Modeling Human Concepts with Subspaces in Deep Vision Models

ANNA BAVARESCO, CIMeC, Center for Mind/Brain Sciences, University of Trento, Trento, Italy and Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, Holland
NHUT TRUONG and URI HASSON, CIMeC, Center for Mind/Brain Sciences, University of Trento, Trento, Italy

Improving the modeling of human representations of everyday semantic categories, such as animals or food, can lead to better alignment between AI systems and humans. Humans are thought to represent such categories using dimensions that capture relevant variance, in this way defining the relationship between category members. In AI systems, the representational space for a category is defined by the distances between its members. Importantly, in this context, the same features are used for distance computations across all categories. In two experiments, we show that pruning a model's feature space to better align with human representations of a category selects for different model features and different subspaces for different categories. In addition, we provide a proof of concept demonstrating the relevance of these findings for evaluating the quality of images generated by AI systems.

CCS Concepts: • **Applied computing** → **Psychology**; • **Human-centered computing**;

Additional Key Words and Phrases: Pruning, Explainable AI, Subspaces, Representation, Generative AI, Similarity

ACM Reference format:
Anna Bavaresco, Nhut Truong, and Uri Hasson. 2025. Modeling Human Concepts with Subspaces in Deep Vision Models. *ACM Trans. Interact. Intell. Syst.* 15, 4, Article 25 (December 2025), 25 pages.
https://doi.org/10.1145/3768340

1 Introduction

The current generation of AI models for language and vision holds a unique status. As opposed to computational systems in which humans pre-specify the features that code for each object, these models learn the relevant features during training. Interestingly, these features appear to have psychological relevance, and one area where these models perform well is predicting human representational spaces, i.e., the degree of similarity between objects as perceived by humans. This has been repeatedly demonstrated for both words and images [for a review, see 40]. Importantly,

Anna Bavaresco—The work reported was carried out as part of A.B's Master's thesis at CIMeC, Università degli Studi di Trento, and completed at her current institution, Institute for Logic, Language and Computation, University of Amsterdam.
Authors' Contact Information: Anna Bavaresco, CIMeC, Center for Mind/Brain Sciences, University of Trento, Trento, Italy and Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, Holland; e-mail: a.bavaresco@uva.nl; Nhut Truong, CIMeC, Center for Mind/Brain Sciences, University of Trento, Trento, Italy; e-mail: leminhnhut.truong@unitn.it; Uri Hasson (corresponding author), CIMeC, Center for Mind/Brain Sciences, University of Trento, Trento, Italy; e-mail: uri.hasson@unitn.it.

ACM 2160-6463/2025/12-ART25
https://doi.org/10.1145/3768340

this was found when using object embeddings from computational models that are not trained to predict human behavior, but on traditional image-classification or word-prediction tasks. Some have taken such findings to suggest that human-like knowledge organization can emerge from statistical training alone, without negative examples, which justifies treating vision **Deep Neural Networks (DNNs)** as potential explanations of behavior [12].

The capacity to predict human perception of object similarity is fundamental to modeling human knowledge, which is strongly based on categories that have a graded-membership structure [35]. Human semantic categories can be thought of as being organized according to latent dimensions (lower-level features), which determine how category members stand in relation to each other. This also determines which category members are more or less typical. Thus, in principle, the core ability to predict object-pair similarity from AI systems opens the door to large-scale modeling of human knowledge.

Initial studies demonstrated that object embeddings extracted from deep computational models can predict human similarity and typicality ratings [e.g., 11, 19, 25, 30, 34]. Recent approaches take a more nuanced position in which these models learn representations that do not inherently approximate those of humans, but can be fine-tuned or otherwise altered to produce better alignment with human behavior. In one of these approaches, *feature reweighting* is used to learn an adjustment of the model's feature weights so that object distances computed from the model better predict human similarity judgments [e.g., 2, 18, 30]. A limitation of reweighting is its explainability: since it is implemented via regularized regression methods, it is difficult to understand what changes in representation are responsible for the improved prediction. In a different approach, *supervised pruning* [e.g., 42, 43] is applied to structurally prune parameters (nodes or entire feature maps) from the original, full model so that the pruned model outperforms the full model in predicting human representations. Because pruning identifies a subset of a feature space without changing the activation of retained features, it allows applying explainability techniques.

In this study, we build on the success of supervised pruning as an approach for modeling and predicting human knowledge of different categories. Departing from prior studies, our aim is to use pruning in order to understand whether the human representation of different categories maps onto different dimensions or subspaces in a pre-trained computational model. From the perspective of constructing intelligently interacting information systems, knowing that different content domains or concepts require focusing on subsets of the latent dimensions in the model is highly important. It can lead to more effective interactions through a more precise modeling of the user's own representational space. In the context of generative AI, this can be used not only to improve alignment with user expectations but also to intentionally introduce heterogeneity into generated content, as we discuss below.

Our study consists of two experiments. In Experiment 1, we used supervised pruning to identify a subset of the model's features that optimize the prediction of human similarity judgments for a given category. These categories included animals, fruits, vegetables, transportation, and furniture. The feature subsets retained for different categories were then analyzed using two general approaches that focus, respectively, on the model features and the basis itself.

First, we studied the similarity of selected feature sets when pruning by different categories. Though intuitive, this approach does not rule out that different feature sets still describe the same representation (basis), as information is distributed redundantly among features in vision models. Therefore, in a second approach, we directly studied the basis described by sets of retained features to understand if these different sets identify different subspaces. We found that the prediction of human representations for different concepts does indeed rely on different subspaces in the original model. We also used feature visualization to describe what types of information strongly or weakly activate different feature subsets.

Our findings from the first experiment prompted us to conduct a second experiment to examine a related question: Do comparisons between highly-similar and weakly-similar objects within a given category rely on different bases? To address this issue, we applied the same pruning and analysis approaches as Experiment 1, but split each category into a subset of high- and low-similarity objects (as defined from human ratings). Here too, the findings were highly systematic: Within each category, the model features retained to predict high- and low-similarity objects were distinct and captured different representational spaces. Experiment 2, therefore, shows that modeling human representation of a semantic category benefits from identifying specific model subspaces that best capture the dimensions relevant to that category, and that different categories select for different representational subspaces.

Finally, we demonstrate the relevance of our findings for generative AI in an additional small-scale analysis, which we intend as a proof of concept. In this analysis, we focused on generated images belonging to a specific semantic category and passed them through a network pruned against human similarity judgments from the same category as well as through a non-pruned network. We then investigated whether descriptive statistics computed from pruned embeddings could better distinguish between generated images preferred and dispreferred by human users. Our findings suggest that the human knowledge captured by pruned networks indeed makes them more sensitive to human preferences, and therefore a promising tool for evaluating generated images.

2 Related Work

2.1 Modeling Human Representation Using DNNs

Evaluating the alignment between the representational space of a computational model and that of humans is not trivial. The model's space is typically derived from the object embeddings, which produces an object-by-feature table. If the human space is also described using an object-by-feature table, the two spaces can be compared using a multitude of methods, including **Canonical Correlation Analysis (CCA)** or **Partial Least Squares (PLS)** [for review, see 21].

However, obtaining object-by-feature tables from humans—e.g., by asking them to list the features pertinent to one object by choosing them from a pre-defined set—suffers from theoretical limitations. A crucial one is that the initial features are pre-selected by the experimenter, with no guarantee that they are psychologically relevant. In other words, it is possible that two objects highly similar in the pre-defined feature space are still perceived by people as dissimilar, as participants internally represent objects based on different features than those identified by the experimenter.

To avoid this issue, human representational spaces are commonly constructed by relying on perceived object distances, rather than object features. Perceived distances are typically measured by asking participants to judge the similarity of object pairs [23] or arrange a set of objects so that pairwise object distances can be derived from the arrangement [22]. This produces a more valid representation of the psychological space for a given set of objects.

Given this constraint, the standard approach for evaluating the alignment between human and model representations involves determining to what extent the two representations produce similar clustering of objects. The intuition is that an accurate model of human cognition should position more closely objects that humans judge as similar. This approach was initially developed to study the similarity of representations produced by two neural networks with different architectures [24], and was later extended to study human-model alignment within the **Representational Similarity Analysis (RSA)** [23] framework.

RSA is applied as follows: for a set of objects, all pairwise similarity judgments are first obtained from humans. From the computational model, all pairwise object distances are computed using a similarity or distance function, such as cosine similarity or Euclidean distance between object

embeddings on a certain layer. At this point, two distance matrices are produced, sometimes referred to as **Representational Dissimilarity Matrices (RDMs)**. The two RDMs' upper triangles are vectorized and their correlation is computed to produce an R^2 value, with higher values indicating higher isomorphism between the two systems. R^2 values in the range of range of 0.2–0.6 for various image datasets have been reported [see 30].

The fact that human RDMs are already well approximated by RDMs produced from embeddings of computer-vision models trained for classification was initially taken to suggest that these models naturally produce representations similar to those of humans and can be used as in-silico models [5]. From a modeling perspective, their success suggested that in these DNNs, relevant information was distributed among the model features, consistent with the idea that information is presented via distributed, non-modular encoding.

2.2 Reweighting

Feature reweighting was developed as an alternative approach for approximating human similarity judgments from model embeddings. It assumes that the features acquired by a model are miscalibrated in salience with respect to human representation. It follows that human representations can be better predicted by learning a reweighting of the model's features. Informally, reweighting can be viewed as implementing an "equalizer" that adjusts each feature's salience.

In practice, reweighting is implemented by learning a weight that is applied to the product of the feature values in two objects. Given two objects A and B, the human-rated similarity between objects A and B, $SIM(A, B)$ is predicted as follows:

$SIM(A, B) = \sum_{i=1}^{n} w_i \cdot f_i(A) \cdot f_i(B)$; where w_i is the weight associated with the ith feature; $f_i(A)$ and $f_i(B)$ are the values of the ith feature for objects A and B respectively. n is the total number of features. The weights w_i are learned to optimize the similarity measure based on human-rated similarities.

In some applications [18, 30], reweighting is used to learn one weight per product, as indicated above, while more complex implementations learn a weight for each possible pair of features [for variants, see 2]. Reweighting strongly improves out-of-sample prediction of human similarity judgments. For example, when predicting human similarity judgments for images of transportation means, reweighting increased out-of-sample prediction accuracy R^2 from 0.51 to 0.58 [30]. Reweighting has also been applied to predicting similarity judgments for words. In this case, an average improvement of $R = 0.22$ was found across various categories [33].

Zhang et al. [45] introduced another approach for reweighting a network in order to better predict human judgments. Their **Learned Perceptual Image Patch Similarity (LPIPS)** metric consists in fine-tuning a pre-trained network's representation against a set of human similarity judgments. Here, human judgments are presented in the form of triplets (two-alternative forced choice), and the network learns to predict which pair is more similar. In contrast to feature-reweighting described above, where the loss reflects the prediction of a vector of similarity judgments on a continuous scale, LPIPS is based on a triplet loss.

2.3 Pruning

Pruning differs from reweighting in its theoretically guided assumption that not all model features will be relevant for modeling a given human concept. Instead, it operationalizes the idea that the pretrained computational model consists of meaningful subspaces, whose variance is captured by distinct sets of features. Consequently, different concepts are optimally represented through a restricted subset of features. The objective function for pruning is therefore different from reweighting: it maximizes the fit between model and human representations by retaining a subset

of the original features, but importantly, without altering the learned weights of these features. Informally, pruning identifies columns in the embedding matrix whose removal produces improved prediction of human similarity judgments.

Like reweighting, pruning is effective. For example, when predicting human similarity judgments for animal images, pruning increased out-of-sample prediction accuracy R^2 from 0.61 to 0.75, while retaining only 20% of the nodes in VGG-19's [38] penultimate layer [42]. Pruning can also be applied directly at the level of feature maps [43], which also improves out-of-sample prediction of human judgments.

In a study using verbal stimuli, Flechas et al. [8] showed that prediction of human similarity judgments for members of the birds category improved from $R = 0.20$ to $R = 0.37$, using only approximately 20% of GloVe embeddings [29]. They also found that the sets of GloVe features retained by pruning varied strongly when supervised by different concepts. While this does not prove that the supervising concepts were modeled by subspaces in the language model, it is strongly consistent with the possibility. The study also found that there was no core set of features that was consistently retained when modeling different concepts, and similar findings for verb categories were reported by Bao and Hasson [3].

2.4 Modular Structure in DNNs

While the findings from studies of pruning motivate the search for modular subspaces when modeling human conceptual knowledge, they are not unique in this regard. By now, several studies suggest that deep-learning models inherently produce a modular organization of information. An elegant demonstration is provided by Cao et al. [4]. The authors evaluated whether it is possible to optimize a language model for performance on different tasks by learning pruning masks over the pre-trained, full model. They showed they could find subnetworks that optimize performance on different language tasks. In the vision domain, it has been shown that it is possible to learn masks over an initial pretrained model so that each mask improves classification performance for a different topic-specific image set [26]. Taken together with the results of the supervised-pruning literature, such findings suggest it may be possible to find different subspaces in the model that best capture the structure of different categories.

2.5 Human Similarity and Generative AI Systems

The quality of a Generative AI system, especially in vision, is measured in part by how closely generated images resemble the real reference images (the quality of the single image, and general similarity to real-world images are other aspects). This requires an evaluation of the generative system in terms of similarity of the true-reference and generated image distributions. Several state-of-the-art methods for this evaluation are based on passing real and generated images through a DNN to extract embeddings, from which a distance metric is computed. For example, the **Fréchet Inception Distance (FID)** [16] is computed by passing the real and generated image sets through Inception-v3 [41], extracting the sets' feature-map embeddings, and computing a metric based on the divergence between the two distributions' means and covariances. Heusel et al. [16] note that this uses "the coding layer of an Inception model to obtain vision-relevant features." It is clear that in this sense, vision-related features refer to those features that are learned to achieve correct classification. However, in applications focusing on human-AI interactions, where the critical objective is to understand image similarity as perceived by humans, it is important to identify those vision features that are relevant to human cognition. In other words, it is important to know whether a certain model feature (e.g., a feature map) is cognitively relevant, but its importance for achieving the model's original objective function (e.g., classification) is less relevant. Using only a

subset of feature maps selected via human-constrained pruning would produce different means and covariance matrices, consequently resulting in different FID values. The same considerations apply when using LPIPS [45], where the distance between any two single images is computed by the distance between their embeddings as extracted from each layer of the model. We evaluate the potential of selecting cognitively relevant features for human-aligned evaluation of artificially generated images in Section 5.

Furthermore, in applied contexts, when generating diverse image variations conditioned on a base image, a user might want to select a set of images so that the set satisfies certain diversity requirements, such as having the images be sufficiently different from each other beyond a certain minimal distance threshold d_{min}. Currently, this would be achieved by obtaining the embeddings of each image from a pre-trained DNN and computing pairwise image distance using all feature maps. Here too, a more relevant result may be obtained by computing distances considering only network elements that are cognitively relevant.

Finally, supervising pruning by human data can yield AI systems that are more aligned with group-specific notions or knowledge. While this may sometimes be risky or undesirable (e.g., when the human "notion" injected into the system is a harmful bias or stereotype), fine-tuning on safe human data can help tune models to specific cultural perspectives or expert knowledge, thus leading to smoother interactions. As we show here, pruning can also produce a more interpretable feature space. Because pruning often selects small subsets of the original feature space, it is possible to apply explainability approaches to understand the semantics of the retained features, as we have recently demonstrated for both the word and image categories [3, 8, 43].

3 Experiment 1

3.1 Research Aims and Hypotheses

Our main aim in this study was to determine whether pruning a pretrained vision model to optimize prediction of human similarity judgments within specific categories selects for different information in the model's representational space. We examine this both at the levels of the retained features and at the level of the latent dimensions constituting the basis itself. With respect to the features retained, note that two visually distinct categories may still select for the exact same set of retained features, indicated here as features *a*, *b*, *c*, *d*. This can occur if for Category_1 the values of *a*, *b* are consistently higher than those of *c*, *d*, while the opposite holds for Category_2. Thus, the surface-level appearance of images cannot indicate whether two categories will select for similar feature sets during pruning. In our prior work [42], we did not investigate this question and focused on the predictive capability of supervised pruning. That work showed that in some cases, pruning against different sets of human similarity judgments produced substantially different numbers of retained features. However, given that DNNs can learn redundant coding across features, the number of retained features cannot indicate whether two sets of features select for similar or different information.

Our second goal was to determine whether the retained features select for different subspaces in the model. Because we study different image categories, where different feature sets are retained for different sets of images, it is not possible to directly compare the different embedding spaces produced by pruning, as each category contains different images. We therefore evaluated the bases formed by different retained feature sets by applying each set as a view (filter) on an independent set of images (ImageNet50K). We then used several convergent methods for studying the basis of the space, including **Principal Component Analysis (PCA)**, PLS, activation maximization and comparative clustering of the embedding space.

Our hypotheses were:

— *H1*: Applying supervised pruning to optimize the prediction of human similarity judgments within different categories would select for relatively distinct surface-level (model) feature subsets for each category.
— *H2*: The selected feature sets would reflect different subspaces in the model, though it is expected that fruit and vegetable categories may select for similar information.

3.2 Methods

3.2.1 Overview. We applied a supervised pruning procedure to different sets of images belonging to different conceptual categories. Optimizing for the fit with the human judgments, we produced six different sets of pruned embeddings, selecting nodes from the penultimate layer of six different instances of the same pre-trained neural network. We then studied the information coded in these features using a set of convergent analyses as detailed below. A formal description of the pruning algorithm and implementation is given in Appendix A (Research Methods), and below we provide the essential details.

3.2.2 Datasets and Vision Model. The images we used for the supervision of pruning and subsequent computation of feature overlap consisted of six image datasets, with each dataset containing 120 images belonging to a certain high-level category. The datasets were: ANIMALS, TRANSPORTATION, FRUITS, FURNITURE, VEGETABLES, and a VARIOUS category that merged image types from the other sets, but also included other types of images. The images and their similarity ratings were curated by Peterson et al. [30] and kindly shared by those authors. Similarity judgments were collected by presenting participants with pairs of images and asking them to rate their similarity on a scale ranging from 0 to 10 (with 10 being the highest possible similarity). All similarity judgments were collected within-dataset and the image sets did not contain overlapping stimuli, with the exception of four images from VEGETABLES that appeared in VARIOUS.

Image embeddings were derived by passing images through a VGG-19 network [38] pre-trained on ImageNet [37], and extracting the penultimate-layer activations. This represented each image as a 4,096-dimensional vector. The network was kept frozen so the same parameters were used to extract all embeddings.

3.2.3 Pruning Algorithm. We used the pruning algorithm presented in [42], which operates by maximizing the fit between a human **Representational Dissimilarity Matrix (RDM)** and one derived by DNN image embeddings. The human RDM is constructed so that every entry corresponds to the human-rated similarity between two images, whereas the entries of the DNN RDM are Pearson correlation coefficients computed over pairs of image embeddings. The degree of match between the two RDMs is computed as the squared correlation (R^2) between the upper triangles of the two RDMs. Pruning aims to identify a subset of DNN features which yields embeddings that better capture human judgments. That is, deriving embeddings from the pruned network leads to a higher R^2 than deriving them from the non-pruned network.

The pruning algorithm (see Algorithm 1), following the terminology of Guyon and Elisseeff [13], is a wrapper-based method where each feature is treated as a single variable, and its contribution to maximizing the objective function is calculated independently of all other variables. After the filter selects a subset of variables, the algorithm is based on iteratively including variables in descending order of importance and identifying the point where prediction is optimal. In prior work [42], we have already shown that this method robustly generalizes to out-of-sample data for these datasets. For this reason, in the current study, we do not reproduce this result but apply pruning to the entire dataset outside of a cross-validation context.

Algorithm 1: Pruning

Inputs:
SM_{HM}: Similarity Matrix of human similarity judgments
SM_{DNN}: Similarity Matrix of similarity estimations derived from the DNN by computing the Pearson correlation between the embeddings of two images
Step 1: Compute baseline
Compute baseline squared Pearson's correlation, $R^2(SM_{HM}, SM_{DNN})$, from the full set of features.
Step 2: Rank features
Substep 1: Remove the first feature from all the original embeddings and compute the corresponding similarity matrix SM_{DNNRED}
Substep 2: Compute the difference $D = R^2(SM_{HM}, SM_{DNN}) - R^2(SM_{HM}, SM_{DNNRED})$. Higher positive values for D indicate that the removed feature was important
Substep 3: Repeat the step above for all the possible $N - 1$ feature subsets (where $N = 4096$)
Substep 4: Rank the features based on D
Step 3: Construct pruned embeddings
Substep 1: Starting from an empty set of features, construct pruned embeddings by iteratively reinserting one feature at a time, in descending order of importance
Substep 2: Compute squared Pearson's correlation R^2 after each feature reinsertion and store the values in the array a
Substep 3: Compute the maximum of a. Its position (index) within the array delimits the set of features to be included in the pruned embeddings

3.2.4 Concordance between Retained-Feature Sets. We applied pruning independently to each dataset (ANIMALS, TRANSPORTATION, FRUITS, FURNITURE, VARIOUS, VEGETABLES), which produced six different pruned networks. We assessed the overlap between features/nodes retained in the pruned networks by computing the Dice coefficient for each pair of feature sets. Specifically, considering two sets U, V, the Dice coefficient, ranging from 0 to 1, is defined as Dice(U, V) = $(2 * |U \cap V|)/(|U| + |V|)$, with higher values indicating a greater degree of overlap. We used the Dice coefficient because it is a well-established measure of overlap between two sets of features and is particularly suitable for comparing binary feature sets, as in the current case.

3.2.5 Activation Maximization Using Independent Data. To study the information captured in each retained feature set, we conducted an analysis using an independent set of 50,000 images from ImageNet's validation set. We first extracted the embeddings for these images. Then we applied the six feature subsets retained by pruning to these embeddings, in each case limiting our analyses to the features selected through pruning (separately for each dataset; see Figure 1). This produces six different versions of the embeddings of the 50,000 images, each using a distinct subset of features. We refer to these as $Variants50K_{1:6}$.

We first performed an activation-maximization analysis to see which types of images would maximally activate a retained feature set. This analysis is based on the idea that a higher cross-feature sum for an image indicates it contains prominent features that the network has learned to recognize. In our particular application, we were interested in identifying images that very strongly activate each feature set, or conversely, very weakly activate the set. Rather than synthesizing images, we identified those images that produced such effects. Specifically, we used the embeddings of the 50K images in each dataset $Variants50K_{1:6}$ as follows: for each image, we summed the feature activation values over the retained features to identify the top-5 most activating and top-5 least activating images. For example, for a sample variant with 50,000 images and retained feature set A, the strongest activating image would be the one satisfying: $\arg\max_x \sum_{i=1}^{A} x_i$ where x_i is activation of feature i in image x and A is the total number of features.

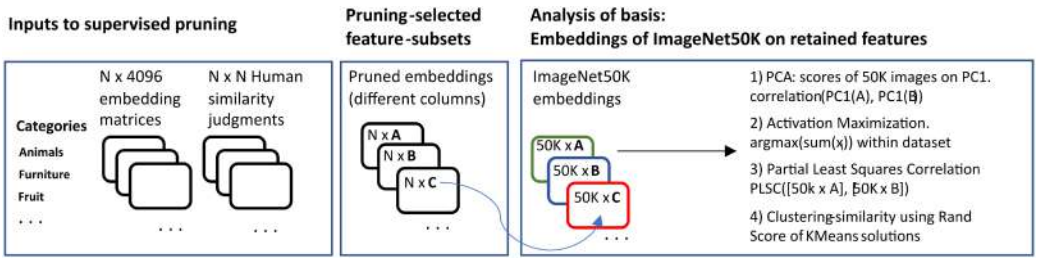

Fig. 1. Workflow used in Experiment 1. The first step consisted of pruning different [$N \times 4{,}096$] image-embedding matrices based on pairwise similarity judgments obtained from humans for those images. This produced different pruning-selected feature subsets. A new set of 50,000 images was then represented using activations on the retained feature sets. The similarity of the representations in these matrices was then assessed using convergent analyses.

3.2.6 Feature Extraction from Independent Data. To understand whether the retained feature sets capture similar latent dimensions, we performed a feature-extraction analysis by applying PCA to $Variants50K_{1:6}$. For each of the six, we focused on the scores of each of the 50K images along the **First Principal Component (PC1)**. After extracting PC1 scores, we quantified the correlation between these score vectors across the six different variants. When applying PCA, we used the scikit-learn [28] implementation, which constrains the left matrix (U) so that its largest coefficient in absolute value is positive. Higher correlations between PC1 score vectors would suggest that a similar primary dimension underlies the features selected.

*3.2.7 Quantifying Alignment Using **Partial Least Squares Correlation (PLSC)**.* While the PCA analysis identified the primary latent factor within each variant of $Variants50K_{1:6}$, PLSC is a dimensionality reduction technique that jointly considers two datasets to identify latent factors that maximize their covariance. It produces a correlation score representing the proportion of covariance explained by a user-specified number of latent dimensions in the PLSC analysis.

We applied the PLS Canonical algorithm (PLS correlation), implemented in scikit-learn, to all pairs of the six variants. This assessed the degree of similarity in the underlying representations between any two sets of retained features (higher scores indicate higher similarity). We repeated the analysis varying the number of latent components requested over 5, 10, 15, 20, 25, and 30.

We did not apply CCA for evaluating the match between the different sets because, in our particular application, it is very likely that different datasets will contain several columns (features) that are identical. CCA would probably identify these identical columns in both datasets and assign them as the first canonical variates.

3.2.8 Clustering of Independent Data Using Retained Features. In this analysis, we clustered the 50,000 images in the independent ImageNet dataset based on the different retained feature sets. Subsequently, we evaluated the similarity of the clustering solutions using the Rand Score [32] and Adjusted Rand Index provided by scikit-learn. If two retained feature sets cluster the images similarly, this suggests that they position objects in relatively similar arrangements in both sets. We used k-means clustering and, as an initial step, determined the optimal number of clusters for each matrix. This was achieved by computing the Silhouette Coefficient [36] for varying numbers of clusters ranging from 2 to 10. Next, we calculated the similarity of each pair of clustering solutions.

3.3 Results

3.3.1 Feature Overlaps between Retained Sets. We used human similarity judgments to supervise the pruning of VGG-19's penultimate layer. As indicated in Table 1, in all cases pruning selected

Table 1. The Impact of Pruning on Prediction of Human Similarity Judgments (R^2 Values)

	Animals	Transportation	Fruits	Furniture	Various	Vegetables
Baseline	0.34	0.29	0.08	0.07	0.19	0.10
Pruned	0.60	0.48	0.19	0.19	0.49	0.22
Features	869	743	626	703	928	543
Baseline	0.34	0.29	0.08	0.07	0.19	0.10
Pruned	0.60	0.48	0.19	0.19	0.49	0.22
Features	869	743	626	703	928	543

Baseline: Prediction using all features. Pruned: Prediction using only features identified by pruning. Features: number of features retained by pruning of the full set of 4,096 features.

Table 2. Similarities between Pruned Feature Sets in the Validation Study: Dice Coefficient between Pruned Feature Sets, and (in Parentheses), the Pairwise Correlation between the Scores of ImageNet 50K on the PC1 of Each Feature Set

	Animals	Transportation	Fruits	Furniture	Various	Vegetables
Animals		0.19 (0.46)	0.20 (0.69)	0.15 (−0.62)	0.25 (0.66)	0.16 (0.63)
Transportation	-		0.13 (0.14)	0.19 (−0.30)	0.25 (0.13)	0.15 (0.18)
Fruits	-	-		0.14 (−0.51)	0.19 (0.81)	0.28 (0.90)
Furniture	-	-	-		0.20 (−0.58)	0.15 (−0.44)
Various	-	-	-	-		0.18 (0.75)

fewer than 25% of the total number (N = 4,096) of features in the layer, with several cases where pruning explained twice the variance as the full model.

The data on feature overlap (Table 2; Dice data) reveals several important patterns. First, it validates the construct validity behind pruning in that it shows that when pruning is applied to categories that are prima facie more similar, the retained feature sets show a stronger overlap. The strongest overlap was found between feature sets selected for FRUITS and VEGETABLES. Additionally, the feature overlaps with VARIOUS tended to be higher compared to other categories, consistent with the fact that this category was constructed by aggregating a mixture of images belonging to the other five categories.

With respect to the magnitude of the Dice coefficients (excluding VARIOUS), these were relatively low, mostly in the range of 0.13–0.18. Nonetheless, assuming independence of feature sets, the theoretical expectation assuming random sampling of features is between 0.04 and 0.1. The fact that the maximum value expected by chance was consistently exceeded indicates an above-chance similarity between the pruned feature sets. An important contributor to these above-chance coefficient values is the fact that approximately 37% of the 4,096 features were never included in any of the retained feature sets, as detailed below.

3.3.2 Feature Distribution across Retained Sets. Though the magnitude of Dice coefficients was modest, there remains the possibility that a core set of features was consistently retained across the different pruned sets. This would indicate a potential core set of semantics that are common to different categories. To evaluate this, for each feature we computed its frequency of occurrence within the pruned sets, excluding VARIOUS from the analysis due to its mixed nature. This assigned each feature a value between 0 and 5. We found that approximately 1,500 of the 4,096 features were not selected in any of the pruned sets. Only one feature was consistently selected in all 5 sets, with

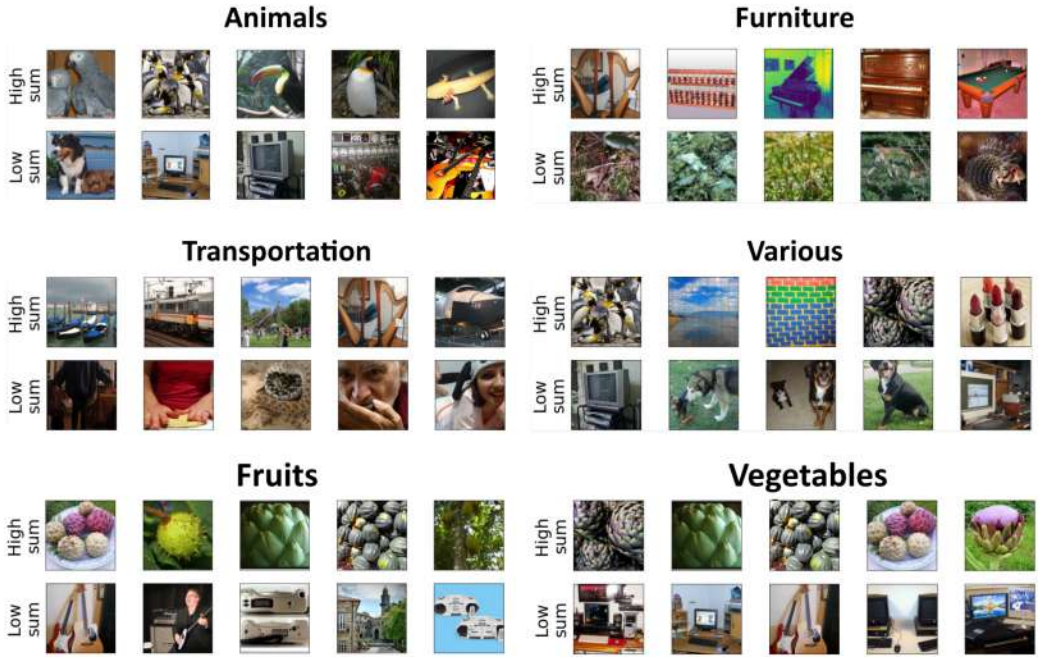

Fig. 2. Visualization of image properties that maximized feature sets retained by pruning, per category. Images identified were from an independent dataset and selected by identifying the five images that maximized (High Sum) or minimized (Low Sum) the activation for different sets of features retained by pruning.

25 features appearing in four sets, 172 in three sets, 713 in two sets, and 1,622 in just one set. Thus, the concept of a core feature set is not supported. Furthermore, a large proportion of features were apparently not relevant to modeling psychological similarity for any of the six datasets.

3.3.3 Bases of Retained Features. We used an independent dataset, which we represented using the different sets of retained features to understand the information they capture. In the first analysis, we identified the images in this dataset that most strongly (and most weakly) activated each set of the feature sets retained by pruning (see Section 3.2.5). These images are presented in Figure 2 (upper rows). For example, we find that the set of features pruned from FRUITS selects for fruits, whereas the set pruned from ANIMALS selects for animals, particularly birds. The images that least activated these sets are also shown, and included artifacts having non-natural straight, parallel lines. We also find that the results for VEGETABLES were similar to those of FRUITS. In contrast, the images identified for FURNITURE features showed the opposite pattern to the natural kinds: the highest-activating images consisted of man-made objects with straight lines, often against a well-defined background, and the lowest-activating images contained animals or extensive vegetation. For TRANSPORTATION, we find that the highest-activating images contained a mixture of transportation devices and other man-made objects, whereas the lowest-activating images consisted mainly of faces or sections of the human body.

Similar conclusions about the basis are suggested by the cross-set correlations between the 1st PC score vectors, presented in Table 2 in parentheses next to each Dice coefficient. The strongest correlation was found between VEGETABLES and FRUITS, consistent with the relatively high Dice coefficient. We further observed correlations with FURNITURE to be low and negative. For TRANSPORTATION, correlations were also low. The PCA analysis additionally shows that a larger Dice

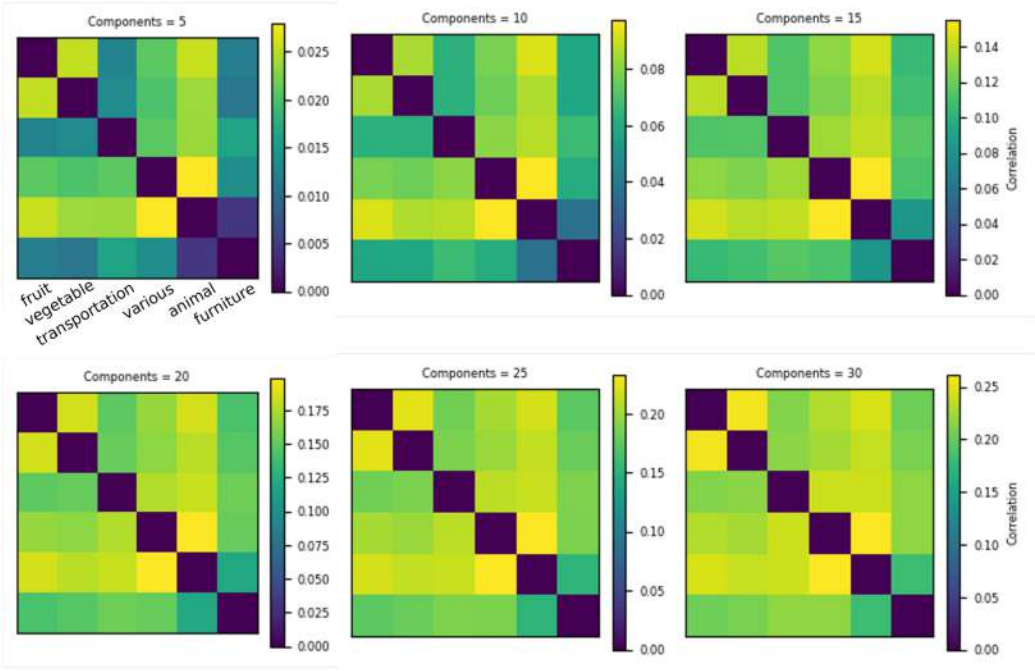

Fig. 3. PLSC between embeddings of ImageNet50K on features retained via pruning. Relatively strong correlations were consistently found between the embedding of the independent images on features retained for Fruits and Vegetables. In contrast, embedding those images on features retained from pruning against similarity judgments for furniture produced a relatively distinct representation with low correlations with others.

coefficient between the features selected by pruning does not necessarily indicate a better match between the latent dimensions they extracted (from the independent dataset). For example, VARIOUS had a higher Dice coefficient with TRANSPORTATION than with FRUITS, but the PCA correlations presented the opposite pattern, showing a higher correlation for the latter.

The results of the PLSC converge with the previous ones, indicating a strong alignment between the FRUITS and VEGETABLE datasets (see Figure 3). The other strong overlap was found between ANIMALS and VARIOUS, which is less interpretable as it may depend on the specific proportion of animal images in the (more heterogeneous) various set. We again find that the features retained for FURNITURE capture a representational structure that differs strongly from the other categories, producing low correlations even when using as many as 30 latent components.

The analysis of the clustering solutions further converges with the PCA analysis. First, we find that the features retained from FRUITS and VEGETABLES produced highly similar clustering solutions (see Figure 4(b)). The matches with FURNITURE were uniformly low and did not differ from chance as indicated by the Adjusted Rand Index. This is also consistent with the results of the PCA analysis.

3.4 Discussion

The study highlights several important findings. First, pruning improved the match between object distances in DNN feature space and human similarity judgments. Second, pruning selected relatively non-overlapping feature sets. A striking result is that only one of the 4,096 features was consistently

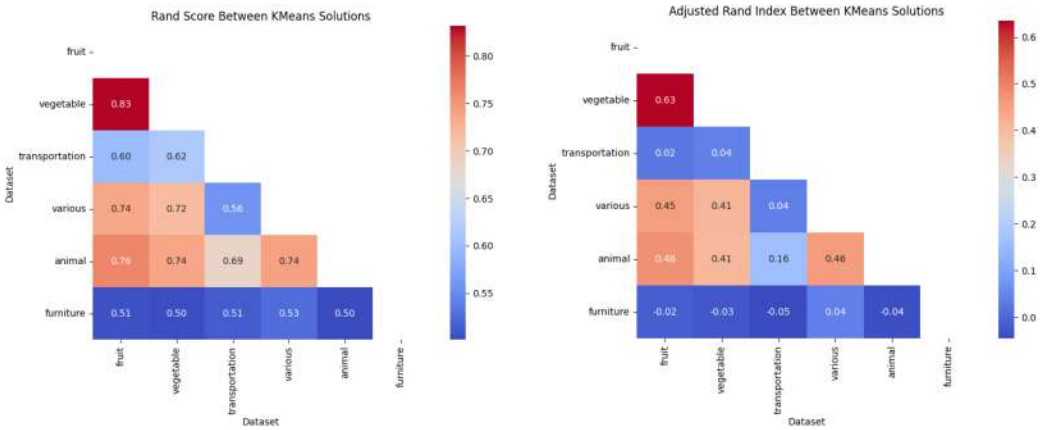

Fig. 4. Similarity of clustering solutions for 50,000 images that were numerically coded on different sets of features retained via pruning (of independent datasets). The left panel presents the Rand Score, and the right panel its adjusted score against chance value.

selected across the retained sets. The analysis of the basis of the feature spaces confirmed that different pruning solutions selected for different information in the computational model. The results of activation maximization, pairwise correlations between PC1 scores, PLS correlation, and similarity between clustering solutions often indicated very low similarity between representational spaces, while confirming that such a similarity was consistently identified for FRUITS and VEGETABLES. In summary, the study confirmed both our hypotheses: Pruning supervised by human representations of different categories can select for different features of the computational model, and importantly, these features identify different variance components in the model's representational space.

4 Experiment 2

4.1 Research Goals and Hypotheses

Experiment 1 demonstrated that learning a human representational space via supervised pruning produces different basis vectors (feature sets) for different categories. These are furthermore associated with different subspaces in the model as indicated by relatively low shared variance (apart from the higher overlap between FRUITS and VEGETABLES).

In the current study, we evaluated whether a similar subspace selection occurs, but at the level of single categories. The six categories we used were broad, superordinate level categories, meaning that while they are highly distinct from each other (producing the effects in Experiment 1), they also subsume a large range of objects that may differ substantially from each other (see Figure B1 in Appendix B for examples of original materials). This allowed us to investigate whether, within each category, comparisons between more similar images select for different dimensions than comparisons between less similar images. Consider, for example, comparing a gorilla to a zebra or to another primate. The former comparison may be based on general body form or environmental features, whereas the latter on more specific dimensions on which the primates are aligned, such as hairiness or facial characteristics.

From the perspective of human-oriented AI, it is important to understand whether human knowledge can be generally approximated from a similarity space defined at the superordinate level or requires a further partitioning into high- vs. low-similarity items within the superordinate category. This distinction is useful for more precise user modeling, and comes at no additional cost as the split into high- and low-similarity objects is based on already collected data.

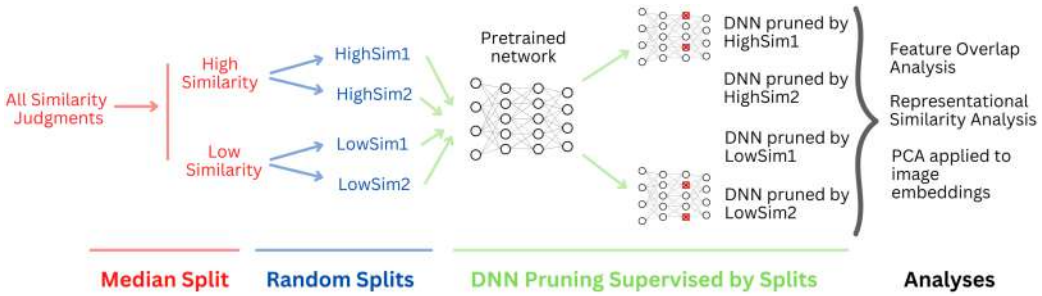

Fig. 5. Overview of analysis workflow in Experiment 2. Human similarity judgments for 120 category members were divided by median split into high-similarity and low-similarity object pairs. These median splits were further randomly divided into two partitions each, producing four non-overlapping splits (partitions) of the initial set of similarity judgments. Each partition was used to supervise the pruning of a pre-trained DNN by optimizing the prediction of the similarity judgments assigned to that partition; this procedure produced four pruned DNNs. The embedding space of these DNNs (pruned nodes marked with red × in figure) was then evaluated in a series of analyses. This entire workflow was applied to six different datasets.

Our hypothesis was:

— *H3*: More- and less-similar objects within a superordinate category are evaluated based on different cognitive dimensions, and select for different subspaces in the computational model.

As detailed below, the main approach to test this hypothesis is evaluating whether pruning by high- and low-similarity judgments produces feature sets that systematically cluster objects in different ways.

4.2 Methods

4.2.1 Overview. Figure 5 presents the overview of the method.

4.2.2 Additional Datasets Used for RSA Applied to Independent Datasets. To study the representational spaces of the pruned networks supervised by high- and low-similarity judgments, we used RSA, as described in Section 4.2.4. For these analyses, we constructed four new independent image datasets approximately matching the content of ANIMALS, FRUITS, FURNITURE and VEGETABLES. We chose not to create new datasets for VARIOUS and TRANSPORTATION because TRANSPORTATION was an exception to the findings of the initial analysis based on feature overlap, and VARIOUS included images of the same type as the other categories, including transportation. The four independent datasets, which we differentiate using a subscript "s," were constructed as follows:

— ANIMALS$_s$ (398 images) was created from ImageNet's validation set [37] by considering only the 398 animal categories and selecting, for each of those, only the best exemplar. We defined each category's exemplar as the image producing the strongest post-softmax confidence in its respective category, following Lake et al. [25].

— FRUITS$_s$ (58 images) was constructed from THINGS [14] by selecting the most and least typical images from 29 fruit categories. We used THINGS instead of ImageNet as the latter only contains 11 fruit categories. Given that our network was pre-trained on ImageNet, it did not have output nodes corresponding to those THINGS labels. We therefore selected the most/least typical images by passing all the images through a VGG-19 network [38], and selecting the image in each category with the lowest/highest mean distance to all other images in the category as computed from the correlation distances between embeddings in the penultimate layer. We chose to include both the most and least typical images because 29 objects would

not have been sufficient to populate an RDM with sufficient variance. Similar considerations also motivated our criteria for creating the two remaining independent datasets.

— FURNITURE$_s$ (56 images) was derived by selecting the most and least typical images from 28 THINGS furniture categories.

— VEGETABLES$_s$ (58 images) was derived by selecting the most and least typical images from 29 THINGS vegetable categories.

4.2.3 Partitions and Feature Overlap. For the same datasets of N items used in Experiment 1, we considered the set of $(N^2 - N)/2$ (upper triangle) unique similarity judgments. Those judgments were divided into the different partitions used to supervise pruning as follows. We first used a median to split the judgments into high-similarity and low-similarity sets. We then further randomly split the HighSim comparison set into two parts, as well as the LowSim set, ultimately producing four partitions of the original similarity matrix: $HighSim_1$, $HighSim_2$, $LowSim_1$, and $LowSim_2$. A summary of the entire workflow is provided in Figure 5.

We identified the DNN image embeddings corresponding to each partition, and then applied pruning separately for each partition. Note that, differently from Experiment 1, here we obtained four different pruned networks per dataset instead of one. This allowed us to compute Dice coefficients between the feature sets obtained from the *same* dataset. To ensure that the Dice values were reliable estimates and not a random product of a chance partition split, we repeated the experiment 100 times. More specifically, while the median value determining the initial High vs. Low split remained the same, the further random split of High and Low judgments into two partitions was repeated in 100 different ways. Here, our prediction was that, for each dataset, the average Dice($HighSim_1$, $HighSim_2$) and Dice($LowSim_1$, $LowSim_2$) values would be higher than the average Dice values for the remaining feature set combinations, i.e. Dice($LowSim_1$, $HighSim_1$), Dice($LowSim_1$, $HighSim_2$), Dice($LowSim_2$, $HighSim_1$) and Dice($LowSim_2$, $HighSim_2$).

Given the findings of Experiment 1, we expected that the overlap between feature sets identified by pruning from low- and high-similarity judgments within a category would still be higher than that found across categories. For this reason, we repeated the analysis performed in Experiment 1, evaluating the consistency of pruned feature sets across categories. The only difference was that here we computed those data separately for sets pruned by low similarity judgments and sets pruned by high similarity judgments.

4.2.4 RSA. In addition, we used RSA [23] to study the representational space of the DNNs pruned by similarity judgments. Our aim was to compare the representational geometries resulting from DNNs pruned against Low1, Low2, High1, and High2 similarity judgments. To do this, we derived the pruned networks from the initial image datasets and then passed the new independent sets of images (described in Section 4.2.2) through the four pruned networks.

To clarify, the procedure applied to each image category, for example ANIMALS, was the following. Four different networks were obtained by pruning VGG-19 against the four partitions of human similarity judgments relative to the dataset ANIMALS already used in Experiment 1, whose properties are described in Section 3.2.2. Then, the images from an independent dataset ANIMALS$_s$, as described in Section 4.2.2, were passed through each of the four pruned networks. This produced four different RDMs, one for each of the four networks. We then calculated the **Second-Order Isomorphism (2OI)** between any two of these RDMs, which provides a measure for the overall alignment of the representational geometries. Thus, different datasets are used to supervise pruning and to construct the RDMs.

Our predictions were as follows. If people base low-similarity judgments on different information than that used for high-similarity judgments, then DNNs pruned by low-similarity judgments should select for certain dimensions, while DNNs pruned by high-similarity judgments should

Table 3. Pruning by High-Similarity or Low-Similarity Judgments Selects Different Feature Sets

Comparison	Animals	Transportation	Fruits	Furniture	Various	Vegetables
1. Low1, High1	0.22 (0.01)	0.28 (0.04)	0.28 (0.03)	0.34 (0.03)	0.21 (0.03)	0.36 (0.03)
2. Low1, High2	0.22 (0.01)	0.28 (0.04)	0.29 (0.03)	0.34 (0.03)	0.21 (0.02)	0.37 (0.03)
3. Low2, High1	0.22 (0.01)	0.28 (0.03)	0.29 (0.03)	0.35 (0.04)	0.21 (0.02)	0.36 (0.02)
4. Low2, High2	0.22 (0.01)	0.27 (0.04)	0.29 (0.03)	0.35 (0.03)	0.21 (0.02)	0.37 (0.03)
Lows vs. Highs (Average of Rows 1:4)	0.22	0.28	0.29	0.35	0.21	0.37
5. Low1, Low2	0.71 (0.04)	0.24 (0.03)	0.48 (0.03)	0.48 (0.03)	0.29 (0.03)	0.52 (0.03)
6. High1, High2	0.69 (0.02)	0.61 (0.03)	0.60 (0.06)	0.65 (0.04)	0.61 (0.01)	0.59 (0.04)

Average Dice coefficient for feature overlap between all the possible combinations of the four partitions (Low1, Low2, High1, and High2), with standard deviations in brackets. Dice values were higher when computed for networks whose pruning was supervised by the same levels of similarity judgment.

select for others. This would then be reflected in the 2OI magnitudes, which we expected to be highest when the two RDMs were derived from networks pruned by low-similarity judgments or when the two RDMs were derived from networks pruned by high-similarity judgments, and lower in the other four cases where one RDM was produced by a network pruned by low-similarity judgments and the other from a network pruned by high-similarity judgments.

4.3 Results

4.3.1 Feature Overlap. The feature-overlap results are reported in Table 3. They provide strong support for our hypothesis that different informational dimensions support the comparison of high-similarity and low-similarity natural images. For five of the six datasets, the Dice coefficient computed for high/high and low/low exceeded that found for low/high combinations. For the TRANSPORTATION dataset, this was found for high/high but not for low/low.[1] That is, the predicted pattern held in 11 of the 12 cases. The strongest result was found for ANIMALS.

To determine the statistical significance of the results reported in Table 3, we conducted eight two-sample *t*-tests for each dataset. Specifically, we tested if the values presented in row 5 were greater than those in rows 1–4 (four comparisons per category), and did the same for row 6 (four additional comparisons). With the single exception of one comparison for TRANSPORTATION, all the other differences were statistically significant, Bonferroni-corrected for 48 tests ($p < 0.001$ for each test). This indicates that the Dice coefficients produced for the low/high pruning combinations (rows 1–4) were significantly lower than those computed for the low/low and high/high combinations (rows 5–6). As Table 3 further shows, the Dice coefficient for the low/low solutions was higher than the low/high cases, with an effect size of approximately 10 standard deviations.

That said, we also observed that the Dice overlap between features selected by supervision from low-similarity judgments tended to be lower than that found when supervising by high-similarity judgments, ranging between 0.24 and 0.71. This could be explained by the recent finding [1] that human similarity judgments for low-similarity objects are noisier than those for high-similarity objects. Importantly though, the Dice coefficients for low/low in row 5 of Table 3 did not reflect a chance overlap or absence of shared dimensions. This can be seen in the fact that the magnitudes of these coefficients always exceeded those computed when determining the overlap between

[1]This dataset primarily consisted of automobiles of different types, but also included other forms of transportation and means of movement, including, e.g., bicycles, wheelbarrows, wheelchairs, skates, and blimps. For this reason, the high-similarity partitions consisted mainly of automobiles, but the low-similarity ones consisted of comparisons between items drawn from diverse categories. This could produce low consistency in features selected for $LowSim_1$ and $LowSim_2$.

Table 4. Low Similarity Judgments Select Different Features for the Different Datasets, and the Same Holds for High-Similarity Judgments

	Animals	Transportation	Fruits	Furniture	Various	Vegetables
(Low1 ∪ Low2) vs. Other Dataset Lows	0.13	0.13	0.18	0.12	0.11	0.13
(High1 ∪ High2) vs. Other Dataset Highs	0.17	0.19	0.16	0.17	0.18	0.18

Values reported indicate Dice-coefficient values.

the retained feature sets *across* datasets (Table 4). For example, for TRANSPORTATION the Dice coefficient between features selected by pruning from (separate sets of) low-similarity judgments was $M = 0.24 \pm 0.03$, but the overlap between those same features and features selected in the same manner for the other datasets was only 0.13. Thus, pruning by low-similarity judgments produced feature sets that are meaningfully consistent within dataset. As we show below, this conclusion is also strongly supported by Representational-Similarity Analysis.

4.3.2 Representational Space: 2OI. As detailed in Section 4.2.2 of the Methods, we studied the representational spaces of the DNNs pruned by $LowSim_1$, $LowSim_2$, $HighSim_1$ and $HighSim_2$ for ANIMALS, FRUITS, FURNITURE and VEGETABLES datasets. Recall that each of the four resulting RDMs reflects a different set of pairwise distances estimated from the DNN embeddings, without overlap. As detailed in the Methods section, after pruning, we passed new *independent* datasets through these pruned networks, and extracted the image embeddings for these images from each of the four pruned DNNs, obtaining four RDMs per dataset. We then quantified the 2OI between these RDMs using RSA, where isomorphism was defined as the squared Pearson's correlation between the upper triangles of any two RDMs.

As Figure 6(d) shows, we found a consistent pattern for all four datasets. Specifically, the two RDMs for $LowSim_1$ and $LowSim_2$ were highly correlated, as were the two RDMs for $HighSim_1$ and $HighSim_2$. This means that high-similarity judgments produced pruned DNNs with well-aligned representational spaces, as did low-similarity judgments. In contrast, the representational spaces of DNNs pruned from mismatching levels of similarity judgments presented a lower match (i.e., lower 2OI). Pruning by low- and high-similarity judgments, therefore, not only systematically selects different features, but produces representational spaces that are more consistent within each level of similarity than across levels of similarity.

This pattern of results is highly unlikely to be found by chance. The probability of the High/High and Low/Low cases having the highest correlation values in a single dataset is $1/15 = 0.066$, and the probability of this occurring independently in all four datasets is extremely low. Furthermore, we calculated the statistical significance for the differences between correlations [7, 39] to determine, within dataset, whether the High/High and Low/Low cases differed significantly from the High/Low combinations. The results were highly significant in all cases ($p < .001$). Finally, for the ANIMALS_s dataset we conducted an additional evaluation where, instead of selecting the image within each category that elicited the highest post-softmax activity, we chose the image that ranked second. The results replicated exactly.

4.4 Discussion

The findings of the second study confirmed the hypothesis that within large superordinate categories of the sort used here, more-similar and less-similar objects will select for different features in the

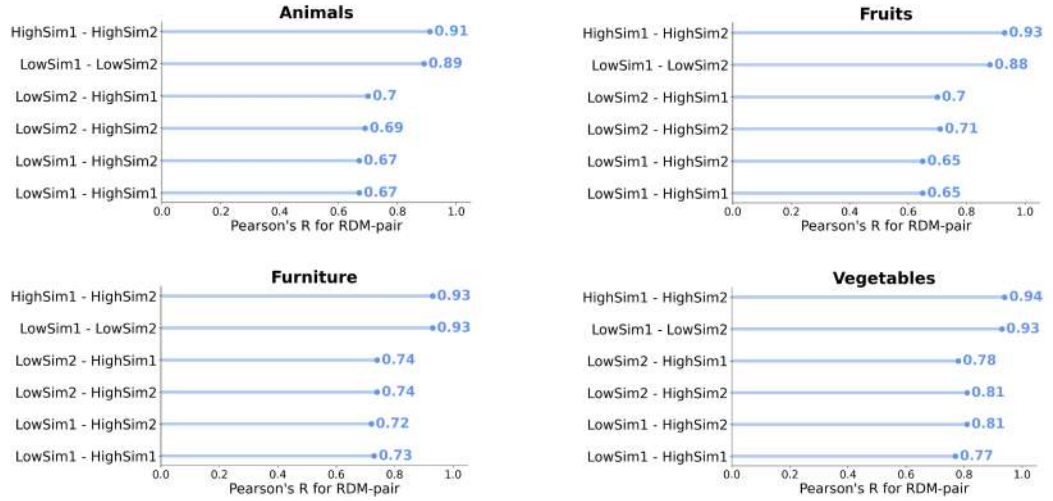

Fig. 6. Pruning produces different representational spaces when supervised by high- or low-similarity judgments. We computed 2OI (Pearson's R) from representational dissimilarity matrices (RDMs) produced by pruned DNNs. Isomorphism was higher when the two RDMs were produced from pruning supervised by low-similarity judgments, or produced from pruning supervised by high-similarity judgments. Isomorphism was weaker in all other cases.

model, which in turn reflect different latent dimensions. Within each category, we found stronger concordance between retained features when computed from two partitions of low-similarity items or two partitions of high-similarity items, as compared to when one partition was pruned from high-similarity and the other from low-similarity judgments (Table 3). Equally important, the features selected by pruning against high- and low-similarity judgments identified different variance components in the representational space. This was demonstrated by the fact that representational similarity was consistently higher when computed from object-by-object dissimilarity matrices (RDMs) where both RDMs were produced from pruning against low similarity judgments, or where both RDMs were produced from pruning against high similarity judgments.

5 Implications for Generative AI Systems: A Proof of Concept

To assess whether the human knowledge captured by pruned models can be relevant for generative AI systems, we conducted a small study on images generated by text-to-image models. Despite the remarkable generation abilities of these models, developing quantitative metrics to assess the quality of their outputs remains challenging.

In this proof of concept, we asked whether networks pruned against human judgments can serve as a tool to systematically compare generated images. More specifically, we analyzed the outputs of a pruned and a non-pruned network using three image sets: one of real images, one of generated images preferred by human users, and one of dispreferred generated images. Importantly, images from all sets belonged to the same semantic category of the similarity judgments used to prune the network.

5.1 Methods

The preferred and dispreferred image sets analyzed in this proof of concept were constructed from the Pick-a-Pic dataset [20], which collects human preferences for images generated by text-to-image

diffusion models. Each dataset instance consists of a textual prompt, two generated images, and an annotation indicating the preferred image or a tie. To ensure that each image appeared in only one comparison, we focused on the "test-unique" and "validation-unique" dataset splits, each containing 500 images.

We then identified images of animals with a string-matching procedure, leveraging the prompts used to generate the images. More specifically, we first identified the images whose prompt contained at least one animal name from a pre-compiled list (see Appendix A.3) and then manually inspected the retrieved images, excluding the non-pertinent ones. The selection process resulted in two subsets: one of 89 preferred animal images and the other of 89 dispreferred animal images. We focused on Animal images due to the limited availability of images from other categories in the Pick-a-Pic dataset, and because we have already shown that a DNN can be effectively pruned on such judgments.

Our analysis was structured in 1,000 simulations. In each simulation, we chose 60 real images of animals from [30] (Peterson_{60}), 60 Pick-a-Pic preferred animal images (Pref_{60}), and 60 Pick-a-Pic dispreferred animal images (Dispref_{60}). Embeddings were obtained from all images to compute FID values. Specifically, we computed for each pair: $FID(\text{Peterson}_{60}, \text{Pref}_{60})$ and $FID(\text{Peterson}_{60}, \text{Dispref}_{60})$.

We then counted how many of the 1,000 simulations satisfied the condition where the FID for preferred images was lower than that for dispreferred images. The total count is therefore given by:

$$\text{Count} = \sum_{i=1}^{1,000} 1(\cdot)\big(FID(\text{Peterson}_{60},\ \text{Pref}_{60}) < FID(\text{Peterson}_{60}, \text{Dispref}_{60})\big).$$

Here, i indexes the simulation and $1(\cdot)$ is the indicator function, which equals 1 if the condition is met. Count, therefore, indicates how often the preferred images are closer to the real-image reference set.

This workflow was applied to both a full (non-pruned) Inception-V3 network [41] and an Inception-V3 network pruned against human similarity judgments on animal images. If pruning effectively captures psychologically relevant features, the count of simulations satisfying the FID criterion should be higher for embeddings derived from the pruned network.

5.2 Results

For the full network, 603 simulations satisfied the condition, while, for the pruned network, the count increased to 650. Both counts deviated significantly from chance ($N = 500$ under $\text{Bin}(1{,}000, 0.5)$). Most importantly, the difference between 650 and 603 was highly significant ($p < .001$ under a binomial test), whether assuming no prior knowledge ($\text{Bin}(1{,}000, 0.5)$ as the sampling distribution) or when evaluated post hoc against the full network's observed distribution ($\text{Bin}(1{,}000, 0.603)$). These results show that computing FID on the cognitively relevant features selected by pruning results in evaluations of generated images that are more aligned with human preferences.

6 General Discussion

6.1 Main Findings and Implications

Our two studies indicate a practical approach to modeling human knowledge of a given category through supervised pruning of pre-trained models. As discussed in the Introduction, while object embeddings in computational models can already predict, to some extent, a human representational space, it has been shown that prediction accuracy can be improved via reweighting or pruning of feature activations.

Our primary contribution is in showing that pruning not only improves prediction but also the explainability of the model itself. We find that modeling different human category representations selects for different features in the model. In Experiment 1, we found that the Dice coefficient could be as low as around 0.13 (in the simplified case of equal set sizes, this means that 90% of each set consists of unique features). Importantly, Experiment 1 shows that pruning selected a relatively small proportion of the total nodes (around 12%–25%) and there was no core set of features consistently selected. A series of analyses of the basis itself, including PCA, PLSC, activation maximization, and clustering-similarity analysis, provided converging evidence supporting the conclusion that different retained sets identified distinct subspaces in the model feature space. Experiment 2 further showed that when dealing with broad, superordinate level categories, high- and low-similarity items may be modeled using different feature subsets, reflecting different variance components.

We found that a pruned network more effectively differentiates between generative AI-produced images that are preferred or dispreferred by humans. This demonstrates that aligning the network with the human representational space can enhance performance on a key technological objective, importantly achieved without task-specific fine-tuning.

Other examples highlight the general advantages of alignment in computational workflows. For instance, Palazzo et al. [27] showed that joint embedding of image representations with human brain activity improved the generation of visual saliency maps. Fong et al. [9] demonstrated that incorporating human behavior into a classification loss function enhanced the performance of image classifiers. Similarly, Peterson et al. [31] found that injecting human uncertainty into neural-network training using soft labels increased robustness to adversarial examples and improved model calibration. Our prior work Tarigopula et al. [42] showed that alignment produced more meaningful clustering of out-of-sample data. Together, these studies indicate that alignment can offer benefits across fundamental machine-learning tasks.

Pruning can also advance objectives central to intelligent interactive systems. By prioritizing those features that are most relevant to a user, it can improve nearest-neighbor searches, which is a key component of recommendation algorithms. Pruning can also support systems where user interaction fine-tunes the embedding space. For example, Zahálka et al. [44] developed an interactive learning visual analytics system in which users assign images to custom categories while the system suggests similar content. This system compresses pre-trained DNN embeddings via binning and quantization, relying on these compressed embeddings for classification and recommendations. In such systems, pruning can be used to provide an a-priori identification of features that structure users' category representations. This can help preserve crucial information during compression to support subsequent learning steps.

Pruning further assigns importance scores to individual feature maps (in convolutional layers) or units (in fully connected layers). These scores indicate each element's contribution to aligning the model with human representational spaces. For example, we have shown in [43] that these scores indicate the image sections critical for distinguishing between visually similar images. By visually presenting which image dimensions are cognitively important, the model offers increased transparency and explainability, helping users understand decisions such as why one image was identified as the nearest neighbor of another. This can, in turn, increase user trust through improved model explainability.

In tandem, it is fundamental that alignment procedures do not compromise performance on the primary task, which can occur through loss of information or the use of joint loss functions. To address this, we propose that in production systems, human-aligned AI could operate as an auxiliary model in the context of ensemble learning. This can balance the benefits of alignment with the need for optimal task-specific performance.

6.2 Limitations

We also consider caveats and limitations. One potential weakness is that the approach relies on obtaining information about human representation by soliciting similarity judgments. Given that the number of pairwise similarity judgments required to populate a full similarity matrix scales quadratically with the number of considered images, this may prove inefficient, especially when dealing with large image sets or modeling similarity notions of single users. However, our method could easily be adapted to different types of human judgments, e.g., odd-one-out judgments, which are more reliable than pairwise ones [17] and require significantly fewer annotations [15].

Another potential weakness is that pruning, which is based on sequential feature selection, may be less flexible than state-of-the-art regression methods. While we have found this is not necessarily the case [42], there may be contexts where reweighting representations with the regression-based approach described in the Introduction produces a better prediction. However, it is unclear how such regression weights (which modify products of feature values rather than the feature activation itself) can be integrated into downstream applications or analyses. In particular, it is unclear how they inform the question of whether different human categories are explained by different subspaces in a model. This is not to suggest that the question is unattainable, but pruning already offers a relatively straightforward path for studying model dimensions that contain information relevant to human comparisons.

Finally, we caution against interpreting the findings as indicating that the features selected via pruning have a direct analogue in human perception, or that the latent dimensions identified have such corollaries. Because a node activation can be expressed as a linear combination of two or more other nodes, different models can produce the same covariance matrices from which we study the basis of the feature space. It is therefore not useful to assume a one-to-one correspondence between psychological features and DNN features. Furthermore, a separate issue is that pruning may select for cognitively irrelevant features, to the extent that these co-vary with cognitively relevant ones. For example, foreground (shape) and background (texture) information in natural scenes may be correlated, and it has been shown that DNNs are more sensitive to texture than to shape content [10]. This means that, while humans may base similarity ratings on foreground/shape features, features coding for backgrounds may be selected via pruning because of correlations between background and foreground elements.

6.3 Conclusions

Our findings support the hypotheses and logic presented in the Introduction, where we suggested that model pruning supervised by human knowledge may offer substantial benefits for generative-AI workflows, as it grounds the comparison of real and generated images in a subspace of the model that is demonstrably relevant to human perception of a given category of interest. This allows embedding real and generated images from such categories in representational spaces that are more similar to that of humans as compared to the single space formed when considering all model features. Furthermore, because different populations understand the world in different ways, it is possible to systematically apply diverse points of view as filters to the generative process itself.

References

[1] Lukas Ansteeg, Frank Leoné, and Ton Dijkstra. 2022. Characterizing the semantic and form-based similarity spaces of the mental lexicon by means of the multi-arrangement method. *Frontiers in Psychology* 13 (2022), 4975.

[2] Ioanna Maria Attarian, Brett D. Roads, and Michael Curtis Mozer. 2020. Transforming neural network visual representations to predict human judgments of similarity. In *Proceedings of the NeurIPS 2020 Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM).*

[3] Wanqian Bao and Uri Hasson. 2024. Identifying and interpreting non-aligned human conceptual representations using language modeling. In *Proceedings of the ICLR 2024 Workshop on Representational Alignment*. Retrieved from https://openreview.net/forum?id=vh56Q8wbGK

[4] Steven Cao, Victor Sanh, and Alexander Rush. 2021. Low-complexity probing via finding subnetworks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 960–966. DOI: https://doi.org/10.18653/v1/2021.naacl-main.74

[5] Radoslaw M. Cichy and Daniel Kaiser. 2019. Deep neural networks as scientific models. *Trends in Cognitive Sciences* 23, 4 (Apr. 2019), 305–317. DOI: https://doi.org/10.1016/j.tics.2019.01.009

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. DOI: https://doi.org/10.1109/CVPR.2009.5206848

[7] Ronald A. Fisher. 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* 1 (1921), 3–32.

[8] Natalia Flechas Manrique, Wanqian Bao, Aurelie Herbelot, and Uri Hasson. 2023. Enhancing interpretability using human similarity judgements to prune word embeddings. In *Proceedings of BlackboxNLP at EMNLP 2023*.

[9] Ruth C. Fong, Walter J. Scheirer, and David D. Cox. 2018. Using human brain activity to guide machine learning. *Scientific Reports* 8, 1 (2018), 5397.

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the International Conference on Learning Representations*.

[11] Iris I. A. Groen, Michelle R. Greene, Christopher Baldassano, Li Fei-Fei, Diane M. Beck, and Chris I. Baker. 2018. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife* 7 (2018), e32962.

[12] Bojana Grujičić. 2024. Deep convolutional neural networks are not mechanistic explanations of object recognition. *Synthese* 203, 1 (2024), 1–28.

[13] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3 (Mar. 2003), 1157–1182.

[14] Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. 2019. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS One* 14, 10 (Oct. 2019), 1–24. DOI: https://doi.org/10.1371/journal.pone.0223792

[15] Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. 2020. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour* 4, 11 (2020), 1173–1185.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30, Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

[17] Nicolas Jones, Armelle Brun, and Anne Boyer. 2011. Comparisons instead of ratings: Towards more stable preferences. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Vol. 1. IEEE, 451–456.

[18] Philipp Kaniuth and Martin N. Hebart. 2022. Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. *NeuroImage* 257 (2022), 119294.

[19] Marcie L. King, Iris I. A. Groen, Adam Steel, Dwight J. Kravitz, and Chris I. Baker. 2019. Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage* 197 (2019), 368–382.

[20] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 36, 36652–36663.

[21] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. 2025. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys* 57, 9 (2025), 1–52.

[22] Nikolaus Kriegeskorte and Marieke Mur. 2012. Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology* 3 (2012), 28167.

[23] Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* 2 (2008), 249.

[24] Aarre Laakso and Garrison Cottrell. 2000. Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology* 13, 1 (2000), 47–76.

[25] Brenden M. Lake, Wojciech Zaremba, Rob Fergus, and Todd M. Gureckis. 2015. Deep neural networks predict category typicality ratings for images. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 37.
[26] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. 2018. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 67–82.
[27] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. 2020. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 11 (2020), 3833–3849.
[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[29] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.
[30] Joshua C. Peterson, Joshua T. Abbott, and Thomas L. Griffiths. 2018. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science* 42, 8 (2018), 2648–2669.
[31] Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9617–9626.
[32] William M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66, 336 (1971), 846–850.
[33] Russell Richie and Sudeep Bhatia. 2021. Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive Science* 45, 8 (2021), e13030. DOI: https://doi.org/10.1111/cogs.13030
[34] Brett D. Roads and Bradley C. Love. 2021. Enriching ImageNet with human similarity judgments and psychological embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3547–3557.
[35] Eleanor Rosch. 1978. Principles of categorization. In *Cognition and Categorization*. Eleanor Rosch and Barbara B. Lloyd (Eds.), Lawrence Erlbaum, Hillsdale, NJ, 27–48.
[36] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20 (1987), 53–65.
[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252. DOI: https://doi.org/10.1007/s11263-015-0816-y
[38] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings.
[39] D. S. Soper. 2023. Significance of the Difference between Two Correlations Calculator [Software]. Retrieved from https://www.danielsoper.com/statcalc
[40] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Jascha Achterberg, Joshua B. Tenenbaum, et al. 2023. Getting aligned on representational alignment. arXiv:2310.13018. Retrieved from https://arxiv.org/abs/2310.13018
[41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.
[42] Priya Tarigopula, Scott Laurence Fairhall, Anna Bavaresco, Nhut Truong, and Uri Hasson. 2023. Improved prediction of behavioral and neural similarity spaces using pruned DNNs. *Neural Networks* 168 (2023), 89–104.
[43] Nhut Truong, Dario Pesenti, and Uri Hasson. 2025. Explaining human comparisons using alignment-importance heatmaps. *Computational Brain & Behavior* 8 (2025), 421–441.
[44] Jan Zahálka, Marcel Worring, and Jarke J. Van Wijk. 2020. II-20: Intelligent and pragmatic analytic categorization of image collections. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 422–431.
[45] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Appendices

A Research Methods

A.1 Preliminaries

— Architecture and dataset: We use VGG-16, a DNN [38], pre-trained on ImageNet [6]. As images we used a dataset provided by Peterson et al. [30], which consists of 720 images divided into six categories of 120 images. The categories were: Animals, Fruits, Furniture, Various, Vegetables

and Automobiles (the latter effectively including any means of transportation including horses, sleds, cranes; Transportation henceforth). Images had a native resolution of 500×500 which was down-scaled to 224×224 to fit the model.

— Human similarity judgments: Let H be a matrix representing the similarity judgments provided by human assessors for n objects. Each entry $H_{i,j}$ in the matrix corresponds to the similarity judgment between objects i and j. We use the upper triangle of matrix H, denoted as H_u.

— Object distances in feature space: Let C be a matrix representing the embeddings of n images onto d features of the penultimate layer of the pre-trained computer vision model, denoted as $C \in \mathbb{R}^{n \times d}$. Specifically, we use VGG-16 with $d = 4{,}096$. Matrix C is obtained by considering all parameters of the pre-trained model, and specifically all 4,096 nodes of the penultimate, fully connected layer. Z_u is the upper triangle of image-pair similarity matrix Z, computed from the Pearson correlation for each pair of rows in C.

— Subspaces in matrix C: We produce a variant of C, denoted as $C^{(\neg k)}$, by excluding a single node k where $k \in \{1, 2, \ldots, 4{,}096\}$. A second variant is produced when using only a subset S of nodes in the penultimate layer. Let $S \subseteq \{1, 2, \ldots, 4{,}096\}$ be a set of selected node indices, and let $C^{(S)}$ be the matrix representing the embedding of n images onto d nodes in the penultimate layer, but when using the subset of feature-maps corresponding to S. Note that in both cases, the image embedding using the pre-trained weights, with the only difference being the set of nodes used in the penultimate layer.

— From the variants of C we derive matching similarity matrices. The first, $Z^{(\neg k)}$, is obtained by computing the Pearson correlation for each pair of rows in $C^{(\neg k)}$. The second, $Z^{(S)}$ is formed using the selected feature indices in $C^{(S)}$.

— As indicated, Z_u and H_u denote the vectorized upper triangles of matrices Z and H respectively. The Pearson's correlation coefficient between the two is denoted as $r(Z_u, H_u)$. We adopt the terminology of referring to this value as a Baseline 2OI between the two domains. Analogously, in some cases we compute $r(Z_u^{(\neg k)}, H_u)$ and $r(Z_u^{(S)}, H_u)$.

A.2 Supervised Pruning: Identifying a Subset of Nodes That Optimizes Prediction of Human Similarity Judgments

We define the **Alignment Importance Score (AIS)** of each node k in terms of its predictive capacity for the human representation H_u. Intuitively, we aim to determine how the removal of each node $k \in \{1, 2, \ldots, 4{,}096\}$ affects the baseline isomorphism, $r(Z_u, H_u)$. The removal of each node produces a modified 2OI score, $r(Z_u^{(\neg k)}, H_u)$. Finally, The AIS of node k is defined in Equation (A1), with positive values indicating a relatively important node, and negative values a less important one. After computing AIS for all nodes, we rank-order them based on their AIS:

$$\text{AIS}_k = r(Z_u, H_u) - r(Z_u^{(\neg k)}, H_u). \tag{A1}$$

We then identify an optimal subset of modes for predicting H_u. In each iteration, one node is added to the subset S in descending order of AIS rank, and we recompute the 2OI, $r(Z_u^{(S)}, H_u)$ using that subset. After these 4,096 iterations, subset S^* ultimately selected is the one that maximizes 2OI.

A.3 Analysis of Generated Images: Identifying Images of Animals

To identify images of animals from the "test-unique" and "validation-unique" splits of the Pick-a-Pic dataset, we used the following manually compiled list of animal names:

Cat dog, fish, sheep, deer, lion, tiger, elephant, giraffe, wolf, bear, monkey, rabbit, horse, cow, pig, chicken, duck, goose, mouse, fox, owl, frog, shark, whale, penguin, seal, koala, panda, zebra, kangaroo, octopus, antelope, bee, butterfly, snake, alligator, crocodile, turtle, lizard, donkey, bat, parrot

Note that here we omit the plural forms of the names to avoid redundancy but those were included in the list used in the experiment.

B Supplementary Figures

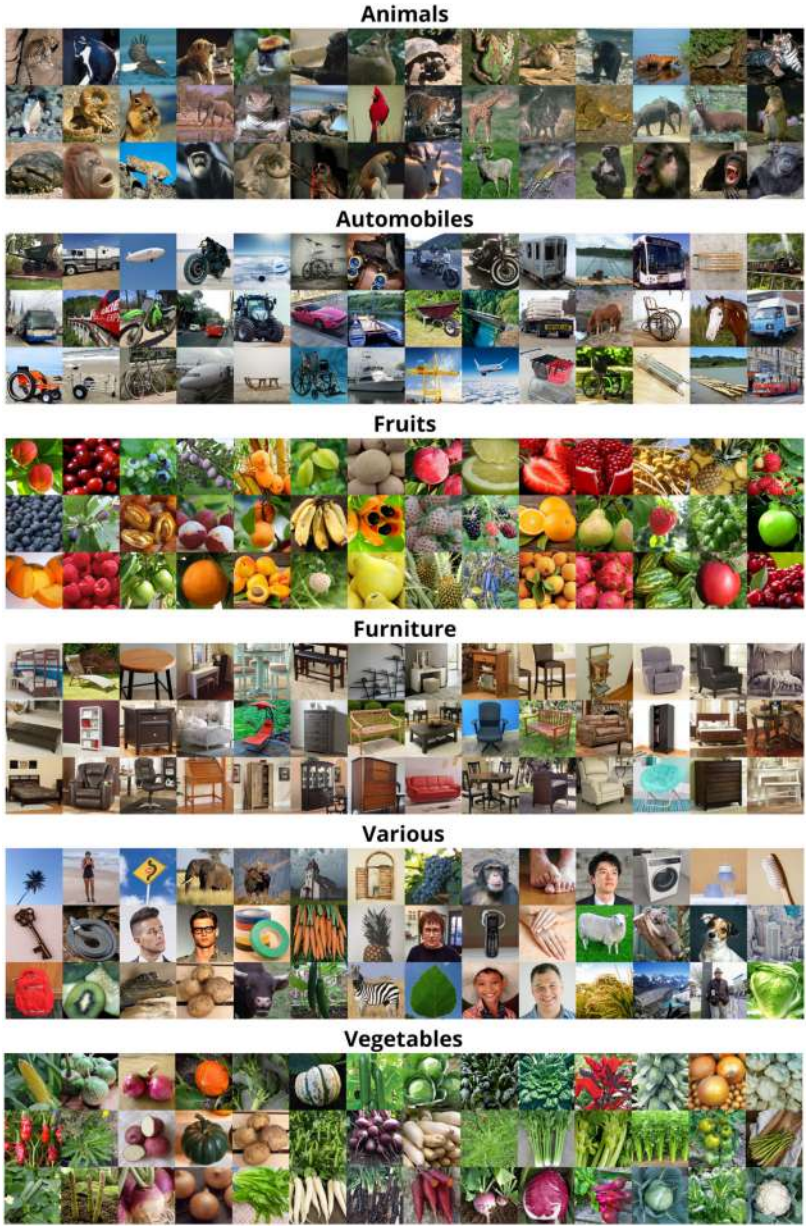

Fig. B1. Example of Images in each dataset. Used with Permission from John Wiley and Sons, License number 5476920331867. From Peterson et al. Cognitive Science, Volume: 42, Issue: 8, Pages: 2648-2669, First published: 03 September 2018, DOI: https://doi.org/10.1111/cogs.12670.

Received 1 May 2024; revised 3 December 2024; accepted 22 August 2025