# Evaluating machine comprehension of sketch meaning at different levels of abstraction

**Kushin Mukherjee[1], Holly Huey[2], Xuanchen Lu[2], Yael Vinker[3], Rio Aguina-Kang[2], Ariel Shamir[5], and Judith E. Fan[2,4]**

[1]University of Wisconsin-Madison, Madison, WI, United States
[2]University of California, San Diego, CA, United States
[3]Tel-Aviv University, Tel-Aviv, Israel
[4]Stanford University, Stanford, CA, United States
[5]Reichman University, Herzliya, Israel

## Abstract

People can reliably understand images that vary in visual abstraction—from detailed illustrations to schematic icons. To what degree are current vision algorithms robust to such variation when attributing meaning to abstract images? We first obtained $> 90K$ human-generated sketches produced under different time limits (4s, 8s, 16s, 32s; $N$=5,563 participants) and AI-generated sketches (Vinker et al., 2022) produced under different ink limits (4, 8, 16, 32 strokes) of 2,048 real-world object concepts spanning 128 categories from the THINGS dataset (Hebart et al., 2019). We then evaluated how well 12 state-of-the-art vision algorithms could (1) predict which concept each sketch was intended to convey and (2) match human performance and response patterns when presented with the same sketches. We found that models achieving generally higher recognition accuracy also tracked human error patterns better, although there remains a sizable gap between human and machine sketch understanding. We also found that, on average, different models expressed similar uncertainty about sketches of the same concept across different levels of abstraction. We hope that public release of this dataset and evaluation protocol will lead to algorithms that display more human-like visual abstraction.

**Keywords:** concepts; drawing; perception; computer vision; benchmark

## Introduction

Humans can use pictures to convey what they perceive and know at varying levels of abstraction—from detailed illustrations to simple sketches. Indeed, the ability to abstract away from the particulars of any given experience to highlight the most important elements is inherent in the act of creating any effective visualization (Viola & Isenberg, 2017; M. Chen, Hauser, Rheingans, & Scheuermann, 2020; McCloud & Manning, 1998; Mi, DeCarlo, & Stone, 2009; Nan et al., 2011). Line drawings present an especially important case study in the capacity for visual abstraction, as demonstrated by the Spanish artist Pablo Picasso in his renown work, *The Bull* (1945), which contains 11 lithographs of bulls, each successively more abstract than the last (Fig. 1). Despite striking variation in their degree of fidelity to the real world, understanding what even the most abstract of these images represent feels effortless for most human viewers.

Such variation is manifest in works of art, but is also pervasive across many domains of human activity. Not only do most cultures produce drawings (Gombrich, 1995), the ability to produce line drawings that capture key aspects of the real world also emerges early in development (Karmiloff-Smith, 1990; Dillon, 2021; Long, Fan, Chai, & Frank, 2021), and the visual properties of these drawings have been linked to children's developing conceptual knowledge (Tversky, 1989; Huey & Long, 2022). Additionally, failures to produce and understand pictures of objects at different levels of abstraction is associated with semantic dementia (Bozeat et al., 2003; Rogers & Patterson, 2007), suggesting links between a robust capacity for visual abstraction and the functional organization of semantic knowledge in the brain. What are the core visual computations that support this ability to grasp the meaning of pictures across so many different levels of visual abstraction?

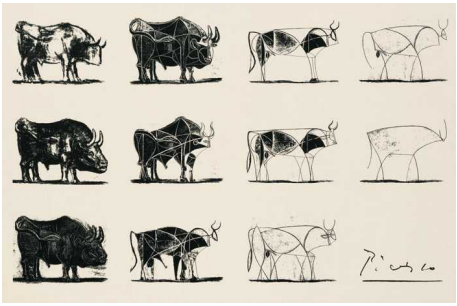

Figure 1: Pablo Picasso. *The Bull*, 1945.

The past several years have seen remarkable progress in uncovering the mechanisms by which the human visual system achieves a robust understanding of the visual world (Yamins et al., 2014; Kriegeskorte, 2015; Zhuang et al., 2021; Konkle & Alvarez, 2022). These mechanistic models now often take the form of trainable neural networks combining several architectural motifs inspired by the primate ventral visual stream (Gross, Rocha-Miranda, & Bender, 1972; Goodale & Milner, 1992; Malach, Levy, & Hasson, 2002; Hung, Kreiman, Poggio, & DiCarlo, 2005). These advances have also recently been applied to the problem of sketch understanding, revealing both the value of these approaches for learning general-purpose perceptual representations to model human visual abstraction (Fan, Yamins, & Turk-Browne, 2018; Yu et al., 2017; Kubilius, Bracci, & Op de Beeck, 2016), as well as persistent challenges in achieving the capacity for robust understanding of visual inputs that vary in their degree of visual abstraction (Baker & Kellman, 2018; Singer, Seeliger, Kietzmann, & Hebart, 2022; Fan, Hawkins, Wu, & Goodman, 2020).

Despite these great strides in the development of high-

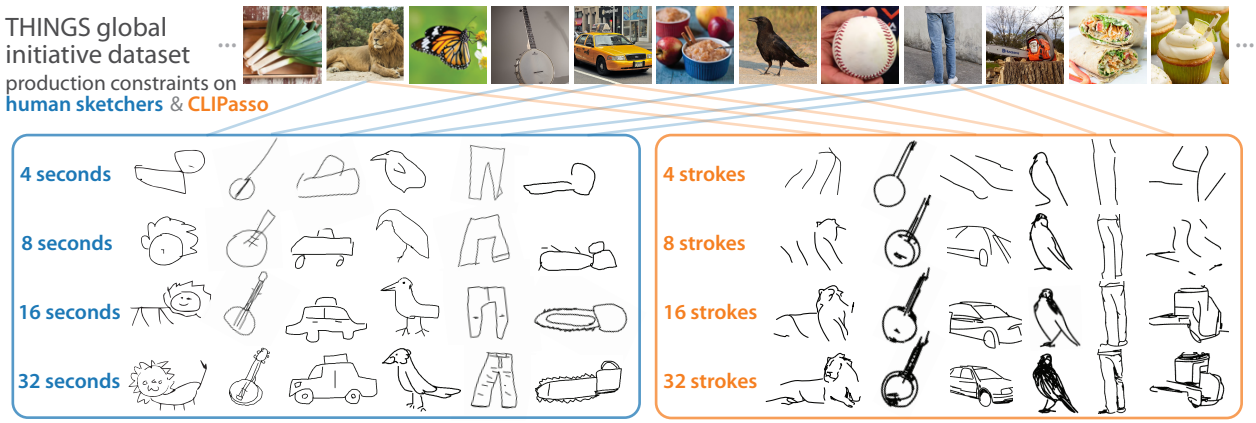


Figure 2: Human sketchers & CLIPasso generated $> 90K$ sketches under different production constraints: drawing time & number of strokes, respectively.

performing vision models, it remains unclear to what degree the *specific* models that have been proposed so far achieve human-like understanding of such a broad range of visual inputs, from natural images to human-generated drawings and symbols. Evaluating this question has been especially challenging given that, while there are several widely used benchmark sketch datasets (Eitz, Richter, Boubekeur, Hildebrand, & Alexa, 2012; Jongejan, Rowley, Kawashima, Kim, & Fox-Gieg, 2017; Sangkloy, Burnell, Ham, & Hays, 2016), none of them systematically vary the degree of detail, a salient axis differentiating depictions of specific instances from more abstract illustrations.

In this paper, we take two major steps towards closing this gap: (1) we develop a large dataset containing human ($N$=5,563 participants) and AI-generated sketches varying in detail, for a representatively wide variety of visual object concepts varying in their degree of abstraction; and (2) we systematically evaluate how well 12 diverse state-of-the-art vision models, varying in their architectures and training methods, represent semantic information in these sketches by benchmarking their recognition performance against human behavior. We build on a growing body of research leveraging a global image dataset generated by the THINGS initiative (Hebart et al., 2019) by sampling 2,048 real-world objects spanning 128 concepts as referents for drawings in our dataset. Our main goals were to test the consistency between models and humans in their ability to recognize the concepts depicted in our sketch dataset, as well as the alignment between distributions of human-generated soft labels ($N$=3,190 participants) and distributions underlying model classification performance (Collins, Bhatt, & Weller, 2022; Peterson, Battleday, Griffiths, & Russakovsky, 2019). Taken together, our work aims to contribute an informative benchmark of human and machine generated sketches spanning varying multiple levels of abstraction. We hope that publicly releasing our datasets and proposed methods for investigating sketch understanding will generate opportunities for future research avenues towards building better computational models of human visual abstraction.

## Method

Our first goal was to generate two parallel large-scale drawing datasets spanning varying levels of abstraction: one produced by humans under varying time limits (4s, 8s, 16s, and 32s), and the other by automatic machine generation varying in the number of strokes per drawing (4, 8, 16, and 32 strokes), using CLIPasso (Vinker et al., 2022) (see Fig. 2 for example sketches). Next, to estimate recognizability of the drawings in both datasets, we conducted an independent recognition study in which participants provided one or more labels for a representative sample of human and machine drawings.

**Stimuli** To generate a diverse large-scale stimulus set of object concepts, we systematically sampled 128 concepts from the database of the THINGS initiative, a global database of 1,854 object concepts (e.g., "lion", "banjo", "car") and naturalistic object images aimed at developing a multi-varied cognitive neuroscience and behavioral metrics on a shared set of objects (Hebart et al., 2019). Building on prior work by Yang and Fan (2021) investigating visual abstraction across different contexts, we selected concepts based on similar parameters spanning four main axes of variation: familiarity, artificiality, animacy, and size. Within each object concept, we randomly sampled 16 object images. Our final stimuli set included 2,048 object images that were used as referents for the human and machine sketching tasks.

### Human Sketch Production Task

**Participants** 5,563 participants (2,870 male; $M_{age}$ = 36.7 years) were recruited from Prolific to produce a series of sketches on a web-based drawing platform. We excluded 104 data sessions from participants, who experienced technical difficulties. In this and all subsequent tasks, participants provided informed consent in accordance with the UC San Diego IRB.

**Procedure** We randomly assigned participants to one of four conditions, each varying in the amount of time that they were permitted to use to generate their drawings: 4, 8, 16, or 32 seconds (Fig. 2, *left*). Each participant produced

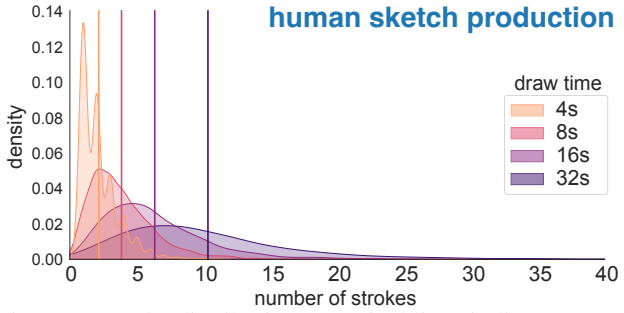

Figure 3: Stroke distributions. Vertical lines indicate means.

16 drawings of different object concepts. During each trial, participants were presented with an object label, a corresponding object photograph (500px x 500px) to give participants a concrete example of what the object looks like, and a drawing canvas (500px x 500px). They were instructed to produce a drawing of the general concept represented by the label, and that, they should not include details specific to the instance of the object in the provided photograph. Next to the canvas, participants were given a countdown timer indicating how many seconds they had left to produce their drawing. A trial ended when the timer ran out or when the participant pressed the "Continue" button if they finished their sketch with time remaining, although they were instructed to try to use as much time as they needed to accurately represent the prompted concept. Participants were were instructed not to include any background context (e.g., grass in a drawing of a "horse"), arrows, or text. Participants also completed a practice trial before the test drawing trials to familiarize themselves to the drawing platform, including the ability to undo their most recent stroke or completely clear their canvas if needed.

## Machine Sketch Generation

We leveraged CLIPasso (Vinker et al., 2022), a recently developed sketch generation algorithm, to also generate sketches of the same 2,048 object images. CLIPasso generates sketches by optimizing the parameters of a set of curves (strokes) to align to a latent representation of an image of an object computed by CLIP (Radford et al., 2021), a neural network model trained on a large corpus of text-image pairs. For each image, we generated sketches at four levels of abstraction using 4, 8, 16, and 32 strokes (Fig. 2, *right*) as an approximately parallel manipulation of the time-restriction paradigm of our human production task.

## Human Sketch Recognition Task

We next aimed to develop a recognizability baseline for each sketch to compare to model classification performance. To accomplish this, we designed a web-based recognition study to crowdsource object labels associated with each sketch in order to capture its perceived meaning. Critically, because sketches often capture a range of semantic meaning that can bring to mind many possible concepts during viewing (Fig. 5, *left*), we allowed participants to submit up to 5 object labels per sketch in order to account for this polysemous property of sketches. Here, we sampled a subset of 8,192 sketches from both our human sketch and CLIPasso sketch datasets. Across the 128 object classes, each had 64 different sketches per dataset with 16 sketches at each level of abstraction.

**Participants** 1,709 Prolific participants (776 male; $M_{age}$ = 39.2 years) were recruited to make judgments about the human sketch dataset and 1,481 Prolific participants (730 male; $M_{age}$ = 41.05 years) were recruited to make judgments about the CLIPasso sketch dataset. We excluded 28 data sessions from participants who experienced technical difficulties. Data collection stopped when each sketch from both datasets had 12 judgments.

**Procedure** Each participant was presented with a randomized series of 64 sketches (300px x 300px) from either the human dataset or CLIPasso dataset. On every trial, they were presented with a single drawing and a text box and asked to provide a label that best represented the drawing by typing their response into the box. Upon beginning to type, participants were presented with a drop-down menu that displayed a subset of the 1,854 THINGS object concepts that had string matches to the participant's typed response. Each label also had words within parentheses to eliminate ambiguity (e.g. *mouse (animal)* was a different label than *mouse (computer)*). The participant could choose from one of these options to label the sketch. Participants could also add additional text boxes to submit multiple labels if they believed a sketch was representative of multiple object concepts, but were not permitted to submit custom label options not included in the 1,854 labels. Prior to test trials, participants completed a practice trial to familiarize themselves with the labeling interface. To assess whether participants were fully engaged with the task, they also completed an attention-check trial in the middle of the experiment that was of the same practice drawing.

## Formalizing abstraction in vision models

Which state-of-the-art vision models display human-like understanding of visual concepts at varying levels of abstraction as seen in sketches? To make progress towards answering this question, we first curated a set of vision models with varied architectural commitments and training procedures. Next, we designed an evaluation protocol such that models' performance could be directly compared to human behavior.

**Models** We evaluated 12 vision models spanning multiple architectures and training paradigms (see Table 1), all of which have demonstrated high performance for object recognition on standard datasets like ImageNet (Deng et al., 2009). We performed all downstream computations on latent features extracted from the deepest (non-fully-connected) layers of these models, which amounted to extracting activation patterns at either the model's final convolution or attention layer.

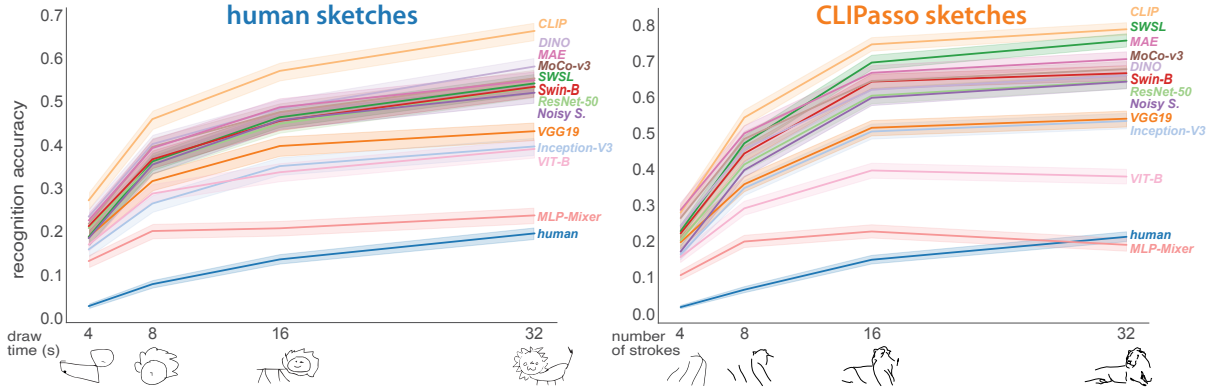


Figure 4: Top-1 recognition accuracy by models of human sketches (*left*) and CLIPasso sketches (*right*). Note that human accuracy reflects performance on a 1854-way classification task, rather than on the 128-way task performed by models.

**Evaluation Protocol** Due to the varied dimensionality and characteristics of the latent model features (outputs of convolution vs. attention layers vs. linear layers), to fairly assess the semantic information decodable from these features, we fit a series of regularized logistic classifiers predicting the class labels of each sketch from the latent features. Classifiers were fit separately to each neural network using 5–fold stratified cross-validation to predict the class label of the sketch. For each sketch, when presented in the test fold, we preserved the full vector of class probabilties corresponding to the 1,854 THINGS object classes.[1] Next, to compare these model data against human recognition, we leveraged the label data generated from our recognition study to compute the number of times each of the 1,854 labels was assigned to a given sketch. By summing these label counts across participants and normalizing them, we generated a human recognition "response" vector for each sketch, representing a distribution over the 1,854 labels as a measure of its perceived polysemy. Importantly, this provided an analogous baseline of comparison to the class probability vectors extracted from the classifiers.

| Models | Architecture | Training Paradigm |
|---|---|---|
| VGG-19 (Simonyan & Zisserman, 2014) | VGG-19 | supervised |
| Inception-V3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016) | Inception-V3 | supervised |
| ResNet-50 (He, Zhang, Ren, & Sun, 2016) | ResNet-50 | supervised |
| ViT-B (Dosovitskiy et al., 2020) | ViT-B | supervised |
| Swin-B (Liu et al., 2021) | Swin-B | supervised |
| MLPMixer-B (Tolstikhin et al., 2021) | MLPMixer-B | supervised |
| MoCo-v3 (X. Chen, Xie, & He, 2021) | ViT-B | self-supervised |
| DINO (Caron et al., 2021) | ViT-B | self-supervised |
| MAE (He et al., 2022) | ViT-B | self-supervised |
| CLIP (Radford et al., 2021) | ViT-B | self-supervised |
| Noisy Student (Xie, Luong, Hovy, & Le, 2020) | EfficientNet-b4 | semi-supervised |
| SWSL (Yalniz, Jégou, Chen, Paluri, & Mahajan, 2019) | ResNet-50 | semi-supervised |

Table 1: Evaluated models and their network architecture backbone and training paradigm.

## Results

Our final drawing dataset contained over 90,000 unique drawings (90,922 human-produced sketches; 8,191 CLIPasso-generated sketches[2]).

[1]Human and machine sketches were generated from only 128 classes, thus many class probability values of the full 1,854 THINGS label set were 0.

[2]We removed one blank sketch from the CLIPasso dataset.

**Shorter drawing times yield sketches with fewer strokes.** To validate our manipulation of drawing duration during human sketch production, we counted the number of unique strokes in a drawing as a measure of detail (Fig. 3) and then fit a mixed-effects linear regression model with random intercepts for category predicting number of strokes as a function of time allotted per drawing. We found that drawings produced under the 4s limit contained the fewest strokes, whereas those produced under the 32s limit contained the most strokes ($\beta = .29$, $SE = 4.95 \times 10^{-3}$, $p < .001$). These results help confirm that the time-restriction manipulation elicited meaningful differences in detail across drawing conditions.

**Production constraints on sketch production impact sketch recognition.** How well do people and vision models recognize the object concepts represented in sketches? And to what extent is the ability to reliably decode the semantic content in a given sketch impacted by the level of abstraction as measured by stroke complexity or time taken to draw it? To evaluate these questions, we fit mixed-effect linear regression models predicting the mean top-1 recognition accuracy in humans and vision models as a function of draw duration for human sketches and number of strokes for CLIPasso sketches. We included model-type as an additional fixed effect to test for performance differences between the vision models. Finally, we included by-concept intercepts and random slopes for draw duration or number of strokes. We found a significant effect of draw duration and number of strokes for both human ($\beta_{draw\,duration} = 5.66 \times 10^{-3}$, $SE = 5.92 \times 10^{-4}$, $p < .001$; $\beta_{strokes} = 6.74 \times 10^{-3}$, $SE = 7.17 \times 10^{-3}$, $p < .001$) and model ($\beta_{draw\,duration} = 9.16 \times 10^{-3}$, $SE = 4.21 \times 10^{-4}$, $p < .001$; $\beta_{strokes} = 1.21 \times 10^{-2}$, $SE = 3.83 \times 10^{-4}$, $p < .001$) recognition performance, indicating that greater amounts of detail in sketches correspond to higher recognition performance by both humans and vision models (Fig. 4). Additionally, we found an effect of model-type when coding for CLIP as the reference category indicating that not all models are equally performant when discerning the semantic structure in both human ($\chi^2(11) = 2712.50$, $p < .001$) and CLIPasso ($\chi^2(11) = 5759.4$, $p < .001$) sketches.

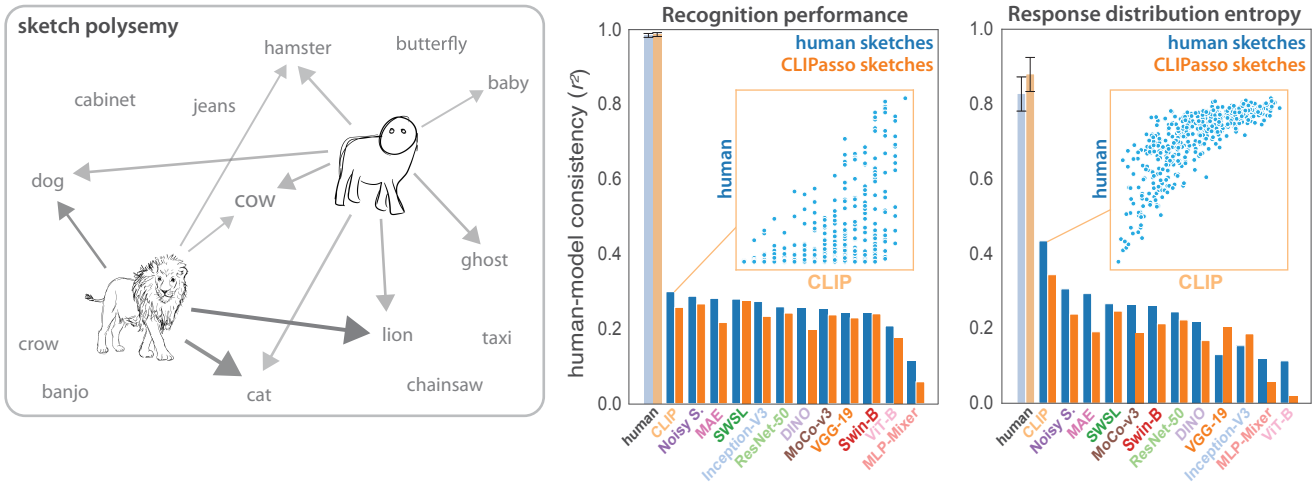


Figure 5: *Left:* Example of sketch polysemy at different levels of abstraction. *Middle:* Proportion of variance explained in top-1 human recognition accuracy by model top-1 accuracy. *Right:* Proportion of variance explained in human response distribution entropy by model class probability entropy. Inset scatterplots show the distribution of entropy values between human judgements and CLIP.

**Vision models vary in their degree of consistency with human recognition performance and response patterns.** Next, we conducted finer-grain analyses to identify which vision models are most consistent with human recognition performance by: (1) comparing the degree of alignment between model and human recognition *accuracy*; and (2) comparing the degree of alignment between model and human recognition *uncertainty*. First, we fit mixed-effect linear regression models predicting mean human top-1 accuracy from vision model top-1 accuracy and model-type with by-concept random slopes and intercepts to account for variability within concepts. We found that model performance was a significant predictor of human recognition performance ( $\beta_{human\ sketches} = 2.96 \times 10^{-1}$ $SE = 2.36 \times 10^{-2}$, $p < .001$; $\beta_{machine\ sketches} = 2.97 \times 10^{-1}$, $SE = 2.34 \times 10^{-2}$, $p < .001$). Furthermore, model comparisons revealed a significant effect of model-type for human ($\chi(11) = 552.92$, $p < .001$) and machine sketches ($\chi(11) = 741.51$, $p < .001$), adding additional support that some vision models are more consistent with human recognition performance than others (Fig. 5, *middle*). Specifically, these analyses revealed that CLIP explains the greatest amount of variance in human recognition accuracy, while MLP-Mixer explains little variance. While there is some discrepancy between which models are more accurate when comparing performance on human sketches vs. CLIPasso sketches, we generally saw that models that have better recognition performance across both datasets (Fig. 4) are also more consistent with human recognition performance.

Second, we computed the entropy of the distributions over object labels for each of the 8,192 sketches across both human sketch and CLIPasso sketch datasets, using the normalized label counts and classifier class probabilities, respectively (Fig. 5, *right*). Next, we fit mixed-effect models predicting human response entropy from model class probability entropy and model-type with by-concept random slopes and intercepts for entropy. Similar to model comparisons to human recognition accuracy, we found that model entropy significantly predicted human response entropy ($\beta_{human\ sketches} = 2.04 \times 10^{-1}$, $SE = 6.57 \times 10^{-3}$, $p < .001$; $\beta_{machine\ sketches} = 2.09 \times 10^{-1}$, $SE = 7.90 \times 10^{-3}$, $p < .001$). Additionally, model comparisons revealed a significant effect of model-type for both human ($\chi^2(11) = 1241.30$, $p < .001$) and machine ($\chi^2(11) = 1509.00$, $p < .001$) sketches, although CLIP was found to be the highest performing model for both human and CLIPasso sketches when using this metric. These evaluations of accuracy and uncertainty reveal that both state-of-the-art model performance and the distributions underlying that performance is predictive of the same constructs for human recognition performance to varying degrees, with some models being more consistent with human behavior than others.

**Sketch recognition behavior is generally consistent across vision models.** To what degree are state-of-the-art vision models consistent in their sensitivity to the semantic information conveyed by sketches containing varying amounts of detail? To evaluate this question, we looked at how consistent different models were in their patterns of average uncertainty over the sketches belonging to each of the 128 object classes and whether this consistency changed as function of stroke count or draw duration. For each model, we first computed 128-dimensional *entropy vectors* to capture the uncertainty expressed by the model for each of the object classes. Each resulting entry in this vector was the average response distribution entropy across sketches for one of the classes. If two models shared similar entropy vectors, it would indicate similar patterns of uncertainty across sketches of each class. To measure how consistent models are across each production constraint

level, we computed separate entropy vectors for each model at each production constraint level for both sketch datasets (number of strokes for CLIPasso sketches and draw duration for human sketches).

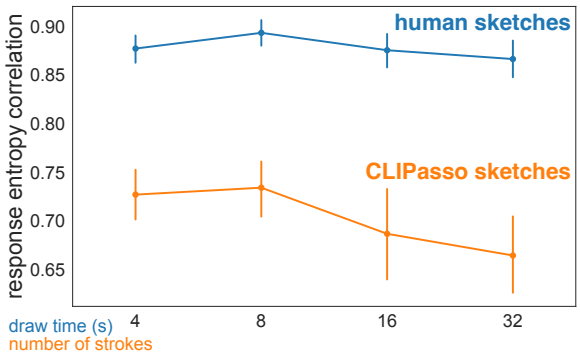


Figure 6: Mean correlation between response entropy vectors across models at different abstraction levels.

We found a high average correlation between entropy vectors across models for each production constraint level and for both sketches generated by human sketchers and CLIPasso (Fig. 6), suggesting that all the models appeared to represent semantic uncertainty in a generally similar manner when classifying sketches of different object classes. Additionally, we conducted model comparisons between two linear regression models predicting response correlation, one with a factor for draw duration and one without. We found that a factor for draw duration did not explain significantly unique variance in the model relative to the intercept-only model ($F(260,3) = 1.78$, $p = .15$). These results indicate that models were equally consistent in their uncertainty when generating classifications of human sketches of different object classes. For sketches generated using CLIPasso, we also compared an intercept-only model to a model with a factor for number of strokes and found a significant effect of number of strokes ($F(3,260) = 3.19$, $p < .05$). Thus, for machine generated sketches, there did appear to be an effect of stroke number on the consistency between models in their expressed semantic uncertainty with models being less consistent in their classifications of sketches with more strokes. Taken together, these results suggest that sketch recognition behavior is generally consistent across these 12 state-of-the-art vision models, particularly for sketches produced by humans and despite being drawn at varying levels of detail.

## Discussion

There has been incredible progress in the development of high performing vision models over the past several years–and recently, reaching enough variation in model architecture and training to prime cognitive psychology research with the opportunity to begin explore the extent to which specific models have achieved human-like understanding of abstract symbolic representations, like sketches. To evaluate this expansive range of state-of-the-art vision models and their ability to extract meaning from sketch representations, we introduce a new large-scale drawing dataset of over 90K sketches based on 2,048 real-world objects spanning 128 diverse concept categories. Our dataset combines both human-made and AI-generated sketches produced under varying production constraints (limitations in drawing duration for human sketchers, and limitations in number of strokes for CLIPasso (Vinker et al., 2022)). By systematically varying the level of abstraction produced by these different production constraints, our dataset provides a rich testbed to investigate human and machine visual understanding of human and machine sketches.

In an era where it is becoming increasingly difficult to adjudicate between state-of-the-art vision models in terms of their relevance to human cognition, our dataset provides a novel substrate to tease apart which models are more human-like than others (Golan, Raju, & Kriegeskorte, 2020). We provide initial benchmarking of a representative set of vision models spanning architectures and training techniques against human generated soft-label distributions for a representative set of sketches from both datasets. Across a battery of model evaluations, we broadly find that the models that we investigated are sensitive to the variation in the semantic information produced by different production constraints and that some models, like CLIP, are more consistent with human recognition of sketches than others, like MLP-Mixer. We also found that models generally agree in their expressed uncertainty over object classes when recognizing sketches and that this uncertainty remains largely consistent across varying levels of abstraction and detail, but with models being less consistent across abstraction levels over their uncertainty in machine generated sketches. This points to an intriguing yet critical gap in the fidelity of sketch production algorithms to human sketch behavior, and we hope results such as ours will help open avenues for closing this gap in human-machine sketch production.

Within our present work, we offer object concept classification as an initial benchmarking protocol to evaluate different models. However, we predict future work will seek to benchmark a wider variety of models including those inspired by cognitive neuroscience (H. Chen et al., 2022; Zhuang et al., 2021; Kubilius et al., 2019) and will seek to build more robust metrics of abstraction beyond those tied to classification-based accuracy.

Our work complements ongoing efforts to understand how different components of current AI vision models give rise to their classification behavior (Hermann & Lampinen, 2020; Nguyen, Raghu, & Kornblith, 2020; Schott et al., 2021; T. Chen, Luo, & Li, 2021) and where limitations might arise in mapping model performance to human cognition (Zhou & Firestone, 2019; Bowers et al., 2022; Mahowald et al., 2023). In the long run, progress along these research fronts may shed light upon the computational basis of humans' ability to produce highly abstract but nonetheless meaningful representations, as well as invented symbolic systems for encoding abstract knowledge in pictorial form.

## Acknowledgments

Many thanks to the members of the Cognitive Tools Lab at UC San Diego for their helpful feedback and support. This work was supported by an NSF CAREER Award #2047191 to J.E.F.. J.E.F is additionally supported by an ONR Science of Autonomy award and a Stanford Hoffman-Yee grant.

All code and materials available at:
`https://github.com/cogtoolslab/visual_abstractions_benchmarking_public2023/`

## References

Baker, N., & Kellman, P. J. (2018). Abstract shape representation in human visual perception. *Journal of Experimental Psychology: General*, *147*(9), 1295.

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.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9650–9660).

Chen, H., Venkatesh, R., Friedman, Y., Wu, J., Tenenbaum, J. B., Yamins, D. L., & Bear, D. M. (2022). Unsupervised segmentation in real-world images via spelke object inference. In *Computer vision–eccv 2022: 17th european conference, tel aviv, israel, october 23–27, 2022, proceedings, part xxix* (pp. 719–735).

Chen, M., Hauser, H., Rheingans, P., & Scheuermann, G. (2020). *Foundations of data visualization*. Springer.

Chen, T., Luo, C., & Li, L. (2021). Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, *34*, 11834–11845.

Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9640–9649).

Collins, K. M., Bhatt, U., & Weller, A. (2022). Eliciting and learning with soft labels from every annotator. In *Proceedings of the aaai conference on human computation and crowdsourcing* (Vol. 10, pp. 40–52).

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).

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eitz, M., Richter, R., Boubekeur, T., Hildebrand, K., & Alexa, M. (2012). Sketch-based shape retrieval. *ACM Transactions on graphics (TOG)*, *31*(4), 1–10.

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*(1), 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.

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, *117*(47), 29330–29337.

Gombrich, E. H. (1995). *The story of art* (Vol. 12). Phaidon London.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25.

Gross, C. G., Rocha-Miranda, C. d., & Bender, D. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of Neurophysiology*, *35*(1), 96–111.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 16000–16009).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

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.

Hermann, K., & Lampinen, A. (2020). What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, *33*, 9995–10006.

Huey, H., & Long, B. (2022). Developmental changes in the semantic part structure of drawn objects. In *Proceedings of the 44th annual meeting of the cognitive science society.*

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.

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

Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children's drawing. *Cognition*, *34*(1), 57–83.

Konkle, T., & Alvarez, G. A. (2022). A self-supervised

domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1), 1–12.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modelling biological vision and brain information processing. *biorxiv*, 029876.

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.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., ... others (2019). Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, *32*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 10012–10022).

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

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Malach, R., Levy, I., & Hasson, U. (2002). The topography of high-order human object areas. *Trends in Cognitive Sciences*, *6*(4), 176–184.

McCloud, S., & Manning, A. (1998). Understanding comics: The invisible art. *IEEE Transactions on Professional Communications*, *41*(1), 66–69.

Mi, X., DeCarlo, D., & Stone, M. (2009). Abstraction of 2d shapes in terms of parts. In *Proceedings of the 7th international symposium on non-photorealistic animation and rendering* (pp. 15–24).

Nan, L., Sharf, A., Xie, K., Wong, T.-T., Deussen, O., Cohen-Or, D., & Chen, B. (2011). Conjoining gestalt rules for abstraction of architectural drawings. *ACM Transactions on Graphics (TOG)*, *30*(6), 1–10.

Nguyen, T., Raghu, M., & Kornblith, S. (2020). Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., & Russakovsky, O. (2019). Human uncertainty makes classification more robust. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9617–9626).

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. T., & Patterson, K. (2007). Object categorization: reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology: General*, *136*(3), 451.

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.

Schott, L., Von Kügelgen, J., Träuble, F., Gehler, P., Russell, C., Bethge, M., ... Brendel, W. (2021). Visual representation learning does not generalize strongly within the same domain. *arXiv preprint arXiv:2107.08221*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2818–2826).

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., ... others (2021). Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, *34*, 24261–24272.

Tversky, B. (1989). Parts, partonomies, and taxonomies. *Developmental Psychology*, *25*(6), 983.

Vinker, Y., Pajouheshgar, E., Bo, J. Y., Bachmann, R. C., Bermano, A. H., Cohen-Or, D., ... Shamir, A. (2022). Clipasso: Semantically-aware object sketching. *arXiv preprint arXiv:2202.05822*.

Viola, I., & Isenberg, T. (2017). Pondering the concept of abstraction in (illustrative) visualization. *IEEE transactions on visualization and computer graphics*, *24*(9), 2573–2588.

Xie, Q., Luong, M.-T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 10687–10698).

Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., & Mahajan, D. (2019). Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*.

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*.

Yu, Q., Yang, Y., Liu, F., Song, Y.-Z., Xiang, T., & Hospedales, T. M. (2017). Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision*, *122*, 411–425.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature communications*, *10*(1), 1334.

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.