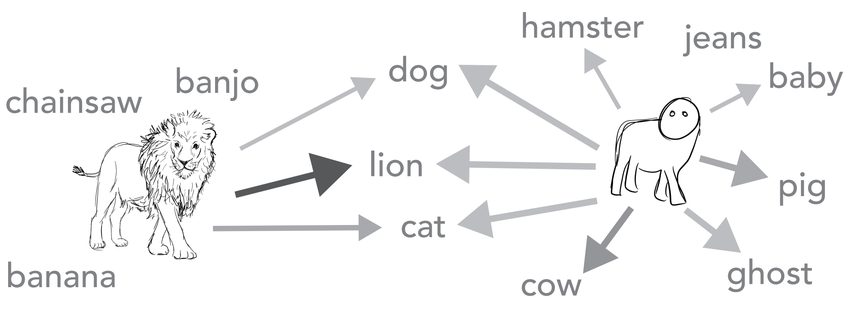
Evaluating machine comprehension of sketch meaning at different levels of abstraction

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

[1]University of Wisconsin-Madison, [2]University of California San Diego, [3]Tel-Aviv University, [4]Reichman University, [5]Stanford University

QUESTION

How well do vision models exhibit human-like understanding of sketches that vary in semantic ambiguity?



METHODS

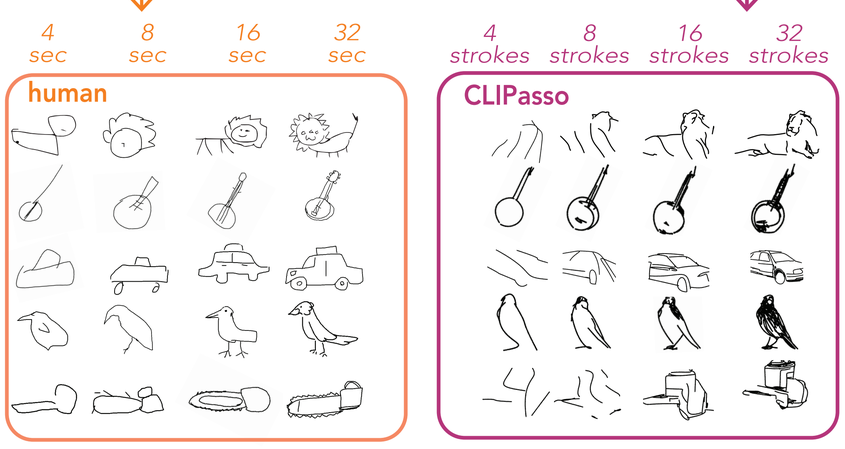
1. Vision Model Selection

	supervised	self-supervised	semi-supervised
convnet	Inception-V3 VGG-19 ResNet-50 ECOSET CORnet-S		Noisy Student SWSL
transformer & MLP	ViT-B Swin-B Harmonization MLP-Mixer-B	MoCo-V3 SimCLR DINO IPCL CLIP MAE	

2. Sketch Dataset Generation

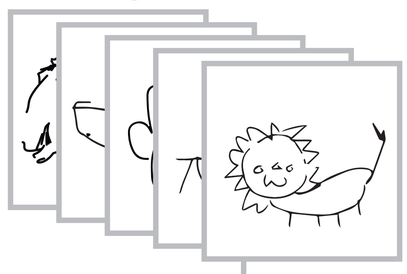
2048 photos from 128 classes from the THINGS database.

N = 5,563



Over 90K sketches including sketches of each concept at 4 levels of detail made by humans and CLIPasso.

3. Measuring Sketch Comprehension



What does this sketch represent?

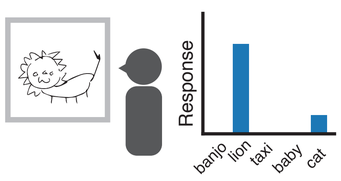
neural network encoders

128-way classification

performance

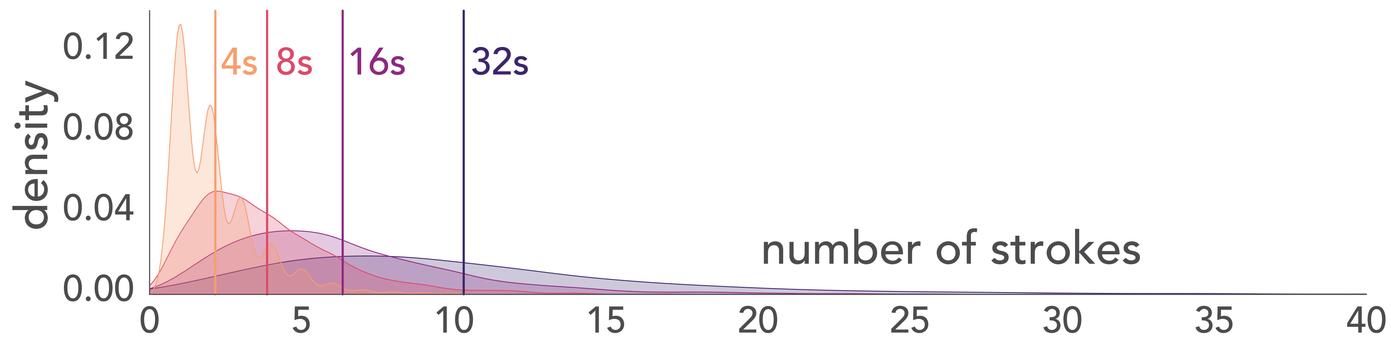
Lion ✓
Cat ✗
Taxi ✗
Dog ✗
...

alignment with human behavior



RESULTS

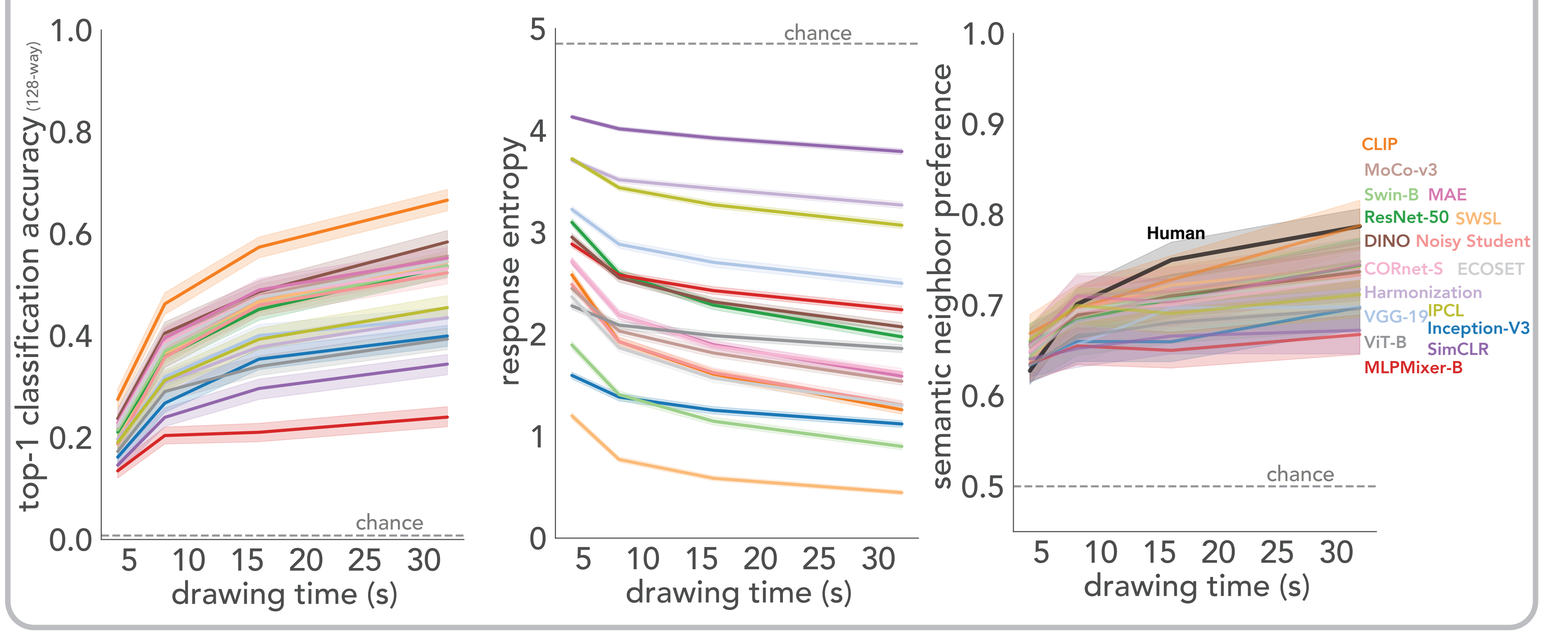
Humans produce sparser sketches under stronger time constraints.



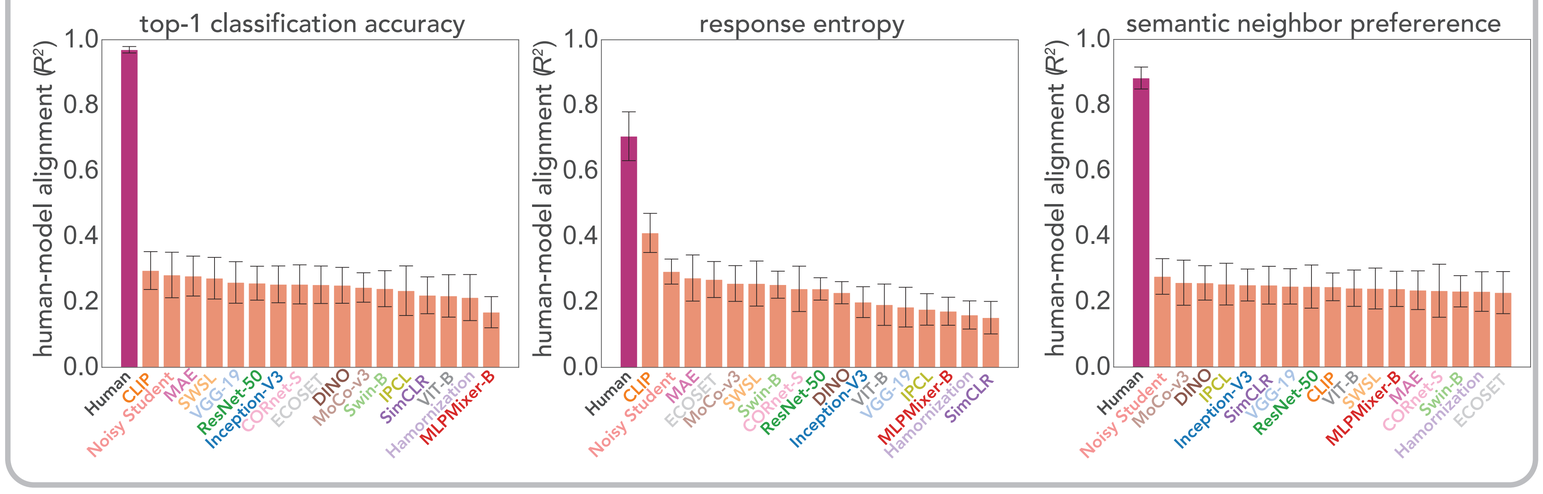
Sparser sketches more semantically ambiguous for models and humans.

Sketches that are more detailed elicit ...

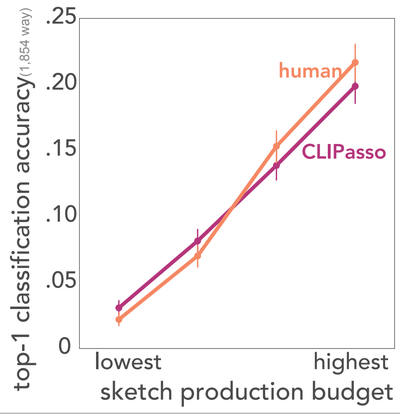
...higher classification accuracy. **...less variable response labels.** **...guesses that are semantically close to the true label when incorrect.**



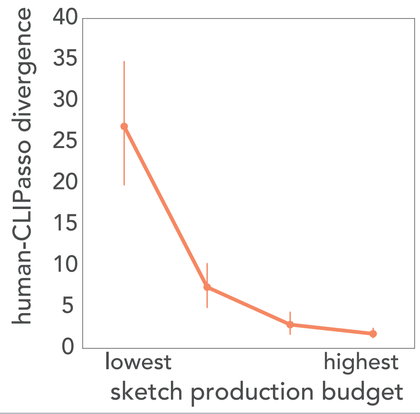
Models vary in their degree of alignment to human behavior but a large gap remains between human and model sketch understanding.



A CLIP-based sketch generation algorithm emulates human sketches at greater levels of detail



Human recognition performance on human-made and CLIPasso-made sketches is comparable.



Labels elicited by humans for human and CLIPasso sketches are more similar for more detailed sketches.

TAKEAWAYS

We introduce a new dataset of >90K sketches at varied abstraction levels made by humans & CLIPasso, an AI-sketch generation model.

State-of-the-art vision models are sensitive to variation in the semantic information conveyed by sketches under different production budgets.

A large alignment gap still remains between the most performant vision models and humans.

CLIPasso-generated and human-made sketches elicit similar responses at greater levels of detail.

correspondence to:
kmukherjee2@wisc.edu

data & materials available at:
https://github.com/cogtoolslab/visual abstractions benchmarking_public2023

paper: