Self-supervised transformers predict dynamics of object-based attention in humans

Hossein Adeli[1], Seoyoung Ahn[2], Nikolaus Kriegeskorte[1], Gregory J. Zelinsky[2,3]

[1]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, USA
[2]Department of Psychology, [3]Department of Computer Science, Stony Brook University, New York, USA

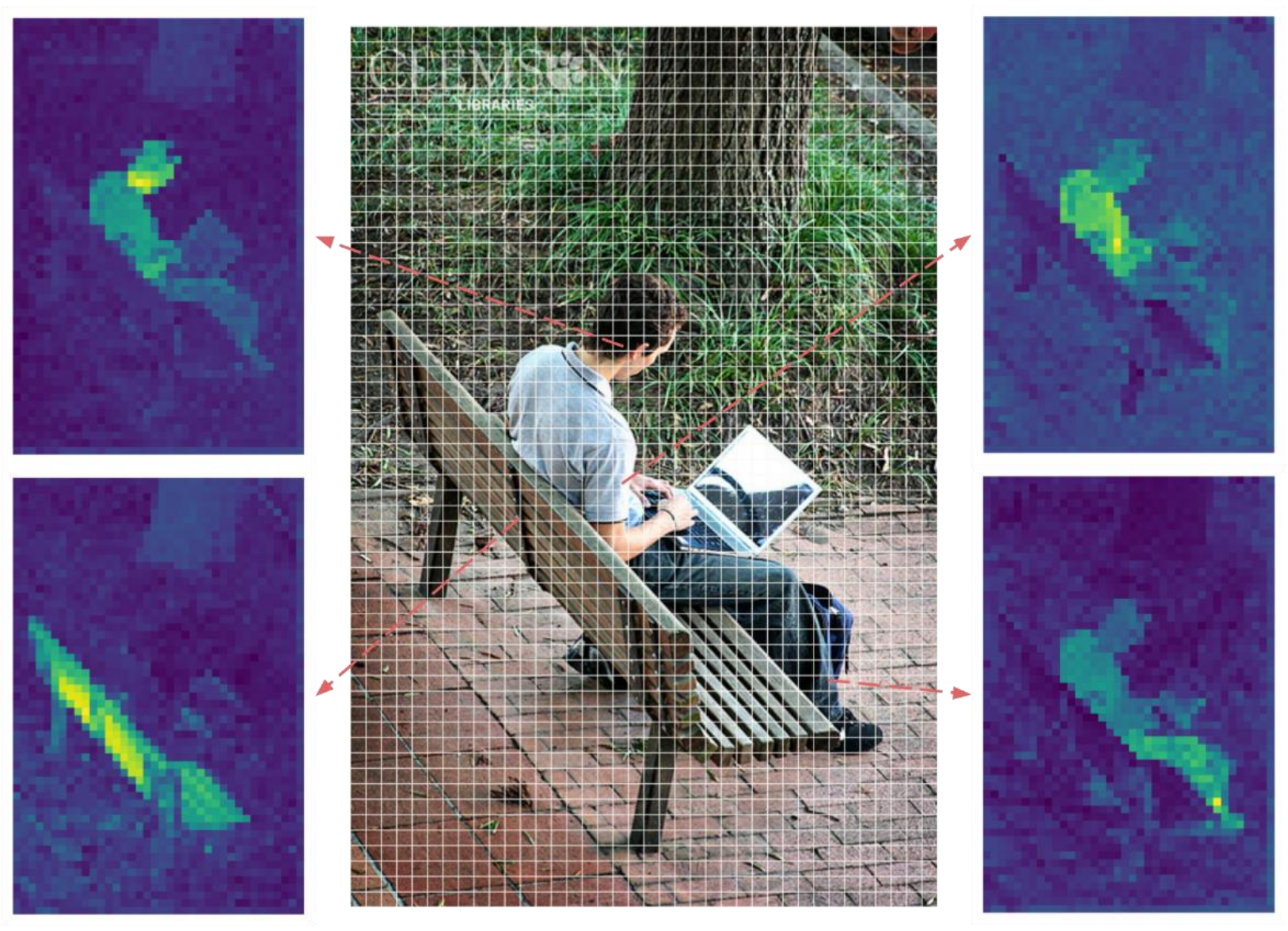
Code/Dataset/Poster at
github.com/Hosseinadeli/affinity_attention

Introduction

- **Spread of attention within objects** has been proposed as a mechanism for how humans group features to segment objects.
- However, such a mechanism has not yet been implemented and tested in naturalistic images.
- Here, we leverage the feature maps from self-supervised vision transformers and propose a model of human object-based attention spreading and grouping.

Affinity-based approach

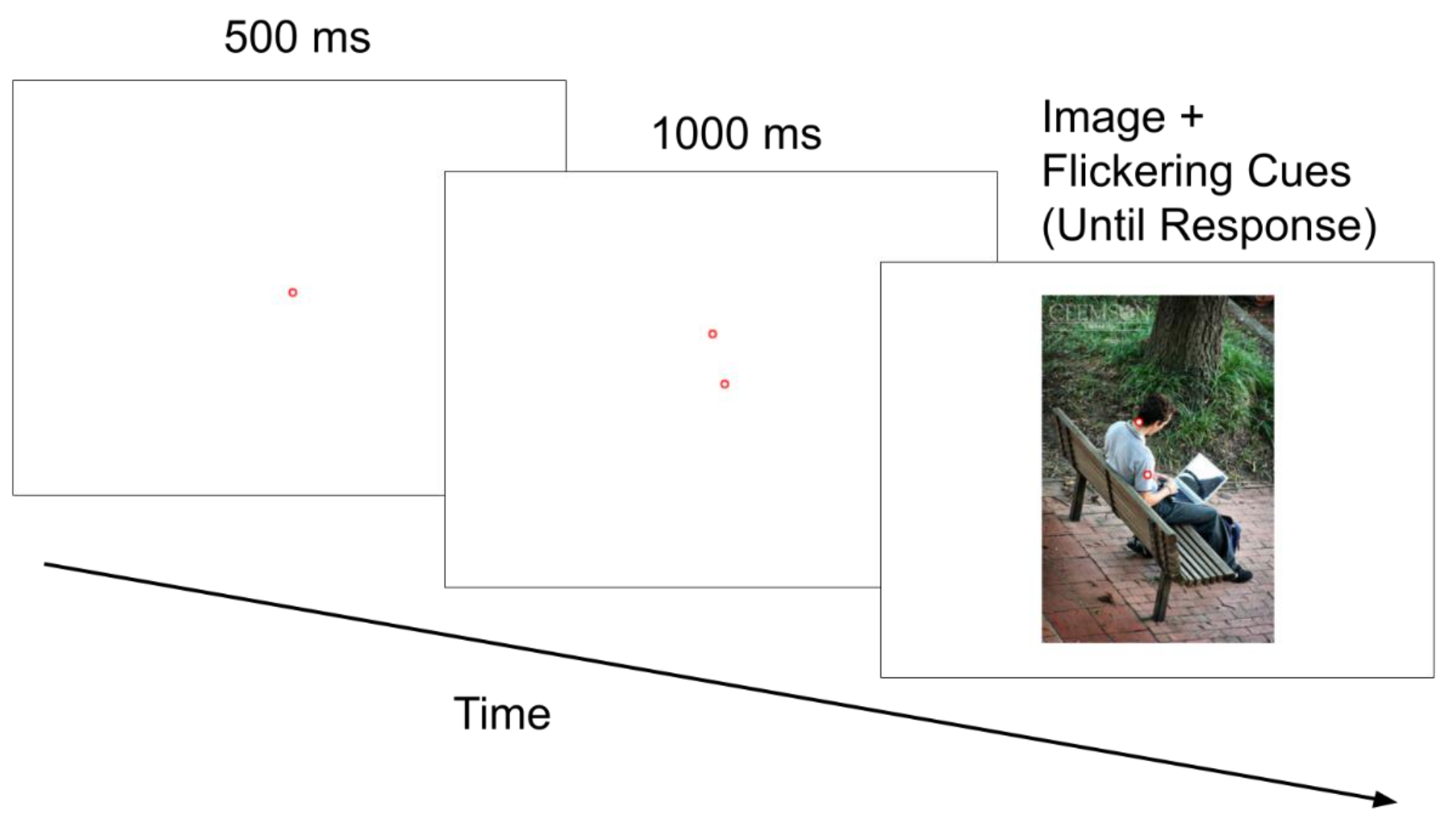
sample affinity maps for a few patches

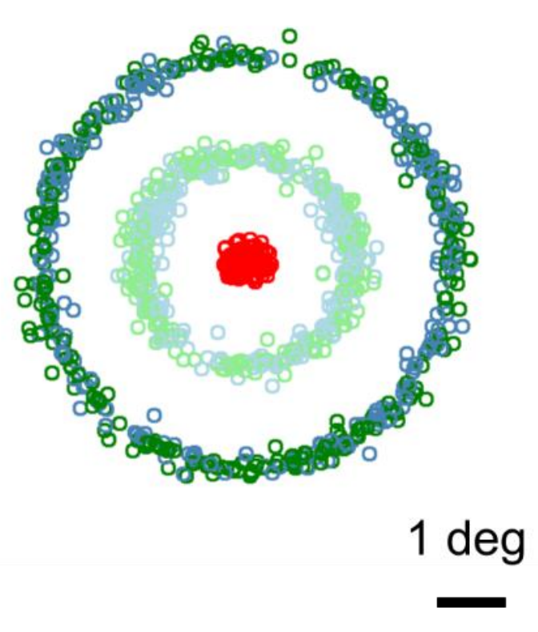


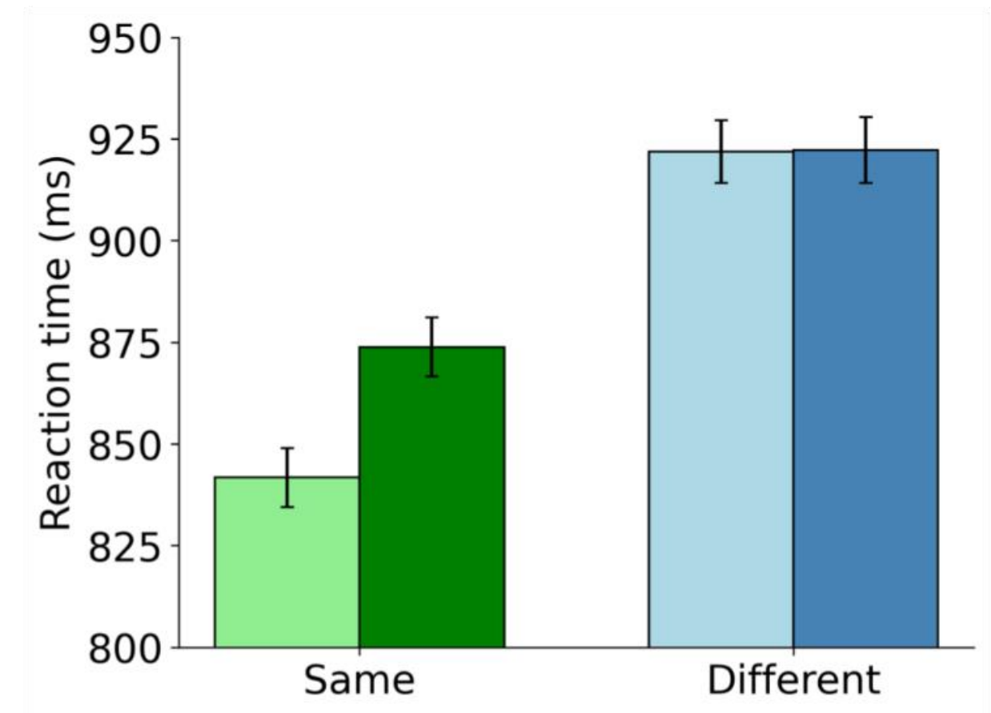
- Long-range horizontal connections in the retinotopic visual cortex link distant points of the visual input. They mediate the formation of maps with contextual connections between neurons, often referred to as association fields.
- **We model these contextual connections with affinity, based on the feature similarity between different patches of the image.**
- Features for patches are extracted from self-supervised transformers trained either on distillation (DINO) or reconstruction (MAE).
- These transformer models first divide the image into patches and then process them through multiple layers. By taking a dot product of one patch's features with all the other patches we can find the affinity map for all the patches.

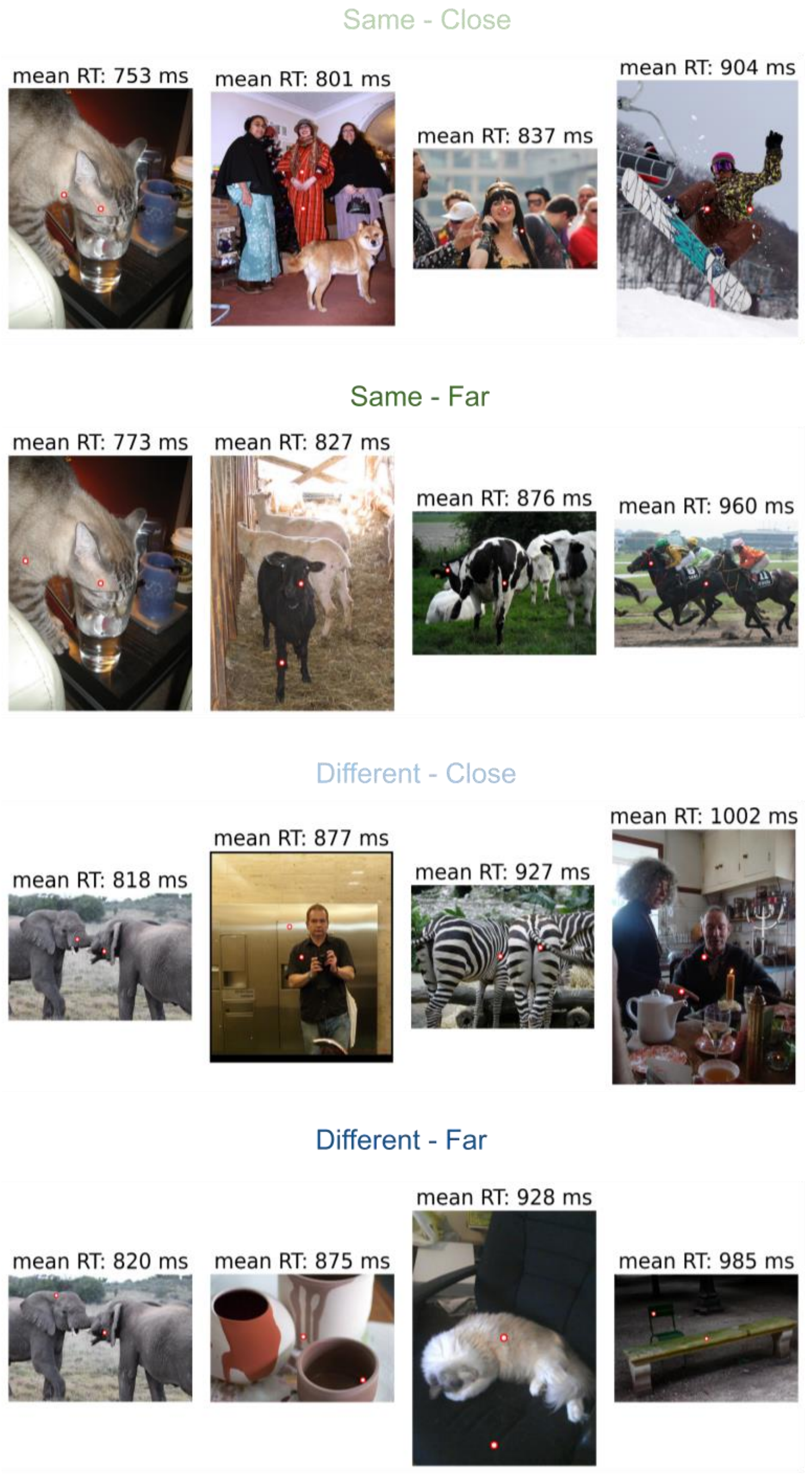
Behavioral Experiment

- 72 subjects
- 255 images
- 255 x 4 = 1020 unique trials
- While holding center fixation, subjects responded (by button press) whether the two dots are on the same or different objects (7% of trials removed due to breaking fixation)

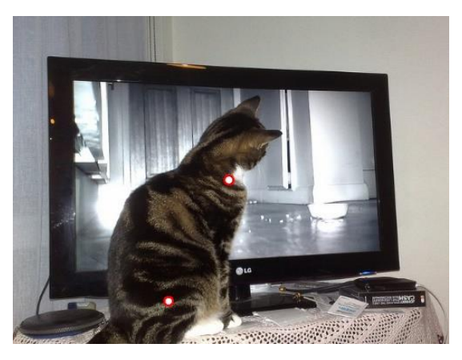




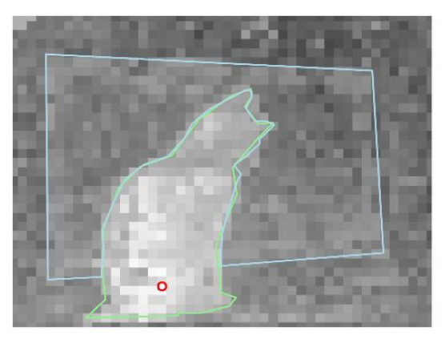




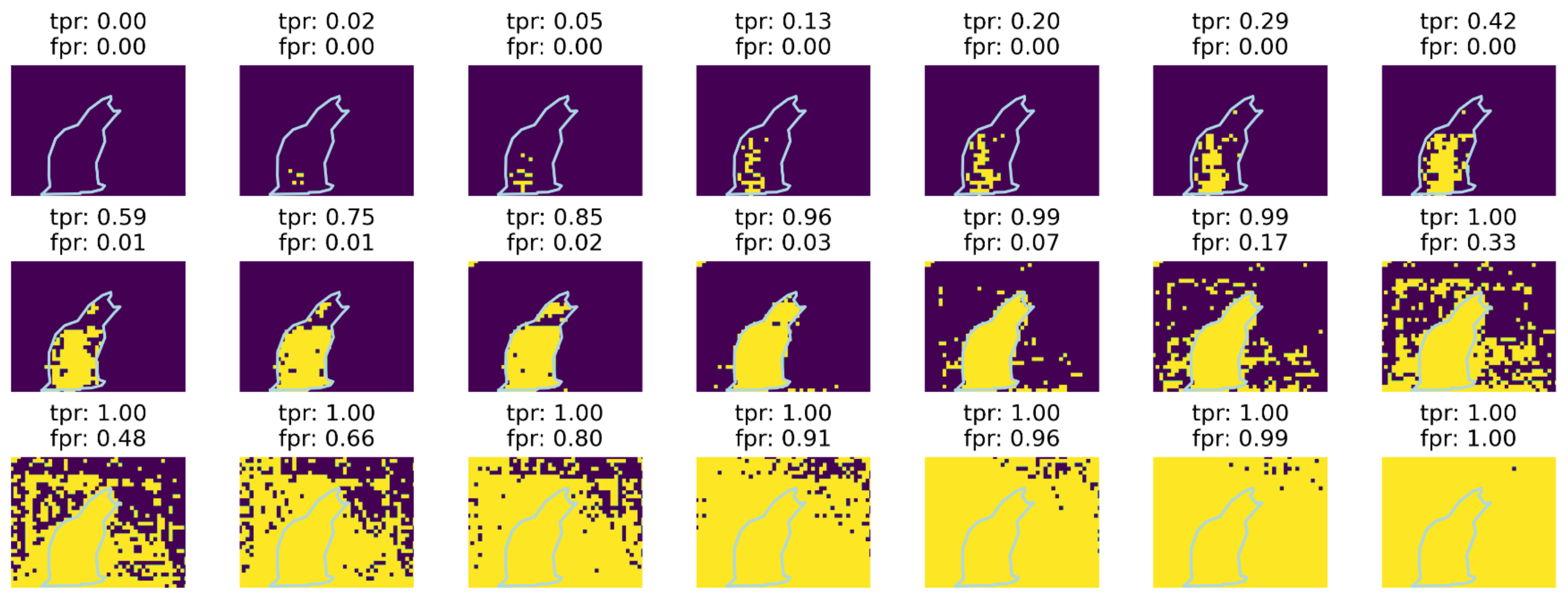
Measuring the object-centric component of Affinity

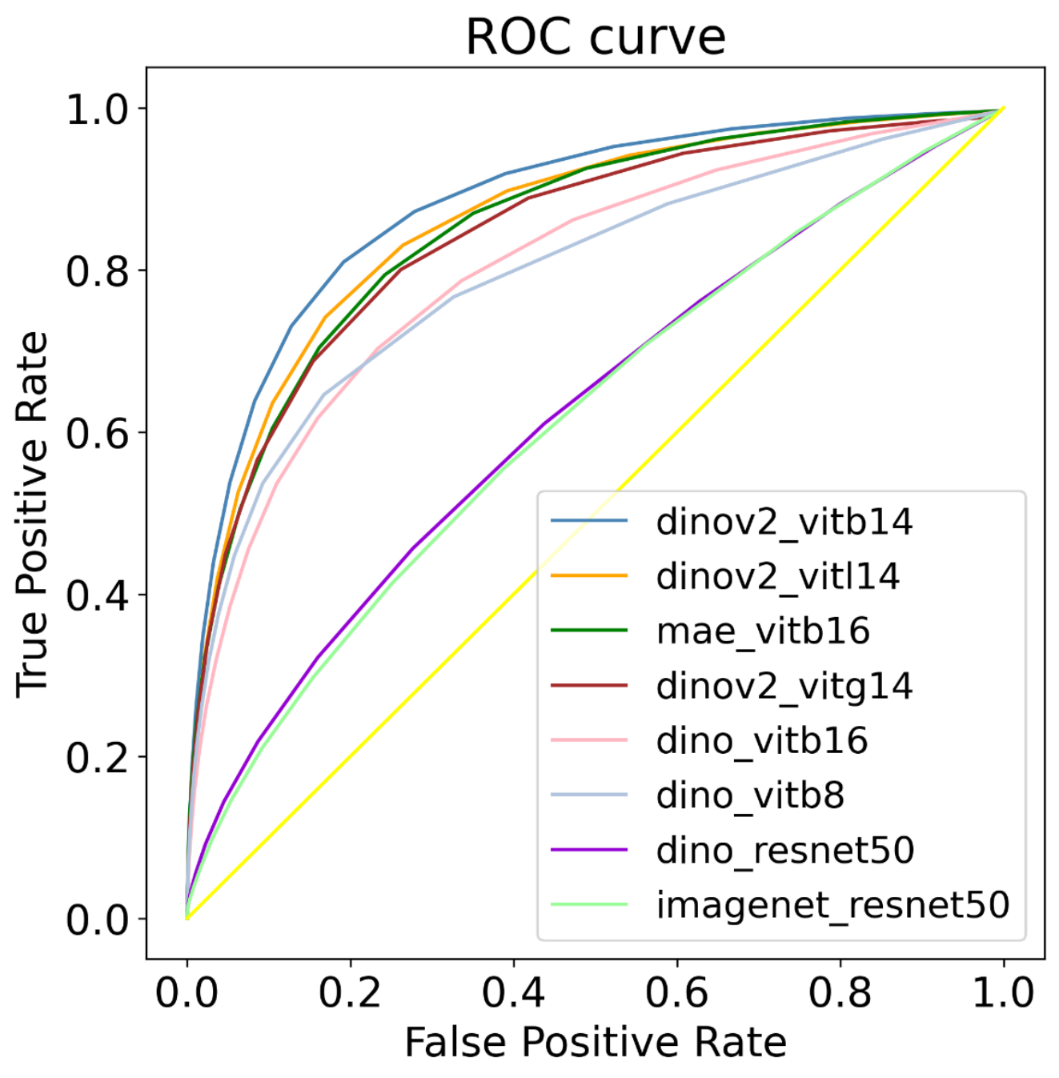


Affinity map for the peripheral dot overlaid on the ground-truth segmentation

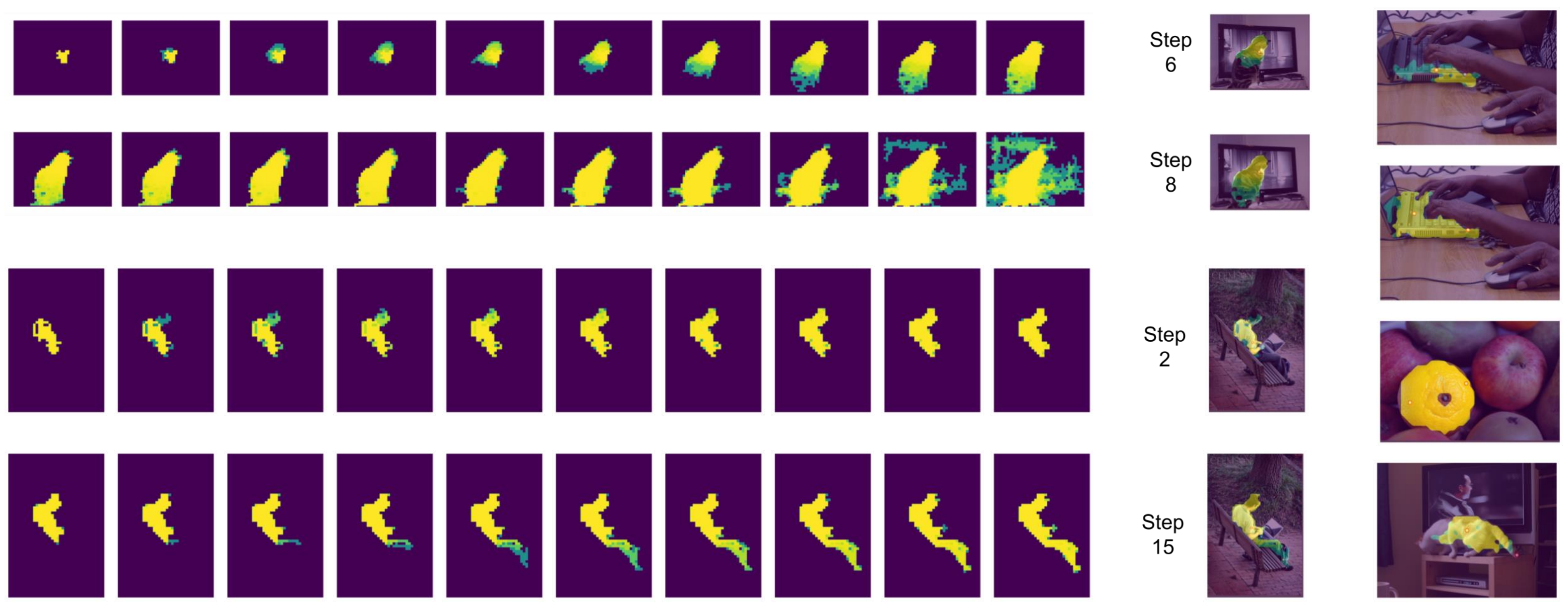


We evaluated different feature types (key, query, value, or conv) from models with different architectures (VIT and resnet) and sizes (base, large, and giant) using different patch sizes (8, 14, and 16) and with different training objectives (DINO, MAE, and recognition).

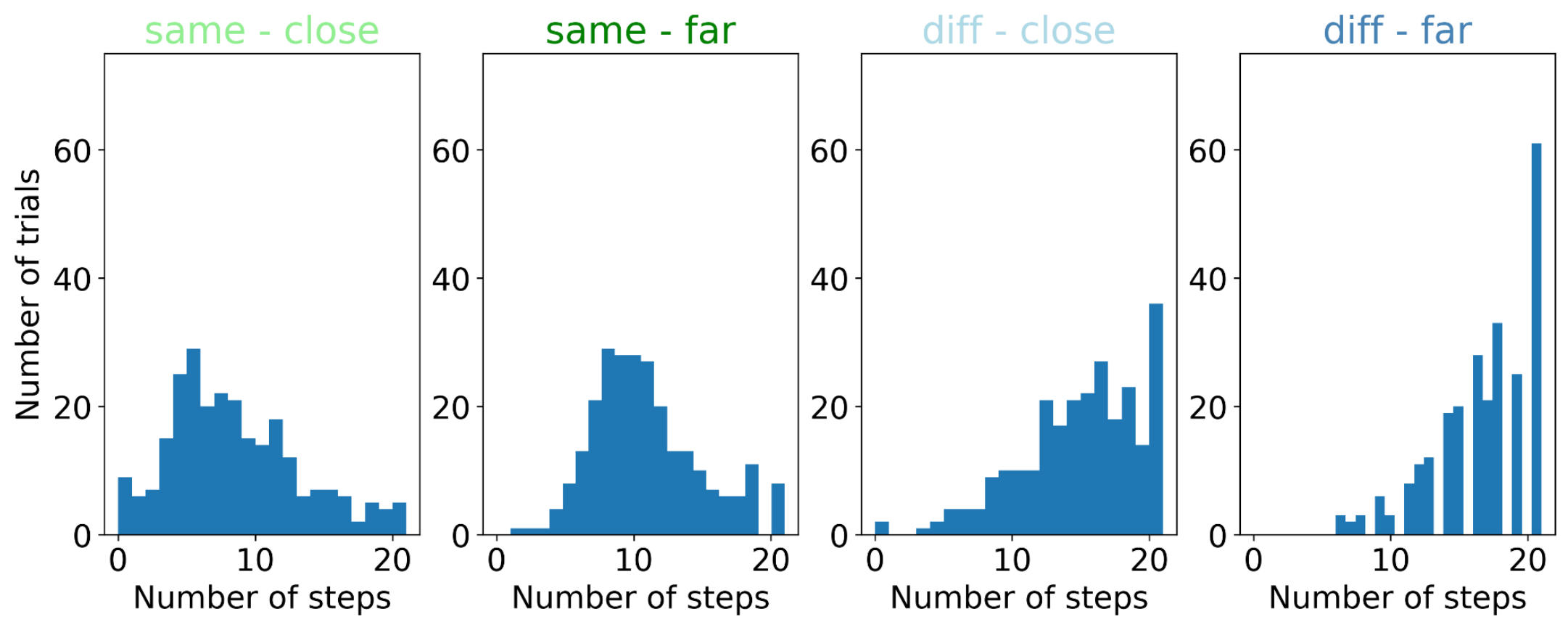


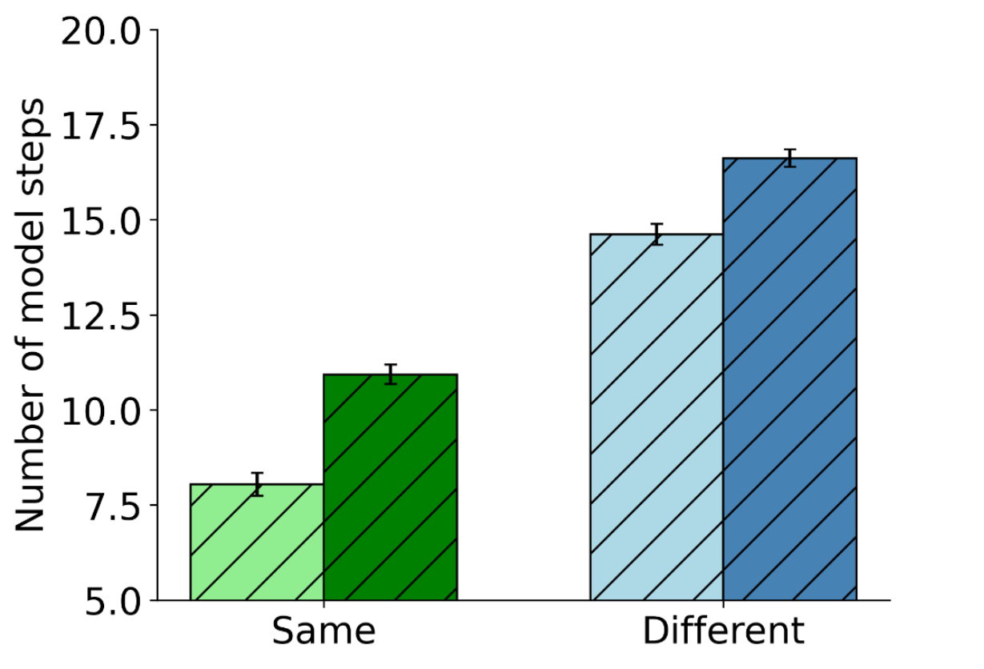


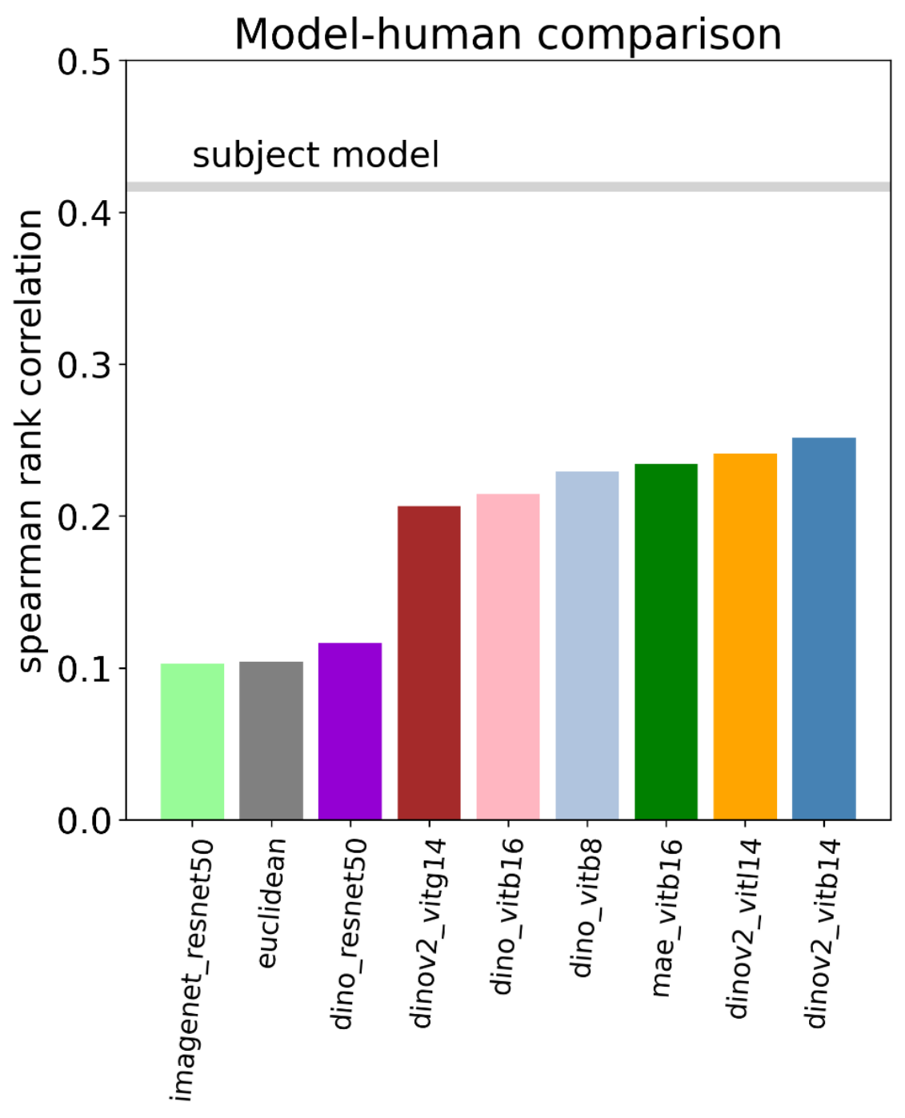
Affinity Spread



- Affinity spread models built on feature maps from the self-supervised transformers show significant improvement over baseline and CNN based models, despite not being trained on the same/different task or with any other object labels.
- Models with better object-centric affinity signals generally better predict human behavior in this task, a trend that we expect to continue.
- There is still a gap between model behavior and humans making this model comparison a useful benchmark for evaluating future developments.







Conclusions and Insights

- We provide **a mechanistic model of how visual grouping in humans is implemented as a recurrent process of spreading attention in natural scenes**:
  - ➢ The role of recurrence is to align the feature vectors of different patches (represented by different neural groups) with one another. The recurrent computation is driven by affinity, the neuronal groups that have similar representations (high affinity) are likely to excite one another. The result is that their vector representations would be more aligned.
  - ➢ Attention is believed to aid in segmentation by tagging the neurons that are likely to be on the same object with increased firing rate. Affinity based approach would posit that the vector representations for patches in each object would align with one another through affinity-based recurrent computation, tagging neuronal groups for each object.
- Our work demonstrates that transformers provide a plausible feature backbone for attention modulated perceptual grouping of features into objects. This extends their value as models of human vision beyond core object recognition.
- Our affinity spread method, building on self-supervised representation learning, does not require a large number of labeled samples for training, making this a more plausible mechanism for how the primate visual system learns to group features and perceive objects

References

- Roelfsema, P. R. (2023). Solving the binding problem: Assemblies form when neurons enhance their firing rate—they don’t need to oscillate or synchronize. Neuron, 111(7), 1003–1019.
- Caron, M., Touvron, H., Misra, I., J´egou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the ieee/cvf international conference on computer vision (pp. 9650–9660).
- Chen, H., Venkatesh, R., Friedman, Y., Wu, J., Tenenbaum, J. B., Yamins, D. L., & Bear, D. M. (2022). Unsupervised segmentation in real-world images via spelke object inference. In Computer vision–eccv 2022: 17th european conference, tel aviv, israel, october 23–27, 2022, proceedings, part xxix (pp. 719–735).
- He, K., Chen, X., Xie, S., Li, Y., Doll ´ ar, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 16000–16009).
- Jeurissen, D., Self, M. W., & Roelfsema, P. R. (2016). Serial grouping of 2d-image regions with object-based attention in humans. Elife, 5, e14320.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13 (pp. 740–755).

Acknowledgement

- Supported by NSF awards IIS-2123920 and IIS-1763981 and NIH award 1R01EY030669.