The organization of high-level visual cortex is aligned with visual rather than abstract linguistic information

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

*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

Recent studies show that linguistic representations predict the response of high-level visual cortex to images, suggesting an alignment between visual and linguistic information. Here, using iEEG, we tested the hypothesis that such alignment is limited to textual descriptions of the visual content of the image and would not appear for their abstract textual descriptions. We generated two types of textual descriptions for images of famous people and places: *visual-text*, describing the visual content of the image, and *abstract-text*, based on their Wikipedia definitions, and extracted their relational-structure representations from a large language model. We used these linguistic representations, along with visual representation of the images based on deep neural network, to predict the iEEG responses to images. Neural relational-structures in high-level visual cortex were similarly predicted by visual-images and visual-text, but not abstract-text representations. These results demonstrate that visual-language alignment in high-level visual cortex is limited to visually grounded language.

Introduction

High-level visual cortex shows selective response to different categories, including objects, scenes, faces, and bodies[1–7]. These category selective responses are localized in distinct cortical regions, with animate stimuli, such as faces and bodies, more laterally, and inanimate objects and scenes more medially (for review see[3], but see[8,9] for a different perspective). The question of whether the neural response of high-level visual cortex is driven by visual, pictorial properties or by semantic, information has been the subject of extensive discussion (for a recent review, see[10]).

Recent investigations using deep learning algorithms offer new tools to explore this question (for recent perspectives supporting this approach see [11–13]). In the past decade, deep learning algorithms have reached human level performance in object and face recognition. Moreover, the relational structure of their embeddings (which will be termed here-*representation*) generated by these algorithms across their different layers are correlated with the relational structure of the neural response to the same images in early and high-level areas of visual cortex[12,14–17]. Interestingly, recent studies show that the neural response to visual stimuli can be predicted not only by visual deep learning algorithms but also by the linguistic representation of their textual descriptions by large language models (LLMs)[18–21]. This surprising finding has been suggested to support the notion that the representations of images in high-level visual brain areas are aligned with semantic information[19,22,23].

For example, Conwell et al (2023) used visual and language deep learning algorithms to predict the fMRI response of participants viewing images from the natural scene dataset[24]. In this dataset, each image is accompanied by verbal descriptions of the content of the image (e.g., “A horse carrying a large load of hay and two people sitting on it”). This enabled them to ask whether the neural response to the images can be predicted not only by the visual representation of the images based on visual deep convolutional neural networks (DCNNs), but also by LLMs’ embedding of the verbal descriptions of the images. Results showed that the response of the occipito-temporal cortex to images was similarly predicted by the visual representations of visual DNNs and the linguistic representations of the LLMs[20]. Using the same dataset, Doerig and colleagues (2024) used a searchlight analysis and found that LLM representations accounted for brain responses to images across the entire visual system, particularly in higher-level areas[19]. Additionally, they showed that using text descriptions about people, scenes, or places, as input to their LLM brain encoder, predicted the response of brain areas that are visually selective to these categories (e.g., sentences about people/places/food were correlated with face/place/food-selective ROIs, respectively). Accordingly, they concluded that the ventral visual cortex represents semantic information

about visual input. They further suggested an alignment between visual and linguistic representations in predicting the response of high-level visual cortex to images.

In contrast to these findings, in a recent iEEG study that examined visual and linguistic contributions to the neural response to familiar face and place images, we found no such alignment[25]. In this study, we found that high-level visual cortex showed high correlation with visual DCNN representations, but low correlations with LLMs representations[25]. An important difference between the type of textual descriptions that we used in our study was that they conveyed conceptual information about the images which is independent of their pictorial visual content, such as a person's name or profession and the location of a given place. Such conceptual/abstract descriptions are available for familiar stimuli, whereas the NSD dataset used by[18–20], includes images of unfamiliar scenes, which were textually described only based on their visual content. We therefore propose that such linguistic form of visual information is aligned with visual content of the images, whereas the abstract textual description is not.

Accordingly, we distinguish between two types of linguistic descriptions of visual images: a linguistic description of the image's purely visual content, which we call *visual text*, and a linguistic description of conceptual knowledge that cannot be derived from the immediate visual attributes of the image (such as the name of the profession or the geographical location of a familiar person/place), which we call *abstract text* (see Fig 2A for an example). The first represents pictorial visual information coded in a linguistic form, whereas the second represents conceptual/abstract knowledge that cannot be derived from the image's immediate visual content and necessitates further linguistic knowledge[26]. We hypothesize that *visual text*, but not *abstract text*, will predict the correlation of neural responses in high-level visual cortex and LLM embeddings.

To test this hypothesis, we used visual and language deep learning algorithms to predict the iEEG responses to images of familiar faces or places (Figure 1). Familiar stimuli offer a convenient category in which visual and abstract/conceptual information can be clearly separated. We generated two types of textual descriptions, *visual text*, which describes content that can be derived from visual information in the image, and *abstract text*, based on the Wikipedia definition of the familiar image, which is independent of its visual content (see Figure 2; see supplementary material for visual and abstract textual descriptions of each image). We hypothesized that the response of high-level visual cortex to the familiar images would be aligned with LLM representations of their visual textual descriptions, replicating previous findings[19,24], but not with the LLM representations of abstract textual descriptions of the same images, in line with[25]. To further validate the visual and conceptual nature of the

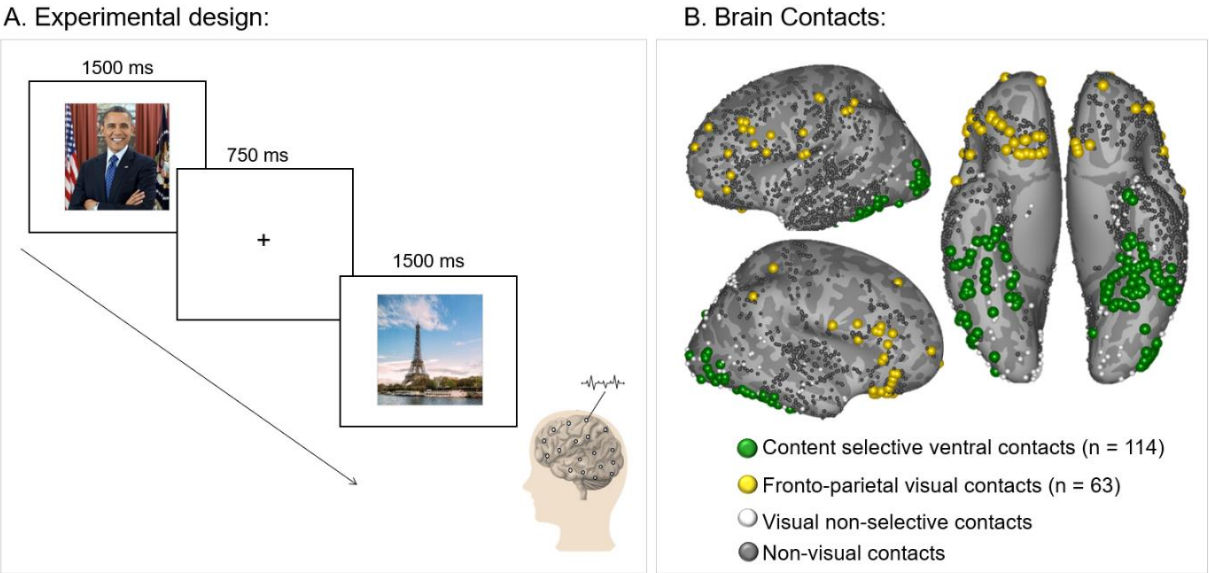


Figure 1: Experimental design: **A.** Participants viewed 28 images of familiar faces and places while their neuronal activity was recorded via intracranial EEG. **B.** Contact locations: yellow - Fronto-parietal visual contacts, green – Content-selective ventral contacts, white – visual non-selective contacts, gray- non-visual contacts.

DNNs representations, we also collected human similarity ratings. We asked human participants to rate the visual similarity of the same images, or the semantic similarity of their name labels and examined the correlations of these ratings with the visual and linguistic DNN representations.

Results:

Experimental design and intra-cranial contacts information:

Our study involved intra-cranial recordings, conducted for clinical purposes in 13 patients (10 females, mean age 34.7 ± 9.6) covering 2571 recording sites. During the experiment the participants viewed 28 images of well-known individuals and places (14 of each category), Each image was presented 4 times in semi-random order (for further details see Methods and [27,28]).

Figure 1 shows the experimental design (Fig. 1A) and recording sites (Fig. 1B) that were examined in the present study. First, we identified the visually responsive contacts (n=377, out of the 2571 total contacts) and then categorized them based on their anatomical locations (ventral, fronto-parietal) and functional characteristics, into different regions of interest (ROIs) (see methods of [27] for details). To test our hypothesis, we focused on the ventral visual ROI and used the fronto-parietal contacts as a control ROI.

The relational structure of the DNN representation of visual images is correlated with visual text but not abstract text.

For each of the familiar images we generated two types of textual descriptions: one that describes pictorial information that can be derived directly from the visual content of the image (visual text) (see Methods for further details) and another, abstract description, was obtained through their Wikipedia definition (abstract text). Careful reading of the Wikipedia-based texts verified their strong bias towards abstract meaning that cannot be derived from their visual appearance alone (See Figure 2A, and full set of textual descriptions in the Supplementary material). We then used visual and language DNNs to construct the visual and abstract representation of the stimuli, respectively. The representations of the images were derived from the final fully connected layer of an ImageNet trained DNN (VGG-16[29]). The representations of the visual and abstract text were extracted using the final layer of GPT2[30] (See Methods section for details).

We then generated a representational dissimilarity matrix (RDM) for each representation, by computing the cosine distance between the embeddings of the 28 images/textual descriptions for each input. This resulted in three different DNN-generated RDMs, one image-based (DCNN-Images), one visual-text based (LLM-Visual), and one abstract-text based (LLM-abstract). We first asked, by comparing the RDMs of the different networks, if the network-generated representations of the images are more similar to their network-generated visual-text representations compared to the representation generated by the abstract text descriptions. The zero order correlation between the representation of the

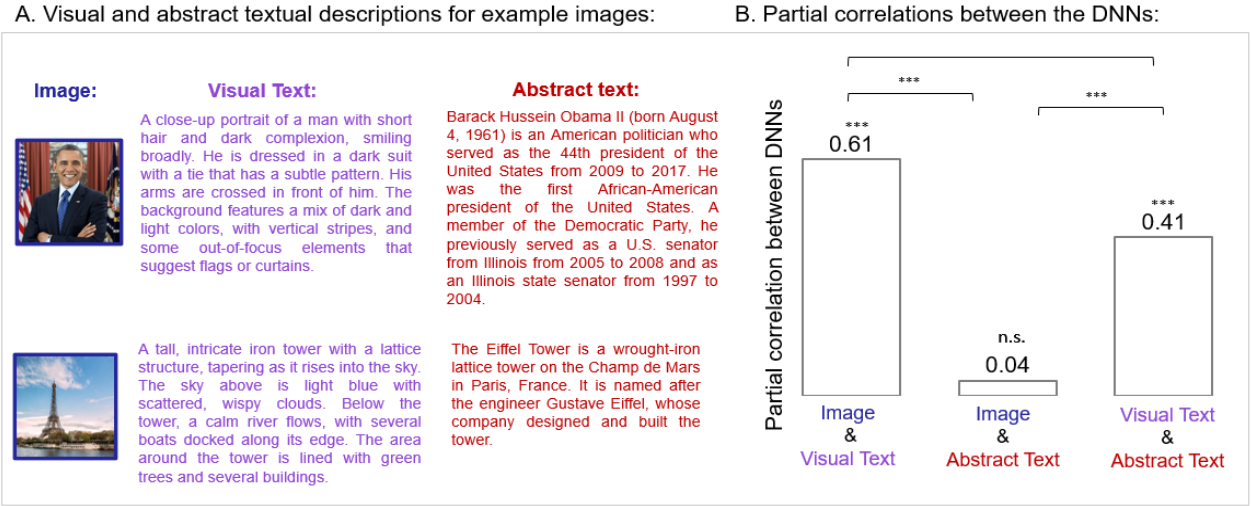


Figure 2: Examples of two images and their visual- and abstract-textual descriptions, and the partial correlations between their visual and language DNNs representations: A. Examples of images and their visual (purple) and abstract (red) textual descriptions: B. A pre-trainedobject DCNN (VGG-16) was used to extract the embeddings of the images (DCNN-image), and a large language model (GPT-2) was used to extract the embeddings of the visual and abstract textual descriptions. The partial correlations between the RDMs of the DNNs were high between images and visual text but low between images and the abstract text.

images with the visual text (r=0.69) was significantly higher than the correlation between the representations of the images with the abstract text (r=0.41), and the correlations between the two types of texts (r=0.55). To measure the unique contribution of each DNN, we calculated the partial correlations between each pair of DNNs representations when the third DNN is held constant. As shown in Figure 2B, the partial correlation between the images and the abstract text (when removing their correlation with the visual text) is not significant (0.04), indicating that the abstract textual descriptions are independent of the visual content of the images. The partial correlation of the visual text was higher with the images (0.61) than with the abstract text (0.41). These findings demonstrate a significant correspondence in the artificial networks between the representations of images and their visual textual descriptions, but no association between the representations of images and their abstract textual descriptions.

Visual ventral cortex is highly correlated with DNN's image based and visual-text based representations, but not with abstract-text based representation.

To examine the similarity between the organization of the three representations generated by DNNs and the representation of ventral visual cortex of the same images, we created a representational dissimilarity matrix (RDM) based on the iEEG responses for the different stimuli. For each stimulus, we calculated the average neural response (amplitude of the broad band high frequencies of the iEEG signal in each contact [28]) over the four stimulus repetitions, within the 0.1 to 0.4 second window after stimulus onset, in each contact in the ROI. We selected this time window because it captures the peak neural visual response[28]. We used the averaged iEEG derived population vector as the response pattern of each stimulus within the ROI. We then measured the Pearson correlations between the activation patterns of each pair of stimuli. The dissimilarity between the neuronal representations of two stimuli was represented as one minus the Pearson correlation value (thus lower values indicate greater similarity).

Next, we compared the relational-structure similarities between the response of the visual ventral cortex to the images, and the three artificial representations generated by the DNNs for the same stimuli seen by the patients. The similarities between the brain-based and DNN-based relational structures were computed as the Pearson correlation between the brain and the respective DNN-based RDMs. To test for statistical significance, we computed the same correlations using a permutation test (shuffled pairs labels, see methods for further details). We tested both the significance of each predictors' (DNNs) correlation individually

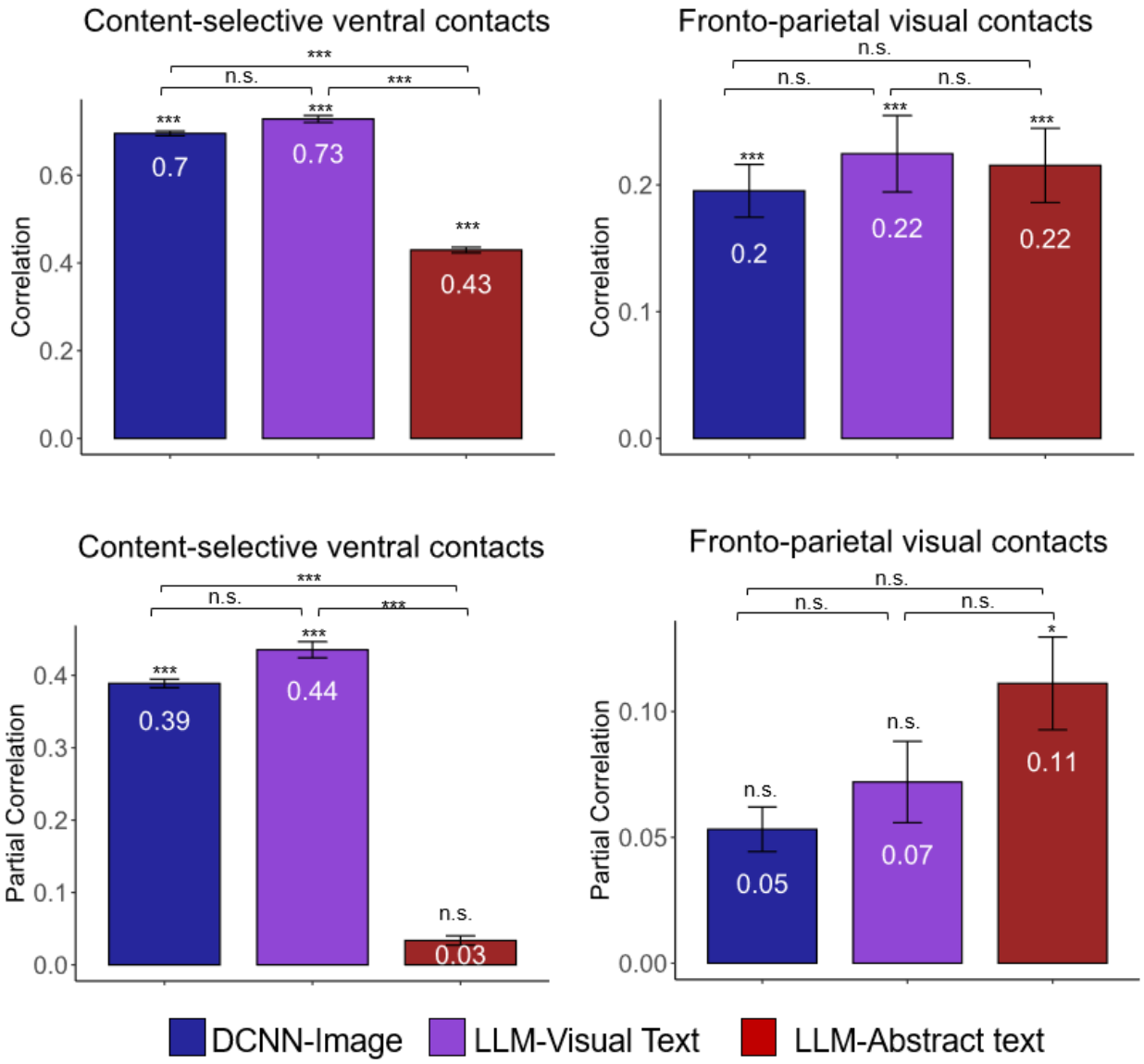


Figure 3: Correlations between the representations of ventral content-selective and fronto-parietal visual contacts to images and the representations of the images by: visual DNN (DCNN-image), visual textual descriptions (LMM-visual text), and abstract textual descriptions (LMM-abstract text): Zero order correlations of each DNN RDM's with RDMs from brain ROIs are presented in top panels; partial correlations with the RDMs of brain activity patterns, when the other two DNNs are held constant, are presented in bottom panels. Left: the correlations and partial correlations with the Content selective ventral contacts (n=114). Right: correlations with the 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 two-sided permutation test (10000 permutations). Reported *p* values are FDR corrected in each ROI separately. Note the clear drop in correlation from visual DCNN and LLM-Visual to LLM-Semantic in the ventral contacts. $* \, p_{FDR} < 0.05$, $** \, p_{FDR} < 0.01$, $*** \, p_{FDR} < 0.001$

and the difference between each correlation in each ROI. P-values were FDR corrected for each ROI and each analysis (individual predictors and differences) separately.

Our analysis shows that the neural response of the visual ventral cortex was similarly correlated with the visual representation of the image (DCNN-Image), and the visual-text representation (LLM-visual) but showed a significantly lower correlation with the abstract text representation (LLM-abstract) (Fig. 3 left top panel). The same analysis with the responses of the frontal contacts to the images showed significant correlations with all three DNN-based representations, but no significant difference between them. To examine the unique contribution of each DNN to the response of the ventral visual cortex to the images, we calculated the partial correlations between each DNN-based representation and the RDM derived from the visual cortex, when the two other representations are held constant.

Both the image-based (DCNN-Image) and the visual-text based (LLM-Visual) representations uniquely predicted the ventral visual cortex response to the images, whereas the abstract text did not account for any additional information. There was no significant difference between the contributions of the image and the visual-text, and they were both significantly higher than the abstract text (Fig. 3 left bottom panel). The response of the frontal visual contacts showed the reverse pattern, where only the abstract-text representation significantly accounted for unique information in their neural response (Fig. 3 right bottom panel).

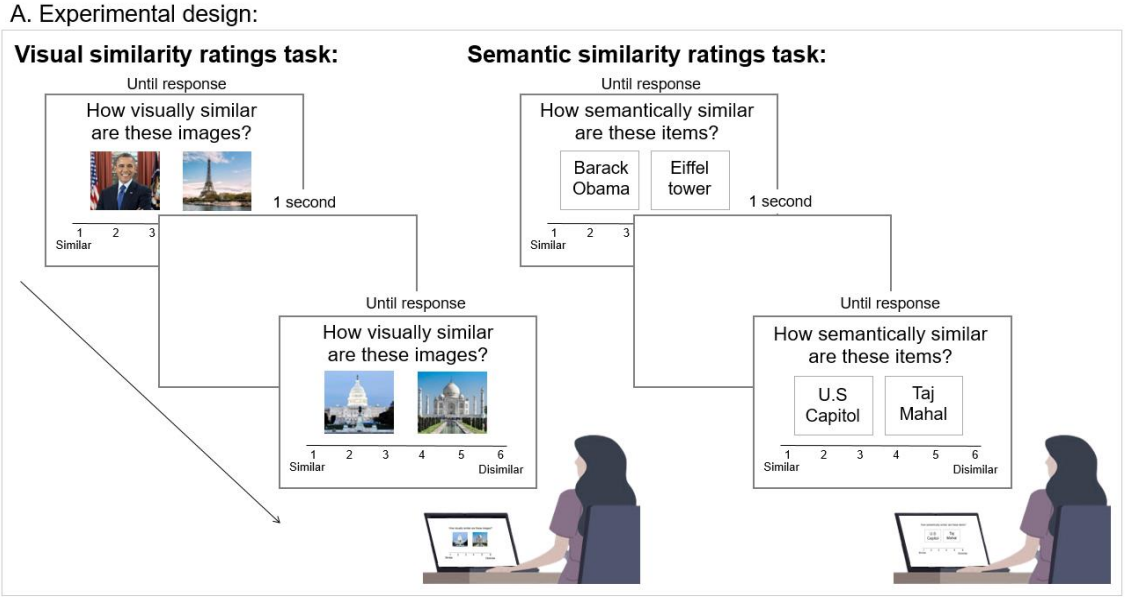
Visual-text representation is more correlated with visual than semantic similarity ratings

To further examine if the visual and abstract textual descriptions represent subjects' visual or semantic information about the images, we asked human participants to rate the visual or semantic similarity of all possible pairs of the 28 stimuli. One group of participants (n = 20) was asked to rate the visual similarity between each pair of images based on their visual appearance. Another group of participants (n = 20) was presented with the names of the stimuli (the name of the identity or place presented in the image) and was asked to rate the semantic similarity between each pair of items based on the conceptual knowledge they had about them (see Fig. 4A; see Methods for further details). For each participant we created an RDM of the stimuli, based on the participant's similarity ratings. We calculated the correlation between each of the participant's RDM and each of the three representations generated by the DNNs. All correlations were fisher transformed. To test the significance of each correlation we performed a one sample t-test (two-sided) and used FDR to correct for multiple comparisons (the Fisher transformed correlations were used to infer statistics, Fig 4 shows raw correlations).

As shown in Fig. 4, the representation of the visual and visual-text DNNs were more correlated with human visual than semantic similarity rating, whereas the representation of the abstract-text DNN showed no such difference.

To test the statistical significance of these pattern of correlations, we performed a mixed ANOVA with DNN-representation (image, visual-text, abstract-text) as within-participant factor, and task (visual similarity rating, semantic similarity rating) as between-participants factor, on the Fisher transformed correlations. The White correction was used to adjust for heteroscedasticity. Analysis revealed a significant main effect for DNN ($F(1.19, 42.9) = 55.27$, $p < 0.001$, $\eta_p^2 = 0.60$) and for Task ($F(1,36) = 5.39$, $p = 0.026$, $\eta_p^2 = 0.13$), and a significant interaction between them $F(1.19, 42.9) = 39.28$, $p < 0.001$, $\eta_p^2 = 0.52$). Post-hoc comparison between visual and semantic rating tasks for DNN-based representation (FDR corrected) revealed that visual and visual-textual representations were more correlated with visual than semantic similarity ratings ($t (108) = 5.31$, $p < 0.001$), $t (108) = 2.51$, $p = 0.013$, respectively), but abstract-textual representations showed no significant difference between visual and semantic rating tasks ($t (108) = 1.60$, $p = 0.12$). Post-hoc comparison between DNNs correlations within each similarity task (FDR corrected) revealed that human visual similarity ratings were more correlated with the image-based representation, which was significantly higher than the visual-text ($t (108) = 3.35$, $p = 0.001$) and abstract-text descriptions ($t(103) = 7.57$, $p < .001$). The correlation with the abstract text descriptions was significantly lower than with the visual text descriptions $t(108) = 4.22$, $p < .001$). Post-hoc comparison did not reveal any difference between the correlations of the DNNs' representations and the human semantic ratings. These results are in line with our conclusion that, similar to the visual-image representation, visual-text representations are aligned with visual rather than semantic information.

A. Experimental design:



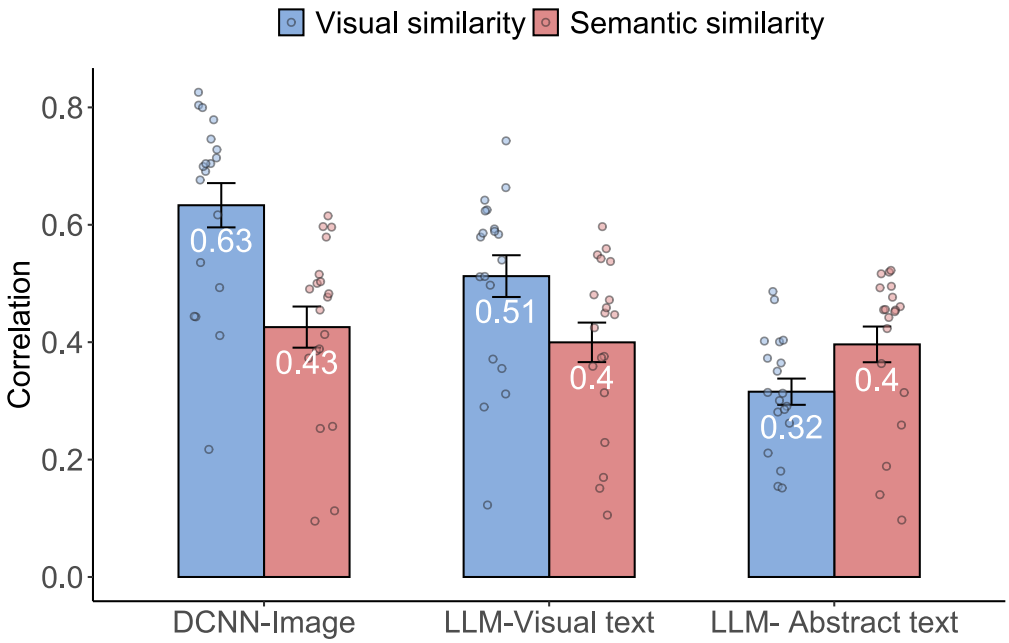


Figure 4: Experiment design and results of human similarity ratings: a. Experimental design: Participants rated the visual similarity (n = 20) of famous faces and places while presented with their images that were used in the iEEG study (left) or the semantic similarity (n = 20) between the names of the same stimuli presented in the images (right) ***b.*** Correlations between DCNN-images, LLM-visual text and LLM-abstract text representations with human visual similarity ratings of the images (blue) and semantic similarity ratings of names (red). Error bars are s.e.m. between participants. *p* values were derived from a single correlation two-sided t-test and post hoc compression between task and predictors. Reported *p* values are FDR corrected.$** \; p_{FDR} < 0.01 \; *** \; p_{FDR} < 0.001$

Discussion

Recent fMRI studies show that linguistic representations generated by large language models predict the neural response to images in high level visual cortex[19–21]. These findings suggest a significant alignment between visual and linguistic representations. However, these studies leave open the question concerning the nature of linguistic information. Does it reflect purely visual, pictorial aspects of the images or also more abstract, conceptual knowledge? This is an important question, as it determines the interpretation of this surprising result, regarding the information guiding the organization of high-level visual cortex. Here we address this question by separately examining two types of linguistic representations, one that describes information that can be extracted from the image and the other that is independent of the pictorial visual content of the image and conveys abstract knowledge about it that can only be gained through non-visual channels (Figure 2). We hypothesized that only the former is aligned with neural responses in visual cortex. To test this hypothesis, we examined the neural responses to images of famous faces and places. For this kind of stimuli, abstract information, such as the name or occupation of the person or the name and location of a place, is independent of the visual appearance of a specific image (see Fig 2A and Supplementary Table 1). We then used a visual DCNN to generate an image-based representation, and a LLM to represent the visual and abstract textual descriptions; next, we tested the alignment of each with the high-level visual cortex representations. Our findings show that visual and visual-textual representations similarly predicted the neural response of high-level visual cortex to images, whereas abstract textual representations showed significantly lower correlation and did not explain additional variance. These findings support our hypothesis that the alignment between linguistic information and the high-level visual cortex response to images, is restricted to textual descriptions of visual information. It further validates the important distinction between visual and abstract linguistic representations.

To further examine if the visual, visual-textual and abstract-textual representations encode the subjects' visual or semantic information, we asked human participants to rate the visual or semantic similarity of the same images. Our findings show that visual similarity ratings were more correlated with visual and visual-textual representations than with abstract-textual representations (Figure 4). This pattern of correlations corresponds to the pattern of correlation in high-level visual cortex (Figure 1B). Moreover, the correlation of visual-textual representation was more correlated with visual similarity ratings than with semantic similarity ratings, which further emphasizes the visual nature of this representation.

Our findings that visual textual representations are aligned with the neural response to images are consistent with recent findings that show that linguistic descriptions of scenes predict the neural response of high-level visual cortex to the corresponding images[18–21]. These verbal descriptions depict categorical information that can be extracted from visual inspection of the image. The question of whether high-level visual cortex represents visual features or semantic categories has been extensively debated (for reviews see [3,10]). Several studies have suggested that low level visual features such as eccentricity, spatial frequency[31] aspect ratio[32] or curvature[33] underlie the large-scale functional organization of high-level visual cortex. A main challenge in dissociating between visual features and semantic categories is that they are not orthogonal. For example, most fruits are round or curvy whereas most tools are long and straight. To disentangle between shape and category, Bracci and Op De Beeck (2016)[34] measured the fMRI response to objects from different categories (e.g., animate, musical instruments, tools, food) that are matched as much as possible in shape. They found that the organization of the visual cortex transitions from shape, in posterior low level visual areas, to category representations, in lateral and ventral occipito-temporal visual brain areas. They further conjecture that selectivity to visual features in these regions may primarily serve a possible goal of the system to classify images to different categories.

While our iEEG study emphasizes visual over semantic organization principles in the local organization of high order visual areas this finding does not necessarily contradict a role for such information at coarser, large scale, levels e.g. those revealed in fMRI studies. For example, in an fMRI study, Carlson et al (2014)[22] measured the response of the visual cortex when participants viewed images of various categories (human/animal faces, natural/man-made objects). They used several measures of semantic representations to predict the brain response to the images, such as human semantic similarity judgments about the same objects and the co-occurrences of the names of these objects in a large corpus of texts (the frequency with which two words appear together in a corpus of text). Results showed that word co-occurrences and behavioral semantic similarity judgments predicted the representational geometry of the neural response to the coarse distinction between cortical areas preferring animate vs. inanimate categories in the inferior temporal cortex. Popham and colleagues (2021)[23] extended these findings by showing participants movies depicting natural scenes of people, animals, objects, and more. They measured semantic similarity between scenes based on word co-occurrence of the names of objects and actions depicted in each movie. They discovered that for each category-selective visual ROI in the visual cortex, there is an adjacent category-selective semantic ROI for the same category, suggesting visual-semantic alignment at the level of category selective cortical areas.

However, it is important to emphasize that such large scale studies do not address the local relational structures within each individual areas- which show exemplar specificity within a category (e.g faces [16,25])

Our partial correlation analysis further shows that visual and visual-textual information contributed distinct proportion of variance in the neural response to images. This suggests that in addition to their aligned representations they also capture different aspects of the information contained in the images. Thus, our findings suggest that there is additional unique information that visual and linguistic representations convey about the visual content of an image beyond their aligned representations. This finding is relevant to the question of whether sensory and linguistic information about the physical world reflects overlapped or distinct types of information, which has been highly debated. Embodied cognition theories posit that all information is grounded by sensory information[35,36], whereas dual-code models suggest a distinction between visual and language inputs[37]. The findings that the representational geometry of color in congenitally blind individuals is very similar to the representation of sighted individuals in specific parietal cortical regions demonstrates that visually-aligned linguistic representations can be generated in the absence of visual input in specialized cortical regions[37].

An interesting issue, supported by our current study, is the intriguing convergence into a unified relational structure of visual information delivered via images on the one hand and the same information delivered through textual means. This result is in line with the recent study of Conwell and colleagues (2024) [18] that found that pure language models highly predicted the ventral visual cortex activity of macaque, a species that has no language. Their findings demonstrate that relational structures of LLMs converge across linguistic and visual dimensions.

Our distinction between visual and abstract linguistic information may correspond to the distinction between concrete/grounded vs. abstract words, respectively, which has been extensively studied for several decades in the language domain[38,39]. These studies reveal substantial cognitive and neural differences between the representations of concrete/grounded words, such as names of objects and abstract words, such as justice and love. However, the comparison between concrete and abstract words inherently involves comparing very different concepts, whereas the visual and abstract textual descriptions used in the current study are linguistic descriptions of the same stimulus. Future studies that concern the difference between grounded and abstract linguistic information [40–42] may use our approach to compare the neural representations of these two types of verbal information that describe the same concept.

In conclusion, our study suggests that the alignment between the neural response of high-level visual cortex to images and linguistic representations is limited to pictorial visual information presented in a linguistic code. Specifically, we demonstrate a clear neural distinction between visual-textual and abstract-textual representations. Visual-textual information was strongly aligned with neural relational structures in high-level visual cortex, whereas abstract-textual information was not. These findings demonstrate the specialized role of the visual cortex in processing information directly derived from visual stimuli.

Methods

Stimuli:

28 color images of famous identities and famous places, 14 of each category. See Supplementary Material for the list of all identities and places.

iEEG data: experimental design and intra-cranial contacts information:

We describe here the participants tasks and stimuli as well as ROI definitions. Additional details can be found in [27].

Participants

Electrophysiological data from 13 patients (10 females, mean age 34.7 ± 9.6) were collected via intracranial EEG during pre-surgical evaluation for drug-resistant epilepsy at North Shore University Hospital, NY. As part of the clinical assessment, subdural or depth contacts were implanted in all patients. All patients gave fully informed consent, including consent for publication of the results. No clinical seizures occurred during the experiment. The Feinstein Institute for Medical Research's IRB review board monitored the study, and the study followed the latest declaration of Helsinki and US National Institute of Health guidelines.

Experimental task and stimuli

The experiment was composed of two runs in which the participants viewed 28 images of famous faces and places, 14 images of each category. Each run started with a 200-second resting phase with closed eyes. Then, the participants were presented with the images, each image was presented for 1500 ms, with 750 ms inter-stimulus intervals. In each run, the participants saw 14 different images (7 of each category), and each image was presented four times in a semi-random order to avoid consecutive repeats. The participants were instructed to memorize the details of the images, as they will be required to recall the images and describe their visual features in a later recall phase. The recall phase is not analyzed in this study. It is noted that the participants performed well, and no participants were excluded based on performance (Further details can be found in [27]).

Definition and grouping of visually responsive contacts:

The analyses in this study were conducted only on visually responsive recording sites (contacts), subdivided into different ROIs. Therefore, we extracted out of all 2571 recording sites the visually responsive contacts, which we defined as electrodes displaying statistically significant activations in response to the images. The data from all 2571 recording sites was preprocessed and transformed to the high-frequency broadband (HFB) signal (see [28] for further details). We averaged the HFB response across the 100 to 500 ms period following stimulus onset and compared each contact's post-stimulus response to the baseline response prior to stimulus onset. The pre-stimulus baseline was then averaged across the -400ms to -100ms interval prior to stimulus onset. We compared the response to the stimuli and the baseline response, using a two-tailed Wilcoxon signed-rank test, that was FDR corrected for all contacts (from all patients) together. We defined visually responsive contacts as contacts demonstrating a significant HFB difference (p_{fdr} < 0.05), resulting in 377 visual contacts.

Then we divided the visually responsive contacts based on anatomical and functional characteristics, to content-selective ventral and fronto-parietal visual contacts. To define the content selective contacts, we examined the visually responsive contacts in six anatomical regions across the ventral visual hierarchy, based on the Desikan Killiany atlas. These regions included the lateral occipital cortex (LO), inferior temporal gyrus (ITG), lingual gyrus, parahippocampal gyrus (PHG), fusiform gyrus, and entorhinal gyrus, excluding early visual contacts. Then, for each contact, we calculated the averaged HFB response to each stimulus during the 100-500ms post-stimulus window. Contacts in the selected regions that showed significant content selectivity, defined as a difference of at least 3.5 SD between the top 10 preferred images (averaged response) and bottom 10 images (averages response), were included in the content-selective contact set (n = 114). 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 1B). Any visually responsive contacts not falling into these specified categories were termed "other visually-responsive contacts" and are depicted in white in Figure 1B, while non-visual contacts are depicted in gray. Further details can be found in [28].

Visual, visual-text and abstract-text representations using DNNs:

We used three different representations that are based on different inputs:

1. Visual (images) - the visual representation of the image stimuli based on a DCNN pre-trained on ImageNet.
2. Visual descriptions (text) - visual descriptions of the image stimuli. The visual descriptions were generated by Chat-GPT[43]: We uploaded each image to chat GPT and used the following prompt: "Please describe the image based on its visual features". In case the text included the names or the location of the image, we removed it from the text (or changed it to he/she/it). In addition, we removed text that described the image overall atmosphere. A total of 4 such sentences were removed from the descriptions of all 28 images. For visual descriptions of the images see Fig 2B (for an example) and Supplementary material (for all visual descriptions).
3. Abstract descriptions (text) – the first paragraph of the Wikipedia descriptions of the identities\places depicted in each image (see Fig 2B for an example and Supplementary material for all abstract descriptions).

We extracted the representation of each input type using the following models:

The representations of the images were extracted using a visual deep learning algorithm - VGG-16 DCNN[29] pre-trained on ImageNet[44], which includes 300000 images from 1000 object categories. VGG is a commonly used network that was shown to be correlated with behavioral and neural responses to the same images[45–48]. We then extracted the embeddings of each image based on the feature vector representation in the penultimate- i.e. the last hidden layer (fc7) just before the output layer of the network, which is the representation that is used for the classification. The similarity between each pair of images was computed based on the cosine distance between these feature vectors.

The representations of the text input were extracted using a large language model - GPT2: A transformer-based language algorithm[30]. 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. We retrieved the embeddings 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. We extracted GPT representation once based on the visual textual descriptions and once based on the abstract textual descriptions, resulting in two different RDMs that were generated with GPT.

Behavioral data: Human similarity ratings:

<u>Participants</u>

A total of 40 participants were recruited for this study from the Prolific platform. Twenty participants were assigned to each of the two experimental conditions: Visual similarity based

on images and semantic similarity based on names (mean age 29 (sd = 2.9), 20 females, 1 preferred not to say). The participants were paid 6 GBP for their participation in the experiment (9 GBP/hour). They gave informed consent prior to the experiment. The study was approved by the ethics committee of Tel Aviv University. We chose this sample size based on previous similarity tasks (Shoham et al)

Procedure

Participants rated the visual or semantic similarity of all possible pairs of the 28 items (378 pairs).

Visual similarity rating: Each trial presented one pair of images of two different items, and the participants were asked to rate the visual similarity between the items. The pairs were presented on the screen one at a time, above a similarity scale (1 (very similar) - 6 (very dissimilar)) until response. Above the images, the participants saw the question “How visually similar are these images?”. The participants selected the similarity score with the mouse. The next pair was presented 1 second after their response. The participants had a forced break for a minimum of 10 seconds after 80 pairs were presented in a row (a total of 4 breaks during the experiment). After they rated all 378 pairs, the participants were asked to indicate for each image whether they were familiar with it before the experiment. The experiment lasted about 40 minutes.

Semantic similarity rating: Participants were asked to rate the similarity between the names of the stimuli based on the conceptual/semantic knowledge they had about them. The timing and procedure were identical to the visual similarity task, beside that in the semantic rating task the instruction was only presented at the beginning and there was no line of instruction in each trial.

Trials’ exclusion:

We excluded trials if the participant was unfamiliar with one of the items. Participants who were unfamiliar with 30% (or more) of the identities were excluded from the analysis. We excluded trials shorter than 200 ms and longer than 30,000 ms (30 seconds) based on the assumption that participants did not perform the task well on such trials. In the visual similarity tasks, 23% of the trials were excluded (33 trials due to RT longer than 30 seconds and the rest based on the familiarity criterion). In the semantic similarity rating task, 15% trials were excluded (50 trials due to RT longer than 30 seconds and the rest based on the familiarity criterion).

Data Analysis:

We used Representational Similarity Analysis (RSA) to predict the neural and behavioral measures based on the DNNs representations. We created a dissimilarity matrix (RDM) of all possible stimuli pairs within the pool of the 28 stimuli (a total of 378 pairs), separately for

high-level visual cortex and the fronto-parietal ROIs, for the visual, visual-text and abstract-text representations and for each participant in the visual and semantic behavioral ratings tasks.

Correlation between the DNN representations and brain ROIs:

We created a representational dissimilarity matrix (RDM) based on the iEEG responses for the different stimuli. For each stimulus, we calculated the average response (mean amplitude of the broad band high frequencies of the iEEG signal in each contact over the four stimulus repetitions, within the 0.1 to 0.4 second window after stimulus onset) for each contact in the high-visual content-selective and fronto-parietal ROI[28]. We used the averaged iEEG derived population vector as the response pattern of each stimulus within each ROI. The dissimilarity between the neuronal representations of two stimuli was represented as one minus the correlation between their activation patterns. We first computed the Pearson correlation between the DNNs' RDMs and the RDMs of the ventral and frontal visual contacts. Then, to further assess the unique contribution of each DNN-based representation 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 value of a single network in a specific ROI we used a permutation test, in which the similarity scores of the neural responses were shuffled and the DNNs' similarity scores (between pair of stimuli) were held constant. 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 was the proportion of iterations in which the original correlations/partial correlations (based on the real similarities) were greater than the value that was calculated in each iteration. P values were adjusted according to Phipson and Smyth[49] 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 to infer significance. We also tested the significance of the difference between each pair of two DNN-based representations. To do so we used a permutation test in which the networks based 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. We first calculated the correlation/partial correlation (respectively) of each network with the neural RDM (based on the shuffled similarities). We then calculated the difference between each two network representations 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 the same direction) 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[49] 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 to infer significance.

Correlation between the DNNs representations and behavioral similarity ratings:

We created a representational dissimilarity matrix (RDM) for each participant based on their similarity ratings. We computed the correlations between the RDM of each participant and the RDM of the DNNs representations. All correlations were fisher transformed. To test for significance, ANOVA, post hoc comparisons and one-sample t-tests were performed on the Fisher's z-transformed correlations using two-sided null-hypothesis tests. FDR was applied to correct for multiple comparisons for all post-hoc tests.

References:

1. Op de Beeck, H. P., Haushofer, J. & Kanwisher, N. G. Interpreting fMRI data: maps, modules and dimensions. *Nat. Rev. Neurosci.* **9**, 123–135 (2008).
2. Grill-Spector, K. & Malach, R. The human visual cortex. *Annu Rev Neurosci* **27**, 649–677 (2004).
3. Grill-Spector, K. & Weiner, K. S. The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* **15**, 536–548 (2014).
4. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* **17**, 401–412 (2013).
5. Peelen, M. V. & Downing, P. E. Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia* **105**, 177–183 (2017).
6. Pitcher, D., Walsh, V. & Duchaine, B. The role of the occipital face area in the cortical face perception network. *Exp. Brain Res.* **209**, 481–493 (2011).
7. Duchaine, B. & Yovel, G. A Revised Neural Framework for Face Processing. *Annu. Rev. Vis. Sci.* **1**, 393–416 (2015).
8. Levy, I., Hasson, U., Avidan, G., Hendler, T. & Malach, R. Center–periphery organization of human object areas. *Nat. Neurosci.* **4**, 533–539 (2001).

9. Contier, O., Baker, C. I. & Hebart, M. N. Distributed representations of behaviour-derived object dimensions in the human visual system. *Nat. Hum. Behav.* (2024) doi:10.1038/s41562-024-01980-y.
10. Bracci, S. & Op De Beeck, H. P. Understanding Human Object Vision: A Picture Is Worth a Thousand Representations. *Annu. Rev. Psychol.* **74**, 113–135 (2023).
11. Simony, E., Grossman, S. & Malach, R. Brain–machine convergent evolution: Why finding parallels between brain and artificial systems is informative. *Proc. Natl. Acad. Sci.* **121**, e2319709121 (2024).
12. Kriegeskorte, N. Deep neural networks: a new framework for modelling biological vision and brain information processing. Preprint at https://doi.org/10.1101/029876 (2015).
13. Doerig, A. *et al.* The neuroconnectionist research programme. *Nat. Rev. Neurosci.* **24**, 431–450 (2023).
14. 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.
15. Dobs, K., Martinez, J., Kell, A. J. E. & Kanwisher, N. Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci Adv* **8**, 8913 (2022).
16. Grossman, S. *et al.* Deep Convolutional modeling of human face selective columns reveals their role in pictorial face representation. Preprint at https://doi.org/10.1101/444323 (2018).
17. Jozwik, K. M., Kietzmann, T. C., Cichy, R. M., Kriegeskorte, N. & Mur, M. Deep neural networks and visuo-semantic models explain complementary components of human ventral-stream representational dynamics.
18. Conwell, C. *et al.* Is visual cortex really “language-aligned”? Perspectives from Model-to-Brain Comparisons in Human and Monkeys on the Natural Scenes Dataset. *J. Vis.* **24**, 1288 (2024).
19. Doerig, A. *et al.* Visual representations in the human brain are aligned with large language models.

20. Conwell, C., Prince, J., Alvarez, G. & Konkle, T. The Unreasonable Effectiveness of Word Models in Predicting High-Level Visual Cortex Responses to Natural Images. in *2023 Conference on Cognitive Computational Neuroscience* (Cognitive Computational Neuroscience, Oxford, UK, 2023). doi:10.32470/CCN.2023.1642-0.
21. Garcia, K., McMahon, E., Conwell, C., Bonner, M. F. & Isik, L. Modeling dynamic social vision highlights gaps between deep learning and humans. Preprint at https://doi.org/10.31234/osf.io/4mpd9 (2024).
22. Carlson, T. A., Simmons, R. A., Kriegeskorte, N. & Slevc, L. R. The Emergence of Semantic Meaning in the Ventral Temporal Pathway. *J. Cogn. Neurosci.* **26**, 120–131 (2014).
23. Popham, S. F. *et al.* Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nat. Neurosci.* **24**, 1628–1636 (2021).
24. Allen, E. J. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).
25. Shoham, A., Broday-Dvir, R., Yaron, I., Yovel, G. & Malach, R. Text-related functionality of visual human pre-frontal activations revealed through neural network convergence. *bioRxiv* 2024–04 (2024).
26. Reilly, J. *et al.* What we mean when we say semantic: Toward a multidisciplinary semantic glossary. *Psychon. Bull. Rev.* (2024) doi:10.3758/s13423-024-02556-7.
27. Norman, Y. *et al.* Hippocampal sharp-wave ripples linked to visual episodic recollection in humans. *Science* **365**, (2019).
28. 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).
29. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. (2014) doi:10.48550/arXiv.1409.1556.
30. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving Language Understanding by Generative Pre-Training.

31. Davidesco, I. *et al.* Spatial and object-based attention modulates broadband high-frequency responses across the human visual cortical hierarchy. *J. Neurosci.* **33**, 1228–1240 (2013).
32. Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
33. Tsao, D. Y., Cadieu, C. & Livingstone, M. S. Object recognition: physiological and computational insights. *Primate Neuroethol.* 471 (2008).
34. Bracci, S. & Beeck, H. Dissociations and associations between shape and category representations in the two visual pathways. *J. Neurosci.* **36**, 432–444 (2016).
35. Foglia, L. & Wilson, R. A. Embodied cognition. *Wiley Interdiscip. Rev. Cogn. Sci.* **4**, 319–325 (2013).
36. Leitan, N. & Chaffey, L. Embodied cognition and its applications: A brief review. (2014).
37. Wang, X., Men, W., Gao, J., Caramazza, A. & Bi, Y. Two Forms of Knowledge Representations in the Human Brain. *Neuron* **107**, 383-393.e5 (2020).
38. Paivio, A. Dual coding theory: Retrospect and current status. *Can. J. Psychol. Can. Psychol.* **45**, 255 (1991).
39. Sadoski, M. & Paivio, A. *Imagery and Text: A Dual Coding Theory of Reading and Writing*. (Routledge, 2013).
40. Barsalou, L. W. Grounded Cognition. *Annu. Rev. Psychol.* **59**, 617–645 (2008).
41. Leshinskaya, A. & Caramazza, A. For a cognitive neuroscience of concepts : Moving beyond the grounding issue. *Psychon Bull Rev* **991–1001**, (2016).
42. Martin, A. GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychon. Bull. Rev.* **23**, 979–990 (2016).
43. OpenAI. ChatGPT. (2023).
44. 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.

45. 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).
46. 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).
47. 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).
48. 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).
49. 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).