# Distinct representation of navigational action affordances in human behavior, brains and deep neural networks

Clemens G. Bartnik[1*], Christina Sartzetaki[1], Abel Puigseslloses Sanchez[1], Elijah Molenkamp[1], Steven Bommer[1], Nikolina Vukšić[1], Iris I. A. Groen[1,2]

[1] Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[2] Psychology Research Institute, University of Amsterdam, Amsterdam, The Netherlands
[*] c.g.bartnik@uva.nl

## Abstract

Humans can use a variety of actions to navigate their immediate environment. To decide how to move, the brain must determine which navigational actions are afforded by the current environment. Here, we demonstrate that human visual cortex represents navigational action affordances of complex natural scenes. Behavioral annotations of possible navigational actions (e.g., walking, cycling) show that humans group environments into distinct affordance clusters using at least three separate dimensions. Representational similarity analysis of multi-voxel fMRI responses in scene-selective visual cortex regions shows that perceived affordances are represented in a manner that is only partly explained by other scene properties (e.g. contained objects), and independent of the task performed in the scanner. Visual features extracted from deep neural networks (DNNs) pretrained on a range of other visual understanding tasks fail to fully account for behavioral and neural representations of affordances. While DNN predictions of human-perceived affordances improve when training directly on affordance labels, the best human-model alignment is observed when visual DNN features are paired with rich linguistic representation in a multi-modal large-language model. These results uncover a new type of representation in the human brain that reflect action affordances independent from other scene properties.

## Significance

To navigate the world around us, we can use different actions, such as walking, swimming or climbing. How does our brain compute and represent such environmental affordances? Here, we show that high-level visual regions in the human brain represent navigational action affordances independent from other visual elements such as surface materials and objects, and do so in an automatic manner. We also demonstrate that commonly used models of visual processing in human brains, namely object- and scene-classification trained deep neural networks, do not strongly represent this information. In fact, our best prediction of human perception of action affordances is derived from prompting a multi-modal large language model. Our results suggest that human action affordance perception relies on specialized neural representations different from those used for other types of visual understanding tasks.

## Introduction

Humans navigate their local environments with remarkable ease, whether cycling through busy city streets or hiking along rugged trails. This highlights the ability of the human brain to not only effortlessly process visual information, but also identify appropriate navigational actions from a wide range of potential actions. The recognition of ecologically relevant behaviors in the immediate environment is commonly referred to as *affordance perception* (Gibson, 1977). Theoretical frameworks of affordance perception are long-standing in fields such as ecological psychology (Gibson, 1977; Rietveld and Kiverstein, 2014; Osiurak et al., 2017), but the neural mechanisms mediating this ability are only just beginning to be explored. Here, we test whether the human brain distinctly represents different navigational actions afforded by real-world scenes.

While initial research on scene perception emphasized scene-defining objects and their relationships (e.g. Biederman et al., 1982; Aminoff et al., 2013), later work included global features such as spatial layout as well as action possibilities afforded by the environment (e.g. Oliva and Torralba, 2001; Greene and Oliva, 2009); see Bartnik and Groen (2023) for a review. Indeed, action possibilities strongly affect human scene categorization (Greene et al., 2016). However, it is unclear how the brain computes this information. Functional magnetic resonance imaging (fMRI) has identified three brain regions — the Parahippocampal Place Area (PPA), Occipital Place Area (OPA), and Retrosplenial Complex (RSC) — involved in processing complex visual scenes (Epstein and Baker, 2019). Of these, OPA has been suggested to represent navigable space (pathways) in indoor environments that afford walking (Bonner and Epstein, 2017), possibly facilitated by representations of scene layout and floor elements (Bonner and Epstein, 2018; Dwivedi et al., 2021b). However, other work found no evidence for representation of affordances such as sports and transportation in scene-selective regions, despite their strong influence on perceived scene similarity (Groen et al., 2018).

A priori, it seems likely that disambiguating *different types* of affordances requires perceiving not only spatial properties of scenes (e.g., the presence of navigable surfaces such floors or roads) but also surface materials (asphalt for driving or cycling) and contained objects (the rocks for climbing, a body of water for swimming). Scene-selective regions have been implicated in the representation of all of these properties; for example, PPA has been shown to represent the spatial layout of scenes, (Kravitz et al., 2011; Park et al., 2011), but also texture (Henriksson et al., 2019; Cant and Xu, 2012) and objects (Janzen and van Turennout, 2004; Marchette et al., 2015; Harel et al., 2013). However, it is unclear if a combination of these properties is sufficient to predict the perceived affordances of a scene, and their corresponding representation in the human brain, or whether additional (possibly non-visual) information is involved. Moreover, to our knowledge, no prior study has actively probed participants to report perceived scene affordances during fMRI measurements, which may impact neural representations of features diagnostic for affordance recognition.

Here, we close these gaps by testing how well various scene properties predict affordance representation assessed with both behavioral and fMRI measurements. Using a novel set of natural scenes spanning six common navigational actions, we specifically test the hypothesis that human scene-selective cortex distinctly represents affordances that pertain to navigation. To probe the underlying computations, we assert whether affordance representations are task-dependent, and to what extent they can be predicted by features extracted from a wide variety of DNNs, including supervised object- and scene-classification trained CNNs, self-supervised models (DINO), language-supervised models (CLIP) and vision transformers (ViTs). Collectively, our results provide novel evidence for action affordance representation in the scene-selective cortex that is not easily captured by other scene properties or by DNNs optimized to predict those properties.

## Results

We investigated how the human brain processes visual information from real-world scenes to perceive action possibilities, referred to as navigational action affordances. We first curated a novel set of scene stimuli, spanning indoor, outdoor man-made, and outdoor natural environments (**Fig. 1A**). Human participants ($n$ = 152) then annotated these scenes on six different actions that people can use to move in the immediate environment (**Fig. 1B**, left). The same scenes were also annotated for four visual scene properties thought

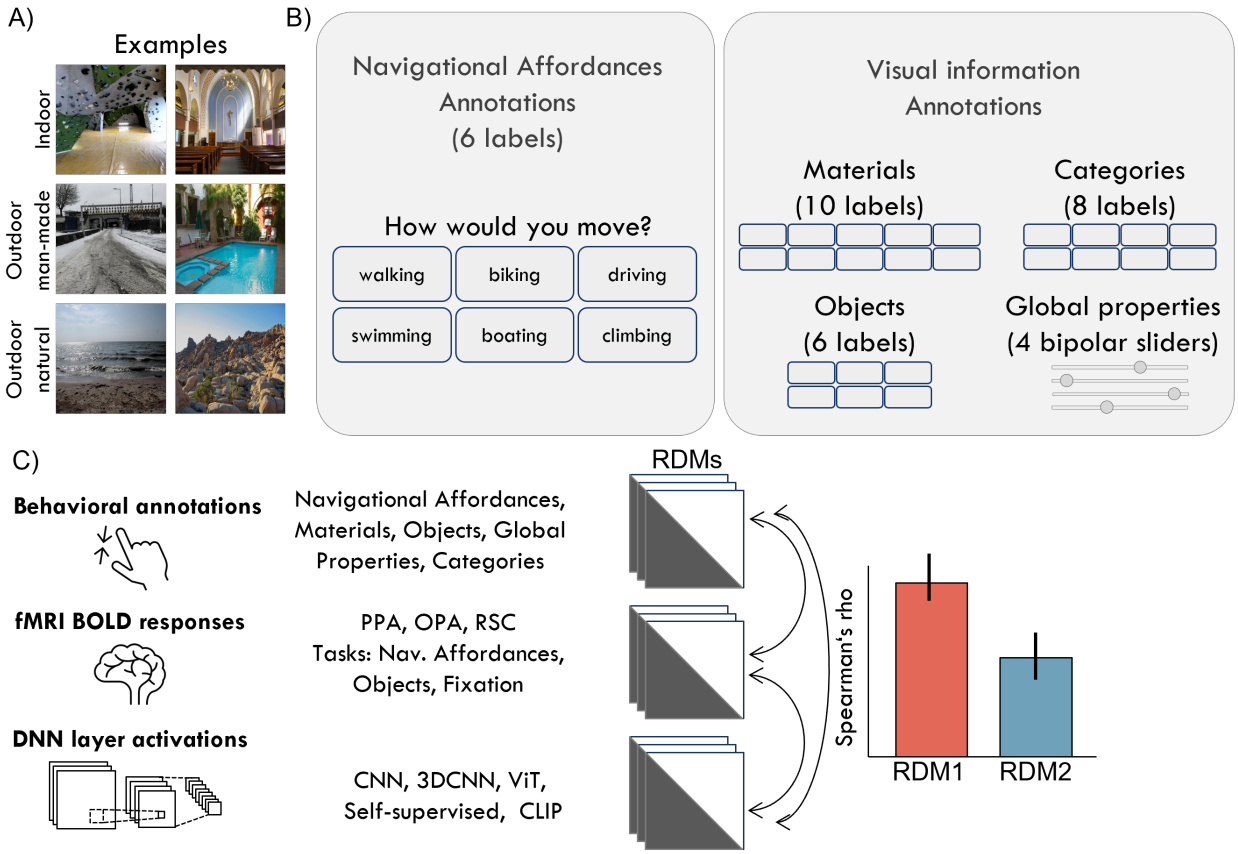



**Figure 1: Stimuli, behavioral tasks and experimental design (A)** Examples from our newly curated stimulus set to measure navigational affordance representation consisting of 231 high-resolution photographs (collected from Flickr containing (CC0 1.0, CC BY 2.0 and CC BY-SA 4.0 licensed), evenly distributed across indoor, outdoor man-made, and outdoor natural environments. **(B)** Overview of the five labeling tasks used to collect behavioral annotations. In addition to annotating 6 navigational affordances depicted on the left, participants also annotated the scenes on perceived materials (labels: Water, Wood, Vegetation, Dirt/Soil, Sand, Stone/Concrete, Snow/Ice, Pavement, Carpet, Metal), scene categories (labels: Room, Inside city, Hallway, Forest, Desert, Open country, Mountain, Coast/River), objects (labels: Building/Wall, Tree/Plant, Road/Street, Furniture, Body of water, Rocks/Stones) and four global properties (Man-made/Natural, Open/Closed, Near/Far, Navigable/Non-navigable). **(C)** Methods overview: Behavioral annotations, multi-voxel BOLD patterns in scene-selective cortex under three different task instructions (navigational affordance labeling; object labeling; orthogonal task at fixation), and DNN activations were collected in response to the new stimulus set. Representational similarity analysis (RSA) was then used to compare the resulting representational spaces, by computing representational dissimilarity matrices (RDMs) for each type of measurement and correlating them to one another.

to be important for scene perception (scene materials, scene categories, contained objects, and global properties; **Fig. 1B**, right). We computed pairwise dissimilarities between images separately for each annotation task, resulting in five representational dissimilarity matrices (RDMs), that we compared with one another. We used a subset of these images and tasks in an fMRI experiment, where we compared navigational affordance and object representations to fMRI BOLD responses in scene-selective cortex, as well as visual feature activations extracted from a large variety of DNNs (**Fig. 1C**).

## Navigational affordances form a distinct representational space

To understand how humans represent navigational affordances, we sorted the RDM obtained from the affordance labeling task (**Fig. 2A**, left) in two different ways: by the type of environment (indoor, outdoor-manmade, outdoor-natural) and after applying hierarchical clustering (**Fig. 2A**, right). The RDM sorted on environment type (**Fig. 2A**, left) shows that indoor scenes are most similar to one another in terms of the navigational actions they afford, while outdoor scenes are more dissimilar. This is likely because most indoor environments primarily afford walking, while outdoor environments typically afford multiple actions

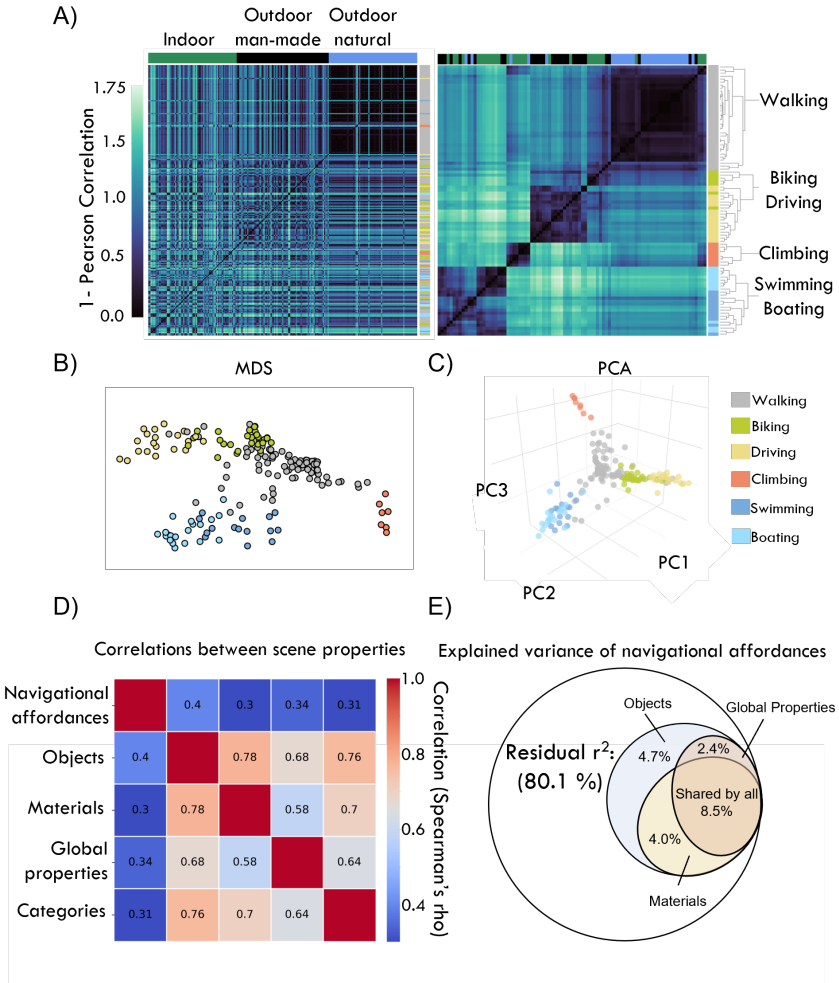



**Figure 2: Navigational affordances form a representational space that is distinct from other scene properties. (A)** RDM derived from navigational affordance annotations, sorted on environment type (left) or after hierarchical clustering (right). Row legend colors indicate different actions and are determined based on the predominant affordance label (see Materials and Methods). **(B)** Multi-dimensional scaling (MDS) of the navigational affordance RDM shows that clusters extend in three directions. A high-resolution version of this figure showing stimulus thumbnails is provided in **Supplementary Figure 1**. **(C)** Principal component analysis (PCA) applied to the navigational affordance annotations reveals a three-dimensional structure. **(D)** Annotations of other scene properties show high correlations between objects, materials, global properties, and scene categories, while navigational affordances correlate notably lower with these scene properties, highlighting their unique representational space. Comparisons using alternative distance metrics are provided in **Supplementary Figure 2**. RDMs for each behavioral annotation task, as well as a comparison with automated labeling methods, are provided in **Supplementary Figure 3**. **(E)** Variance partitioning of navigational affordance annotations by the three highest correlating scene properties. The Euler diagram shows the unique and shared variance in navigational affordances explained by objects, materials, and global properties. In total, 80.1% of the total variance in the navigational affordances representational space remains unexplained (note that the residual circle is not drawn to scale).

(e.g., walking and cycling). Indeed, applying hierarchical clustering (see Materials and Methods) on the affordance RDM yields at least four major clusters, the biggest of which refers to walking, mostly containing indoor scenes (**Fig. 2A**, right). However, this cluster also contains images from other environments, and the remaining clusters also span multiple environment types, indicating there is no one-to-one mapping of environment type to navigational affordance.

To better understand the clustering of navigational affordances, we applied multidimensional scaling (MDS) (**Fig. 2B**) to visualize the similarities in behavioral affordance annotations for our stimuli in a 2D space. This visualization shows that walking, as the most common category, forms a central hub surrounded by clusters for the other actions, one extending towards biking and driving, another towards swimming and boating, and

a third towards climbing. This three-dimensional structure is also evident when visualizing the first 3 dimensions of a PCA performed on the affordance annotations directly (**Fig. 2C**). These results suggest that the navigational affordance labels provided by our human participants form a highly structured representational space spanning at least three distinct dimensions.

We quantified the relation between navigational affordances and other scene properties by correlating the affordance RDM with RDMs derived from annotations of objects, materials, scene categories and global properties, obtained in the same behavioral experiment (**Fig. 2B**, right). Affordances show lower correlations with other scene properties (average $\rho$ = 0.34, SD = 0.04) compared to how correlated these are with one another (average $\rho$ = 0.69, SD = 0.07) (**Fig. 2D**). This pattern was robust across multiple distance metrics for quantifying representational dissimilarity (**Supplementary Figure 2**). A multiple linear regression analysis further shows that even when combined, other scene properties only account for a small amount (19.9%) of the total variance in navigational affordances, with the only unique contribution coming from objects (unique $r^2$ = 4.7%). Materials and objects share 4.0% of the variance in affordance annotations, global properties and objects share 2.4%, and most of the explained variance (8.5%) is shared by all three scene properties (**Fig. 2E**). This highlights not only inherent covariance of visual scene properties (Malcolm et al., 2016), but also suggests that in human behavior, navigational affordances form a distinct representational space that is not trivially predicted by a linear combination of other scene properties.

The limited correspondence between action affordance representations and other scene properties is somewhat surprising, given our *a priori* intuition that we use (a combination of) readily available scene features to compute affordances, as well as the clear separation of action affordance annotations into distinct clusters separating, for example, water-based (swimming, boating) from land-based activities (**Fig. 2B,C**). One possibility is that our annotations of other scene properties, consisting of a handful of labels in each task, were not exhaustive enough to fully capture the representation of navigational affordances. To test this, we also constructed RDM based on outputs of classifiers trained on the SUN Attribute Database (Patterson et al., 2014), ADE20K (Zhou et al., 2017b) and Places365 (Zhou et al., 2017a), which have larger sets of labels of materials, global properties, categories, and objects (e.g., 150 objects in ADE20k; 37 materials in SUN Attribute Database; 365 scene categories in Places365). Despite constituting a more exhaustive sampling, these automated labels yielded similarly low predictions of our navigational affordance annotations (average $\rho$ = 0.21, SD = 0.09; **Supplementary Figure 3**). Moreover, classifier outputs of function labels (a subset of the labels included in the SUN Attribute Database) which include several navigational actions (e.g., cycling, driving) similarly correlated comparatively low with automated labels of other scene properties (average $\rho$ = 0.28, SD = 0.14). These results suggest that the low predictability of human navigational affordance perception by other scene properties is not driven by our precise choice of annotation labels, but rather reflects a persistent difference in representational spaces.

Together, these behavioral results suggest that our selection of six navigational action affordances of visual scenes occupy a three-dimensional representational space that weakly correlates with other scene properties and environmental types, with the strongest alignment observed with object representations.

## Scene-selective cortex uniquely represents navigational affordances

Above, we used behavioral annotations on a large set of scenes to establish the distinct representational space of affordances for six ways people can move in the world. Next, we tested our hypothesis that these navigational affordances are represented in the human visual system. Brain activity was measured using a 3T MRI scanner while participants ($n$ = 20) viewed a strategically sampled subset of 90 images that maintained the three-dimensional structure of the behavioral affordance space (see Materials and Methods and **Supplementary Figure 4B**). To examine the influence of task instructions on affordance representation in the brain, all participants performed three distinct tasks on each of the 90 images whilst in the scanner: affordance annotation, object annotation, and an orthogonal task at fixation (see Material and Methods).

We first established that our behavioral tasks transferred well to the scanner environment by analyzing the behavior of the fMRI participants on the affordance and object annotation tasks. Average RDMs de-

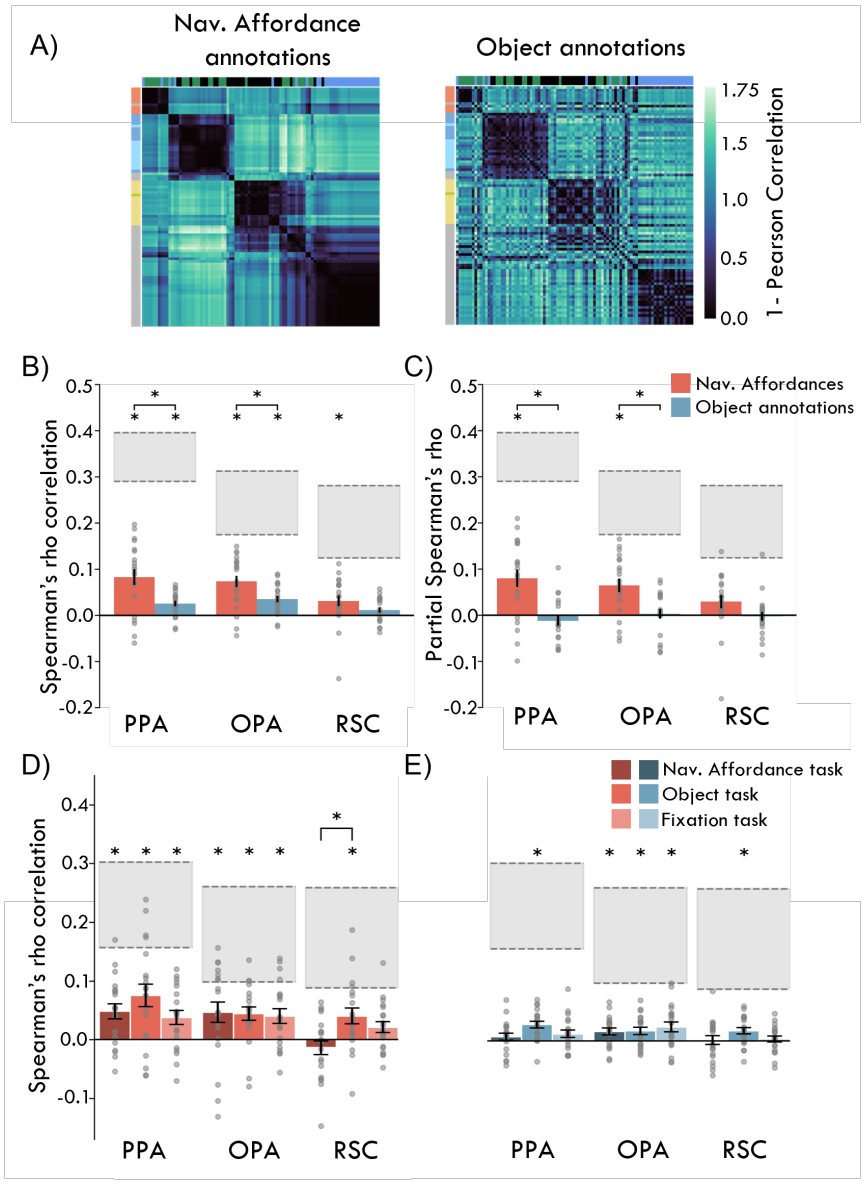



**Figure 3: Scene-selective regions represent navigational affordances (A)** RDMs derived from in-scanner navigational action affordance (left) and object (right) annotations, sorted by hierarchical clustering of the action affordance RDM. **(B)** Correlation of response patterns in scene-selective brain regions with navigational action affordance (red) and object annotations (blue) RDMs. Bars represent averages for individual participants (gray dots). Shaded areas delineate upper and lower noise ceilings, reflecting RDM similarity between individual participants and the group mean. Error bars indicate standard error of the mean (SEM) across participants. Asterisks (*), indicate one-sample t-tests against zero $p<0.05$ corrected for multiple comparisons using Bonferroni correction. Pairwise comparisons between action and object spaces are signified by horizontal brackets when $P<0.05$ in paired-sample t-tests (Bonferroni-corrected). **(C)** Correlation between action affordance and fMRI RDMs, when partialling out the object space, and vice versa. Plot elements as in (B). **(D)** Task-specific correlations with navigational action affordances RDM (dark shades: navigational affordance task; medium shades: object task; light shades: fixation task) **(E)** Task-specific correlations with object annotations RDM. Shades as in (D).

rived from in-scanner task responses (**Fig. 3A**) correlated strongly with those derived from annotations obtained in the behavioral experiment (affordances: $\rho = 0.85$; objects: $\rho = 0.82$), while still maintaining distinct characteristics (correlation between affordances and object RDMs; $\rho = 0.44$; **Supplementary Figure 5A**). In-scanner affordance annotations also exhibited high agreement among subjects (mean $\rho = 0.52$; **Supplementary Figure 5B**), showing that action affordance representations are not only stable across testing environments but also across participants, similarly to or even more than object annotations (mean $\rho = 0.33$).

To investigate what scene properties are represented in scene-selective brain areas, we first constructed individual subject RDMs based on pairwise correlation distances of multi-voxel activity patterns in PPA, OPA and RSC, averaged across all tasks (see **Supplementary Figure 6**) and calculated the correlation of these RDMs with the average behavioral RDMs derived from the in-scanner affordance and object annotations, respectively (**Fig. 3B**). We found evidence for significant navigational affordance representations in PPA (mean $\rho$ = 0.08, SEM = 0.017; one-sample t-test against zero, two-sided [$t(19)$ = 4.81, $P<0.001$, Bonferroni corrected]), OPA ($\rho$ = 0.07, SEM = 0.012); [$t(19)$ = 5.73, $P<0.001$] and RSC ($\rho$ = 0.03, SEM = 0.012); [$t(19)$ = 2.59, $P$ = 0.018]. Both PPA and OPA, but not RSC also represented object information (PPA; $\rho$ = 0.03, SEM = 0.006); [$t(19)$ = 4.18, $P<0.001$], OPA; ($\rho$ = 0.03, SEM = 0.007); [$t(19)$ = 4.83, $P<0.001$], RSC; $\rho$ = 0.01, SEM = 0.006); [$t(19)$ = 1.86, $P$ = 0.078]). Furthermore, through the use of paired samples t-tests, we established that responses in both PPA [$t(19)$ = 4.06, $P<0.001$] and OPA [$t(19)$ = 2.61, $P$ = 0.002] correlated significantly higher with the navigational affordance space than with the object space. In contrast, the RSC [$t(19)$ = 1.28, $P$ = 0.214] did not show significant differences between the two spaces. Similar results were obtained when comparing fMRI responses to individual participant behavior (**Supplementary Figure 7A**).

The stronger correlation with the action affordance respresentational space compared to object space suggests that navigational affordances are indeed represented in scene-selective cortex. However, given the substantial correlation between the affordance and object annotation-derived RDMs (**Fig. 3A**), some of the variance in fMRI response patterns may be equally well described by both spaces. To determine the degree of unique representation of each space, we computed partial correlations (**Fig. 3C**) and found significant unique correlations with the affordance space in PPA [$t(19)$ = 4.12, $P$ = 0.001] and OPA [$t(19)$ = 4.3,$P<0.001$], but not RSC [$t(19)$ = 2.02, $P$ = 0.058], and no significant partial correlation with the object space in any ROI (PPA: [$t(19)$ = -1.15, $P$ = 0.263]; OPA: [$t(19)$ = 0.29, $P$ = 0.773], RSC: [$t(19)$ = -0.26, $P$ = 0.797]). Paired t-tests furthermore show that the partial correlation with the affordance space is significantly higher than the object space for both PPA [$t(19)$ = 3.17, $P$ = 0.005] and OPA [$t(19)$ = 2.49, $P$ = 0.022] but not RSC [$t(19)$ = 1.33, $P$ = 0.198]. These results suggest that PPA and OPA, but not RSC, represent navigational action affordances of real-world scenes independent of their contained objects.

Comparisons of the fMRI response patterns with all five representational spaces derived from the independent behavioral experiment (**Fig. 1B**) furthermore show that affordance annotations exhibit the highest correlation in all three ROIs, followed by global properties (**Supplementary Figure 7B**). Unlike objects, affordances and global properties *both* exhibit a unique correlation with PPA and OPA (**Supplementary Figure 7C**). Overall, these results suggest that scene-selective PPA and OPA represent different navigational action affordances, and that these representations are at least partly distinct from other scene properties.

## Representation of action affordances in scene-selective regions is task-independent

Having established unique action affordance representations in scene-selective cortex, we next examined how task instructions affect the strength of these representations in the brain. To do so, we generated RDMs based on fMRI responses separated by the three different tasks conducted by subjects in the scanner (navigational affordance labeling, object labeling, fixation cross color labeling) and correlated each of these with the RDMs derived from in-scanner behavioral annotations (**Fig. 3A**). If navigational affordances are only represented in the brain when task-relevant, we expect to find significant correlations with the navigational affordance space only during the navigational affordance labeling task, but not the object labeling task or the fixation task. Alternatively, if navigational affordances are computed automatically, significant correlations with the navigational affordance space are expected in all three tasks. A third possibility is that navigational affordances are represented when participants respond to the scene images (i.e., during active labeling of either affordances or objects), but not when they respond to the fixation cross (fixation task).

The results, shown in **Fig. 3D**, are consistent with task-independent computation of navigational affordances in PPA and OPA, but not RSC. Both PPA and OPA exhibited significant correlations with the affordance space across all three tasks (PPA navigational affordance task ($\rho$ = 0.05, SEM = 0.013); [$t(19)$ = 3.71, $P<0.01$], PPA object task ($\rho$ = 0.03, SEM = 0.006); [$t(19)$ = 3.8, $P<0.01$], PPA fixation task ($\rho$ = 0.04, SEM = 0.012);

[$t(19)$ = 3.14, $P$ = 0.01]; OPA navigational affordance task ($\rho$ = 0.02, SEM = 0.006); [$t(19)$ = 2.62, $P$ = 0.02], OPA object task ($\rho$ = 0.04, SEM = 0.011); [$t(19)$ = 3.82, $P<0.01$], OPA fixation task ($\rho$ = 0.02, SEM = 0.008); [$t(19)$ = 3.17, $P$ = 0.01]). For RSC we find a significant correlation with the affordance space during the object task ($\rho$ = 0.04, SEM = 0.013); [$t(19)$ = 2.93, $P$ = 0.01], but not the navigational affordance task ($\rho$ = -0.01, SEM = 0.013); [$t(19)$ = -1.15, $P$ = 0.26] or fixation task ($\rho$ = 0.02, SEM = 0.009); [$t(19)$ = 2.24, $P$ = 0.04]. Moreover, paired tests between correlations with the affordance space across tasks in OPA and PPA were non-significant [all $t(19) < 0.580$] and [$P > 0.206$]. This suggests that the representations of navigational affordances within these scene-selective regions remain consistent regardless of the task at hand. The RSC shows a different, unexpected pattern, with a significantly lower correlation with the action affordance space during the action affordance task compared to the object task [$t(19)$ = -3.217, $P$ = 0.005].

Representations of objects (**Fig. 3E**) appear to be comparatively more task-dependent in PPA and RSC, which only show significant correlations with the object space during the object task PPA ($\rho$ = 0.03, SEM = 0.006); [$t(19)$ = 4.8, $P<0.01$], RSC ($\rho$ = 0.02, SEM = 0.005); [$t(19)$ = 3.13, $P$ = 0.01]. In contrast, the OPA again demonstrates significant correlations in all tasks (navigational affordance task ($\rho$ = 0.02, SEM = 0.006); [$t(19)$ = 2.77, p=0.01], object task ($\rho$ = 0.02, SEM = 0.006); [$t(19)$ = 2.76, $P$ = 0.01]; fixation task ($\rho$ = 0.02, SEM = 0.008);[$t(19)$ = 2.82,$P$ = 0.01]. Similar results were obtained when comparing partial correlations across tasks (**Supplementary Figure 8**). Together, these results suggest that the representations for navigational affordances are automatically extracted by scene-selective regions OPA and PPA.

## Pre-trained DNNs show comparatively weaker alignment with action affordances

To further investigate the computations underlying action affordance perception, we again used RSA to examine to what extent deep neural networks (DNNs) are representationally aligned (Sucholutsky et al., 2023) with human-perceived navigational affordances. We sample multiple architectures (both CNNs and ViTs), task objectives (object/scene classification, scene segmentation, video classification) and training procedures (supervised learning using labels, contrastive learning using CLIP, self-supervised learning using DINO) and focus on the comparative ability of these models to predict the similarity between human annotations of perceived navigational affordances versus objects. We computed RDMs from the activations of a subset of layers in each model, and compared these to RDMs derived from participant behavior.

All DNNs demonstrate significant correlations with both navigational affordance and object representations (all [$t > 17.87$, all $P < 0.001$]; **Fig. 4A-B**). However, across all models, correlations were substantially lower with navigational affordances (average $\rho$ = 0.19, SD = 0.06) than with objects (average $\rho$ = 0.31, SD = 0.07) [pairwise $t(12)$ = 10.23, $P < 0.001$]. Classic CNNs optimized for object classification on ImageNet correlated the least with human behavioral action affordances, but alignment increased for CNNs trained on scene datasets, with the highest correlation for the ResNet50 trained on semantic segmentation (ADE20k: $\rho$ = 0.21, SEM = 0.008). Video-trained 3DCNNs for action recognition (X3D and SlowFast) and CNNs trained with alternative training objectives (DINO or CLIP) show comparable correlations as these scene-trained models, while Vision Transformers (ViTs) exhibit the highest correlations with navigational affordances (maximum for ViT Base Patch 16 with $\rho$ = 0.29, SEM = 0.001). Human object annotations (**Fig. 4B**) also showed the highest correlation with the ViT Base Patch 16 ($\rho$ = 0.38, SEM = 0.02).

These results suggest that, unlike human scene-selective cortex, the DNNs we tested generally exhibit higher correspondence with object representations than with navigational action affordances. Indeed, even the best correlating DNN with in-scanner affordance annotations (ViT Base Patch 16) still correlated significantly higher with objects [pairwise $t(19)$ = 3.66, $P$ = 0.0017]. Moreover, the relatively large gap between best-correlating model and the noise ceiling for navigational affordance behavior shows that a substantial amount of variance in this highly stable representational space in humans is unexplained by DNNs. We find the same pattern when comparing this ViT with all behavioral annotations and alternative spaces for the full set of images (**Fig. 4C**): both navigational affordance spaces, obtained either from human behavioral annotations or automatic labeling, show significantly lower correlations with the DNN than with all other representational spaces (with the exception of scene categories from Places365).

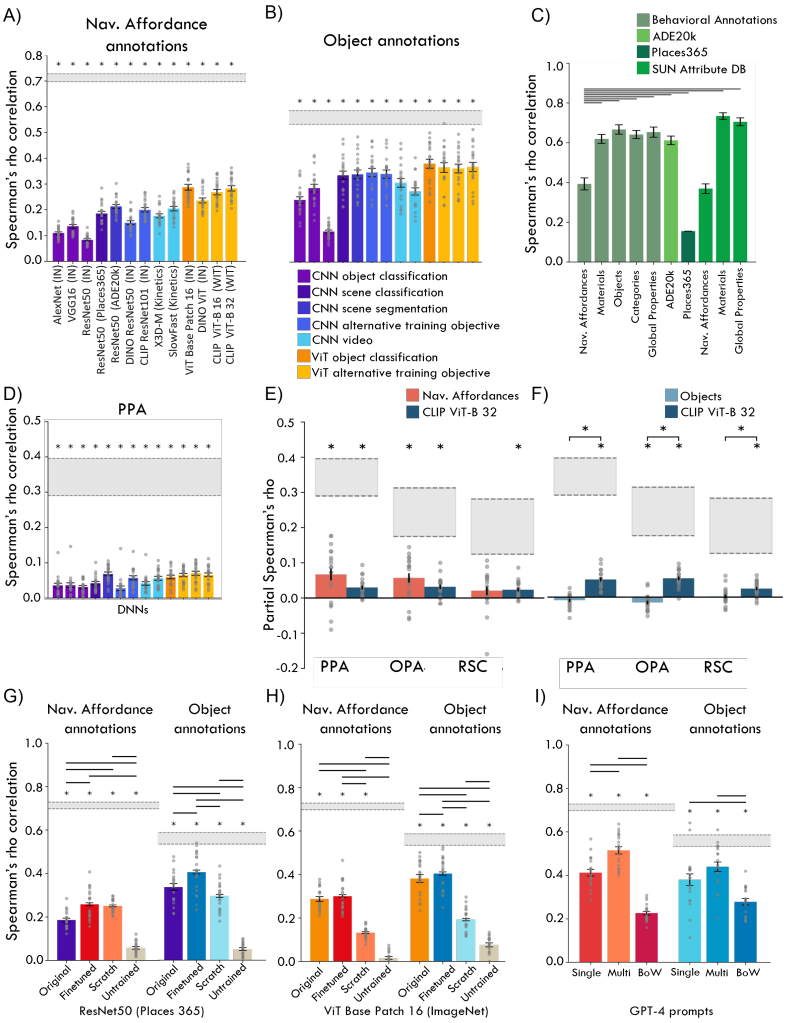



**Figure 4: Comparison of DNNs with human behavioral annotations and fMRI responses in scene-selective cortex. (A)** Average correlation (Spearman's $\rho$) between in-scanner navigational affordance action annotation RDMs and RDMs derived from feature activations from DNNs with different architectures and training objectives. Bars represent average correlations across participants for the best correlating layer in each model (see **Supplementary Table 2**) and gray dots indicate individual participants. Significant t-tests of the average correlation against zero are marked by asterisks (*), indicating p<0.05 corrected for multiple comparisons using Bonferroni correction. The shaded area delineates the upper and lower noise ceiling and error bars indicate the standard error of the mean (SEM) across participants. **(B)** Correlation of various DNNs with in-scanner object annotation RDMs. Plot elements as in (A). **(C)** Average correlation of the best correlating layer of the ViT Base Patch 16 DNN with the annotations for the full set of images from the online behavioral experiment and automatic labeling outputs. Black horizontal lines indicate significant pairwise comparisons with a two-tailed sign test, corrected for multiple comparisons using p <0.05 FDR correction, Error bars indicate 95% confidence intervals derived from a bootstrap distribution (n = 1000) bootstraps). **(D)** Average correlation of each DNN's best correlating layer with representational similarity in PPA (averaged across tasks). Results for OPA and RSC are reported in **Supplementary Figure 10**. Plot elements as in (A). **(E)** RSA correlations between scene-selective regions and navigational affordances when partialling out the CLIP ViT-32 representational space, and vice versa. **(F)** Correlation between scene-selective regions and object annotations when partialling out the CLIP ViT-32 representational space, and vice versa. **(G)** Average correlation (Spearman's $\rho$) between human behavioral RDMs and RDMs derived from various ResNet50 models. Original: ResNet50 (Places365) same as in (A-B). Finetuned: Last fully connected layer fine-tuned on function labels of Scene Attribute Database. Scratch: All layers of the ResNet50 are retrained using function labels. Untrained: a randomly initialized model. Plot elements as in (A). Black horizontal lines indicate significant pairwise comparisons at p <0.05, Bonferroni-corrected for multiple comparisons. **(H)** Average correlation (Spearman's $\rho$) between human behavioral RDMs and RDMs derived from various ViT Base Patch 16 models. Plot elements as in (G). **(I)** Average correlation between RDMs derived by prompting OpenAIs GPT-4 and in-scanner human behavioral responses, using three types of prompts: choosing a single (Single) or multiple (Multi) affordance/object option of the same six possible options given to human participants; or list a maximum of 50 possible actions/objects in the scene (bag-of words (BoW; see **Supplementary Table 5**). Plot elements as in (G).

Together, these results demonstrate significantly lower alignment of DNNs with navigational affordances compared to objects. While this is unsurprising for DNNs explicitly trained to classify objects, we note that this result also holds for models trained on other tasks such as scene classification and scene segmentation, and models trained with self-supervised (DINO) and contrastive learning objectives (CLIP).

## Affordance representation in scene-selective cortex is independent of DNN features

To determine how well the different types of DNNs are aligned with human brain responses to our novel stimulus set, we also used RSA to compare their feature activations with response patterns in PPA (**Fig. 4D**, and OPA and RSC (**Supplementary Figure 10**). All models show significant correlations with the scene ROIs (PPA: all [$t(19) > 4.08$, $P < 0.001$], OPA: all [$t(19) > 4.51$, $P < 0.001$] and RSC: all [$t(19) > 4.71$, $P < 0.001$]), with ViTs showing numerically highest correlations in all three ROIs (PPA, CLIP ViT-B 32, $\rho$ = 0.070, SEM = 0.006; OPA, CLIP ViT-B 32, $\rho$ = 0.061, SEM = 0.004; RSC, CLIP ViT-B 16 ($\rho$ = 0.043, SEM = 0.004). However, model correlations are substantially below noise ceiling, and models show comparatively less differentiation than when they are compared to behavioral measurements (**Fig. 4A-B**).

Partial correlation analyses furthermore show that the alignment of DNN features with brain activity primarily reflects object-related information rather than affordance-related information. Both PPA and OPA exhibit significant unique correlations with navigational affordances (PPA: $\rho$ = 0.067, SEM = 0.017); [$t(19)$ = 3.87, $P$ = 0.001]; OPA: $\rho$ = 0.057, SEM = 0.013); [$t(19)$ = 4.17, $P$ = 0.001]) when partialling out the best correlating DNN (CLIP ViT-B 32), with the DNN itself also explaining unique variance in these ROIs (PPA: $\rho$ = 0.03, SEM = 0.006); [$t(19)$ = 5.11, $P < 0.001$]; OPA: ($\rho$ = 0.03, SEM = 0.006); [$t(19)$ = 4.9, $P < 0.001$], **Fig. 4E**). In contrast, neither PPA or OPA show significant positive correlations with object annotations when partialling out DNN features (PPA: ($\rho$ = -0.009, SEM = 0.004); [$t(19)$ = -2.05, $P$ = 0.054]; OPA: ($\rho$ = -0.015, SEM = 0.005); [$t(19)$ = -2.94, $P$ = 0.008], **Fig. 4F**). RSC again shows a deviant pattern, exhibiting no significant unique correlations with either the navigational affordance space ($\rho$ = 0.020, SEM = 0.012, [$t(19)$ = 1.61, $P$ = 0.125] or the object space ($\rho$ = 0.002, SEM = 0.005); [$t(19)$ = 0.38, $P$ = 0.707]). These results suggest that DNN features do not fully account for navigational affordances representation in human PPA and OPA, instead capturing object representations in these regions.

## DNN affordance representation trained with supervision and in multi-modal LLMs

Our results so far point to a consistent gap in DNN's ability to explain object versus navigational affordance representations, reflected in human behavioral and brain measurements. In this final section, we explored two ways in which DNN alignment with human affordance perception could potentially be improved.

First, while we so far tested a variety of DNNs, none of them was *explicitly* trained to perform navigational affordance perception. Therefore, we assessed if DNN alignment with human behavior increases when using explicit supervision with affordance labels from the SUN Attribute Database subset of function labels (see Materials & Methods and **Supplementary Table 4**). We find that both end-to-end training and fine-tuning improve correlations with human-perceived affordances relative to a ResNet50 trained on scene classification (**Fig. 4G**) [pair-wise t-test on fine-tuned vs. original $\Delta\rho$ = 0.07, $t(19)$ = 6.89, $P < 0.001$] and [end-to-end training vs. original $\Delta\rho$ = 0.06, $t(19)$ = 7.5, $P < 0.001$]. However, for the already better aligned ViT Base Patch 16 pretrained on ImageNet, fine-tuning only marginally improved the behavioral alignment with affordances, and training from scratch reduced correlations relative to the original model (**Fig. 4H**). This suggests that with direct mapping to affordance labels, visual features extracted from (pre-trained) DNNs can approximate human-perceived affordances to some degree. However, regardless of architectures and training method, a gap in alignment relative to object representations still remains (between navigational affordances and objects for ResNet50: difference in $\rho$ = 0.147 and ViT Base Patch 16: difference in $\rho$ = 0.104).

Second, the emergence of multi-modal large language models such as GPT-4 (OpenAI et al., 2023) offers the possiblity to test DNNs on tasks that they have not been trained for. To explore if such a model can predict navigational affordances, we conducted an analogous behavioral experiment with GPT-4 as we did with human participants. We presented the images to GPT-4 with a few different prompts, asking either for a

single or multiple actions/objects from the same list of response options as presented to human participants, or to name all possible actions and objects in a given scene to create a bag of word (BoW) representation of the scene (see **Supplementary Table 5** for exact prompts used). RDMs were created using pairwise 1-Pearson correlation of the one-hot encoded possible navigational affordances/objects (see Methods for details). As shown in **Fig. 4I**, the multi-modal LLM generates behavioral output that now aligns with navigational affordances (all [$t(19) > 25.05$, $P < 0.001$]) to a similar extent as object annotations (all [$t(19) > 13.97$, $P < 0.001$]), although a gap relative to the noise ceiling still remains. While it is unclear what features ChatGPT uses to generate these outputs, this result suggests that pairing visual DNN features with linguistic representations may help predict navigational affordance perception at an near-human level.

Collectively, our findings show that DNNs trained on object or scene classification and other commonly-used visual understanding tasks capture only some aspects of human scene representations. Specifically, we show that alignment with object representations does not automatically translate into good alignment with navigational affordances. However, utilizing direct supervision or rich linguistic representations may allow DNN models to more adequately capture navigational affordance perception in humans.

# Discussion

The aim of our study was to assess whether the human visual system represents different navigational actions of complex natural scenes, and if these representations are distinct from other scene properties. The notion of action affordances shaping visual perception is long-standing in psychology (Gibson, 1977), but the presence of distinct action affordance representations for scene environments in the human brain has so far not been demonstrated. Utilizing behavioral and fMRI measurements alongside computational deep neural network models, we discovered that human perceived navigational affordances - reflecting different ways we can move through our local environment - form a unique representational space that is not trivially predicted from other perceived properties such as objects or global layout. Moreover, scenes associated with specific navigational actions, such as driving or swimming, evoked similar multi-voxel activation patterns in PPA and OPA (but not RSC), and by accounting for variation in other visual scene properties, we find that these navigational affordances are indeed independently represented in the brain. Importantly, these representations proved to be task-independent, suggesting that OPA and PPA encode distinct navigational action affordances automatically even when participants passively view the images or perform a non-navigational task. While pretrained DNNs partially explained activity in scene-selective regions, their alignment with navigational affordances was lower than with object representations, and increasing this alignment may require direct supervision or pairing with richer linguistic representations as found in large language models.

Prior fMRI research has linked navigational affordance perception primarily to OPA, with evidence accumulating that it encodes egocentric information relevant to navigate the immediate environment (Persichetti and Dilks, 2016; Dilks et al., 2011; Julian et al., 2016; Kamps et al., 2016b,a). Several studies have found that OPA is involved in the perception of walkable pathways within indoor visual scenes (Bonner and Epstein, 2017; Park and Park, 2020), possibly mediated by the representation of extended surfaces and floor elements in the lower visual field (Bonner and Epstein, 2018; Dwivedi et al., 2021b). Our results extend these findings by considering multiple types of environments and navigational actions other than walking, and our results suggest that OPA may additionally represent *different types* of navigational actions that one could perform on such navigable surfaces. A recent study by Jones et al. (2023) investigated if OPA differentiated other locomotive actions than walking, specifically crawling and flying - neither or which were included in our behavioral annotations - and found selective responses to videos taken only during walking. This could mean that neural representation of distinct action affordances may be restricted to a relatively narrow range of navigable actions, although methodological differences (the use of videos and analysis of univariate rather than multivariate responses) could also explain the discrepancy with our study.

In addition to OPA, we here also find representation of distinct action affordances in PPA, which has been suggest to represent a wide range of scene properties, including scene category (Walther et al., 2009), overall spatial layout (Epstein and Kanwisher, 1998; Kravitz et al., 2011; Park et al., 2011), but also texture

(Cant and Xu, 2012; Henriksson et al., 2019) and objects (Janzen and van Turennout, 2004; Harel et al., 2013; Marchette et al., 2015). In fact, the wide range of proposed roles of PPA in visual scene representation (see Epstein and Baker, 2019; Bartnik and Groen, 2023, for recent reviews) and the inherent correlations between different scene properties (Malcolm et al., 2016; Greene and Hansen, 2020) motivated us to collect multiple types of annotations for our scenes in order to directly compare these against our new operationalization of action affordance representation in scene-selective cortex. As shown in **Supplementary Figure 7B**, we find that both navigational affordances and global properties are consistently represented in PPA, while scene categories and objects are less consistently represented. Importantly, representations of global properties do not 'explain away' affordances in PPA (or OPA) (**Supplementary Figure 7B**), and results were robust across multiple ways of assessing affordance representation (in-scanner behavior, separate online experiment, and automatic labeling using classifiers). In contrast, the lack of representation of navigational action affordances in RSC is consistent with a substantial body of work showing that RSC is involved in navigating the broader, rather than the immediate environment (Epstein and Vass, 2014; Park and Chun, 2009), linking visual inputs to cognitive maps in hippocampal systems (Vann et al., 2009; Dilks et al., 2022).

In addition to expanding the operationalization of navigational affordances to different types of actions, we also assessed the task-dependency of these representations by measuring brain responses to our scenes under three different task instructions. Prior evidence on task effects on ventral visual cortex representations is mixed, with some studies finding modulations of representational patterns depending on in-scanner tasks in object-selective (Harel et al., 2012; Hebart et al., 2018) and scene-selective cortex (Lowe et al., 2016; Groen et al., 2018), while others observed representations of spatial layout in PPA (Kravitz et al., 2011) or pathways in OPA (Bonner and Epstein, 2017) even when not task-relevant. A direct comparison between a categorization and navigation task on neural representations of indoor walkways found task-specific effects in PPA, but not OPA (Persichetti and Dilks, 2018). Here, we find that both OPA and PPA show task-independent representations of our different action affordances, while object representations in these regions were more task-dependent. Task-independent representations of navigational affordances in scene-selective cortex are consistent with classic psychophysical findings by Greene and Oliva (2009) highlighting that perception of visual properties relevant to navigation is processed within very brief glimpses, as well as M/EEG studies showing early modulation of neural responses by scene properties (Groen et al., 2013; Harel et al., 2016), including affordance-related information (Harel et al., 2022; Djebbara et al., 2021).

DNNs have recently emerged as a popular tool used to study computations underlying human visual perception (Kietzmann et al., 2019; Storrs and Kriegeskorte, 2019), allowing for mapping of representations of specific DNN features to different brain regions (e.g., Khaligh-Razavi and Kriegeskorte, 2014; Dwivedi et al., 2021a). Scene-selective regions were initially associated with mid-level DNN layers in object- and scene-classification trained CNNs (Bonner and Epstein, 2018; Groen et al., 2018; King et al., 2019), but Dwivedi et al. (2021b) pointed out that DNNs trained on scene segmentation ("scene parsing") show a higher alignment with scene-selective regions; we observe a similar increase in our data (**Figure 4D**). However, while we find that DNNs exhibit some overlap with navigational affordance features, we find that they are consistently more closely aligned with object representations - the same distinction was observed across multiple DNN architectures trained with a variety of tasks, datasets and learning rules. Unlike Bonner and Epstein (2018) we therefore argue that affordance representations do not emerge 'automatically' from DNN features for the purpose of object- or scene classification. However, we also do not mean to suggest that affordances are not in some way linked to visual scene information: our results using DNN fine-tuning (**Figure 3G-H**) show that it is possible to achieve enhanced prediction of (behaviorally perceived) affordances using direct supervision with affordance labels (but we note that results were not robust across DNN architectures). Interestingly, we find the best alignment with humans when utilizing GPT-4-vision, a large language model that can analyze visual inputs. This could potentially be taken to indicate that vision DNNs extract visual features that are adequate for navigational affordance perception but still need higher-level knowledge in the form of rich linguistic representations to match human-level navigational affordance perception. While more research is needed to substantiate this claim, this notion is consistent with normative frameworks that state that visual features need to be combined with prior knowledge to enable affordance perception (Rietveld and Kiverstein, 2014), as well as recent findings showing that CNNs trained with richer semantic embedddings yield more accurate predictions of human visual cortex activity (Wang et al., 2023; Doerig et al., 2022).

Our key finding is that different navigational action affordances are distinctly represented in behavioral and neural measurements of human visual perception. We performed extensive comparisons to demonstrate this space can be reliably measured across different experimental paradigms and in fMRI. We also explored the relation of these representations with different types of visual information derived from either human behavioral annotations, automatically derived labels using classifcation, from feature activations in a large set of DNN models, and by training or prompting such models explicitly. Our results so far suggest human representations are not fully captured using any of these, thus offering a new human-alignment challenge for computational models of visual processing. Indeed, several key questions remain for future work.

First, we only sampled a subset of pretrained DNNs in the current study, and it is possible that a DNN we did not consider here will provide better predictions of human affordance perception. However, our goal here was to explore relative differences in alignment of affordances relative to other scene properties, not to provide a systematic comparison of how different DNN aspects (architectures, training datasets, task objectives) affect alignment with human perception (Conwell et al., 2022). However, such a comprehensive assessment is an important avenue for future research, and to facilitate future model comparisons, we make our stimuli, behavioral and neural data freely available, allowing testing of more models as well as comparison with other public datasets. Second, our analyses involving fine-tuning and end-to-end training DNNs on affordance labels show inconsistent results, with clear improvement for ResNets but not for ViTs (**Figure 3G-H**). This could be related to many differences between these models (pretraining dataset, differences in how 'data-hungry' the models are) but could also reflect imperfections in the supervision signal we used (Sun Attribute Database labels by Patterson and Hays (2012); Patterson et al. (2014)). These attribute labels were not collected for the purpose of studying navigational affordances per se, and notably do not include the label 'walking' which as shown in (**Figure 1A**) strongly drives human representational similarity of scene affordances. A bigger and better large scale-dataset is likely needed to more precisely assess how well human affordance perception can be mimicked using DNNs trained via direct supervision on affordance labels. Third, our comparison of task effects may be weakened by our experimental design in which participants were only probed in a small subset of trials in each fMRI run; although response probing was randomized across trials and therefore unpredictable for the participants (thereby encouraging them to perform the task on each trial), stronger task effects may be observed when participants are consistently probed on each trial.

Finally, we here tested navigational action affordance perception on static images, whereby participants indicated how they would move in a scene image using simple button presses. This is arguably still substantially different from the flow of input that people experience when actively navigating the world. Future work should ideally combine more immersive tasks (Gregorians and Spiers, 2022; Zhang and Gallant, 2020) with rich characterizations of the visual inputs using a range of computational models, in order to pinpoint the visuo-semantic transformations underlying human perception of distinct navigational affordances.

# Materials and Methods

## Stimuli

We created a novel stimulus set of 231 high-resolution color photographs (1024×1024 pixels) of visual environments. To prevent overlap with DNN training sets, images were collected from Flickr (freely available under either CC0 1.0, CC BY 2.0 and CC BY-SA 4.0 licenses) rather than any open large-scale image database commonly used to train DNNs (e.g., ImageNet, Places365, MS COCO). Images portrayed typical everyday environments devoid of prominent objects, humans, or animals. Each photo was taken from a human-scale, eye-level perspective, ensuring ecological validity (refer to **Fig. 1A** for exemplars). Diverging from previous research that predominantly utilized indoor imagery (e.g., Bonner and Epstein, 2017), our stimulus set contains images from three superordinate scene classes: indoor, outdoor-natural, and outdoor-man-made. This broad division is predicated on the premise that different environmental contexts afford

distinct functionalities (Zhou et al., 2017a). A t-distributed Stochastic Neighbor Embedding (t-SNE) visualization of our stimulus set, displayed on the attribute space of over 12,000 images from the SUN Attribute Database (Patterson and Hays, 2012) can be found in **Supplementary Fig. 9**, demonstrating that our stimulus set covers a much wider range of stimuli compared to previous fMRI studies. Images were matched in luminance to the mean across the stimulus set using the SHINE toolbox (Willenbockel et al., 2010).

## Behavioral Annotations

Humans naturally engage in a diverse range of actions within natural environments. However, the visual features within these environments often co-vary with each other, complicating the identification of the essential features for specific tasks (Malcolm et al., 2016; Greene and Hansen, 2020). To determine the relation between navigational affordances and various visual scene features, we ran an online study in which we collected behavioral navigational affordance annotations along with annotations of multiple readily identifiable scene properties: contained objects, materials, global properties, and scene category (see 'Behavioral task descriptions' below). Within the same experimental session, participants performed two additional tasks that are not further analyzed in the current paper.

### Subjects

The online experiment initially involved 167 participants (81 female) recruited from prolific.ac (Palan and Schitter, 2018). Of these, 15 participants did not complete the study, resulting in a final count of 152 participants. The participants had a median age of 25 years (range 18 - 74 years, SD = 14). We used selection criteria provided by Prolific to screen for reliable participants (Highest education level completed: High school diploma/A-levels; normal or corrected-to-normal vision: yes; First Language: English; Approval Rate: 90–100). On average, participants spent about 39 minutes completing the experiment (SD = 18 minutes). For their involvement in the study, each participant received a compensation of 7.5 pounds (GBP). Participants provided informed consent by selecting a checkbox on a digital consent form. The Ethical Committee of the Computer Science Department at the University of Amsterdam approved the experiment.

### Behavioral task descriptions

The experiment was designed using the Gorilla Experiment Builder (Anwyl-Irvine et al., 2020) and began with an introductory screen that informed participants about the study, instructions on how to switch into full-screen mode, and the tasks they would undertake. The experiment featured seven distinct tasks, of which the five following tasks are included in this study. In the *Take Action* task, participants selected from six options (walking, biking, driving, swimming, boating, climbing) to indicate how they would navigate through a scene. In the *Scene objects* task participants selected from six options (Building/Wall, Tree/Plant, Road/Street, Furniture, Body of water, Rocks/Stones) the presence of certain objects in the scene. This selection of objects was inspired by the objects that can be perceived in a single fixation (Fei-Fei et al., 2007). The *Scene Materials* task involved indicating the presence of ten different materials (Water, Wood (Not tree), Vegetation, Dirt/Soil, Sand, Stone/ Concrete, Snow/Ice, Pavement, Carpet, Metal), which were inspired by the material labels used in the SUN attributes database (Patterson et al., 2014). The *Scene Categories* task asked participants to select from eight types of scene categories (Room scene, Inside city scene, Hallway scene, Forest scene, Desert scene, Open country scene, Mountain scene, Coast/River scene) which ones applied to the scene. In all of these tasks, participants were allowed to select one or more annotations without any time limit. In the task *Scene Attributes*, participants used sliders on various bipolar dimensions (Open/Closed, Man-made/Natural, Near/Far, Navigable/Non-navigable) which were inspired by prior research by Kravitz et al. (2011) and Greene and Oliva (2009). The two tasks that were not further analyzed in this paper were *Path Drawing* (involving tracing potential pathways on an image) and *Scene Description* (asking participants to encapsulate each scene in a single word using a text box).

### Behavioral task procedure

Participants were divided into one of five groups to counterbalance the task sequence and ensure that each participant viewed each image only once in one of the tasks. In each task, members of a specific group viewed the same set of 33 images and performed the required annotations. This approach guaranteed that each image received ratings from a similar number of participants. The order of images within each task was randomized for each subject to prevent bias. After completing the seven tasks, participants filled out a feedback form with questions about their experience during the experiment. Each of the 231 real-world scene images was rated across all tasks by at least 20 participants and on average 21.7 (SD = 0.87) participants. The time taken for each annotation task was relatively consistent, with a median duration of 4.3 minutes (ranging from 3.32 to 7.95 minutes).

### Behavioral data analysis

For the behavioral annotations, the button presses for the different tasks were recorded. Representational dissimilarity matrices (RDMs) were computed using 1-Pearson correlation distances between the proportion of participants that annotated the presence or possibility of a given label (e.g., the proportion of participants that annotated the scene as affording walking). These features were first standardized across the images by removing the mean and scaling to unit variance. Similar results also can be obtained using when not scaling or using different distance metrics (see **Supplementary Fig. 2**). This resulted in five symmetric 231x231 RDMs, one for each annotation task.

## Automatic labeling

### Places365

To capture a richer space of possible scene categories we used a PyTorch implementation of a Wider-ResNet18 trained for scene classification on Places365 (Zhou et al., 2017a) to classify each of our images to one of the 365 scene classes. This model is part of the unified code provided by CSAILVision `https://github.com/CSAILVision/places365/blob/master/run_placesCNN_unified.py`. The resulting one-hot vector were compared between all pairs of images to compute RDMs using Euclidean distance metric. We used Euclidean distance here since each image was assigned a single class label, rendering the computation of 1-Person correlation inapplicable.

### SUN Attribute Database

To better understand how the choice and number of available labels affected the relation between navigational affordance representations and other scene properties, we utilized the Sun Attribute Database (Patterson et al., 2014). We again employed the PyTorch implementation of a Wider-ResNet18 trained for scene classification on Places365 (Zhou et al., 2017a). The provided script extracts features from the model's last convolutional layer (512 feature values) and applies a linear transformation to 102 attribute dimensions. The 102 scene attributes are grouped into 4 superordinate groups, namely functions (37 labels), materials (37 labels), spatial envelope (16 labels), and surface properties (12 labels). Notably, scene functions contain navigational affordance descriptions that partially overlap with the ones used in our behavioral experiment (e.g. cycling, swimming) but also more diverse activities such as socializing and gaming. Therefore we also created an additional subspace closer to our navigational affordance/action labels by only including the functions related to navigation (9 labels). We passed all our 231 images through the classifier and created RDMs based on a subset of the resulting attribute vector for each image standardized across the images and used 1-Pearson correlation as distance metric. This resulted in three additional RDMs representing navigational affordances, materials and global properties (spatial envelope).

### ADE20K

To automatically obtain a representational space based on the co-occurrences of objects in scenes, we utilized a scene segmentation network model consisting of a ResNet50dilated encoder and a PPM-deepsup

(PPM + deep supervision trick) decoder that classifies each pixel to belong to one of 150 object classes (Zhou et al., 2017b). We passed the images through the network model, and extracted the unique predicted object labels for each image as one-hot vectors, e.g. if an object such as a "wall" is present in the scene or not. Again the feature vectors were standardized across the images and 1-Pearson correlation was used to obtain the dissimilarity in the labeled object classes resulting in one RDM representing the object/co-occurrence representational space.

## fMRI experiment

### Stimuli

A subset of the stimuli from the online experiment were used for the fMRI study. To select a representative subset, we utilized Principal Component Analysis (PCA) on the behavioral action affordance annotations from the online experiment to extract the underlying dimensions. As walking was for most images the predominant label, we balanced the selection by labeling each image with the next most often selected navigational action, if it was labeled by over half of the participants. Subsequently, we selected a subset of 90 images using two selection criteria. Firstly, we ensured an equal distribution of images from each of the three environmental categories – indoor, outdoor-natural, and outdoor-man-made – allocating 30 images to each type of environment. Further, we balanced the number of images for each action affordance across the environments to the maximum extent possible, selecting images with high scores on the three principal components in the navigational affordance PCA (**Figure 2C**). The selection of the fMRI subset relative to the full image set is depicted in **Supplementary Figure 4**.

### Subjects

In total 20 healthy participants (13 female) with a median age of 24 years (range 18 - 33 years, SD = 3.52) completed the fMRI experiment, which consisted of four scanning sessions: an initial functional localizer session and three subsequent main experiment sessions in which participants were presented with scene stimuli under three different task instructions (see 'Experimental design' below). Three additional participants did not complete all sessions and are therefore not included in the data analysis. All participants had normal or corrected-to-normal vision, reported no neurological disorders nor used related medication. Before participating in each session participants filled out prescreening forms for MR safety and gave written informed consent. The Ethical Committee of the Psychology Department at the University of Amsterdam approved the experiment. Participants were compensated 25 euros per session.

### Experimental design

Each session consisted of 6 runs in each of which all 90 images were presented in random order in an event-related design, lasting in total 243.6 seconds each. Each scene run began with an introduction screen explaining the task to perform (navigational affordances, object, or fixation). Stimuli were presented at a resolution of 1017x1017 pixels such that they subtended $\sim$20 degrees of visual angle. Each image (independent of task) was overlaid with a fixation cross, randomly selected from one of six colors. Each image was presented for 1s followed by an ISI (randomized: 2.2, 3.8, or 5.4 seconds) with a gray screen and a white fixation cross. In each run, a response screen was displayed for a subset of 15 images prompting participants to indicate their selection with a button press. These response-prompted images were pseudo-randomly-selected such that one behavioral response for each image was collected across all 6 runs. The response screen was presented for 3.8 seconds displaying the six response options specific to each task. The three tasks were adapted from the previously conducted behavioral study. In the navigational affordance task, participants were instructed to press the same response options as in the behavioral experiment (Walking, Biking, Driving, Climbing, Swimming, Boating) during the response trials. In the object task, participants were instructed to choose from six response options (Building, Plant, Furniture, Road, Body of Water, Stones). In the fixation task, participants were instructed to select the color (Blue, Red, Orange, Purple, Yellow, Cyan) of the fixation cross. When pressing a button, the color of the text label would turn grey indicating a pressed button. As in the previous behavioral study, participants were allowed to select

multiple response options per trial in the navigational affordance and in the object task. While a response was only required for a subset of trials in each run, response screens appeared in random order ensuring participants needed to perform the task during each trial. The task session orders were counterbalanced across subjects, and all subjects completed all four sessions each scheduled on different days without any specific time frame between the sessions. Every experimental session took approximately 1.5 hours.

### MRI acquisition

All participants were scanned on a Philips Achieva 3T MRI scanner and a 32-channel SENSE head coil. Each scanning session started with a survey scan for spatial planning of subsequent scans, followed by a structural T1-weighted scan acquired using a T1TFTE pulse sequence (TR = 8.262 ms, FOV = 240 X 220 X 188 mm (AP x FH x RL) with a resolution of 1 x 1 x 1 mm voxel size). During localizer scans (session 1, see 'Functional localizers' below), functional images were acquired using a gradient echo, echo-planar pulse sequence (TR = 2,000 ms; TE = 27.63 ms; FA = 76.10; 36 sagittal slices with interleaved acquisition; 3 x 3 x 3 mm voxel size; 80 x 80 matrix; 240 x 240 x 118.5 mm FoV) covering the whole brain. A total of 140 volumes were recorded. In the same session, a set of population receptive field mapping runs were collected (not further analyzed in this study) using the same sequence, whereby a total of 150 volumes were recorded. In the following three main experimental task sessions, another T1-weighted scan was acquired in the beginning of the session followed by functional T2$^*$-weighted sequences using a multiband gradient echo EPI sequence (TR = 1,600 ms; TE = 30 ms; FA = 70; 56 sagittal slices with the interleaved acquisition; 2 x 2 x 2 mm voxel size; 112 x 112 matrix; 224 x 224 x 123 mm FoV) covering the whole brain. A total of 281 volumes were recorded per run.

### Functional localizers

To identify category-selective regions of interest (ROIs), participants completed 6 functional mapping runs, consisting of a block design, each lasting 6.9 minutes (208 TRs) (same stimuli and paradigm as in Silson et al. (2022, 2019)). Twenty color images (568 x 568 pixels) from one of six categories were displayed in each block: Faces, Objects, Body Parts, Buildings, Scenes, and Scrambled Images. Participants performed a one-back repetition detection task while stimuli were shown against a gray background for 300 ms with gaps of 500 ms in blocks of 16 seconds duration. Blocks were separated by 4-second short fixation breaks and 16-second long fixation breaks at the start and end of the runs. Category block orders were counterbalanced within and between runs to ensure equal occurrence and spacing in a mirror-symmetrical order.

### fMRI preprocessing

Preprocessing was performed using *fMRIPrep* 23.1.4 ((Esteban et al., 2019); (Esteban et al., 2018, RRID: SCR_016216) which is based on *Nipype* 1.8.6 ((Gorgolewski et al., 2011); (Gorgolewski et al., 2018, RRID: SCR_002502)). The 4 T1-weighted (T1w) images collected for each subject were corrected for intensity non-uniformity (INU) with `N4BiasFieldCorrection` (Tustison et al., 2010), distributed with ANTs (Avants et al., 2008, RRID:SCR_004757). The T1w-reference was then skull-stripped with a *Nipype* implementation of the `antsBrainExtraction.sh` workflow (from ANTs), using OASIS30ANTs as target template. An anatomical T1w-reference map was computed after registration of the 4 T1w images (after INU-correction) using *mri_robust_template* (FreeSurfer 7.3.2, Reuter et al., 2010). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with `antsRegistration` (ANTs), using the brain-extracted versions of both T1w reference and the T1w template. The following template was were selected for spatial normalization and accessed with *TemplateFlow* (23.0.0, Ciric et al., 2022): *ICBM 152 Nonlinear Asymmetrical template version 2009c* (Fonov et al., 2009, RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym).

### Functional data preprocessing

For each of the functional runs per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom method-

ology of *fMRIPrep*. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) were estimated before any spatiotemporal filtering using *mcflirt* (FSL, Jenkinson et al., 2002). A B0-nonuniformity map (or fieldmap) was estimated based on one echo-planar imaging (EPI) references with *topup* ((Andersson et al., 2003); FSL None). The estimated fieldmap was then aligned with rigid-registration to the target EPI (echo-planar imaging) reference run. The field coefficients were mapped onto the reference EPI using the transform. The BOLD reference was then co-registered to the T1w reference using *bbregister* (FreeSurfer) which implements boundary-based registration (Greve and Fischl, 2009). Co-registration was configured with six degrees of freedom. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. Gridded (volumetric) resamplings were performed using *antsApplyTransforms* (ANTs), configured with Lanczos interpolation (Lanczos, 1964). All resamplings were performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Many internal operations of *fMRIPrep* use *Nilearn* 0.10.1 (Abraham et al., 2014a, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see `https://fmriprep.readthedocs.io/en/latest/workflows.html`.

### Localizer fMRI analysis and ROI definitions

General linear models (GLM) were created for each individual subject and each individual run using the implementation provided by *Nilearn/Nistats* (`https://nilearn.github.io`; Abraham et al., 2014b, version 0.9.0). This GLM included 16s block regressors for each of the six categories (scenes, buildings, objects, faces, bodies, scrambled) which were convolved with a canonical hemodynamic response function (Glover, 1999) and fitted using AR1 autocorrelation correction and smoothed with a Gaussian 3mm FWHM filter. Subject-specific scene-selective regions were identified by first generating thresholded ($p < 0.05$, FDR) z-scored t-statistic activation maps to a number of contrasts that tested for higher activity to scenes than to objects, faces and scrambled image blocks (e.g. 'scenes' >'faces' & 'objects' and 'scenes' >'scrambled') which were subsequently intersected with an atlas of scene-selective parcels from Julian et al. (2012). In the last step, all contrast activation maps were unified to create subject-specific ROIs for PPA, OPA, RSC. The obtained ROIs contained on average 71% (PPA), 78% (OPA) and 66% (RSC) of the voxels in the atlas parcels, resulting in an average of 915 (PPA), 302 (OPA) and 1868 (RSC) voxel-sized ROIs.

### Experimental task sessions fMRI data analysis

The fMRI time-series data from the main experimental sessions were fitted with subject- and run-specific GLMs which contained one regressor for each of the 90 images convolved with a canonical hemodynamic response function (Glover, 1999). Additionally, 3 translational motion regressors, 3 rotational motion regressors and 3 low-frequency cosine DCT-basis regressors calculated by fMRIPrep were included as covariates. GLMs were fitted using an AR1 autocorrelation correction and smoothed with a Gaussian 3mm FWHM filter, after which one contrast per image was computed and resulting beta-estimates were masked by the subject-specific ROIs. The resulting voxel patterns were averaged across runs and used to compute subject-specific RDMs for each ROI using 1-Pearson correlation distances between each pair of images.

### In-scanner behavioral data analysis

For each participant, individual RDMs were calculated using 1-Pearson correlation distance and standardized features consisting of the annotations made by each participant indicating the possibility for the six navigational actions, the presence of objects, or the color of the fixation cross.

## DNN feature activations

To derive RDMs based on deep neural network features, we extracted activations from a number of pre-trained DNN models for our full image set using the Net2brain Python package (`https://github.com/`

cvai-roig-lab/Net2Brain; Bersch et al., 2022, version 0.1.0). We selected 13 commonly used DNNs (see **Supplementary Table 2** for an overview), divided in seven different groups (see legend of **Figure 4A**). We extracted feature activations for the layers pre-selected by Net2Brain and computed RDMs of the standardized features per layer (removing the mean and scaling to unit variance) using 1-Pearson correlation distance metric.

## End-to-end training and fine-tuning on affordances

Given the limited alignment of pre-trained DNNs with action affordances, we also attempted to directly train models on scene affordances and fine-tune pre-trained models specifically for this task. We used the functions subset of the SUN Attribute Dataset (Patterson et al., 2014), which provides labels for different scene attributes. Each label is annotated with a number of votes ranging from zero to three: a score of zero indicates the absence of an affordance, while a score of three signals its guaranteed presence. Here, an affordance was considered present in an image if it had received at least two votes by the annotators, and absent if it had received zero votes. Affordances that received a single vote were excluded due to their ambiguity. We also omitted affordances with fewer than 300 instances from our training set, as this was considered insufficient for training deep convolutional neural networks effectively. This process resulted in a dataset comprising of 8,147 images and 24 of 37 affordance labels (see **Supplementary Table 4 for overview**. For the models, we utilized a ResNet50 and the Vision Transformer (ViT) Base Patch 16, both implemented in PyTorch 1.11.0. The ResNet50 was imported from the Net2Brain Toolbox (Bersch et al., 2022) and the ViT Base Patch 16 from the "PyTorch Image Models" repository by (Steiner et al., 2021). We adjusted both models to feature a final layer with 24 output units, each representing an affordance label, and included a sigmoid activation function and an Adam optimizer. We employed Binary Cross Entropy (BCE) Loss since each label was treated as an independent binary classification task (presence vs absence). While for end-to-end training the models learned entirely from the training data, for the fine-tuning approach we loaded pre-trained weights and only updated the final fully connected layer while keeping all other layers frozen. We applied 5-fold cross-validation to ensure the robustness of our optimization and utilized early stopping based on validation loss to prevent overfitting. Our training protocol used a batch size of 64, and we optimized the models using a two-phase hyperparameter tuning process: we first employed a random search to quickly identify effective parameter ranges, followed by a grid search to fine-tune these parameters and select the best-performing models. **Supplementary Table 3** shows some of the parameter configurations of the trained and fine-tuned models.

## GPT-4 annotations

We presented each of our 231 images to GPT-4 OpenAI et al. (2023) to generate behavioral responses similar to those generated by the human participants for navigational affordances and objects, using a variety of prompts (see **Supplementary Table 5**). These tasks were to pick a single affordance/object, multiple affordances/objects and create a bag of words representation of all possible affordances and objects in the image. RDMs were created based on a one-hot encoding of the resulting responses using Euclidean distance metric for the single annotation behavior and 1-Pearson correlation distance for the multiple and BoW annotations.

## Representational Similarity Analysis

RDMs were compared to one another by computing Spearman's $\rho$ as proposed by the RSA toolbox (Kriegeskorte, 2008; Nili et al., 2014) (lower triangle, excluding the diagonal; Ritchie et al. (2017)). This was done on the subject average in-scanner behavior and single subject ROI RDMs. We compute layer-wise comparisons of DNN feature activation RDMs with single-subject ROI RDMs and single-subject in-scanner behavior RDMs always reporting the highest correlating layer (see **Supplementary Table 1** and **Supplementary Table 2** for all extracted layer)s. One-sample t-tests were used to test whether the average correlation was significantly different from zero, and paired sample t-tests were used to test for significant differences between correlations. The obtained p-values were corrected for multiple comparisons using Bonferroni correction (2

for fMRI: 0.05/2 = 0.025, 3 for fMRI task-effect: 0.05/3 = 0.017, 13 DNN models: 0.05/13 = 0.003). Pairwise comparisons were also corrected for multiple comparisons using Bonferroni correction when comparing different tasks and DNN models to one another. When comparing DNN features with behavioral annotation RDMs obtained from the online behavioral experiment and the automatic labeling, we used (n = 1000) bootstrapping (sampling rows/columns with replacement) to obtain the 95 % confidence intervals. The differences between the bootstrapped distributions were statistically compared using two-tailed sign test and FDR-corrected for multiple comparisons.

### Code availability

Stimuli, behavioral annotations, MRI-data and extracted DNN feature activations will be made available on the OSF `https://osf.io/v3rcq/`. Python scripts containing preprocessing and analysis code to reproduce the results will be available at `https://github.com/cgbartnik`.

## Acknowledgements

This work was supported by an VENI grant (VI.Veni.194030) from the Netherlands Organisation for Scientific Research (NWO) to IIAG. We thank Lukas Muttenthaler and Steven Scholte for helpful feedback.

## References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014a). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8.

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014b). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8.

Aminoff, E. M., Kveraga, K., and Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends in Cognitive Sciences*, 17(8):379–390.

Andersson, J. L., Skare, S., and Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage*, 20(2):870–888.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52(1):388–407.

Avants, B., Epstein, C., Grossman, M., and Gee, J. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41.

Bartnik, C. G. and Groen, I. I. A. (2023). Visual Perception in the Human Brain: How the Brain Perceives and Understands Real-World Scenes. In *Oxford Research Encyclopedia of Neuroscience*. Oxford University Press.

Bersch, D., Dwivedi, K., Vilas, M., Cichy, R. M., and Roig, G. (2022). Net2Brain: A Toolbox to compare artificial vision models with human brain responses.

Biederman, I., Mezzanotte, R. J., and Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177.

Bonner, M. F. and Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, 114(18):4793–4798.

Bonner, M. F. and Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLOS Computational Biology*, 14(4):e1006111.

Cant, J. S. and Xu, Y. (2012). Object ensemble processing in human anterior-medial ventral visual cortex. *Journal of Neuroscience*, 32(22):7685–7700.

Ciric, R., Thompson, W. H., Lorenz, R., Goncalves, M., MacNicol, E., Markiewicz, C. J., Halchenko, Y. O., Ghosh, S. S., Gorgolewski, K. J., Poldrack, R. A., and Esteban, O. (2022). TemplateFlow: FAIR-sharing of multi-scale, multi-species brain models. *Nature Methods*, 19:1568–1571.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., and Konkle, T. (2022). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, pages 2022–03.

Dilks, D. D., Julian, J. B., Kubilius, J., Spelke, E. S., and Kanwisher, N. (2011). Mirror-image sensitivity and invariance in object and scene processing pathways. *Journal of Neuroscience*, 31(31):11305–11312.

Dilks, D. D., Kamps, F. S., and Persichetti, A. S. (2022). Three cortical scene systems and their development. *Trends in cognitive sciences*, 26(2):117–127.

Djebbara, Z., Fich, L. B., and Gramann, K. (2021). The brain dynamics of architectural affordances during transition. *Scientific Reports*, 11(1):2796.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., and Charest, I. (2022). Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*.

Dwivedi, K., Bonner, M. F., Cichy, R. M., and Roig, G. (2021a). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLOS Computational Biology*, 17(8):e1009267.

Dwivedi, K., Cichy, R. M., and Roig, G. (2021b). Unraveling representations in scene-selective brain regions using scene-parsing deep neural networks. *Journal of Cognitive Neuroscience*, 33(10):2032–2043.

Epstein, R. A. and Baker, C. I. (2019). Scene perception in the human brain. *Annual Review of Vision Science*, 5(1):373–397.

Epstein, R. A. and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676):598–601.

Epstein, R. A. and Vass, L. K. (2014). Neural systems for landmark-based wayfinding in humans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1635):20120533.

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., Isik, A. I., Erramuzpe, A., Kent, James D. andGoncalves, M., DuPre, E., Sitek, K. R., Gomez, D. E. P., Lurie, D. J., Ye, Z., Poldrack, R. A., and Gorgolewski, K. J. (2018). fmriprep. *Software*.

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe Aliaga, A., Kent, J., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S., Wright, J., Durnez, J., Poldrack, R., and Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16:111–116.

Fei-Fei, L., Iyer, A., Koch, C., and Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1):10.

Fonov, V., Evans, A., McKinstry, R., Almli, C., and Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47, Supplement 1:S102.

Gibson, J. J. (1977). The theory of affordances. *Hilldale, USA*, 1(2):67–82.

Glover, G. H. (1999). Deconvolution of Impulse Response in Event-Related BOLD fMRI1. *NeuroImage*, 9(4):416–429.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., and Ghosh, S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, 5:13.

Gorgolewski, K. J., Esteban, O., Markiewicz, C. J., Ziegler, E., Ellis, D. G., Notter, M. P., Jarecka, D., Johnson, H., Burns, C., Manhães-Savio, A., Hamalainen, C., Yvernault, B., Salo, T., Jordan, K., Goncalves, M., Waskom, M., Clark, D., Wong, J., Loney, F., Modat, M., Dewey, B. E., Madison, C., Visconti di Oleggio Castello, M., Clark, M. G., Dayan, M., Clark, D., Keshavan, A., Pinsard, B., Gramfort, A., Berleant, S., Nielson, D. M., Bougacha, S., Varoquaux, G., Cipollini, B., Markello, R., Rokem, A., Moloney, B., Halchenko, Y. O., Wassermann, D., Hanke, M., Horea, C., Kaczmarzyk, J., de Hollander, G., DuPre, E., Gillman, A., Mordom, D., Buchanan, C., Tungaraza, R., Pauli, W. M., Iqbal, S., Sikka, S., Mancini, M., Schwartz, Y., Malone, I. B., Dubois, M., Frohlich, C., Welch, D., Forbes, J., Kent, J., Watanabe, A., Cumba, C., Huntenburg, J. M., Kastman, E., Nichols, B. N., Eshaghi, A., Ginsburg, D., Schaefer, A., Acland, B., Giavasis, S., Kleesiek, J., Erickson, D., Küttner, R., Haselgrove, C., Correa, C., Ghayoor, A., Liem, F., Millman, J., Haehn, D., Lai, J., Zhou, D., Blair, R., Glatard, T., Renfro, M., Liu, S., Kahn, A. E., Pérez-García, F., Triplett, W., Lampe, L., Stadler, J., Kong, X.-Z., Hallquist, M., Chetverikov, A., Salvatore, J., Park, A., Poldrack, R., Craddock, R. C., Inati, S., Hinds, O., Cooper, G., Perkins, L. N., Marina, A., Mattfeld, A., Noel, M., Snoek, L., Matsubara, K., Cheung, B., Rothmei, S., Urchs, S., Durnez, J., Mertz, F., Geisler, D., Floren, A., Gerhard, S., Sharp, P., Molina-Romero, M., Weinstein, A., Broderick, W., Saase, V., Andberg, S. K., Harms, R., Schlamp, K., Arias, J., Papadopoulos Orfanos, D., Tarbert, C., Tambini, A., De La Vega, A., Nickson, T., Brett, M., Falkiewicz, M., Podranski, K., Linkersdörfer, J., Flandin, G., Ort, E., Shachnev, D., McNamee, D., Davison, A., Varada, J., Schwabacher, I., Pellman, J., Perez-Guevara, M., Khanuja, R., Pannetier, N., McDermottroe, C., and Ghosh, S. (2018). Nipype. *Software*.

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., and Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, 145(1):82–94.

Greene, M. R. and Hansen, B. C. (2020). Disentangling the independent contributions of visual and conceptual features to the spatiotemporal dynamics of scene categorization. *The Journal of Neuroscience*, 40(27):5283–5299.

Greene, M. R. and Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, 20(4):464–472.

Gregorians, L. and Spiers, H. J. (2022). Affordances for Spatial Navigation. In Djebbara, Z., editor, *Affordances in Everyday Life*, pages 99–112. Springer International Publishing, Cham.

Greve, D. N. and Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72.

Groen, I. I. A., Ghebreab, S., Prins, H., Lamme, V. A. F., and Scholte, H. S. (2013). From image statistics to scene gist: Evoked neural activity reveals transition from low-level natural image structure to scene category. *Journal of Neuroscience*, 33(48):18814–18824.

Groen, I. I. A., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., and Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife*, 7:e32962.

Harel, A., Groen, I. I. A., Kravitz, D. J., Deouell, L. Y., and Baker, C. I. (2016). The temporal dynamics of scene processing: A multifaceted EEG investigation. *eNeuro*, 3(5):ENEURO.0139–16.2016.

Harel, A., Kravitz, D., and Baker, C. (2012). Task-dependent representations of visual objects. *Journal of Vision*, 12(9):273–273.

Harel, A., Kravitz, D. J., and Baker, C. I. (2013). Deconstructing visual scenes in cortex: Gradients of object and spatial layout information. *Cerebral Cortex*, 23(4):947–957.

Harel, A., Nador, J. D., Bonner, M. F., and Epstein, R. A. (2022). Early electrophysiological markers of navigational affordances in scenes. *Journal of Cognitive Neuroscience*, 34(3):397–410.

Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., and Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *eLife*, 7:e32816. Publisher: eLife Sciences Publications, Ltd.

Henriksson, L., Mur, M., and Kriegeskorte, N. (2019). Rapid invariant encoding of scene layout in human OPA. *Neuron*, 103(1):161–171.e3.

Janzen, G. and van Turennout, M. (2004). Selective neural representation of objects relevant for navigation. *Nature Neuroscience*, 7(6):673–677.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841.

Jones, C. M., Byland, J., and Dilks, D. D. (2023). The occipital place area represents visual information about walking, not crawling. *Cerebral Cortex*, 33(12):7500–7505.

Julian, J., Fedorenko, E., Webster, J., and Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage*, 60(4):2357–2364.

Julian, J., Ryan, J., Hamilton, R., and Epstein, R. (2016). The occipital place area is causally involved in representing environmental boundaries during navigation. *Current Biology*, 26(8):1104–1109.

Kamps, F. S., Julian, J. B., Kubilius, J., Kanwisher, N., and Dilks, D. D. (2016a). The occipital place area represents the local elements of scenes. *NeuroImage*, 132:417–424.

Kamps, F. S., Lall, V., and Dilks, D. D. (2016b). The occipital place area represents first-person perspective motion information through scenes. *Cortex*, 83:17–26.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, 10(11):e1003915.

Kietzmann, T. C., McClure, P., and Kriegeskorte, N. (2019). Deep Neural Networks in Computational Neuroscience. In *Oxford Research Encyclopedia of Neuroscience*. Oxford University Press.

King, M. L., Groen, I. I., Steel, A., Kravitz, D. J., and Baker, C. I. (2019). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, 197:368–382.

Kravitz, D. J., Peng, C. S., and Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: It's the spaces more than the places. *Journal of Neuroscience*, 31(20):7322–7333.

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.

Lanczos, C. (1964). Evaluation of noisy data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, 1(1):76–85.

Lowe, M. X., Gallivan, J. P., Ferber, S., and Cant, J. S. (2016). Feature diagnosticity and task context shape activity in human scene-selective cortex. *NeuroImage*, 125:681–692.

Malcolm, G. L., Groen, I. I. A., and Baker, C. I. (2016). Making sense of real-world scenes. *Trends in Cognitive Sciences*, 20(11):843–856.

Marchette, S. A., Vass, L. K., Ryan, J., and Epstein, R. A. (2015). Outside looking in: Landmark generalization in the human navigational system. *Journal of Neuroscience*, 35(44):14896–14908.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, 10(4):e1003553.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, , Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, , Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Peres, F. d. A. B., Petrov, M., Pinto, H. P. d. O., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2023). GPT-4 Technical Report.

Osiurak, F., Rossetti, Y., and Badets, A. (2017). What is an affordance? 40 years later. *Neuroscience & Biobehavioral Reviews*, 77:403–417.

Palan, S. and Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Park, J. and Park, S. (2020). Coding of navigational distance and functional constraint of boundaries in the human scene-selective cortex. *The Journal of Neuroscience*, 40(18):3621–3630.

Park, S., Brady, T. F., Greene, M. R., and Oliva, A. (2011). Disentangling scene content from spatial boundary: Complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *Journal of Neuroscience*, 31(4):1333–1340.

Park, S. and Chun, M. M. (2009). Different roles of the parahippocampal place area (PPA) and retrosplenial cortex (RSC) in panoramic scene perception. *NeuroImage*, 47(4):1747–1756.

Patterson, G. and Hays, J. (2012). SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758, Providence, RI. IEEE.

Patterson, G., Xu, C., Su, H., and Hays, J. (2014). The SUN Attribute Database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81.

Persichetti, A. S. and Dilks, D. D. (2016). Perceived egocentric distance sensitivity and invariance across scene-selective cortex. *Cortex*, 77:155–163.

Persichetti, A. S. and Dilks, D. D. (2018). Dissociable neural systems for recognizing places and navigating through them. *The Journal of Neuroscience*, 38(48):10295–10304.

Reuter, M., Rosas, H. D., and Fischl, B. (2010). Highly accurate inverse consistent registration: A robust approach. *NeuroImage*, 53(4):1181–1196.

Rietveld, E. and Kiverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology*, 26(4):325–352.

Ritchie, J. B., Bracci, S., and Op De Beeck, H. (2017). Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. *NeuroImage*, 148:197–200.

Silson, E. H., Groen, I. I. A., and Baker, C. I. (2022). Direct comparison of contralateral bias and face/scene selectivity in human occipitotemporal cortex. *Brain Structure and Function*, 227(4):1405–1421.

Silson, E. H., Steel, A., Kidder, A., Gilmore, A. W., and Baker, C. I. (2019). Distinct subdivisions of human medial parietal cortex support recollection of people and places. *eLife*, 8:e47391.

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. (2021). How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers.

Storrs, K. R. and Kriegeskorte, N. (2019). Deep Learning for Cognitive Neuroscience. Publisher: arXiv Version Number: 1.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Grant, E., Groen, I., Achterberg, J., Tenenbaum, J. B., Collins, K. M., Hermann, K. L., Oktar, K., Greff, K., Hebart, M. N., Jacoby, N., Zhang, Q., Marjieh, R., Geirhos, R., Chen, S., Kornblith, S., Rane, S., Konkle, T., O'Connell, T. P., Unterthiner, T., Lampinen, A. K., Müller, K.-R., Toneva, M., and Griffiths, T. L. (2023). Getting aligned on representational alignment.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320.

Vann, S. D., Aggleton, J. P., and Maguire, E. A. (2009). What does the retrosplenial cortex do? *Nature Reviews Neuroscience*, 10(11):792–802.

Walther, D. B., Caddigan, E., Fei-Fei, L., and Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of Neuroscience*, 29(34):10573–10581.

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., and Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., and Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods*, 42(3):671–684.

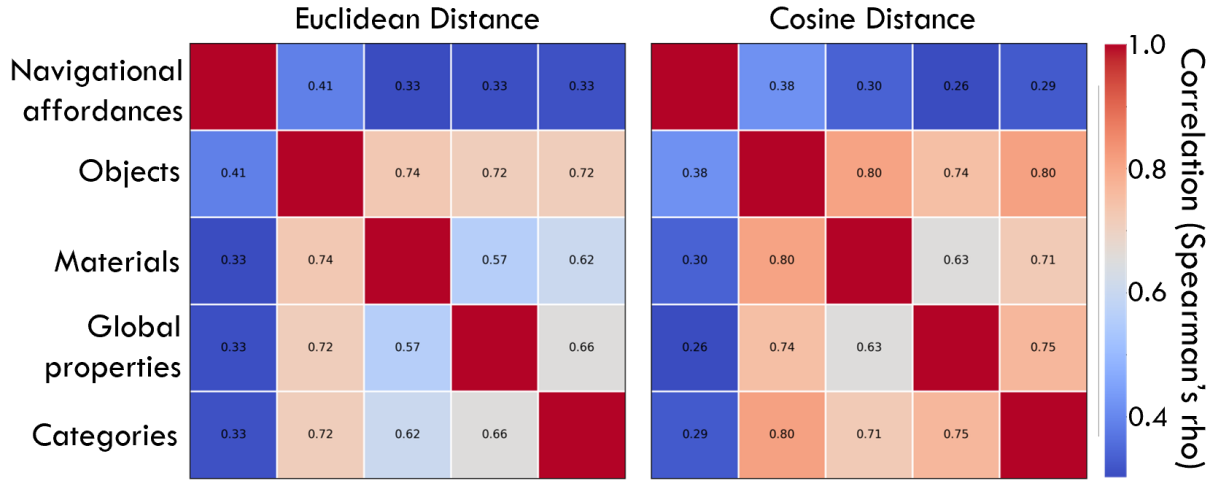
Zhang, T. and Gallant, J. L. (2020). A naturalistic navigation task reveals rich distributed representations of information across the human cerebral cortex. *Journal of Vision*, 20(11):462.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017a). Places: A 10 Million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464.

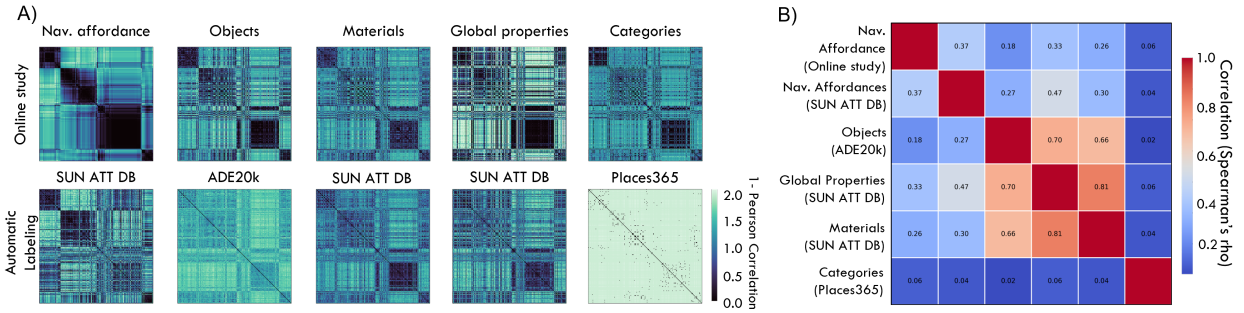
Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. (2017b). Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130.

# Supplementary Information

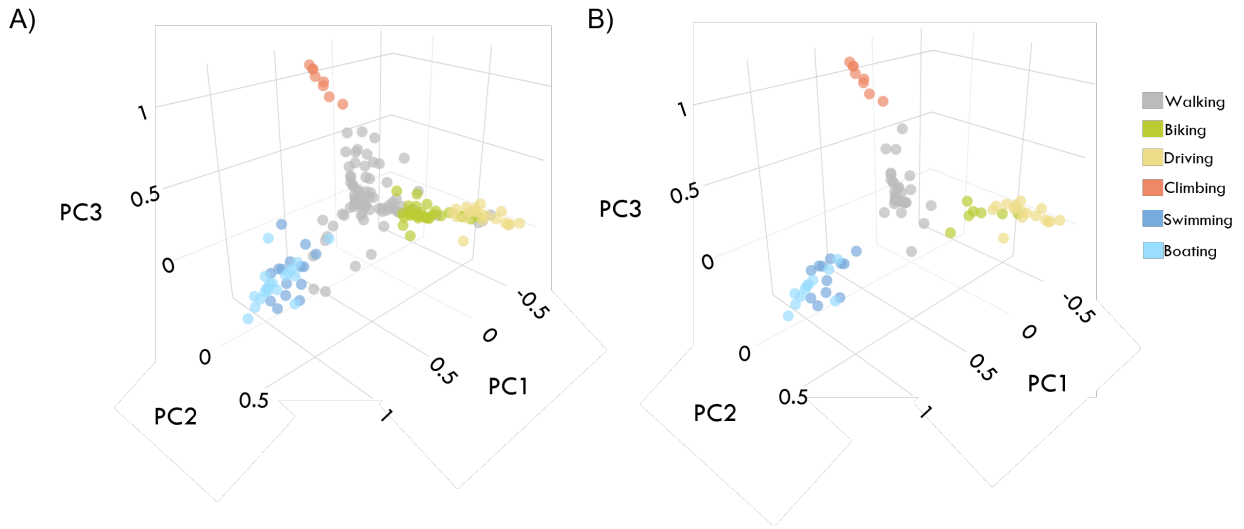
**Supplementary Figure 1: Two-dimensional multidimensional scaling of navigational affordance representational space.** This figure displays the same spatial arrangement of our complete 231 stimulus set as displayed in **Figure 2** but with stimulus thumbnails each framed by a colored border. The color coding corresponds to the most frequently chosen navigation action associated with each image (see Materials and Methods for details)



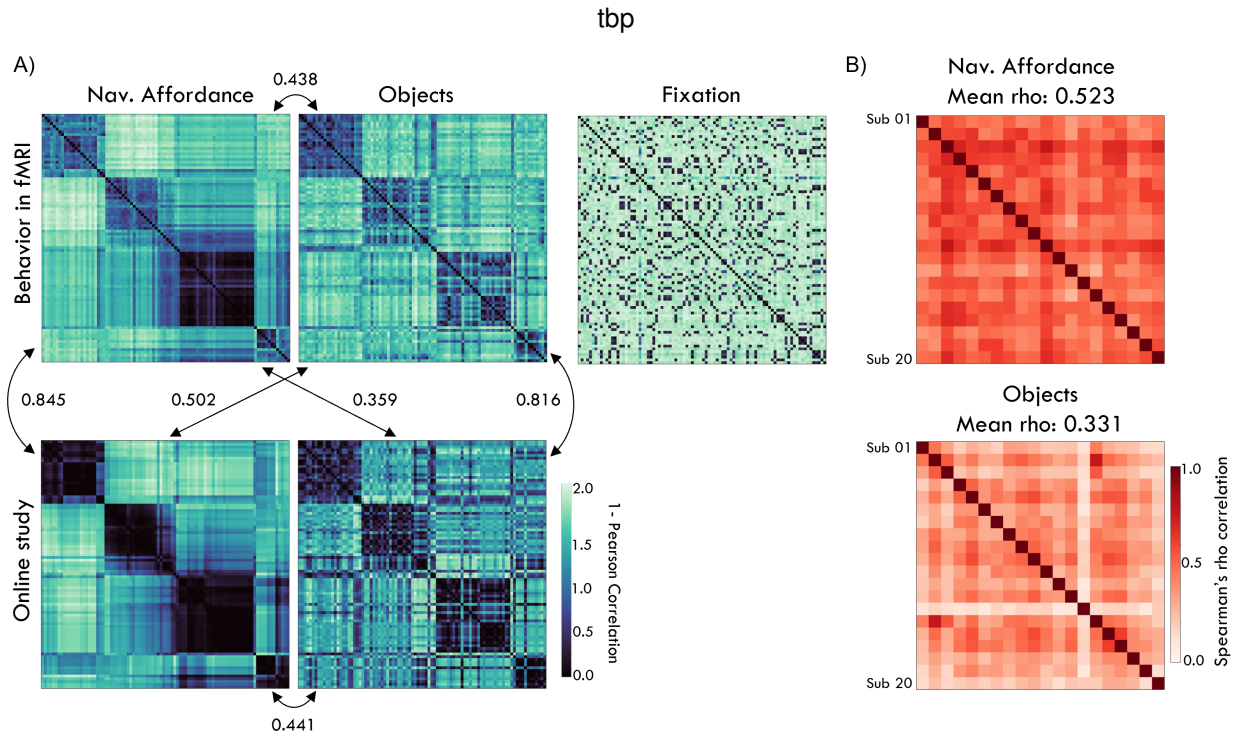

**Supplementary Figure 2: Correlation between navigational affordances and other scene properties using different distance measures to compute RDMs.** We consistently find that navigational affordances are a unique space and that the other visual scene properties highly correlate with each other, also when behavioral RDMs are constructed using Euclidean or cosine distance.



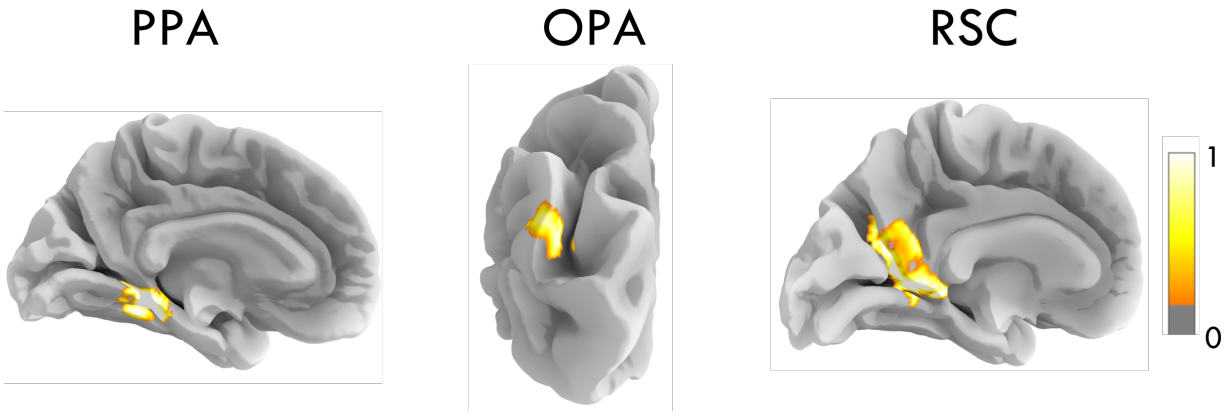

**Supplementary Figure 3: Comparison of RDMs for all scene properties obtained using behavioral annotations or automated labeling methods. (A)** Upper row, RDMs of the representational space of various scene properties for which we collected annotations in an online experiment. Bottom row, RDMs of similar spaces that were created using automatic labeling by utilizing large-scale image datasets and computational models. These were used to evaluate if the interrelation of the RDMs that we found in our behavioral experiment were a product of our limited selection of possible labeling options or if this is also true in more unrestricted cases. **(B)** Correlations between RDMs for for all spaces computed from automatic labels, and our behavioral annotations of navigational affordances. We find a similar pattern as for annotations obtained through human behavior: navigational affordances correlate relatively low with other scene properties which highly correlate with each other. Note that category labels here are an exception, correlating poorly with all other spaces (including navigational affordances), unlike in our behavioral experiment. This is likely due to the high sparsity of the Place365-derived category RDM.



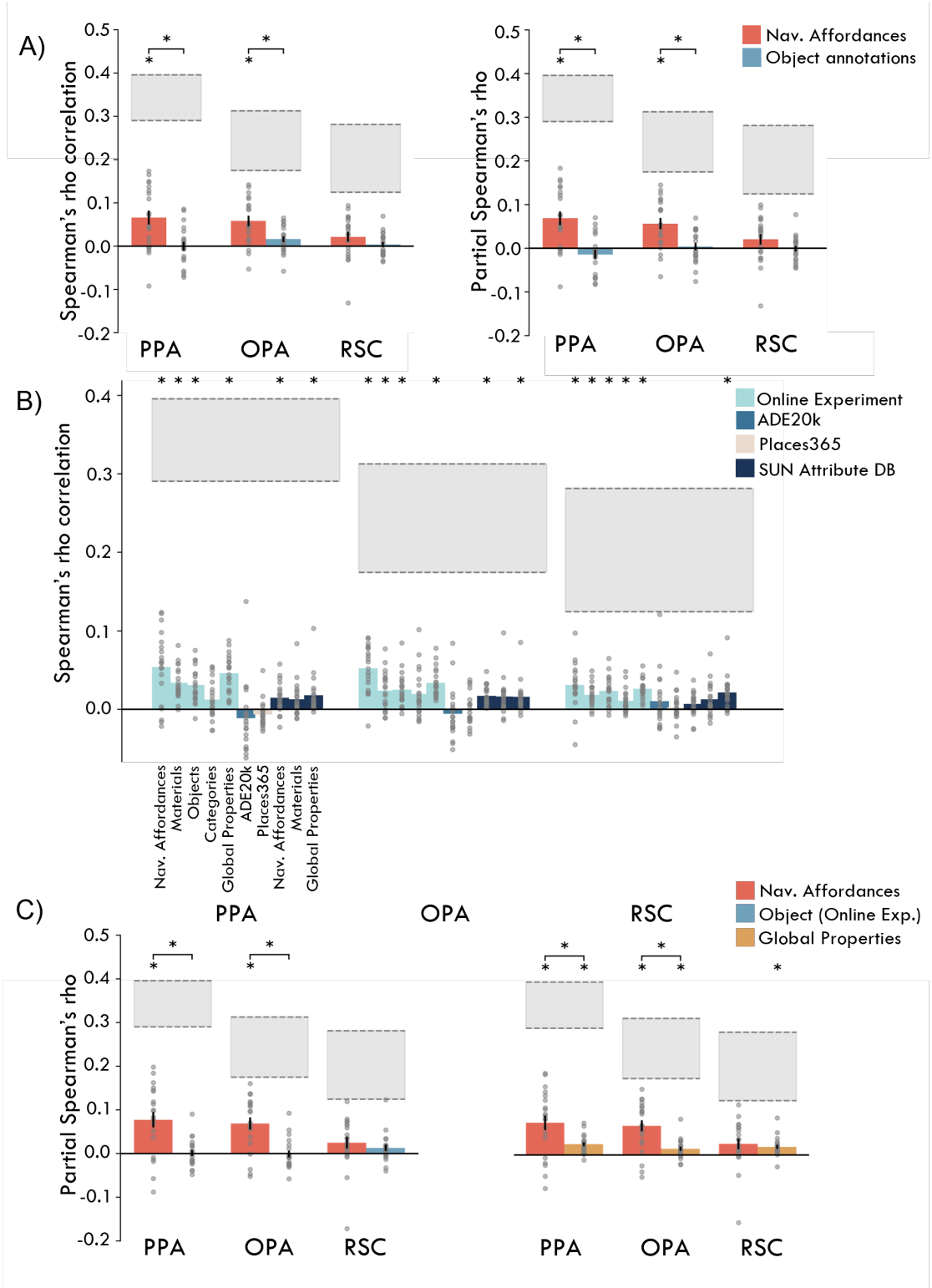

**Supplementary Figure 4: Using Principal Component Analysis for strategically sampling a stimulus subset. (A)** PCA of the behavioral navigational affordance RDM from the online study including all 231 images. The color coding corresponds to the most frequently chosen navigation action associated with each image (see Materials and Methods for details) **(B)** Highlighting the strategically chosen subset of 90 images employed in the fMRI study in the same PCA space. This subset is composed of an equal number of indoor, outdoor-natural, and outdoor-manmade scenes (30 each), selected to capture the the three-dimensional navigational affordance structure.



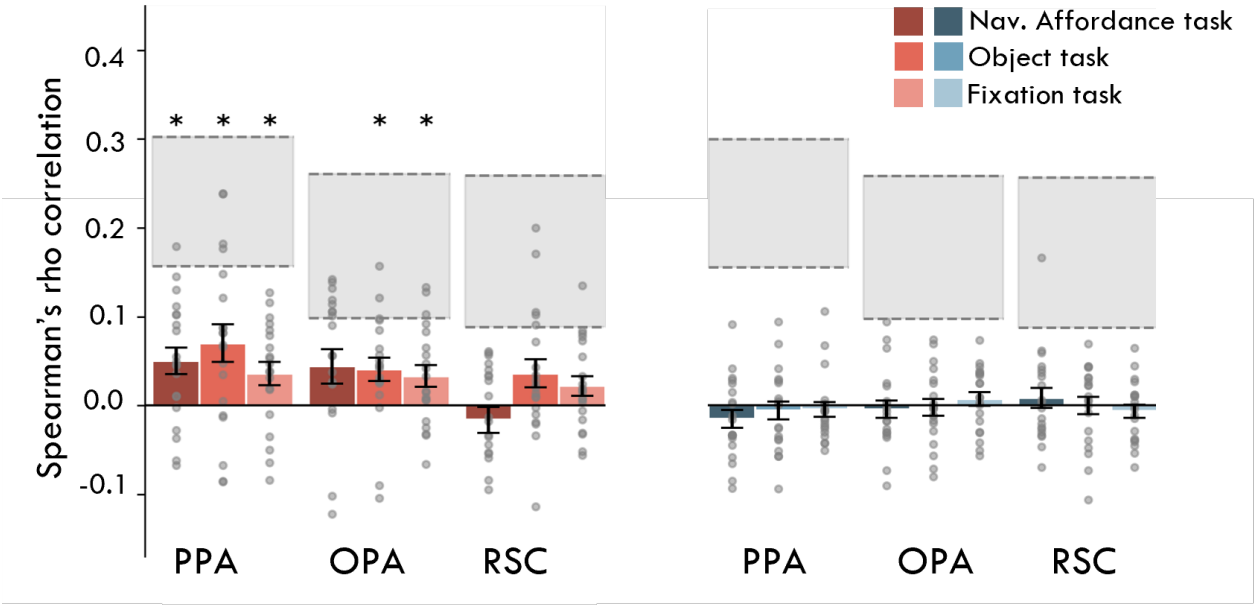

**Supplementary Figure 5: Comparison of behavioral data obtained in the online study versus fMRI experiment. (A)** The upper row contains the average RDMs computed based on three behavioral annotation tasks (navigational affordances, objects and color of fixation cross) performed by the participants in the fMRI scanner. The bottom row shows the corresponding RDMs from the online study. These results show that the task replicates well in the scanner, showing a high correlation between behavioral annotations from the same task in the scanner and in the online experiment, while the correlations between tasks are similar between online and fMRI scanner participants. Behavior for the orthogonal fixation task (top right) is completely unrelated to behavioral annotations of either navigational affordance or objects, as intended. **(B)** Intercorrelation of RDMs for each behavioral tasks across the fMRI participants. Participant performance was more consistent during the navigational affordance task than during the object task.



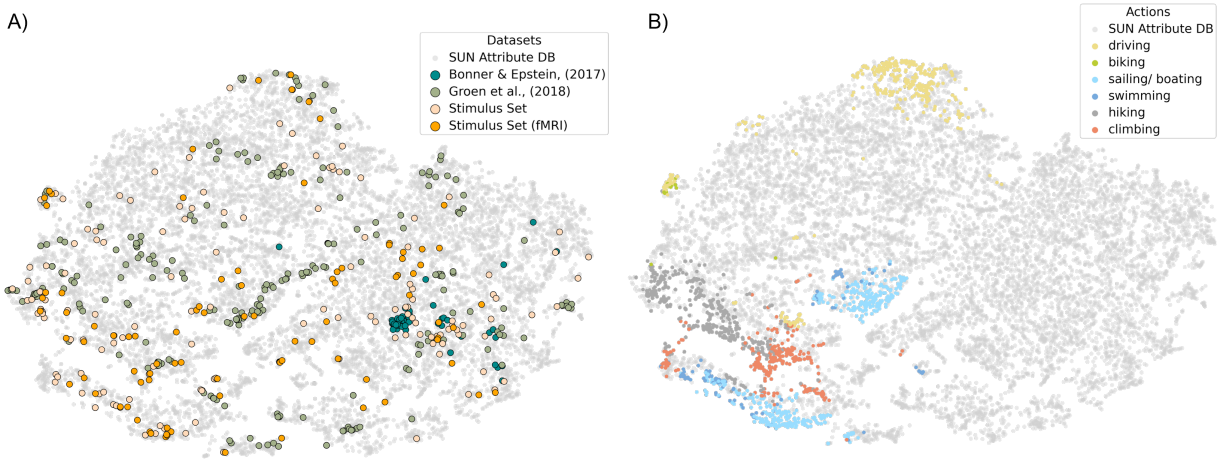

**Supplementary Figure 6: Scene-selective ROIs.** Scene-selective ROI (Parahippocampal Place Area, PPA; Occipital Place Area, OPA; and Retrosplenial Cortex, RSC) masks of subject 004 in MNI-152 template space (see Materials and Methods for details).



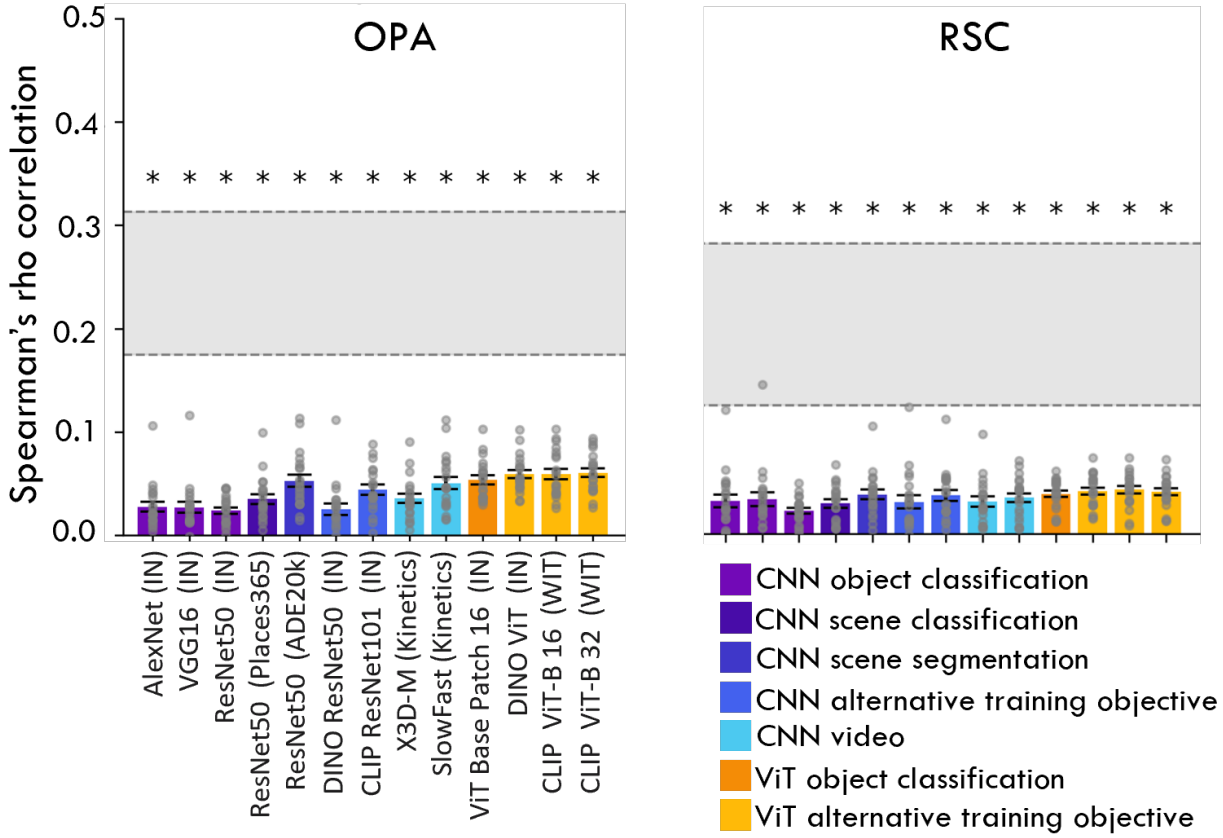

**Supplementary Figure 7: Navigational affordance representations in PPA and OPA for individual participants, online behavior and automatic labels.** **(A)** Full (left) and partial (right) correlations of the fMRI RDMs with in-scanner individual subject behavioral RDMs. **(B)** Correlations of the fMRI RDMs with the online behavioral and the automatic labeled spaces. **(C)** Partial correlations analysis with online behavioral RDMs show unique representation of navigational affordances versus objects (left) and global properties (right). Error bars indicate the standard error of the mean (SEM) across participants. Asterisks (*), indicating one-sample t-test against zero $p<0.05$ corrected for multiple comparisons using Bonferroni correction.




**Supplementary Figure 8: Task-specific partial correlations of fMRI RDMs with behavioral RDMs of affordance and object annotations.** Partial correlations of RDMs derived from fMRI response patterns in scene-selective brain regions with navigational affordances (red) and object annotations (blue) RDMs during a navigational affordance labeling task (darker colors), and object labeling task (middle colors) and an orthongal fixation task (light colors). Bars represent averages for individual participants (gray dots). Shaded areas delineate upper and lower noise ceilings, reflecting the range of RDM similarity between individual participants and the group mean. Error bars indicate the standard error of the mean (SEM) across participants. Asterisks (*), indicating one-sample t-test against zero $p<0.05$ corrected for multiple comparisons using Bonferroni correction.




**Supplementary Figure 9: Visualization of scene representational space and associated navigational affordances (A)** t-SNE visualization of the SUN Attribute Database representational space containing over 12.000 images with various datasets from prior research (Bonner and Epstein, 2017) and (Groen et al., 2018) and our image set (whole and fMRI subset) embedded in the representational space. **(B)** t-SNE visualization of the SUN Attribute Database with coloring based if a scene was labeled to afford a specific navigational function. These judgments are based of the labeling in the SUN Attribute Database.




**Supplementary Figure 10: Average correlation of each DNN's best correlating layer with representational similarity in OPA and RSC.** Bars represent average correlations across participants and tasks for the best correlating layer in each model and gray dots indicate individual participants. Significant t-tests of the average correlation against zero are marked by asterisks (*), indicating $p<0.05$ corrected for multiple comparisons using Bonferroni correction. The shaded area delineates the upper and lower noise ceiling and error bars indicate the standard error of the mean (SEM) across participants.

**Supplementary Table 1:** Selection of commonly used DNNs reporting the number of extracted layers and the highest correlating layer with in-scanner behavior and scene-selective ROIs. See **Supplementary Table 2** for an overview of all extracted layers in each model.

| Model | Dataset | Extracted layers (Number) | Nav. Affor-dances | Objects | PPA | OPA | RSC |
|---|---|---|---|---|---|---|---|
| AlexNet | ImageNet | 6 | features6 | features10 | features6 | features6 | features6 |
| VGG16 | ImageNet | 14 | features0 | features26 | features0 | features0 | features0 |
| ResNet50 | ImageNet | 5 | layer2 | layer3 | layer2 | layer4 | layer2 |
| ResNet50 | Places365 | 6 | model9 | model7 | model9 | model9 | model7 |
| ResNet50 | Ade20K | 6 | decoder ppm last conv2 | decoder conv last1 | decoder ppm last conv2 | decoder ppm last conv2 | decoder ppm last conv2 |
| DINO ResNet50 | ImageNet | 5 | layer1 | fc | layer1 | layer1 | layer1 |
| CLIP ResNet101 | ImageNet | 5 | layer3 | layer4 | layer3 | layer3 | layer3 |
| X3D-m | Kinetics | 11 | blocks.4 | blocks.4 | blocks.4 | blocks.3 | blocks.4 |
| Slowfast ResNet101 | Kinetics | 13 | blocks.5 | blocks.6 | blocks.5 | blocks.3_fast | blocks.5 |
| DINO ViT Base Patch 16 | ImageNet | 12 | blocks7attn | blocks8attn | blocks9 | blocks9 | blocks9 |
| ViT Base Patch 16 | ImageNet | 13 | blocks.11 | blocks.11 | blocks.11 | blocks.8.attn | blocks.7.attn |
| CLIP ViT-B 16 | WiT | 13 | resblocks9 | resblocks10 | resblocks7 | resblocks9 | resblocks9 |
| CLIP ViT-B 32 | WiT | 13 | resblocks10 | resblocks10 | resblocks10 | resblocks8 | resblocks8 |

**Supplementary Table 2:** Deep Neural Network Models and the extracted layers.

| Model | Layers |
|---|---|
| AlexNet | features.0, features.3, features.6, features.8, features.10 |
| VGG16 | features.0, features.2, features.5, features.7, features.10, features.12, features.14, features.17, features.19, features.21, features.24, features.26, features.28 |
| ResNet50 (ImageNet) | layer1, layer2, layer3, layer4 |
| ResNet50 (Places365) | model.4, model.5, model.6, model.7, model.9 |
| ResNet50 (ADE20k) | convlast.1, fpnout.2.0.2, ppmconv.2.2, ppmconv.3.2, ppmlastconv.2 |
| DINO ResNet50 | layer1, layer2, layer3, layer4, fc |
| CLIP ResNet101 | layer1, layer2, layer3, layer4 |
| X3D-m | blocks.1, blocks.1slow, blocks.2, blocks.2slow, blocks.3, blocks.3slow, blocks.4, blocks.4slow, blocks.5, blocks.5slow |
| Slowfast ResNet101 | blocks.1fast, blocks.1slow, blocks.2fast, blocks.2slow, blocks.3fast, blocks.3slow, blocks.4fast, blocks.4slow, blocks.5, blocks.5slow, blocks.6, blocks.6slow |
| DINO ViT Base Patch 16 | blocks.0.attn, blocks.1.attn, blocks.2.attn, blocks.3.attn, blocks.4.attn, blocks.5.attn, blocks.6.attn, blocks.7.attn, blocks.8.attn, blocks.9.attn, blocks.10.attn, blocks.11.attn |
| ViT Base Patch 16 | blocks.0, blocks.1, blocks.2, blocks.3, blocks.4, blocks.5, blocks.6, blocks.7, blocks.8, blocks.9, blocks.10, blocks.11 |
| CLIP ViT-B 16 | resblocks.0, resblocks.1, resblocks.2, resblocks.3, resblocks.4, resblocks.5, resblocks.6, resblocks.7, resblocks.8, resblocks.9, resblocks.10, resblocks.11 |
| CLIP ViT-B 32 | resblocks.0, resblocks.1, resblocks.2, resblocks.3, resblocks.4, resblocks.5, resblocks.6, resblocks.7, resblocks.8, resblocks.9, resblocks.10, resblocks.11 |

**Supplementary Table 3:** Overview of models used for fine-tuning and end-to-end trained of DNNs using affordance labels. The fine-tuned ResNet50 was pretrained on Places365, while the fine-tuned ViT Base Patch 16 was pretrained on ImageNet.

| Model | Goal | Learning Rate | Number of Epochs (5 folds) | Precision | Recall |
|---|---|---|---|---|---|
| ResNet50 | End-to-end training | 0.0015 | 36, 42, 42, 44, 45 | 0.73 (0.03) | 0.20 (0.04) |
| ResNet50 | Fine-tuning | 0.0053 | 33, 28, 31, 31, 34 | 0.85 (0.07) | 0.63 (0.03) |
| ViT Base Patch 16 | End-to-end training | 0.0006 | 25, 31, 28, 32, 30 | 0.68 (0.04) | 0.12 (0.03) |
| ViT Base Patch 16 | Fine-tuning | 0.0005 | 14, 19, 20, 20, 18 | 0.80 (0.01) | 0.67 (0.01) |

**Supplementary Table 4:** List of the 24 of 37 functions (affordances) attribute labels from SUN Attribute Database (Patterson et al., 2014). ** marked functions were excluded as they contained less than 300 images per category.

| | **Functions/affordances** | |
|---|---|---|
| Sailing/boating | Driving | Transporting things or people |
| Biking | Vacationing/touring | Climbing |
| Sunbathing** | Reading | Teaching/training** |
| Hiking | Diving | Bathing** |
| Camping | Cleaning** | Congregating |
| Studying/learning** | Competing | Exercise |
| Research** | Gaming** | Farming |
| Swimming | Shopping | Working |
| Eating | Digging** | Praying** |
| Socializing | Using tools | Conducting business** |
| Waiting in line/queuing | Sports | Playing |
| Spectating/being in an audience | Constructing/building** | Medical activity** |

**Supplementary Table 5:** Overview of prompts used to generate GPT-4 behavioral responses for navigational affordances and objects. Three tasks were used: single affordance/object selection, multiple affordance/object selections, and an unrestricted (50 words) bag of words (BoW).

| **Condition** | **Action Prompt** | **Object Prompt** |
|---|---|---|
| Single | Choose the word from 'walking, boating, biking, swimming, climbing or driving' that is the best action for the image. Your response must be a single word from the list. | What is the object option that is in the scene? Select a single option from this list: 'Building/Wall, Tree/Plant, Road/Street, Furniture, Body of water, Rocks/Stones'. Your response can only be a single option from the list. |
| Multiple | Choose the appropriate actions in the image by selecting one or multiple options from this list: 'walking, boating, biking, swimming, climbing or driving'. Your response must be only options from the list separated by commas. | Choose the appropriate objects in the image by selecting one or multiple options from this list: 'Building/Wall, Tree/Plant, Road/Street, Furniture, Body of water, Rocks/Stones'. Your response must be only options from the list separated by commas. |
| Bag of Words | Describe the action affordances in this image in 50 words.<br>List the action affordances in this image in a word each, separated by commas. | Describe the objects in this image in 50 words.<br>List the objects in this image in a word each, separated by commas. |