Text-related functionality of visual human pre-frontal activations revealed through neural network convergence.

Adva Shoham[1], Rotem Broday-Dvir[2], Itay Yaron[3] ,Galit Yovel*[1,3] & Rafael Malach*[2]

*Equal contribution

[1]School of Psychological Sciences, Tel Aviv University

[2]Department of Brain Sciences, Weizmann Institute of Science

[3]Sagol School of Neuroscience, Tel Aviv University

Corresponding author

Rafael Malach
Department of Brain Sciences
Weizmann Institute of Science,
Rehovot, Israel
rafi.malach@gmail.com

Summary

The functional role of visual activations of human pre-frontal cortex remains a deeply debated question. Its significance extends to fundamental issues of functional localization and global theories of consciousness. Here we addressed this question by comparing potential parallels between the relational structure of prefrontal visual activations and visual and textual-trained deep neural networks (DNNs). The frontal visual relational structures were revealed in intra-cranial recordings of human patients, conducted for clinical purposes, while the patients viewed familiar images of faces and places. Our results reveal that visual relational structures in frontal cortex were, surprisingly, predicted by text and not visual DNNs. Importantly, the temporal dynamics of these correlations showed striking differences, with a rapid decline over time for the visual component, but persistent dynamics including a significant image offset response for the text component. The results point to a dynamic text-related function of visual prefrontal responses in the human brain.

Key Words

Frontal cortex, Visual cortex, Consciousness, deep neural networks, object recognition, ECoG, Visual recognition, Semantic representations

Acknowledgments

The study was funded by a CIFAR BMC fellowship to R. Malach and ISF grant no.917/21 to G. Yovel

Introduction

While a large body of human brain research has centered on characterizing the functional organization of the visual system located at the posterior part of the brain[1], a consistent finding in visual research highlights visual pre-frontal activations as well. These activations, although showing substantial report-related modulations (e.g.,[2]), have been demonstrated also in relatively passive report-free situations such as movie watching [3–7] or viewing static stimuli[8–11]. Furthermore, there is an ongoing debate concerning the functional role of these pre-frontal visual activations in human mental processes and behavior. Of particular significance, the hypothetical function assigned to pre-frontal activations and their temporal dynamics has played a major role in global theories of consciousness [12]. For example, "localist" theories (e.g., [13]), positing that different contents of awareness should be linked to information in local cortical regions will argue that the visual pre-frontal regions are likely linked to non-perceptual aspects- typically attributed to the frontal lobe- such as language and decision making. In contrast, "globalist" theories- such as *Global Neuronal Workspace*, or *High Order Theories*[14–17] will attribute to frontal activations a central role in enabling visual perceptual awareness. Moreover, according to the *Global Neuronal Workspace* theory, prefrontal activations are hypothesized to be transient, showing at both the appearance and disappearance of perceived stimuli, reflecting the updating of consciously perceived contents[8,12] (for a recent review of this debate see [18,19]). Thus, highlighting the potential functional role and dynamics of visual activation in human prefrontal cortex, beyond its immediate interest in understanding the functionality of the frontal lobes has also far-reaching theoretical consequences.

In recent years, the striking success of deep neural networks (DNNs) opened a new window for addressing the difficult problem of cortical functionality (see [20,21]). A particularly fruitful research direction has been the search for parallel relational structures in brains and artificial networks[22–27] (for review, see [28]). Relational structures are a proposed representational scheme in which cognitive content, e.g., a visual image, is represented by the set of similarities and differences of this image from other visual images[13,29–31]. Such representations are particularly convenient when comparing different systems and modalities. In the domain of human perception, various recent studies have shown that these algorithms account for a significant proportion of variance in human representations of faces and objects, outperforming any previous computational models[28,32–35]. Given the finding of significant parallels between artificial networks and visual representations at the posterior cortex, it is of interest to search for such parallels in the visually active sites found in the human frontal lobe.

Typically, studies investigating brain activity during perception utilize DNNs that are trained on visual inputs[33,36,37]. In a recent study Shoham et al[38] proposed a novel approach, involving models trained on different types of information (visual or textual) to test whether they explain different components underlying the variance in human perception. In the current study we extended this paradigm to explore brain activity. Specifically, we aimed to discern whether the response of the frontal lobe to visual stimuli is primarily encoding visual or textual information. To achieve this, we employed four distinct models trained on visual visual-textual or textual input and assessed which was preferentially correlated with the frontal lobe activation pattern. To quantify the contributions of visual and textual information to visual activations in the frontal lobe, we used intra-cranially recorded IEEG data in which patients were presented with images of famous individuals and places and were asked to memorize the visual details in each image for a later recognition test (see Figure 1). We employed visual (VGG-16, CLIP-images) and text (CLIP-text, GPT) DNNs to model the brain activity. This method enabled us to investigate whether the processing occurring in the frontal lobe during perception was mainly driven by visual or textual information.

Results:

Experimental design and intra-cranial contacts information:

Our study included intra-cranial recordings, conducted for clinical purposes in 13 patients (10 females, mean age 34.7 ± 9.6) and 2571 recording sites. The participants were presented with images of famous individuals and places. Briefly, participants viewed 28 different images, 14 familiar places and 14 familiar faces, divided to two experimental runs. Each image was presented four times during the run, in a pseudo-random order, ensuring no image was repeated twice consecutively. The images were presented for a sustained time duration of 1500 ms, with 750 ms inter-stimulus fixation intervals. No participants were excluded based on performance (for further details see methods and [39,40]).

Figure 1 depicts the experimental set-up and recording sites in the frontal cortex and posterior visual areas that were examined in the present study. All contacts examined in this study were visually responsive, including the frontal electrodes. We first defined all visually responsive contacts and then divided them, based on their anatomical locations, into the different regions of interest (ROIs) (see methods and Norman et al., 2017[40] for details).

Fronto-parietal visual contacts are correlated with text-based rather than visual-based DNNs.

To explore the potential parallels between the visual activations in the frontal lobe and DNNs, we compared their relational structures as revealed through representational dissimilarity matrices (RDMs, see [32]). For each stimulus, we calculated the average response in each contact within a range of 0.1-0.4 sec (100-400 ms) from the moment the stimulus was presented. We mainly focused on this time window as it encompasses the peak of the response, but we explored the response averaged across other time windows as well (time windows at 0.4-0.7 sec, 0.7-1.0 sec, 1.0-1.3 sec, 1.5-1.8 sec). We defined the contacts' mean response (averaged across 4 repetitions) population vectors as the stimulus activation patterns in each ROI and measured the Pearson correlation between each two stimuli activation patterns. One minus correlation represented the dissimilarity score between the two stimuli (lower = more similar) in this ROI. This provided us with the RDM for each ROI (see Figure 1).

Next, we compared the relational-structure similarities between the frontal contacts and the output layers of four different deep convolutional networks: two networks were text-

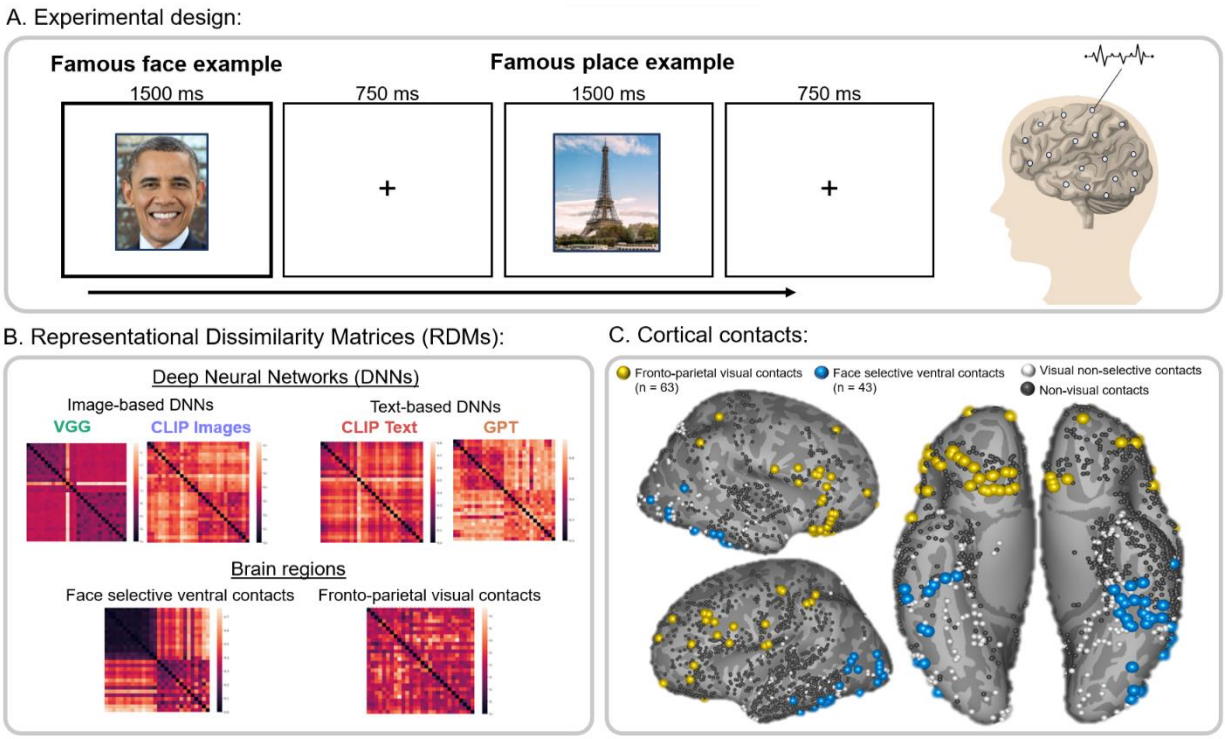

Figure 1: Experimental design and general methods: **a.** Experimental design: participants viewed images of familiar faces or places. **b.** Representational Dissimilarity Matrices (RDMs) that were used for relational structure analysis. Top – the RDMs of two image-based and two text-based artificial networks. Bottom – brain ROIs RDMs. **c.** Contact groups: yellow - fronto-parietal visual contacts, blue – face selective ventral contacts, white – visual non-selective contacts, gray- non-visual contacts.

based (GPT and CLIP-Text) and two networks were image-based (CLIP-Image and VGG). For further details about the different DNNs see methods.

In each DNN, the similarity among all stimuli was computed based on the cosine distance between the extracted feature vectors. One minus the cosine similarity represented the dissimilarity score between the stimuli (lower = more similar). This provided us with the RDM of each DNN that was based on its output (or penultimate) layer (see Figure 1). The correspondence between the patients' frontal-visual ROI RDMs (or the ventral face-selective ventral ROI) and the artificial networks' layers were computed using Pearson correlations: We computed the correlation between the frontal ROI RDM with each network RDM. Then, to test the unique similarity between each network representation with the frontal lobe we calculated the partial correlations between each DNN and the patients' RDM, when all other DNNs are held constant. To test for statistical significance, we computed the same correlations using a permutation test (shuffled pairs labels, see methods for further details) and bootstrap analyses (leaving one participant/contact/stimuli/repetition out). We tested both the significance of the predictors' partial correlations individually and the difference between each partial correlation in each ROI. P values were FDR corrected for each ROI and each analysis (individual predictors and differences) separately.

Figure 2 depicts the partial correlations found between each of the different DNNs (warm colors depict the text-related while cold colors the visual-related ones), when the other three DNNs are held constant. The figure shows the occipito-temporal, high order, visual face-related ventral ROI (Panel A) and the fronto-parietal visual contacts (Panel B) in the 0.1-0.4 sec time window. As can be seen, in the Fronto-Parietal ROI, only the text-related DNNs displayed significant partial correlations to the patients' RDMs. The opposite selectivity was found in the face-selective ventral contacts in the visual cortex. Here, only the visual-related VGG DNN showed a significant correlation to the Face-related ventral contacts.

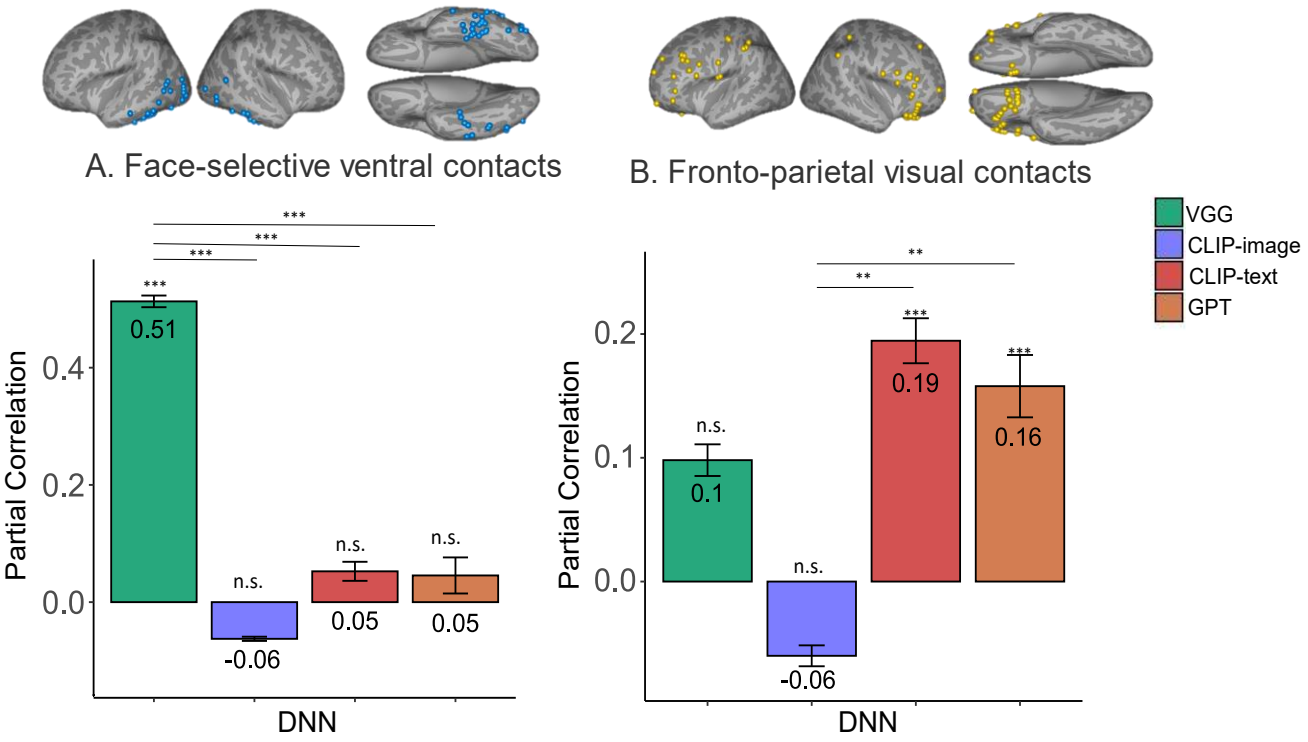

Figure 2: Partial correlations of image and text DNNs in Face-selective ventral and Frontal-parietal visual contacts: **A**. Face-selective occipito-temporal contacts (n=43), Error bars indicate leave one participant out procedure s.e.m. **B.** Fronto-parietal visual contacts (n=63), Error bars indicate leave one participant out procedure s.e.m. All *p* values were derived from a pair-images permutation test (10000 permutations). Reported *p* values are FDR corrected in each ROI separately. Note the clear shift in bias from visual-based to text-based preference from occipito-temporal to frontal-parietal ROIs. $** \; p_{FDR} < 0.01$,

Sustained textual contribution with a stimulus-offset response in frontal-parietal visual responses.

Given that CLIP-text showed the strongest partial correlation with the fronto-parietal visual contacts, we proceeded to investigate this brain-network correlation across different time windows (Figure 3). For comparative analysis, we examined CLIP-text correlations within the same time windows with face-selective ventral contacts as well as the correlations of the ventral and frontal-parietal contacts with the visual DNN (VGG) (Note that analysis shown in Figure 3 displays the correlations and not the partial correlations that are shown in Figure 2). To assess the significance of differences in correlations across time windows, we used a permutation test (see methods for further details). When exploring CLIP-text and fronto-parietal correlations (Figure 3, bottom-right panel), they were all as high as the correlation of the first time window (0.1-0.4 sec) besides the 1.0-1.3 sec,

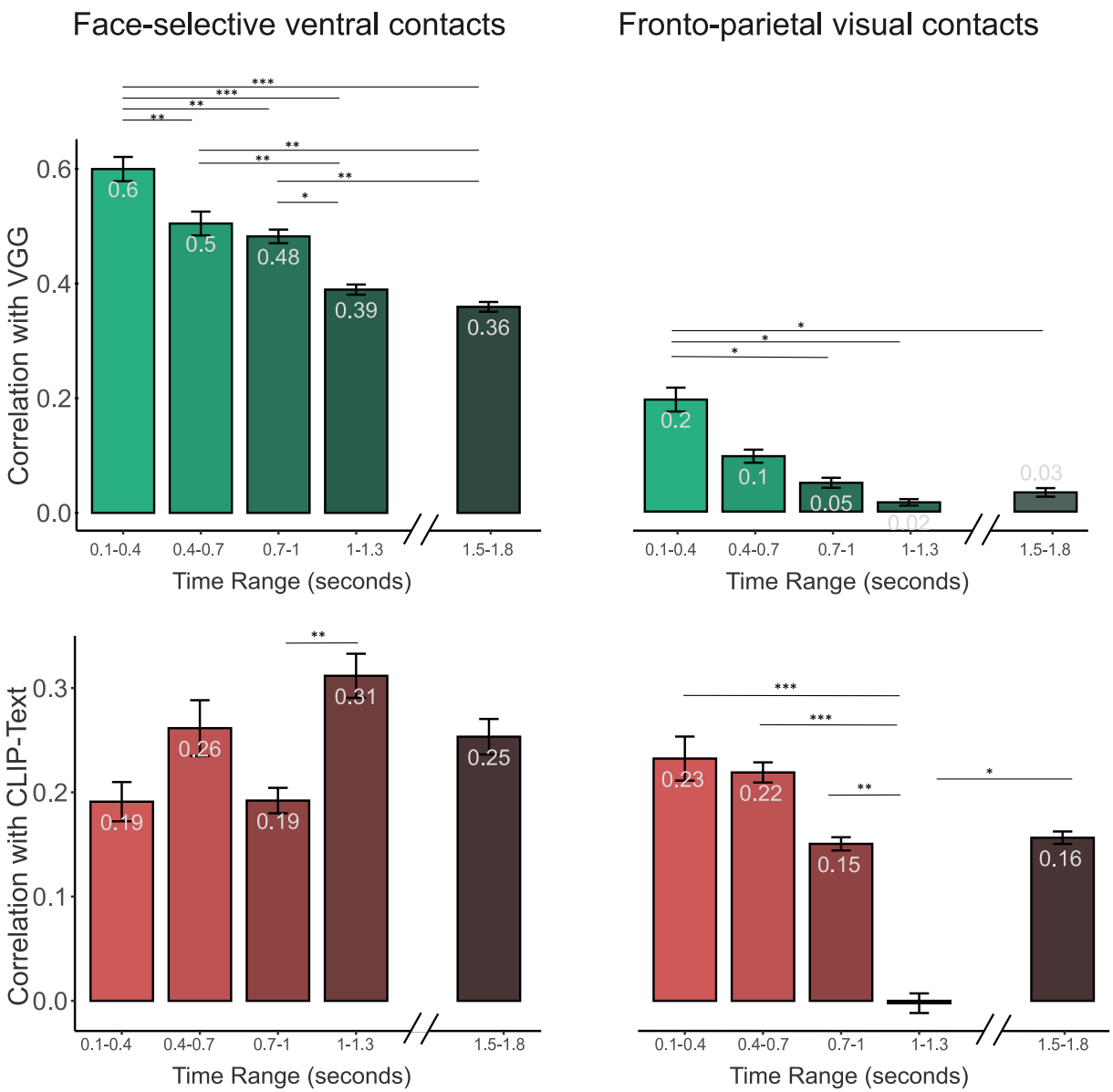

Figure 3: Correlations with VGG or CLIP-text in Face selective ventral and Frontal-parietal visual contacts at different time windows: Top bar plot (green bars) show the correlations with VGG in: face-selective ventral contacts (left), and fronto-parietal visual contacts (right). Bottom bar plot (red bars) shows the correlations with CLIP-text in face-selective ventral contacts (left), and fronto-parietal visual contacts (right). Note the strikingly different dynamics in fronto-parietal contacts between the visual component that declined rapidly and the text component with persisted followed by an offset response. Error bars indicate leave one participant out procedure s.e.m $* \, p_{FDR} < 0.05$, $** \, p_{FDR} < 0.01$, $*** \, p_{FDR} < 0.001$

which showed no correlation. This might suggest a relatively persistent processing pattern during the stimuli presentation until 1 sec. Interestingly, a similar pattern seems to reappear at the stimulus offset (1.5 sec). This was not the case for VGG correlations with the fronto-parietal visual contacts (Figure 3 top-right panel) which showed a steep decline after the first-time window. Nor is the case for the face-selective ventral contacts that showed gradual decline in VGG correlations (Figure 3 top-left panel) and variable CLIP-text correlations with time (Figure 3 bottom-left panel). see Supplementary Figure 1 for a similar analysis with GPT.

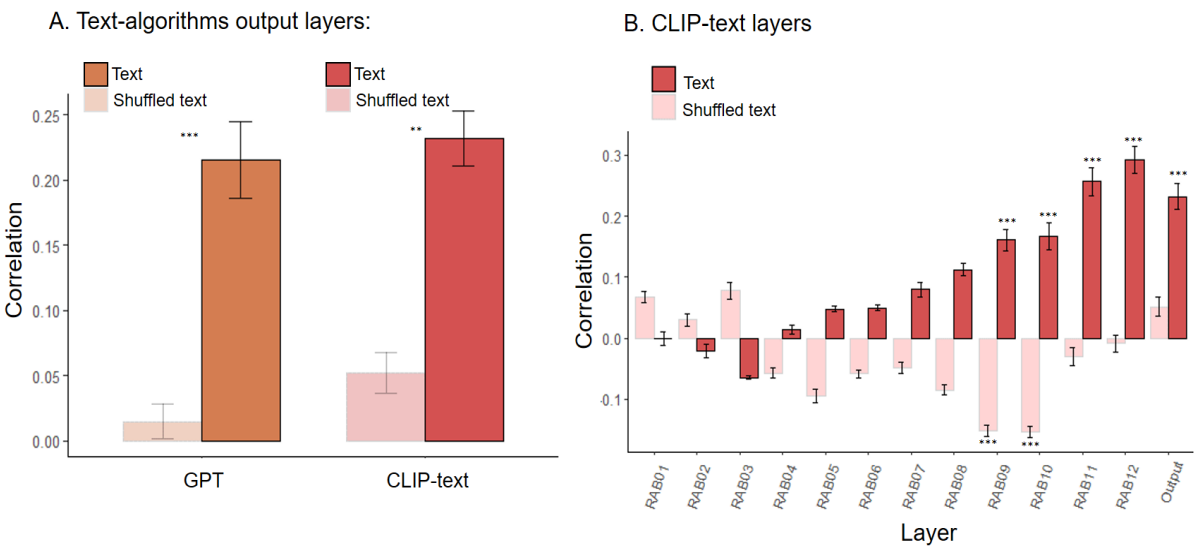

Figure 4: Text DNNs correlations with Fronto-parietal visual contacts when embeddings were based on the original text or on shuffled text: **A.** The correlations with the fronto-parietal visual contacts in the first time window (0.1-0.4 sec) with GPT output layer (orange bars) and CLIP-text output layer (red bars), based on the original text (opaque) or shuffled text (pale). **B.** The correlations with the fronto-parietal visual contacts in the first time window (0.1-0.4 sec) with CLIP-text different layers based on the original text (opaque) or shuffled text (pale). Error bars indicate leave one participant out procedure s.e.m. $** \; p_{FDR} < 0.01 \quad *** \; p_{FDR} < 0.001$

A monotonic increase in the correlations with fronto-parietal visual contacts across the layers of CLIP-text

To further explore the initial-time window (0.1-0.4 sec) we asked whether the frontal correlation is merely due to the inner structure of the DNNs, regardless of its input? Figure 4A shows that the correlation to the text DNNs when presented with shuffled text input instead of the original text drastically reduced the RDMs correlation. We also extracted the RDM according to each layer of CLIP-Text network (see methods for further details). Examining the evolution of the correlation to the Frontal contacts' RDM through the DNN layers (Figure 4B) revealed a clear monotonic increase towards higher order layers, reaching the highest values at the three top layers (Residual attention blocks (RAB) 11,12, and output-13), indicating a specific, functional similarity between the frontal lobe responses and the top CLIP-Text layers. Also note the striking difference between the RDM correlations of the correct and shuffled texts.

To rule out the possibility that these correlations were a result of an outlier patient, contact or stimulus response, we examined how robust was the correlation to CLIP-Text when using only partial data in the analysis. These analyses and results are reported in the Supplementary material. As can be seen in Supplementary Figure 2, the results remained fairly robust to different bootstrap analyses, arguing against an outlier-driven effect.

Correlations of CLIP-text with fronto-parietal visual contacts are maximal for the first image repetition.

Since in the experiment the same visual images were shown four times to each patient, it was of interest to examine how such repetitions affected the observed correlation to the CLIP-Text network. To that end, separate RDMs were constructed for each repetition of the stimuli. Supplementary Figure 3 depicts the RDM correlations across CLIP-Text layers for each of the repetitions. Result show that the correlation levels changed in a complex manner across repetitions. With the novel images (first appearance) failing to show any correlation and a significant correlation appearing for the first repetition followed by a progressive decline in the second and third ones.

Finally, to examine whether the correlation to CLIP-Text changes across different frontal subdivisions, the contacts were divided into two major anatomical subdivisions, orbito-frontal and lateral-frontal contacts (see further detailed in the Supplementary file). As can be seen in Supplementary Figure 4, the overall pattern of correlations across anatomical regions in the frontal cortex remained largely unchanged showing a gradual increase in correlation when moving from low-level to high-level network layers.

Discussion

In this study, we used text-based and image-based DNNs to investigate their similarity to brain activity, while participants viewed images of famous faces and places. Our analysis revealed that visually responsive fronto-parietal contacts' activations were best predicted by text-based networks, suggesting semantic rather than visual-related processing.

The frontal lobe processing in visual perception

Our results reveal, in agreement with a number of prior studies based on both fMRI and intracranial brain recordings[2,3,8–11,41], that the frontal lobes are consistently activated by visual images. Importantly, in the present study the patients were instructed to view and memorize the images rather than report on seeing them or not - so the visual activations cannot be attributed to active reporting but to the perceptual processing of these images.

Though previous studies showed activation in the frontal lobe while viewing visual stimuli[3,8–10,42], the necessity of the frontal lobe in conscious perception is still unclear. Some research highlights the non-visual nature of frontal lobe tasks, with the highest activations during reporting and motor tasks[2,5,18,43], and lesion cases challenge the idea of the prefrontal cortex's essential role in perception[44,45]. It is important to note that identifying the function of cortical responses is a fundamental challenge in all human brain studies, particularly since causal manipulations are rare and necessarily associated with poorly controlled

clinical cases. We and others have recently proposed[20,38,46,47] that insight into the functionality of cortical responses can be gained by searching for “convergent evolution”- i.e. parallels that are found between artificial networks and brain systems[21,32,48]. Here we adopted this strategy to gain insights into the enigma of frontal visual responses. By searching for correlations between relational structures, revealed through RDMs, in frontal cortex and visual and textual deep learning networks, we have uncovered a preference for text-related networks. The correlations to the image-based networks although positive, were far weaker and did not account significantly for the visual frontal response beyond the textual-based networks.

The fact that the more specific correlation was found to a text-based network rather than a similar visually related ones (CLIP-Images) is intriguing, especially considering that the patients' RDMs were based on visual responses and that visual-related DNNs were correlated with visual contacts in high level visual cortex. This finding suggests that the frontal visual responses may be more closely tied to lingual analysis influenced by visual contents rather than visual perception per-se. Notably, robust visual responses associated with non-perceptual functions are well-documented in the human brain. Particularly striking cases are visual activations found in the hippocampus associated with conceptual representations[49,50] and visual activations in human high order somatosensory cortex[51,52]. Given the established role of the frontal lobes in language and thought processing[53], it is not surprising that the visual activations in this study appear to be more linked to linguistic rather than perceptual processing functions.

The temporal dynamics of prefrontal and posterior correlations

Comparing the temporal dynamics of correlations between the RDMs of DNNs and brain activations revealed different patterns of results in posterior compared to frontal ROIs. In the posterior face selective ventral ROI both image-based (VGG) and text-based (CLIP-Text) networks showed consistent correlations across time windows, with higher correlations for VGG. Conversely, activations in the frontal ROI were correlated with VGG only in the earliest time window (0.1- 0.4 sec), in sharp contrast to the correlations profile of CLIP-Text which were found to be significant in all but one of the time windows (1-1.3 sec). These results echo previous findings showing persistent representations of visual contents in posterior regions[8–10], and add novel insight regarding the dynamics of prefrontal activations in two ways: First, the obtained results point to a differential dynamics for the transient, visual-related aspects of the frontal activations vs the persistent text-related ones. Second, by focusing on the linguistic component of the frontal activation, our results demonstrate an intriguing offset response, predicted, albeit for visual processing (300 ms

after stimulus offset), by the Global Neuronal Workspace theory[8]. It is worth mentioning that CLIP-Text and GPT exhibit distinct patterns of correlation with the fronto-parietal RDM when looking at different time windows (Figure 3 and Supplementary Figure 1). These variations might be due to their different input (name or paragraph from Wikipedia) or the way they were trained. CLIP-Text is a multi-modal visual-semantic DNN, while GPT is a unimodal language DNN. Further research is needed to account for these differences.

Dynamic modulation of correlations by stimulus repetition

The experimental design in the present study included four repetitions of each of the stimuli presented. The temporal distance between repetitions was randomized across stimuli- however, grouping the repetitions according to their sequential repetition order revealed a complex dynamic of the correlation magnitude to the CLIP-Text network. Thus, one could discern two, not necessarily linked, changes (see Supplementary Figure 3). The first was a rapid increase from essentially no correlation in the first, novel image presentation to the first time the images were repeated. This suggests that the frontal cortex function that is correlated to the CLIP-Text network may not be an encoding function, but rather related to short-term memory or recall. It is interesting to note that a similar discrepancy between novel and repeated presentations were observed in the hippocampal formation[39] where content-selectivity appeared only on a second presentation, while the first response was a more generalized novelty signal. It is intriguing to speculate that this novelty effect may have been dominant also in the pre-frontal recordings and masked the potential correlation with the CLIP-Text network. The second effect was a gradual decline in the correlation at further repetitions. This type of gradual decline, termed adaptation or repetitions suppression, has been ubiquitously observed in the visual system[9,54–57]. It is interesting to note that what appears as a repetition-suppression effect can be revealed through the window of brain- artificial network correlations.

Sources for the low-level correlation magnitude

Examining the magnitude of the correlation between the RDMs derived from frontal activations and the CLIP-Text layers even when specifically focusing on the optimal conditions, failed to reveal correlation levels higher than 0.3. This is a significantly lower value compared to magnitude of the brain to network correlation values found in high order visual areas – which was substantially higher, ranging from 0.4-0.6[32] (and see VGG correlations in Figure 3). A number of sources, not necessarily mutually exclusive, could account for such relatively low correlation values. The most obvious factor is the relatively low activation level of the frontal contacts. This weak activation and low SNR is typical in report-free paradigms such as the one adopted in the present study (See also [2,9]). Whether

this effect is simply due to smaller signal to noise ratios in the weakly responding contacts or whether it reflects a less selective signal in these contacts remains to be studied.

A second possibility is that as indeed suggested by the preferential correlation to the text-related network the true function of the pre-frontal contacts is not perceptual but language related. Since the task of the patients did not involve explicit language or textual aspects- the visual activation may have merely occurred as an automatic, auxiliary by-product of the visual presentation of a familiar image - a ubiquitous cortical phenomena as has been pointed out above[18,58,59]. Resolving this issue will necessitate additional experiments in which linguistic and decision-making tasks will be manipulated in addition to the visual images.

Finally, one cannot rule out the possibility that more advanced network structures may produce higher and more consistent activation levels. The recent explosive growth in the variety and power of language models will offer rich ground for future explorations in this domain.

To conclude, our findings suggest a closer association between visually responsive fronto-parietal regions and text-based rather than visually related DNNs. These findings highlight the non-visual processing of the frontal lobe during perception. This association was observed in different latencies of the stimuli presentations with a drop and a recovery after the stimulus offset, which is aligned with the Global Neuronal Workspace theory.

Methods:

Human data:

We describe here the participants tasks and stimuli as well as ROI definitions. Additional details can be found in[9].

<u>Participants</u>

Electrophysiological data were acquired from 13 patients (10 females, mean age 34.7 ± 9.6) via intracranial IEEG recordings. The recordings were obtained while the patients underwent pre-surgical evaluation for drug-resistant epilepsy at North Shore University Hospital in NY. As part of the clinical assessment, subdural or depth contacts were implanted in all patients. The study followed the latest version of Declaration of Helsinki, and all patients provided a fully informed consent to participate, including consent to publish the results, according to the US National Institute of Health guidelines. The Feinstein Institute for Medical Research's institutional review board monitored the study. Notably, no clinical seizures occurred during the experiment.

Experimental task and stimuli

The experiment comprised two runs, each starting with a 200-second resting phase with closed eyes. Subsequently, participants viewed 14 different images in each session (7 images of each category – famous faces or famous places). In total, 28 stimuli were used, 14 in each run. Each image was presented four times (for 1500 ms each time, with 750 ms inter-stimulus intervals) in a semi-random order to avoid consecutive repeats. Participants were informed that they would later be required to recall the images and describe their visual features in detail, rather than merely identifying them. Consequently, they were instructed to memorize the details of the images. Participants performed well, and no participants were excluded based on performance (Further details can be found in [40]).

Definition and grouping of visually responsive contacts:

Data from all contacts was inspected, preprocessed, and transformed to the high-frequency broadband (HFB) signal (see [9,39,40] for further details). We extracted visual-responsive contacts by a comparison of each contact's post-stimulus HFB response, averaged across the 100 to 500 ms period following stimulus onset, with its pre-stimulus baseline, averaged across the -400ms to -100ms interval prior to stimulus onset. This comparison was conducted using a two-tailed Wilcoxon signed-rank test, that was FDR corrected for all contacts (from all patients) together. Subsequently, contacts demonstrating a significant HFB response (p_{fdr} < 0.05) were categorized as visually responsive, totaling 377 contacts.

Following this, visually responsive contacts were grouped into subsets based on their anatomical and functional characteristics, namely face-selective ventral and fronto-parietal visual contacts. Face-selective ventral contacts were defined by comparing mean HFB responses between faces and places during the 100-500ms post-stimulus window using a Wilcoxon rank-sum test. Contacts showing significantly greater activation to faces compared to places (p_{fdr} < 0.05), located anatomically beyond early visual regions (but excluding the frontal cortex), were categorized as face-selective ventral contacts (n = 43, depicted in blue in Figure 1C). We defined the fronto-parietal ROI based on the Desikan Killiany atlas locations of the following labels: superior frontal gyrus, rostral middle frontal gyrus, pars orbitalis, pars triangularis, pars opercularis, precentral gyrus, postcentral gyrus, supramarginal gyrus, orbital frontal gyrus, and the anterior cingulate. Visual contacts located within these regions were categorized as the fronto-parietal visual contacts (n = 63, depicted in yellow in Figure 1C). Any visually responsive contacts not falling into these specified categories were termed "other visually-responsive contacts"

and are depicted in white in Figure 1C, while non-visual contacts are depicted in gray. Further details can be found in [9].

Deep neural networks:

Visual network – ImageNet pre-trained VGG-16: To examine a network based visual representation of the presented stimuli, we used the VGG-16 network[60] pre-trained on ImageNet[61], which includes 300000 images from 1000 object categories. VGG is a commonly used network performing object and face categorization at human level[38,62–64]. We then extracted the embeddings of each image based on the feature vector representation in the penultimate- i.e. the hidden layer just below the output (fc7) layer of the network, which is the representation that is used for the classification. The similarity between each images pair was computed based on the cosine distance between these feature vectors.

Text network – openAI GPT2: A transformer-based language algorithm[65]. GPT is trained on a vast corpus of text to carry out various language tasks (e.g. semantic similarity, questions answering, grammatical correction etc.), often involving longer text captions. Therefore, we chose to extract its representations using text from Wikipedia to obtain the representation of each stimulus. Specifically, we retrieved the embeddings of the first paragraph from Wikipedia corresponding to each familiar stimuli (identity or a place), based on the model output layer. Subsequently, we computed the similarity between each stimuli pair based on the cosine distance between the representations of these textual descriptions.

Multi-modal network - CLIP (Contrastive Language-Image pre-training)[66]: CLIP is trained in a self-supervised, contrastive manner to create similar representations for images and their text caption based on a training set of 400M images and their captions from the internet[66]. We extracted the embeddings of each image based on the output layer of trained resnet architecture using the visual and language components of CLIP. It is worth mentioning that while CLIP-Text and CLIP-Images networks are trained by pairing images and text in a contrastive leaning protocol, each network has a separate embedding space. For each component (text or images) we extracted the representation based on its output layer.

CLIP Visual Component: We extracted the representations of the visual images that were presented to the patients. We then computed the similarity between each object pair based on the cosine distance between these visual representations. We will refer to these representations as CLIP-images.

CLIP Text component: CLIP textual component is trained with short captions that are used as labels of images on the internet. Moreover, CLIP-text was limited to short text input (up to 77 characters). Therefore, we extracted its representations based on the names of the stimuli. We extracted the representations based on the names of the familiar exemplar (identity or a place) presented in each stimulus (e.g. “Eiffel Tower”). We computed the similarity between each object pair based on the cosine distance between these name representations. We will refer to these representations as CLIP-text.

This provided us with the RDM of each network that is based on its output (or penultimate) layer. In addition, we also extracted the RDMs of each layer of the network.

Data Analysis:

We first computed the Pearson correlation between the RDMs of each DNN and the RDMs of the ventral and frontal visual contacts. Then, to assess the unique contribution of each DNN to the neural responses we computed the partial correlation of each DNN while holding the other DNNs constant.

To test the significance of a correlation/partial correlation of a single network in a specific ROI We used a permutation test in which the networks’ similarity scores (between pair of stimuli) were held constant, but we shuffled the similarity scores that are brain based. For each analysis and each ROI, we shuffled the similarity between pairs of stimuli in 10000 iterations. Then we calculated the correlation/partial correlation (respectively) of the network with the patients’ RDM (based on the shuffled similarities). P-value assigned to each network and layer was the proportion of iterations in which the original correlations/partial correlation (based on the real similarities) was greater (or smaller) than the value that was calculated in each iteration. P values were adjusted according to Phipson and Smyth[67] correction for permutation tests, and then they were FDR corrected for all networks that were tested in the same analysis, for each ROI separately. Then we used a two-tailed test (alpha 0.025) to infer significance. We also tested the significance of the difference between each pair of two DNNs or two time windows; To do so we used a permutation test in which the networks’ similarity scores were held constant, but we shuffled the similarity scores that are brain based. For each analysis and each ROI, we shuffled the similarity between pairs of stimuli in 10000 iterations. Then we calculated the correlation/partial correlation (respectively) of each network/time window with the neural RDM (based on the shuffled similarities), and proceeded to calculate the difference between each two networks/time windows correlations/partial correlations (with the neural RDMs). P value assigned to each comparison of a pair of networks was the proportion of iterations in which the original difference in correlations/partial correlation (based on the

real similarities) was greater (or smaller) than the value that was calculated in each iteration. P values were adjusted according to[67] correction for permutation tests, and then they were FDR corrected for all networks that were tested in the same analysis, for each ROI separately. Then we used a two-tailed test (alpha 0.025) to infer significance.

References:

1. Grill-Spector, K. & Malach, R. The human visual cortex. *Annu Rev Neurosci* **27**, 649–677 (2004).
2. Noy, N. *et al.* Ignition's glow: Ultra-fast spread of global cortical activity accompanying local "ignitions" in visual cortex during conscious visual perception. *Conscious. Cogn.* **35**, 206–224 (2015).
3. Golland, Y. *et al.* Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation. *Cereb. Cortex* **17**, 766–777 (2007).
4. Kronemer, S. I. *et al.* Human visual consciousness involves large scale cortical and subcortical networks independent of task report and eye movement activity. *Nat. Commun.* **13**, 7342 (2022).
5. Frässle, S., Sommer, J., Jansen, A., Naber, M. & Einhäuser, W. Binocular rivalry: frontal activity relates to introspection and action but not to perception. *J. Neurosci.* **34**, 1738–1747 (2014).
6. Bellet, J. *et al.* Decoding rapidly presented visual stimuli from prefrontal ensembles without report nor post-perceptual processing. *Neurosci. Conscious.* **2022**, niac005 (2022).
7. Kapoor, V. *et al.* Decoding internally generated transitions of conscious contents in the prefrontal cortex without subjective reports. *Nat. Commun.* **13**, 1535 (2022).
8. Consortium, C. *et al.* An adversarial collaboration to critically evaluate theories of consciousness. *BioRxiv* 2023–06 (2023).
9. Broday-Dvir, R., Norman, Y., Harel, M., Mehta, A. D. & Malach, R. Perceptual stability reflected in neuronal pattern similarities in human visual cortex. *Cell Rep.* **42**, (2023).

10. Vishne, G., Gerber, E. M., Knight, R. T. & Deouell, L. Y. Distinct ventral stream and prefrontal cortex representational dynamics during sustained conscious visual perception. *Cell Rep.* **42**, (2023).
11. Chan, A. W.-Y. & Downing, P. E. Faces and eyes in human lateral prefrontal cortex. *Front. Hum. Neurosci.* **5**, 51 (2011).
12. Melloni, L., Mudrik, L., Pitts, M. & Koch, C. Making the hard problem of consciousness easier. *Science* **372**, 911–912 (2021).
13. Malach, R. Local neuronal relational structures underlying the contents of human conscious experience. *Neurosci. Conscious.* **2021**, niab028 (2021).
14. Brown, R., Lau, H. & LeDoux, J. E. Understanding the higher-order approach to consciousness. *Trends Cogn. Sci.* **23**, 754–768 (2019).
15. Lau, H. & Rosenthal, D. Empirical support for higher-order theories of conscious awareness. *Trends Cogn. Sci.* **15**, 365–373 (2011).
16. Mashour, G. A., Roelfsema, P., Changeux, J.-P. & Dehaene, S. Conscious processing and the global neuronal workspace hypothesis. *Neuron* **105**, 776–798 (2020).
17. Dehaene, S. & Changeux, J.-P. Experimental and theoretical approaches to conscious processing. *Neuron* **70**, 200–227 (2011).
18. Malach, R. The role of the prefrontal cortex in conscious perception: the localist perspective. *J. Conscious. Stud.* **29**, 93–114 (2022).
19. Michel, M. Conscious perception and the prefrontal cortex a review. *J. Conscious. Stud.* **29**, 115–157 (2022).
20. Simony, E., Grossman, S. & Malach, R. Brain-Machine convergent evolution: a window into the functional role of neuronal selectivity. (2023).
21. Kanwisher, N., Khosla, M. & Dobs, K. Using artificial neural networks to ask ‘why’questions of minds and brains. *Trends Neurosci.* **46**, 240–254 (2023).

22. Bomatter, P. *et al.* When pigs fly: Contextual reasoning in synthetic and natural scenes. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 255–264 (2021).
23. Vinken, K., Boix, X. & Kreiman, G. Incorporating intrinsic suppression in deep neural networks captures dynamics of adaptation in neurophysiology and perception. *Sci. Adv.* **6**, eabd4205 (2020).
24. Ponce, C. R. *et al.* Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* **177**, 999–1009 (2019).
25. Ritchie, J. B. *et al.* Untangling the animacy organization of occipitotemporal cortex. *J. Neurosci.* **41**, 7103–7119 (2021).
26. Wardle, S. G. & Baker, C. I. Recent advances in understanding object recognition in the human brain: deep neural networks, temporal dynamics, and context. *F1000Research* **9**, (2020).
27. Bankson, B. B., Hebart, M. N., Groen, I. I. A. & Baker, C. I. The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *Neuroimage* **178**, 172–182 (2018).
28. Kriegeskorte, N. *Deep Neural Networks: A New Framework for Modelling Biological Vision and Brain Information Processing*. http://biorxiv.org/lookup/doi/10.1101/029876 (2015) doi:10.1101/029876.
29. Edelman, S., Grill-Spector, K., Kushnir, T. & Malach, R. Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology* **26**, 309–321 (1998).
30. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 249 (2008).
31. Lau, H., Michel, M., LeDoux, J. E. & Fleming, S. M. The mnemonic basis of subjective experience. *Nat. Rev. Psychol.* **1**, 479–488 (2022).

32. Grossman, S. *et al.* Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* **10**, 1–13 (2019).
33. Dobs, K., Martinez, J., Kell, A. J. E. & Kanwisher, N. Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci Adv* **8, eabl8913**, (2022).
34. Groen, I. I. A. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife* **1–26**, (2018).
35. Hasson, U., Nastase, S. A. & Goldstein, A. Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron* **105**, 416–434 (2020).
36. Carlin, J. D. & Kriegeskorte, N. Adjudicating between face-coding models with individual-face fMRI responses. *PLoS Comput Biol* **13**, 1–28 (2017).
37. Schyns, P. G., Snoek, L. & Daube, C. Degrees of algorithmic equivalence between the brain and its DNN models. *Trends Cogn Sci* **26**, 1090–1102 (2022).
38. Shoham, A., Grosbard, I. D., Patashnik, O., Cohen-Or, D. & Yovel, G. Using deep neural networks to disentangle visual and semantic information in human perception and memory. *Nat. Hum. Behav.* 1–16 (2024).
39. Norman, Y. *et al.* Hippocampal sharp-wave ripples linked to visual episodic recollection in humans. *Science* **365**, (2019).
40. Norman, Y., Yeagle, E. M., Harel, M., Mehta, A. D. & Malach, R. Neuronal baseline shifts underlying boundary setting during free recall. *Nat. Commun.* **8**, 1301 (2017).
41. Goldberg, I. I., Harel, M. & Malach, R. When the brain loses its self: prefrontal inactivation during sensorimotor processing. *Neuron* **50**, 329–339 (2006).
42. Hesselmann, G., Hebart, M. & Malach, R. Differential BOLD activity associated with subjective and objective reports during “blindsight” in normal observers. *J. Neurosci.* **31**, 12936–12944 (2011).

43. Tsuchiya, N., Wilke, M., Frässle, S. & Lamme, V. A. No-report paradigms: extracting the true neural correlates of consciousness. *Trends Cogn. Sci.* **19**, 757–770 (2015).
44. Stuss, D. T. & Benson, D. F. Neuropsychological studies of the frontal lobes. *Psychol. Bull.* **95**, 3 (1984).
45. Corbetta, M., Kincade, M. J., Lewis, C., Snyder, A. Z. & Sapir, A. Neural basis and recovery of spatial attention deficits in spatial neglect. *Nat. Neurosci.* **8**, 1603–1610 (2005).
46. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci* (2016) doi:10.1038/nn.4244.
47. Cohen, U., Chung, S., Lee, D. D. & Sompolinsky, H. Separability and geometry of object manifolds in deep neural networks. *Nat. Commun.* **11**, 746 (2020).
48. Doerig, A. *et al.* The neuroconnectionist research programme. *Nat. Rev. Neurosci.* **24**, 431–450 (2023).
49. Quiroga, R. Q. Concept cells: the building blocks of declarative memory functions. *Nat. Rev. Neurosci.* **13**, 587–597 (2012).
50. Quiroga, R. Q., Mukamel, R., Isham, E. A., Malach, R. & Fried, I. Human single-neuron responses at the threshold of conscious recognition. *Proc. Natl. Acad. Sci.* **105**, 3599–3604 (2008).
51. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G. & Malach, R. Intersubject synchronization of cortical activity during natural vision. *science* **303**, 1634–1640 (2004).
52. Pitcher, D., Garrido, L., Walsh, V. & Duchaine, B. C. Transcranial magnetic stimulation disrupts the perception and embodiment of facial expressions. *J. Neurosci.* **28**, 8929–8933 (2008).
53. Berkovich-Ohana, A. *et al.* Inter-participant consistency of language-processing networks during abstract thoughts. *NeuroImage* **211**, 116626 (2020).
54. Grill-Spector, K. *et al.* Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron* **24**, 187–203 (1999).

55. Grill-Spector, K., Henson, R. & Martin, A. Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* **10**, 14–23 (2006).
56. Gerber, E. M., Golan, T., Knight, R. T. & Deouell, L. Y. Cortical representation of persistent visual stimuli. *Neuroimage* **161**, 67–79 (2017).
57. Golan, T. *et al.* Human intracranial recordings link suppressed transients rather than'filling-in'to perceptual continuity across blinks. *elife* **5**, e17243 (2016).
58. Block, N. What is wrong with the no-report paradigm and how to fix it. *Trends Cogn. Sci.* **23**, 1003–1013 (2019).
59. Block, N. Finessing the bored monkey problem. *Trends Cogn. Sci.* **24**, 167–168 (2020).
60. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. (2014) doi:10.48550/arXiv.1409.1556.
61. Deng, J. ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition 248–255* (IEEE, 2009). doi:10.1109/CVPR.2009.5206848.
62. Jacob, G., Pramod, R. T., Katti, H. & Arun, S. P. Qualitative similarities and differences in visual object representations between brains and deep networks. *Nat Commun* **12**, 1872 (2021).
63. Abudarham, N., Grosbard, I. & Yovel, G. Face recognition depends on specialized mechanisms tuned to view-invariant facial features: insights from deep neural networks optimized for face or object recognition. *Cogn Sci* **45**, 13031 (2021).
64. Yovel, G., Grosbard, I. & Abudarham, N. Deep learning models challenge the prevailing assumption that face-like effects for objects of expertise support domain-general mechanisms. *Proc. R. Soc. B* **290**, 20230093 (2023).
65. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving Language Understanding by Generative Pre-Training.

66. Radford, A. Learning transferable visual models from natural language supervision. in *Proc. 38th International Conference on Machine Learning* (eds. Meila, M. & Zhang, T.) 8748–8763 (2021).
67. Phipson, B. & Smyth, G. K. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.* **9**, (2010).