Emergent dimensions underlying human understanding of the reachable world

Emilie L. Josephs[1], Martin N. Hebart[2], and Talia Konkle[3]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA
[2]Vision and Computational Cognition Group, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[3]Psychology Department, Harvard University, Cambridge, USA
[*]ejosephs@mit.edu

ABSTRACT

Near-scale, reach-relevant environments, like work desks, restaurant place settings or lab benches, are the interface of our hand-based interactions with the world. How are our conceptual representations of these environments organized? For navigable-scale scenes, global properties such as openness, depth or naturalness have been identified, but the analogous organizing principles for reach-scale environments are not known. To uncover such principles, we obtained 1.25 million odd-one-out behavioral judgments on image triplets assembled from 990 reachspace images. Images were selected to comprehensively sample the variation both between and within reachspace categories. Using data-driven modeling, we generated a 30-dimensional embedding which predicts human similarity judgments among the images. First, examination of the embedding dimensions revealed key properties that distinguish among reachspaces, relating to their structural layout, affordances, visual appearances and functional roles. Second, clustering analyses performed over the embedding revealed four distinct interpretable classes of reachspaces, with separate clusters for spaces related to food, electronics, analog activities, and storage or display. Finally, we found that the similarity structure among reachspace images was better predicted by the function of the spaces than their locations, suggesting that reachspaces are largely conceptualized in terms of the actions they are designed to support. Altogether, these results reveal the behaviorally-relevant principles that that structure our internal representations of reach-relevant environments.

Introduction

While we may never know how a raven is like a writing desk [1], we can confidently articulate how a writing desk is like a library desk, and not like a spaceship control panel. What knowledge supports this judgment? Judgments of similarity emerge in part because the world is structured and predictable: entities can be divided into types, and entities of the same type will share a set of properties[2,3]. To date, extensive research has uncovered much of the organization of object and scene properties[4–15]. However, the representations underlying the rich, near-scale environments in which we perform hand-based actions have only recently begun to be explored[16–18].

Consider the desktop environment where you type an email, the tabletop where you enjoy a meal, or the kiosk where you check in for a flight. These reach-relevant spaces (hereafter "reachspaces") are highly behaviorally-relevant environments, which support many of our hand-based tasks and activities and form the backdrop to many of our day-to-day behaviors (see Figure 1 for examples). They differ from both singleton objects and navigable-scale scenes: they encompass spatial extent and multiple objects, but they require coordination of the hands among graspable objects, rather than transportation of the body through an enclosing space. Recently, evidence has emerged that reachspace images have distinct visual statistics from object or scene images[17,19] and elicit distinct topographies of activity in the brain, with particularly strong recruitment of parietal regions[17,20]. However, the factors that structure human knowledge of these reach-relevant environments have not been systematically mapped.

One way to understand the structure of internal representations is to probe the similarity among many exemplars of a concept[21,22]. In representational similarity analysis, the similarity (or dissimilarity) among items is conceptualized geometrically as the distance between them in a multi-dimensional feature space and is often expressed as a matrix of pairwise distances[23–27]. These similarity measurements can be leveraged to discover a low-dimensional embedding space for a set of concepts, revealing the dimensions along which concepts vary. These dimensions indicate properties that are relevant to humans, and it has been proposed that they form the mental axes along which categorization and generalization operate[2,4,28–30].

[1]In Alice's Adventures in Wonderland by Lewis Carroll[1], the Mad Hatter poses a riddle: "Why is a raven like a writing desk?". Famously, he never provides the answer, and generations of readers have been left to guess for themselves.

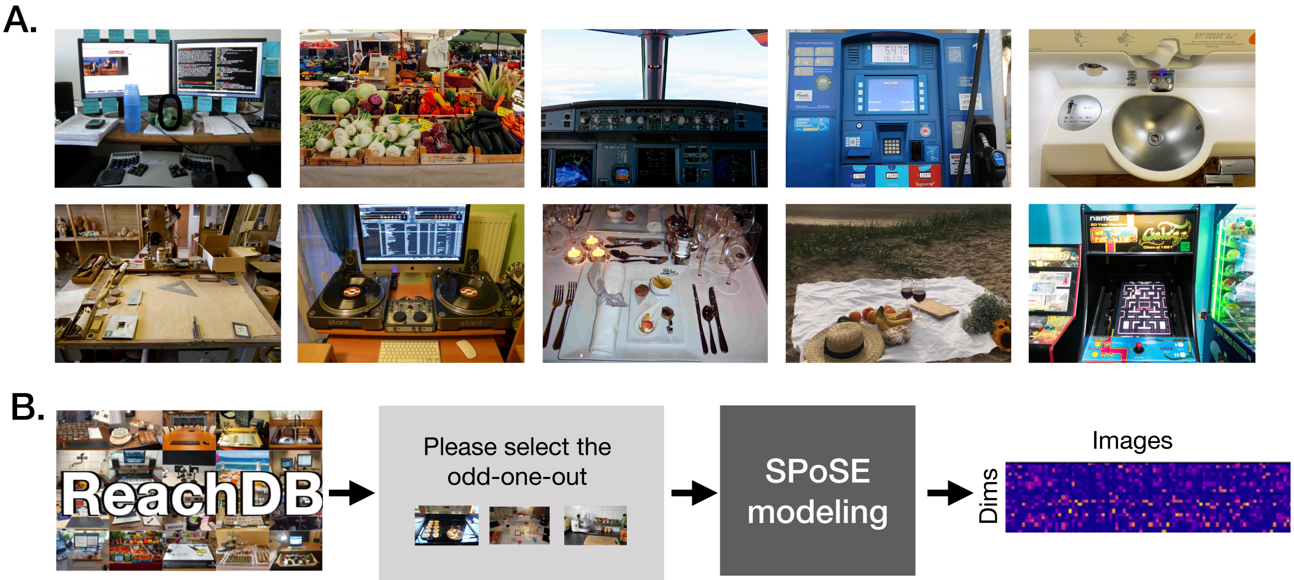


Figure 1. Stimuli and Methods. A) Examples of reachspaces: reach-relevant environments that support task-oriented behavior using the hands. B) Steps of the modeling procedure: 990 images were selected from the Reachspace Database (ReachDB), with 3 exemplars from each of 330 categories. Odd-one-out judgments were collections on 1.25 million triplets, and modeled using the Sparse Positive Similarity Embedding approach (Hebart et al, 2020), to derive a low-dimensional embedding that predicted human similarity judgments.

Here, we use large-scale crowdsourcing and computational modeling to reveal the similarity structure underlying our knowledge of reachable environments and derive key properties underlying this structure. Studies in applied areas, such as ergonomics, human factors engineering, and environmental psychology, have suggested some properties which distinguish among reachspaces but have generally explored this question within a very narrow scope. For example, some distinctions have been proposed based on the action demands of the space, including digital versus traditional media workspaces (e.g.[31]), or workspaces which support precision work versus strength-based work[32]. Other distinctions have been proposed based on the people using the space (e.g. experts versus novices[33]; individual versus collaborative work[34,35]). However, these divisions have been explored piecemeal, by making a-priori distinctions within a prescribed kind of workspace, for example by testing for differences between personal and collaborative spaces within the circumscribed category of digital office workstations. Thus, it remains an empirical question how our knowledge of the broader reachable world is structured, and whether distinct classes can be identified in a data-driven manner from a large and comprehensive sample of everyday environments.

In the present work, we collected 1.25 million behavioral similarity judgments on a set 990 images of reach-relevant environments and used computational modeling to capture the representational structure of these judgments. Broadly, we find that a 30-dimensional space can account for the similarity structure in the judgements and that this space can be divided into four distinct classes. Additionally, we find that the similarity judgments among pairs of reachspaces is predicted more strongly by their respective functions than their locations. Altogether, this work reveals the structure of internal representations of reach-relevant spaces and highlights the broader importance of function for organizing knowledge about the world.

Results

A set of 990 reachspaces images, spanning a diverse collection of 330 different categories, were selected from the Reachspace Database[36]. All images depicted near-scale views of spaces that afford hand-based actions, captured from the point of view of an agent performing a task in the space (See Figure 1A for examples). The depicted reachspaces generally consisted of extended surfaces, oriented horizontally or vertically, populated with objects or other elements that afford interaction, such as tabletops, countertops, or digital kiosks. Image categories were selected to widely sample the reachable world, including places where we work, play, study, eat, shop, create, perform music, store things, and more. These categories are highly granular, dividing up reachspace types according to a combination of the settings they are found in (e.g. home, hotel), the rooms they belong to (e.g. office, dining room), the locus of interaction in the space (e.g. desk, counter), and specific actions associated with the depicted reachspace (e.g. working, eating, cake decorating). Three different images were included per category, chosen to be as different from each other as possible while remaining good examples of the category.

A similarity space was derived for these images by modeling behavioral judgments. We used the Sparse Positive Similarity Embeddings approach, or SPoSE[7], which has two notable features. First, and most importantly for our purposes, it creates a dimensional model of the similarity space, inferring a set of axes which underlie the variation among images and assigning each image a score on each of these dimensions. Second, it allows us to infer this similarity space from a fraction of all possible combinations of images, reducing data collection to tractable levels.

In a first step, 1.25 million behavioral similarity judgments were obtained on images from the stimulus set, using a triplet "odd-one-out" task (Figure 1B). On each trial, participants were shown three images and asked to indicate which was the most different from the other two. A single triplet judgment yields information about three pairs of objects: it indicates that the selected image has low similarity with each of the two non-selected images, and the non-selected images are similar to each other. In this approach, the third image acts as a minimal context within which to evaluate the similarity of the other two images. Across many trials, a given pair of images is thus evaluated across many different contexts, allowing us to measure the probability that two images will be considered similar, marginalized across all possible contexts. Images were randomly selected from the stimulus set, with the constraint that every possible pair of images was sampled at least once.

Next, these judgments were modeled using the SPoSE approach. SPoSE works by randomly initializing each image as a point in a high-dimensional feature space, and then tuning image weights along these dimensions to derive an embedding which can predict similarity judgments obtained in the triplet task. This model makes two theoretical assumptions. First, it assumes that the dimensions of this embedding space are sparse — that is, that each reachspace only has some dimensions, but not all (e.g. a labbench would have low weights on dimensions relating to food or leisure). Second, it assumes that the dimensions are positive, such that they can only add up but not cancel out (e.g. a food-related property and a seating-related property should not cancel each other out). This also means that the weight of an image on a dimension can be interpreted as the amount of the corresponding property present in the image.

As the SPoSE modeling approach is stochastic, we ran 50 iterations of the model, yielding 50 embeddings. We found that the solutions were largely stable: the number of discovered dimensions ranged from 27 to 32, and there was very low variance in triplet prediction accuracy on the separate test set across the 50 embeddings (mean: 60.55%, stdev: 0.04%). All further analyses were conducted on the embedding that was most representative of all 50 SPoSE iterations (see Methods). This model contained 30 dimensions.

Before proceeding, we confirmed that the dimensions in the selected embedding were replicable, that is, that they appeared consistently across different iterations of the model. Specifically, we derived a replicability score for each dimension in the

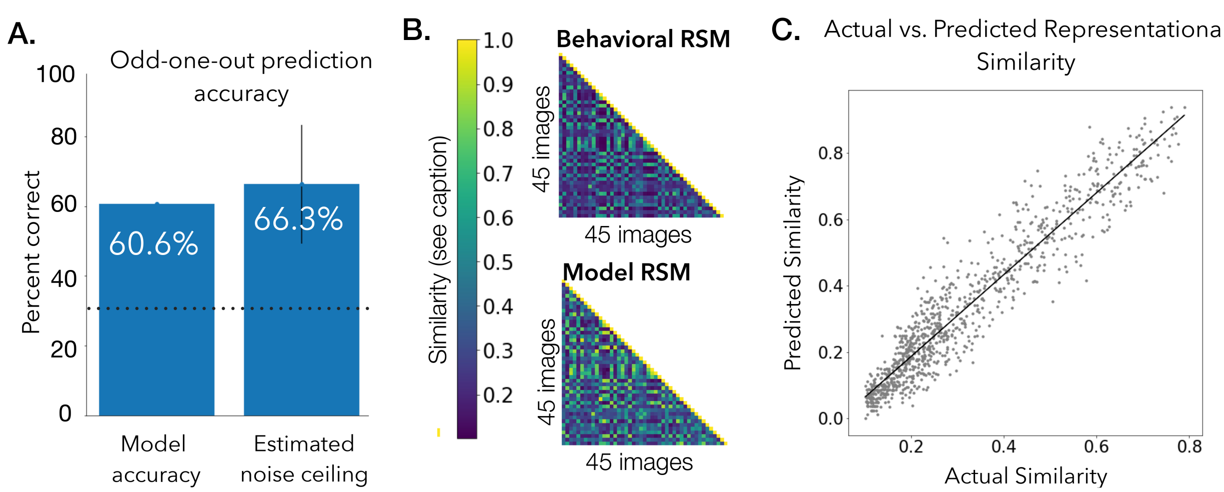


Figure 2. Validation of the model embedding derived from similarity judgments over 990 reachspace images. A) Model performance on odd-one-out prediction for held-out test set. The noise ceiling of the behavioral data was estimated from a separate behavioral sample, and represents the average inter-rater reliability over 1000 triplets. B) Representational similarity matrices for a 45-image subset of the stimulus set, created by fully sampling all possible triplets in a validation behavioral experiment (top) and by estimating similarity based on the model embedding (bottom). Here, the similarity between two images is operationalized as the proportion of times they are judged to be similar, across all trials. C) Correlation between actual and predicted similarity between all image pairs in (B).

selected embedding by calculating the average Pearson correlation between it and its closest analog (i.e. most correlated dimension, controlled for using cross-validation, see Methods) in each of the remaining 49 initializations of the model which were not selected for analysis. All dimensions had replicability scores greater than r=0.70, with the exception of one dimension corresponding to outdoor arrays of objects, whose score was r= 0.40 (20/30 dimensions had replicability scores >0.90, and 26/30 had >0.80). Altogether, this suggests that the following results are not specific to individual model iteration we examine.

Model validation

Did the model successfully learn to predict odd-one-out judgments? After training, the model predicted odd-one-out choices from a held-out set of triplets with 60.6% accuracy (Figure 2, chance = 33.3%). To contextualize this performance, we estimated the behavioral noise ceiling of these data. In a separate behavioral sample, odd-one-out judgments were collected for 1000 triplets, randomly sampled from the image set (mean number of judgments per triplet: 28.6, range 18-36). The average consistency in judgments across participants was 66.3%. Thus, with its 60.6% performance, the SPoSE model could predict human behavior up to 82.1% of the noise ceiling (see Methods).

One promise of SPoSE is to accurately capture the representational space underlying triplet similarity judgments on a large number of images, while sampling only a subset of all possible triplets. To validate this prediction, we conducted a smaller-scale behavioral experiment. We collected similarity judgments for all possible triplets from a subset of 45 images (26,588 total trials, see Methods). A behavioral representational similarity matrix (RSM) was created from the participants' responses by computing the proportion of times a given pair was considered similar, across all triplets in which they appeared together. Similarly, a model RSM was created from the embeddings for these 45 images objects, by computing the predicted probability that each pair would be similar, across all triplets (see Methods). Overall, behavioral and model RSMs were highly correlated (Figure 2C, r=0.95). Thus, the model's representational space was able to accurately reconstruct the space derived from a full sampling of triplets. This supports the claim that the representational space derived from the embedding reflects the similarity space underlying human similarity judgments over these items.

Overall, the SPOSE model was highly successful both at predicting triplet similarity judgments and at reconstructing the representational space underlying these judgments. This model yielded two outputs: 1) a set of 30 dimensions that summarize axes along which reachspace can differ, and 2) an RSM that quantifies the similarity among pairs of reachspace images along these dimensions. In the following sections, we examined these dimensions and this RSM to reveal key factors that underlie similarity judgments among our reachspace images.

The embedding dimensions are interpretable and informative

A significant benefit of the SPoSE modeling approach is that it yields an embedding with accessible and interpretable dimensions. We first examined these dimensions, to gain insights about the properties that are salient to observers when making similarity judgments on reachable environments.

Each dimension was visualized by ordering the images according to their weights on the dimension (Figures 3, 4, and 5 show the top 6 images per dimension, with word clouds depicting participant-generated labels for the dimension). While the dimensions emerged independently in the model, here we discuss them in pairs or groups, to better highlight some of the concepts they capture. First, some of the dimensions pertained to global properties of the space: for example, separate dimensions emerged for cluttered versus clear spaces (Fig 3A). Additionally, many of the dimensions captured complex combinations of semantic category and physical affordance information about a space. For example, two dimensions emerged for musical instruments, but they distinguished between keyed instruments and non-keyed instruments (Fig 3B). Likewise, two dimensions emerged for outdoor spaces, distinguishing those containing multiple small objects from single large objects (Fig 3C), and separate dimensions emerged for workshop-related spaces where the primary surfaces was oriented vertically vs horizontally (Fig 3D). Third, some dimensions also captured information about the intended user of the space: separate dimensions emerged for children's games versus adults' games (i.e. gambling, Fig 4A) and for electronic spaces used by everyday consumers versus those requiring expertise (Fig 4B). Fourth, some dimensions captured information about the physical properties of the space: craft-related spaces had separate dimensions for arts which use wood vs other media (Fig 4C) and other dimensions emerged for spaces with ceramic, paper, or stainless steel components (Fig 5). Additionally, multiple dimensions emerged relating to the storage of items, with distinctions between storage at home, in retail, with portability constraints, and for travel (Fig 5, Fig 3D).

Note that any given dimension can be characterized in multiple ways. Here we discuss just one interpretation per dimension to highlight the kinds of concepts they measure. Overall, dimensions discovered by the model were interpretable and revealed fine-grained distinctions within this large set of reachable environments.

Evidence for conceptually distinct reachspace classes

We next characterized the global structure of reachspace similarity in this 30-dimensional space. Similar large-scale work on objects[7], showed global organization into well-known classes (animate/inanimate, natural/human-made). The current study

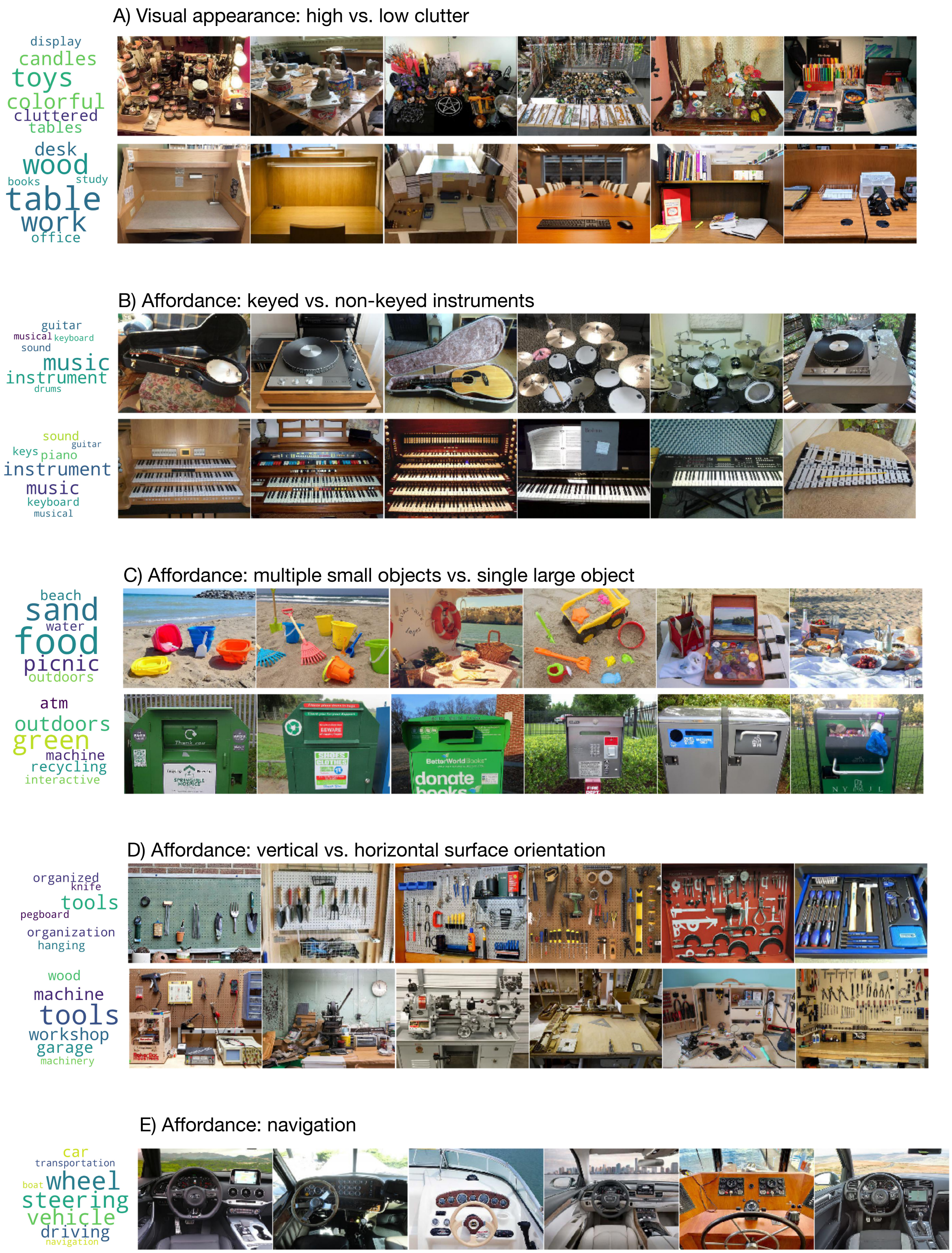


Figure 3. With Figures 4 and 5, this figure shows the dimensions forming the embedding. Each dimension is illustrated with the top 6 images of the dimension (largest weight), and word clouds show responses from 50 participants asked to judge what is captured by the dimension. For illustrative purposes, we divided the dimensions into groups to highlight the subtle distinctions they are sensitive to.

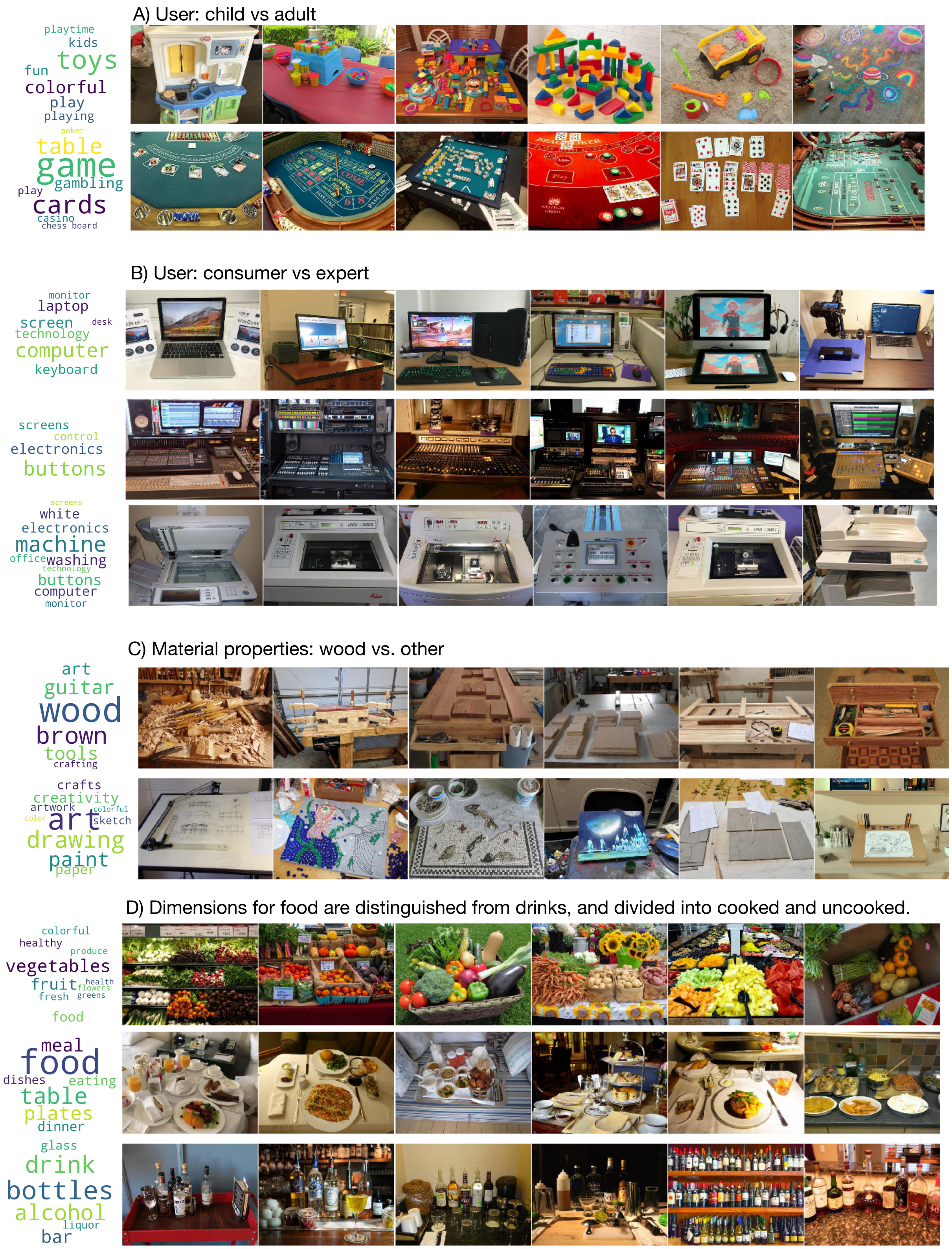


Figure 4. With Figures 3 and 5, this figure shows the dimensions forming the embedding.

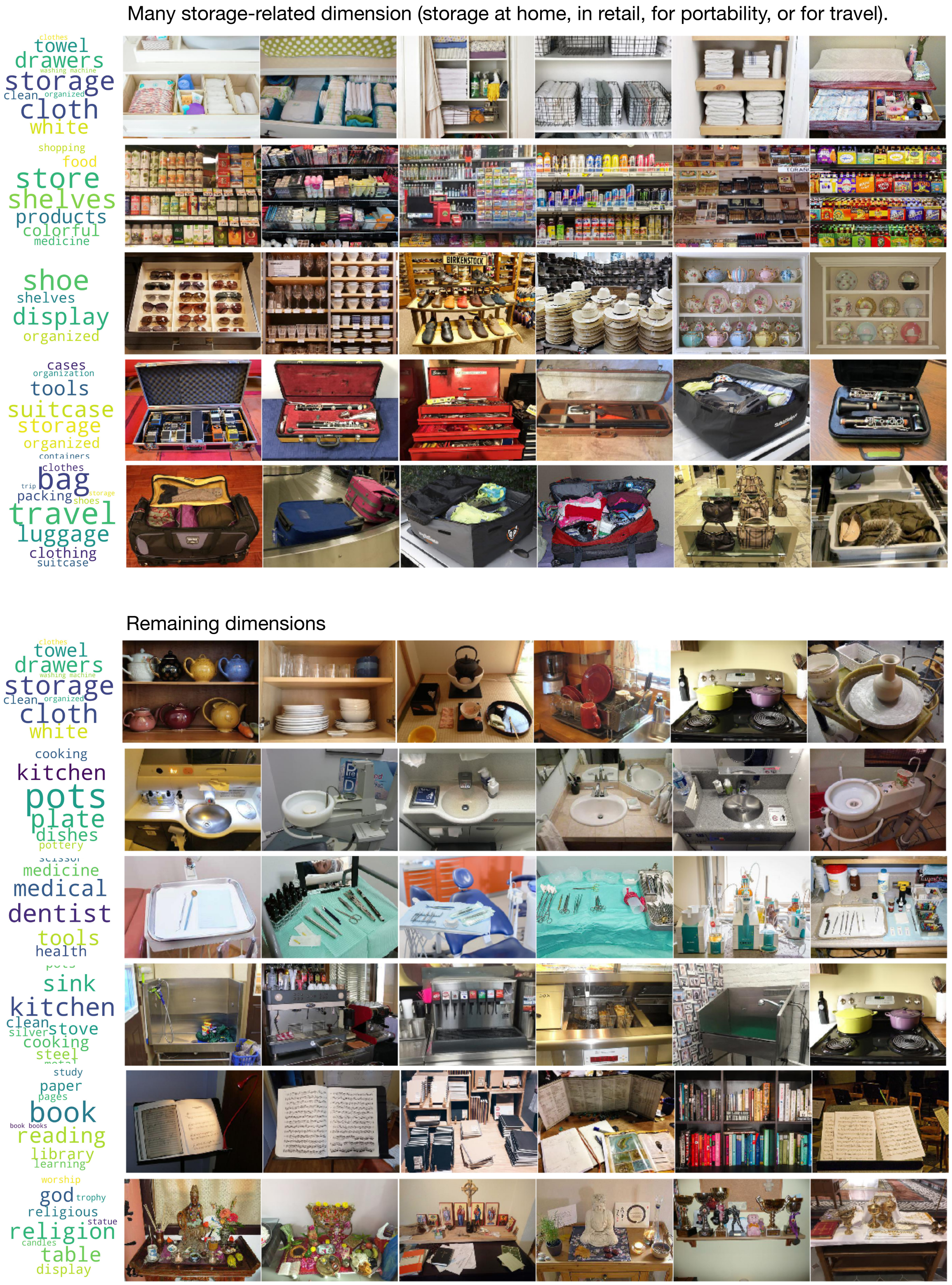


Figure 5. With Figures 3 and 5, this figure shows the dimensions forming the embedding.

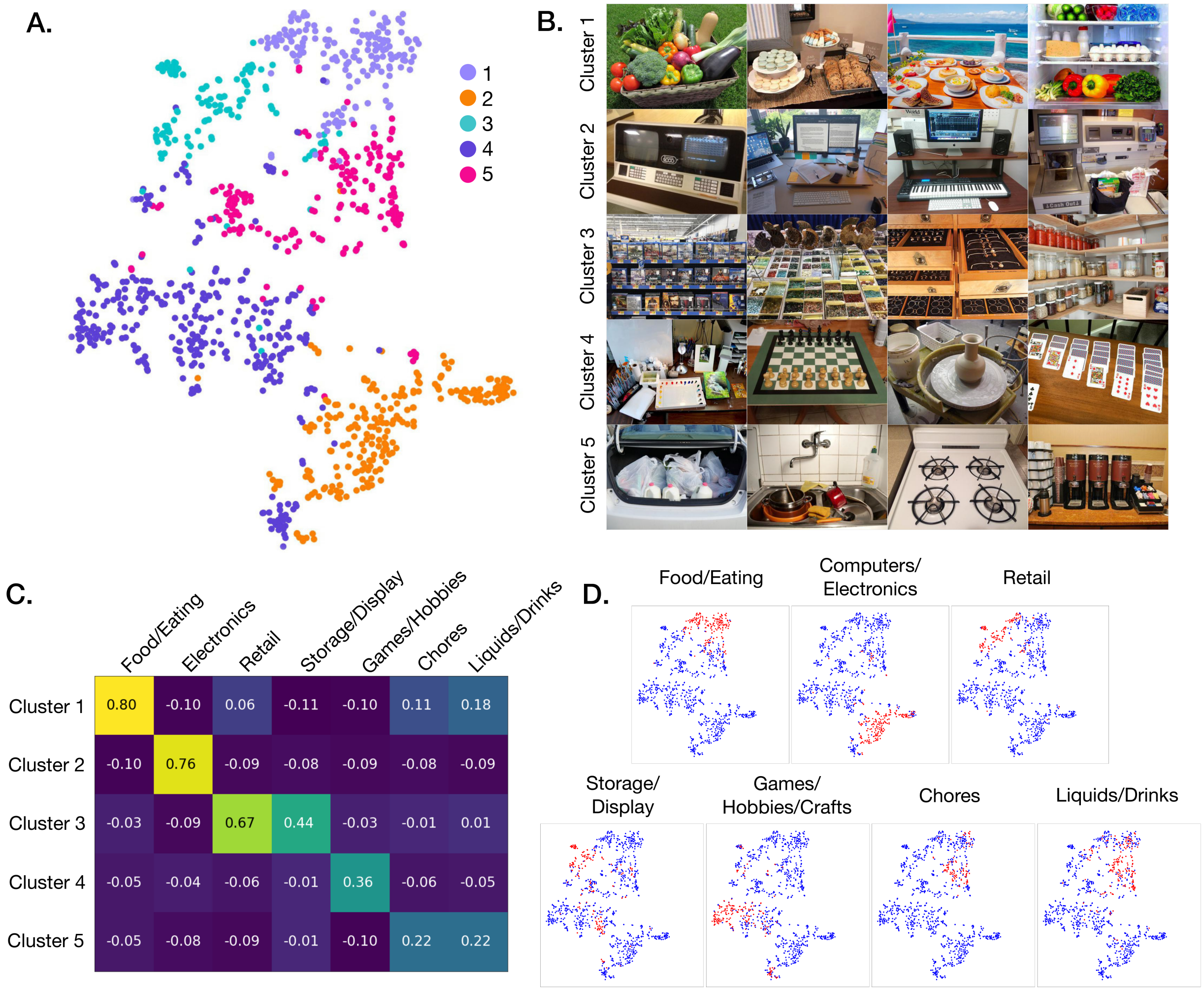


Figure 6. Data driven discovery of large-scale divisions in the representational space of reachspaces. A) 2-D projection of the representational space using MDS-initialized t-SNE. Dots correspond to images, and are colored according to their clustering assignment in k-means clustering (k=5). B) Four example images from each cluster. C) Adjusted Rand Index measuring the image-wise correspondence between cluster assignment and labels derived from behavioral ratings. D) Visualization of the embedding space with each behaviorally-derived label shown. Red dots indicate the images that were judged to fit the given labels in a behavioral tasks.

drew from 330 different reachspace categories, but it is possible that these are conceptually clustered into a smaller number of classes. How many classes were participants sensitive to in these images, and what concepts do they correspond to?

Taking a data-driven approach, we applied k-means clustering to group the images according to their similarity in the embedding space; this procedure yields both clusters and the corresponding cluster centroids, which indicate what parts of the embedding space these images occupy. As images are grouped into more and more clusters, the clusters become smaller and closer together in the embedding space, and the similarity of the cluster centroids increases until it reaches a plateau. We focused our analyses on the clustering solution right before the plateau in mean cluster center similarity, k=5, to preserve distinctiveness among the clusters (see Supplemental Figure 1). Figure 6A shows a 2D projection of the representational space for the 990 images, with cluster assignment indicated by color (the projection was obtained using MDS-initialized t-SNE, to capture both global and local structure). Visual inspection of the images in each cluster (Figure 6B) suggests that they relate to 1) food and eating, 2) computers and electronics, 3) spaces for storage, retail and display (excluding food), and 4) entertainment, hobbies, and handicrafts. The fifth cluster was less interpretable and contained a mix of spaces related to drinks or liquids, and household chores. Thus, human similarity judgments suggest the existence of about 4-5 broad types of reachspaces within our sample.

To validate these possible cluster identities, we collected behavioral ratings for the images on Mturk (see Methods) and assessed the correspondence between ratings and cluster assignments. For each possible cluster identity described above, participants were presented with a brief description (see Supplemental Table 1 for task wording) and asked to indicate for each image whether it matched the concept or not. Correspondence between these conceptual labels and the k-means clusters was assessed using the Adjusted Rand Index (ARI), which calculates the proportion of times both solutions agreed about whether a pair of images was in the same cluster or not (ARI=0 indicates chance, ARI=1 indicates perfect agreement between the solutions).

Results are shown in Figure 6C and 6D. A correspondence matrix between cluster identity (cluster 1-5) and conceptual label confirmed that each of the hypothesized cluster identities accounted for different clusters (6C). The clearest clusters corresponded to food-related and electronics-related reachspaces, with near perfect alignment between images with those attributes and clusters 1 and 2 (ARI= 0.80 and 0.76, respectively). Cluster 3 aligned moderately-well with retail spaces (ARI = 0.67) and with the broader concept of spaces designed for storage and display (ARI = 0.44). Cluster 4 corresponded moderately-well with spaces for games, hobbies, art and handicrafts (ARI = 0.36). The final cluster was less interpretable and showed slight correspondence to drinks- and liquids-related reachspaces (ARI=0.22) and slight correspondence to spaces for household chores (ARI=0.22).

Overall, this analysis shows that the reachspaces in our sample could be divided into a relatively small number of broadly distinct classes. Food and electronics accounted for the most clearly distinct clusters, and the remaining reachspaces showed distinctions among spaces for storage, for arts, hobbies or entertainment, and for chores.

Reachspace similarity judgments reflects shared function more than shared location

Recent studies of navigable-scale environments found that the function of a place plays a large role in determining what other places it will be considered similar to (see Greene, Baldassano, Esteva, Beck & Fei-Fei, 2016[14]). Are reach-relevant environments likewise grouped by function? One the one hand, reachspaces are designed to support specific activities, so it is reasonable to suppose that they would show strong conceptual organization by function. However, reachspaces are related to the broader environment in lawful ways and tend to belong to particular locations (e.g. a kitchen counter is in a kitchen, while a bedside table is in a bedroom). Thus, they may be better grouped according to the locations they occupy. We next tested which of these principles – function or location – best accounted for similarity judgments among reach-relevant environments.

As in Greene et al (2016), we defined the function of the space as the action it affords. Thus, we labelled each reachspace according to the specific action that would be performed there (e.g. shopping, chopping vegetables, cake decorating, working, embroidering). We captured these actions at a high level of specificity (e.g. "chopping vegetables" and "rolling dough" rather than the more general "cooking"), as this is a better reflection of the precise activity, object array, and motor plans associated with a given reachspace.

The location of each reachspace was operationalized in three different ways. At the broadest level, location was indexed by the setting, or general environment type, that the reachspace is located in (e.g. office building, home, hotel, hospital). At a more specific level, location was operationalized as the room containing the reachspace (e.g. kitchen, office, bedroom). Finally, location was also specified in terms of the interaction locus (hereafter locus) of the space, i.e. the type of object or surface that forms the primary structure of the reachspace (e.g. desk, table, pegboard, counter, control panel). Note that while these three models describe context at different scales of spatial inclusion, they are not nested hierarchically (a table can be found in many different rooms, and rooms such as offices can be found across many different settings) and should instead be thought of as different, partially independent ways of slicing across the images.

Overall, reachspaces were divided into 38 Settings, 122 Rooms, 151 Loci, and 131 Actions. These labelling schemes were highly independent from each other (average Adjusted Rand Index among the different labelling schemes was 0.14, see Supplemental Figure 2). All labels were based on the Reachspace Database annotations for each image and were validated by the experimenters (see Methods). We quantified whether sharing a label under each of these schemes predicted greater similarity than having different labels. Over 10,000 iterations, we randomly selected one reference image and 2 comparison images, with the constraint that one comparison image shared a label with the reference and the other did not. We then measured the proportion of times the reference-comparison pair which shared a label was more similar (i.e. lower Euclidean distance) than the pair that did not.

Results are shown in Figure 7. Sharing an Action label predicted greater similarity 75.1% of the time. In comparison, sharing a location label at the Setting, Room and Locus level predicted greater similarity 62.3%, 67.8%, 65.1% of the time, respectively (chance = 50%, confirmed by simulation). Thus, both location and afforded action accounted for some of the structure in the representational space of reachspaces, but action was the better predictor of representational similarity. Overall, this suggests that human judgments of similarity among reachspaces relate more to the function they serve than the places they occupy.

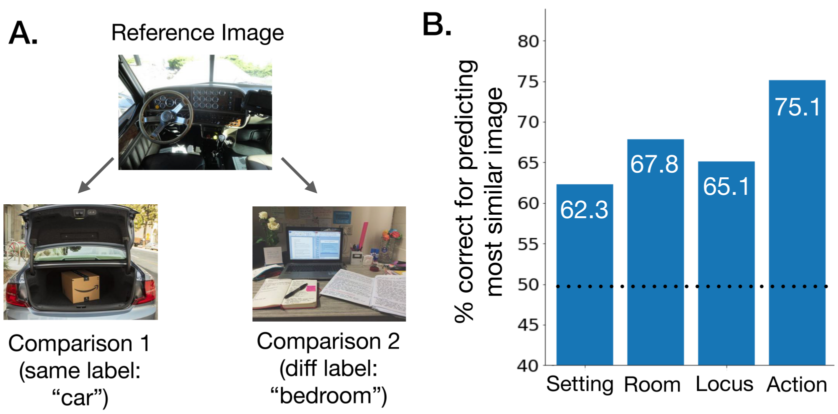


Figure 7. Evaluating the relative influence of location and function on reachspace similarity judgments. A. Method: Given a reference image, participants were asked to select two comparison images, one which shares a label with the reference and another that doesn't. Four sets of labels were used, indexing the Setting, Room, or Locus the reachspace belongs to and the Action it supports. B. Results: Bars indicate the percent of time that the reference image had higher similarity to the comparison image which shared a label. The dotted line indicates chance (50%).

Discussion

Here, we used 1.25 million human similarity judgments to derive an embedding for 990 images of reach-relevant environments, and examined this embedding to characterize key factors that organize our conceptual representations of the reachable world. We found that human similarity judgments can be predicted with a 30-dimensional representational space and that the embedding dimensions capture information relating to the content, layout, purpose, and even typical user of the space (i.e. adult vs child). We described the global structure of this similarity space, finding evidence for a small number of broad reachspace classes, and we found that function is a better determinant of conceptual similarity than location. Altogether, this work reveals the conceptual structure of reach-relevant environments, in a large scale, data-driven manner.

Dimensions of reachspace similarity space

What kind of information is captured by the dimensions? In general, dimensions discovered by the model appeared to capture multiple attributes (e.g. "vertically oriented storage in a workshop" in Fig 4D). This reflects the consistent finding that attributes in the world are "clumped" and covary with each other, rather than being uniformly distributed[4]. Additionally, the discovered dimensions capture both low-level visual information and high-level semantic information (e.g. the dimension for child-related spaces also featured bright colors and mid to high levels of clutter). For interpretability, we have applied relatively high-level labels to the dimensions, but it is possible that similarity judgments rely equally on the covarying lower-level, perceptual features. Indeed, scene perception research suggests that the low- and mid-level visual appearance of an environment is diagnostic of, and to some extent inseparable from, its higher level identity and function[37].

It is important to note that the precise content of the dimensions is shaped by methodological and stimulus choices. Our aim was to capture a general, intuitive similarity space, so we used a triplet task with minimal instructions. Different similarity spaces would emerge if participants were asked to judge similarity on specific bases. Likewise, the number of dimensions depend in part on the stimulus set. If some areas of the stimulus space were oversampled, this could lead to an inflated number of dimensions spanning this part of the space, and conversely the model could not learn dimensions for areas of the stimulus space that are undersampled. In spite of this, several design choices increase the chance that the results will generalize to other samples: the sparsity constraints on the model encourages it to discard spurious dimensions, and the set of reachspaces was carefully sampled to constitute a comprehensive set, both within and across categories.

How do the dimensions underlying reachspace similarity judgments compare to those for objects? The SPoSE approach was initially applied to object images[7], so we can compare the embeddings from the two studies. Comparing the dimensions from our Figures 3,4, and 5 to the dimensions in Extended Data Fig 2 from Hebart et al (2020), we find that some of the dimensions found here showed some correspondence to dimensions for objects, most notably those for electronics and food. However, there are some major differences. First, objects dimensions in Hebart et al. (2020) showed more reliance on simple features like color (e.g. black, red) or shape (e.g. round, disc-like, long and thin). In contrast, dimensions for reachspaces show evidence of integrating over more complex feature combinations, as discussed above. Second, some dimensions that appeared for objects are enriched with contextual information for reachspaces. For example: objects have one dimensions for "clothes", but reachspace dimensions distinguish whether the clothes are in an environment relating to retail, storage, or travel (Figure 6). Overall, while there is some overlap in the relevant concepts, the representational space of reachspaces cannot be reduced to that of individual objects.

It is more difficult to assess the generalization to scenes, as the SPoSE approach has not yet been applied to scenes. Previous work extracting scene attributes from text descriptions[13] found that scenes properties relate to their functions (e.g. sports), prominent materials (e.g sand, foliage), material properties (e.g. rusty, glossy), and spatial envelope (e.g. open, enclosed). However, no dimensionality reduction was applied to those results, so the attribute list is large (102 attributes) and does not capture the correlation structure among them. Going forward, it will be important to test objects, reachspaces, and scenes in the same paradigms to discover the attributes that are common across them and those that are specific to different scales of experience.

Major classes of reachspaces

A clustering analysis discovered four identifiable classes of reachspaces, corresponding to food-related spaces, electronics-related spaces, hobby/craft/entertainment-related spaces, and storage/retail/display spaces. While these labels provide a description of the reachspace categories in each cluster, we suggest the following broader interpretation of the classes: 1) food-related spaces, 2) digital or electronic spaces, 3) analog spaces intended for active engagement with objects, and 4) analog spaces intended for passive storage or display of objects. There was an additional cluster which was ambiguous, showing weak correspondence to either household chores or drinks/liquids, but due to its ambiguity, we do not provide a broader interpretation for this or include it in the list of possible reachspace classes.

Why might these four classes emerge in our internal representations of reachspaces? One possibility is that human agents interact with each of these spaces in generally different manners. Acting in analog spaces usually requires manipulating multiple objects to change the physical state of the reachspace (e.g. moving them around), and requires reasoning about the location and relations of objects across time. Active vs passive analog spaces require different amounts of interaction and monitoring over time, and food-related spaces involve additional reasoning about physiological states like hunger or appetite. In contrast, in electronic spaces, events are largely invisible and instantaneous, without physical grounding, and agents must act on simple symbols (e.g. cursors, buttons), whose function are given by learned arbitrary input-output mappings. Some differences also exist in the components of environments from the different classes. Analog spaces have objects that can be moved or manipulated independently, while electronic spaces often feature components that are attached to a main structure, such as buttons, keys and switches. Thus, these four classes may reflect differences in the representations required to behave in the different environments.

Function as a major determiner of intuitive similarity

One major implication of this work is that function is a salient factor organizing our knowledge of reach-relevant environments. This is reminiscent of the “design stance” which human adopt toward artifacts, in which objects are understood in terms of what they were designed to do[38]. According to this theory of conceptual formation, the underlying nature of an artifact is related to its intended function, which will constrain its form and materials, and provide the best explanatory variable for its appearance and construction. The present results suggest that this stance can explain reasoning about environments as well as artifacts. Indeed, function has also been found to act as an organizing principle for conceptual information about navigable-scale scenes[14,24], suggesting that this is a general feature of our conceptual representation of things and spaces in the world. Altogether this result points to the possibility that the function of a space places strong constraints on its content and appearance, and to the general importance of goals for organizing our understanding of the world.

Conclusion

Altogether, these findings point to distinct classes within the domain of reach-relevant environments. It is still an open question whether these distinctions are reflected in other aspects of reachspace perception. Future work is needed to establish whether these distinctions show additions dissociations in behavior, in their emergence across development, in their susceptibility to disruption following neurological events or cognitive decline, or in the large-scale neural activity the elicit. These results also have implications for the continuing study of reachspaces: as we develop theories of how reach-relevant environments are perceived and represented, it will be important to consider these distinctions and account for how they shape representations.

Methods

Participants:

A total of 4,269 Amazon Mechanical Turk (MTurk) workers were recruited across these studies (mean age was 37.78, 45.2% female, 0.3% non-binary). All workers were based in the US and had MTurk performance approval ratings over 90%, with a minimum of 500 HITs completed. Workers gave informed consent and were compensated for their participation. All procedures were approved by the Harvard University Human Subjects Institutional Review Board. For the odd-one-out judgments, 3,075 Workers participated in the main task, 322 Workers participated in a 45-image validation task, and 376 Workers participated in the reliability sample. An additional 447 Workers participated in image rating studies, and 49 Workers provided labels for dimensions in a naming task.

Stimuli:

The stimulus set consisted of 990 images of reachspaces, selected from a beta version of the Reachspace Database[36]. There were 330 reachspace categories, with 3 images each. The Reachspace Database provides tags for each reachspace image, which index setting it belongs to (the broader location type, e.g. hotel, home, office building, the outdoors), the room or site it occupies (e.g. dining room, conference room, campsite), the primary structure supporting the interaction with the environment ("interaction locus", e.g. surfaces such as tables and shelves, or large interactable objects like control panels and digital kiosks), and the specific action it affords (e.g. cake decorating, titrating). Image display size could vary according to individual Worker computer parameters, but images always maintained an aspect ratio of 4:3, and the maximum display size was 400x300 pixels for the triplet odd-one-out task, and 200x150 in the other tasks.

Behavioral tasks:

Triplet odd-one-out task: On each odd-one-out trial, participants were shown three images side by side and asked to indicate with a click which image was the odd one out. At the beginning of the task, participants were told to "imagine yourself in the environments: where are you standing, what are you holding, what are you doing?", but given no additional guidance. This task was conducted in sets of 20 trials, and Workers could perform as many sets as they wanted up to 250 sets. Image triplets were randomly selected, with the constraint that every possible pair of images showed up together a minimum of 2 times.

Image rating task: To assign labels to clusters in the representational space, we collected correspondence ratings for all images on experimenter-generated labels. For the rating task, participants were given a description, such as "For this task, please indicate which images are related to electronic equipment: that is, images that are related to electronics, computers, and other digital equipment", and indicated which images corresponded to the label. Each trial consisted of a five by five array of images, and participants clicked to select images fitting the labels. Selected images were highlighted with a red border, and could be unselected by clicking again. To prevent the task from being too long, the image set was divided across three separate task sets, and participants could perform as many sets as they wanted. Each image was seen only once per set, with the exception of 20 duplicate images which were used for quality assurance (subjects with <75% agreement on these duplicate images were excluded from analysis).

Dimension naming: Common-sense labels were obtained for each of the 30 dimensions in a simple naming task. Participants saw a 4-by-3 array of reachspace images and were asked to name what was shown in the images. Each array consisted of images selected from the top of one dimension from the SPoSE embedding. Arrays were created by randomly selecting 12 of the top 20 images for that dimension. To ensure that our dimension labels were not influenced by the exact images included, 5 such arrays were created per dimension, yielding 5 different random samples of 12. A given participant saw only one array for a given dimension. Participants were asked to type up to 5 possible labels that described the images in the array, keeping them to 1-2 words in length. Dimensions were presented in random order to minimize order effects. There were 32 trials, one for each dimension, and a compliance-assessment trial consisting of a 4-by-3 array of beach scenes.

SPoSE computational modeling:

Behavioral data were trimmed to enforce quality according to the following criteria before modeling and analysis: 1) all individuals with 60 trials or more who used the same response for more than 40% of HITs were removed, 2) all HITs where participants responded with more than three consecutive sequences (e.g. position 1 then 2 then 3, repeated 3 times) were removed. Both of these criteria were set a-priori based on simulations of expected data patterns. Finally, all trials with reaction times more than 3 standard deviations from the mean were removed (reactions times were log transformed prior to trimming to account for the right-skew of reaction time distributions). In total, 1,251,823 trials passed quality assurance and were used in the modelling.

An embedding for the images was derived from these data following the procedure from Hebart et al. 2020. Briefly, a model embedding is initialized with random weights (range 0,1) on 90 latent dimensions for each of the images, yielding a 990-by-90 matrix. The embedding is then tuned using stochastic gradient descent trained on the human odd-one-out responses. Training proceeds by making odd-one-out predictions based on the embedding, and the model is tested on a withheld set of behavioral responses (10% of the dataset) after every epoch (max 500 epochs), until the model converges. Lambda (the parameter enforcing sparsity, λ=0.007) and learning rate (lr = 0.0005) were tuned in a pre-analysis parameter-optimization step performed on approximately 80% of the data and were selected to yield the lowest final loss. Finally, all dimensions with weights below 0.1 were removed to yield the final embedding.

This model procedure is stochastic, so different initializations will give slightly different results. To select the most stable solution, we ran 50 random initializations of the model and selected the version which yielded the embedding with the highest average correlation to all other embeddings, in a split-half cross-validated analysis. First, half of the data (i.e. weights on all dimensions for half the images) was used to identify corresponding dimensions across embeddings. This was done by identifying, for each dimension in a given embedding, the dimension in each remaining embedding which was most highly correlated with it. Next, the remaining data was used to assess correlations among corresponding dimensions. These correlations were then averaged together over all dimensions within an embedding, and the embedding with the highest average correlation was selected for further analysis.

Noise ceiling and embedding validation:

A noise ceiling was derived for the behavioral data by estimating the average participant agreement on a given answer for a given triplet. One thousand random triplets were selected, and an average of 29 ratings were collected per triplet (range: 18 to 36). The consistency of responses for each triplet was estimated as the proportion of the time that the most popular response was chosen (100% = consistent agreement about the odd-one-out, 33% = chance). This value was averaged across all triplets. The noise ceiling was calculated as follows: (performance – chance)/(noise ceiling - chance).

The ability of the embedding to predict human representational similarity was validated by randomly selecting 45 images from the stimulus set and obtaining a separate behavioral sample of odd-one-out judgments on all possible triplets. A behavioral representational similarity matrix (RSM) was derived by scoring the three possible pairs of images in each triplet: pairs involving the odd-one-out were given a "0" ("considered different"), and the remaining pair was given a "1" ("considered similar"). These scores were aggregated across trials, and divided by the total number of trials per cell to yield a matrix of the proportion of times each pair was treated as similar, across all the triplets in which they were encountered. A model representational similarity matrix for the 45 images was derived in a similar way: we first used the embedding to predict trial outcomes for every possible triplet combination, then computed an RSM from these results using the same approach as for the behavioral data. The correspondence between these matrices was assessed by taking the Pearson correlation between them.

In total, 28,177 odd-one-out trials were included for noise ceiling estimation, and 26,588 odd-one-out trials were used for the 45-image validation sample.

Exploring the embedding dimensions:

The replicability of a given dimension was calculated as its average correlation across the 50 SPoSE iterations. For each dimension in the final embedding, the closest match was identified in each of the embeddings from the remaining 49 iterations using Pearson correlation. The average value of this correlation across all embeddings was taken as the replicability of the dimension. Names for each of the dimensions were derived in 2 ways. First, we provide concise labels corresponding to the concept we felt was most clearly illustrated in each dimension. For increased objectivity, we additionally solicited dimensions names using the behavioral procedure described above. Dimension names were collected from 50 participants, then aggregated. All labels appearing more than 3 times were retained and displayed using word clouds.

Assessing clustering in the representational space:

Divisions were identified among the images using k-means clustering. We tested cluster numbers ranging from 2 to 8. For each clustering solution, the average correlation among all cluster centers was obtained. The optimal number of clusters was determined to be that at which this value starts to plateau. This method of cluster number selection yields the highest number of clusters where cluster centers show more than minimal distinctions from each other. Possible labels for each cluster were derived by looking at the images in the cluster and naming the concepts that they appeared to correspond to. The validity of these labels was assessed using the image rating task described above. Worker responses were turned into binary scores: all images selected for a given label more than 50% of the time was considered to match the labels, all remaining images were not. Finally, the match between each concept and the k-means clusters was assessed by taking the Adjusted Rand Index between the vector of cluster assignments and the vector of label assignments for a given concept.

Assessing the organizing principles of the representational space:

We tested whether similarity judgments on reachspaces reflected shared setting, room, interaction locus or action. The setting, room, locus and action for each reachspace was drawn from the category name given to each image in the Reachspace Database. Prior to the analysis, author EJ confirmed that the labels matched the images. The contribution of each of these factors was assessed with a similarity prediction score: for each of 10,000 draws, a randomly selected reference image was compared to two other images: one sharing a label with the reference (according to the given factor) and the other having a different label. Apart from the constraint imposed by the labels, comparison images were randomly selected. The similarity between the reference and each comparison was assessed using the Euclidean distance between the images' embedding weights, and the similarity prediction score was calculated as the proportion of times that image which shared a label with the reference had the higher similarity.

References

1. Carroll, L. *Alice's Adventures in Wonderland* (Broadview Press, Ontario, 1832-1898 (2000)).

2. Mervis, C. B. & Rosch, E. Categorization of natural objects. *Annu. Rev. Psychol.* **32**, 89–115, DOI: 10.1146/annurev.ps.32.020181.000513 (1981).

3. Shepard, R. N. & Arabie, P. Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychol. Rev.* **86**, 87 (1979).

4. Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. Basic objects in natural categories. *Cogn. psychology* **8**, 382–439 (1976).

5. McRae, K., Cree, G. S., Seidenberg, M. S. & McNorgan, C. Semantic feature production norms for a large set of living and nonliving things. *Behav. research methods* **37**, 547–559 (2005).

6. Murphy, G. *The big book of concepts* (MIT press, 2004).

7. Hebart, M. N., Zheng, C. Y., Pereira, F. & Baker, C. I. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nat. Hum. Behav.* **4**, 1173–1185, DOI: 10.1038/s41562-020-00951-3 (2020).

8. Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).

9. Konkle, T. & Oliva, A. Canonical visual size for real-world objects. *Neuron* **37**, 1114–1124, DOI: 10.1037/a0020413 (2012).

10. Caramazza, A. & Shelton, J. R. Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *J. Cogn. Neurosci.* **10**, 64.

11. Greene, M. R. & Oliva, A. Natural scene categorization from conjunctions of ecological global properties. *Proc. Annu. Meet. Cogn. Sci. Soc.* **28**, 7.

12. Greene, M. R. & Oliva, A. Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cogn. psychology* **58**, 137–176 (2009).

13. Patterson, G., Xu, C., Su, H. & Hays, J. The sun attribute database: Beyond categories for deeper scene understanding. *Int. J. Comput. Vis.* **108**, 59–81, DOI: 10.1007/s11263-013-0695-z (2014).

14. Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M. & Fei-Fei, L. Visual scenes are categorized by function. *J. Exp. Psychol. Gen.* **145**, 82–94, DOI: 10.1037/xge0000129 (2016).

15. Oliva, A. & Torralba, A. Building the gist of a scene: The role of global image features in recognition. *Prog. brain research* **155**, 23–36 (2006).

16. Previc, F. H. The neuropsychology of 3-d space. *Psychol. bulletin* **124**, 123 (1998).

17. Josephs, E. L. & Konkle, T. Perceptual dissociations among views of objects, scenes, and reachable spaces. *J. Exp. Psychol. Hum. Percept. Perform.* **45**, 715–728, DOI: 10.1037/xhp0000626 (2019).

18. Josephs, E. L. & Konkle, T. Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex. *Proc. Natl. Acad. Sci.* **117**, 29354–29362, DOI: 10.1073/pnas.1912333117 (2020).

19. Torralba, A. & Oliva, A. Depth estimation from image structure. *IEEE Transactions on pattern analysis machine intelligence* **24**, 1226–1238 (2002).

20. Bartolo, A. *et al.* Contribution of the motor system to the perception of reachable space: an fmri study. *Eur. J. Neurosci.* **40**, 3807–3817 (2014).

21. Edelman, S. Representation is representation of similarities. *Behav. Brain Sci.* **21**, 449–467, DOI: 10.1017/S0140525X98001253 (1998).

22. Shepard, R. Toward a universal law of generalization for psychological science. *Science* **237**, 1317–1323, DOI: 10.1126/science.3629243 (1987).

23. Jozwik, K. M., Kriegeskorte, N., Storrs, K. R. & Mur, M. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Front. Psychol.* **8**, 1726, DOI: 10.3389/fpsyg.2017.01726 (2017).

24. Groen, I. I. *et al.* Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife* **7**, e32962, DOI: 10.7554/eLife.32962 (2018).

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

26. Dobs, K., Isik, L., Pantazis, D. & Kanwisher, N. How face perception unfolds over time. *Nat. communications* **10**, 1–10 (2019).

27. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. systems neuroscience* **2**, 4 (2008).

28. Tversky, B. Parts, partonomies, and taxonomies. *Dev. Psychol.* **25**, 983–995, DOI: 10.1037/0012-1649.25.6.983 (1989).

29. Landau, B., Smith, L. B. & Jones, S. S. The importance of shape in early lexical learning. *Cogn. Dev.* **3**, 299–321, DOI: 10.1016/0885-2014(88)90014-7 (1988).

30. Goldstone, R. L. The role of similarity in categorization: Providing a groundwork. *Cognition* **52**, 125–157 (1994).

31. Morris, M. R., Brush, A. B. & Meyers, B. R. Reading revisited: Evaluating the usability of digital display surfaces for active reading tasks. In *Second Annual IEEE International Workshop on Horizontal Interactive Human-Computer Systems (TABLETOP'07)*, 79–86 (IEEE, 2007).

32. Das, B. & Sengupta, A. K. Industrial workstation design: a systematic ergonomics approach. *Appl. ergonomics* **27**, 157–163 (1996).

33. Kirsh, D. The intelligent use of space. *Artif. intelligence* **73**, 31–68 (1995).

34. Scott, S. D., Carpendale, M. S. T. & Inkpen, K. Territoriality in collaborative tabletop workspaces. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, 294–303 (2004).

35. Potvin, B., Swindells, C., Tory, M. & Storey, M.-A. Comparing horizontal and vertical surfaces for a collaborative design task. *Adv. Human-Computer Interact.* **2012** (2012).

36. Josephs, E. L., Zhao, H. & Konkle, T. The world within reach: An image database of reach-relevant environments. *J. Vis.* **21** (2021).

37. Groen, I. I., Silson, E. H. & Baker, C. I. Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Philos. Transactions Royal Soc. B: Biol. Sci.* **372**, 20160102 (2017).

38. Kelemen, D. & Carey, S. The essence of artifacts: Developing the design stance. *Creat. mind: Theor. artifacts their representation* 212–230 (2007).

Acknowledgements (not compulsory)

Work reported in this publication was supported by the National Eye Institute of the National Institutes of Health under Award Number R21EY031867, by a Pershing Square Fund grant for Research on the Foundations of Human Behavior, and by the Harvard Data Science Initiative. Thanks to the Konklab writing club for their feedback on early drafts of this paper.

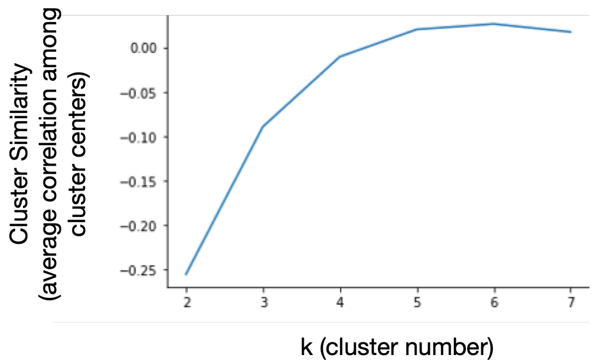
Author contributions statement

E.J. and T.K. conceived the experiments, M.H. developed and shared the modeling procedure, E.J. collected the data and analyzed the results, M.H. assisted on analyses. All authors reviewed the manuscripts.

Supplemental Materials: Emergent dimensions underlying human understanding of the reachable world

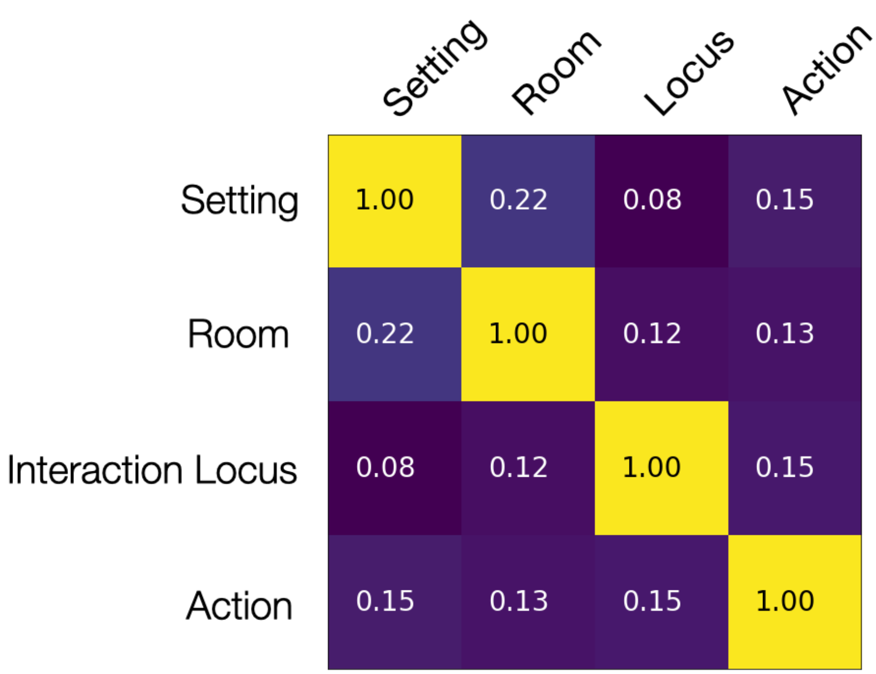
Supplemental Figure 1: Selecting k for k-means cluster analysis

Here we show the cluster similarity (operationalized as the average correlation among cluster centers) for k 2 through 7. We selected k = 5 for analysis, as the value of k at which similarity plateaus, and cluster centers become highly similar to each other.



Supplemental Figure 2: Correspondence among Setting, Room, Locus and Action Labels

The correspondence between the different labelling schemes were assessed using the Adjusted Rand Index, where 1 = perfect match, and 0 = no correspondence.



Supplemental Table 1: Task wording for validating cluster identities

Food	For this task, please indicate which images are related to FOOD. That is, images related to eating, preparing food, and food items
Electronics	For this task, please indicate which images are related to ELECTRONIC EQUIPMENT. That is, images that are related to electronics, computers, and other digital equipment
Retail	For this task, please indicate which spaces are related to RETAIL. That is, related to buying things, shopping or browsing.
Storage/Display	For this task, please indicate which images are related to STORAGE OR DISPLAY. That is, images related to putting objects away while you're not using them, or selecting which objects to interact with.
Hobbies/Crafts/Entertainment	For this task, please indicate which images show analog spaces related to CRAFTS OR ENTERTAINMENT. That is, images that are related to hobbies, creative activities, handiwork, arts, crafts, making things or playing games. Please do not include electronic spaces.
Drinks/Liquid	For this task, please indicate which images are related to DRINKS OR LIQUIDS. That is, images related to water, liquids, and objects or containers that involved liquids
Chores	For this task, please indicate which images show appliances or spaces that support cleaning, cleanup, food or drink preparation, and food or drink dispensing. Please do not include pictures related to food or eating.