Orientation Bias and Abstraction in Working Memory: Evidence from Vision Models and Behavior

Fabio Bauer[1,2,*] and Or Yizhar[1,2] and Bernhard Spitzer[1,2]
[1]Group Adaptive Memory and Decision Making, Max Planck Institute for Human Development, Berlin, Germany
[2]Technische Universität Dresden, Dresden, Germany
[*]bauer@mpib-berlin.mpg.de

Abstract

Remembering visual orientations involves systematic biases in human working memory. We tested whether vision models exhibit analogous orientation biases when exposed to rotated real-life objects and if they represent orientations independent of object identity. Using representational similarity analysis (RSA), we compare the representational geometries from eight vision models and human behavioral reports to theoretical patterns of orientation encoding and bias. Our analysis differentiated between two representational domains: the 180° space (2-fold rotationally symmetric), and the 360° space (distinguishing 'up' vs. 'down'). To examine the extent to which orientation representations generalized, we compared artificial neural network (ANN) activations within- and between-objects. We found that vision models showed orientation encoding in 180° space and exhibited a pronounced attraction bias, unlike the characteristic repulsion effects observed in human participants. Further, the vision models display limited 360° orientation encoding, with inadequate cross-object generalization. In contrast, human working memory reports readily reflected orientations in 360° in a generalized fashion. Thus, while contemporary vision models can represent stimulus-specific orientation information, they fail to replicate abstract object-independent orientation encoding and bias that humans effortlessly achieve. Our findings underscore critical limitations of current vision models for studying visual working memory processing.

Introduction

During Visual Working Memory (WM) remembered orientations are repelled away from cardinal axes. These biases are well-documented for grating stimuli within 180° space (Tomassini, Morgan, & Solomon, 2010; de Gardelle, Kouider, & Sackur, 2010; Girshick, Landy, & Simoncelli, 2011; Appelle, 1972). WM maintenance of orientation information has been shown to involve lower-level visual loci (Chunharas, Hettwer, Wolff, & Rademaker, 2023; Dake & Curtis, 2025; Iamshchinina, Christophel, Gayet, & Rademaker, 2021; Sheehan & Serences, 2022; Duan & Curtis, 2024; Harrison & Tong, 2009). ANN models of vision have been used to emulate representations found in the ventral stream, offering insights into how humans process and represent visual information (Conwell, Prince, Kay, Alvarez, & Konkle, 2024; Schrimpf et al., 2018). Repulsion biases for grating stimuli also replicate in vision models and might originate from the prevalence of horizontal and vertical visual stimuli in the environment (Henderson & Serences, 2021). However, grating stimuli are symmetric and thus only cover 180° of rotational space. We recently found that cardinal repulsion bias extends to real-world objects in 360° orientation space even with stimuli of "orthogonal" spatial dimensions (e.g. table vs. tower)(Linde-Domingo & Spitzer, 2024; Yizhar, Bauer, Pont-Sanchis, Bröhl, & Spitzer, 2025). This indicates a high level of abstraction, where remembered orientations are generalized, independent of the object's physical characteristics. In this work, we want to explore if pretrained contemporary vision models are plausible computational frameworks for the described visual working memory phenomena. Our analyses address four questions: (i) Can deep vision models encode the orientation of real-world objects in 360° orientation (beyond visual gratings)? (ii) Do vision model representations show cardinal repulsion biases for rotated real-life objects, analogous to human orientation biases? (iii) Do the models' learned orientation representations generalize cross-objects?

Methodology

We analyzed the layer activations of current vision models used in cognitive neuroscience, including three brain-inspired CNNs (CORNet-S/RT/Z) (Kubilius et al., 2018), established feedforward CNNs (AlexNet, VGG19, ResNeXt-101) (Krizhevsky, Sutskever, & Hinton, 2017; Xie, Girshick, Dollár, Tu, & He, 2017; Simonyan & Zisserman, 2014), and two vision transformers (ViT-B, SLIP) (Dosovitskiy et al., 2021; Mu, Kirillov, Wagner, & Xie, 2021), viewing images of rotated objects (Fig 1a). Human behavioral reports were collected employing a retro-cued Visual Working Memory task, similar to Linde-Domingo (2024) with a free object rotation task and a subset of nine images (n=40). Using RSA (Kriegeskorte, 2008), we conducted two complementary analyses: the Within-Object analysis compared different orientations of the same object, creating RDMs for responses at various orientations, while the Between-Object analysis compared responses to different objects across multiple orientations. The resulting vision model and human behavioral RDMs were compared for similarity with theoretical templates representing 180° or 360° orientational space and biases.

Results and Conclusion

i) Weak 360° Orientation Representations in Vision Models: In the Within-Object analysis, we found that the tested vision models do form some representation of an object's full

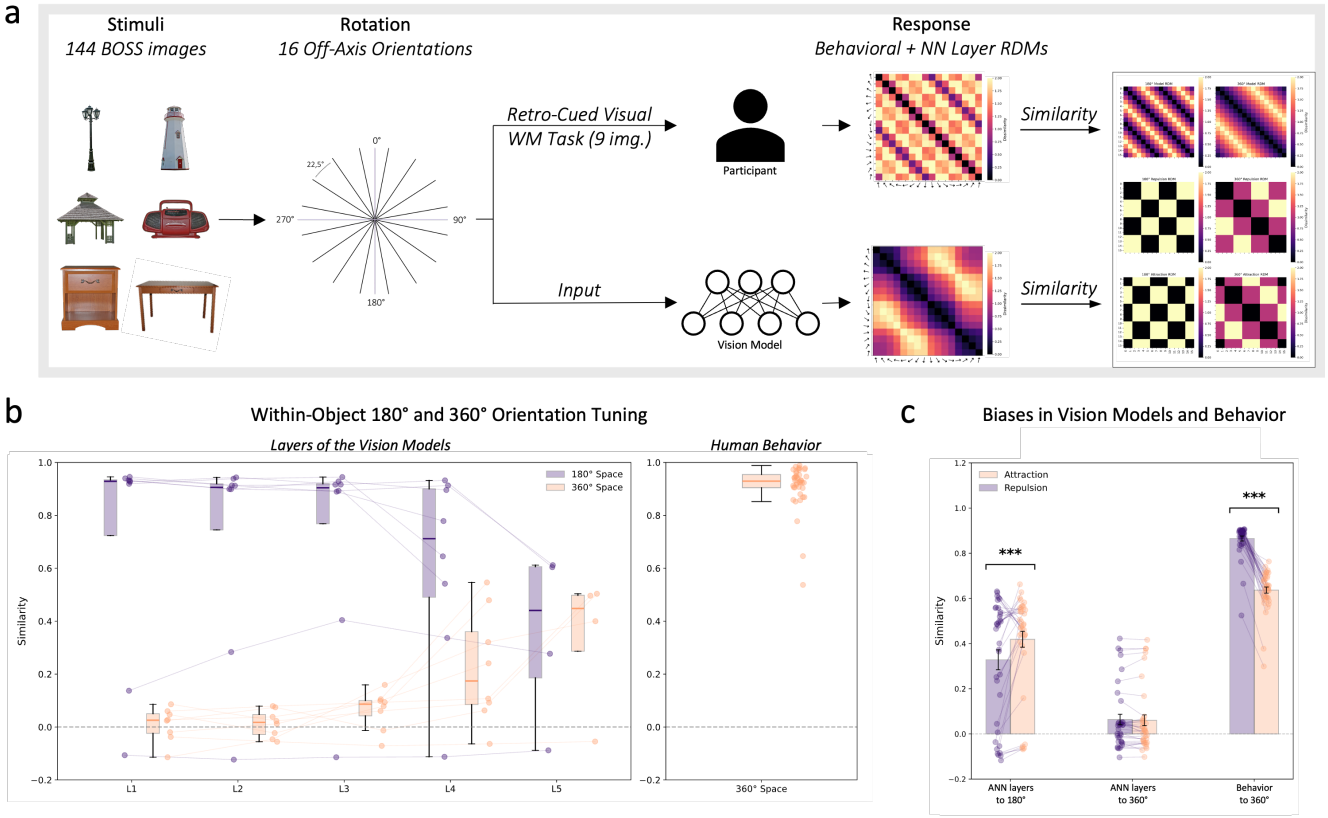

Figure 1: Data collection of vision model activations and human behavior, Orientation Space evaluation, and Bias evaluation. (a) Rotated real-life object images were presented to eight vision models (3 CNNs, 3 brain-inspired CNNs, 2 Vision Transformers) and 40 human participants. Resulting orientation-based RDMs were compared to six theoretical RDMs to determine whether orientation encoding was tuned for 180° or 360° and systematically biased. (b) Within-object analysis of similarity (Spearman correlation) between vision models' layer activation RDMs and theoretical RDMs shows 180° encoding in early to mid layers, shifting to 360° representation in late layers. Human behavior exhibits a 360° structure. (c) Within-Object attraction and repulsion bias across NN layers and human behavior for 180°/360° spaces, revealing systematic deviations from uniform encoding and opposing bias between models and human behavior.

orientation (0–360°). Still, this weak pattern only emerges in the deepest layers. Early to mid-layer representations in the models were invariant to a 180° flip (e.g. treating an object at 0° and 180° as similarly oriented regarding activations). Only the later layers carried any information distinguishing upright from inverted object views (Fig. 1b). Here, object-dependent similarities for the two orientation domains are similar, however only four of the eight models were deep enough to consider a fifth layer. Importantly, this finding pertains only to orientation-encoding for the same object (Within-Object analysis). The models' weak late-layer representations suggests that, unlike humans (Yizhar et al., 2025), vision models trained on object recognition do not preserve complete 360° orientation information throughout their hierarchy. **ii) Absence of Human-Like Orientation Biases:** Next, we asked whether vision models exhibit biases in their encoding of object orientation comparable to systematic human biases. We found no evidence of a 360° orientation bias, as expected if the vision models do not possess a strong 360° representation of object orientation. However, the 180° representation bias analysis revealed a robust preference for attraction over repulsion models, indicating an attraction bias rather than a repulsion bias to cardinals. This contrasts with human behavioral reports, where a 360° bias is characterized by a stronger repulsion from cardinal axes. Interestingly, the vision models' bias was stronger in Within-Object than in Between-Object comparisons. Suggesting that the model's systematic deviation from uniform encoding is tied to object identity.

iii) No Generalization of Orientation Encoding Across Objects: We investigated whether the 360° orientation information in the models is abstract enough to generalize across different objects. The difference between the Within- and the Between-Object analysis can be interpreted as the abstraction of orientation independent of object identity because the Within-Object analysis isolates the effect of orientation on the same object representation in contrast the between-object analysis examines how orientation changes that representation across different object identities. Within-object analysis shows a 2-fold rotationally symmetric 180° representation encoding in early to mid layers, shifting to 360° representation in late layers. The vision models' representational space significantly matches 180° encoding across all layers (between

objects), with no significant 360° representation. Interestingly, the 180° similarity is much weaker between objects than within, suggesting that orientation is entangled with object identity. Thus the models show a generalized "sense of direction" in 180° tied to object identity, but fail to display any abstract orientation representation across objects. In summary, our findings highlight a significant gap between human and vision model orientation encoding. While vision models demonstrate limited 360° orientation encoding and exhibit an attraction bias in 180° space, they fail to generalize across objects or replicate the human-like 360° repulsion bias. These results emphasize the limitations of current vision models in achieving robust, object-independent orientation encoding. This mismatch may impair the "out-of-the-box" utility of vision models as computational frameworks for abstraction and bias in human visual working memory.

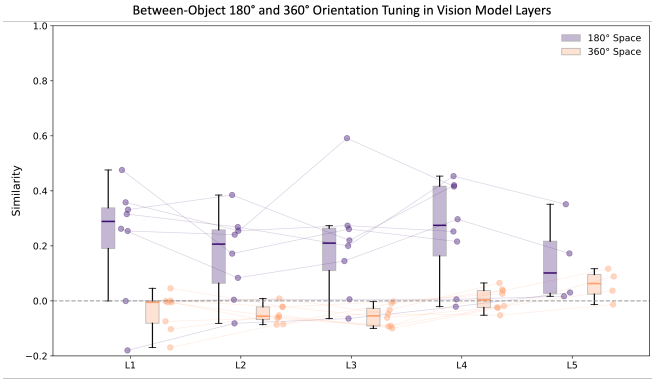

Figure 2: Between-Object analysis of similarity: Spearman correlation between vision models' layer activation RDMs and orientation RDMs show strong alignment with 180° encoding across layers.

Author Note

F.B. was supported by the International Max Planck Research School on Computational Methods in Psychiatry and Ageing Research (COMP2PSYCH, https://www.mps-ucl-centre.mpg.de/comp2psych; participating institutions: Max Planck Institute for Human Development, University College London) as a pre-doctoral fellow.

References

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, *78*(4), 266–278. doi: 10.1037/h0033117

Chunharas, C., Hettwer, M. D., Wolff, M. J., & Rademaker, R. L. (2023). A gradual transition from veridical to categorical representations along the visual hierarchy during working memory, but not perception. *bioRxiv*. doi: 10.1101/2023.05.18.541327

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, *15*(1), 9383. doi: 10.1038/s41467-024-53147-y

Dake, M., & Curtis, C. E. (2025, March). Perturbing human v1 degrades the fidelity of visual working memory. *Nature Communications*, *16*(1). Retrieved from http://dx.doi.org/10.1038/s41467-025-57882-8 doi: 10.1038/s41467-025-57882-8

de Gardelle, V., Kouider, S., & Sackur, J. (2010). An oblique illusion modulated by visibility: Non-monotonic sensory integration in orientation processing. *Journal of Vision*, *10*(10), 6. doi: 10.1167/10.10.6

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations*.

Duan, Z., & Curtis, C. E. (2024). Visual working memories are abstractions of percepts. *eLife*, *13*, RP94191. doi: 10.7554/eLife.94191.3

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*, 926–932. doi: 10.1038/nn.2831

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635. doi: 10.1038/nature07832

Henderson, M., & Serences, J. T. (2021, August). Biased orientation representations can be explained by experience with nonuniform training set statistics. *Journal of Vision*, *21*(8), 10. Retrieved from http://dx.doi.org/10.1167/jov.21.8.10 doi: 10.1167/jov.21.8.10

Iamshchinina, P., Christophel, T. B., Gayet, S., & Rademaker, R. L. (2021). Essential considerations for exploring visual working memory storage in the human brain. *Visual*

Cognition, *29*(7), 425–436. doi: 10.1080/13506285.2021.1915902

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. Retrieved from http://dx.doi.org/10.3389/neuro.06.004.2008 doi: 10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. doi: 10.1145/3065386

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *bioRxiv*. doi: 10.1101/408385

Linde-Domingo, J., & Spitzer, B. (2024). Geometry of visuospatial working memory information in miniature gaze patterns. *Nature Human Behaviour*, *8*(2), 336–348. doi: 10.1038/s41562-023-01737-z

Mu, N., Kirillov, A., Wagner, D., & Xie, S. (2021). Slip: Self-supervision meets language-image pre-training. *arXiv*. doi: 10.48550/arXiv.2112.12750

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2018, September). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*. Retrieved from http://dx.doi.org/10.1101/407007 doi: 10.1101/407007

Sheehan, T. C., & Serences, J. T. (2022). Attractive serial dependence overcomes repulsive neuronal adaptation. *PLOS Biology*, *20*(9), e3001711. doi: 10.1371/journal.pbio.3001711

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv*. doi: 10.48550/arXiv.1409.1556

Tomassini, A., Morgan, M. J., & Solomon, J. A. (2010). Orientation uncertainty reduces perceived obliquity. *Vision Research*, *50*(5), 541–547. doi: 10.1016/j.visres.2009.12.005

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (pp. 1492–1500). Honolulu, HI. doi: 10.1109/CVPR.2017.634

Yizhar, O., Bauer, F., Pont-Sanchis, I., Bröhl, F., & Spitzer, B. (2025). *Abstracted representation of object orientation during working memory in early visual cortex.* (Manuscript in preparation)