High-level visual cortex representations are organized along visual rather than abstract principles.

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

Short title: Visual cortex encodes visual, not abstract, information

Abstract:

A fundamental question dominating the study of human visual cortex is whether it is organized along visual or semantic information. This question is unresolved, and the controversy has been rekindled by the recent report that surprisingly, revealed that the representations of textual description of images by linguistic artificial networks' successfully predict the response of high-level visual cortex to visual images. These findings appear to support a linguistic, abstract, organizing principle of human visual cortex. Here, using iEEG recordings from high level visual cortex in patients, we contributed to this debate, by testing the hypothesis that this linguistic alignment is restricted to textual descriptions of the visual content of the images (visual text) and does not extend to abstract textual descriptions (abstract text). We selected images that depict familiar faces and places, as these images allow for the best dissociation between these two types of text and generated their visual and abstract (e.g., name and biography of a person) textual descriptions. We then predicted the relational structures of the iEEG response to the images using their textual representations based on a large language model and the image representation based on a convolutional neural network. Neural relational-structures in high-level visual cortex were similarly predicted by images and visual-text but not abstract-text representations. Abstract text best predicted responses of the fronto-parietal cortex to the images. 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 responses 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). While there is substantial consensus with regards to the anatomical locations of the various visual representations- the principles by which these representations are organized remain controversial. Specifically- it is still debated whether the neural responses of high-level visual cortex are organized according to visual, pictorial properties or along semantic, abstract information. This question reflects on the general issue of cortical domain specificity and modularity and consequently 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 representation of the neural response to the same images in early and high-level areas of visual cortex[12,14–17]. Intriguingly, 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].

In particular, 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 as well as the linguistic representations of the LLMs[20]. Using the same dataset, Doerig and colleagues (2024) used a searchlight analysis on the fMRI -derived maps 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 ventro-temporal 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.

However, the textual descriptions used in these studies[18–20] were largely limited to the visual content of the images. Yet, visual images can also be described using textual information that is independent of their visual content, conveying conceptual or abstract knowledge acquired about them. Thus, it remains unclear whether the visual-language alignment observed in high-level visual cortex may be restricted to visual text or could extend also to abstract textual descriptions that are independent of the visual content of the images. In the present study we distinguish between two types of linguistic descriptions of visual images: a linguistic description of the image's purely visual content, which we term *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 term *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[25]. We hypothesize that *visual text*, but not *abstract text*, will predict the correlation of neural responses in high-level visual cortex and LLM embeddings.

Our hypothesis is partially supported by a recent iEEG study [26] that used visual DCNN representations and linguistic LLM representations to predict the neural responses to familiar faces and places. In this study, the linguistic descriptions were based on Wikipedia entries for the familiar stimuli (see Fig. 2A), which primarily contained conceptual information unrelated to the visual content of the images. The results showed a strong correlation between neural responses to images and visual DCNN representations but a weak correlation with LLM representations of abstract text[26]. Here we hypothesize that the linguistic form of visual information is aligned with visual content of the images, whereas the abstract textual description is likely represented in fronto-parietal visual areas. To test this hypothesis, both visual and abstract textual descriptions should be generated for the same images and used to predict their neural responses in the visual cortex.

To that end, we used visual and language deep learning algorithms to predict the iEEG responses obtained from high order visual areas in a unique data set containing images of familiar faces or places (Figure 1). Familiar faces and places are advantageous to test our hypothesis as they enable us to generate abstract descriptions that are independent of the visual appearance of the image. 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[26]. To further validate the visual and conceptual nature of the DNNs representations, we also collected human similarity ratings for the same set of images. 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. To extend these findings beyond the limited set of stimuli used in the iEEG study, we replicated the similarity rating study with another set of visual images.

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]). It should be noted that the IEEG method, while providing superior spatio-temporal resolution of the neuronal signals, imposes strict limits on the duration of the experiments- hence the limited number and categories of visual images that could be used in the study.

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 (ventro-temporal, 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 ventro-temporal visual ROI and used the fronto-parietal visual 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). To assure that the textual descriptions are uniquely associated with each of the images we asked participants to match each text to one of the images from

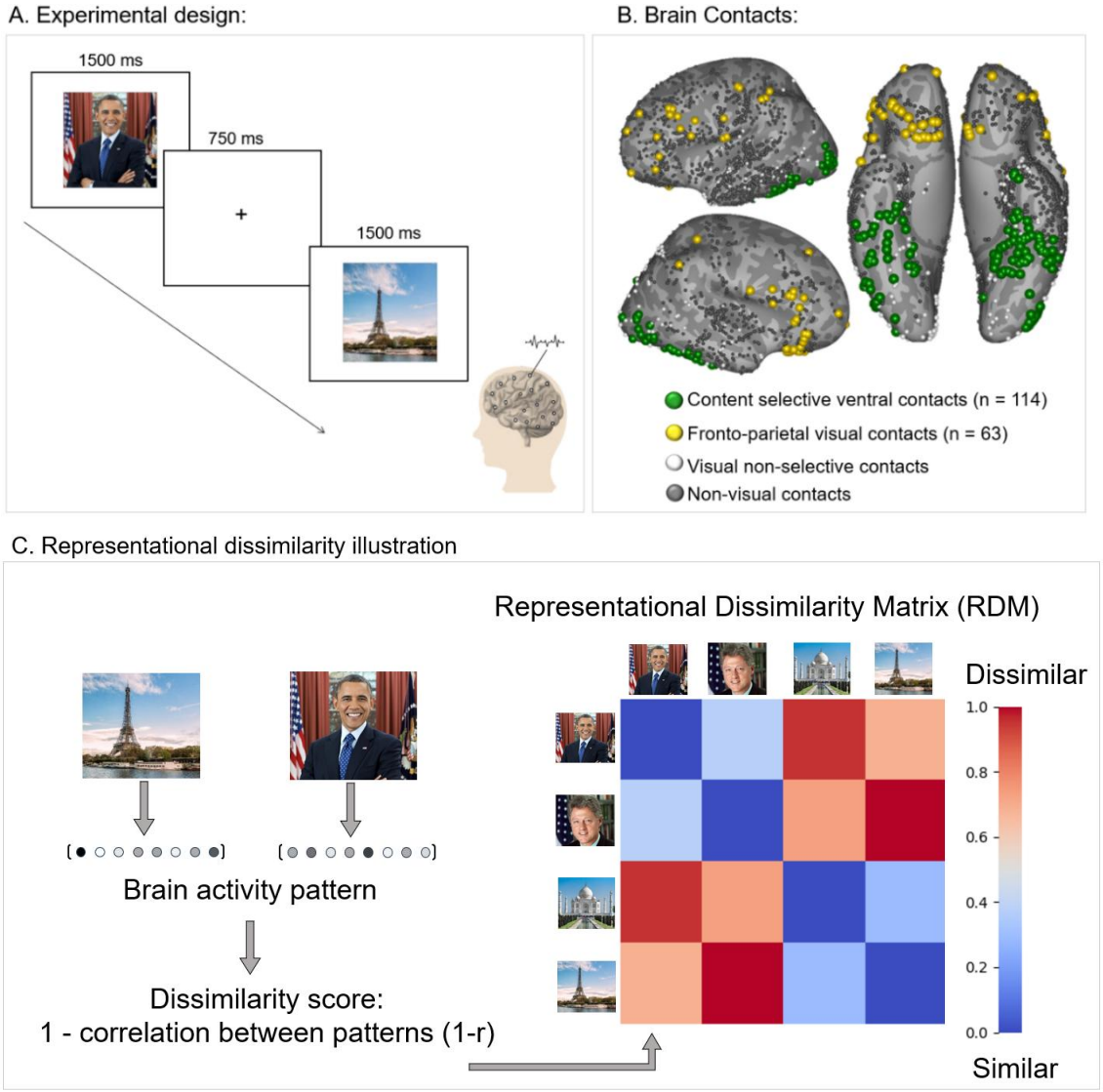
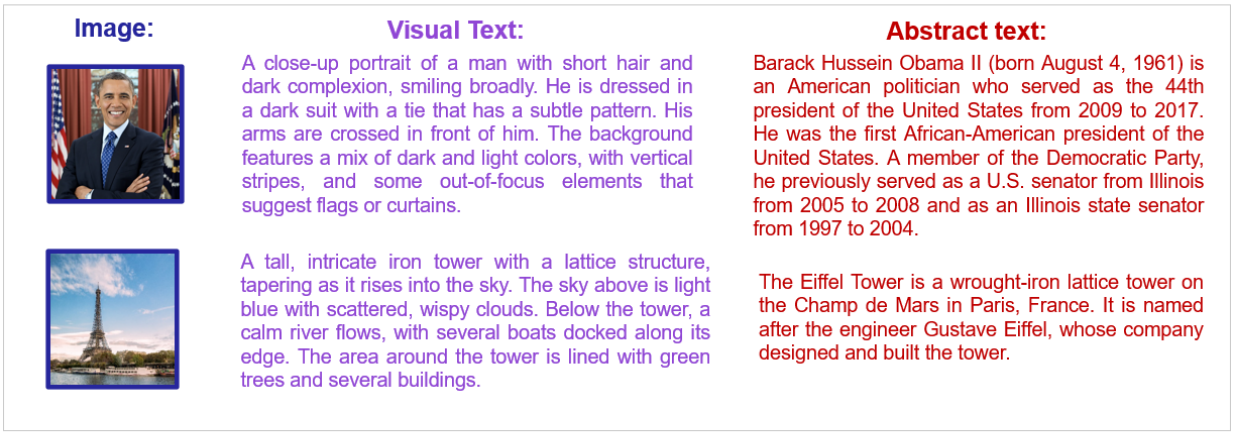


Figure 1: Experimental design: **A.** Participants viewed 28 images of familiar faces and places while their neuronal activity was recorded via intracranial EEG. **B.** Contact locations: yellow - Fronto-parietal visual contacts, green – Content-selective ventro-temporal contacts, white – visual non-selective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

the same category. Both text types were strongly associated with the images: for visual descriptions, the text was correctly matched by an average of 93.6% of participants (mode = 100%, minimum = 70%), and for abstract descriptions, by an average of 95% (mode = 100%, minimum = 80%). 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 DNNs, if the DNN-generated representations of the images are more similar to their DNN-generated

A. Visual and abstract textual descriptions for example images:



B. Correlations and partial correlations between the DNNs representations:

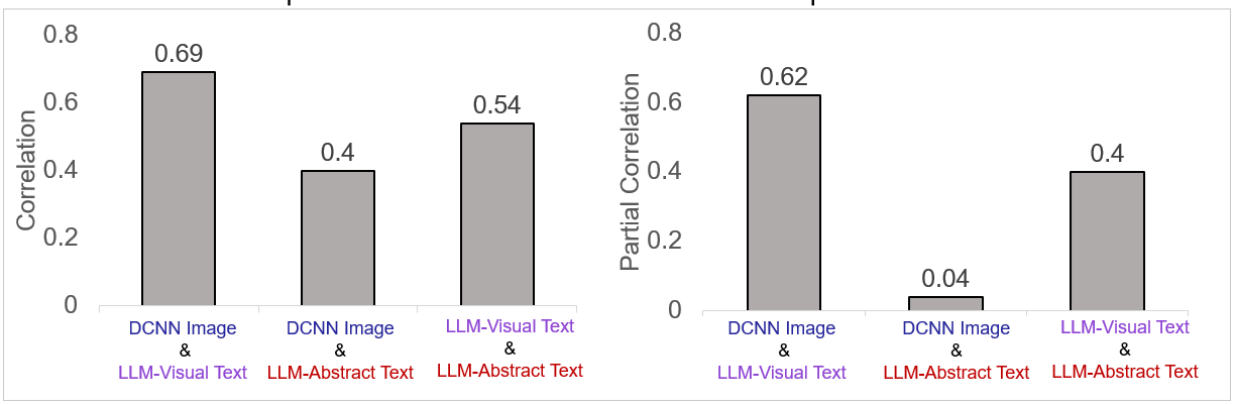


Figure 2: Examples of two images and their visual- and abstract-textual descriptions, and the partial correlations between their visual and language DNNs representations: **A**. Examples of images and their visual (purple) and abstract (red) textual descriptions: **B.** A pre-trained object DCNN (VGG-16) 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 correlations and partial correlations between the RDMs of the DNNs were high between images and visual text but low between images and the abstract text. All images in the figure were replaced by licensed images with a similar appearance from Freepik (https://freepick.com).

visual-text representations compared to the representation generated by the abstract text descriptions. The zero order correlation between the representation of the 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 assess the unique association between each pair of DNNs, independent of their shared variance with the third 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), suggesting 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 ventro-temporal 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 high order ventro-temporal 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 (i.e. RDM matrices) similarities between the response of the visual ventro-temporal 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-derived 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 and the difference between each correlation

184 in each ROI. P-values were FDR corrected for each ROI and each analysis (individual
185 predictors and differences) separately.

186 Our analysis shows that the neural response of the visual ventro-temporal cortex was
187 similarly correlated with the visual representation of the image (DCNN-Image), and the
188 visual-text representation (LLM-visual) but showed a significantly lower correlation with the
189 abstract text representation (LLM-abstract) (Fig. 3 left top panel). The same analysis with the
190 responses of the fronto-parietal contacts to the images showed significant correlations with

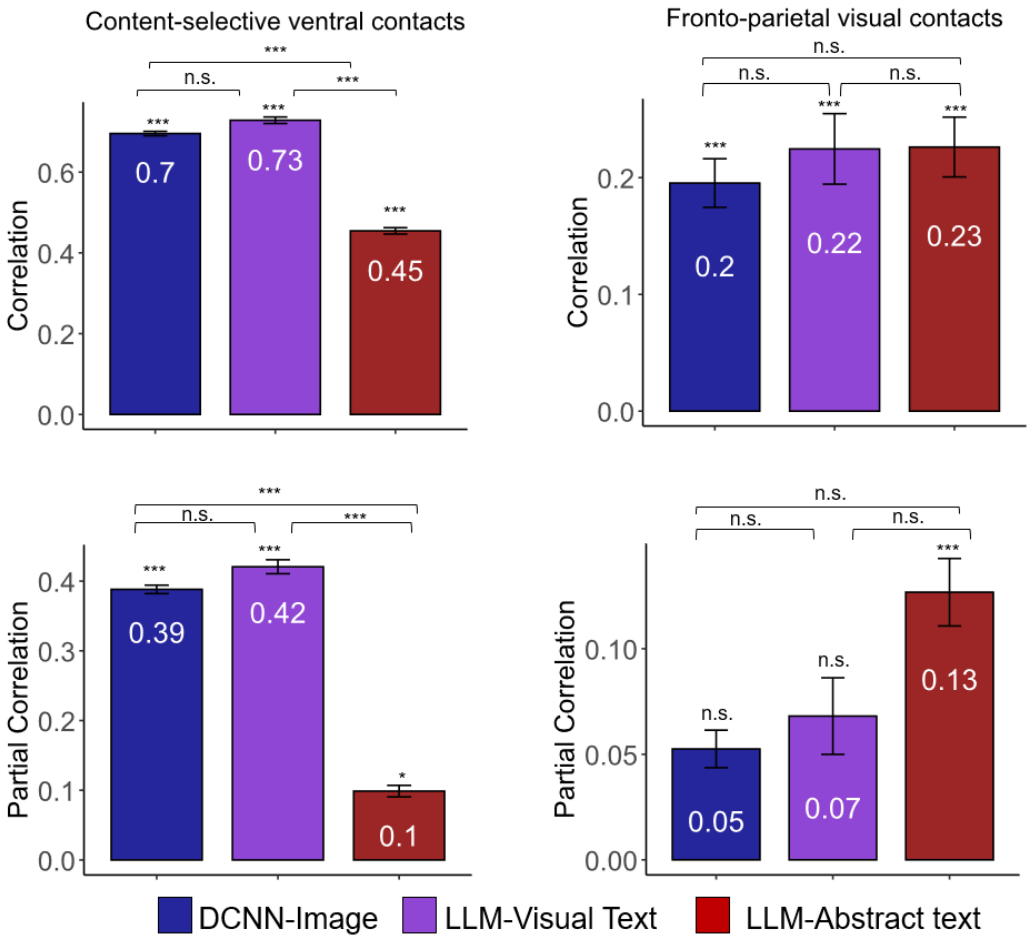


Figure 3: Correlations between the representations of ventro-temporal content-selective and fronto-parietal visual contacts to images and the representations of the images by: visual DNN (DCNN-image), visual textual descriptions (LMM-visual text), and abstract textual descriptions (LMM-abstract text): 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. $* \, p_{FDR} < 0.05$, $** \, p_{FDR} < 0.01$, $*** \, p_{FDR} < 0.001$

all three DNN-based representations, but no significant difference between them. To examine the unique similarity of each DNN representation to the response of the ventro-temporal visual contacts to the images, we calculated the partial correlations between each DNN-based RDM and the RDM derived from the visual cortex, when the two other DNN representations are held constant. The same correlational analysis was performed on the response of the fronto-parietal contacts.

Both the image-based (DCNN-Image) and the visual-text based (LLM-Visual) representations uniquely predicted the ventro-temporal visual cortex response to the images, whereas the abstract text did not account for any additional information. There was no significant difference between the correlations 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 fronto-parietal 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).

Temporal dynamics show that over time the response of the visual ventro-temporal cortex is mainly explained by the image representation

The iEEG recordings enable us to also examine the temporal dynamics of the visual and linguistic representations. We ran the same correlational and partial correlation analysis of the DNNs with the neural responses across time (see Figure 4 panels A and B). Findings show that correlations with the visual DNN representation persists for longer duration until stimulus offset, whereas the DNN representation of the visual text is short lasting (see Figure 4 panel B).

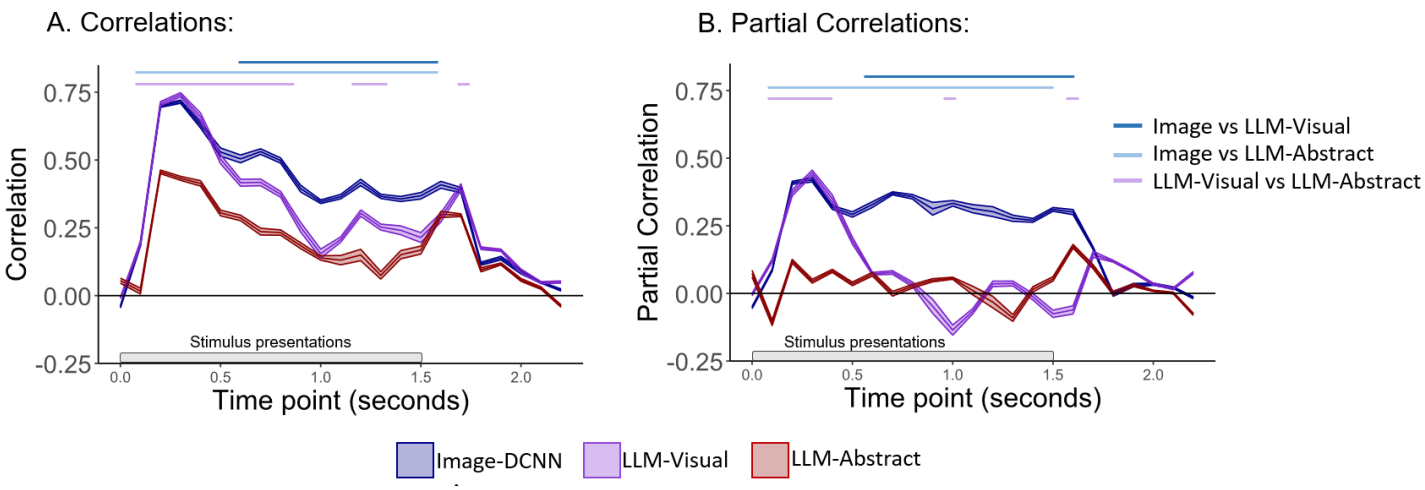
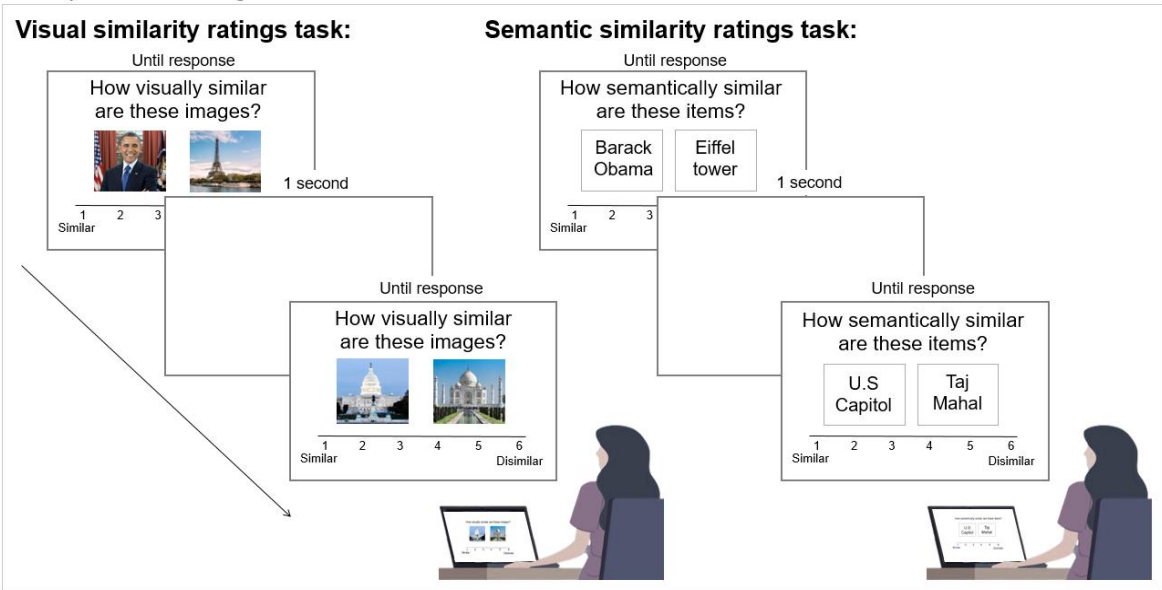


Figure 4: Correlations and partial correlations over time, between the representations of ventral content-selective contacts to images and the representations of the images by: visual DNN (DCNN-image), visual textual descriptions (LMM-visual text), and abstract textual descriptions (LMM-abstract text): **A.** Zero order correlations of each DNN RDM's with the RDM of the ventral content-selective contacts' activity (n=114) **B.** Partial correlations with the RDM of brain activity patterns, when the other two DNNs are held constant. Ribbon indicates leave one participant out procedure s.e.m. The lines above represent significance, all *p* values were derived from a pair-images two-sided permutation test (10000 permutations).

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

To further examine if the visual and abstract textual descriptions represent 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

A. Experimental design:



B. Partial Correlations of the different DNNs with Visual and Semantic human similarity ratings:

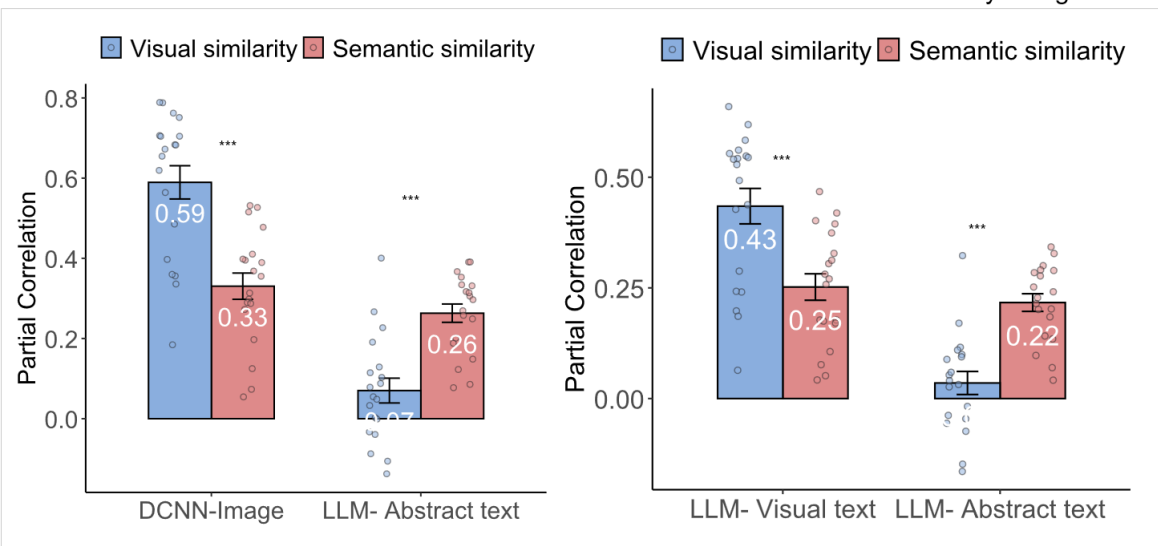


Figure 5: Experiment design and results of human similarity ratings: A. Experimental design: Participants rated the visual similarity (n = 20) of famous faces and places while presented with their images that were used in the iEEG study (left) or the semantic similarity (n = 20) between the names of the same stimuli presented in the images (right) ***B.*** Partial correlations with human similarity ratings: Left panel- When including only the DCNN-Image and the abstract-text representations right panel- When including only the visual-text and the abstract-text representations. Both figures depict significant interaction and similar results. Error bars are s.e.m. between participants. p values were derived from a post hoc compression between task and predictors. Reported p values are FDR corrected.$** \; p_{FDR} < 0.01 \;\; *** \; p_{FDR} < 0.001$. All images in the figure were replaced by licensed images with a similar appearance from Freepik (https://freepick.com).

similarity between each pair of items based on the conceptual knowledge they had about them (see Fig. 5A; see Methods for further details). For each participant we created an RDM of the stimuli, based on the participant's similarity ratings. To examine the unique contribution of each of these representations, we computed their partial correlations. We performed this analysis once when including only the DCNN-Image and the abstract-text representations, and once when including only the visual-text and the abstract-text representations.

As shown in Fig. 5, these two analyses resulted in a very similar pattern of results with the representation of the visual and visual-text DNNs resulting in higher partial correlation with human visual than semantic similarity rating, whereas the representation of the abstract-text showed the reversed pattern. To test the statistical significance of these findings, the partial correlations were transformed to fisher z-scores (the figures show the original partial correlation before fisher-z transformation) and a mixed ANOVA was performed with DNN-representation (image and abstract-text or visual-text and abstract-text, respectively) as within-participant factor, and task (visual similarity rating, semantic similarity rating) as between-participants factor, on the Fisher transformed partial correlations. The ANOVA including DCNN-Image and abstract-text representations revealed a significant main effect for DNN ($F(1, 36) = 75.7$, $p < 0.001$, $\eta_p^2 = 0.68$) and for Task ($F(1,36) = 4.13$, $p = 0.049$, $\eta_p^2 = 0.10$), and a significant interaction between them $F(1, 36) = 46.24$, $p < 0.001$, $\eta_p^2 = 0.56$). Post-hoc comparison between visual and semantic rating tasks for DNN-based representation (FDR corrected) revealed that image representations show higher partial correlation with visual than semantic similarity ratings ($t (72) = 6.29$, $p < 0.001$), but abstract-textual representations show higher partial correlation with the semantic rather than the visual rating tasks ($t (72) = 3.44$, $p < 0.001$). The ANOVA including visual-text and abstract-text representations revealed a significant main effect for DNN ($F(1, 36) = 43.8$, $p < 0.001$, $\eta_p^2 = 0.55$) but not for Task ($F(1,36) = 0.33$, $p = 0.57$, $\eta_p^2 = 0.009$), and a significant interaction between them $F(1, 36) = 30.54$, $p < 0.001$, $\eta_p^2 = 0.46$). Post-hoc comparison between visual and semantic rating tasks for DNN-based representation (FDR corrected) revealed that visual-text representations show higher partial correlation with visual than semantic similarity ratings ($t (72) = 4.64$, $p < 0.001$), but abstract-textual representations show higher partial correlation with the semantic rather than the visual rating tasks ($t (72) = 3.92$, $p < 0.001$).

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.

Using a different set of stimuli revealed a similar pattern of correlations between the DNNs and the human similarity ratings

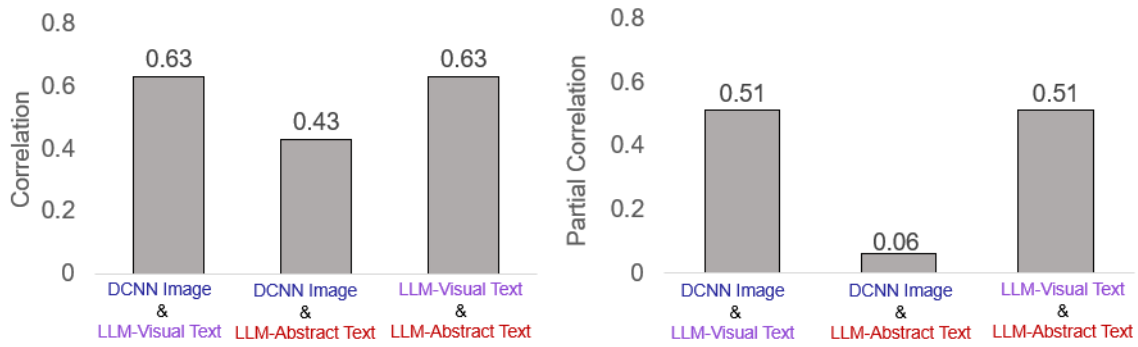
As mentioned above, visual and abstract texts can be clearly dissociated for familiar images, as our acquaintance with them involves acquiring non-visual verbal knowledge. However, to extend our findings beyond familiar images, we also ran the same visual and semantic similarity task with unfamiliar scenes. This also allowed us to overcome a shortcoming of the current study in which, due to the constraint imposed by the unique clinical conditions of the IEEG recordings, the number and types of images that we used were limited. It was therefore important to validate the relationship between visual text and abstract text to visual and semantic representations with an additional set of images. To that end, we used 15 images of different indoor and outdoor scenes taken from Bainbridge et al[31]. We extracted visual and abstract descriptions of the images (see Figure 6A for an example of a kitchen and a playground) as we did for the first set of images. Then we used a DCNN to extract the images representations and an LLM to extract the representation of both the visual and the abstract textual descriptions (see methods for further details). Figure 6B shows that the correlation and partial correlation of the images are much higher with visual text than with abstract text representations. We asked participants to make visual or semantic similarity ratings of all possible pairs of the 15 stimuli belonging to the new set. As with the original set of images, 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 scene presented in the image, for example “Kitchen”) and was asked to rate the semantic similarity between each pair of items based on the conceptual knowledge they had about them (see Methods for further details). Then we calculated the partial correlation between the participants’ visual and semantic similarity ratings with the DNNs representations, between DNN pairs, holding one DNN constant in each calculation. The results, as shown in Figure 6C, were very similar to the first set of stimuli (Figure 5B). The partial correlation used for the analysis were transformed to fisher z-scores (the figures show the original partial correlation before fisher-z transformation).

The ANOVA including DCNN-Image and abstract-text representations revealed a significant main effect for DNN ($F(1, 38) = 4.85$, $p = 0.034$, $\eta_p^2 = 0.11$) and for Task ($F(1,38) = 4.29$, $p = 0.045$, $\eta_p^2 = 0.10$), and a significant interaction between them $F(1, 38) = 52.54$, $p < 0.001$, $\eta_p^2 = 0.58$). Post-hoc comparison between the visual and semantic rating tasks for DNN-based representation (FDR corrected) revealed that image representations show higher partial correlation with visual than semantic similarity ratings ($t (76) = 4.74$, $p < 0.001$), but abstract-textual representations higher partial correlation with the semantic rather than the visual rating tasks ($t (76) = 7.09$, $p < 0.001$). The ANOVA including visual-text and abstract-text representations revealed a significant main effect for DNN ($F(1, 38) = 8.5$, $p = 0.006$, $\eta_p^2 =$

A. Visual and abstract textual descriptions for example images:

Image:	Visual Text:	Abstract text:
	A spacious space with sage green cabinetry featuring intricate molding and glass-paneled doors. The backsplash consists of white herringbone tiles, and the countertops are light wood. A large wooden island with matching stools is in the center, topped with a glass vase filled with eucalyptus leaves. A white wreath hangs above the stove, which has a copper kettle and cooking essentials nearby.	a room or part of a room used for cooking and food preparation in a dwelling or in a commercial establishment. A modern middle-class residential is typically equipped with a stove, a sink with hot and cold running water, a refrigerator, and worktops and cabinets arranged according to a modular design
	A vibrant outdoor area with a multi-level structure that includes slides, ladders, and climbing frames. A roundabout with green seats and a nearby bench are also visible. The space is enclosed by a green fence and surrounded by lush trees.	a place designed to provide an environment for children that facilitates play, typically outdoors

B. Correlations and partial correlations between the DNNs representations:



C. Partial correlations between the DNNs representations and human similarity ratings

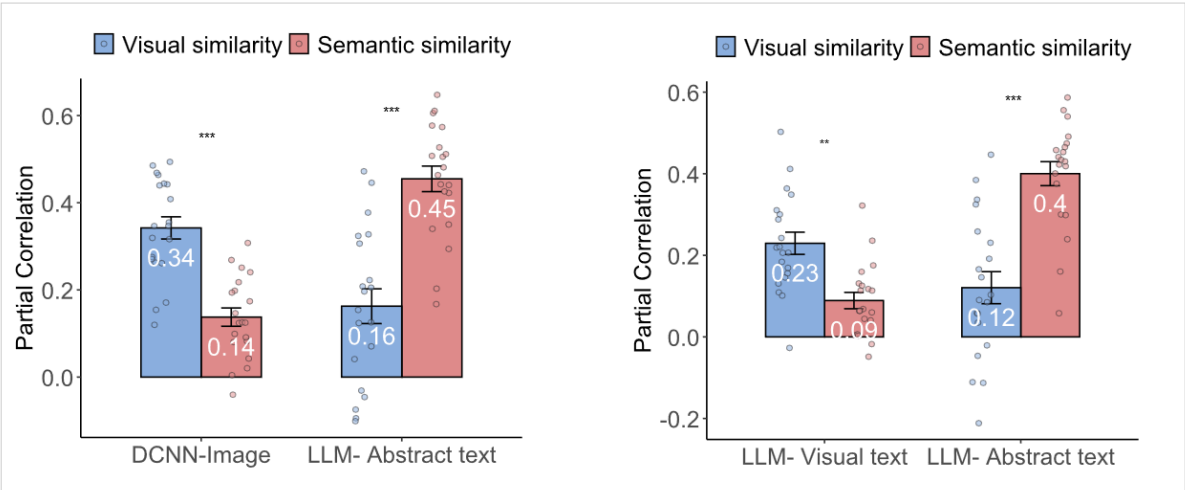


Figure 6: The results using a different set of images: A. Examples of two scene images, a kitchen and a playground, and their visual (purple) and abstract (red) textual descriptions: **B.** A pre-trained object DCNN (VGG-16) was used to extract the embeddings of the images (DCNN-image), and a large language model (GPT-2) was used to extract the embeddings of the visual and abstract textual descriptions. The correlations and partial correlations between the RDMs of the DNNs were high between images and visual text but low between images and the abstract text. **C.** Partial correlations with human visual and semantic similarity ratings when including only the DCNN-Image and the abstract-text representations (left) or when including only the visual-text and the abstract-text representations (right). Both figures in panel c depict significant interaction and similar results. Error bars are s.e.m. between participants. *p* values were derived from a post hoc compression between task and predictors. Reported *p* values are FDR corrected.$** \; p_{FDR} < 0.01 \; *** \; p_{FDR} < 0.001$. All images in the figure were replaced by licensed images with a similar appearance from Freepik (https://freepick.com).

between them $F(1, 38) = 33.34$, $p < 0.001$, $\eta_p^2 = 0.46$). Post-hoc comparison between visual and semantic rating tasks for DNN-based representation (FDR corrected) revealed that visual-text representations show a higher partial correlation with visual than semantic similarity ratings ($t(76) = 3.24$, $p = 0.001$), but abstract-textual representations display a higher partial correlation with the semantic rather than the visual rating tasks ($t(76) = 6.75$, $p < 0.001$). These results further support our conclusion that visual-text representations are aligned with visual rather than semantic information, whereas abstract text representations are correlated with semantic rather than visual information.

Discussion

Many neuroimaging and IEEG studies have convincingly shown that the representations generated for images by DCNNs trained for recognizing images are strongly correlated with neural responses in the high-level visual cortex[4,14–16,32]. Interestingly, recent fMRI studies reveal that linguistic representations of textual descriptions generated by large language models also achieve similarly high predictions of the neural response to images in high level visual cortex[19–21]. These findings suggest a strong alignment between visual and linguistic representations. However, the question of the nature of this linguistic information was largely left unaddressed. Does it reflect purely visual and pictorial aspects of the images or also more abstract, conceptual knowledge, that cannot be derived directly from the visual images themselves? This is an important question, as it determines the interpretation of this intriguing finding, regarding the information guiding the organization of high-level visual cortex. To address this question, we generated two types of linguistic representations to the same images, one that describes information that can be extracted directly from the images 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 employed direct iEEG recordings of neural responses to images of famous faces and places, obtained during clinical diagnostic procedure in epilepsy patients (see methods). In this type 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 DNN representations of images and of visual-text similarly predicted the neural response of high-level visual cortex to images, consistent with recent reports[19–21], whereas abstract textual representations showed significantly lower

correlations and did not explain additional variance beyond the visual information. 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 pictorial, 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 DNN representations encode the subjects' visual or semantic perceived information, we asked human participants to rate the visual or semantic similarity of the same images. Our findings show that visual-text was more correlated with visual similarity ratings than with semantic similarity ratings (Figures 5 and 6). This pattern of correlation corresponds to the pattern of correlation in high-level visual cortex (Figure 1B), and further emphasizes the visual nature of this representation. This same pattern of findings was found for another set of images (Figure 6), supporting the notion that results that were found for the images used in the iEEG study are not specific to this image-set.

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 results were even replicated in a monkey fMRI study that used the same stimuli, further indicating that these representations do not reflect the operation of a language system[18]. 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[33] aspect ratio[34] or curvature[35] 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)[36] 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 ventro-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 defining major cortical subdivisions such as revealed in fMRI studies. For example, in an fMRI study, Carlson et al (2014)[22] 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 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 [16,26]). For such detailed analysis, higher spatial resolution methods such as iEEG are especially informative[37].

The partial correlation analysis further shows that visual and visual-textual information contributed distinct proportion of variance in the neural response to images. Analysis of the temporal dynamics of the representation also indicates that the visual representation shows higher correlation for longer duration till stimulus offset, whereas the visual linguistic representation is short lived. This suggests that in addition to their aligned representations the neural responses 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 reflect overlapped or distinct types of information, which has been highly debated. Embodied cognition theories posit that all information is grounded in sensory information[38,39], whereas dual-code models suggest a distinction between visual and language inputs[40]. 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[40].

The 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[41,42]. These studies reveal substantial cognitive and neural differences between the representations of concrete/grounded words, such as names of objects as opposed to 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 [43–45] 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 supports the notion 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.

Materials and Methods

<u>iEEG study Familiar faces and places:</u>

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. During the iEEG recordings the participants did not indicate whether they were familiar with the identities and places before the experiment. Therefore, to ensure that the images used are indeed generally familiar items, in a different study we asked 40 participants to indicate whether they are familiar with each face or place. Across all participants, the images were recognized by an average of 90% (mode = 97%), confirming that these faces and places were highly familiar stimuli.

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].

<u>Participants</u>

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 stimuli-evoked response to 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 parameters, to content-selective ventro-temporal and fronto-parietal visual contacts. To define the content selective contacts, we examined the visually responsive contacts in six anatomical regions across the ventro-temporal 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 (averaged 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[46]: 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[47], 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[48–51]. 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 of the last word in each description. 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 studies:

<u>Human similarity ratings – faces and places:</u>

<u>Stimuli:</u>

The same stimuli that were used for the iEEG recording.

<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)

<u>Procedure</u>

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).

Text image matching task – faces and places:

Stimuli:

The same stimuli that were used for the iEEG recording.

Participants

A total of 40 participants were recruited for this study from the Prolific platform. Twenty participants were assigned to each of the two conditions: visual text matching (mean age 33 (sd = 3.4), 5 females) or abstract text matching (mean age 32 (sd = 3.4), 10 females). The participants were paid 2.25 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

Procedure

Participants matched each visual or abstract description of all 28 images to one image out of 14 images (the images from the same category, faces or places). They were presented with one text at a time and were asked to choose the image that best matches each text. The images were presented simultaneously, and the participants selected the number of images that best matched the text. After making their choice they were presented with the next text trial.

At the end of the study, they were asked to indicate whether they were familiar with the faces/places before the study. The experiment lasted about 15 minutes.

We did not exclude trial based on any criteria.

Human similarity ratings – scenes:

15 color images of indoor and outdoor scenes (see Supplementary Materials for all scene images and their textual descriptions).

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) - Similarly to set 1, the visual descriptions were generated by Chat-GPT[46]. In case the text included the name of the scene, or the words “the image shows” we removed it from the text. For visual descriptions of the images see Fig 6A (for an example) and Supplementary material (for all visual descriptions).
3. Abstract descriptions (text) – the first paragraph of the Wikipedia descriptions of the scene depicted in each image (see Fig 6A 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 VGG-16 DCNN[29] pre-trained on ImageNet[47] (similar to the familiar face and place stimuli).

The representations of the text input were extracted using a large language model – SGPT[52]: GPT Sentence Embeddings for Semantic Search is a recent natural language processing (NLP) algorithm that is first pre-trained to predict the next word in a sentence similar to other NLP algorithms and use contrastive fine-tuning to create similar representations for pairs of sentences that describe the same content[52]. This model was chosen because the correlations of its representations with semantic similarity ratings (r = xxx) of the stimuli was higher than the representations generated by GPT-2 (r = xxx). We extracted the embeddings of the Wikipedia definition based on the 1.3B parameters bi-encoder's output layer and computed the similarity between each scenes pair based on the cosine distance between these representations.

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 30 (sd = 2.9), 18 females). 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 15 items (105 pairs). The timing and procedure of both tasks was identical to the behavioral study described for the familiar faces and places described above except that the instructions were presented at the beginning and with no reminder on each trial.

Trials' exclusion:

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, 0.4% of the trials were excluded. In the semantic similarity rating task, 0.4% of the trials were excluded.

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. We performed the same analyses for the scene images (Figure 6). (15 Images, 105 pairs).

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] (for the time analysis reported in Figure 4, we did not average across time and used each time point separately, apart from that, we calculated everything the same). 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 ventro-temporal and fronto-parietal 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[53] 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[53] 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. For the time analysis reported in Figure 4, we performed the permutation test for each time point separately, and then FDR corrected for all time points together.

<u>Correlation between the DNNs representations and behavioral similarity ratings:</u>

We created a representational dissimilarity matrix (RDM) for each participant based on their similarity ratings. We computed the partial 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 were performed on the Fisher's z-transformed partial correlations using two-sided null-hypothesis tests. FDR was applied to correct for multiple comparisons for all post-hoc tests in each analysis separately.

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. 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.
26. 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).
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. Bainbridge, W. A., Hall, E. H. & Baker, C. I. Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nat. Commun.* **10**, 5 (2019).
32. Kanwisher, N., Khosla, M. & Dobs, K. Using artificial neural networks to ask 'why'questions of minds and brains. *Trends Neurosci.* **46**, 240–254 (2023).
33. 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).

34. Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
35. Tsao, D. Y., Cadieu, C. & Livingstone, M. S. Object recognition: physiological and computational insights. *Primate Neuroethol*. 471 (2008).
36. Bracci, S. & Beeck, H. Dissociations and associations between shape and category representations in the two visual pathways. *J. Neurosci.* **36**, 432–444 (2016).
37. Privman, E. *et al.* Enhanced category tuning revealed by intracranial electroencephalograms in high-order human visual areas. *J. Neurosci*. **27**, 6234–6242 (2007).
38. Foglia, L. & Wilson, R. A. Embodied cognition. *Wiley Interdiscip. Rev. Cogn. Sci.* **4**, 319–325 (2013).
39. Leitan, N. & Chaffey, L. Embodied cognition and its applications: A brief review. (2014).
40. 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).
41. Paivio, A. Dual coding theory: Retrospect and current status. *Can. J. Psychol. Can. Psychol.* **45**, 255 (1991).
42. Sadoski, M. & Paivio, A. *Imagery and Text: A Dual Coding Theory of Reading and Writing*. (Routledge, 2013).
43. Barsalou, L. W. Grounded Cognition. *Annu. Rev. Psychol.* **59**, 617–645 (2008).
44. Leshinskaya, A. & Caramazza, A. For a cognitive neuroscience of concepts : Moving beyond the grounding issue. *Psychon Bull Rev* **991–1001**, (2016).
45. 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).
46. OpenAI. ChatGPT. (2023).

47. 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.

48. 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).

49. 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).

50. 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).

51. 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).

52. Muennighoff, N. SGPT: GPT Sentence Embeddings for Semantic Search. (2022).

53. 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).

Acknowledgments

The study was funded by a CIFAR BMC fellowship to R. Malach, ISF grant no.917/21 to G. Yovel and the Sieratzki Institute for Advances in Neuroscience scholarship to A. Shoham.

Supplementary data: Table 1: Familiar faces and places images:

Image	Visual Text	Abstract Text
Image link	A man with short hair and a dark complexion, wearing a dark suit, white shirt, and black tie. His expression is one of bemusement, with his lips pursed and one eyebrow slightly raised. He is standing in front of a blurred background with some abstract shapes and colors.	Barack Hussein Obama II (born August 4, 1961) is an American politician who served as the 44th president of the United States from 2009 to 2017. He was the first African-American president of the United States. A member of the Democratic Party, he previously served as a U.S. senator from Illinois from 2005 to 2008 and as an Illinois state senator from 1997 to 2004.
Image link	A man with gray hair saluting with his right hand. He is wearing a dark suit with a white shirt and red striped tie. The background features anAmerican flag. The man has a slight smile, and his expression is friendly.	William Jefferson Clinton (born August 19, 1946) is an American politician who served as the 42nd president of the United States from 1993 to 2001. He previously served as governor of Arkansas from 1979 to 1981 and again from 1983 to 1992, and as attorney general of Arkansas from 1977 to 1979. A member of the Democratic Party, Clinton became known as a New Democrat, as many of his policies reflected a centrist Third Way political philosophy. He is the husband of Hillary Clinton, who was a senator from New York from 2001 to 2009, secretary of state from 2009 to 2013 and the Democratic nominee for president in the 2016 presidential election.

Image link	A muscular man with short hair and camouflage face paint. He is holding a large, double-barreled weapon over his shoulder. The background is out of focus, emphasizing his intense expression and the weapon.	Arnold Alois Schwarzenegge(born July 30, 1947) is an Austrian-American actor, film producer, businessman, former bodybuilder and politician who served as the 38th governor of California between 2003 and 2011. As of 2022, he is the most recent Republican governor of California. Time magazine named Schwarzenegger one of the 100 most influential people in the world in 2004 and 2007
Image link	A young woman with blonde hair styled in pigtails. She is wearing a white blouse and a gray cardigan. The background is blurred, focusing attention on her face. She has a confident expression, with slightly parted lips and direct eye contact.	Britney Jean Spears (born December 2, 1981) is an American singer, songwriter, dancer, and actress. Often referred to as the Princess of Pop, she is credited with influencing the revival of teen pop during the late 1990s and early 2000s. Regarded as a pop icon, Spears has sold nearly 150 million records worldwide, including over 70 million in the United States alone, making her one of the world's best-selling music artists.
Image link	A man with a shaved head and intense expression. He has some scratches and bruises on his face. The background is a mix of warm tones, suggesting an action scene. He is wearing a rugged outfit, enhancing his tough appearance.	Walter Bruce Willis (born March 19, 1955) is a retired American actor. His career began on the off-Broadway stage in the 1970s. He achieved fame with a leading role on the comedy-drama series Moonlighting (1985 1989) and appeared in over a hundred films, gaining widespread recognition as an action hero after his portrayal of John McClane in the Die Hard franchise

		(1988 2013) and other roles.
Image link	A man with short dark hair holding a microphone. He is wearing a brown blazer and a blue shirt with a floral tie. The background is dark with a red curtain visible on the left. The man appears to be speaking or performing.	Jerome Allen Seinfeld (born April 29, 1954) is an American stand-up comedian, actor, writer, and producer. He is best known for playing a semi-fictionalized version of himself in the sitcom Seinfeld, which he created and wrote with Larry David. The show aired on NBC from 1989 until 1998, becoming one of the most acclaimed and popular American sitcoms of all time. As a stand-up comedian, Seinfeld specializes in observational comedy. In 2004, Comedy Central named him as the 12th-greatest stand-up comedian of all time
Image link	A man with slicked-back blonde hair holding a glass of champagne. He is wearing a black tuxedo with a bow tie. The background is dark with sparkling lights. The man is smiling and looking directly at the camera.	Leonardo Wilhelm DiCaprio (born November 11, 1974) is an American actor and film producer. Known for his work in biopics and period films, he is the recipient of numerous accolades, including an Academy Award, a British Academy Film Award, and three Golden Globe Awards. As of 2019, his films have grossed over $7.2 billion worldwide, and he has been placed eight times in annual rankings of the world's highest-paid actors.

Image link	A woman with curly black hair smiling widely. She is wearing a white top with exposed shoulders. The background is solid red. The woman has turquoise earrings and is facing the camera.	Oprah Gail Winfrey (January 29, 1954) is an American talk show host, television producer, actress, author, and philanthropist. She is best known for her talk show, The Oprah Winfrey Show, broadcast from Chicago, which was the highest-rated television program of its kind in history and ran in national syndication for 25 years, from 1986 to 2011.Dubbed the Queen of All Media, she was the richest African-American of the 20th century was once the world's only black billionaire, and the greatest black philanthropist in U.S. history.By 2007, she was sometimes ranked as the most influential woman in the world.
Image link	A bald man with a mischievous expression on his face. He is wearing a gray suit with a high collar. His left hand is raised with the pinky finger touching his lips. The background is plain and gray, focusing attention on his face.	Douglas Powers commonly known as Dr. Evil, is a fictional character portrayed film series. He is the main antagonist and Austin Powers and secret twin brother). He is James Bond villains, primarily Ernst Stavro Blofeld. Dr. Evil routinely hatches schemes to terrorize and take over the world, and is usually accompanied by Number Two, a goon who fronts his evil corporation Virtucon Industries, his cat Mr. Bigglesworth and his sidekick Mini-Me, a dwarf clone of himself.

Image link	A woman with long blonde hair playing an acoustic guitar. She is sitting in front of a microphone. The woman is wearing a floral-patterned blouse. The background features a sign with "Central Perk" written on it.	Phoebe Buffay is one of the six main characters from the American sitcom Friends. She is created by David Crane and Marta Kauffman and portrayed by actress Lisa Kudrow. In the series' universe, Phoebe was born on February 16 and is the daughter of Frank and Lily Buffay. Her biological mother's name was Phoebe Abbott, whom she was named after. Phoebe has a twin sister, Ursula, a waitress who is also portrayed by Kudrow. Phoebe can speak several languages, including French and Italian. She appeared in all of the show's 236 episodes.
Image link	A woman with short black hair and bangs lying on a bed. She is wearing a black top and has red lipstick. The woman is holding a cigarette in one hand. The background is slightly blurred, focusing attention on her face and expression.	Pulp Fiction is a 1994 American black comedy crime film written and directed by Quentin Tarantino, who conceived it with Roger Avary. Starring John Travolta, Samuel L. Jackson, Bruce Willis, Tim Roth, Ving Rhames, and Uma Thurman, it tells several stories of crime in Los Angeles. The title refers to the pulp magazines and hardboiled crime novels popular during the mid-20th century, known for their graphic violence and punchy dialogue.

Image link	An animated character with tan skin and black hair, wearing a red fez and a purple vest. He has a broad, friendly smile and is looking slightly upwards. The background is a gradient of soft pastel colors.	Aladdin is a fictional character and the titular protagonist of Walt Disney Pictures' 31st animated feature of the same name based on Aladdin, a folk tale of Middle Eastern origin. He is voiced by American actor Scott Weinger, while his singing voice is provided by Brad Kane. He also stars in the two direct-to-video sequels The Return of Jafar (1994) and Aladdin and the King of Thieves (1996), as well as the animated television series based on the film. Mena Massoud played a live-action version of the character in a live action adaptation of the 1992 film.
Image link	An animated character with yellow skin and spiky hair. He is riding a skateboard, wearing a red shirt and blue shorts. The background is an orange gradient with flames, enhancing the action scene. He has a mischievous expression, with wide eyes and a grin.	Bart Simpson is a fictional character in the American animated television series The Simpsons and part of the Simpson family. He is voiced by Nancy Cartwright and first appeared on television in The Tracey Ullman Show short Good Night on April 19, 1987. Cartoonist Matt Groening created and designed Bart while waiting in the lobby of James L. Brooks' office. Groening had been called to pitch a series of shorts based on his comic strip, Life in Hell, but instead decided to create a new set of characters. While the rest of the characters were named after Groening's family members, Bart's name is an anagram of the word brat. After appearing on The Tracey Ullman Show for two years, the Simpson family received its

		own series on Fox, which debuted December 17, 1989. Bart has appeared in every Simpsons episode except Four Great Women and a Manicure.
Image link	An animated duck character with a white body and yellow beak. He is wearing a blue sailor suit and a blue hat with a black ribbon. The background is plain white, focusing attention on the character. He has a cheerful expression with a wide open mouth and raised eyebrows.	Donald Fauntleroy Duck is a cartoon character created by The Walt Disney Company. Donald is an anthropomorphic white duck with a yellow-orange bill, legs, and feet. He typically wears a sailor shirt and cap with a bow tie. Donald is known for his semi-intelligible speech and his mischievous, temperamental, and pompous personality. Along with his friend Mickey Mouse, Donald was included in TV Guide's list of the 50 greatest cartoon characters of all time in 2002 and has earned a star on the Hollywood Walk of Fame. He has appeared in more films than any other Disney character, and is the most published comic book character in the world outside of the superhero genre.

Image link	A tall clock tower with a large clock face. The tower is part of a larger building with gothic architectural details.A bridge spans across a river in the foreground. The sky is clear blue, highlighting the intricate details of the structure.	Big Ben is the nickname for the Great Bell of the striking clock at the north end of the Palace of Westminster in London, England, and the name is frequently extended to refer also to the clock and the clock tower. The official name of the tower in which Big Ben is located was originally the Clock Tower, but it was renamed Elizabeth Tower in 2012 to mark the Diamond Jubilee of Elizabeth II.
Image link	A large, reflective bean-shaped sculpture. The sculpture is set in an urban plaza with people walking around it. Tall buildings surround the plaza, providing a cityscape background. The reflection on the sculpture distorts the surrounding scene.	Cloud Gate is a public sculpture by Indian-born British artist Anish Kapoor, that is the centerpiece of AT&T Plaza at Millennium Park in the Loop community area of Chicago, Illinois. The sculpture and AT&T Plaza are located on top of Park Grill, between the Chase Promenade and McCormick Tribune Plaza & Ice Rink. Constructed between 2004 and 2006, the sculpture is nicknamed the Bean because of its shape, a name Kapoor initially disliked, but later grew fond of. Made up of 168 stainless steel plates welded together, its highly polished exterior has no visible seams. It measures 33 by 66 by 42 feet (10 by 20 by 13 m), and weighs 110 short tons (100 t; 98 long tons).

Image link	A tall, iron lattice tower standing in the center of the image. The sky is a gradient of blue and orange, indicating either sunrise or sunset. The tower is surrounded by illuminated pathways and fountains. There is a reflection of the tower in the water below.	The Eiffel Tower is a wrought-iron lattice tower on the Champ de Mars in Paris, France. It is named after the engineer Gustave Eiffel, whose company designed and built the tower. Locally nicknamed La dame de fer (French for Iron Lady), it was constructed from 1887 to 1889 as the centerpiece of the 1889 World's Fair and was initially criticised by some of France's leading artists and intellectuals for its design, but it has become a global cultural icon of France and one of the most recognisable structures in the world.The Eiffel Tower is the most visited monument with an entrance fee in the world; 6.91 million people ascended it in 2015. The Tower was made a Monument historique in 1964 and named part of UNESCO World Heritage Site (Paris, Banks of the Seins) in 1991.
Image link	A massive, triangular pyramid made of large stone blocks. The structure is set against a clear, blue sky. The base of the pyramid is surrounded by a sandy terrain. Small figures of people are visible near the base, providing scale.	The Great Pyramid of Giza is the larges Egyptian pyramid and tomb of Khufu. Built in the 26th century BC during a period of around 27 years, it is the oldest of the Seven Wonders of the Ancient World, and the only one to remain largely intact. As part of the Giza pyramid complex, it borders present-day Giza in Greater Cairo, Egypt.

Image link	A sidewalk with embedded stars in a pattern. The stars are pink with a gold circle in the middle. The sidewalk is black and speckled.A road runs parallel to the sidewalk with parked cars and a bus. There are trees and buildings lining the street.	The Hollywood Walk of Fame is a historic landmark which consists of more than 2,700. five-pointed terrazzo and brass stars embedded in the idewalks along 15 blocks of Hollywood Boulevard and three blocks of Vine Street¼ hollywood California. The stars are permanent public monuments to achievement in the entertainment industry, bearing the names of a mix of musicians, actors, directors, producers, musical and theatrical groups, fictional characters, and others. The Walk of Fame is administered by the Hollywood Chamber of Commerce and maintained by the self-financing Hollywood Historic Trust. It is a popular tourist destination, with an estimated 10 million annual visitors in 2010. The Hollywood Chamber of Commerce holds trademark rights to the Hollywood Walk of Fame.
Image link	A large concrete dam situated between rugged cliffs.A wide river or reservoir of blue water behind the dam. The dam has a curved structure. There are roads and bridges on both sides of the dam. Mountains are visible in the background.	Hoover Dam is a concret arch-gravity dam in the Black Canyon of the Colorado River, on the border between the U.S. states of Nevada and Arizona. It was constructed between 1931 and 1936 during the Great Depression and was dedicated on September 30, 1935, by President Franklin D. Roosevelt. Its construction was the result of a massive effort involving thousands of

		workers, and cost over one hundred lives. It was referred to as Hoover Dam after President Herbert Hoover in bills passed by Congress during its construction, but was named Boulder Dam by the Roosevelt administration. The Hoover Dam name was restored by Congress in 1947.
Image link	A tall, cylindrical tower leaning significantly to one side. The tower features multiple levels with small arches and columns. Next to the tower is a rectangular building with a flat roof and arched windows. Both structures are made of light-colored stone.A grassy area is in the foreground, with people walking or standing near the base of the tower.	The Leaning Tower of Pisa (Italian: torre pendente di Pisa), or simply, the Tower of Pisa, is the campanile, or freestanding bell tower, of Pisa Cathedral. It is known for its nearly four-degree lean, the result of an unstable foundation. The tower is one of three structures in the Pisa's Cathedral Square (Piazza del Duomo), which includes the cathedral and Pisa Baptistry.

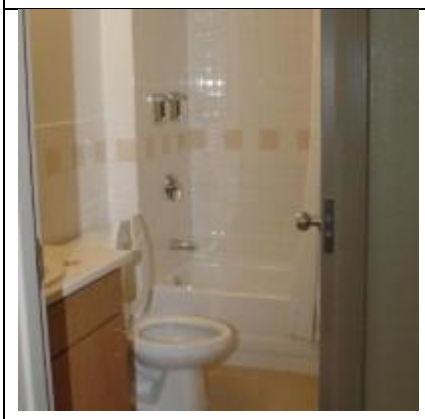
Image link	A large, jutting rock formation extending over a vast landscape. A small lion cub stands at the edge of the rock, held by a baboon. The background shows a wide, open savannah with scattered trees. Sunlight streams through clouds, creating dramatic rays of light. The scene is animated and vividly colored, capturing a sense of grandeur.	The Lion King is a 1994 American animated musical drama film produced by Walt Disney Feature Animation and released by Walt Disney Pictures. It is the 32nd Disney animated feature film and the fifth animated film produced during a period known as the Disney Renaissance. The Lion King was directed by Roger Allers and Rob Minkoff (in their feature directorial debuts) and produced by Don Hahn, from a screenplay by Irene Mecchi, Jonathan Roberts, and Linda Woolverton. Its original songs were written by composer Elton John and lyricist Tim Rice, with a score by Hans Zimmer. The film features an ensemble voice cast that includes Matthew Broderick, James Earl Jones, Jeremy Irons, Jonathan Taylor Thomas, Moira Kelly, Nathan Lane, Ernie Sabella, Rowan Atkinson, Robert Guillaume, Madge Sinclair (in her last film role), Whoopi Goldberg, Cheech Marin, and Jim Cummings.

Image link	A large red suspension bridge spanning over water. The bridge has two tall towers with cables connecting them. The background shows a cityscape with lights. The sky is a deep blue, suggesting dusk or dawn.	The Golden Gate Bridge is a suspension bridge spanning the Golden Gate, the one-mile-wide (1.6 km) strait connecting San Francisco Bay and the Pacific Ocean. The structure links the U.S. city of San Francisco, California the northern tip of the San Francisco Peninsula to Marin County, carrying both U.S. Route 101 and California State Route 1 across the strait. It also carries pedestrian and bicycle traffic, and is designated as part of U.S. Bicycle Route 95. Being declared one of the Wonders of the Modern World by the American Society of Civil Engineers, the bridge is one of the most internationally recognized symbols of San Francisco and California. It was initially designed by engineer Joseph Strauss in 1917.
Image link	A grand white marble mausoleum with a large central dome. Four tall, slender minarets stand at each corner of the platform. The building features intricate carvings and symmetrical arches. A long, reflective pool leads up to the entrance, flanked by gardens. The sky is clear, providing a stark contrast to the white marble structure.	The Taj Mahal is an ivory-white marble mausoleum on the right bank of the river Yamuna in the Indian city of Agra. It was commissioned in 1632 by the Mughal emperor Shah Jahan (r. 1628 1658) to house the tomb of his favourite wife, Mumtaz Mahal; it also houses the tomb of Shah Jahan himself. The tomb is the centrepiece of a 17-hectare (42-acre) complex, which includes a mosque and a guest house, and is set in formal gardens bounded on three sides by a crenellated wall.

Image link	A bustling city scene at night with bright, colorful billboards and neon lights. Tall buildings covered with large advertisements. The street is filled with people and various light sources.	Times Square is a major commercial intersection, tourist destination, entertainment center, and neighborhood in Midtown Manhattan, New York City. It is formed by the junction of Broadway, Seventh Avenue and 42nd Street. Together with adjacent Duffy Square, Times Square is a bowtie-shaped space five blocks long between 42nd and 47th Streets.
Image link	A large white neoclassical building with a prominent dome. The building is illuminated against a dark blue sky.A statue of a horseman stands in front of the building. There are stairs and pathways leading up to the entrance.	The United States Capitol, often called The Capitol or the Capitol Building, is the meeting place of the United States Congress and the seat of the legislative branch of the U.S. federal government. It is located on Capitol Hill at the eastern end of the National Mall in Washington, D.C. Though no longer at the geographic center of the federal district, the Capitol forms the origin point for the district's street-numbering system and the district's four quadrants.
Image link	A grand white building with tall columns at the entrance.A circular fountain with green water in front of the building. The building is surrounded by a well-maintained lawn.AnAmerican flag flies from the top of the building. Bare trees are visible, suggesting a winter or early spring setting.	The White House is the official residence and workplace of the president of the United States. It is located at 1600 Pennsylvania Avenue NW in Washington, D.C., and has been the residence of every U.S. president since John Adams in 1800. The term White House is often used as a metonym for the president and his advisers.

Image link	Two tall, identical skyscrapers standing side by side. Each tower has a simple rectangular design with a grid of windows. The towers dominate the skyline, surrounded by smaller buildings. The sky is clear and blue, enhancing the sharp outlines of the buildings.	The original World Trade Center (WTC) was a large complex of seven buildings in the Financial District of Lower Manhattan, New York City, United States. It opened on April 4, 1973, and was destroyed in 2001 during the September 11 attacks. At the time of their completion, the Twin Towers—the original 1 World Trade Center (the North Tower) at 1,368 feet (417 m); and 2 World Trade Center (the South Tower) at 1,362 feet (415.1 m)—were the tallest buildings in the world. Other buildings in the complex included the Marriott World Trade Center (3 WTC), 4 WTC, 5 WTC, 6 WTC, and 7 WTC. The complex contained 13,400,000 square feet (1,240,000 m2) of office space.

Table 2: Scenes images: Taken from Bainbridge et al[31].

Image	Visual Text	Abstract Text
	a section of a tall wooden structure with red lettering spelling 'CYCLONE' and a smaller sign above it reading 'ASTROLAND.' At the top of the structure, a small group of people in a vehicle-like form with arms raised is visible against a gray sky.	a type of amusement ride employing a form of elevated railroad track that carries passengers on a train through tight turns, steep slopes, and other elements designed to produce a thrilling experience.
	a small bathroom with a white toilet, a rectangular sink with a cabinet below it, and a tiled bathtub in the background. The walls and tiles are neutral in color, and a silver faucet and showerhead are visible on the tiled surface.	a room, typically in a home or other residential building, that contains either a bathtub or a shower (or both).
	a bedroom with a four-poster bed covered in light bedding and decorative pillows, surrounded by furniture including a small bench and a wooden cabinet. The room has a warm color palette, and there is a tall potted plant and framed artwork on the wall in the background.	a room situated within a residential or accommodation unit characterised by its usage for sleeping

	a long bridge illuminated with blue and yellow lights, supported by a central vertical structure with cables extending outward. The bridge spans towards a horizon with a darkening sky transitioning from deep blue to orange, and a water tower is visible in the distance.	a structure built to span a physical obstacle (such as a body of water, valley, road, or railway) without blocking the way underneath. It is constructed for the purpose of providing passage obstacle
	a conference room with U-shaped wooden tables surrounded by red chairs, illuminated by fluorescent ceiling lights. The walls display posters, and there are indoor plants and a whiteboard in the background.	a room provided for singular events such as business conferences and meetings. Room It is commonly found at large hotels and convention centers though many other establishments
	a rectangular concrete structure with water passing beneath it and metallic gates positioned across the opening. The surrounding landscape features sparse vegetation, rocks, and a clear blue sky above.	a barrier that stops or restricts the flow of surface water or underground stream
	an industrial complex with tall chimneys, cylindrical structures, and a network of pipes and walkways, surrounded by sparse vegetation. The scene is set against a cloudy sky, with a mixture of rusted and metallic colors on the structures.	an industrial facility, often a complex consisting of several buildings filled with machinery, where workers manufacture items or operate machines which process each item into another.

	a cascading stone fountain with water flowing over its edges, set within a paved courtyard surrounded by greenery. The fountain is adorned with two white globes on either side and a background of yellow flowers and tall, dark trees.	source or spring, is a decorative reservoir used for discharging water. It is also a structure that jets water into the air for a decorative or dramatic effect
	a small red building with white trim, a steep roof, and a white door at the center. The building is surrounded by a hedge in the foreground and is set against a clear sky.	a single-unit residential building. It may range in complexity from a rudimentary hut to a complex structure of wood, masonry, concrete or other material
	A wooden counter occupies the foreground, topped with a decorative figure and surrounded by chairs with curved backs. The background features white cabinets, a light-blue wall, and windows adorned with patterned curtains.	a room or part of a room used for cooking and food preparation in a dwelling or in a commercial establishment
	a room with bright blue furniture, a white and multicolored circular-patterned carpet, and a wall-mounted piece of art in shades of blue.	a room for relaxing and socializing in a residential house or apartment

	a fenced playground area with a colourful plastic play structure featuring slides, surrounded by wood chips, and additional play equipment in the background.	a place designed to provide an environment for children that facilitates play, typically outdoors
	a large rectangular pool with clear blue water reflecting the sky above, surrounded by tiled flooring with a geometric pattern. In the background, there are palm trees, white lounge chairs, and a white structure with an arch and canopy, all set under a cloudy sky.	a structure designed to hold water to enable swimming or other leisure activities. Pools can be built into the ground or built above ground
	Multiple buildings are visible, with one prominently featuring dark, rectangular windows set in a reddish-brown grid pattern. In front, lighter-colored, ornate facades with intricate roof details contrast with the larger, modern structure.	an enclosed structure with a roof and walls, usually standing permanently in one place, such as a house or factory.
	a straight, wide road stretching into the horizon under a bright blue sky dotted with fluffy white clouds. Both sides of the road are bordered by sparse vegetation and barren land, with distant mountains faintly visible in the background.	a thoroughfare for the conveyance of traffic that mostly has an improved surface for use by vehicles and pedestrians.