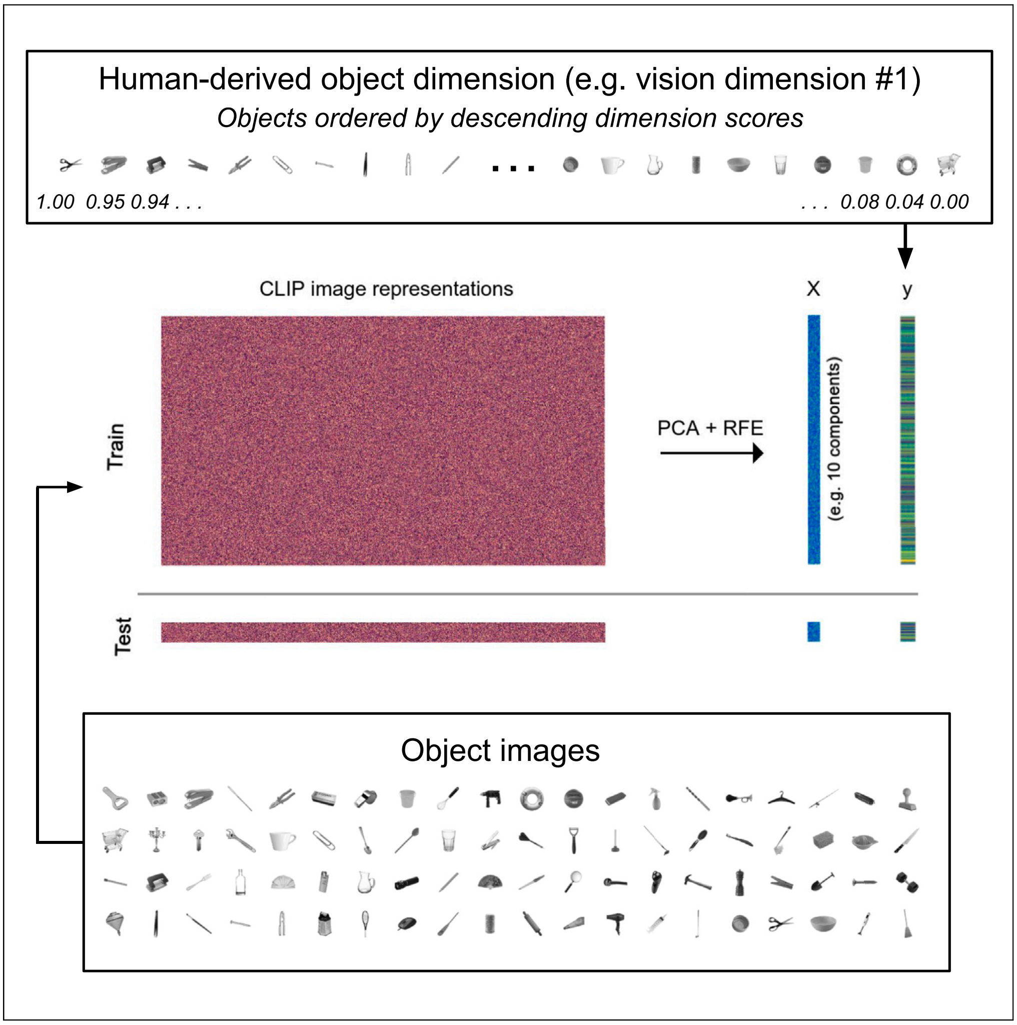
iScience

CellPress
OPEN ACCESS

Article

Fine-grained knowledge about manipulable objects is well-predicted by contrastive language image pre-training


Human-derived object dimension (e.g. vision dimension #1)
Objects ordered by descending dimension scores
1.00 0.95 0.94 . . .
. . . 0.08 0.04 0.00
CLIP image representations
X
y
Train
Test
PCA + RFE
(e.g. 10 components)
Object images

Jon Walbrin, Nikita Sossounov, Morteza Mahdiani, Igor Vaz, Jorge Almeida

jon.walbrin@gmail.com

Highlights

Human-derived manipulable object dimensions are well-predicted by CLIP

CLIP outperforms several other widely used object recognition networks (e.g., VGG16)

Multimodal pre-training on a large, diverse dataset likely drives CLIP performance

Walbrin et al., iScience *27*, 110297
July 19, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.110297

iScience

CellPress OPEN ACCESS

Article

Fine-grained knowledge about manipulable objects is well-predicted by contrastive language image pre-training

Jon Walbrin,[1,2,4,*] Nikita Sossounov,[1,2] Morteza Mahdiani,[3] Igor Vaz,[1,2] and Jorge Almeida[1,2]

SUMMARY

Object recognition is an important ability that relies on distinguishing between similar objects (e.g., deciding which utensil(s) to use at different stages of meal preparation). Recent work describes the fine-grained organization of knowledge about manipulable objects via the study of the constituent dimensions that are most relevant to human behavior, for example, vision, manipulation, and function-based properties. A logical extension of this work concerns whether or not these dimensions are uniquely human, or can be approximated by deep learning. Here, we show that behavioral dimensions are generally well-predicted by CLIP-ViT - a multimodal network trained on a large and diverse set of image-text pairs. Moreover, this model outperforms comparison networks pre-trained on smaller, image-only datasets. These results demonstrate the impressive capacity of CLIP-ViT to approximate fine-grained object knowledge. We discuss the possible sources of this benefit relative to other models (e.g., multimodal vs. image-only pre-training, dataset size, architecture).

INTRODUCTION

Correctly recognizing and using everyday manipulable objects (e.g., a hammer) is a critical human ability, and uncovering the organization of human object knowledge is a long-standing research aim. Coarse-grain, category-specific characteristics of human object knowledge are shown via category-selective brain responses (e.g., distinct cortical activations for tools, faces, animals[1–3] although these regions are typically identified using visual tasks, the representational contents of many of these areas reveal that they are sensitive to semantic properties beyond vision[4–11]). Category-specific distinctions are important, but everyday human behavior is often more reliant on finer-grain, within-category discriminations. For example, when preparing a meal, selecting and using the correct kitchen utensils is more useful than distinguishing the kitchen utensil from an animal. One powerful means of measuring finer-grained distinctions is via the study of the representational similarities between objects.[12] For example, the similarity of a given pair of objects is measured by their distance from each other within a high-dimensional representational space (e.g., a space defined by a set of object features, such as elongation, size[13]). However, the specific features, or *dimensions* that most strongly drive similarities, are often not immediately apparent.

Indeed, objects can be described by a rich set of dimensions, for instance, shape, grip type, usage context, and recent work has begun to reveal the central dimensions that govern object knowledge. For example, Hebart et al.[14] (see also[15,16]) used human behavioral judgments to develop a computational model that identifies a set of human-interpretable object dimensions. These results are complemented by other work that describes the conceptual space spanned by objects via behavioral and neural measures.[17–19] These studies used large stimulus sets that reveal dimensions that describe known categorical distinctions such as body parts, food, and animals. Recently, Almeida et al.,[20] adopted a similar approach to identify the *within-category* multidimensional structure of manipulable/hand-held objects. This work demonstrates that multimodal dimensions - vision, manipulation, and function-related dimensions - are extractable from human behavioral judgments, and that these dimensions can in turn be learned by naive participants, and are neurally relevant (i.e., they explain responses to manipulable objects across different brain areas that are consistent with the human tool network).

In brief, these dimensions provide insight into the fine-grained structure of human knowledge about manipulable objects. One important advancement of this work lies in determining whether these dimensions are uniquely "human" or whether they can be approximated by deep learning. Human object knowledge arises from multiple factors and constraints that may not be grasped in many deep learning scenarios. For example, the organization of object responses in the brain is strongly influenced by structural and functional connectivity.[7,21–23] Developmental work shows that object and conceptual knowledge undergoes continual refinement across childhood[24–29] and much of this learning depends on human mechanisms that are not explicitly implemented in most pre-trained neural networks, for example, complex sensory

[1] Proaction Laboratory, Faculty of Psychology and Educational Sciences, University of Coimbra, Coimbra, Portugal
[2] CINEICC, Faculty of Psychology and Educational Sciences, University of Coimbra, Coimbra, Portugal
[3] Mila (Quebec AI Research Institute), Montreal, QC, Canada
[4] Lead contact
*Correspondence: jon.walbrin@gmail.com

https://doi.org/10.1016/j.isci.2024.110297

experience such as visuo-motor grasping behavior. However, neural networks may still acquire some aspects of human object knowledge. Studies that probe the representational contents of pre-trained neural networks reveal evidence of "incidental learning" of information that is not explicitly intended: For instance, convolutional neural networks (CNNs) trained on an a simple image classification task can learn certain semantic associations - for example, image representations for bicycles are more similar to helmets than semantically unrelated objects, such as forks,[30] and fish are more similar to underwater scenes than land scenes.[31]

These associations arise from statistical regularities that are present in training data, for example, via repeated exposure to the co-occurence of objects in images. Indeed, there is also evidence that statistical regularities contribute to human object knowledge. For example, expectations that arise from day-to-day experience with objects are shown to influence behavioral and neural responses to objects, such as faster recognition of an airplane when presented in the upper-rather than than lower visual field, or for groups of semantically related objects relative to unrelated objects.[32,33] In short, neural networks trained on human generated data - that implicitly learn some of the statistical regularities that are inherent to human object knowledge - may capture certain aspects of human object knowledge. However, to capture finer-grained details, more powerful models are likely needed.

Recently, deep representation learning approaches have made impressive gains in terms of approximating human behavior (e.g., GPT, BERT). For example, contrastive language image pre-training (CLIP)[34] jointly trains image and text embeddings from a large dataset (e.g., > 2 billion image-text pairs) to generate semantically elaborate representations that transfer well to other tasks, such as zero-shot image classification. Beyond the exposure to a vast training dataset that may facilitate the learning of statistical regularities, the combination of visual and linguistic information employed by CLIP allows for the acquisition of deeper semantic representations: This likely accounts for recent demonstrations that CLIP outperforms comparison networks in the prediction of neural object responses.[35–37] Accordingly, CLIP is potentially well-suited to learn relatively subtle information about objects - such as manipulable object dimensions - beyond simpler semantic distinctions already shown by neural networks.[30,31]

The present study investigates whether human derived object dimensions are well-approximated by deep learning. To do so, we cast each dimension as an outcome variable in a separate regression analysis where neural network (e.g., CLIP) features serve as predictors. Modeling human-to-neural-network correspondence in this way may have several advantages over representational similarity analysis (RSA) - a common approach for human-to-neural-network comparisons.[30,31,38,39] First, RSA measures the overall similarity of two measures or systems by comparing their multi-dimensional representational spaces, but does not consider the constituent dimensions that structure these spaces: The potential interpretational value of these dimensions is overlooked - for example, understanding that human-network correspondence is driven by a dimension consistent with "elongation" is more informative than broadly construed "visual properties." Second, most RSA approaches are agnostic to the weighting of individual input features such that all features from a neural network layer are assumed to contribute equally (although some modified approaches overcome this problem.[40] In brief, modeling dimensions directly may allow a more direct and detailed comparison of human knowledge and deep learning.

Here, we show that CLIP affords relatively good predictions of human-derived object dimensions, and is able to predict behavioral categorization results over novel (unseen) objects based on these object-related dimensions. This demonstrates an impressive standard for the fine-grained representational correspondence of object information between humans and artificial neural networks.

RESULTS

Overview of dimension extraction: Almeida et al. (2023)

We investigated whether human-derived object dimensions can be reliably predicted by a CLIP network, along with a set of comparison neural networks. These dimensions are identical to those obtained by Almeida et al.,[20] and we start with a brief overview of how they were extracted (see Figure 1 for a visualization of these dimensions; see STAR Methods for further details). Subjects completed a computerized "word piling" task where they arranged a set of word icons - that depict 80 manipulable objects - into piles based on their judged similarity, that is, similar objects in the same pile. This task was performed for 3 knowledge types - vision, manipulation, and function based properties of the objects. Group level representational dissimilarity matrices that describe the pairwise dissimilarities for the object set)were obtained for each knowledge type. Non-metric MDS (using the *mdscale* function in MATLAB 2019b) with Kruskal's normalized stress criterion was applied to each of these matrices to obtain a set of dimensions that explain most of the variance for each knowledge type. This resulted in 15 dimensions overall - 5 vision dimensions, 6 manipulation dimensions, and 4 function dimensions. Importantly, each dimension describes a continuous ordering of the 80 objects, such that the numerical score of each object indicates where it is situated along the dimension (from high to low extremes).

These dimensions were extracted in a data-driven manner, and so are not forced to be cognitively interpretable, however a free-labeling task in an independent set of subjects revealed that they can be labeled consistently. For example, the first vision, manipulation, and function dimensions are consistent with the concepts of "metal vs. other materials," "power vs. precision grip," and "cooking vs. sports," respectively (see Figure 1 caption for a full description of these labels). The behavioral relevance of these dimensions was demonstrated in separate subjects that were able to accurately learn the ordering of objects along dimensions in an alternative forced choice task. Finally, these dimensions also explain the variance of brain areas (via parametric analysis) that typically show tool-preferring responses.

Human-derived dimensions are well-predicted by contrastive language image pre-training-ViT representations

To test whether human-derived dimensions can be recovered by neural network image representations, cross-validated regression was performed for each dimension separately (see Figure 2D). On each fold - 10 in total - predictions were generated from a training set that consisted

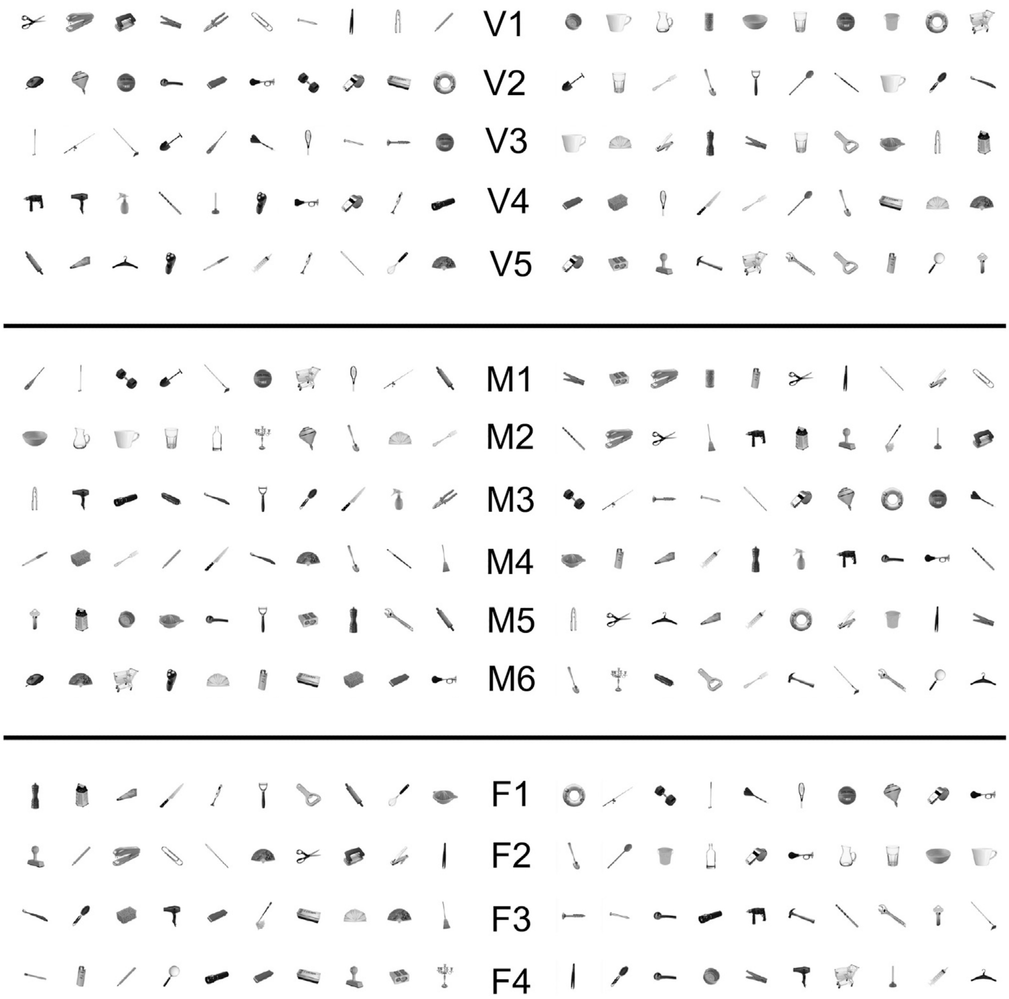


Figure 1. Visualization of objects for each of the 15 behavioral dimensions (obtained by Almeida et al.)

Vision (V1-V5), manipulation (M1-M6) and function dimensions (F1-F4) are shown. For more intuitive visualization, the 10 highest and 10 lowest objects on either extreme of the dimension are shown on the left and right of the row, respectively. The descriptive labels given for dimensions by subjects from 20, are as follows. Vision dimensions: Metal vs. other materials (V1), elongated vs. round (V2), size (V3), material properties (V4), sharp/tapering/pointy (V5). Manipulation dimensions: Power vs. precision (M1), dexterity vs. force (M2), hand vs. finger movements (M3), press/squeeze (M4), rotation (M5), fine vs. coarse movements (M6). Function dimensions: Cooking vs. sports (F1), kitchen vs. office (F2), cleaning vs. construction (F3), studying vs. household chores (F4).

of 90% of object images and subsequently tested on the remaining 10% of images. That is, from an original image set of 800 images that consisted of 10 exemplars per each of the 80 objects (see Figure 2A for example images of the 80 objects), 720 images (9 sets of exemplars) were used to train on each fold, with the remaining held-out image set of 80 object images used for testing. To achieve this, image representations were extracted from a given neural network, and principal component analysis (PCA) and recursive feature analysis (RFE) were performed to identify the final sets of features for each cross-validation fold that were used for making predictions (X = components derived from PCA + RFE; y = dimension scores for the objects). Model fits (e.g., R^2) were used to determine goodness-of-fit for each fold, and then averaged together to obtain a final model fit. For each dimension, significance testing of model fits was performed by comparing each final model fit value against a permutation distribution of model fits generated from 5000 "random dimensions" where dimension scores were shuffled (see STAR Methods for full details).

We first present results for the best performing model - CLIP-ViT - the H-14 vision transformer variant based CLIP network (based on output representations of the image encoder). We then present comparison analyses with several other networks. As described previously, CLIP-ViT was pre-trained on multimodal information (image-text pairs for >2 billion stimuli) and is shown to rival or exceed the performance of CNNs on a variety of human object-related tasks.[34,37,41]

Indeed, we show here that most behavioral dimensions are relatively well-predicted by CLIP-ViT (see Figure 3A; results shown when modeling with the best 10 components per dimension). Specifically, model fits are significantly better than a chance for each respective dimension across the 3 knowledge types - vision, manipulation, and function related knowledge. R^2 values range between 0.80 and 0.65

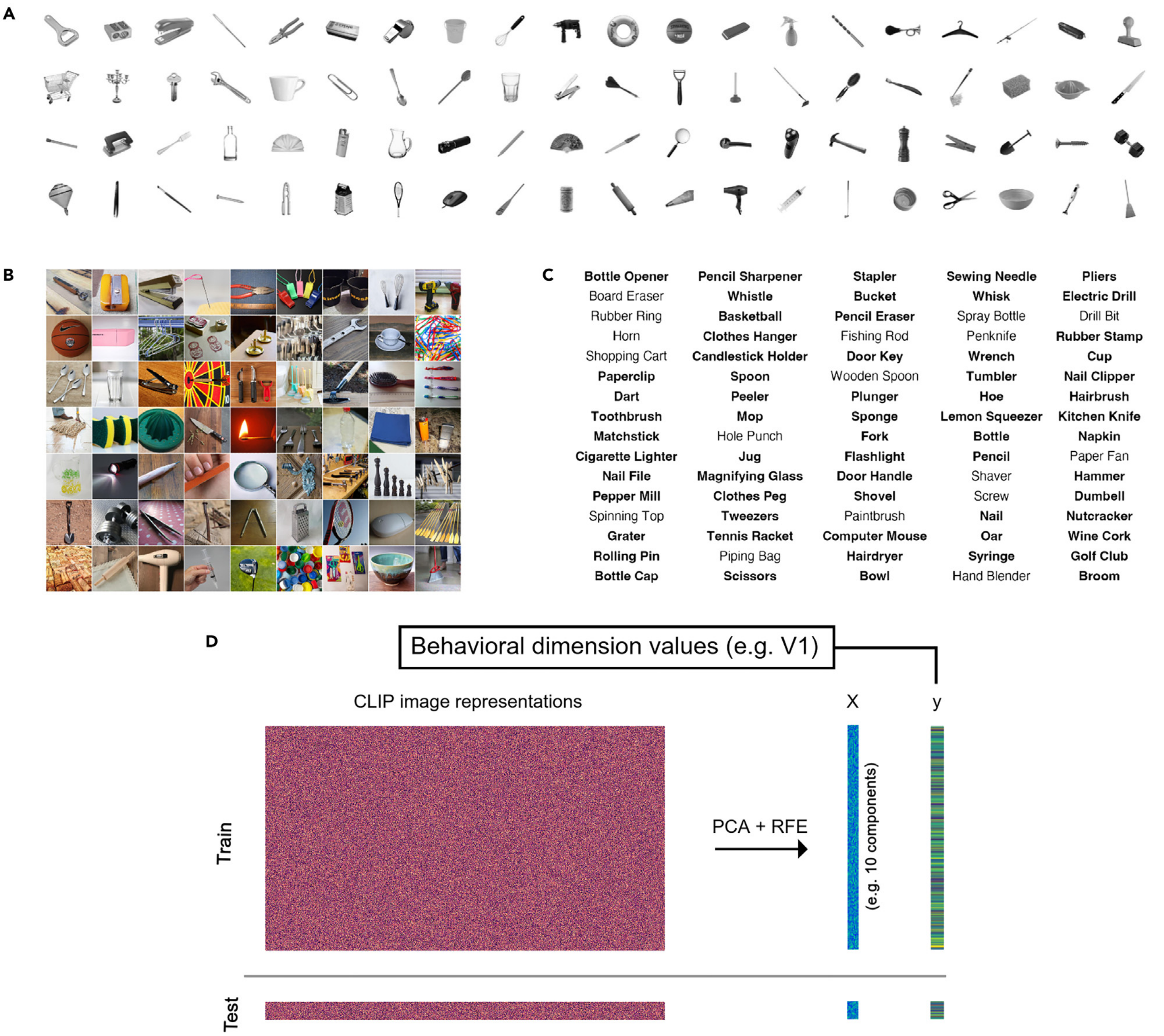


Figure 2. Stimulus sets and overview of the cross-validation approach

(A) Example images for the 80 objects image set.

(B) Example images for THINGS image set (63 object identities that are also in the 80 object set).

(C) Text labels for each of the 80 objects (objects that are also in the THINGS image set are shown in bold).

(D) Schematic of the cross-validation approach for the main analysis. In this example, CLIP image representations are depicted as rows in the large red-ish matrix (e.g., 9 exemplar sets of 80 objects = 720 training samples), with the remaining images depicted in the matrix below (80 objects for the "held-out" exemplar set = 80 test samples). Cross-validated PCA & RFE was applied to reduce these representations to a set of components (e.g., 10 components; blue matrices) that served as predictors (X) for the scores of a given dimension (y; green-blue vectors).

for the first dimension of each knowledge type. Interestingly, the strength of model fit of dimensions shows some correspondence with their ordering, with later dimensions showing weaker model fits. Importantly, this may relate to the fact that these latter dimensions explain less variance of the original similarity data.[20]

We next tested the generalizability of these predictions for a different image set - namely a subset of images taken from the THINGS image database[42] (see Figure 2B for example images). This consisted of 63 objects that were also in the 80 objects set. Unlike the original 80 objects images that were visually controlled (gray-scaled and segmented on a white background), images from the THINGS set depict objects in a more naturalistic context (e.g., full color with background scenery), and so comparisons with this image set serves to further validate the predictive capacities of CLIP-ViT. As before, we implemented cross-validated regression by training each model on representations derived from the 80 object image set, and testing on those associated with the THINGS objects (see STAR Methods for further details). As before, model fits

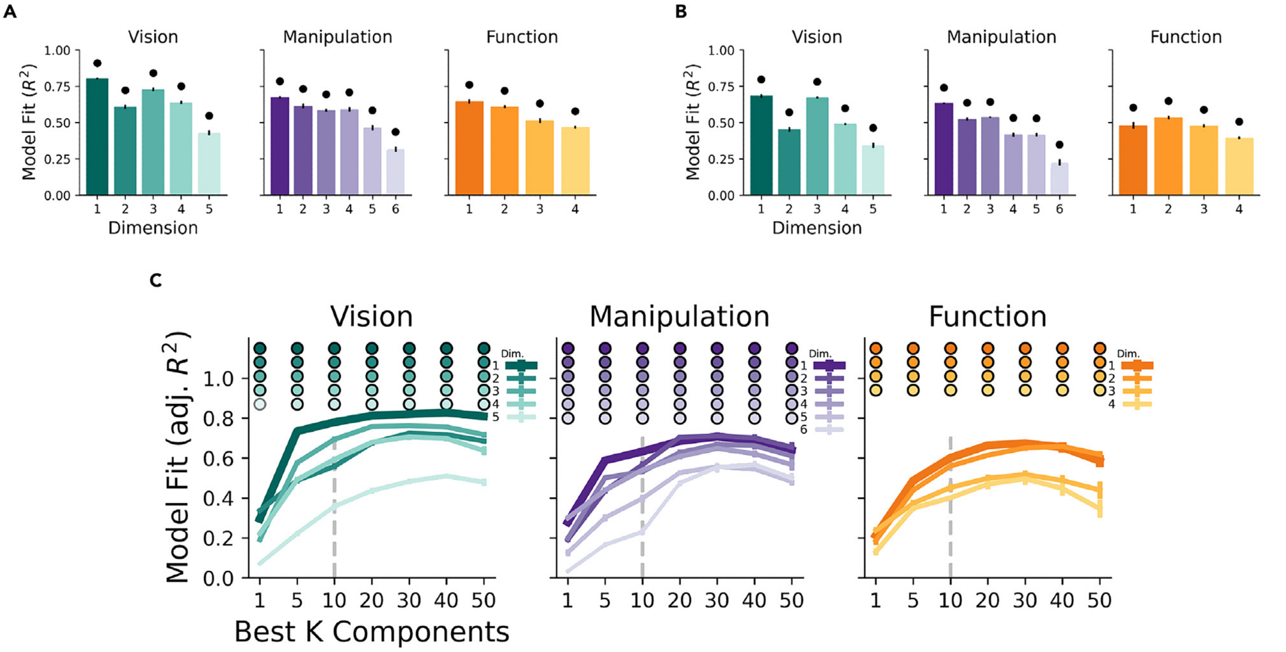


Figure 3. Regression model fits for CLIP-ViT, for each behavioral dimension

(A) Model fits for cross-validation on the original 80 object image set.

(B) Model fits when generalizing to the THINGS image set (training on 80 object images, testing on THINGS database images). These results are based on modeling with the 10 best components (per cross-validation fold) as predictors (X) for each respective target dimension (y). Bars show the mean model fit (R^2) across folds, error bars are SEM of fits across folds. Above-chance model fits based on 5000 permutations are indicated with black circles ($p \leq 0.001$).

(C) Regression model fits for each behavioral dimension when considering different sized sets of best k components. Colored lines are mean model fits for each dimension across folds, error bars are SEM of fits across folds. Model fit is adjusted R^2 to correct for inflated model fit estimates with increasingly large component sets. Vertical gray dashed line indicates best 10 components as shown in sub-figures A & B. Colored dots show significance for each dimension and best k components value - opaque and transparent circles denote significance at $p \leq 0.001$, and $p < 0.05$, respectively.

are well-above chance for most dimensions, and in most cases, of comparable magnitude (see Figure 3B; results shown when modeling with the best 10 components per dimension).

The preceding results show modeling performance when selecting the best 10 components as regression predictors; that is, on each training fold, selecting the 10 components that yield the highest regression coefficients for the training data, and then independently testing them on held-out data. We also present model fits (adjusted R^2) across a range of "best k component" values (1, 5, 10, 20, 30, 40, 50 components) for each dimension to explore how model fits change as a function of component set size. Model fits are above-chance for all dimensions and best component sets (see Figure 3C). The sharpest increase in model fit occurs for most dimensions between 1 and 5 components, with more gradual improvements up until 30–40 components, with small reductions in model fits occurring thereafter. Notably, the first and second dimensions of each knowledge type - V1, M1, F1 - tend to show the highest model fits across component sets, although later dimensions (i.e., 3 onwards) tend to show relatively similar model fits for larger sets of components. Together these results reveal the striking capacity of CLIP-ViT to capture human-derived object dimensions.

Contrastive language image pre-training-ViT component consistency analyses

The previous analyses were performed with cross-validation and so the definition and selection of CLIP-ViT components on each fold was free to vary - that is, the definition and selection of components potentially varied across folds. We therefore explored which components were most likely to be selected, and whether the components that were selected were "chosen" in a consistent fashion across folds. Two analyses were performed. First, we inspected the selection frequencies of the 10 modal best components per dimension - that is, the 10 components that were most frequently selected across all cross-validation folds, as shown in Figure 4A. We note 2 observations. 1) The first few dimensions for each of the 3 knowledge types rely strongly on "early" components (i.e., components 1–5); that is, 2 or more of the first 5 CLIP-ViT components are consistently selected for the first few behavioral dimensions. By contrast, the remaining dimensions tend to draw more upon later components (component 6 onwards). We note that, although the definition of these components may vary slightly across cross-validation folds, we assume that these differences are relatively minimal given the large amounts of training data in each fold. Nevertheless, it is likely that there is less across-fold variability in the definition of early components relative to later components, by virtue of earlier components explaining more variance of the original data than later components. This may contribute to the relatively higher selection frequencies observed for earlier

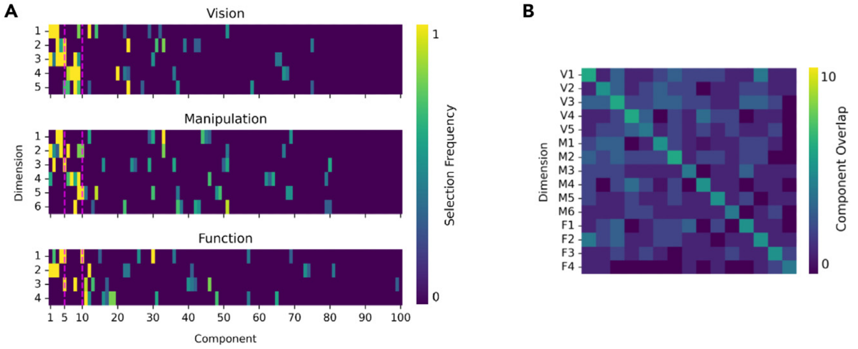


Figure 4. CLIP-ViT component analyses

(A) Selection frequencies of the modal 10 best components across 10 cross-validated folds are shown for each behavioral dimension. Color intensity (0–1) indicates the proportion of folds for which that component was included across folds. This analysis shows that the first 2–3 dimensions for each knowledge type rely strongly on the first 5 components.

(B) Conjunction (overlap) of the 10 best selected components across each of the 10-folds. Within-dimension overlap is shown on the main diagonal of the matrix, off-diagonals are between-dimension overlap. Color intensity (0–10) represents the median overlap of selected components across all folds. This analysis shows a relatively high across-fold consistency of the best 10 selected components selected for each dimension.

components. 2) Within the first 5 components, there is rarely more than one overlapping component across dimensions (however, for the first vision dimension, there are 2 overlapping components with the third vision dimension, and 3 overlapping components with the second function dimension). This demonstrates that the prediction of different dimensions largely depends on independent sets of components.

Next, we examined the consistency of component selection across folds within each dimension versus between dimensions. Specifically, we calculated the selection overlap (conjunction) of the best 10 components for each of the 10-folds - that is, the number of components that are commonly selected across the 10 fold-wise selections of best 10 components. In Figure 4B, matrix diagonals depict within-dimension overlap and off-diagonals show between-dimension overlap. Within-dimension overlap is unequivocally higher than between-dimension overlap (median within-dimension overlap ranges between 5 and 6 components; overlap was slightly lower - 4 components - for the last dimension of each knowledge type, along with the third manipulation dimension). Despite greater within-than between dimension overlap across leave-one-object-out folds, these results also suggest moderate differences in the components selected across folds that ostensibly result from slightly different exemplar images used per fold. In short, these results suggest that component selection is relatively stable across folds, and that each dimension draws from a relatively distinct set of components.

Human behavior and contrastive language image pre-training-ViT representations reveal generalizable dimension predictions

We next tested two critical aspects of object dimension knowledge with a set of 20 previously unseen objects. Examples of these images are shown in Figure 5A. First, can humans generalize their learned understanding of the dimensions to held-out objects that were not included in the original definition of these dimensions? Second, can CLIP-VIT representations predict human behavioral judgments about these held-out objects? To do so, we obtained both *model predicted dimension values* from CLIP-ViT representations and *human predicted dimension values* from a behavioral task described as follows.

This task entailed three phases (see[20] for a similar task; see STAR Methods for more details). In the first two phases (training phases), adult participants (n = 10, 10, & 11 subjects per V1, M1, & F1 dimensions, respectively) completed a computerized task. This entailed passively learning (via on-screen feedback, no response given) whether each target object (represented with a word token) was more similar to an anchor object situated at either extreme of the dimension: That is, extreme A corresponds to the object with the highest score on the dimension, and extreme B corresponds to the lowest scored object, according to the rank order of the dimension values for each object, respectively. For the second phase, subjects made a binary decision as to whether the target object in each trial was closer to the anchor object on extreme A or B of the dimension, and received feedback for each response. A subset of 50 of the original 80 objects were used as target objects.

For the third phase (test phase), participants were tested on their ability to generalize their understanding of each primary dimension (V1, M1, F1) to 20 unseen objects that were not included in either the original definition of the dimensions or the preceding training phases. Again, this entailed judging whether each unseen object was closer to an object on extreme A or B (24 trials per object; no feedback was given). Importantly, these 20 objects do not have dimension scores: Instead, behavioral proxy values for these objects were obtained by calculating the percentage of trials that each object was classified as closer to extreme A (scores ranged from 0 to 100%). The average values per object across trials and subjects were used as human predicted dimension values for subsequent comparison with predictions from the CLIP-ViT model.

Before comparing human and CLIP-ViT predictions, behavioral generalization performance was assessed by inspecting the average responses for the 20 unseen objects. Importantly, if subjects failed to generalize their dimension knowledge to these objects, the mean

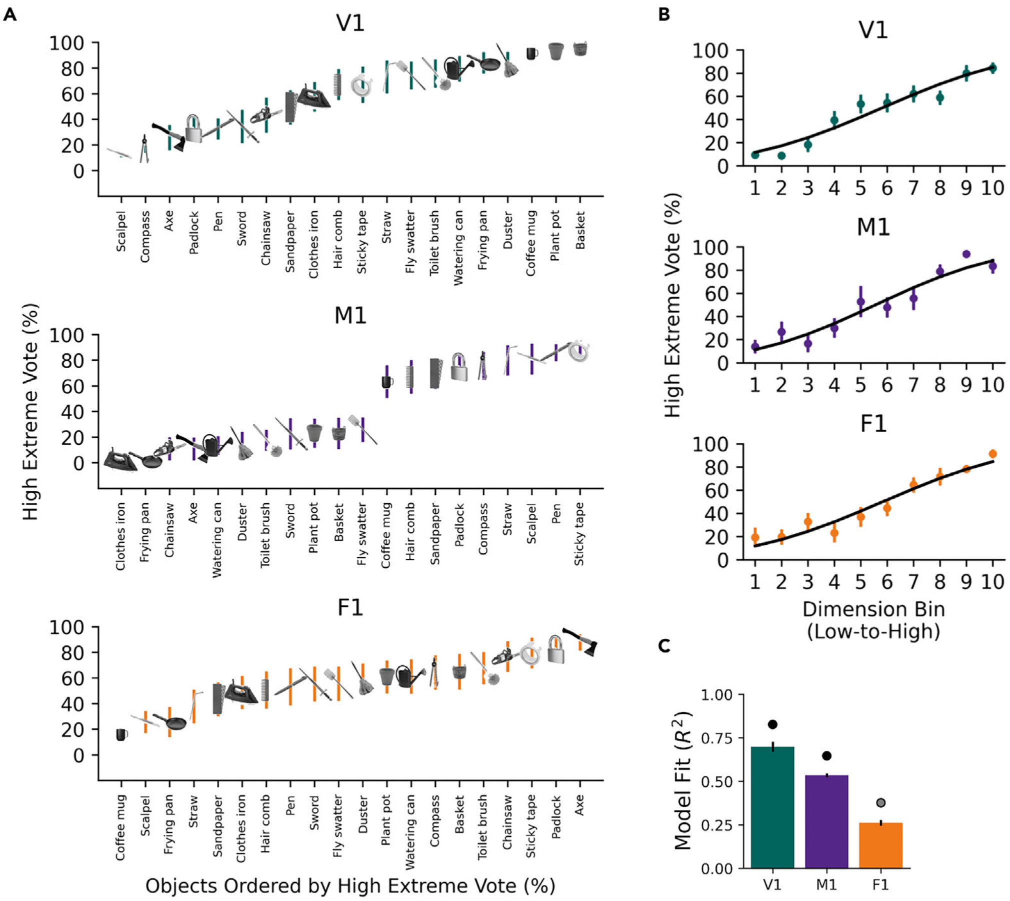


Figure 5. Both humans and CLIP-ViT representations make generalizable predictions about previously unseen objects for the primary dimensions (V1, M1, F1)

(A) Humans can generalize their understanding of the primary dimensions to a set of 20 previously unseen objects (that were not used in the original definition of the dimensions). Objects are plotted as a function of the percentage of trials where they are considered more similar to the highest scored object on the dimension (high extreme). Error bars show SEM of judgments across individuals.

(B) Comparison behavioral data for a set of 30 untrained objects from the original set of 80 objects but distinct from the 50 objects used for the training phase of the task. Markers and error bars are the mean and SEM values for "binned" objects along the dimension, plotted as an average across 8 consecutive objects, per decile bin along the extent of the dimension. Black lines are cumulative Gaussian curve fits.

(C) Human behavior and CLIP-ViT representations show generalizable predictions about the 20 unseen objects. Bars show the mean model fit (R^2) values (between the behavioral predictions (y) and model predictions), error bars are SEM of model fits across folds. Black and gray circles indicate permutation significance ($p \leq 0.001$ & $p \leq 0.01$, respectively).

percentage of extreme A votes should be close to chance level across all items - that is, all objects categorized as extreme A approximately 50% of the time. However, if subjects could correctly generalize what they had learned, object judgments should vary as a function of their similarity to extreme A objects - that is, those considered highly similar or dissimilar to extreme A should have responses close to 100% or 0%, respectively.

Subjects showed a clear ability to generalize their knowledge of the primary dimensions to completely unseen objects (see Figure 5A): The ordering generated for each dimension closely follows the expected pattern of the original dimension. For example, the arrangement of objects follows what is generally shown for the original dimensions illustrated in Figure 1, for example, "power vs. precision" grip for the first manipulation dimension. In addition to the 20 unseen objects, responses to 30 untrained objects that were not included in the training phases were also measured (3 trials per object; note that these objects were not used during the training phases, but they were used in the original set of 80 objects used to define the dimensions). Subjects' object judgments for these 30 objects (see Figure 5B) clearly vary as a function of proximity to extreme A - that is, objects that are closer to the high extreme of the dimension are judged as more similar to it than items on extreme B.

Finally, we demonstrate that CLIP-ViT representations show relatively good prediction of human dimension judgments for the 20 unseen objects (using cross-validated regression, similar to earlier analyses; see STAR Methods for full details). For vision and manipulation dimensions, there is good correspondence between the dimension predictions that both humans and CLIP-ViT make about these 20 unseen objects

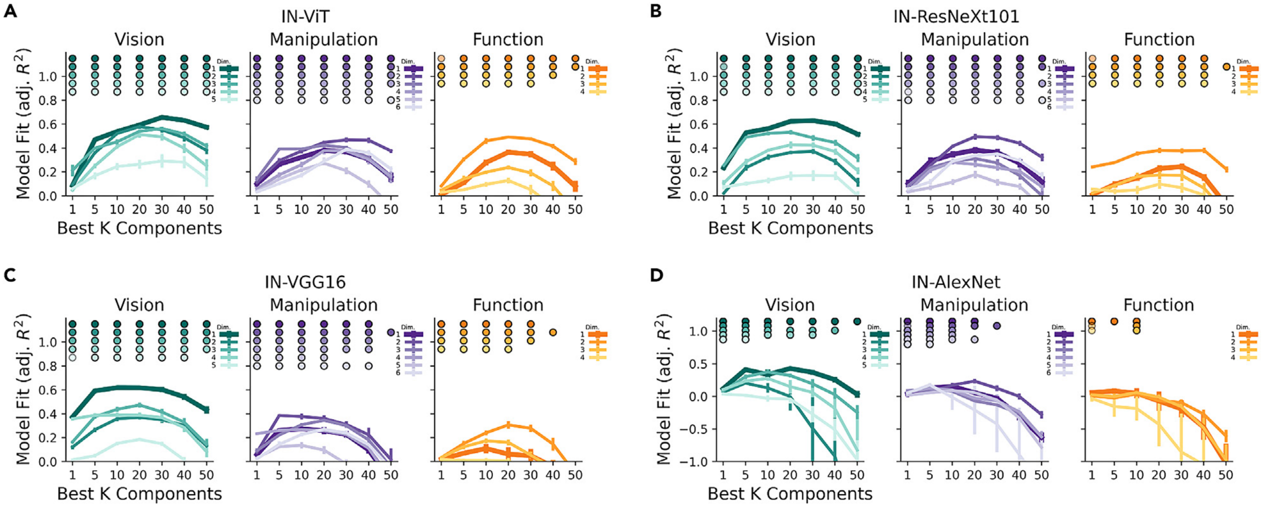


Figure 6. Regression model fits (adjusted R^2) for ImageNet pre-trained neural networks

Colored lines represent mean model fits across folds for each dimension, per best k selected components. Error bars are SEM of model fits across folds.

(A) IN-ViT.

(B) IN-ResNeXt101.

(C) IN-VGG16.

(D) IN-AlexNet (y axis re-scaled to capture poorer model fits). Colored dots show significance for each dimension and best k components value (opaque and transparent circles denote significance at $p \leq 0.001$, and $p < 0.05$, respectively).

(see Figure 5C; R^2 values: V1: 0.71; M1: 0.55). For the function dimension this correspondence is also significant, but appreciably weaker in strength (F1: 0.24; see discussion for further interpretation). In short, we show that CLIP-ViT representations make relatively generalizable predictions about human behavior that extend to objects outside of the original definition of the behavioral dimensions.

Comparison network analyses

To better contextualize the preceding CLIP-ViT results, we next tested whether other models that have been pre-trained for visual object recognition showed similar results. Specifically, we tested a set of models that were not trained with a CLIP protocol, but instead on an ImageNet protocol.[43] Briefly explained, ImageNet is a large image database that is commonly used for pre-training vision-based neural networks with an image classification task - that is, learning to associate 1000 image categories with their corresponding class labels. We expected that CLIP-ViT would substantially outperform ImageNet models due to several factors: That is, these networks were pre-trained with images only, with a simpler image classification objective, and using a smaller, less diverse training dataset.

Four networks were tested: 1) IN-ViT (visual transformer, variant H-14): This was chosen due to an identical transformer architecture as the CLIP-ViT model. Importantly, comparisons with this model allow for the determination of whether any potential performance advantage for CLIP-ViT arises from how it is pre-trained above-and-beyond its underlying architecture. 2) IN-ResNeXt101: At the time of testing, this was one of the most performant CNNs on brainscore.org (in terms of highest correspondence with neural and behavioral benchmarks). 3) IN-VGG16, and 4) IN-AlexNet: These CNNs were chosen due to their versatility and long-standing popularity (they have been widely adopted across object recognition literature and achieve good performance across many test scenarios). For each network, output features were extracted from the last layer before the classification head, and used analogously to CLIP-ViT representations.

Model fits (adjusted R^2) for each dimension and best k component set, for each network, are shown in Figure 6. Three main trends are noted. 1) For IN-ViT, IN-ResNeXt101, and IN-VGG16 networks, above-chance model fits are shown for all dimensions, across most best k component sets. 2) For these three networks, model fits tend to be highest for vision dimensions, followed by manipulation, then function dimensions. 3) IN-AlexNet performs relatively weakly compared to the other networks: Model fits are above-chance in some cases, especially for the first dimension of each knowledge type, for the best 10–20 component sets before model fits fall-off. Fold-wise estimates are also extremely noisy in some cases (see large error bars in Figure 6D; indeed, we observed that model fits were especially bad for 1-fold and so we visualize results for the 9 remaining folds to avoid excessive re-scaling of the y axis). Additionally, we provide supplementary evidence that modified image pre-training (EcoSet pretraining[44]) does not afford an appreciable improvement on standard ImageNet pre-training (see Figure S1).

Finally, we ran direct comparisons between CLIP-ViT and the four other networks. Specifically, for each dimension and best k components set, we calculated the unique variance explained by CLIP-ViT over-and-above each of the four other networks. For example, in the case of the CLIP-ViT > IN-ViT comparison, this was calculated as the model fit of both models combined minus the model fit of IN-ViT ([CLIP-ViT + IN-ViT R^2] - IN-ViT R^2). Three trends are noted. 1) CLIP-ViT explains substantially more unique variance than the other models do (see Figure 7;

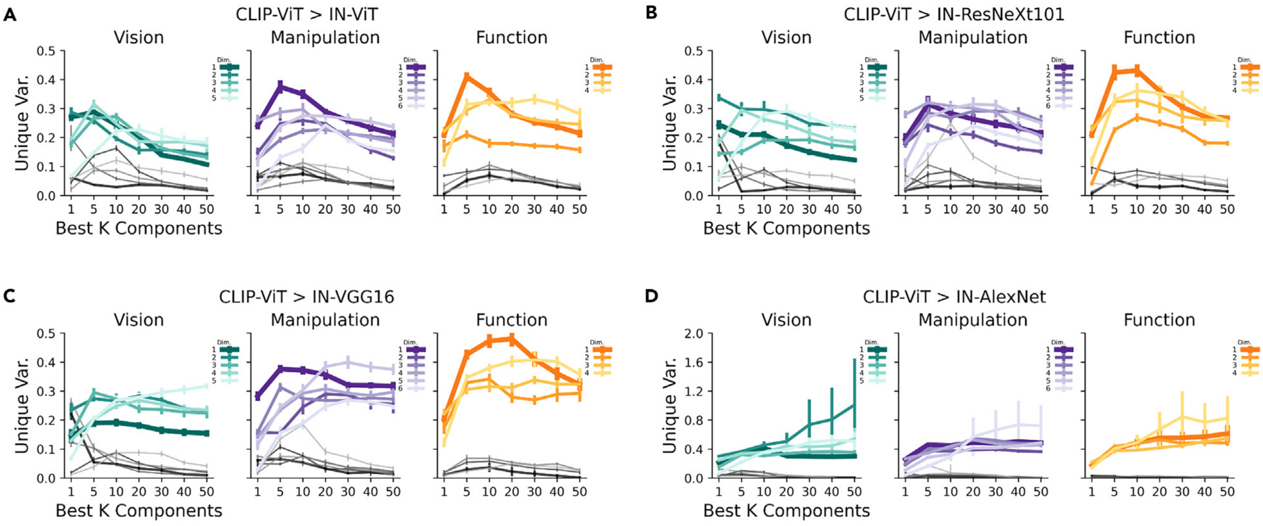


Figure 7. Unique variance explained for CLIP-VIT relative to ImageNet networks

(A) IN-ViT as the comparison network.

(B) IN-ResNeXt101 as the comparison network.

(C) IN-VGG16 as the comparison network.

(D) IN-AlexNet as the comparison network (y axis is re-scaled to account for high unique variance estimates resulting from very poor model fits, as shown in Figure 6D). Colored lines represent unique variance explained by CLIP-VIT over-and-above the comparison network, for each dimension, across different best k component sets. Error bars are SEM of exemplar image sets. For reference, unique variance explained by the comparison network over-and-above CLIP-ViT is plotted with gray scale lines (first dimension plotted in black with each successive dimension plotted in lighter grays).

colored lines show unique variance of CLIP-ViT, and for reference, the "other" network is shown with gray-scaled lines). 2) Unique variance for CLIP-ViT is often most pronounced for best k component sets 5–10. 3) Unique variance explained by CLIP-ViT is generally highest for the function dimensions, followed by manipulation-, then vision dimensions. In summary, CLIP-ViT shows superior prediction of the behavioral dimensions than the other 4 networks, despite reasonable model fits in many cases.

DISCUSSION

We present 3 main findings here. First, human-derived dimensions that describe different knowledge properties of manipulable objects - that is, visual, manipulation, functional dimensions - are relatively well-predicted by CLIP. Second, this model can generate good predictions of human behavior for previously unseen objects. Third, the predictive capacity of this model is substantially greater than comparison models, and that most of this benefit ostensibly results from several differences in pre-training regime (i.e., learning objective, data modality, and amount of training data). These results extend well-established demonstrations that deep neural networks represent broad characteristics of visual object recognition.[30,37–39,41,45–50]

Indeed, recent work demonstrates that CLIP affords excellent predictions of human brain activity. For example, Wang et al.[36] used CLIP image features as inputs to voxelwise encoding models that generate good predictions of fMRI responses to objects and natural scenes. Similarly, a recent study by Contier et al.[35] used CLIP image features to encode and predict fMRI responses to object dimensions derived from the THINGS object set: This powerful approach revealed well-known cortical topologies, for example the organization of category-selective cortex and low-level feature turning in early visual cortex. These findings, along with the present results, reveal the striking capacity of CLIP to predict high-level human behavior and brain responses.

Notably, we observed that CLIP-ViT vastly outperforms a set of comparison networks, and that this benefit potentially results from several differences between CLIP and ImageNet pre-training regimes, namely: Data modality (image & text vs. Image only), learning objective (joint-learning of image-text representation pairs vs. image classification), and training dataset size (larger vs. smaller). We do not make strong claims about any specific factor alone, but consider that these factors collectively contribute to the superior performance of CLIP-ViT. We do however show evidence that ViT architecture - in itself - does not drive this performance benefit: CLIP-ViT substantially outperforms the ImageNet version of ViT. Indeed, recent fMRI work shows a very minimal effect of network architecture for CLIP models when predicting high-level neural responses to objects (e.g., CLIP models with ViT vs. ResNet as image encoders[36,37]). Similarly, other fMRI work demonstrates that the influence of pre-training regime is a much more significant factor on downstream task performance than network architecture.[44]

Indeed, this pre-training advantage benefits the prediction of dimensions across each of the three tested knowledge domains. Strong performance for vision dimensions is relatively unsurprising given the vast exposure to image information during pre-training. Notably, functional information about objects is also, to some extent, accessible via visual information: This is shown to some extent with simpler pre-training schemes for CNNs that can learn contextual associations between objects.[30,31] Beyond visual information, the linguistic contributions of CLIP likely contribute to the improved approximation of functional information relative to vision-only models (see Figure 7). Manipulation dimensions were also relatively well-predicted. In lieu of the human motoric experience, CLIP may indirectly learn from the presence of cues such as hand-posture, or relative size information, conveyed during pre-training by images as well as their associated text descriptions (e.g., "using tweezers to extract a splinter" indicates fine, precise manipulation).

The predictive capacity of CLIP-ViT has important consequences for the interpretation and understanding of how human knowledge about manipulable objects might develop. Clearly, there are stark differences between human learning and artificial deep learning. For instance, the organization of human object knowledge in the brain is constrained by connectivity,[7,9,21–23,51–53] and is refined via lived experience, as shown by developmental[24–29] and visual expertise work.[54–57] Other experiential human factors such as visuo-motor behavior, and the acquisition of learned statistical regularities about objects - for example, repeated exposure to the co-occurrence of objects, for example, spoon and bowl - undoubtedly contribute to this knowledge too. While definitively elucidating the individual contributions of these sources of knowledge toward human-based object dimensions is beyond the scope of the current study, the present findings suggest that the statistical regularities that CLIP exploits, along with the semantic boost afforded by its linguistic inputs, are crucial factors. Importantly, the fact that CLIP out-performed all other models strongly suggests that low-level (visual) regularites are not enough to explain and determine object representations and related information.

We note that, while many dimensions have relatively good model fits, these results are not at ceiling - the best R^2 values are around 0.80, and not all dimensions are equally well-predicted. In our view, this implies that, while modal and amodal statistical learning is an important factor in determining fine-grained knowledge about manipulable objects, we expect that other critical human factors - such as those outlined above - make important contributions too. Determining the relative contributions of these knowledge sources is certainly a worthy future research aim.

Although comparison networks show poorer predictions of dimensions than CLIP-ViT overall, there are several notable aspects of these results in three of the ImageNet pre-trained networks that show many above-chance model fits (IN-ViT, IN-ResNeXt101, & IN-VGG16). These models show relatively good prediction of vision dimensions, suggesting that the visual tuning of these models is sufficient for reasonable approximations of these dimensions. These networks also show above-chance, albeit weaker, prediction of manipulation and function dimensions, suggesting that image-only pre-training may partially capture these kinds of information, ostensibly via statistical regularities or contextual information that is incidentally captured across training (e.g., as similarly shown by[30,31]). By contrast, IN-AlexNet shows substantially poorer performance than all other models. This is likely due to its relatively shallow architecture - 5 convolutional layers in total - that may not be sufficient to capture information about dimensions that deeper comparison networks are able to.

We also consider relative differences between dimensions within each knowledge domain, for CLIP-ViT. One notable trend is that the best model fits are always shown for the first or second dimensions. However, these differences are not always strongly pronounced and so we do not wish to overstate this trend. For example, for larger component sets, the first 2 manipulation dimensions are only slightly better than dimensions 3 & 4 (see Figure 3C). This suggests that in most cases, the first few dimensions are well-approximated with smaller component sets, but later dimensions (e.g., dimension 3 onwards) provide good predictions when considering larger component sets.

We also ran exploratory analyses that potentially provide more detail on why these first few dimensions perform well: The first few dimensions of each knowledge domain draw strongly from early CLIP-ViT components (first 5 components). As mentioned in the results, this likely reflects the possibility that early components "benefit" from less variability across folds, by virtue that early components explain more variance than later components, and as such, the definition of these components on each fold is likely to be more similar than for later components. Additionally, this result might also demonstrate loose commonalities between the original representational spaces from which behavioral dimensions and CLIP-ViT components are extracted; that is, that early dimensions/components extracted from each respective space seem to show some broad similarity in terms of the variance that they capture. However, we are cautious not to overstate a like-for-like correspondence between dimensions and components, given supplementary evidence that individual CLIP-ViT components do not show a clear resemblance to any behavioral dimension (see Figure S4 for a visualization of the object ordering for each of the first 10 CLIP-ViT components). Moreover, as the main analyses show (see Figure 3C), combinations of multiple components better predict individual behavioral dimensions than any individual component does.

We observed largest unique variance gains for CLIP-ViT for function dimensions, and in particular F1, indicating that this network provides an especially strong advantage over the other networks in terms of predicting function knowledge. We also show that although F1 predictions seem to generalize to other object sets - both THINGS images and 20 unseen object images - this is to a lesser extent than V1 and M1 dimensions. In the case of THINGS images, these images contain relatively more contextual information such as background information and the presence of other objects that is not present in the 80 objects images that have been segmented and presented on a white background. These differences may account for slightly poorer generalization where these kinds of contextual information might be especially informative. In the case of the 20 unseen objects analysis, these objects were pre-selected before analyzing and interpreting the ordering of objects along the F1 dimension: As shown in Figure 1, this dimension broadly depicts cooking vs. sports-related objects on the two extremes, but the chosen set of 20 unseen objects contains many instances that do not clearly correspond to either of these extremes. This may contribute to poorer generalization performance for F1, relative to M1 and V1 where intermediate items can be more intuitively situated on the dimension.

In summary, the present findings are in broad agreement with recent work that reveals strong correspondence between CLIP and higher-level conceptual object knowledge.[15,35–37] This work represents a necessary step toward a better understanding of the fine-grained organization of human object knowledge.

Limitations of the study

The dimensions that we focus on here are not exhaustive and those extracted from other cognitive domains are necessary for a deeper understanding of high-level human behavior. Future work may address different formulations of dimensions. For example, here we obtained manipulation dimensions by probing declarative knowledge about how objects are used, but this potentially differs in some ways from dimensions extracted directly from motor behavior.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - ○ Image sets
  - ○ Neural network feature extraction
  - ○ Component selection
  - ○ Regression
  - ○ Behavioral task
  - ○ Analysis of unique variance explained
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110297.

ACKNOWLEDGMENTS

This work was supported by: European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program Starting Grant number 802553; "ContentMAP'' awarded to JA; European Research Executive Agency Widening programme under the European Union's Horizon Europe Grant 101087584 "CogBooster" awarded to JA; Portuguese Foundation for Science and Technology (FCT) with: a) An Individual Scientific Employment Stimulus Grant awarded to JW (2021.03185.CEECIND); b) An Advanced Computing Project Grant awarded to JW (2021.09761.CPCA).

AUTHOR CONTRIBUTIONS

Jon Walbrin: *conceptualization, software, formal analysis, investigation, writing - original draft, writing - review and editing, and visualization.* Jorge Almeida: *conceptualization, resources, and funding acquisition.* Nikita Sossounov: *software, formal analysis, investigation, visualization, and writing - review and editing.* Morteza Mahdiani: *formal analysis and investigation.* Igor Vaz: *formal analysis and investigation.*

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 7, 2023
Revised: February 23, 2024
Accepted: June 14, 2024
Published: June 17, 2024

REFERENCES

1. Downing, P.E., Chan, A.W.Y., Peelen, M.V., Dodds, C.M., and Kanwisher, N. (2006). Domain specificity in visual cortex. Cerebr. Cortex *16*, 1453–1461.
2. Grill-Spector, K., and Weiner, K.S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. Nat. Rev. Neurosci. *15*, 536–548.
3. Peelen, M.V., and Downing, P.E. (2017). Category selectivity in human visual cortex: Beyond visual object recognition. Neuropsychologia *105*, 177–183.
4. Capitani, E., Laiacona, M., Mahon, B., and Caramazza, A. (2003). What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. Cogn. Neuropsychol. *20*, 213–261.
5. Caramazza, A., and Shelton, J.R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. J. Cognit. Neurosci. *10*, 1–34.
6. Eick, C.M., Kovács, G., Rostalski, S.M., Röhrig, L., and Ambrus, G.G. (2020). The occipital face area is causally involved in identity-related visual-semantic associations. Brain Struct. Funct. *225*, 1483–1493.

7. Mahon, B.Z., Anzellotti, S., Schwarzbach, J., Zampini, M., and Caramazza, A. (2009). Category-specific organization in the human brain does not require visual experience. Neuron *63*, 397–405.
8. Mahon, B.Z., and Caramazza, A. (2009). Concepts and categories: a cognitive neuropsychological perspective. Annu. Rev. Psychol. *60*, 27–51.
9. Mahon, B.Z., and Caramazza, A. (2011). What drives the organization of object knowledge in the brain? Trends Cognit. Sci. *15*, 97–103.
10. Ratan Murty, N.A., Teng, S., Beeler, D., Mynick, A., Oliva, A., and Kanwisher, N. (2020). Visual experience is not necessary for the development of face-selectivity in the lateral fusiform gyrus. Proc. Natl. Acad. Sci. USA *117*, 23011–23020.
11. Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A.L., McGettigan, C., and Garrido, L. (2021). FFA and OFA encode distinct types of face identity information. J. Neurosci. *41*, 1952–1969.
12. Kriegeskorte, N., Mur, M., and Bandettini, P.A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4.
13. Valério, D., Hussain, A., and Almeida, J. (2023). Semantic feature production norms for manipulable objects. Preprint at bioRxiv. https://doi.org/10.1101/2023.04.24.537452.
14. Hebart, M.N., Zheng, C.Y., Pereira, F., and Baker, C.I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nat. Human Behav. *4*, 1173–1185.
15. Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R.A., and Kornblith, S. (2022). Human alignment of neural network representations. Preprint at arXiv. https://doi.org/10.48550/arXiv.2211.01201.
16. Josephs, E.L., Hebart, M.N., and Konkle, T. (2023). Dimensions underlying human understanding of the reachable world. Cognition *234*, 105368.
17. Binder, J.R., Conant, L.L., Humphries, C.J., Fernandino, L., Simons, S.B., Aguilar, M., and Desai, R.H. (2016). Toward a brain-based componential semantic representation. Cogn. Neuropsychol. *33*, 130–174.
18. Fernandino, L., Tong, J.Q., Conant, L.L., Humphries, C.J., and Binder, J.R. (2022). Decoding the information structure underlying the neural representation of concepts. Proc. Natl. Acad. Sci. USA *119*, e2108091119.
19. Huth, A.G., Nishimoto, S., Vu, A.T., and Gallant, J.L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron *76*, 1210–1224.
20. Almeida, J., Fracasso, A., Kristensen, S., Valério, D., Bergström, F., Chakravarthi, R., Tal, Z., and Walbrin, J. (2023). Neural and behavioral signatures of the multidimensionality of manipulable object processing. Commun. Biol. *6*, 940.
21. Kamps, F.S., Hendrix, C.L., Brennan, P.A., and Dilks, D.D. (2020). Connectivity at the origins of domain specificity in the cortical face and place networks. Proc. Natl. Acad. Sci. USA *117*, 6163–6169.
22. Peelen, M.V., He, C., Han, Z., Caramazza, A., and Bi, Y. (2014). Nonvisual and visual object shape representations in occipitotemporal cortex: evidence from congenitally blind and sighted adults. J. Neurosci. *34*, 163–170.
23. Saygin, Z.M., Osher, D.E., Norton, E.S., Youssoufian, D.A., Beach, S.D., Feather, J., Gaab, N., Gabrieli, J.D.E., Kanwisher, N., and Kanwisher, N. (2016). Connectivity precedes function in the development of the visual word form area. Nat. Neurosci. *19*, 1250–1255.
24. Bova, S.M., Fazzi, E., Giovenzana, A., Montomoli, C., Signorini, S.G., Zoppello, M., and Lanzi, G. (2007). The development of visual object recognition in school-age children. Dev. Neuropsychol. *31*, 79–102.
25. Huber, L.S., Geirhos, R., and Wichmann, F.A. (2023). The developmental trajectory of object recognition robustness: children are like small adults but unlike big deep neural networks. J. Vis. *23*, 4.
26. Jüttner, M., Müller, A., and Rentschler, I. (2006). A developmental dissociation of view-dependent and view-invariant object recognition in adolescence. Behav. Brain Res. *175*, 420–424.
27. Nishimura, M., Scherf, S., and Behrmann, M. (2009). Development of object recognition in humans. F1000 Biol. Rep. *1*, 56.
28. Scherf, K.S., Behrmann, M., Humphreys, K., and Luna, B. (2007). Visual category-selectivity for faces, places and objects emerges along different developmental trajectories. Dev. Sci. *10*, F15–F30.
29. Walbrin, J., Almeida, J., and Koldewyn, K. (2023). Alternative brain connectivity underscores age-related differences in the processing of interactive biological motion. J. Neurosci. *43*, 3666–3674.
30. Aminoff, E.M., Baror, S., Roginek, E.W., and Leeds, D.D. (2022). Contextual associations represented both in neural networks and human behavior. Sci. Rep. *12*, 5570.
31. Bracci, S., Mraz, J., Zeman, A., Leys, G., and Op de Beeck, H. (2023). The representational hierarchy in human and artificial visual systems in the presence of object-scene regularities. PLoS Comput. Biol. *19*, e1011086.
32. Kaiser, D., and Cichy, R.M. (2018). Typical visual-field locations facilitate access to awareness for everyday objects. Cognition *180*, 118–122.
33. Kaiser, D., Quek, G.L., Cichy, R.M., and Peelen, M.V. (2019). Object vision in a structured world. Trends Cognit. Sci. *23*, 672–685.
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In International conference on machine learning (PMLR), pp. 8748–8763.
35. Contier, O., Baker, C.I., and Hebart, M.N. (2023). Distributed representations of behaviorally-relevant object dimensions in the human visual system. Preprint at bioRxiv. https://doi.org/10.1101/2023.08.23.553812.
36. 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. Nat. Mach. Intell. *5*, 1415–1426.
37. Zhou, Q., Du, C., and He, H. (2022). Exploring the brain-like properties of deep neural networks: a neural encoding perspective. Mach. Intell. Res. *19*, 439–455.
38. Xu, Y., and Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. Nat. Commun. *12*, 2065.
39. Zeman, A.A., Ritchie, J.B., Bracci, S., and Op de Beeck, H. (2020). Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. Sci. Rep. *10*, 2453.
40. Kaniuth, P., and Hebart, M.N. (2022). Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. Neuroimage *257*, 119294.
41. Tuli, S., Dasgupta, I., Grant, E., and Griffiths, T.L. (2021). Are convolutional neural networks or transformers more like human vision?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2105.07197.
42. Hebart, M.N., Dickter, A.H., Kidder, A., Kwok, W.Y., Corriveau, A., Van Wicklin, C., and Baker, C.I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. PLoS One *14*, e0223792.
43. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (IEEE), pp. 248–255.
44. Mehrer, J., Spoerer, C.J., Jones, E.C., Kriegeskorte, N., and Kietzmann, T.C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. Proc. Natl. Acad. Sci. USA *118*, e2011417118.
45. Bracci, S., and Op de Beeck, H.P. (2023). Understanding human object vision: a picture is worth a thousand representations. Annu. Rev. Psychol. *74*, 113–135.
46. Cichy, R.M., and Kaiser, D. (2019). Deep neural networks as scientific models. Trends Cognit. Sci. *23*, 305–317.
47. Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F.A., and Brendel, W. (2021). Partial success in closing the gap between human and machine vision. Adv. Neural Inf. Process. Syst. *34*, 23885–23899.
48. Kubilius, J., Bracci, S., and Op de Beeck, H.P. (2016). Deep neural networks as a computational model for human shape sensitivity. PLoS Comput. Biol. *12*, e1004896.
49. Lee, D., and Almeida, J. (2021). Within-category representational stability through the lens of manipulable objects. Cortex *137*, 282–291.
50. Mukherjee, K., and Rogers, T. (2023). Using Drawings and Deep Neural Networks to Characterize the Building Blocks of Human Visual Similarity.
51. Amaral, L., Bergström, F., and Almeida, J. (2021). Overlapping but distinct: distal connectivity dissociates hand and tool processing networks. Cortex *140*, 1–13.
52. Walbrin, J., and Almeida, J. (2021). High-level representations in human occipito-temporal cortex are indexed by distal connectivity. J. Neurosci. *41*, 4678–4685.
53. Walbrin, J., Downing, P.E., Sotero, F.D., and Almeida, J. (2024). Characterizing the discriminability of visual categorical information in strongly connected voxels. Neuropsychologia *195*, 108815.
54. Bilalić, M., Grottenthaler, T., Nägele, T., and Lindig, T. (2016). The faces in radiological images: fusiform face area supports radiological expertise. Cerebr. Cortex *26*, 1004–1014.
55. Duyck, S., Martens, F., Chen, C.Y., and Op de Beeck, H. (2021). How visual expertise changes representational geometry: A

behavioral and neural perspective. J. Cognit. Neurosci. *33*, 2461–2476.

56. Harley, E.M., Pope, W.B., Villablanca, J.P., Mumford, J., Suh, R., Mazziotta, J.C., Enzmann, D., Engel, S.A., and Engel, S.A. (2009). Engagement of fusiform cortex and disengagement of lateral occipital cortex in the acquisition of radiological expertise. Cerebr. Cortex *19*, 2746–2754.

57. Martens, F., Bulthé, J., van Vliet, C., and Op de Beeck, H. (2018). Domain-general and domain-specific neural changes underlying visual expertise. Neuroimage *169*, 80–93.

58. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. Adv. Neural Inf. Process. Syst. *35*, 25278–25294.

59. Mahmoudpour, S., and Schelkens, P. (2023). On the Agreement of Deep Neural Networks with the Brain in Encoding Visual Stimuli: Implications for Image Quality Assessment. In 2023 24th International Conference on Digital Signal Processing (DSP) (IEEE), pp. 1–5.

60. Muttenthaler, L., and Hebart, M.N. (2021). THINGSvision: a Python toolbox for streamlining the extraction of activations from deep neural networks. Front. Neuroinf. *15*, 679838.

61. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Data and code for reproducing results	This paper	github.com/jwalbrin/OK_CLIP
Software and algorithms		
Python 3.11	Python software foundation	python.org/downloads/release/python-3110/
Scikit-learn 1.3.2	Scikit-learn	scikit-learn.org/stable/
Numpy 1.26.3	Numpy	numpy.org
Other		
Dimension data from Almeida et al. 2023	Almeida et al. 2023	osf.io/jzuf3/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Jon Walbrin (jon.walbrin@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data required to re-produce analyses from this paper is available on github (github.com/jwalbrin/OK_CLIP).
- All code required to re-produce analyses from this paper is available on github (github.com/jwalbrin/OK_CLIP).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Participants were 31 undergraduate students from the University of Coimbra (Portugal), aged between 18-40 years, 50% female, and of white-european background. Participants completed a computer-based behavioral experiment. This data was used as both a behavioral validation task, and to test the correspondence with predictions generated from a neural network. No explicit sample size calculation was performed, but the sample here is in line with sufficient sample sizes in related work.[20] No between sex analyses were performed as there was no a-priori reason to do so. These experiments were approved by the ethics committee of the Faculty of Psychology and Educational Sciences of the University of Coimbra, and followed all ethical guidelines. Participants provided written informed consent, and were compensated for their time (either with course credit or financial compensation).

METHOD DETAILS

Image sets

The main image set consisted of 80 objects (400 * 400 pixels) presented in gray-scale on a white background, as used by Almeida et al.[20] (see Figure 2A for example images and Figure 2C for object names). There were 10 exemplar images per object (800 images total). Similar images were generated for an additional set of 20 objects (gray-scaled, white background; 400 * 400 pixels; 10 exemplars each, 200 images total) for use in generalization analyses - examples of these images are shown in Figure 5A. Another set of images were selected from the THINGS image database[42] that depict color images of objects in natural contexts (variable image sizes; see Figure 2B for example images); we inspected the database and identified 63 objects that were also contained in the main 80 objects image set, and randomly selected 10 exemplars each (630 images total).

Neural network feature extraction

We extracted image representations, rather than text representations, for the CLIP-ViT network because: a) The current results follow on from the study by 20, and so we wanted to extract representations from images used in that study; b) The 80 object images that we did use could be easily compared with other existing image sets, such as the THINGS image set[42]; c) this allowed for direct comparisons with several other image-only neural networks that show good performance on object recognition tasks.

We extracted image representations from CLIP-ViT - a pre-trained vision transformer trained with CLIP.[34] Briefly explained, this network is pre-trained with pairs of images and text descriptions (e.g. "ugly Christmas sweaters with cats"). This entails a contrastive learning procedure that jointly trains (updates the weights of) an image encoder and text encoder to generate representations that maximize the cosine similarity between correct image-text pairs. This network (obtained from pypi.org/project/timm/) was pretrained on a dataset consisting of ~2.3 billion image-text pairs (LAION2B subset of the LAION5B dataset.[58] For feature extraction, we obtained the final output representations that consisted of 1280 features from the class token embeddings of the image encoder/vision transformer, similar to previous work.[37,59] We used the THINGSvision toolbox[60] for extraction.

For the comparison ImageNet pre-trained networks, features were extracted as follows: a) IN-ViT (1280 features from the class token embedding layer of H-14 variant of vision transformer pre-trained on ImageNet 21K; pypi.org/project/timm); b) IN-ResNeXt101 (2048 features from the average pooling layer of ResNeXt101 CNN pre-trained on ImageNet 1K; pytorch.org); c) IN-VGG16 (4096 features from the last activation layer of VGG16 CNN pre-trained on ImageNet 1K; pytorch.org), and d) IN-AlexNet (4096 features from last activation layer of AlexNet CNN pre-trained on ImageNet 1K; pytorch.org).

Component selection

We used principal component analysis (PCA; implemented with Scikit-learn Python package[61]) to reduce the image representations of each network to an initial set of 100 components. This set of initial components was the same for each dimension tested (but the selected components were determined with cross-validation, per dimension). We used dimensionality reduction for 3 reasons: a) To alleviate differences in feature set size between the networks (e.g. 1280 for CLIP-ViT vs. 2048 for ResNeXt101); b) To ensure that predictors were orthogonal to each other, and; c) Because components individually explain more variance than individual features in the original representation. Recursive feature elimination (RFE; implemented with Scikit-learn Python package[61]) was used to determine the best performing components for each behavioral dimension (k components). The ranking of components was determined by obtaining their regression coefficients, iteratively removing the single weakest component until k features have been selected. As described in the results section, component definition and selection (PCA & RFE) was performed with cross-validation such that test samples were never used for the definition and selection of components. We also present supplementary analyses that demonstrate virtually identical results when using nested cross-validation with PCA and RFE for component selection (see Figures S2 and S3).

Regression

For each analysis, cross-validated principal component regression was implemented (the same cross-validation scheme was applied to PCA and RFE beforehand). For the original analysis (see Figure 2D), the training partition of each fold contained 9 exemplar image sets of 80 objects (720 samples), with the left out exemplar set used for the test partition (80 samples). We note that we previously obtained similar results with a leave-one-item-out regression scheme (i.e. training on all data for 79 objects (790 samples), and testing on all data for the held out object (10 samples)), but we opt for the "leave-one-exemplar-out" scheme described above as this attenuated overfitting in some cases (e.g. very poor model fits were shown when testing larger sets of components).

Model fit values (e.g. R^2) were calculated for the 80 held out objects on each fold. This resulted in 10 model fits - one for each of the 10 sets of exemplar images. Note, that the assignment of an image to a given exemplar set is arbitrary - exemplar sets are used merely to distinguish 10 non-identical sets of images for the 80 objects, but the assignment of specific images to each fold was consistent across all reported analyses. The mean and standard errors of these R^2 values are plotted in the main figures. Permutation analyses were performed by calculating model fits for 5000 random dimensions, simply by shuffling the order of the original values of a given dimension. Note that the assumption of exchangeability is met. For each permutation iteration, the permuted dimension scores were used to estimate a model fit for each of the cross-validated prediction folds; that is, between the predictions for a single set of 80 objects, and the 80 shuffled dimension values. These model fits were then averaged across folds to create a final permutation model fit. P-value significance was determined as (C + 1) / (5000 + 1), where C is the number of permutations that exceeded the test score. For example, p = .001 is roughly equivalent to the test score being within the top 5 scores overall.

For generalization to THINGS images, a similar scheme was adopted. This entailed training a model on images from the 80 objects set, and then generating predictions for samples from the THINGS image set. Notably, the THINGS images here consist of 63 objects that were also included in the original 80 set of images. We therefore used the full set of 80 objects to train each model, but only generate predictions for 63 objects, and then compare these against the corresponding 63 values on a given dimension. This was implemented as follows.

Two loops were used on each iteration: a) An outer training loop with the original 80 objects images (i.e. 10 folds of 720 images each (9 exemplar sets of 80 objects)); b) An inner testing loop with the 63 THINGS images (i.e. 10 folds of 567 images each (9 exemplar sets of 63 THINGS images)). A model fit was calculated for each of the 100 iterations (i.e. 10 training x 10 testing iterations). Mean model fits were calculated as the average across all iterations. For error bar plotting, and to be consistent with the main analysis, a set of 10 model fits were obtained - one per training iteration - by mean-averaging across all inner loop iterations. Permutation analyses were performed in an analogous manner to before. 5000 random permuted model fits were obtained - first by obtaining them for each of the 100 iterations as described above, and then averaging to a single permuted model fit value.

To test the generalization of 80 objects to the 20 unseen objects, a similar approach was taken. That is, models were trained on the 80 objects data, and predictions were generated for the 20 unseen objects. However, because model fits using only 20 values could potentially yield unstable model fits (due to low number of datapoints), each model fit was performed for all 10 exemplars of the 20 objects (200 values).

Thus, for each iteration (10 folds in total), 720 training samples from the 80 object data were used, and predictions were generated for all 200 unseen object samples. A model fit was generated between the 200 predictions and corresponding behavioral proxy values for each dimension. As before, a final model fit was calculated as the mean fit across the 10 folds. Permutations were conducted in an analogous manner. Importantly, exchangeability was enforced by ensuring that the shuffling of behavioral proxy values was performed in a group-wise fashion - that is, the same shuffle ordering was applied to each of the 10 exemplar sets, such that different exemplars of the same object always received the same value.

Behavioral task

As reported by Almeida et al.,[20] a behavioral dimension learning task was performed with independent participants from the behavioral dimension definition task described above. Adults completed a task with three phases (N = 31; 10-11 subjects each for testing generalization of the V1, M1, and F1 dimensions with the 20 unseen objects; note that these relatively small subject counts were deemed sufficient, as the final measures were based on a the mean-average scores across subjects, and this was shown to be sufficient for similar subject counts in the original study). In the first phase, subjects completed a computerized task where they passively learned (no responses) to associate each target object (represented with a word token) with items at either extreme of a given dimension (extreme A or extreme B, respectively). For the second phase, a set of 50 training items were identified by dividing a given dimension into 10 decile bins - each containing consecutive sets of 8 items along the extent of the dimension, as determined by their dimension scores. 5 items from each of the 10 bins were used as target items during this training phase. On each trial, subjects made a binary decision as to whether the target was close to the anchor item on extreme A or B of the dimension, and received feedback for each response.

In the third phase (test phase), subjects again made binary decisions for each target item (closer to extreme A or B; no feedback given). Here, target items consisted of 20 completely unseen objects (24 trials each; distinct from the original 80 objects). Additionally, responses to the 30 untrained objects - from the original set of 80 objects, but distinct from the 50 objects used during training - were also tested for comparison (3 trials each; these were the 3 remaining items from each of the 10 bins along the extent of the dimension). These data are plotted for comparison in Figure 5B. Cumulative gaussian curves were fitted to the decile bin means:

$$y = \frac{1}{2} erfc\left(\frac{-\boldsymbol{\sigma}(x - \boldsymbol{\mu})}{\sqrt{2}}\right)$$

where y is the percent responses, x is decile bin, σ is the slope of the cumulative Gaussian, μ is the midpoint of the curve, and *erfc* is the complementary error function. The cumulative Gaussian spans from 0% to 100% (lower and upper asymptote, respectively).

Analysis of unique variance explained

We directly compared the predictive capabilities of CLIP-ViT to each comparison network by calculating the unique variance explained (e.g. see 36, for similar analysis). Unique variance of CLIP-ViT was calculated for each set of best k components as follows. First, model fits (R^2, as described from the main analysis, above) were calculated for a combined set of components from the two networks for a given comparison. For example, for the best 5 components estimate, this involved the best 5 components of each of the respective models (10 components total). Next model fits for each respective model in isolation were calculated (e.g. for best 5 components each). Finally, unique variance of CLIP-ViT was calculated as the combined model fit minus the fit of the other model - that is, all of the combined variance that the other model cannot explain. An identical approach was taken to calculate the unique variance for the other model (i.e. all variance that CLIP-ViT cannot explain).

QUANTIFICATION AND STATISTICAL ANALYSIS

Analyses were performed with Python (version 3.11) using various packages (e.g. scikit-learn, numpy). Principal component analysis, recursive feature elimination and linear regression methods were used. Results (e.g. means, SEMs) and indications of significant tests are presented in the manuscript figures. Permutation significance was determined as stated in the method details section.