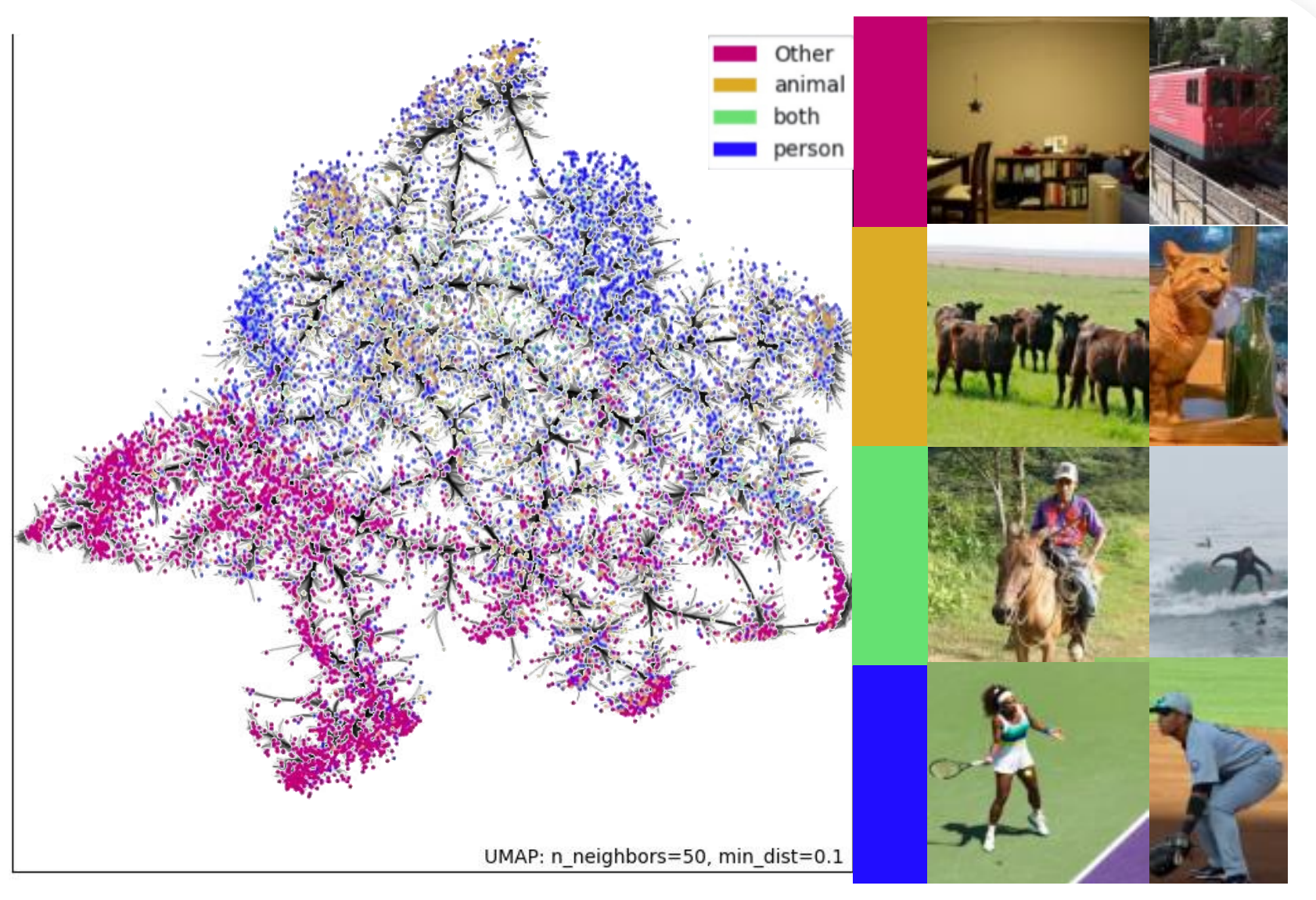
Method

Through pretrained multi-modal transformer, using its image encoder and Encoding Interface Network, VISION gained power to estimate fMRI responses like brain.



Results

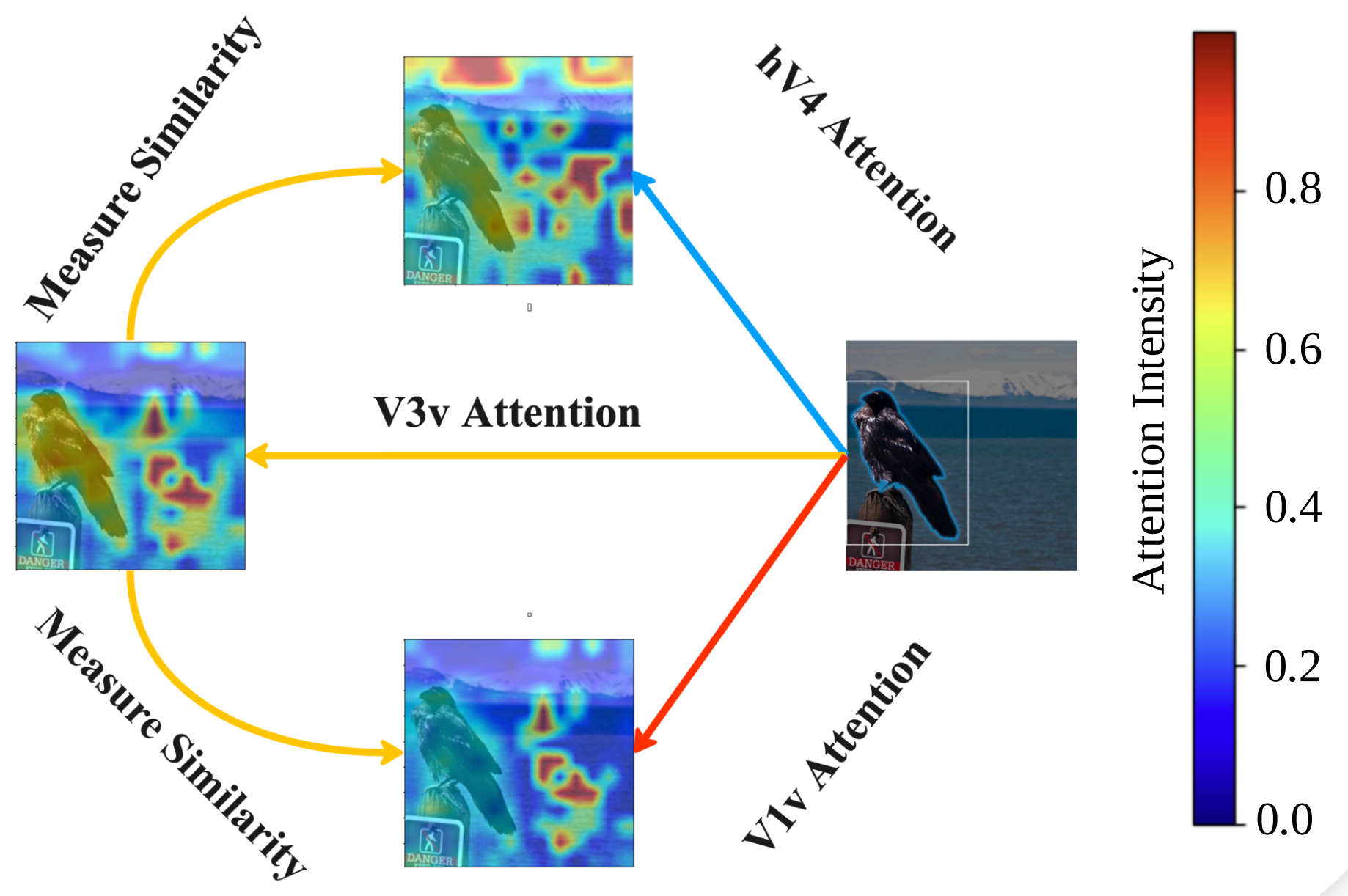




Benchmarking: We have compared both the quantitative and qualitative outcomes of our models against two established baselines (i.e., vanilla by Allen et al. and baseline model by Gifford et al.)

Feature Space Visualization: As illustrated in figure below, response patterns within all participants' ours models tend to group semantically. Typically, such distinct clustering, where identical categories are close-knit while being separated from dissimilar ones, emerges in networks that have been specifically trained and regularized to differentiate between these categories. Moreover, when compared with previous works, this clustering performance is improved.

Attention Visualization: Following the traditional use of ScoreCam on BLIP's vision encoder's last normalization layer, we have selected hV4, V3v, and V1v as testing regions. According to KL divergence, hV4 and V3v are 3.72 times more similar than hV4 and V1v, which could also be visually inferred in figure. This is consistent with functional connectivity analysis derived from resting state fMRI and current findings in figure which futher shows VISION model resemblance with the biological visual cortex.



Ruixing Liang*, Xiangyu Zhang*, Qiong Li, Lai Wei, Hexin Liu, Avisha Kumar, Kelley M Kempski Leadingham, Joshua Punnoose, Leibny Paola Garcia, Amir Manbachi

JOHNS HOPKINS MEDICINE
JOHNS HOPKINS WHITING SCHOOL of ENGINEERING
HEPIUS
PennState College of Engineering