Drawings of THINGS: A large-scale drawing dataset of 1,854 object concepts

Kushin Mukherjee[1,2,3], Holly Huey[4], Laura M. Stoinski[3,5,6], Martin N. Hebart[3,7,8], Judith E. Fan[9], and Wilma A. Bainbridge[10,11]

[1]Department of Psychology, University of Wisconsin-Madison

[2]Wisconsin Institute for Discovery, University of Wisconsin-Madison

[3]Max Planck Institute for Human Cognitive and Brain Sciences

[4]Department of Psychology, University of California, San Diego

[5]University of Leipzig

[6]International Max Planck Research School on Cognitive NeuroImaging

[7]Department of Medicine, Justus Liebig University

[8]Center for Mind, Brain and Behavior, Universities of Marburg, Giessen, and Darmstadt

[9]Department of Psychology, Stanford University

[10]Department of Psychology, University of Chicago

[11]Neuroscience Institute, University of Chicago

Author Note

Correspondence should be sent to kushinm11@gmail.com

Abstract

The development of large datasets of natural images has galvanized progress in psychology, neuroscience, and computer science. Notably, the THINGS database constitutes a collective effort towards understanding of human visual knowledge by accumulating rich data on a shared set of visual object concepts across several studies. In this paper, we introduce **`Drawing of THINGS`** (**`DoT`**), a novel dataset of 28,627 human drawings of 1,854 diverse object concepts, sampled systematically from concrete picturable and nameable nouns in the American English language, mirroring the structure of the THINGS image database. In addition to data on drawings' stroke history, we further collected fine-grained recognition data for each drawing, along with metadata on participant demographics, drawing ability, and mental imagery. We characterize people's ability to communicate and recognize semantic information encoded in drawings and compare this ability to their ability to recognize real-world images of the same visual objects. We also explore the relationship between drawing understanding and the memorability and typicality of the objects contained in THINGS. In sum, we envision **`DoT`** as a powerful tool that builds on the THINGS database to advance understanding of how humans express knowledge about visual concepts.

Keywords: drawing, big data, visual concepts, object recognition, memory, typicality

Drawings of THINGS: A large-scale drawing dataset of 1,854 object concepts

Introduction

A central goal in the cognitive sciences is to understand how semantic knowledge is organized in the mind and brain. *Visual* semantic knowledge (S. Thompson-Schill, Aguirre, Desposito, & Farah, 1999; Warrington & Shallice, 1984) in particular is implicated in object recognition (DiCarlo, Zoccolan, & Rust, 2012; De Lange, Heilbron, & Kok, 2018; Gauthier & Tarr, 2016; Huth, Nishimoto, Vu, & Gallant, 2012; T. Rogers, Hodges, Patterson, & Lambon Ralph, 2003), the encoding of objects in memory (T. T. Rogers et al., 2004; Konkle, Brady, Alvarez, & Oliva, 2010; Cunningham & Wolfe, 2014), and organizing perceived similarity among objects (Mur et al., 2013; Charest, Kievit, Schmitz, Deca, & Kriegeskorte, 2014; Tversky, 1977; Rosch, 1975; Muttenthaler et al., 2022; Hebart, Zheng, Pereira, & Baker, 2020). Recent years have seen important advances in the elucidation of the representational structure of visual semantic knowledge — how concepts are situated with respect to each other and the features that shape this organization (Hebart et al., 2020; Fu et al., 2023; Caplette & Turk-Browne, 2024; Mur et al., 2013). This has been facilitated by two key advances: (1) the development of large image datasets that are representative of the object concepts that people are familiar with (Hebart et al., 2019, 2023; Mehrer, Spoerer, Jones, Kriegeskorte, & Kietzmann, 2021; Allen et al., 2022) and (2) the dense annotation and augmentation of the items within these datasets with behavioral and neural measurements (Stoinski, Perkuhn, & Hebart, 2023; Grootswagers, Zhou, Robinson, Hebart, & Carlson, 2022; Kramer, Hebart, Baker, & Bainbridge, 2023; Hansen & Hebart, 2022; Vanasse et al., 2022). Experimental results based on such datasets are valuable insofar as conclusions derived from them are not restricted to a handful of experimenter-curated stimuli and are much more likely to generalize broadly. The use and development of these datasets have also contributed to a more cohesive body of research with methods and data that are interoperable across studies, a strategy that has been fruitful in other fields, especially computer science (Deng et al., 2009; Schuhmann et al.,

2022; Everingham, Van Gool, Williams, Winn, & Zisserman, 2010; Sharma, Ding, Goodman, & Soricut, 2018; Lin et al., 2014).

Here we harness the study of *drawing production and recognition* at scale to advance understanding of visual semantic knowledge. While open-ended verbal production, such as generating feature lists for concepts, has been central to investigations of semantic cognition for many decades (McRae, Cree, Seidenberg, & McNorgan, 2005; Devereux, Tyler, Geertzen, & Randall, 2014; De Deyne et al., 2008), there has been a rapid recent development of methods for using visual production to investigate the content and organization of conceptual knowledge (Mukherjee & Rogers, 2024; Mukherjee, Hawkins, & Fan, 2019; Long, Fan, Huey, Chai, & Frank, 2024; Fan, Bainbridge, Chamberlain, & Wammes, 2023). Drawing tasks have been used in studies of perception (Biederman & Ju, 1988; Sayim & Cavanagh, 2011; Yang & Fan, 2021), development (Long, Fan, Chai, & Frank, 2021; Long et al., 2024; Dillon, 2021), learning (Fan, Yamins, & Turk-Browne, 2018; Chamberlain, Kozbelt, Drake, & Wagemans, 2021), memory (Bainbridge, Pounder, Eardley, & Baker, 2021; Bainbridge, Hall, & Baker, 2019; Megla, Rosenthal, & Bainbridge, 2025; Bozeat et al., 2003), and others. The importance and utility of using line drawing-based pictorial representations of objects as a tool for understanding visual and semantic cognition can be seen in the widespread adoption of even early datasets such as the one curated by (Snodgrass & Vanderwart, 1980) in the study of human (S. L. Thompson-Schill, D'Esposito, Aguirre, & Farah, 1997; Patterson, Nestor, & Rogers, 2007; Alvarez & Cavanagh, 2004; ?, ?; Singer, Cichy, & Hebart, 2023) and machine (Kubilius, Bracci, & Op de Beeck, 2016; Singer, Seeliger, Kietzmann, & Hebart, 2022; Mukherjee et al., 2024) vision. Due to their abstract nature and divergences in low-level visual features relative to real-world objects, drawings have also served as key test items for the cognitive benchmarking of artificial vision and learning systems (Singer et al., 2022; Mukherjee et al., 2024; Mukherjee & Rogers, 2024; Boutin et al., 2023; Wang, Ren, & Zemel, 2021; Lake, Salakhutdinov, & Tenenbaum, 2015). Thus, drawings provide ways to

probe human visual knowledge using open-ended responses that complement the closed-set response space used in standard recognition and discrimination tasks. This open-ended nature of drawing is important because it can provide insight into the contents of human mental representation in trial-efficient ways. For example, Bainbridge et al. (2019) used a drawing-based approach to ask participants to reproduce visual scenes from memory. The use of drawing as a medium allowed the researchers to detect when participants failed to reproduce specific objects or when they mis-remembered where objects were located within the scene. Drawing tasks have also been used to characterize the enrichment of visual concept knowledge throughout middle childhood (Long et al., 2024), and to characterize what parts of objects people deem relevant to include in a drawing depending on their communicative goals (Fan, Hawkins, Wu, & Goodman, 2020; Mukherjee et al., 2019; Huey, Lu, Walker, & Fan, 2023).

Several drawing datasets already exist that were developed using web-based tools (De Leeuw, 2015; Bainbridge, 2022) and crowdworking platforms, such as Amazon Mechanical Turk (AMT). Sheepmarket (Koblin, 2009), a dataset of over 10,000 drawings of sheep collected from such a set of crowd workers, was an early proof of concept that reasonably high quality drawing data can be collected through this medium. Later, Eitz, Hays, and Alexa (2012) developed the TU-Berlin dataset representing 250 classes, but with over 20,000 drawings in the dataset drawn by non-experts on Amazon Mechanical Turk (AMT), increasing the overall semantic diversity and scale of drawing datasets. A similar approach of collecting crowdsourced drawings was adopted by (Sangkloy, Burnell, Ham, & Hays, 2016), who collected over 75,000 unique drawings of 125 categories to build the Sketchy dataset. A unique aspect of this latter dataset was that that each drawing was based on a real-world photograph allowing for comparisons of human and machine visual recognition of the common objects across modalities. Google's Quick! Draw game (Jongejan, Rowley, Kawashima, Kim, & Fox-Gieg, 2017), an online game where people could make drawings and have an AI system guess a label for it, led to the development of

a dataset of over 50 million sketches of 345 categories, the first system to allow for training data-intensive deep learning models (Ha & Eck, 2017). Long et al. (2024) recently introduced a drawing dataset of over 37,000 drawings of 48 object classes made by children between the ages of 2 and 10, allowing for investigations into changes in the structure of visual concepts across development. These datasets have been instrumental in galvanizing progress towards understanding how humans flexibly deploy their visual knowledge to create visual representations that abstract and distill from their real-world counterparts, yet nevertheless often suffice for effective communication (Hawkins, Sano, Goodman, & Fan, 2023). While these prior approaches have laid critical theoretical and methodological groundwork, they are nevertheless limited in several ways. First, despite the largest of these datasets being on the scale of several million individual drawings, the diversity of the object classes in Quick Draw! falls short of the range of visual concepts represented in modern investigations of visual concepts (Hebart et al., 2023) nor do they capture the *kinds* of object categories that align with human visual experience (Mehrer et al., 2021). Second, it is difficult to relate findings on drawings of the visual objects in these datasets to findings about the same objects more generally due to the absence of commensurate large-scale data either on the drawing or real-world image side. A path forward is to integrate drawing dataset collection efforts within a larger ecosystem of research findings grounded in shared test items.

Here, we introduce **Drawings of THINGS** (**DoT**), a dataset of over 28,000 drawings of 1,854 visual objects that adopts such an integrative approach. **DoT** differentiates itself from prior datasets most saliently along the axis of semantic diversity. A critical feature of this dataset lies in the variety of semantic categories represented, leveraging the richness of THINGS (Hebart et al., 2019, 2023), a database of 26,107 images of 1,854 object concepts with psychologically relevant metadata. Importantly, the objects in THINGS were carefully sampled from an extensive set of concrete nouns in American English, ensuring a selection that is both broad and distinctive, while focusing on the most common and easily

identifiable concrete concepts (e.g., dog, sweater, acorn). Accompanied by extensive metadata and object property ratings (Stoinski et al., 2023; Hebart et al., 2023), THINGS supports the precise selection of stimuli, conditions, and control variables. Additionally, the growing number of publicly available datasets and studies utilizing the THINGS concepts (Hebart et al., 2023; Grootswagers et al., 2022; Gifford, Dwivedi, Roig, & Cichy, 2022; Kramer et al., 2023; Mukherjee et al., 2024; Dobs, Martinez, Kell, & Kanwisher, 2022; Hebart et al., 2020; Benchetrit, Banville, & King, 2023; Papale, Wang, Self, & Roelfsema, 2025) facilitates comparisons of findings across different disciplines, methods, and species. As a result, `DoT` constitutes a new semantically diverse drawing dataset that supports asking a multitude of questions about visual knowledge, while dovetailing with the rich THINGS database ecosystem that allows us to connect each drawing to corresponding object photographs, metadata, and neural data.

In addition to the drawings themselves, we also collected fine-grained human recognizability data on each drawing and on photographs of the same objects in the THINGS database. For each drawing, we also collected stroke-order information that could be used as a proxy for what information is prioritized first or last in drawing a given object. This form of temporal stroke data has been critical in building computational models of human concept learning (Lake et al., 2015). We also include data on each sketcher's demographics, drawing ability, imageability, and educational background. We demonstrate how this dataset can be used to test hypotheses about visual semantic knowledge and its relationship to other aspects of visual cognition. Lastly, we make available all our data for the community to develop their own studies and test a larger array of hypotheses.

Methods

Experiment 1: Drawing production study

Our first goal was to collect a large-scale dataset of drawings of the 1,854 objects present in the THINGS database. Similar to recent studies employing drawing-based methods (Yang & Fan, 2021; Bainbridge et al., 2019; Hawkins et al., 2023; Fan et al., 2020;

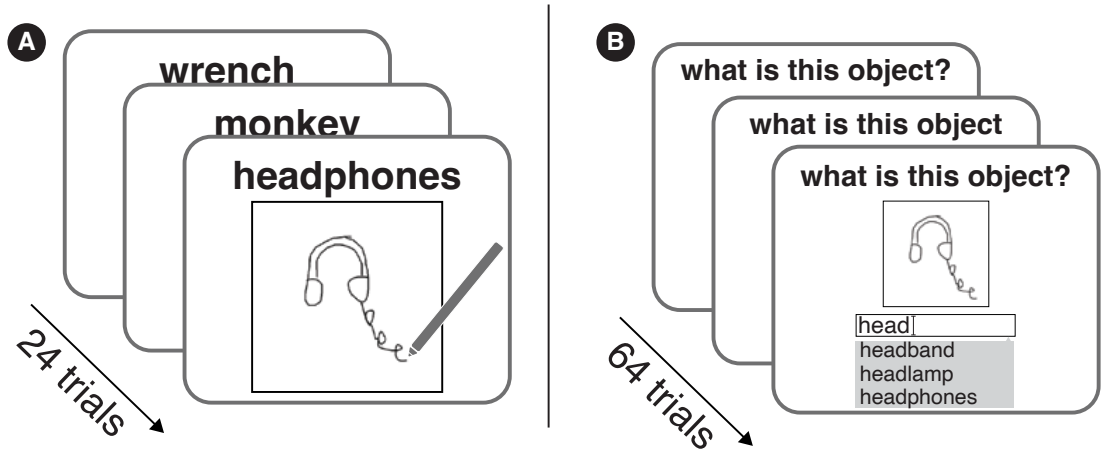

Figure 1

(A) In experiment 1, each participant made drawings of 24 unique objects. (B) In experiment 2, separate participants provided up to 5 labels for each of 64 drawings sampled from the drawings made in experiment 1.

Megla, Rosenthal, & Bainbridge, 2024; Mukherjee et al., 2024), we used Prolific, an online crowdsourcing platform, to collect multiple drawings of each of the THINGS objects.

Participants

1,316 participants were recruited on the online experimental platform Prolific. Participants were allowed to complete multiple sessions contingent on maintaining satisfactory ratings. We excluded data from 2 participants due to technical errors in data collection, leaving a sample size of 1,316 participants (735 Male, 511 Female, 68 Other/Did not wish to say [1]; M_{age}= 37.04). All participants provided informed consent in accordance with the University of California San Diego Institutional Review Board (IRB) and were compensated at $15/hour.

[1] Of the 1,314 participants in the drawing production study, 50 did not complete all trials including the demographic information trial. Our gender and age breakdowns do not include data from these participants.

Procedure

The experiment was designed using jsPsych version 7 (De Leeuw, 2015) using a variant of its sketchpad plugin. On each trial of the experiment, participants were presented with a blank 550px × 550px canvas surrounded by a black border of size 2px and an object word above the canvas. Object words were obtained from the names of the 1,854 objects contained within the THINGS database, which capture most concrete nameable objects in human experience (e.g., apple, chinchilla, laptop). In cases where the label was ambiguous (such as 'bat' the animal or 'bat' the sporting good), additional disambiguating text, obtained from the THINGS+ dataset (Stoinski et al., 2023), was presented in parentheses (e.g., bat (animal)). We appended such disambiguating text to 204 out of the 1,854 concepts. The participants' task was to use their touch pad, mouse, or any other cursor device to make a drawing of the prompted object such that a naive viewer would be able to guess the object's label from looking at the drawing. Participants could make their drawings only in a single color (black) and were given the option to undo their most recent stroke, redo the stroke in case they accidentally undid an action, and also to clear the entire canvas if they wanted to restart their drawing. The undo, redo, and clear actions were implemented as buttons near the bottom of the canvas. We only allowed participants to participate through a desktop or laptop computer, asking them to retry if they logged into our task using a phone or tablet. Participants were free to spend as much time on each drawing as they wanted and had to at least make one stroke to be allowed to proceed to the next trial. There was a button to 'submit' the drawing and continue on to the next trial when they were done with each drawing. In addition to saving the image of the drawings made by participants, we also saved the order of strokes made by participants (Figure 2).

After each drawing trial, participants were asked whether they knew what the prompted object looked like in real life, which they could answer with either 'yes' or 'no'. This was to flag potential low quality drawings that were a result of the sketcher not being familiar with the queried object. Each participant completed 24 drawing trials, with the

median time taken to complete the experiment being 21 minutes. Before completing the experiment, participants completed a questionnaire where they reported their demographic information including age, gender, ethnicity, state of residence (optional) and level of education. They were also asked to answer 7-point Likert scale questions on self-reported drawing skill and mental imagery capabilities ("*When you try to form a mental picture, it is usually:*", where the endpoints of the scale were 'no image' and 'very clear').

Filtering for high quality drawings

We collected a total of 29,876 drawings with an average of 16.2 drawings per THINGS object ($SD = 4.69$), with the smallest number of drawings collected for an object being 12 and the largest being 41. In order to filter out drawings that were inappropriate, not relevant to the prompted object label, or blank canvases, we had a team of 6 internal annotators rate the validity of each drawing given the prompted object. Not all annotators saw each drawing but each drawing was shown to at least 3 annotators. The annotators were asked to indicate if a given drawing was 'valid' or 'invalid' according to the following specific criteria – invalid drawings were those that consisted of (1) illegible scribbles, (2) text written out in the form of a drawing, or (3) offensive imagery. Since there is some degree of subjectivity to these criteria, we collected at least 3 validation ratings for each of the 29,876 drawings. We chose to only include drawings in our final dataset that were rated as 'valid' by the majority of the annotators, i.e., at least 2 out of 3. This criterion allowed for screening out of obviously bad drawings while remaining conservative in terms of data exclusion. This led us to remove 1,249 drawings (4.18% of the total number of drawings), which yielded a final dataset of 28,627 drawings.

Experiment 2: Drawing recognition study

After collecting the dataset of drawings, we ran a second study with the aim of characterizing patterns of recognition behavior for these drawings. To this end, we conducted an experiment where a new cohort of participants had to provide object labels for each of the drawings in our dataset.

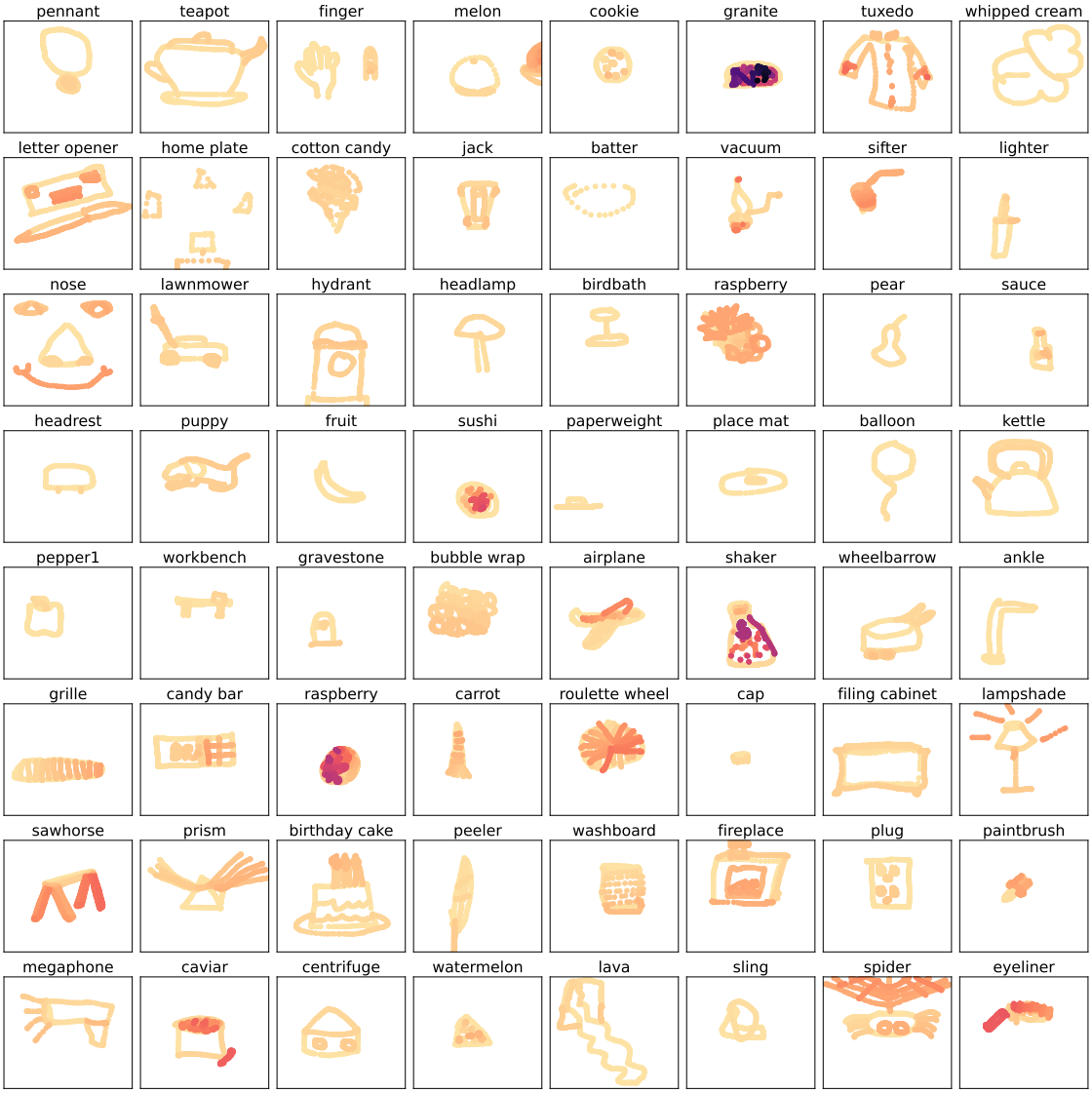

Figure 2

Sample Drawings from ***DoT*** *rendered stroke-by-stroke. Darker strokes indicate those made later during the drawing process.*

Participants

1,578 participants were recruited on Prolific to complete the drawing recognition study. Participants were allowed to complete multiple sessions contingent on maintaining satisfactory ratings. 13 participants either did not complete even a single recognition trial or faced technical difficulties during the experiment. We excluded an additional 8 participants who failed to provide reasonable labels for a catch trial embedded within normal recognition trials, which depicted a drawing of a cat. Our final sample consisted of 1,557 participants (744 Male, 674 Female, 32 other, 6 did not wish to say [2]; M_{age}= 39.96). All participants provided informed consent in accordance with the UC San Diego and University of Chicago IRBs. Participants were compensated at $8/hour.

Stimuli

The stimuli were the 28,627 drawings in our dataset that remained after exclusion of poor quality drawings. The set of all possible labels were each of the THINGS 1,854 object labels (with the appropriate disambiguating text in parentheses for the relevant objects).

Procedure

Participants were tasked with providing semantic labels (i.e., object names) for the drawings. The labels they could provide were restricted to objects in the THINGS database. On each trial of the experiment, participants were provided with a random drawing from experiment 1 and text above the drawing asking 'What is this object?'. There was an empty text box below the drawing where the participant could type an answer. As the participant typed, several candidate objects would be populated in a drop-down menu based on partial string matches to the THINGS objects. The participant had to select one of the possible drop-down labels and could not enter a label that was not in the set. Given that some drawings could be ambiguous or may evoke more than one

[2] Of the 1,557 participants in the recognition study, 101 did not complete all trials including the demographic information trial. Our gender and age breakdowns do not include data from these participants.

semantic label, we allowed participants to provide up to five labels, although only one label was required to progress to the next trial. Before starting the experiment, participants completed a practice trial to familiarize themselves with the interface. All participants were shown the same researcher-made drawing of a cat during this practice trial. While we did not exclude any participants based on whether they correctly labeled this practice trial, 95.89% of the participants provided the exact label 'cat' and 99.73% of participants provided a label that was either 'cat', 'kitten', 'fox', or 'dog' showing that our interface was intuitive to use and participants generally understood the task. The same drawing was presented once more during the actual experimental trials as a 'catch trial'. Only 8 participants (0.5%) failed to pass this check. To begin the experiment, participants had to correctly answer three comprehension check questions that tested that they knew not to enter multiple labels of the same object, not to leave the response text box empty, and to answer with their best guess when they weren't sure of the object depicted in the drawing. Each participant completed 64 recognition trials where they were never shown more than one drawing of each THINGS object. Before completing the experiment, participants completed a questionnaire where they reported their demographic information including age, gender, ethnicity, and highest level of education. They also reported their mental imagery using the scale described in experiment 1.

Drawings from each object were labeled by an average of 55.67 participants, with each individual drawing being labeled by an average of 3.63 participants. On average, participants provided 1.52 ($SD = 0.57$) labels per drawing with a maximum of 5 labels provided for a single drawing. Participants spent an average of 18.01 seconds per recognition trial (SD=2.65 seconds).

Experiment 3: Image recognition study

We additionally collected a dataset of recognizability scores for all photographs included in the THINGS database (Hebart et al., 2019; Stoinski et al., 2023). Similar to the drawing recognition study, participants were instructed to select the most fitting label

for the object depicted in each photograph from a drop-down menu containing the 1,854 THINGS object labels. The goal of this experiment was to provide complementary recognition data on the same visual objects when depicted using photographs as opposed to line drawings.

Participants

A total of 13,158 participants were recruited via the crowdsourcing platform Amazon Mechanical Turk. All participants resided in the US and provided informed consent in compliance with the Ethics Committee of the Medical Faculty of Leipzig University, Germany. Participants received a small compensation for each completed HIT and could choose to participate in as many HITs as they wished.

To ensure data quality, several exclusion criteria were applied during and after data collection. Participants were flagged as potentially non-compliant if they completed five trials in under 1,100 ms, or all ten trials in under 1,300 ms. Additionally, participants were flagged as non-compliant if, in five or more trials, they selected one of the top drop-down options after typing only a single letter in the drop-down menu search bar. Individuals flagged as potentially non-compliant on two occasions were prevented from further participation using a custom function embedded in the experimental script. Each HIT included one easily recognizable catch image. After data collection, participants were flagged as non-compliant if they mislabeled the catch images in 50% or more of their HITs. Workers flagged as non-compliant had all their data excluded from the analysis. Following these exclusions, 8,094 individuals remained (4,607 female, 3,395 male, 92 diverse), ranging in age from 18 to 98 years ($M = 37.81$, $SD = 11.25$). On average, participants completed 6.10 HITs ($SD = 58.94$).

Stimuli

Participants provided labels for all 26,107 images in the THINGS dataset (Hebart et al., 2019), along with the 1,854 public domain images provided in THINGS+ (Stoinski et al., 2023).

Procedure

A single image recognizability experiment (human intelligence task, HIT) consisted of 10 trials. In each trial, participants were presented with an image and asked to select the most appropriate label for the depicted object from a drop-down menu listing the 1,854 THINGS objects (Hebart et al., 2019). Similar to the drawing recognition experiment, the labeling interface allowed participants to begin typing with the dropdown auto-suggesting labels based on partial matches. For objects with ambiguous meanings, additional context was provided in parentheses. Here, participants provided only a single label per image.

Each HIT included one easily nameable catch image in a randomly assigned trial, assuming that participants who failed to label this image likely were noncompliant in completing the experiment. As catch images, we selected 2,796 unique THINGS images that were correctly named in 100% of cases during the free label generation task of the THINGS+ nameability experiment (Stoinski et al., 2023). Initially, each image was sampled 20 times. After excluding trials from noncompliant participants, we conducted a follow-up collection to fill up the number of samples per image back up to 20. As every HIT had to include one catch image, some catch images were sampled slightly more often. After data exclusion, each individual image had been labeled between 5 and 23 times (M = 14.70, SD = 2.40). Averaged across the 13 or more image examples per object, each object was labeled between 165 to 532 times (M = 221.71, SD = 41.44).

Results

The final **`DoT`** dataset consisted of 28,627 drawings belonging to 1,854 object classes. Each drawing is associated with a 550px × 550px raster image representation of the drawing, a history of strokes used to make the drawing, which can be used to re-render the drawing stroke-by-stroke, the amount of time the participants took to complete the drawing, the mean recognizability of that drawing [3], and whether the sketcher was familiar

[3] We report recognizability both based on the first guess (top-1 guess) or the best guess amongst all labels entered by a participant (top-k guess).

with the object or not. We additionally provide embeddings for each drawing obtained from SigLIP (Zhai, Mustafa, Kolesnikov, & Beyer, 2023), a state-of-the-art multimodal transformer model as described in the following section.

Dataset Diversity and Semantic Selectivity

Computer vision models trained for object recognition have been shown to be good models of human vision across a variety of behavioral and neural metrics (Yamins et al., 2014) (cf. (Bowers et al., 2022)). Recent years have seen the emergence of alternative self-supervised training objectives such as visual contrastive learning (Konkle & Alvarez, 2022; Zhuang et al., 2021; Chen, Kornblith, Norouzi, & Hinton, 2020) and image-language contrastive learning (Radford et al., 2021), and transformer-based vision models are competitive with and even surpass convolutional neural networks at aligning with human behavior despite sharing fewer inductive biases with primate vision (Conwell, Prince, Kay, Alvarez, & Konkle, 2024). We leveraged such a transformer-based multimodal model SigLIP (Zhai et al., 2023), which was trained on web-scale corpora of images and text data, and computed embeddings for each of the drawing images to test the diversity of their representations. For simplicity, below, we report results on drawings belonging to the 27 superordinate 'core' categories identified by Hebart et al. (2019) (e.g., food, furniture, clothing). We first selected drawings of objects that belonged to the 27 categories. This resulted in a set of 13,802 drawings that we show in Figure 3 A. Each point corresponds to the position of a single object obtained by averaging the SigLIP feature vectors for all drawings of that object and then projecting the high dimensional vectors into a 2-dimensional space using the t-stochastic neighbor embedding algorithm. The mean drawing-to-drawing similarity expressed by SigLIP reveals that drawings of some objects appear highly clustered (e.g., animals, birds, clothing, and weapons) while drawings of other categories appear more dispersed (e.g., tools and food). Overall, it does appear that `DoT` qualitatively spans a reasonably large degree of visual variety while organizing objects in a semantically coherent manner.

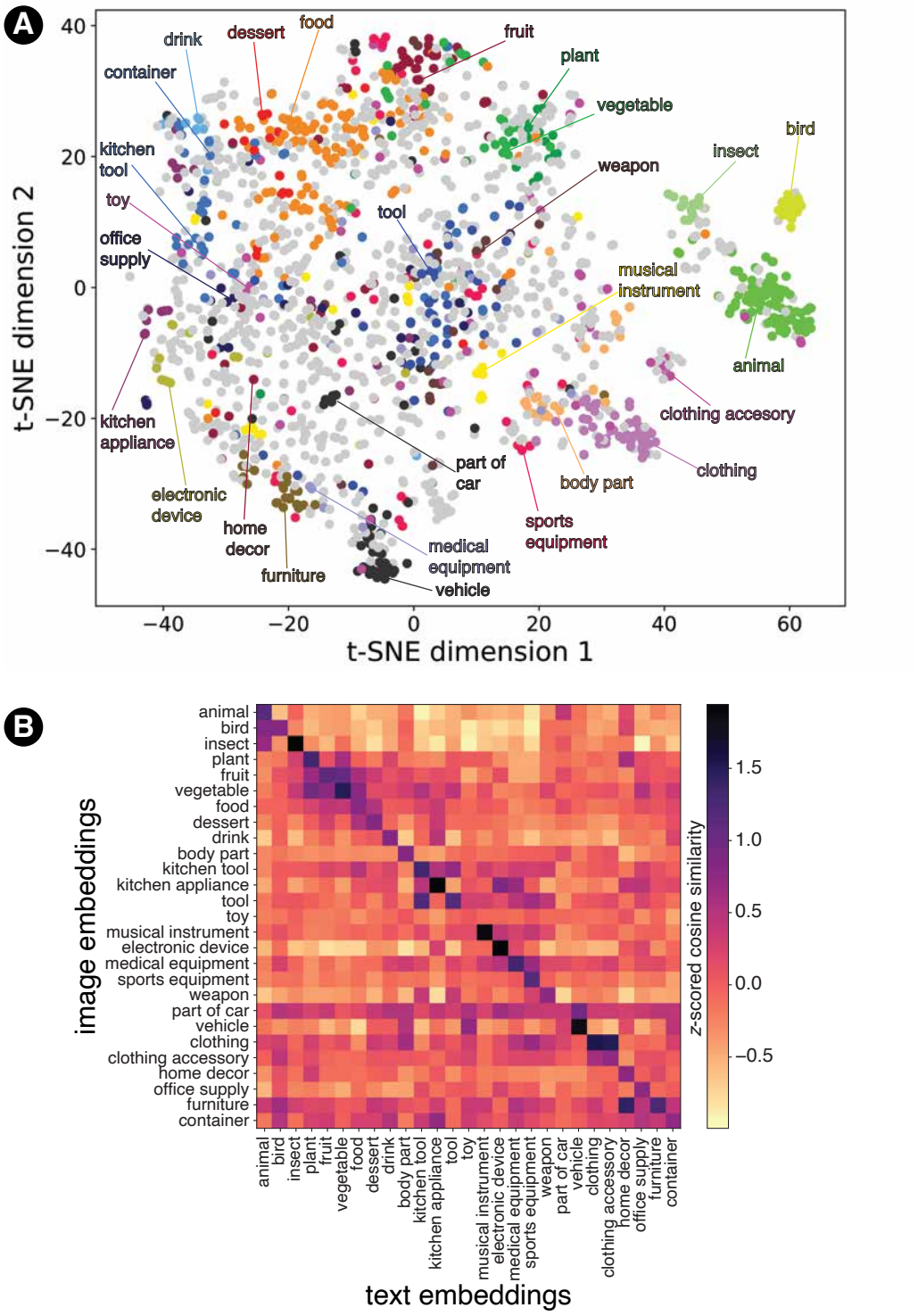

Figure 3

(A) t-SNE Visualization of the mean SigLIP representation of drawings of each of the 1,854 THINGS categories. Objects belonging to the 27 core categories are highlighted in different colors. (B) Similarity between the mean SigLIP image embeddings of drawings from the 27 core categories with each of the text embeddings for the 27 core category labels.

To quantify this semantic coherence and measure the extent to which the drawings serve as good candidates for the core categories they belong to (without looking at the human recognition data), we once again used SigLIP to estimate the extent to which the visual embeddings of drawings belonging to each category aligned with the text embeddings for its category label. We conducted these analyses at the level of the core categories to align closely with the analyses in Hebart et al. (2019), where the authors showed that semantic embeddings of the THINGS objects showed category selectivity for the core categories. For each drawing, we computed the dot product between its SigLIP image embedding and the SigLIP text embedding of each of the 27 core categories. This allowed us to compute the difference between image-to-label similarities between each drawing and its target category and the mean similarities to the remaining 26 categories. If this difference is positive, it constitutes a measure of how selectively that drawing evoked the target core concept relative to the others. We conducted a permutation test where we scrambled the category labels and repeated the procedure 1,000 times to estimate statistical significance. Figure 3 B shows the cosine similarities between the average image embedding for each category (averaged over all drawings belonging to that category) and the text embeddings for each of the category labels. Similarities were z-scored to put them on the same scale. We found that, for the majority of categories, their drawings were significantly selective for the true label except for 'part of car' ($p < 0.001$ for 26 out of 27 categories after Bonferroni corrections for multiple comparisons).

Thus, using pure machine vision metrics, we show that **`DoT`** exhibits a high degree of category selectivity indicating that the drawings visually evoke the core categories they are made to evoke. For the category that didn't show high selectively for its true labels (part of car), this is perhaps expected given that many parts of cars were closer to the label 'vehicle' relative to the true label. In subsequent sections we investigate to what extent this selectivity holds true for human observers and whether they can discern semantic structure in drawings at even finer-grained levels of analysis.

Recognizability of object drawings in `DoT`

Having established that **`DoT`** is visually diverse and that individual drawings broadly adhere to the semantic categories of the objects they depict, we turned our attention to investigating the degree to which human observers could recognize the semantic information that sketchers had encoded in their drawings. A representative measure of the semantic information within a drawing is the object label assigned to it by participants in our drawing recognition study (experiment 2). Thus, we began by measuring how often participant-provided labels were exact matches to the true label for each drawing. We compared the object-level recognizability scores derived from drawings to recognizability scores derived from photographs of the same objects (experiment 3) in order to characterize the extent to which drawings of objects were more or less efficient at conveying semantic information relative to real-world exemplars.

In order to measure how often the object label inferred by an observer matched the label provided to the sketcher during the production task, we first computed the proportion of guessed labels that were an exact match to the label shown to the sketcher of each drawing. We measured this at two levels - (1) by testing if the first label provided by each participant in the recognition experiment matched the true label (top-1 accuracy) and (2) by testing whether *any* of the labels provided by each participant matched the true label (top-k accuracy). For the image recognition experiment, we computed an accuracy metric where we allowed homonyms of the correct label to also be considered 'correct', assuming that they were likely selected by mistake; for example, if 'button (clothes)' was chosen instead of the intended label 'button (device)'. We report results using the latter score.

We found a high degree of variability among the recognizability of drawings of objects, with the top-1 accuracy spanning the entire range of possible values (M=21.79%; SD=22.84% max = 96.15%, min = 0%). We observed a similar trend for top-k accuracy with an expected slight increase in accuracy (M=26.50%; SD=24.39%; max=100%, min = 0). Moving forward, when we refer to recognizability for drawings, we are generally

referring to the top-k metric unless otherwise specified. While drawings of objects like 'ladder','snowman', and 'cactus' were almost always recognized and labeled correctly, others like those of 'mongoose', 'mint', and 'moccasin' failed to *ever* elicit the correct label (Figure 5).

Since sketchers self-reported their familiarity with each prompted object label during the drawing trials, we calculated the mean object recognizability separately for trials where they recognized the label and for those where they did not. We found that on average drawings of objects where the sketcher was familiar with the object label were more recognizable ($M = 27.42\%$, $SD = 24.67\%$) than drawings of objects where the sketcher was not familiar with the object label ($M = 11.49\%$, $SD = 22.98\%$). Thus, perhaps unsurprisingly, familiarity with the prompted object led sketchers to produce more recognizable drawings. The fact that recognizability for drawings for the 'unfamiliar' group was not 0 indicates that some sketchers may have either misunderstood what the question was asking or that even not knowing what the object looks like in 'real life' did not preclude them from possessing the requisite knowledge to make a drawing that nevertheless conveyed the visual concept.

In order to evaluate whether the low recognizability scores could be attributed to the fact the drawings were made by non-expert artists and simply provided insufficient visual information to inform a correct guess, we compared the recognizability of each object as measured using drawings with scores for the same objects when assessed using images. Overall, recognition accuracy for images of objects ($M = 52.65\%$, SD= 24.41%) was higher than recognition accuracy for drawings of objects ($M = 26.50\%$, SD= 24.39%) ($t(1,853)$=39.49, $p < 0.001$). However, since there was comparable variance in recognizability for both domains, the key question was whether the objects that had low recognizability scores for drawings were the same ones that had low recognizability scores for photographs. To test this, we first estimated a noise ceiling for how reliable the rank-ordering of object recognizability was. This was done by first generating two samples

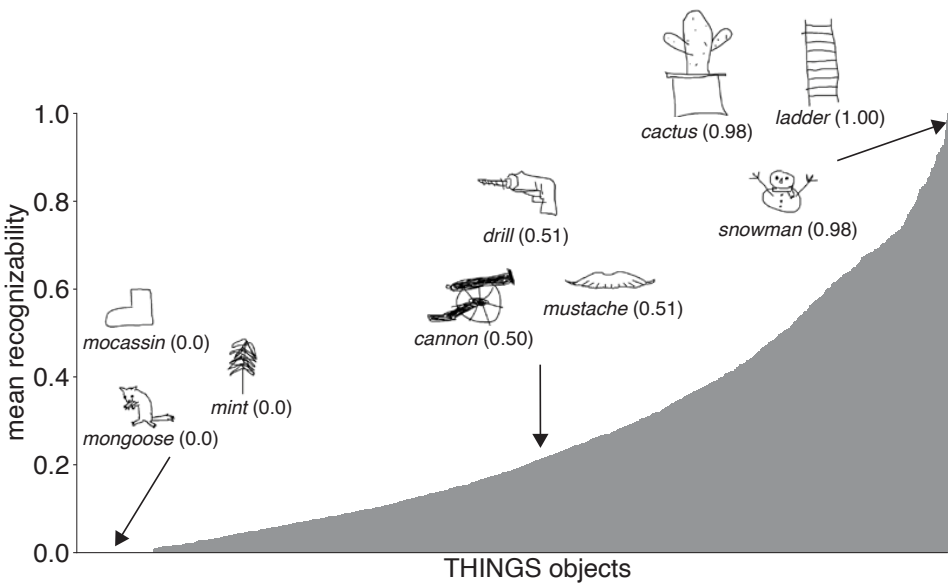

Figure 4

Distribution of mean drawing recognizability across all 1,854 THINGS concepts sorted from least recognizable to most recognizable. Numbers in parentheses next to highlighted examples correspond to mean top-k accuracy for that object.

of the image recognizability data by sampling 20 recognition trials for each object with replacement and computing the Kendall's τ between the rank ordering of objects in terms of their mean recognizability scores. This procedure was repeated 1,000 times using different splits of the data to estimate a mean Kendall's τ ($M = 0.89$, $SD = 0.002$). This constitutes a reliability measure of the rank ordering of objects based on photograph recognizability and the best that we can expect for the rank ordering of object recognizabilty between drawings and photographs. Using this measure, we next computed a 'noise-corrected' rank-order correlation between objects' drawing recognizability scores and image recognizability scores. Generally we found that objects that were difficult to recognize as drawings were also difficult to recognize in photographs (corrected Kendall's $\tau(1,853) = 0.42$, $p < 0.001$) (Figure 5 A).

However, there were several cases that diverged from this general trend. The points

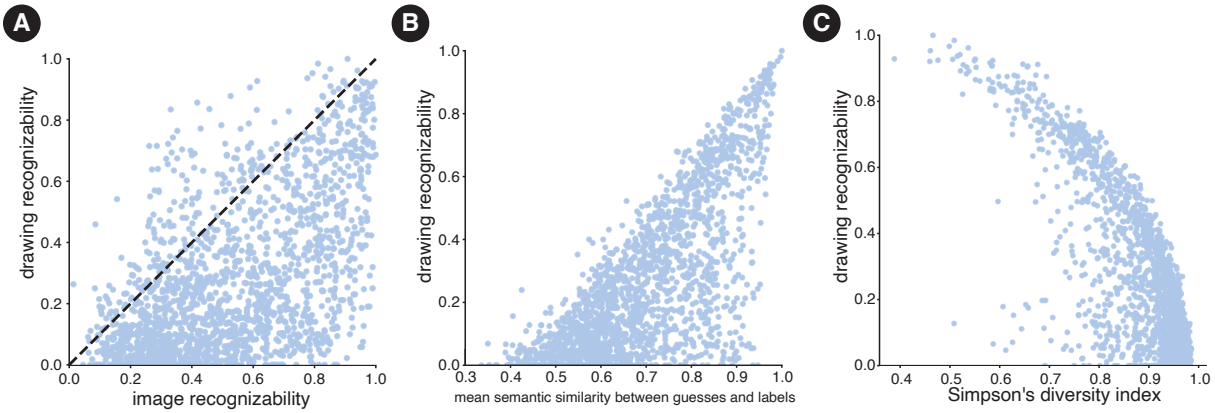

Figure 5

Each point in each of the three plots corresponds to a single THINGS object. (A) Recognizability of drawings of objects as a function of recognizability of images of the same objects (B) Drawing recognizability as a function of the mean similarity between guessed labels and the true labels, averaged over drawings for each object (C) Drawing recognizability as a function of label diversity.

above the identity line in Figure 5 A correspond to such objects that were more easily recognized as drawings than as photographs. To quantify the degree to which objects were easier to recognize as drawings relative to images, we computed a recognizability difference score by subtracting the average recognizability score for images of an object from the recognizability score for drawings of that object ($recognizability_{drawing} - recognizability_{photo}$). Objects that had a positive score for this metric were those that were easier to recognize as drawings relative to photographs.

Objects such as 'ticktacktoe' (0.48), 'reindeer' (0.44), 'palm tree' (0.44), and 'popsicle' (0.43) were among the objects that had the highest drawing − image recognizability scores. Overall, 211 of the 1,854 (11.38%) objects were more easily recognized as drawings relative to photographs. At the same time several objects were *harder* to recognize when rendered as drawings as opposed to photographs (e.g., 'chocolate'(-0.89), 'tiger' (-0.68), and 'frog'(-0.56)).

Thus, even when considering all guesses that the participants provided for each

drawing, we found a high degree of variability in terms of how easily object drawings were recognized. We additionally found that, while the recognizability of real-world instances of objects are predictive of drawing recognizability, there exist objects that are more efficiently communicated using drawings over photographs than others. We next considered whether the incorrect labels might also contain signatures of people's visual knowledge of the queried objects.

Diversity of guessed labels predicts lower drawing recognizability

The previous section established that there was variability in the degree to which drawings of objects were accurately labeled. We measured whether this diversity in labels generated by participants for drawings of objects was predictive of that object's recognizability. The rationale was that to the extent that label diversity does *not* predict recognizability, this would suggest that some drawings systematically convey the wrong object. Alternatively, if greater label diversity was predictive of lower recognizability then drawings might be evoking many different objects in the eyes of an observer (e.g., drawings of dogs, a four-legged animal, might be labeled as other four legged mammals such as cow, horse, zebra, etc.).

For each drawing, we operationalized the label diversity using Simpson's diversity index (SDI) (Simpson, 1949). Concretely, the SDI for each drawing was given by —

$$SDI = 1 - \frac{\sum_i n_i(n_i - 1)}{N(N-1)}$$

where N refers to the total number of labels provided for the sketch across participants, including the optional 2nd to 5th guesses, and n_i refers to the number of guesses made for the ith object. The more dispersed the guesses for an object are across all the possible labels (i.e., the more participants disagree on what is depicted in the drawing), the closer the SDI is to 0. Conversely, the more concentrated the guesses are around a few labels (i.e., the more participants agree on what is depicted in the drawing), the closer it is to 1. In general, we found that guesses tended to be concentrated around a few labels (SDI $M =$

0.88, SD= 0.09), but that there was still variability across objects (max SDI = 0.99 ; min SDI = 0.38). We found a strong negative correlation between an object's SDI and its recognizability (Pearson $r(1,853)$ = -0.80, $p < 0.001$) (Figure 5 C). This supports the notion that when observers converge on a single label for a given drawing it is often the correct label that was shown to the sketcher during the production task. However, it leaves open the question of why drawings with high label diversity exist.. For cases where observers are generating multiple labels for drawings of a given object and still getting them 'wrong', there are two broad possibilities — (1) the drawing lacked the visual information to allow an observer to correctly identify the intended object and thus they responded randomly, or (2) the drawing did convey enough information for the observer to recognize it, but they failed to use the appropriate label given the semantic diversity of 1,854 objects in THINGS, many of which are semantic 'neighbors' (e.g., kitten and cat). In order to better adjudicate between these two possibilities, in the following section, we looked beyond exact-match metrics to a semantic matching approach where we measured whether guessers produced labels that were in the correct 'semantic neighborhood' of the true labels.

Label guesses semantically approximate the true label

In order to establish a graded metric for how similar different objects were, we leveraged *semantic embeddings* for each of the 1,854 object categories collected by Hebart et al. (2023). These embeddings position each of the THINGS objects in a common 66-dimensional space, where the inter-item vector distance reflects the item-to-item similarity as assessed using 4.6 million human triplet odd-one-out judgments (Zheng, Pereira, Baker, & Hebart, 2019; Stoinski et al., 2023; Hebart et al., 2023). Since these embeddings were computed based on similarities perceived between photographs of the THINGS objects as opposed to drawings, it was first important to validate the generalizability of these embeddings for this particular stimuli class of drawings. In order to do so, we tested whether the similarity between the true labels for drawings of a given

object and observers' guesses for drawings of that object in embedding space was predictive of the likelihood of that object being labeled correctly.

For each drawing, we computed how similar a participant's guess was to the *true* label by computing the cosine similarity between semantic embeddings of the guess and true label. In cases where there were multiple guesses, we adopted two strategies. First, we computed the similarity between the true label and the mean embedding of the guessed labels. Overall, we found that an object's *mean guess similarity*, that is how similar guesses were to the true label, was predictive of its recognizability (Pearson $r(1,853)= 0.71$, $p<0.001$) (Figure 5 (B)). This approach would penalize a participant if one of the multiple guesses was semantically distal from the true label. So, as an alternative approach we computed the similarity between the embedding of the true label and the *closest guess*. We found that this *best guess similarity* was slightly more predictive of an object's recognizability (Pearson $r(1,853) = 0.74$, $p < 0.001$). Thus the similarity structure expressed by real-world photographs of visual objects generalizes and is able to capture how recognizable visually abstract drawings of the same objects are.

Having established that item-wise similarity between labels and guesses in semantic embedding space is predictive of whether drawings of objects are recognizable, we next sought to use these semantic embeddings to measure whether *incorrect* guesses were nevertheless semantically related to the true label. Concretely, we used the similarity of incorrect label guesses to the target label to derive such a measure of the object's *semantic neighbor preference* (Mukherjee et al., 2024). In order to compute semantic neighbor preference (SNP) for a given object's drawings, we first determine a rank ordering of the 1,853 *off*-target labels in terms of their similarity to the true target label (based on their semantic embeddings). The higher the rank of a given object label, the closer a neighbor it is to the true object label for that drawing. In order for a drawing to have high SNP, the rank of all its off-target labels should be reasonably high (that is, they should be close to 1). We next computed the cumulative proportion of off-target labels as a function of their

rank. A drawing that elicits labels neighboring the true label will have most of its assigned labels ranked highly (i.e., lower numerical ranks), causing the cumulative proportion to rise steeply toward 1. In contrast, if many assigned labels are semantically distant, their ranks will be lower (i.e., higher numerical values), leading to a more gradual increase. This rate of increase is captured by the area under the curve (AUC) of the cumulative proportion vs. label rank plot. This AUC value corresponds to the SNP metric. A SNP value closer to 1 indicates that guesses were in the same 'semantic neighborhood' of the true label and thus that drawing has a high semantic neighbor preference. In contrast, an SNP value close to 0.5 would indicate that the off-target labels were indeed random and that there is no semantically meaningful structure in the incorrect guesses. We computed SNP values for each object, by averaging the SNP values for each drawing belonging to the object. We generally found that most objects had a high semantic neighbor preference (M=0.78, SD= 0.12) relative to a uniform baseline (0.5), indicating that even when participants do not provide the exact true label, their guesses are systematic and are semantically related to the true label. Even after accounting for the fact that different labels may nevertheless 'point' to the same underlying object, our SNP scores continued to span a considerable range (max = 0.99, min = 0.36). In addition to computing SNP based on the first guess a participant made, we also computed a 'top-k' variant in a manner similar to what we previously did for computing recognizability. In cases where a participant was incorrect and provided multiple labels, we considered the label that was closest to the true label in semantic embedding space as the guess and computed the rank of that label. Considering the 'top-k' guesses led to an overall higher SNP (M=0.81; SD = 0.10; max = 0.99, min = 0.42).

Despite this range, the vast majority of objects (1,846 of 1,854; except 'bagpipe','bomb', 'boomerang', 'chest' (furniture), 'chest' (container), 'football', 'mouse' (device), and 'ring') showed an SNP value greater than 0.5, supporting the notion that generally objects' off-target labels were semantically meaningful. We also estimated the SNP using the photograph recognizability data and found that photographs yielded even

higher SNP scores ($M = 0.90$, $SD = 0.08$; max = 0.99, min = 0.44) (summarized in Figure 6 A). The SNP scores between drawings (top-k guesses) and photographs (top-k guesses) were also moderately correlated (Pearson $r(1,853) = 0.37$, $p < 0.001$) indicating that observers make similar patterns of off-target guesses when presented with visual object concepts in both drawing and photographic form. It is not surprising that drawings and images are not perfectly correlated in terms of SNP since drawings, by nature, are more abstract and might thus evoke a broader variety of visual concepts relative to less abstract images.

For subsequent analyses, we utilized SNP as a measure of observers' sensitivity to semantic information in drawings, as it reflects guesses within the semantic neighborhood of the correct label. This shift can be understood as a gradual expansion of what qualifies as an "acceptable" label. Moving from top-1 to top-k accuracy broadens the range of participant responses considered correct based on predefined criteria. In contrast, transitioning from accuracy to SNP represents a shift in the researcher-defined target itself, focusing on the semantic proximity of responses rather than strict correctness. The series of analyses in the present section raises the question as to whether cognitively circumscribed properties of these visual objects may determine the extent to which people are able to produce and consequently recognize drawings of them. To illustrate the potential of `DoT` as a standardized dataset that interfaces with existing large-scale resources germane to the cognitive sciences, in the following section we present an analysis of our data that utilizes data from existing THINGS datasets.

Typicality but not memorability of concepts yield more recognizable drawings.

One overarching goal of the original THINGS dataset was to provide a representative sample of objects, covering oft ignored object classes, to help explain the factors that influence recognition and recall of these objects in naturalistic behavior (Hebart et al., 2019). Here, we explore whether two candidate cognitive factors –memorability (Bainbridge, 2019) and typicality (Rosch & Mervis, 1975; Rosch, Simpson,

& Miller, 1976) – might help explain the pattern of results we find with respect to drawing and image recognition. Memorability is a temporally consistent image-specific property that is predictive of one's ability to remember that stimulus. Sensitivity to memorability develops early in childhood (Guo & Bainbridge, 2024), is not purely determined by low-level visual features of images (Bainbridge & Rissman, 2018), and is largely independent from other top-down cognitive processes such as cognitive control (Bainbridge, 2020). While work investigating the underlying features of memorability on naturalistic images is nascent (Kramer et al., 2023), here we consider that a visual concept might inherit the memorability of its exemplars, including drawings, leading to shared memorability between photographs and drawings of the same object (Han, Rezanejad, & Walther, 2023). Since it has been shown that semantic properties are the most predictive of concept-level memorability (Kramer et al., 2023), to the extent that recognizing drawings of objects recruit the same features used during recognition of real objects (Fan et al., 2018), we might expect memorability to be a predictor of drawings' recognizability. Additionally, the specific features that make a concept memorable might be the same features sketchers implicitly choose to include in their drawings of the concept.

To test this hypothesis, we correlated each object's semantic neighbor preference with memorability scores for each of the THINGS objects collected by Kramer et al. (2023). These scores were measured in a behavioral experiment utilizing a continuous recognition memory task, where participants viewed a stream of images from the THINGS database and had to indicate with a button press whenever they remembered a repeat image from earlier in the stream. This resulted in over 1-million human memory ratings. The memorability for each image was then computed as the difference between the hit rate (HR), the proportion of correct identifications of repeats, and false-alarm rate (FAR), the proportion of incorrect detections. This resulted in a corrected recognition score, which has been shown to be highly similar across participants–people tended to consistently remember and forget the same images (Kramer et al., 2023). We found no significant

correlation between memorability and semantic neighbor preference ($r(1,853)$=-0.028, p=0.21), which shows that, while memorability is critically important for organizing semantic knowledge, it does not account for people's ability (or inability) to recognize abstract sketches of concepts (Figure 6 B).

Typicality is a salient property of members of a category that influences how quickly and easily the members are named (Rosch & Mervis, 1975; Rosch, 1975), how early they are acquired in development (Mervis & Pani, 1980), and how susceptible they are to miscategorization (McCloskey & Glucksberg, 1978). While one can judge how typical a given image is of a category, it is also possible to judge how typical a given category is of a higher-order superordinate category. Concepts that are generally more typical might support easier retrieval and facilitate recognition. We hypothesized that concepts that were more typical might be more easily accessible and thus might be both easier to draw and easier to recognize in sketches. We leveraged a set of typicality ratings for the THINGS concepts developed by Stoinski et al. (2023), and correlated SNP scores with typicality scores. Briefly, the typicality scores in Stoinski et al. (2023) were collected by asking participants to rank the typicality of a random subset of THINGS concepts with respect to a common parent category. For example, participants were shown concept words such as "dog", "parrot", "zebra" etc. and asked to rank them in terms of how representative they were of the category 'animal'. The final typicality score for each concept reflected how often that concept was ranked highly in the subset in which it was presented. Concepts that were more typical of their superordinate categories had scores closer to 1 and atypical concepts were closer to 0. We found a moderately positive correlation between semantic neighbor preference and typicality ($r(1,853) = 0.35$, p<0.001) supporting the notion that drawings of concepts that were more typical tended to contain visual features that led to semantically meaningful guesses from observers (Figure 6 C).

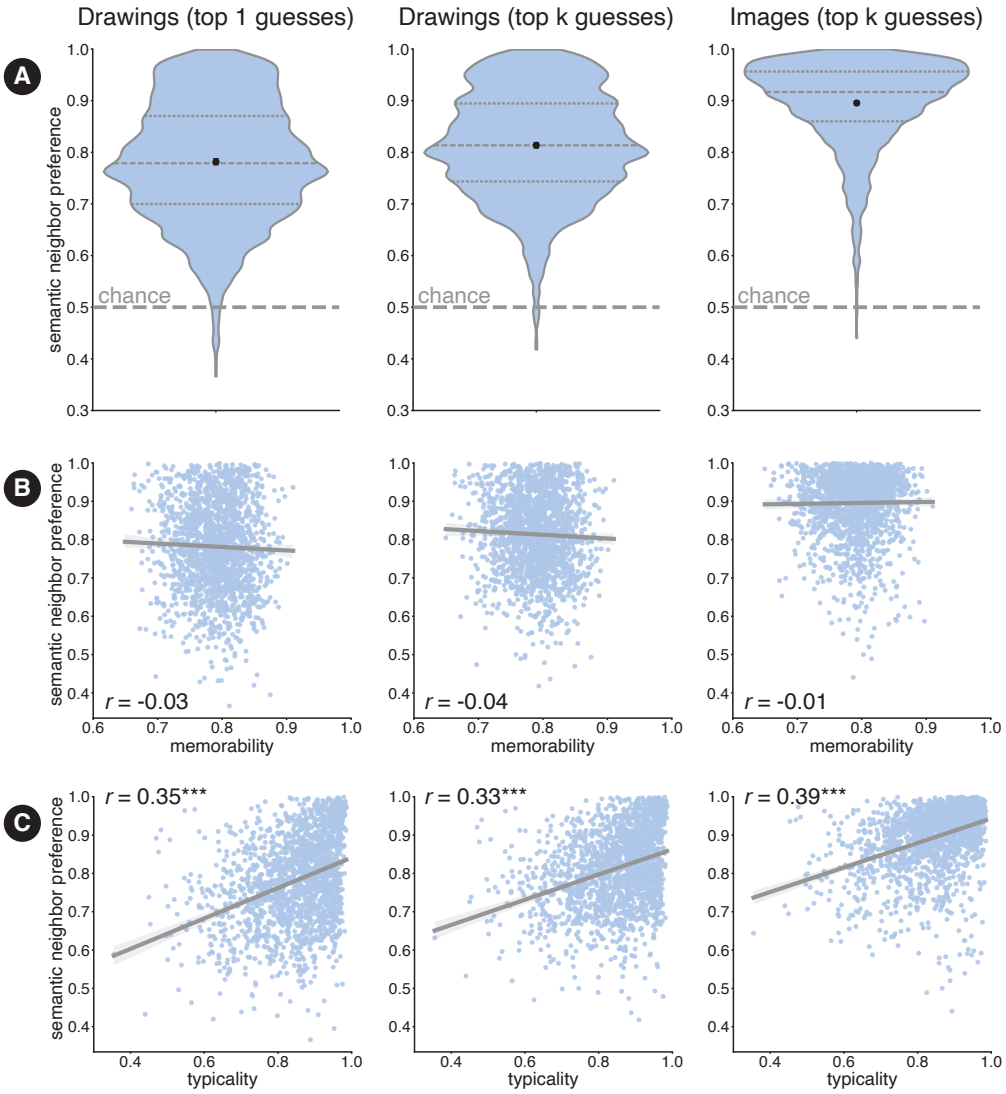

Figure 6

(A) Distribution densities of semantic neighbor preference scores across objects. Dashed lines within violin plots indicate medians and quartiles and the black dots indicate means. (B) Semantic neighbor preference as a function of object memorability. (C) Semantic neighbor preference as a function of object typicality. Asterisks indicate statistical significance: *** $p < 0.001$

General Discussion

Here, we introduced **DoT**, a large-scale dataset of over 28,000 drawings of 1,854 visual object concepts representative of the human visual experience inherited from the THINGS database (Hebart et al., 2019, 2023). We focused on collecting high-quality drawings that nevertheless reflected the strategies and skill level of the average person without artistic training. We also collected dense multilabel recognition data on each drawing, allowing for rich comparisons with existing metadata on the THINGS concepts. This dataset builds on the tradition of using drawings to probe visual semantic knowledge (Fan et al., 2018; Yang & Fan, 2021; Mukherjee et al., 2019; Sangkloy et al., 2016; Eitz et al., 2012; Bainbridge et al., 2019), while addressing a key gap in existing datasets - semantic diversity. To demonstrate the utility of our dataset, we characterize the range of behaviors covered by our production and recognition experiment and present several analyses relating recognition performance to other aspects of cognition.

First, we note substantial variability in the recognizability of concepts when rendered as drawings, with some concepts being consistently identified correctly while others were never labeled accurately. Highlighting the interoperability of our dataset with sister datasets, we use recognition data on photographs of the THINGS concepts to show that there exist cases where recognizability of drawings of concepts diverges from that of photographs of the same concepts. Interestingly, there were some concepts (11.38% of the database) where drawings were better recognized than their photographic counterparts, highlighting the potential of abstract visual representations to capture and communicate the diagnostic and essential features of visual object concepts that may not be as salient in real-world photographs. In contrast, cases where drawings fail to communicate the target concept might help delineate the limitations of hand-drawn renderings for communication in the absence of additional shared context. However, more generally, concepts that were difficult to recognize as drawings tended to also be challenging to identify in photographs, suggesting alignment in the features that make objects recognizable across different visual

representations (Singer et al., 2023).

Our analysis of label diversity and semantic similarity provides evidence that even when participants do not provide the exact correct label for a drawing, their guesses tend to be semantically related to the true concept. This is reflected in the high semantic neighbor preference (SNP) scores observed for most concepts. Thus, even in cases where drawings fail to perfectly capture the target label, observers assign labels in a manner that is systematic and in the correct semantic neighborhood. While concept memorability did not significantly predict recognition performance, a concept's typicality with respect to superordinate categories showed a moderate positive correlation with semantic neighbor preference. This indicates that more typical exemplars of a category may be easier to both draw and recognize, possibly due to their more accessible and well-defined features in semantic memory. Drawings allow for capturing high-fidelity information about people's visual semantic knowledge that is impossible to capture using traditional measures. The use of drawings as the medium of production and as stimuli for a recognition study helped jointly show that people's ability to visually communicate is highly varied and that this variation can potentially be explained as a function of concept-level properties.

While we present our analyses as a starting point, the **DoT** dataset opens up several avenues for future research including further fine-grained investigations into human semantic memory, the faculties that support visual abstraction, and the computational principles that support sketch recognition and generation, potentially leading to better models of how humans represent and communicate visual concepts. In conclusion, **drawings of THINGS** provides a rich resource for investigating the structure and organization of visual semantic knowledge. By combining a diverse set of object concepts with both production and recognition data, this dataset enables researchers to ask nuanced questions about how humans represent, communicate, and interpret visual information.

Declarations

Dataset & Code Availability

The experiments reported in this manuscript were not preregistered.

Data for the experiments reported can be accessed at the folliwng OSF repository:

https://osf.io/s9yta/

All pre-processing and analyses reported will be available at:

github.com/cogtoolslab/drawings-of-THINGS-public

Acknowledgements & Funding

We would like to thank Sam Rosenthal, Esther Goldberg, Rio Aguina-Kang, Ivette Colon, Claire Peplinksi, and Katya Müller for their help in testing the experimental interfaces and helping with data preprocessing. K.M. was supported by the McPherson Eye Research Institute at the University of Wisconsin-Madison. W.A.B. was supported by the National Eye Institute (R01-EY034432). J.E.F. was supported by NSF CAREER #2436199, NSF DRL #2400471, the Stanford Human-Centered Institute for Artificial Intelligence (HAI), and the Stanford Accelerator for Learning. M.N.H. was supported by a Max Planck Research Group grant of the Max Planck Society, the ERC Starting Grant COREDIM (ERC-StG-2021-101039712) ,and the Hessian Ministry of Higher Education, Science, Research and Art (LOEWE Start Professorship and Excellence Program "The Adaptive Mind").

Competing interests/Conflicts of Interest

N.A.

Ethics approval

All studies were conducted in accordance with the respective institution IRBs. Refer to the methods sections for each study for study-specific IRB details.

Consent for publication & Consent for publication

All participants provided informed consent in accordance with institutional IRBs. All participant data are anonymized and thus no personally identifiable information is included in this publication. Refer to the methods sections for each study for study-specific IRB details.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... others (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological science*, *15*(2), 106–111.

Bainbridge, W. A. (2019). Memorability: How what we see influences what we remember. In *Psychology of learning and motivation* (Vol. 70, pp. 1–27). Elsevier.

Bainbridge, W. A. (2020). The resiliency of image memorability: A predictor of memory separate from attention and priming. *Neuropsychologia*, *141*, 107408.

Bainbridge, W. A. (2022). A tutorial on capturing mental representations through drawing and crowd-sourced scoring. *Behavior Research Methods*, *54*(2), 663–675.

Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature communications*, *10*(1), 5.

Bainbridge, W. A., Pounder, Z., Eardley, A. F., & Baker, C. I. (2021). Quantifying aphantasia through drawing: Those without visual imagery show deficits in object but not spatial memory. *Cortex*, *135*, 159–172.

Bainbridge, W. A., & Rissman, J. (2018). Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval. *Scientific Reports*, *8*(1), 8679.

Benchetrit, Y., Banville, H., & King, J.-R. (2023). Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812*.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, *20*(1), 38–64.

Boutin, V., Fel, T., Singhal, L., Mukherji, R., Nagaraj, A., Colin, J., & Serre, T. (2023).

Diffusion models as artists: are we closing the gap between humans and machines? *arXiv preprint arXiv:2301.11722*.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... others (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 1–74.

Bozeat, S., Lambon Ralph, M. A., Graham, K. S., Patterson, K., Wilkin, H., Rowland, J., ... Hodges, J. R. (2003). A duck with four legs: Investigating the structure of conceptual knowledge using picture drawing in semantic dementia. *Cognitive neuropsychology*, *20*(1), 27–47.

Caplette, L., & Turk-Browne, N. B. (2024). Computational reconstruction of mental representations using human behavior. *Nature Communications*, *15*(1), 4183.

Chamberlain, R., Kozbelt, A., Drake, J. E., & Wagemans, J. (2021). Learning to see by learning to draw: A longitudinal analysis of the relationship between representational drawing training and visuospatial skill. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(1), 76.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, *111*(40), 14565–14570.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, *15*(1), 9383.

Cunningham, C. A., & Wolfe, J. M. (2014). The role of object categories in hybrid visual and memory search. *Journal of Experimental Psychology: General*, *143*(4), 1585.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., &

Storms, G. (2008). Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. *Behavior research methods*, *40*, 1030–1048.

De Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in cognitive sciences*, *22*(9), 764–779.

De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*, 1–12.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, *46*, 1119–1127.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Dillon, M. R. (2021). Rooms without walls: Young children draw objects but not layouts. *Journal of Experimental Psychology: General*, *150*(6), 1071.

Dobs, K., Martinez, J., Kell, A. J., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science advances*, *8*(11), eabl8913.

Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects? *ACM Transactions on graphics (TOG)*, *31*(4), 1–10.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, *88*, 303–338.

Fan, J. E., Bainbridge, W. A., Chamberlain, R., & Wammes, J. D. (2023). Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, *2*(9), 556–568.

Fan, J. E., Hawkins, R. D., Wu, M., & Goodman, N. D. (2020). Pragmatic inference and

visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior*, *3*, 86–101.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive science*, *42*(8), 2670–2698.

Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., & Isola, P. (2023). Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*.

Gauthier, I., & Tarr, M. J. (2016). Visual object recognition: Do we (finally) know more now than we did? *Annual review of vision science*, *2*(1), 377–396.

Gifford, A. T., Dwivedi, K., Roig, G., & Cichy, R. M. (2022). A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, *264*, 119754.

Grootswagers, T., Zhou, I., Robinson, A. K., Hebart, M. N., & Carlson, T. A. (2022). Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, *9*(1), 3.

Guo, X. H., & Bainbridge, W. A. (2024). Children develop adult-like visual sensitivity to image memorability by the age of 4. *Journal of Experimental Psychology: General*, *153*(2), 531.

Ha, D., & Eck, D. (2017). A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*.

Han, S., Rezanejad, M., & Walther, D. B. (2023). Memorability of line drawings of scenes: the role of contour properties. *Memory & Cognition*, 1–21.

Hansen, H., & Hebart, M. N. (2022). Semantic features of object concepts generated with gpt-3. *arXiv preprint arXiv:2202.03753*.

Hawkins, R. D., Sano, M., Goodman, N. D., & Fan, J. E. (2023). Visual resemblance and interaction history jointly constrain pictorial meaning. *Nature Communications*, *14*(1), 2199.

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., ...

Baker, C. I. (2023). Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, *12*, e82580.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, *14*(10), e0223792.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173–1185.

Huey, H., Lu, X., Walker, C. M., & Fan, J. E. (2023). Visual explanations prioritize functional properties at the expense of visual fidelity. *Cognition*, *236*, 105414.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224.

Jongejan, J., Rowley, H., Kawashima, T., Kim, J., & Fox-Gieg, N. (2017). *The quick, draw! dataset.*

Koblin, A. M. (2009). The sheep market. In *Proceedings of the seventh acm conference on creativity and cognition* (pp. 451–452).

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1), 1–12.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological science*, *21*(11), 1551–1556.

Kramer, M. A., Hebart, M. N., Baker, C. I., & Bainbridge, W. A. (2023). The features underlying the memorability of objects. *Science advances*, *9*(17), eadd2981.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, *12*(4), e1004896.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13* (pp. 740–755).

Long, B., Fan, J., Chai, Z., & Frank, M. C. (2021). Parallel developmental changes in children's drawing and recognition of visual concepts.

Long, B., Fan, J. E., Huey, H., Chai, Z., & Frank, M. C. (2024). Parallel developmental changes in children's production and recognition of line drawings of visual concepts. *Nature Communications*, *15*(1), 1191.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547–559.

Megla, E., Rosenthal, S. R., & Bainbridge, W. A. (2024). Drawings reveal changes in object memory, but not spatial memory, across time. *bioRxiv*.

Megla, E., Rosenthal, S. R., & Bainbridge, W. A. (2025). Drawings reveal changes in object memory, but not spatial memory, across time. *Cognition*, *254*, 105988.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118.

Mervis, C. B., & Pani, J. R. (1980). Acquisition of basic object categories. *Cognitive psychology*, *12*(4), 496–522.

Mukherjee, K., Hawkins, R. X., & Fan, J. W. (2019). Communicating semantic part information in drawings. In *Cogsci* (pp. 2413–2419).

Mukherjee, K., Huey, H., Lu, X., Vinker, Y., Aguina-Kang, R., Shamir, A., & Fan, J. (2024). Seva: Leveraging sketches to evaluate alignment between human and machine visual abstraction. *Advances in Neural Information Processing Systems*, *36*.

Mukherjee, K., & Rogers, T. T. (2024). Using drawings and deep neural networks to characterize the building blocks of human visual similarity. *Memory & Cognition*, 1–23.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-it object representation. *Frontiers in psychology*, *4*, 128.

Muttenthaler, L., Zheng, C. Y., McClure, P., Vandermeulen, R. A., Hebart, M. N., & Pereira, F. (2022). Vice: Variational interpretable concept embeddings. *Advances in Neural Information Processing Systems*, *35*, 33661–33675.

Papale, P., Wang, F., Self, M. W., & Roelfsema, P. R. (2025). An extensive dataset of spiking activity to reveal the syntax of the ventral stream. *Neuron*.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature reviews neuroscience*, *8*(12), 976–987.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Rogers, T., Hodges, J., Patterson, K., & Lambon Ralph, M. (2003). Object recognition under semantic impairment: The effects of conceptual regularities on perceptual decisions. *Language and Cognitive Processes*, *18*(5-6), 625–662.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological review*, *111*(1), 205.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, *104*(3), 192.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573–605.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human perception and performance*, *2*(4), 491.

Sangkloy, P., Burnell, N., Ham, C., & Hays, J. (2016). The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, *35*(4), 1–12.

Sayim, B., & Cavanagh, P. (2011). What line drawings reveal about the visual brain. *Frontiers in human neuroscience*, *5*, 118.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., ... others (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, *35*, 25278–25294.

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2556–2565).

Simpson, E. H. (1949). Measurement of diversity. *nature*, *163*(4148), 688–688.

Singer, J. J., Cichy, R. M., & Hebart, M. N. (2023). The spatiotemporal neural dynamics of object recognition for natural images and line drawings. *Journal of Neuroscience*, *43*(3), 484–500.

Singer, J. J., Seeliger, K., Kietzmann, T. C., & Hebart, M. N. (2022). From photos to sketches-how humans and deep neural networks process objects across different levels of visual abstraction. *Journal of vision*, *22*(2), 4–4.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of*

experimental psychology: Human learning and memory, *6*(2), 174.

Stoinski, L. M., Perkuhn, J., & Hebart, M. N. (2023). Thingsplus: New norms and metadata for the things database of 1854 object concepts and 26,107 natural object images. *Behavior Research Methods*, 1–21.

Thompson-Schill, S., Aguirre, G., Desposito, M., & Farah, M. (1999). A neural basis for category and modality specificity of semantic knowledge. *Neuropsychologia*, *37*(6), 671–676.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proceedings of the National Academy of Sciences*, *94*(26), 14792–14797.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Vanasse, T. J., Boly, M., Allen, E. J., Wu, Y., Naselaris, T., Kay, K., ... Tononi, G. (2022). Multiple traces and altered signal-to-noise in systems consolidation: Evidence from the 7t fmri natural scenes dataset. *Proceedings of the National Academy of Sciences*, *119*(44), e2123426119.

Wang, A., Ren, M., & Zemel, R. (2021). Sketchembednet: Learning novel concepts by imitating drawings. In *International conference on machine learning* (pp. 10870–10881).

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*(3), 829–853.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yang, J., & Fan, J. E. (2021). Visual communication of object concepts at different levels of abstraction. *arXiv preprint arXiv:2106.02775*.

Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). Sigmoid loss for language image pre-training. In *Proceedings of the ieee/cvf international conference on computer*

vision (pp. 11975–11986).

Zheng, C. Y., Pereira, F., Baker, C. I., & Hebart, M. N. (2019). Revealing interpretable object representations from human behavior. *arXiv preprint arXiv:1901.02915*.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118.