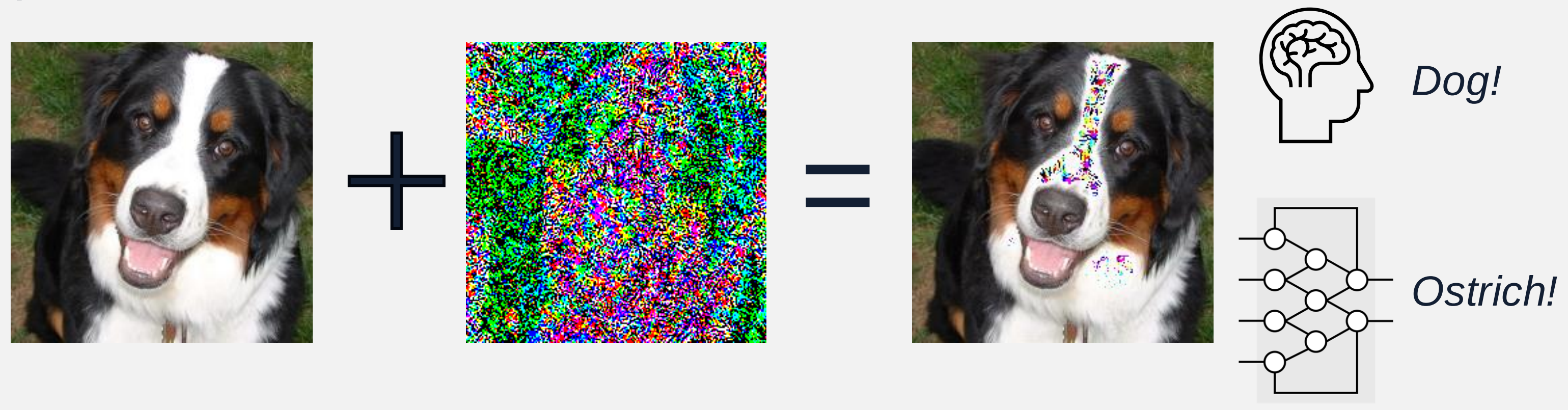
Increasing robustness of ventral visual cortex revealed by neurally-guided deep neural networks

Zhenan Shao[1,2], Linjian Ma[2], Bo Li[2,3], Diane M. Beck[1]
[1]Department of Psychology, Beckman Institute, University of Illinois at Urbana-Champaign
[2]Department of Computer Science, University of Illinois at Urbana-Champaign
[3]Department of Computer Science, University of Chicago

INTRODUCTION

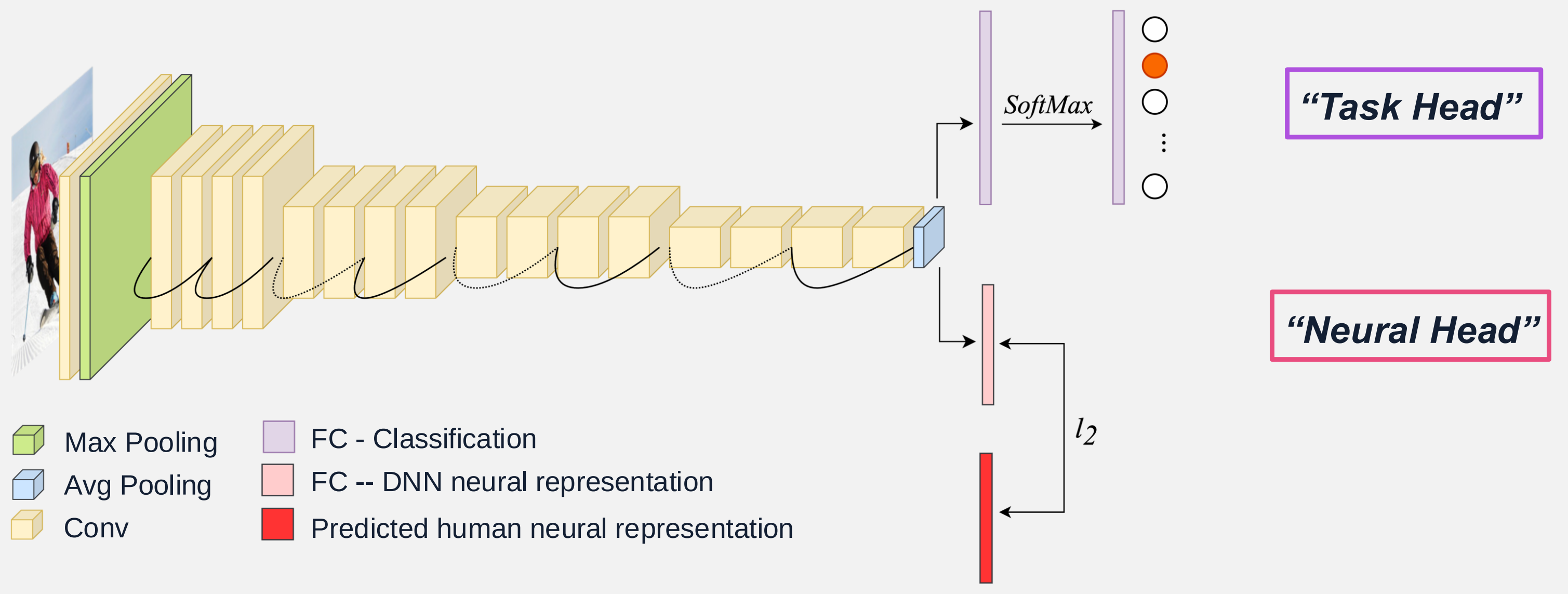
The human visual system is remarkably robust to identity-preserving changes to the image (i.e. changes in viewpoint, illumination, noise), an achievement thought to evolve across successive stages of the ventral visual stream (VVS)[1,2].

Although deep neural networks (DNNs) can achieve human-level performance on many visual tasks, they have been shown to be vulnerable to "*adversarial attacks*" — subtle image perturbations (see below) that drastically reduce DNNs' performance[3].



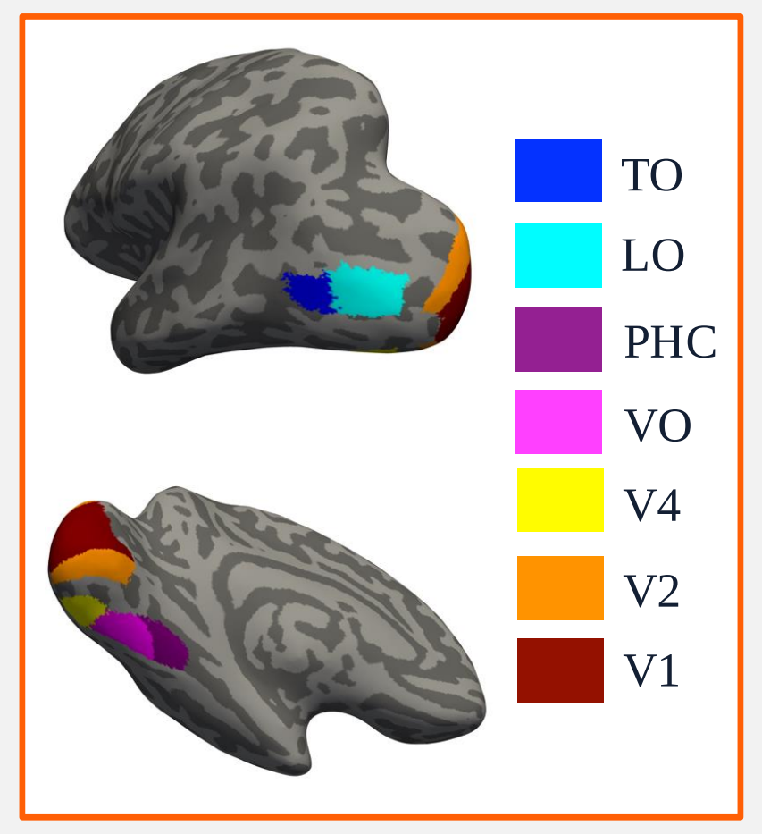
Hypothesis: Guiding DNNs to learn neural representations from successive stages of the VVS should result in successive increases in robustness to adversarial attack.

METHODS



Neural-guidance: Similar to previous work[4,5], we employ a two-headed ResNet18 architecture that simultaneously learns a 50-category ImageNet classification task ("*task head*") while aligning the model's penultimate layer with the neural representations ("*neural head*") from a specific region of interest (ROI).

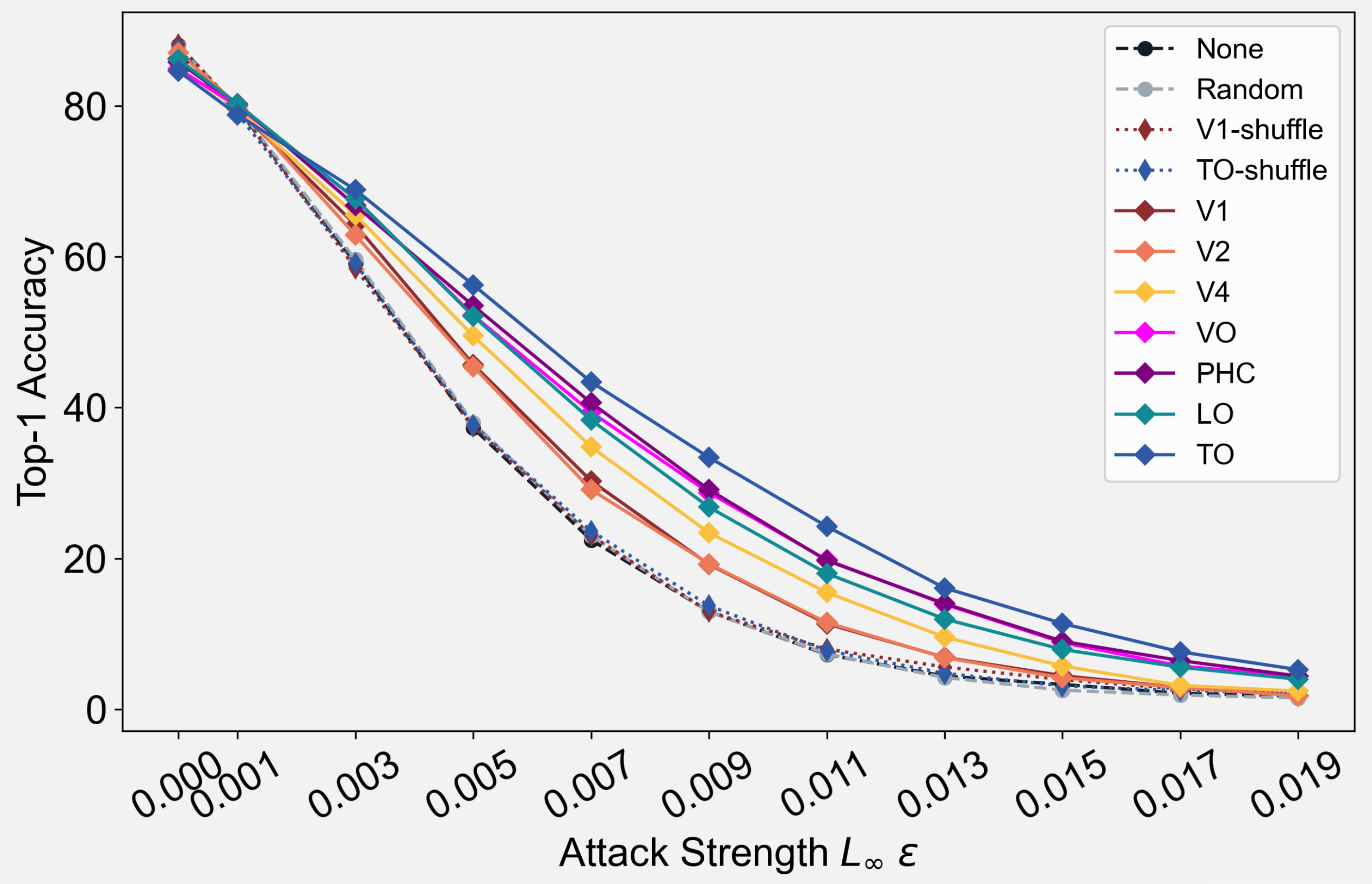
Neural representation: We used human 7T fMRI data from the NSD[6] dataset, extracting seven ROIs to capture the evolving representational space. "*Neural predictors*" were trained as surrogates for each ROI.

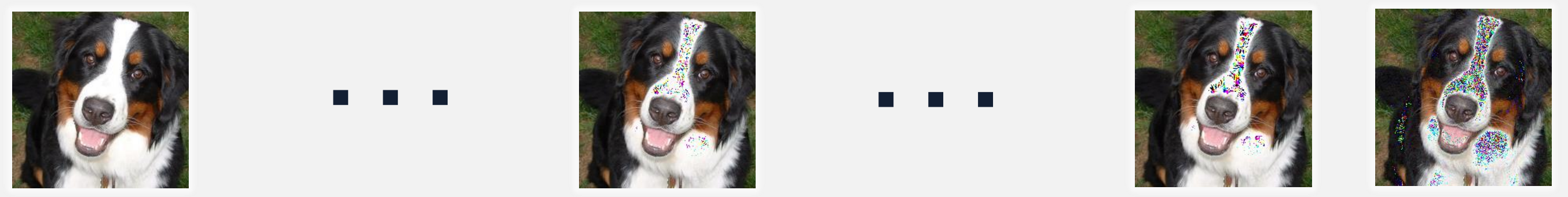


Control conditions: Four baseline models were included: "*None*", "*Random*", "*V1-shuffle*", and "*TO-shuffle*", each representing different alternative hypotheses.

Hierarchical Robustness Gain

l_∞-based adversarial attack: $\max_{||\tau||_p < \epsilon} l(f_\theta(x+\tau), y)$

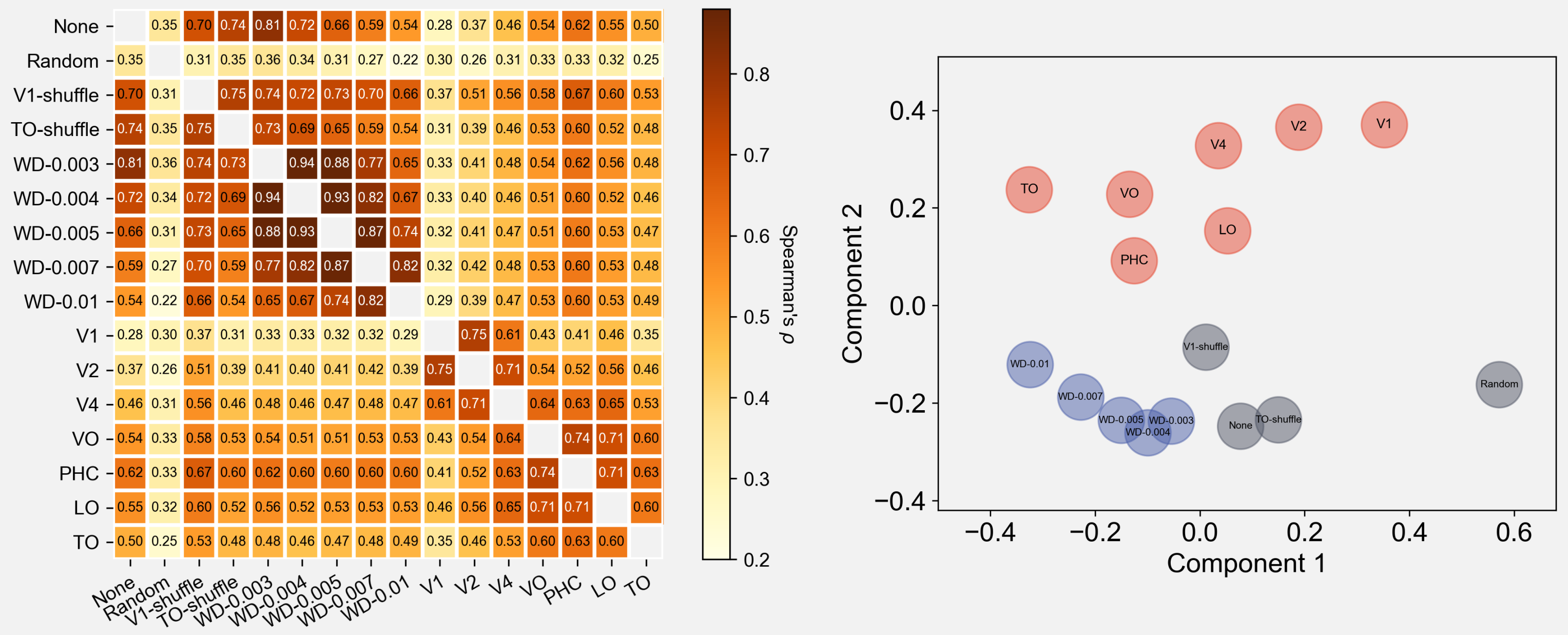




Neural guidance improved DNN robustness, with a clear hierarchical pattern of increasing improvement when using progressively later brain regions. This pattern was consistent across multiple human subjects, adversarial attack benchmarks, datasets, and tasks.

Distinct Representational Spaces

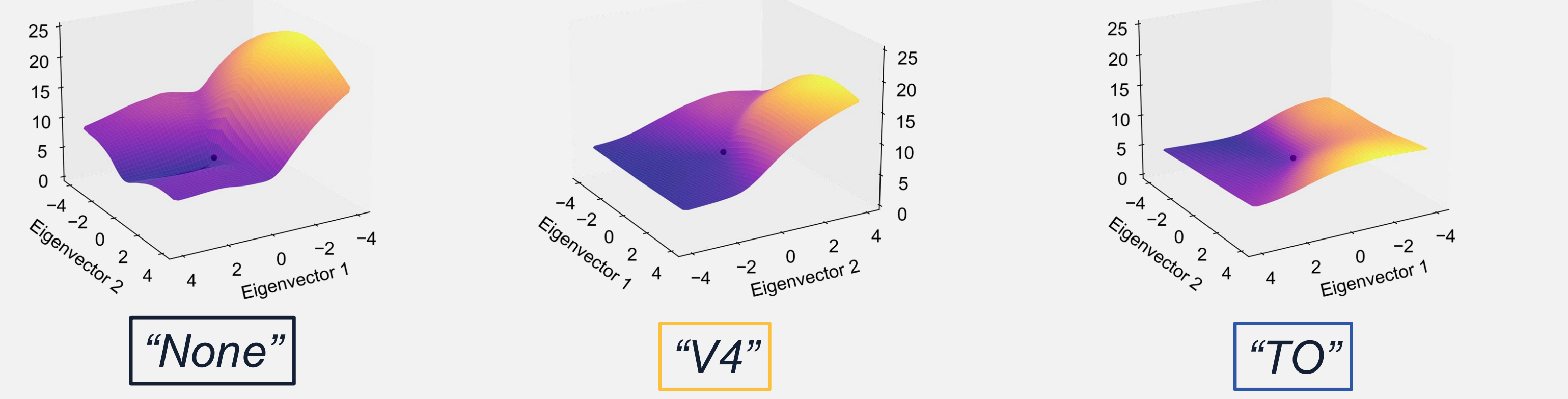
Representational Similarity Analysis (RSA)[7] revealed a representational shift for neurally-guided models (red circles) away from conventional models (blue and gray circles). These distinct neural representational space may contribute to the improved robustness in neurally-guided models.



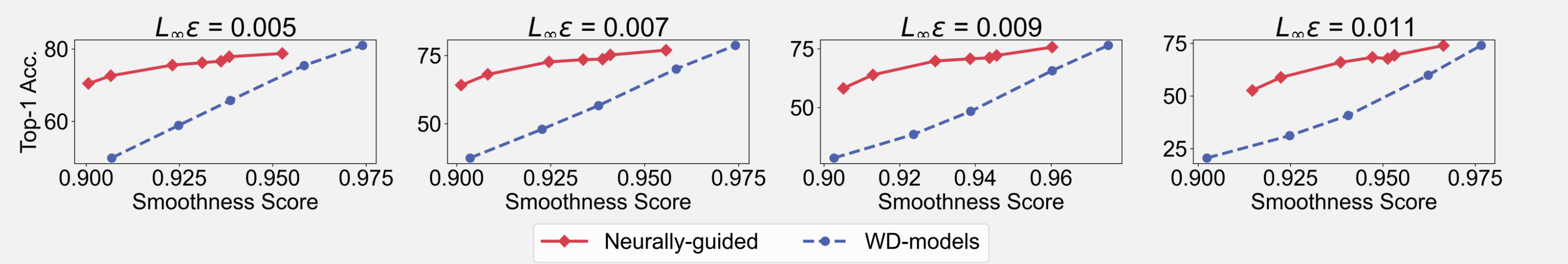
Multidimensional Scaling (MDS) plot (right) of RSA matrix (left) is provided for a clearer 2D visualization.

Decision Surface Smoothness and Uniqueness

Surface smoothness: DNNs guided by successive regions of the VVS exhibited increasingly smoother loss landscapes, corresponding to the increasing robustness gains.



Surface uniqueness: Neurally-guided models resulted in decision spaces more resistant to adversarial examples transferred from standard ResNet models, compared to models smoothed by conventional methods[8].



CONCLUSIONS & IMPLICATIONS

Conclusions:

- Neural guidance from successive ROIs in VVS leads to hierarchical improvements in DNN adversarial robustness.
- Neurally-guided DNNs developed distinct representational spaces that are smoother and resistant to transfer attacks.

Implications:

- Robustness emerges from the evolving representational space along the ventral visual stream
- Potential for understanding human representational space and advancing DNN architectural developments

REFERENCES

[1][Dicarlo, & Cox, 2007] [2][Isik et al., 2014] [3][Szegedy et al., 2014] [4][Li et al., 2019] [5][Dapello et al., 2022] [6][Allen et al., 2022] [7][Kriegeskorte, 2008] [9][Rosca et al., 2020]

Paper

Code

This work used NCSA Delta GPU through allocation SOC230011 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.