Cognitive and Computational Models of Face Recognition

Chen Damoze

Introduction

A fundamental principle of the human visual system is its hierarchical organization, consisting of distinct cortical pathways for processing different aspects of visual information. Processing begins in the early visual cortex, which extracts elementary features such as edges and orientations from the retinal input (Hubel & Wiesel, 1968). From there, information is processed along two primary streams. The ventral visual stream progresses from simple feature combinations in posterior areas to full object representations in anterior regions, supporting object recognition (Grill-Spector & Weiner, 2014). In parallel, the dorsal visual stream extends to the parietal lobe and is specialized for processing spatial information, motion, and guiding actions (Ayzenberg & Behrmann, 2022; Goodale & Milner, 1992). This architecture shows a progression from simple visual features in early stages to more complex, conceptual information in higher-level areas (Felleman & Van Essen, 1991), raising the question of whether high-level visual cortex is organized according to visual-perceptual features or abstract semantic categories.

Recent studies showed that large language models (LLMs) could predict neural responses to visual images (Conwell et al., 2024; Doerig et al., 2022) but left unclear whether the visual cortex aligns with visually descriptive or abstract conceptual language. Shoham et al. (2024) distinguished between visual text describing the visual content of an image and abstract text referring to non-visual conceptual properties, finding that the high-level visual cortex was explained by visual, not abstract, representations. While they focused on familiar people and places using iEEG, the current study extends this framework by examining whether similar distinctions hold across a broader set of visual categories, including unfamiliar objects, and by employing both fMRI and MEG to capture complementary spatial and temporal dynamics.

Accordingly, we asked whether the organization of the high-level visual cortex representations reflects visual or abstract properties. We hypothesized that this organization is primarily grounded in visual features rather than abstract information. To test it, we used Representational Similarity Analysis (RSA) to compare fMRI and MEG data with three models: DCNN-Image, LLM-Visual Text, and LLM-Abstract Text. We hypothesized that in the ventral stream, both DCNN-Image and LLM-Visual Text would show stronger correlations than LLM-Abstract Text. The dorsal stream was expected to align primarily with DCNN-Image, reflecting its spatial processing role (Vossel et al., 2014).

Methods

Dataset

The dataset consisted of 92 object images obtained from the study by Cichy et al. (2016). Each image presented a single object on a uniform white background. The set included faces, bodies, organs, places, animals, plants, and inanimate items.

Neural Data

We used publicly available neural data from the study by Cichy et al. (2016), combining MEG and fMRI to characterize the spatio-temporal dynamics of visual object processing. The dataset included recordings from healthy adult participants, 16 in the MEG experiment and 15 in the fMRI experiment (10 female; M = 25.87, SD = 5.38). Each participant viewed the same set of real-world object images while performing a simple vigilance task to ensure attention.

The MEG data provided millisecond-level temporal resolution, capturing time-resolved neural responses to each visual stimulus. These responses were summarized as Representational Dissimilarity Matrices (RDMs), which reflect the dissimilarity between neural activation patterns elicited by different images over time.

fMRI scans were used to estimate neural dissimilarity patterns across cortical voxels and regions of interest (ROIs) in occipital and temporal cortices. Representational dissimilarity was defined as the degree of pattern difference between voxel activation profiles for each image pair. For further details on the neural data, see Cichy et al. (2016).

To characterize the spatial extent of brain-model correspondence, ROIs were defined using the probabilistic atlas from the study by Wang et al. (2015) and grouped into three large-scale visual networks. Early visual cortex (contained V1d, V1v, V2d, V2v, V3d, and V3v), Ventral visual regions (contained LO1, LO2, VO1, VO2, PHC1, PHC2, hV4), and Dorsal visual regions (contained IPS0-IPS5, SPL1, FEF, V3a, V3b, MST, hMT).

Computational Representations

We constructed three types of model-based representations: visual (images), visual-text, and abstract-text.

Visual representations

Visual embeddings were extracted from each image using the CLIP model. Each image was encoded into a high-dimensional feature vector capturing its visual characteristics. Pairwise dissimilarities between all images were calculated as (1-cosine similarity), forming a visual RDM.

Visual-text representations

For each image, a detailed textual description of its visual features was generated through an iterative optimization process. An initial visual description was generated using the GPT-4o model with a prompt to focus only on perceptual details (for example, shapes, colors) while avoiding identity or context. This process was repeated for 10 iterations. In each step, feedback was provided to the model based on two metrics: the CLIP similarity between the generated text and the image, and the SGPT similarity between the generated text and the abstract-text description (see below). The goal of the feedback was to guide the model toward generating descriptions that were more visually aligned with the image while being semantically distinct from the abstract description. From the 10 generated descriptions for each image, the one with the highest CLIP score was selected for the final set.

Abstract-text representations

To capture semantic-level representations, we collected short textual descriptions from external sources such as Wikipedia or dictionary entries corresponding to the entities or objects shown in the images. For faces, the descriptions focused on emotional expressions. These texts represent conceptual knowledge rather than direct visual information.

Textual embeddings for both the visual-text and abstract-text descriptions were obtained using the SGPT model.
For each set of embeddings, an RDM was computed as (1-cosine similarity). Each of the three RDMs (Visual, Visual-text, Abstract-text) was then vectorized by extracting its upper triangular portion, resulting in a single dissimilarity vector per model.
Five images whose visual sentences included self-referential elements in every iteration during the process of creating visual text were excluded from the analysis. Self-referential responses occurred when the model failed to generate a visual sentence.

Data Analysis

Relationships between computational representations

To examine the structural correspondence between the three representational domains (visual, visual-text, and abstract-text), we calculated Pearson correlations between their vectorized RDMs. Zero-order correlations were computed to measure the overall similarity between any two representations. Partial correlations were used to assess the unique association between two representations while controlling for the influence of the third.

Brain-Representational Similarity Analysis

To link computational models with neural data, we applied Representational Similarity Analysis (RSA) separately to the fMRI and MEG datasets. We measured brain-model correspondence by computing Pearson correlations between the vectorized upper triangle of each neural RDM (from fMRI or MEG) and the vectorized RDM from each of the three computational models. For the fMRI data, this resulted in a correlation value for each participant, ROI, and model. In addition to zero-order correlations, partial correlations were computed to estimate the unique variance explained by each model while controlling for the other two models. For the MEG data, time-resolved RDMs were calculated. The correlation was computed between each model and the MEG RDM at each millisecond, resulting in a time course of brain-model similarity for each participant.

Statistical testing

Statistical significance for the difference between model correlations was assessed using the two-sided Wilcoxon signed-rank test. For the fMRI data, comparisons between two models were conducted within each of the three large-scale visual networks (Early, Ventral, Dorsal). For each participant, the correlation values (R) for all ROIs within a given network were averaged, creating a single mean correlation value per participant for that network. The two-sided Wilcoxon signed-rank test was then performed on these paired, participant-level mean values. In addition, a one-sample Wilcoxon signed-rank test was then applied to these participant-level mean values to assess whether correlations were significantly above zero. For the MEG data, tests were computed across participants independently at each time point. All resulting p-values from these multiple comparisons were corrected using the False Discovery Rate (FDR) procedure, and were considered a significant preference only when $p<0.05$.

Results

Comparing Representational Similarity Across Image, Visual-Text, and Abstract-Text DNNs

We examined the correspondence between the three RDMs (DCNN-Image, LLM-Visual, and LLM-Abstract) to determine whether the image-based representations were more closely aligned with the visual-text representations than with those derived from abstract descriptions. At the zero-order level, all correlations were significant. The correlation between visual-text and abstract-text representations (r = .69) was higher than the

correlation between image and visual-text representations (*r* = .60) and the correlation between image and abstract-text representations (*r* = .57). These results suggest that abstract descriptions capture some, but not all, of the visual structure. In addition, it indicates that linguistic representations are strongly related to each other. The partial correlation between visual text and abstract text remained relatively high (*r* = .52), indicating a strong link even when the images were controlled for. As shown in Figure 1, the unique correspondence between images and visual text descriptions was moderate (r = .36), while images and abstract text descriptions showed a lower unique correlation (*r* = .26). This suggests that visual text captures more of the unique visual structure than abstract text does.

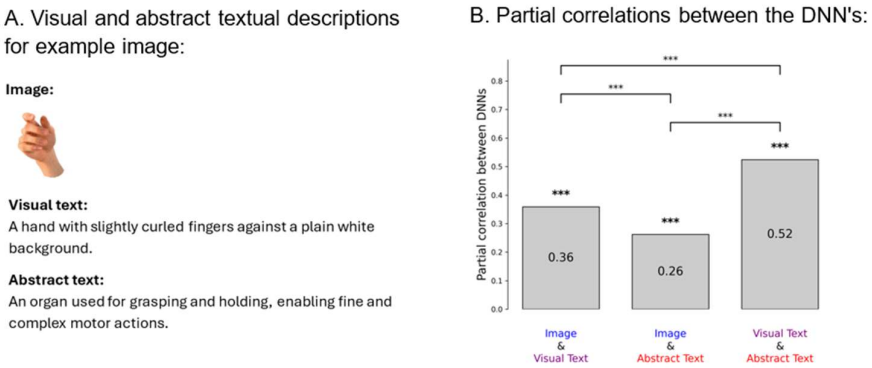

Figure 1: Examples of an image and its visual and abstract textual description, and the partial correlations between visual and language DNNs representations*:* A. Example of an image with its visual and abstract textual descriptions. B. A pre-trained DCNN (CLIP) was used to extract image embeddings (DCNN-image), and a large language model (SGPT) was used to extract embeddings of the visual and abstract textual descriptions.

Representational Spaces of the Computational Models

To visualize the representational structure of each model, we used Uniform Manifold Approximation and Projection (UMAP). The resulting 2D UMAP plots illustrate how different categories of objects are organized in each model's representational space (see Figure 2). The UMAP results suggest that the three representations capture different aspects of image similarity. The Abstract Text and Image representations appear to organize images into discrete, category-based clusters (for example, grouping all animals), leading to high cluster agreement. The Visual Text representation, while highly correlated with both other models

(see Figure 1), captures more fine-grained, continuous visual features that cut across categorical boundaries.

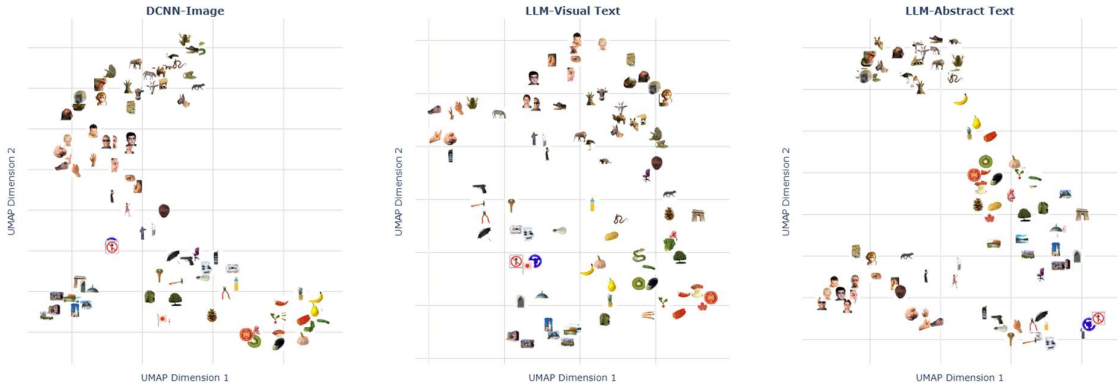

Figure 2: Uniform Manifold Approximation and Projection (UMAP) of Object Representations Across Three Models. The figure displays two-dimensional UMAP projections for 87 objects, illustrating the similarity structure within three distinct embedding spaces. The left panel shows the projection based on image embeddings. The middle panel is based on embeddings of visual descriptions. The right panel is based on embeddings of abstract descriptions. All embeddings were generated using the CLIP text encoder.

fMRI Results

Our analysis of fMRI data revealed patterns of model performance across different cortical networks. We calculated both zero-order and partial correlations for each model within three pre-defined cortical networks: the ventral, the dorsal, and the early visual cortex (EVC). For further details, see methods. Only statistically significant results were reported, and non-significant findings regarding paired tests were excluded.

Early Visual Cortex

In the zero-order correlation analysis, the LLM-Abstract Text model showed a significant correlation with EVC activity (r = .03, $pFDR$ = .023). In contrast, the DCNN-Image and LLM-Visual Text models did not show significant correlations ($pFDR$ = .974 and $pFDR$ = .075, respectively). Paired comparisons revealed that the correlations of both linguistic models (Visual Text and Abstract Text) were significantly stronger than the DCNN-Image model ($pFDR$ = .012 for both comparisons). In the partial correlation analysis, only the LLM-Abstract Text model showed a significant correlation (r = .04, $pFDR$ = .012). The unique

contributions of the DCNN-Image and LLM-Visual Text models were not significant (*pFDR* = .999 and *pFDR* = .348, respectively). Paired comparisons showed that the unique contributions of both text-based models were significantly stronger than those of the DCNN-Image model (Images vs. Visual Text: *pFDR* = .023; Images vs. Abstract Text: *pFDR* = .012). (see Figure 3)

Ventral Visual Cortex

The zero-order correlation analysis indicated that all three models had a significant correlation with brain activity, DCNN-Image (*r* = .03, *pFDR* = .012), LLM-Visual Text (*r* = .04, *pFDR* = .005), and LLM-Abstract Text (*r* = .04, *pFDR* = .005). A paired comparison showed that the correlation of the LLM-Visual Text model was significantly stronger than the DCNN-Image model (*pFDR* = .023). In the partial correlation analysis, only the LLM-Visual Text model had a significant unique contribution (*r* = .02, *pFDR* = .005). The DCNN-Image and LLM-Abstract Text models did not have significant unique contributions (*pFDR* = .857 and *pFDR* = .051, respectively). The paired comparison showed a significant difference only between the DCNN-Image and LLM-Visual Text models (*pFDR* = .025), indicating a higher unique contribution from the LLM-Visual Text model. (see Figure 3)

Dorsal Visual Cortex

All three models showed significant zero-order correlations with brain activity. DCNN-Image (*r* = .02, *pFDR* = .012), LLM-Visual Text (*r* = .02, *pFDR* = .012), and LLM-Abstract Text (*r* = .01, *pFDR* = .046). In the partial correlation analysis, only the DCNN-Image model maintained a significant unique correlation with brain activity (*r* = .02, *pFDR* = .017). The unique contributions of the LLM-Visual Text and LLM-Abstract Text models were not significant (*pFDR* = .081 and *pFDR* = .925, respectively). (see Figure 3)

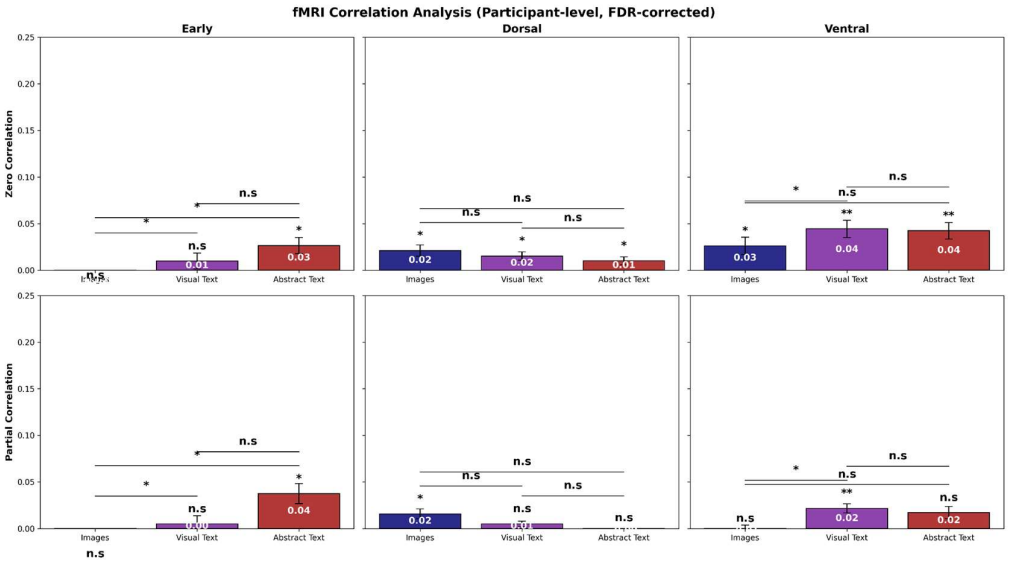

Figure 3: Correlations between fMRI activity patterns and computational model representations across three visual networks. The figure shows correlations between the representational dissimilarity matrices (RDMs) of brain activity in the Early, Dorsal, and Ventral visual networks and the RDMs of three models: a visual DNN (DCNN-Image), visual descriptions (LLM-Visual Text), and abstract descriptions (LLM-Abstract Text). The top panels display the zero-order correlations for each model RDM with the brain activity RDMs. Bottom panels display the partial correlations, reflecting the unique correlation of each model's RDM with brain activity patterns while controlling for the other two models. Error bars indicate the standard error of the mean (s.e.m) across participants. All reported p-values are FDR corrected.* pFDR<0.05, ** pFDR<0.01, *** pFDR<0.001 (Statistical details are described in the Methods section).

MEG Results

The MEG analysis revealed the temporal dynamics of how these representations mapped onto neural activity. The zero-order correlations for all three models began to rise shortly after stimulus onset (0 ms), peaked at approximately 150-200 ms, and then gradually declined (see Figure 4). The Visual Text model showed the strongest correlation with neural patterns, reaching a peak correlation (r = .14) at 170 ms. As shown in Figure 4, the Image model and the Abstract Text model showed lower peak correlations (r = .11 and r = .10, respectively). Statistical comparisons revealed several findings. The Visual Text model showed a significantly stronger correlation than the Abstract Text model in sustained time windows from approximately 116 ms to 560 ms, and again around 716-790 ms. The Image model showed a significantly different correlation than the Abstract Text model in a

sustained window from approximately 58 ms to 83ms. The Visual Text model showed a significantly stronger correlation than the Image model, primarily in windows from approximately 150 ms to 230 ms and around 245-400 ms, while the Image model showed a significantly stronger correlation in a later window (1150-1170 ms). The results indicated that the Visual Text representation showed the strongest overall correlation with neural activity throughout most of the session. The Image-based representation showed advantages over Abstract Text early in visual processing (starting around 60 ms). The Visual Text model surpassed both other models during peak processing times (150-230 ms) and maintained superiority over Abstract Text for an extended period (see Figure 4).

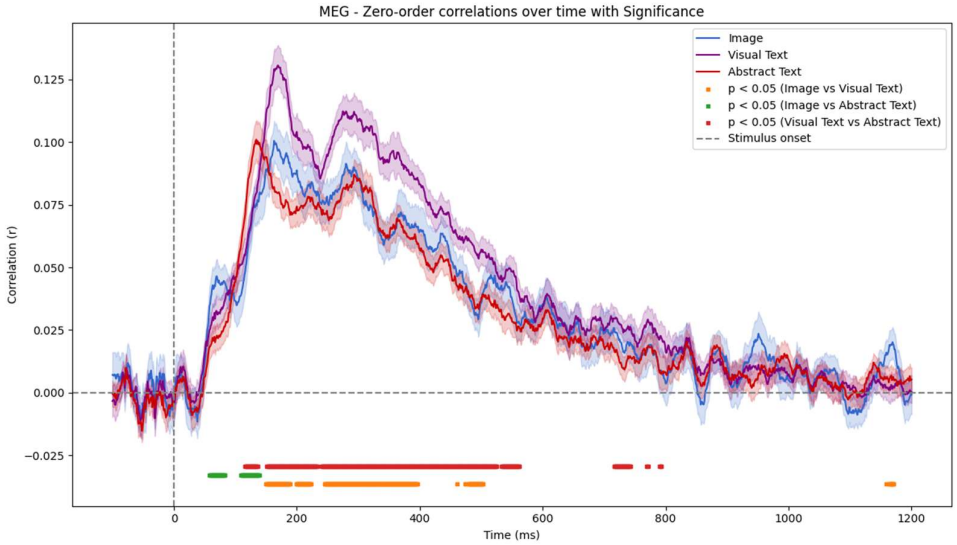

Figure 4: Temporal dynamics of model correlations with MEG data. Correlation values are plotted across time (ms), with the vertical dashed line at 0 ms indicating stimulus onset. Colored lines represent the three models: Image (blue), Visual Text (purple), and Abstract Text (red). Shaded areas indicate standard errors of the mean. Colored horizontal bars below the x-axis denote time windows with significant differences between model pairs: Image vs Visual Text (orange), Image vs Abstract Text (green), and Visual Text vs Abstract Text (red).

Discussion

The current study examined whether the organization of the high-level visual cortex representations reflects visual or abstract principles. This distinction was first demonstrated at the computational level, where we found that image-based representations (DCNN-Image) shared more unique variance with visual-text descriptions than with abstract ones. Consistent with our hypothesis, the fMRI findings showed that the ventral visual stream

aligned more with visual-descriptive representations than with abstract ones, and the dorsal stream aligned more with visual than with visual-descriptive representations. Although the overall correlations were modest, the results were significant. Only the LLM-Visual Text model showed a significant unique contribution to ventral stream activity after controlling for the other two models. In contrast, and in line with its role in spatial processing (Freud et al., 2016), only the DCNN-Image model maintained a significant unique correlation with dorsal visual stream activity. These findings support the claim that the advantage of the visual-text model is not a global effect but specifically reflects the ventral cortex's role in object recognition, emphasizing its integration of perceptual and linguistic information.

The MEG findings provide temporal support for this conclusion. The LLM-Visual Text model not only showed the highest overall correlation, but it also performed significantly better than the other two models during the central processing window for objects. This time window is associated with peak activity in the ventral stream (Cichy et al., 2014), indicating that the ventral stream favors visual-text over abstract during object processing. The UMAP analysis helps explain why the visual-text model correlates better with neural activity in the ventral stream. While the abstract text model organized objects into discrete, category-based clusters, the visual text model organized them based on continuous features that cut across categorical boundaries. This continuous structure aligns with views that the ventral stream represents rich, continuous visual features rather than just abstract categorical labels (Baldassi et al., 2013).

In conclusion, our results show that the high-level visual cortex, particularly the ventral stream, aligns with visually grounded linguistic representations rather than abstract conceptual descriptions, which highlight the importance of distinguishing between visual and abstract linguistic representations. These findings extend the work of Shoham et al. (2024), who observed similar visual-linguistic alignment using iEEG. Future studies could build on this approach by comparing both familiar and unfamiliar images within the same object categories, while combining fMRI and MEG, with high-resolution local recordings such as iEEG. This approach would offer a more complete understanding of when and where visually grounded representations emerge in the ventral cortex, and whether the preference for visual grounding reflects a general principle of high-level visual processing or is influenced by stimulus familiarity.

References

Ayzenberg, V., & Behrmann, M. (2022). The dorsal visual pathway represents object-centered spatial relations for object recognition. *Journal of Neuroscience*, *42*(23), 4693-4710.

Baldassi, C., Alemi-Neissi, A., Pagan, M., DiCarlo, J. J., Zecchina, R., & Zoccolan, D. (2013). Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLoS computational biology*, *9*(8), e1003167.

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature neuroscience*, *17*(3), 455-462.

Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-based fusion of MEG and fMRI reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, *26*(8), 3563-3579.

Conwell, C. et al. Is visual cortex really “language-aligned”? Perspectives from Model-toBrain Comparisons in Human and Monkeys on the Natural Scenes Dataset. J. Vis. 24, 1288 (2024).

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). Visual representations in the human brain are aligned with large language models.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, *1*(1), 1-47.

Freud, E., Plaut, D. C., & Behrmann, M. (2016). ‘What’ is happening in the dorsal visual pathway. *Trends in cognitive sciences*, *20*(10), 773-784.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, *15*(1), 20-25.

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536-548.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215-243.

Shoham, A., Broday-Dvir, R., Malach, R., & Yovel, G. (2024). High-level visual cortex representations are organized along visual rather than abstract principles. bioRxiv, 2024-11.

Vossel, S., Geng, J. J., & Fink, G. R. (2014). Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. *The Neuroscientist*, *20*(2), 150-159.

Wang, L., Mruczek, R. E., Arcaro, M. J., & Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. *Cerebral cortex*, *25*(10), 3911-3931.