COCO-Search18: A Dataset for Predicting Goal-directed Attention Control

Yupei Chen[1], Zhibo Yang[2], Seoyoung Ahn[1], Dimitris Samaras[2], Minh Hoai[2], and Gregory Zelinsky[1,2,*]

[1]Department of Psychology, Stony Brook University
[2]Department of Computer Science, Stony Brook University
[*]gregory.zelinsky@stonybrook.edu

ABSTRACT

Attention control is a basic behavioral process that has been studied for decades. The currently best models of attention control are deep networks trained on free-viewing behavior to predict bottom-up attention control—saliency. We introduce COCO-Search18, the first dataset of laboratory-quality *goal-directed behavior* large enough to train deep-network models. We collected eye-movement behavior from 10 people searching for each of 18 target-object categories in 6202 natural-scene images, yielding ∼300,000 search fixations. We thoroughly characterize COCO-Search18, and benchmark it using three machine-learning methods: a ResNet50 object detector, a ResNet50 trained on fixation-density maps, and an inverse-reinforcement-learning model trained on behavioral search scanpaths. Models were also trained/tested on images transformed to approximate a foveated retina, a fundamental biological constraint. These models, each having a different reliance on behavioral training, collectively comprise the new state-of-the-art in predicting goal-directed search fixations. Our expectation is that future work using COCO-Search18 will far surpass these initial efforts, finding applications in domains ranging from human-computer interactive systems that can anticipate a person's intent and render assistance to the potentially early identification of attention-related clinical disorders (ADHD, PTSD, phobia) based on deviation from neurotypical fixation behavior.

Keywords: Goal Dataset, Attention Dataset, Fixation Dataset, Gaze Dataset, Visual Search, Inverse-Reinforcement Learning

Introduction

Our control of visual attention comes broadly in two forms. One is bottom-up, where control is exerted purely by the visual input [1,2]. This is the usage of "attention" captured by saliency models, which exploded in popularity in the behavioral fixation-prediction and computer-vision literature [1,3–5]. The other form of control is top-down, where behavioral goals rather than bottom-up salience control the allocation of visual attention. Goal-directed attention control underlies all the things that we *try* to do, and this diversity makes its prediction vastly more challenging than predicting bottom-up saliency, and more important. In addition to its basic research value, a better understanding of goal-directed attention could lead to the development of biomarkers of neurotypical attention behavior against which clinical conditions can be quantitatively compared to advances in intelligent human-computer interactive systems that can anticipate a user's visual goals and render real-time assistance [6–8].

Goal-directed attention has been studied for decades [9–16], largely in the context of visual search. Search is arguably the most basic of goal-directed behaviors; there is a target object and the goal is to find it, or conclude its absence. Goals are extremely effective in controlling the allocation of gaze. Imagine two encounters with a kitchen, first with the goal of learning the time from a wall clock and again with the goal of warming a cup of coffee. These "clock" and "microwave" searches would yield two very different patterns of eye movement, as recently demonstrated in a test of this gedanken experiment [17], and understanding this goal-directed control has been a core aim of search theory. The visual search literature is voluminous (see reviews [18–20]). Given our modeling goals, we focus here on the prediction of image locations that people fixate as they search for objects, and how the selection of these fixations depends on the target goal.

Visual search has a mature literature, rich with many hugely influential theories and models [11–16]. Yet despite its success, over the last years progress has stalled. Our premise is that this is due to the absence of a dataset of search behavior sufficiently large to train deep network models. Our belief is based on observation of what occurred in the bottom-up attention-control literature during the same time. The prediction of fixations during free viewing, the task-less cousin of visual search, has become an extremely active research topic, complete with managed competitions and leaderboards for the most predictive models [21]. The best of these saliency models are all deep networks, and to our point, all of them were trained on large datasets of labeled human behavior [22–26]. For example, one of the best of these models, DeepGaze II [22], is a deep network trained on SALICON [24]. SALICON is a crowd-sourced dataset consisting of images that were annotated with human mouse clicks indicating attentionally-salient image locations. This model of saliency is therefore trained on image locations that were labeled by people to be salient. Without SALICON, DeepGaze II, and models like it [23–26], would not have been possible, and our understanding of free-viewing behavior, widely believed to reflect bottom-up attention control, would be greatly diminished. For the

task of visual search, there is nothing remotely comparable to SALICON [24]. Here we will describe COCO-Search18, the largest dataset of goal-directed fixation behavior in the world. COCO-Search18 was recently introduced at CVPR2020 [27], and our aim in this paper is to elaborate on this dataset and highlight its richness for researchers modeling attention.

Methods

Behavioral Data Collection

COCO-Search18 is built from Microsoft COCO, Common Objects in Context [28]. COCO consists of over 200,000 images of scenes that have been hand-segmented into 80 object categories. This ground truth labeling of objects in images makes COCO valuable for training computer vision models of object detection [28–32]. However, in order for COCO to be similarly valuable for training models of goal-directed attention, these images would also need to be labeled with the locations fixated by people searching for different target-object goals. COCO-Search18 fills this niche by providing these training labels for search behavior.

The dataset consists of a large-scale annotation of a subset of COCO, 18 of its 80 object categories, with goal-directed search fixations. Each of 10 participants searched for each of 18 target-object categories (blocked) in 6,202 COCO images, mostly of indoor scenes. This effort required an average of 12 hours per participant, distributed over 6 days. This substantial behavioral commitment makes it possible to train models of individual searchers[27], although our focus is on group behavior in the current paper. The eye position of each participant was sampled every millisecond using a high-quality eye-tracker under controlled laboratory conditions and procedure, resulting in 70,000,000 gaze-position samples in total. These raw gaze samples were clustered into 299,037 search fixations (∼30,000 per participant), which dropped to 268,760 fixations after excluding those from incorrect trials. See Figure 1 for representative images and fixation data for each target category. See SM1 for details about: category, image, and fixation selection criteria, the eye tracker and eye tracking procedure, participant instruction, and a comparison between COCO-Search18 and existing datasets of search behavior.

Search-Relevant COCO-Search18 Statistics

Figure 2A shows several search-relevant characterizations of the COCO-Search18 images. The distribution of target-object bounding-box sizes (left) skewed toward smaller targets, with the range constrained by image selection to be between 1% and 10% of the image size (see SM1). The mean visual angle of the targets, based on the square root of its bounding box size, was 8.4°, about the size of a clenched fist at arm's length. The middle panel shows the distribution of initial target eccentricities, which is how far the target appeared in peripheral vision based on central fixation at the start of search. Target eccentricities ranged from 10° to 25°, with a mean of ∼15° eccentricity. The distribution of the number of "things" in each image (right), based on the COCO object and stuff labels[33], shows that some images depicted only a handful of objects whereas others depicted 20 or more (keeping in mind that this labeling was coarse). This is potentially important because search efficiency is known to degrade with the number of items in a search array[34], and a similar relationship has been suggested for the number of objects appearing in scenes[35–37]. See SM2 for these analyses broken down by target category, and correlated with search efficiency.

Search Procedure and Metrics

The paradigm used for data collection was speeded categorical search [38–40]. The participant's task was to indicate whether an exemplar of a target category appeared in an image of a scene (Figure S3). They did this by making a target present/absent judgment as quickly as possible while maintaining accuracy. The target category was designated at the start of a block of trials. Half of the search images depicted an exemplar of a target (target-present, TP), and the other half did not (target-absent, TA).

Our measure of goal-directed attention control is the efficiency that a person's gaze moves to a target in a search task. Because the target was an object category, the specific term for this measure of search efficiency is *categorical target guidance* [38,40], here defined as the controlled direction of gaze to a target-category goal. We consider multiple measures of target guidance in Figure 3, but here we focus on the cumulative probability of fixating the target after each search movement [41–44]. A target category that is successful in controlling gaze will be fixated in fewer eye movements compared to a target category having less capacity for target guidance. A desirable property of the target-fixation-probability (TFP) function (Figure 2B and C) is that it is meaningful to compute the area under the TFP curve (TFP-auc), which we suggest as a new metric for evaluating search guidance across target categories and models.

Model Comparison

Now that COCO-Search18 exists, what can we do with it? To answer this question we conducted benchmarking to determine how well current state-of-the-art methods, using COCO-Search18, can predict categorical search fixations. To create a context for this model comparison we considered three very different modeling approaches, which all shared a common backbone model architecture, a ResNet50 pre-trained on ImageNet [45].

Our first approach predicted search fixations using object detectors trained for each of the target categories. We did this by re-training the pre-trained ResNet50 on just the 18 target categories using the COCO labels. Standard data augmentation methods of re-sizing and random crops were used to increase variability in the training samples. We then used these trained detectors to predict search fixations on the test images. For a given target and test image, we obtained a confidence map from the target detector and used it to sample a sequence of fixation locations in decreasing level of confidence. Note that this approach is pure computer vision,

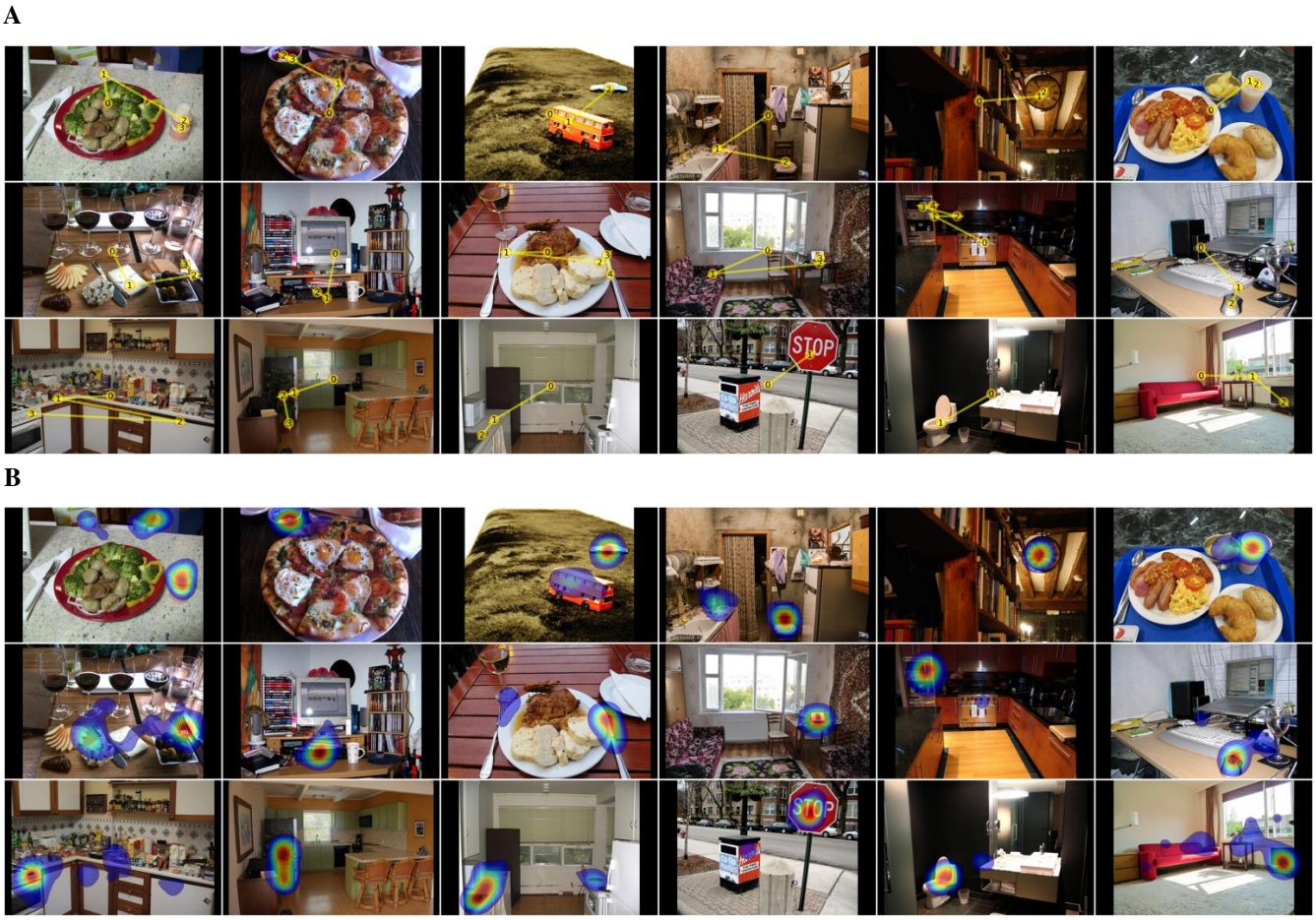


Figure 1. (A). Examples of a target-present image for each of the 18 target categories. Yellow lines and numbered discs indicate a representative search scanpath from a single participant. From left to right, top to bottom: bottle, bowl, car, chair, analog clock, cup, fork, keyboard, knife, laptop, microwave, mouse, oven, potted plant, sink, stop sign, toilet, tv. (B). Examples of fixation density maps (excluding initial fixations at the center) computed over participants for the same scenes.

meaning that it uses the image pixels solely and knows nothing about behavior.

With COCO-Search18, however, it is possible to also train on the search behavior. There are multiple ways of doing this. In our second approach we re-trained the same ResNet50, only this time using the fixations made by searchers viewing the training images as labels. These fixations were represented as feature embeddings consisting of fixation-density maps (FDMs) obtained for the TP training images of each target category. This model is in a sense a search version of models like DeepGaze II [22] in the free-viewing fixation-prediction literature, where FDMs are also embedded as training features. We therefore refer to this model as Deep Search. Deep Search differs from the Target Detector model in that it is trained by search fixation density maps to predict search behavior.

For our third modeling approach we used inverse-reinforcement learning (IRL) [46–48], an imitation-learning method from the machine-learning literature, to simply mimic the search scanpaths observed during training. We chose IRL over other imitation-learning methods because it is based on reward, known to be a powerful driver of behavior [49–51], but we think it is likely that other imitation-learning methods would perform similarly. The IRL model we used[48] works by learning, through an adversarial process playing out over many iterations, how to make model-generated behavior, initially random, become more like human behavior. It does this by rewarding behavior that happens to be more human-like. IRL is therefore very different from a Target Detector, but also different from Deep Search, which also gets to use search behavior in its training. The IRL model learns to imitate the search scanpath, meaning the sequence of fixations made to the search target, whereas Deep Search uses only FDMs that do not represent the temporal order of fixations. Because the IRL model used the most search behavior for training, we hypothesized that it would best predict search behavior in our model comparison. See SM3 for additional details about IRL.

State Comparison

In addition to the model comparison, we also compare several state representations used by the models. In the current

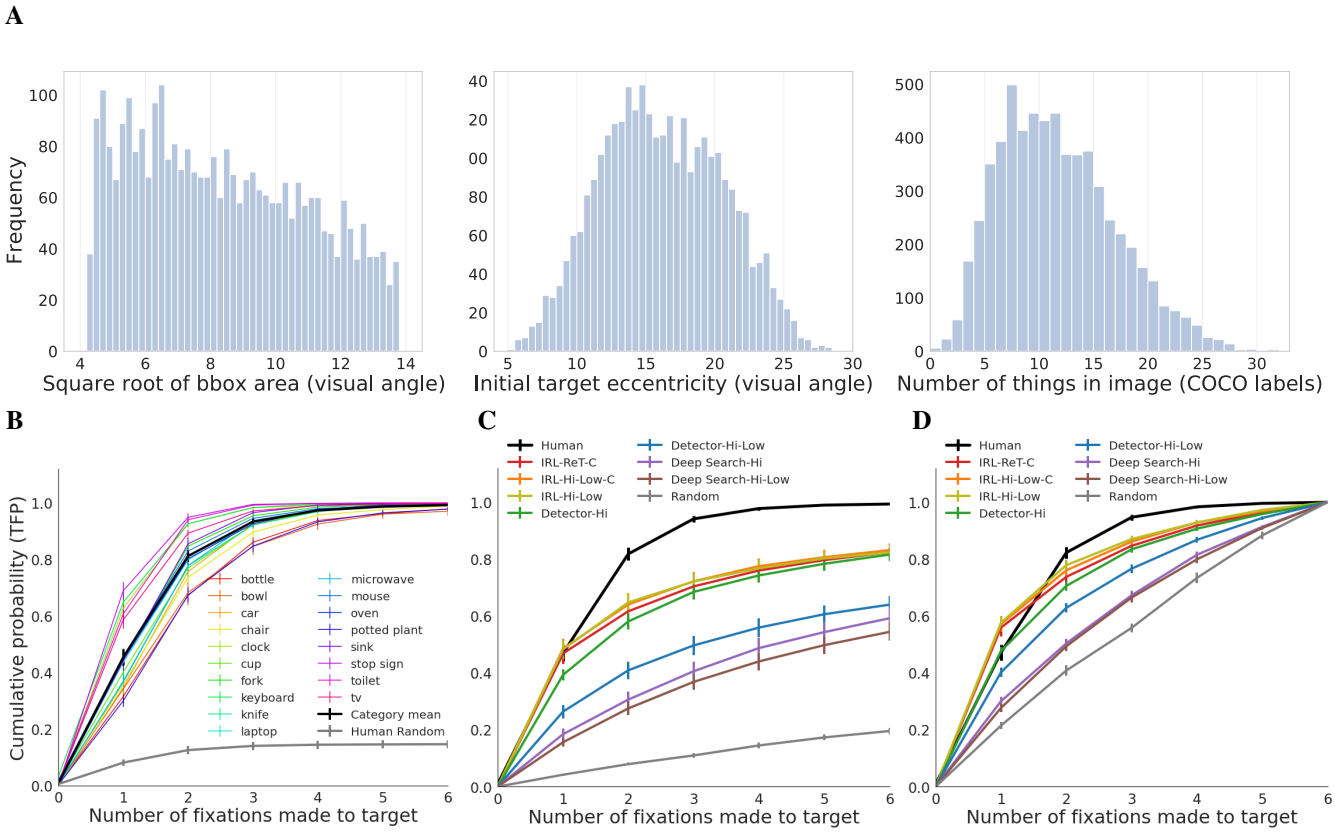


Figure 2. (A). Distributions of target sizes, based on the visual angle of their bounding-box areas (left), and initial target eccentricities (middle), both for the target-present images. The number of things (objects and "stuff" categories, both based on COCO-stuff labels) for all the images (right). (B). Cumulative probability of fixating the target (y-axis) as a function of fixation serial position (0-6, x-axis), shown individually for the 18 categories (color lines) and averaged over target types (bold black line). The bottom-most function is a Human Random baseline obtained by computing target-fixation probability using a scanpath from the same participant searching for the same target category but in a different image. Means and standard errors are first computed over images, and then over participants. (C). TFP functions for all of models on the test images. Names designate a model type (IRL, Detector, Deep Search) and a state representation (ReT, Hi-Low, Hi, C), separated by hyphens. Average behavioral TFP is also plotted in bold black, this time for just the test data (which explains the small differences from the corresponding function in B, which included the training and testing data). This Random baseline was obtained by making six movements of the Hi-Low foveated retina, with ISTs after each, and determining whether any of these movements brought the high-resolution central window to the target. Means and standard errors are first computed over images, and then over categories. (D). A re-plot of C but only including data from trials in which the target was fixated within the first six fixations (i.e., search scanpaths that succeed in locating the target.

context, the state is the information that is available to control search behavior, and essential to this are the features extracted from each search image. We refer to the original images as high-resolution (Hi-Res), in reference to the fact that they were not blurred to reflect retina constraints. Extracting features from a Hi-Res image produces a Hi-Res state, and it is this state that is used by most object-detection models in computer vision where the goal is to maximize detection success. Primate vision, however, is profoundly degraded from this Hi-Res state by virtue of the fact that we have a foveated retina. A foveated retina means that high-resolution visual inputs exist only for a small patch of the image at the current fixation location, and blurred everywhere else. Given our goal to model the fixation behavior of searchers, each of whom had a foveated retina, we included this basic biological constraint in the state to determine its importance in model training and prediction of search behavior[52]. As fundamentally, each new fixation changes the state by allowing high-resolution information to be obtained from the vantage of a new image location, and capturing these spatio-temporal state changes in the context of search was a core goal in the development of COCO-Search18.

We considered several related fovea-inspired states. First, we used the method from Perry and Geisler [53] to compute a Retina-Transformed (ReT) image, a blurred version of the Hi-Res image that approximates the gradual loss in visual acuity that occurs from viewing at increasing eccentricities in peripheral vision. Second, we implemented an even more simplified foveated retina consisting of just a high-resolution central patch ($5^\circ \times 5^\circ$ visual angle) surrounded by low-resolution

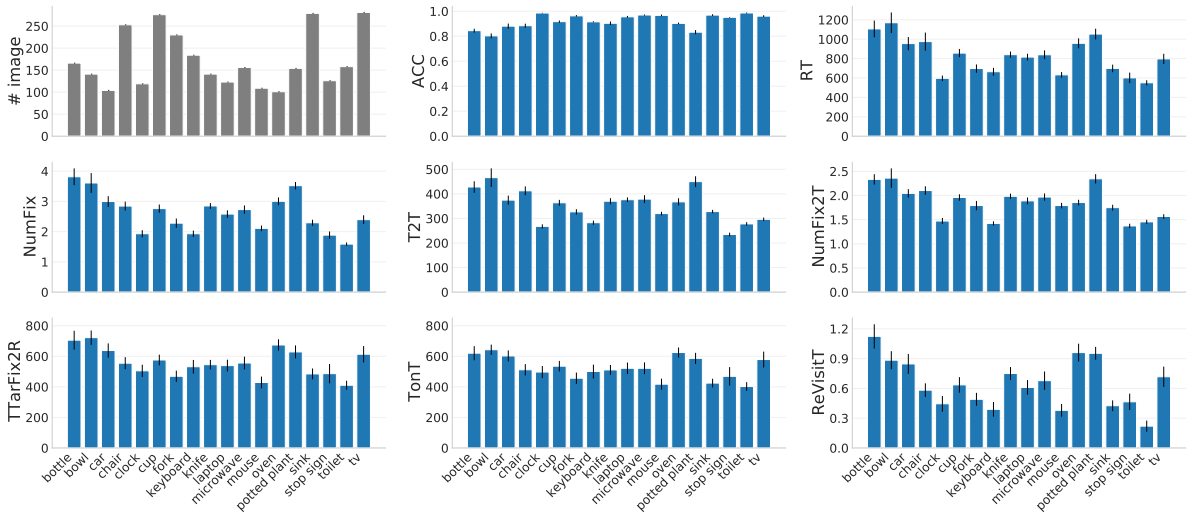


Figure 3. Basic behavioral analyses of the target-present data from COCO-Search18, grouped by the 18 target categories. The gray plot (top left) shows the number of images in each category. The blue plots show eight behavioral analyses. Top row: manual measures of response accuracy (ACC) and reaction time (RT). Middle row: gaze-based analyses of categorical guidance efficiency. Left to right: number of fixations made before the button press (NumFix), time until first target fixation (T2T), and number of fixations made until first target fixation (NumFix2T). Bottom: gaze-based measures of target verification. Left to right: time from first target fixation until response (TTarFix2R), total time spent fixating the target (TonT), and the number of re-fixations on the target (ReVisitT). Values are means over 10 participants, and error bars represent standard errors.

"peripheral" vision elsewhere, with the critical difference from the ReT image being that only a single level of blur was used to approximate the low-resolution periphery (obtained using a Gaussian filter with sigma = 2). Although inclusion of this Hi-Low state was motivated largely by the need to reduce the computational demands created by gradually blurring a ReT image (requiring 10∼15 times the processing time per image), having this condition also enables a needed initial evaluation of how veridically low-level visual-system constraints need to be followed when training deep-network models of human goal-directed behavior. Our third state representation seeks to capture the time-varying changes that occur with each fixation made during search.

A behavioral consequence of having a fovated retina is that we make saccadic eye movements, and we explored two ways of capturing the fixation-contingent changes in state that occur with every saccade. One assumes a high-resolution foveal window that simply moves within a blurred image, such that each change in fixation will bring currently blurred visual information into clearer view. This conception of the oculomotor system is the one that first comes to mind because it follows most closely the movement of a foveated retina over a visual input. We will consider this to be the default state. However, this default state also assumes that information obtained on fixation n is completely lost by fixation $n+1$, and while something like this is true for the high-resolution visual information[54] it is far less clear whether people also forget the objects that they recognized from previous fixations while in the high-resolution states[55]. To address the potential for an object context building over fixations, we use a state that accumulates the high-resolution foveal views obtained at each search search fixation, a state we refer to as Cumulative (-C). Over the course of multi-fixation search, the Hi-Low-C state would therefore accumulate high-resolution foveal snapshots with each new fixation, progressively de-blurring what would be an initially moderately-blurred version of the image. We explore these two extremes of information preservation during search so as to inform future uses of a foveated-retina state representation to train deep network models.

General Model Methods

All of the models followed the same general pipeline. Each 1050×1680 image input was resized to 320×512 pixels to reduce computation. This is what we refer to as the Hi-Res image; the ReT image is computed from this, as is the Low version of the Hi-Low image. These images were passed through the ResNet50 backbone to obtain 20×32 feature map outputs, with the features extracted from these images now reflecting either Hi-Res, ReT, or Hi-Low states. Different models were trained using these features and others, as described in the Model Comparison section, and all model evaluations were based on a 70% training, 10% validation, and 20% testing, random split of COCO-Search18 within each target category. See SM3 for additional details about the training and testing separation, and Figure S15 for how the two compare on search performance measures.

	TFP-AUC ↑	Probability Mismatch ↓	Scanpath Ratio ↑	Sequence Score ↑	MultiMatch ↑			
					shape	direction	length	position
Human	5.200	-	0.862	0.490	0.903	0.736	0.880	0.910
Random	0.744	4.455	0.392	0.283	0.823	0.563	0.805	0.751
Detector-Hi	4.001	1.209	0.680	0.416	0.882	0.587	0.875	0.804
Detector-Low	2.975	2.225	0.601	0.379	0.861	0.585	0.849	0.786
Deep Search-Hi	2.519	2.681	0.579	0.355	0.881	0.573	**0.887**	**0.849**
Deep Search-Low	2.282	2.918	0.546	0.343	0.874	0.573	0.877	0.837
IRL-ReT-C	4.170	1.131	0.731	0.422	0.865	**0.590**	0.853	0.792
IRL-Hi-Low-C	**4.262**	**1.031**	0.747	**0.424**	**0.884**	0.583	0.876	0.814
IRL-Hi-Low	4.245	1.036	**0.753**	0.421	0.877	0.582	0.869	0.813

Table 1. Results from scanpath prediction models (rows) using multiple scanpath metrics (columns) applied to the COCO-Search18 test images. Arrows indicate the direction of better prediction success, and values in bold indicate best predictions across the model comparison. In the case of Sequence Score and Multimatch, "Human" refers to an oracle method whereby one searcher's scanpath is used to predict another searcher's scanpath; "Human" for all other metrics refers to observed behavior. See main text for additional details about the scanpath-comparison metrics, and SM3 for purely spatial comparisons using the AUC, NSS, and CC metrics.

The trained models were used to obtain model-specific priority maps for the purpose of predicting the search fixations in each test image. The priority map for the Target Detector was a map of detector confidence values for each pixel location, and fixations were sampled in decreasing order of confidence.The priority map for Deep Search uses a different objective function, one that maps image features to FDMs obtained from the behavioral search ground-truth. The priority map for the IRL model reflects a maximization of the total reward expected by making a sequence of search fixations to different locations in an image, based on the reward function learned during training to mimic observed search behavior. The IRL-ReT-C model therefore accumulates graded-resolution foveal views of an image with each reward-driven eye movement. Although our aim is to explore as systematically as possible the space of models ' states, for some models a given state representation is not applicable. For example, it makes no sense for the IRL model to use the Hi-Res state. Because that state representation does not change from one search fixation to the next, it would be impossible to learn fixation-dependent changes in state. Similarly, it makes no sense to have a cumulative state for anything but the IRL model, as the others would be unable to use this information. However, it does make sense to test a Target Detector and Deep Search on both Hi-Res and Hi-Low states, which are reflected in the Table 1 conditions.

The IRL models were constrained to have an action space discretized into a 20×32 grid, which again was done mainly to reduce computation time. A given action, here a change in fixation, is therefore a selection of 1 from 640 possible grid cells, which amounts to a limitation imposed on the spatial resolution of the models' oculomotor system. The selected cell was then mapped back into 320×512 image space by upscaling, and the center of this cell became the location of the model's next fixation. The non-IRL models made their action selection in the 320×512 image space, with higher priority values selected with higher probability. For each test image, all of the model×state combinations in our comparison were required to make six changes in fixation. This number was informed by the behavioral data showing that the probability of target fixation was at ceiling by the sixth eye movement (Figure 2B). To produce these 6-fixation scanpaths, we iterated the fixation generation procedure using inhibitory spatial tagging (IST). IST is a mechanism serving the dual functions of (1) breaking current fixation, thereby enabling gaze to move elsewhere, and (2) discouraging refixations of previously searched locations. IST has long been used by computational models of free viewing [56,57] and search [58,59]. Here we enforce IST by setting the priority map to zero after each fixation over a region having a radius of 2.5° visual angle (based on a 3×3 grid within the 20×32 action space). See SM3 for additional details.

Behavioral and Model Performance

Behavioral Performance

We interrogated COCO-Search18 using multiple performance measures. Figure 3 reports many of these analyses, grouped by manual responses (accuracy and response time), gaze-based measures of categorical guidance (number of fixations before the button press, and both the time and number of fixations until first target fixation), and measures of target verification time (time from first target fixation until the button press, time spent fixating the target, and the number of target re-fixations). Figure 4 is a visualization of agreement between these measures. The columns show ranked performance, with dark red indicating more difficult (or less efficient) search and dark blue indicating relatively easy or efficient search. The rows in Figure 4A group the performance data by target category, where clear consistency exists among the measures. For example, the two predominately red rows near the top indicate agreement in that the bottle and bowl objects were difficult

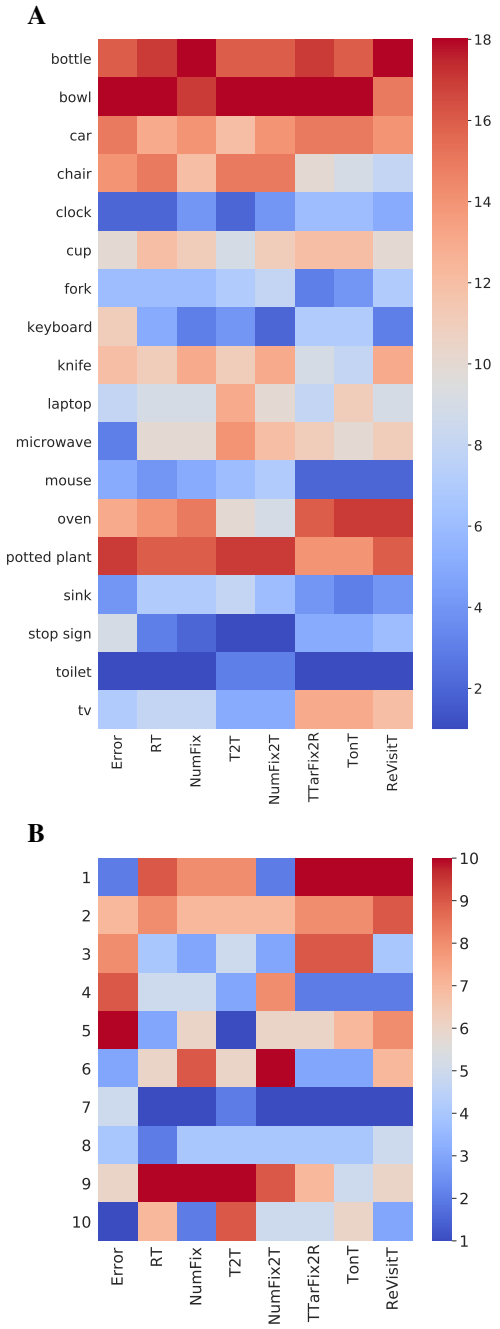


Figure 4. (A). Ranked target-present data [1-18] based on averages over participants, grouped by target category (rows) and shown for multiple performance measures (columns). These measures include: response Error, reaction time (RT), number of fixations (NumFix), time to target (T2T), number of fixations to target (NumFix2T), time from first target fixation until response (TTarFix2R), time spent fixating the target (TonT), and the number of target re-fixations (ReVisitT). Redder color indicates higher rank and harder search targets, bluer color indicates lower rank and easier search. (B). A similar ranking of the target-present data, now averaged over target category and grouped by participant (rows 1-10). Performance measures and color coding are the same as in 3A.

targets, speculatively because these target categories have particularly high variability in their image exemplars. Relatedly, appearing near the bottom are two of the consistently easiest targets, stop signs and toilets, both having relatively well-defined category membership. Figure 4B shows a similar plot, only now performance is averaged over target categories and plotted for individual participants. Here we see that search accuracy and efficiency clearly differ across participants. Participants 7 and 8 are better searchers than Participants 2 and 9, meaning that they tended to find the target faster and with fewer fixations while keeping a low error rate. Differences in search strategy can also be seen from this visualization. Participant 1 searched the display carefully, resulting in few missed targets, but this person's search was not very efficient. In contrast, Participant 4 was quick to find and verify targets, but had relatively low accuracy. See SM2 for parallel analyses of the target-absent data.

However, arguably the gold-standard measure of attention control is the target-fixation-probability (TFP) after each saccade made during search. Figure 2B shows TFP functions for the first six search saccades, averaged over participants and plotted for individual target categories. The mean behavior over targets is indicated by the bold black function, which is replotted in Figure 2C. The noteworthy pattern is that the slope of the group function is far steeper than that of the chance baselines. These random models were obtained by computing TFP either by using a scanpath from a different image but from the same person and target category (Figure 2B) or by generating six movements of a Hi-Low foveated retina, with ISTs, and determining whether the target falls within its high-resolution central window (Figure 2C). On average, about half of the targets were fixated with the very first saccade. By the second saccade TFP jumped to .82, and by the third saccade it reached a performance ceiling of .94, which increased only slightly after saccades 4-6. This high degree of attention control means that, although we aimed to create a search dataset having a moderate level of difficulty, COCO-Search18 skews easy. This is due in part to unexpectedly large practice effects (see SM2 for details). However, it is fortuitous that such a strong attention-control signal exists in the behavioral data, given the challenge faced by even start-of-the-art models in predicting this simple search behavior.

Model Evaluation

Fixation prediction models broadly fall into two groups, those that focus exclusively on the spatial distribution of fixation locations, and those that also predict the spatio-temporal order of the fixations. We evaluated prediction success for both spatial FDMs and search scanpaths. Moreover, because we required all models to generate sequences of fixations for the images in the test dataset, predicted fixations can be directly compared to the fixation behavior of people searching the same scene for the same target. To compare predicted against observed FDMs, we adopted three metrics: Area Under the Receiver Operating Characteristic Curve (AUC), which treats a priority map

as a binary classifier discriminating fixated from non-fixated image locations; Normalized Scanpath Saliency (NSS), the averaged normalized prediction value at behavioral fixation locations; and a Pearson's Correlation Coefficient (CC), capturing the degree of linear relationship between predicted and behavioral FDMs (for additional discussion, see: Borji & Itti [5]; Bylinskii, et al.[60]). Comparing spatio-temporal scanpaths requires different metrics. For this we used: TFP-auc, which is the area under the cumulative target-fixation-probability curve; Probability Mismatch, which sums over each fixation in a scanpath the absolute differences between the behavioral and model TFP; Scanpath Ratio, the Euclidean distance between the initial fixation location (roughly the center of the image) and the location of the target (center of bounding box) divided by the summed Euclidean distances between the fixation locations in the search scanpath[41,61–63]; Sequence Score, which clusters scanpaths into strings and uses a string matching algorithm for comparison [64]; and MultiMatch, which takes a multi-dimensional approach to computing scanpath similarity [65,66]. Table 1 provides an exhaustive evaluation of how each model×state combination fared in fair comparison using these scanpath metrics of prediction success. As we hypothesized, the IRL model outperformed the others, and did so for every metric except MultiMatch, where it tied with Deep Search. See Table S1 for analyses of the FDMs, and SM4 for additional details about the metrics.

To evaluate search efficiency, we compute TFP for each model and directly compare these functions to the behavioral TFP data. Figure 2C shows this for the model×state combinations (in color) and for the mean search behavior (in black). Focusing first on state comparisons, we did not find large differences between the states tested. For example, whether blur was graded or binary appeared not to matter, as indicated by the very similar TFP functions for the ReT and the Hi-Low states using the IRL model. This pattern also appears in Table 1, where the IRL models differ by tiny margins. For this reason, and its far greater computational efficiency, we adopted only the Hi-Low state in the other model comparisons (e.g., there is no Deep Search-ReT model). Similarly, but specific to the IRL model, it made little difference whether or not the state accumulated high-resolution visual information with each fixation in a search scanpath. The inability of an IRL model to use this accumulated visual information is broadly consistent with the view that very little high-resolution information is preserved across saccades [54]. However, it did matter whether the state included a foveated retina or not, as indicated by the difference between Hi-Res and Hi-Low states for the Detector and Deep Search models. This state comparison suggests that future work may want to avoid manipulations of fine-grained retinal blur and assumptions about intersaccadic visual memory, and focus on adding more basic limitations on human visual perception to a model's pipeline, with the inclusion of a Hi-Low foveated retina being one example.

All of the tested models made reasonably good predictions of search behavior in this challenging benchmark, where "good" is liberally defined as bearing greater resemblance to the human behavior than the chance baseline. However, the Deep Search models and the Detector-Low model were clearly less efficient in their search behavior than either human behavior or any of the IRL models. This poor relative performance likely stems from the fact that these models were not capturing the serial order of search fixations, and that this order mattered. A corollary finding is that the IRL models, because they learned these spatio-temporal sequences of search fixations, better predicted search behavior. This is true broadly for the IRL models, which all perfectly predicted search efficiency at the first fixation. Also interesting is the degree that an object detector (Detector-Hi) can predict search behavior, supporting previous speculation [67]. If an application's goal is to predict a person's fixation behavior during search without regard for biological plausibility, a simple object detector will work well based on our testing with COCO-Search18. Another finding from Figure 2C is that none of the models achieved the high level of successful target fixation exhibited in human performance. Performance ceilings after six saccades (termed *fixated-in-6 accuracy*) ranged from .54 (Deep Search-Hi-Low) to .83 (IRL-Hi-Low-C), all well below the near perfect fixated-in-6 accuracy from human searchers. These lower performance plateaus, undoubtedly reflecting limitations in current object recognition methods, means that the models tended either to fixate the target efficiently in the first one or two eye movements (like people), or tended not to fixate the target at all (unlike people). If a model could not represent the features used for target guidance as robustly as people, there may be images for which it generates no guidance signal at all. Search efficiency on these trials would be very poor and typically extend beyond six eye movements, hence the performance plateaus.

These different performance ceilings are problematic in that they conflate limitations arising from object recognition with limitations in effective target prioritization, as measured by search efficiency. As a case in point, a strength of the TFP-auc metric is that it is grounded in the TFP functions from Figure 2C, but this means that it includes the different performance ceilings in its measure and this weakens it as a pure measure of attention control. To address this problem, in Figure 2D we again plot TFP functions, but now for only for trials in which the target was successfully fixated within the first six saccades. By restricting analysis to only trials having perfect fixated-in-6 accuracy, the metric becomes more focused on search efficiency. By this measure, and keeping in mind that the data are now skewed toward easier searches, the IRL-Hi-Low-C and IRL-Hi-Low models remain the most predictive overall, although now all IRL models overestimate slightly the efficiently of the first search saccade. But perhaps the biggest winner in this comparison is the Detector-Hi model, which now predicts TFP perfectly after the first fixation, and has generally improved performance for subsequent fixations. We tentatively conclude that simple prioritization of fixations by an object detector predicts reasonably well the prioritiza-

tion made by attention in visual search. The losers in this comparison were the Deep Search models, which remained less efficient than human behavior even after normalization for fixated-in-6 accuracy.

Discussion

Recent years taught us the importance of large datasets for model prediction. COCO-Search18 is currently the largest dataset of goal-directed attention control. It consists of search fixations in sufficient number to be used as labels for training deep network models. We conducted a systematic (but still incomplete) exploration of models and state representations to provide some initial context for the types of model predictions that are possible using COCO-Search18, given current state-of-the-art (or nearly so). The model comparison focused on the degree search behavior was used during training, ranging from none (Detector), to some (Deep Search), to entire search-fixation scanpaths (IRL). With respect to the IRL model, its use with COCO-Search18 is the first attempt to predict the spatio-temporal movements of goal-directed attention by training on human search behavior. In our state comparison we focused on the different ways that a primate foveated retina, and its movement, might be represented, learned, and used to improve fixation prediction models. We also thoroughly analyzed the dataset properties and search behavior, meaning that COCO-Search18 can be used immediately to start generating new testable hypotheses using multiple behavioral measures and metrics. But likely the greatest contribution of our work is as yet unknown. With a dataset the size and quality of COCO-Search18, opportunities exist to explore new policies or reward functions that might better predict goal-directed control, and hopefully strengthen the bridge between behavioral science and machine learning that has built using the concept of attention.

COCO-Search18 is now part of the MIT/Tuebingen Saliency Benchmark, previously the MIT Saliency Benchmark but renamed to reflect the new group that will be managing the competition. The training, validation, and test images in COCO-Search18 are already freely available as part of COCO[28]. Researchers are also free to see and use COCO-Search18's training and validation search fixations, but the fixations on the test images are withheld. As part of a managed benchmark, in a separate track it will be possible to upload predictions and have them evaluated on the test dataset. We hope that you enjoy using COCO-Search18: https://saliency.tuebingen.ai/datasets/COCO-Search18/

References

1. Itti, L., Koch, C. & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis machine intelligence* **20**, 1254–1259 (1998).
2. Itti, L. & Koch, C. Computational modelling of visual attention. *Nat. reviews neuroscience* **2**, 194–203 (2001).
3. Harel, J., Koch, C. & Perona, P. Graph-based visual saliency. In *Advances in neural information processing systems*, 545–552 (2007).
4. Borji, A., Sihite, D. N. & Itti, L. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Process.* **22**, 55–69 (2012).
5. Borji, A. & Itti, L. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis machine intelligence* **35**, 185–207 (2013).
6. Kurylo, U. & Wilson, J. R. Using human eye gaze patterns as indicators of need for assistance from a socially assistive robot. In *International Conference on Social Robotics*, 200–210 (Springer, 2019).
7. Admoni, H. & Srinivasa, S. Predicting user intent through eye gaze for shared autonomy. In *2016 AAAI Fall Symposium Series* (2016).
8. Krishna Sharma, V., Saluja, K., Mollyn, V. & Biswas, P. Eye gaze controlled robotic arm for persons with severe speech and motor impairment. In *ACM Symposium on Eye Tracking Research and Applications*, 1–9 (2020).
9. Buswell, G. T. *How people look at pictures: a study of the psychology and perception in art.* (Univ. Chicago Press, 1935).
10. Yarbus, A. L. Eye movements during perception of complex objects. In *Eye movements and vision*, 171–211 (Springer, 1967).
11. Treisman, A. & Souther, J. Search asymmetry: A diagnostic for preattentive processing of separable features. *J. Exp. Psychol. Gen.* **114**, 285 (1985).
12. Duncan, J. & Humphreys, G. W. Visual search and stimulus similarity. *Psychol. review* **96**, 433 (1989).
13. Chelazzi, L., Miller, E. K., Duncan, J. & Desimone, R. A neural basis for visual search in inferior temporal cortex. *Nature* **363**, 345–347 (1993).
14. Wolfe, J. M. Guided search 2.0 a revised model of visual search. *Psychon. bulletin & review* **1**, 202–238 (1994).
15. Najemnik, J. & Geisler, W. S. Optimal eye movement strategies in visual search. *Nature* **434**, 387–391 (2005).
16. Torralba, A., Oliva, A., Castelhano, M. S. & Henderson, J. M. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychol. review* **113**, 766 (2006).
17. Zelinsky, G. *et al.* Benchmarking gaze prediction for categorical visual search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0 (2019).
18. Eckstein, M. P. Visual search: A retrospective. *J. vision* **11**, 14–14 (2011).

19. Hollingworth, A. Guidance of visual search by memory and knowledge. In *The influence of attention, learning, and motivation on visual search*, 63–89 (Springer, 2012).

20. Wolfe, J. M. Visual search. In *The handbook of attention*, 27–56 (2015).

21. Judd, T., Ehinger, K., Durand, F. & Torralba, A. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, 2106–2113 (IEEE, 2009).

22. Kummerer, M., Wallis, T. S., Gatys, L. A. & Bethge, M. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 4789–4798 (2017).

23. Jia, S. & Bruce, N. D. Eml-net: An expandable multi-layer network for saliency prediction. *Image Vis. Comput.* 103887 (2020).

24. Jiang, M., Huang, S., Duan, J. & Zhao, Q. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1072–1080 (2015).

25. Liu, N. & Han, J. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *IEEE Transactions on Image Process.* **27**, 3264–3274 (2018).

26. Cornia, M., Baraldi, L., Serra, G. & Cucchiara, R. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Process.* **27**, 5142–5154 (2018).

27. Yang, Z. *et al.* Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 193–202 (2020).

28. Lin, T.-Y. *et al.* Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755 (Springer, 2014).

29. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788 (2016).

30. Liu, W. *et al.* Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37 (Springer, 2016).

31. Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890 (2017).

32. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969 (2017).

33. Caesar, H., Uijlings, J. & Ferrari, V. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1209–1218 (2018).

34. Wolfe, J. M. What can 1 million trials tell us about visual search? *Psychol. Sci.* **9**, 33–39 (1998).

35. Rosenholtz, R., Li, Y. & Nakano, L. Measuring visual clutter. *J. vision* **7**, 17–17 (2007).

36. Neider, M. B. & Zelinsky, G. J. Cutting through the clutter: Searching for targets in evolving complex scenes. *J. Vis.* **11**, 7–7 (2011).

37. Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I. & Sherman, A. M. Visual search for arbitrary objects in real scenes. *Attention, Perception, & Psychophys.* **73**, 1650 (2011).

38. Schmidt, J. & Zelinsky, G. J. Search guidance is proportional to the categorical specificity of a target cue. *Q. J. Exp. Psychol.* **62**, 1904–1914 (2009).

39. Castelhano, M. S., Pollatsek, A. & Cave, K. R. Typicality aids search for an unspecified target, but only in identification and not in attentional guidance. *Psychon. Bull. & Rev.* **15**, 795–801 (2008).

40. Maxfield, J. T., Stalder, W. D. & Zelinsky, G. J. Effects of target typicality on categorical search. *J. vision* **14**, 1–1 (2014).

41. Henderson, J. M., Weeks Jr, P. A. & Hollingworth, A. The effects of semantic consistency on eye movements during complex scene viewing. *J. experimental psychology: Hum. perception performance* **25**, 210 (1999).

42. Brockmole, J. R. & Henderson, J. M. Prioritizing new objects for eye fixation in real-world scenes: Effects of object–scene consistency. *Vis. Cogn.* **16**, 375–390 (2008).

43. Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L. & Dodd, M. D. Examining the influence of task set on eye movements and fixations. *J. vision* **11**, 17–17 (2011).

44. Zhang, M. *et al.* Finding any waldo with zero-shot invariant and efficient visual search. *Nat. communications* **9**, 1–15 (2018).

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

46. Ng, A. Y., Russell, S. J. *et al.* Algorithms for inverse reinforcement learning. In *Icml*, vol. 1, 2 (2000).

47. Abbeel, P. & Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, 1 (2004).

48. Ho, J. & Ermon, S. Generative adversarial imitation learning. In *Advances in neural information processing systems*, 4565–4573 (2016).

49. Schultz, W. Multiple reward signals in the brain. *Nat. reviews neuroscience* **1**, 199–207 (2000).

50. Watanabe, K., Lauwereyns, J. & Hikosaka, O. Neural correlates of rewarded and unrewarded eye movements in the primate caudate nucleus. *J. Neurosci.* **23**, 10052–10057 (2003).

51. Montague, P. R., Hyman, S. E. & Cohen, J. D. Computational roles for dopamine in behavioural control. *Nature* **431**, 760–767 (2004).

52. Akbas, E. & Eckstein, M. P. Object detection through search with a foveated visual system. *PLoS computational biology* **13**, e1005743 (2017).

53. Perry, J. S. & Geisler, W. S. Gaze-contingent real-time simulation of arbitrary visual fields. In *Human vision and electronic imaging VII*, vol. 4662, 57–69 (International Society for Optics and Photonics, 2002).

54. Irwin, D. E. Integrating information across saccadic eye movements. *Curr. Dir. Psychol. Sci.* **5**, 94–100 (1996).

55. Hollingworth, A. & Henderson, J. M. Accurate visual memory for previously attended objects in natural scenes. *J. Exp. Psychol. Hum. Percept. Perform.* **28**, 113 (2002).

56. Parkhurst, D., Law, K. & Niebur, E. Modeling the role of salience in the allocation of overt visual attention. *Vis. research* **42**, 107–123 (2002).

57. Navalpakkam, V. & Itti, L. Modeling the influence of task on attention. *Vis. research* **45**, 205–231 (2005).

58. Wang, Z. & Klein, R. M. Searching for inhibition of return in visual search: A review. *Vis. research* **50**, 220–228 (2010).

59. Zelinsky, G. J. A theory of eye movements during target acquisition. *Psychol. review* **115**, 787 (2008).

60. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A. & Durand, F. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis machine intelligence* **41**, 740–757 (2018).

61. Brockmole, J. R. & Henderson, J. M. Recognition and attention guidance during contextual cueing in real-world scenes: Evidence from eye movements. *Q. journal experimental psychology* **59**, 1177–1187 (2006).

62. Hollingworth, A. Two forms of scene memory guide visual search: Memory for scene context and memory for the binding of target object to scene location. *Vis. Cogn.* **17**, 273–291 (2009).

63. Hout, M. C. & Goldinger, S. D. Target templates: The precision of mental representations affects attentional guidance and decision-making in visual search. *Attention, Perception, & Psychophys.* **77**, 128–149 (2015).

64. Needleman, S. & Wunsch, C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Mol. Biol.* **48**, 443–153 (1970).

65. Anderson, N. C., Anderson, F., Kingstone, A. & Bischof, W. F. A comparison of scanpath comparison methods. *Behav. research methods* **47**, 1377–1392 (2015).

66. Dewhurst, R. *et al.* It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behav. research methods* **44**, 1079–1100 (2012).

67. Zelinsky, G., Peng, Y., Berg, A. & Samaras, D. Modeling guidance and recognition in categorical search: Bridging human and computer object detection. *J. Vis.* **12**, 957–957 (2012).

68. Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A. & Oliva, A. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Vis. cognition* **17**, 945–978 (2009).

69. Gilani, S. O. *et al.* Pet: An eye-tracking dataset for animal-centric pascal object classes. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6 (IEEE, 2015).

70. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. & Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

71. Maxfield, J. T. & Zelinsky, G. J. Searching through the hierarchy: How level of target categorization affects visual search. *Vis. Cogn.* **20**, 1153–1163 (2012).

72. Papadopoulos, D. P., Clarke, A. D., Keller, F. & Ferrari, V. Training object class detectors from eye tracking data. In *European conference on computer vision*, 361–376 (Springer, 2014).

73. Cerf, M., Harel, J., Einhäuser, W. & Koch, C. Predicting human gaze using low-level saliency combined with face detection. In *Advances in neural information processing systems*, 241–248 (2008).

74. Treisman, A. & Gormican, S. Feature analysis in early vision: evidence from search asymmetries. *Psychol. review* **95**, 15 (1988).

75. Neider, M. B. & Zelinsky, G. J. Exploring set size effects in scenes: Identifying the objects of search. *Vis. Cogn.* **16**, 1–10 (2008).

76. Zelinsky, G. J. Tam: Explaining off-object fixations and central fixation tendencies as effects of population averaging during search. *Vis. cognition* **20**, 515–545 (2012).

77. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

Supplementary Materials

SM1: Behavioral Data Collection

Comparable datasets of search behavior

Figure S1 shows how COCO-Search18 compares to other large-scale datasets of search behavior. To our knowledge, there were only three such image datasets that were annotated with human search fixations [17,68,69]. In terms of number of fixations, number of target categories, and number of images, COCO-Search18 is far larger. The PET dataset [69] collected search fixations for six animal target categories in 4,135 images selected from the Pascal VOC 2012 dataset [70], but the search task was non-standard in that participants were asked to "find all the animals" rather than search for a particular target category. This paradigm is therefore search at the super-ordinate categorical level, which is far more weakly guided than basic-level search [71]. Gaze fixations were also recorded for only 2 seconds/image, and multiple targets often appeared in each scene. The microwave-clock search dataset (MCS [17]) is our own work and a predecessor of COCO-Search18. In collecting data for the 18 target categories in COCO-Search18 we had to start somewhere, and our first two categories were microwaves and clocks. Until recently, perhaps the best dataset of search fixations was from[68], but it is relatively small, limited to only the search for people in scenes, and is now a decade old. Note that, whereas there are larger datasets with respect to free-viewing fixations (SALICON [24]) or fixations collected under other visual tasks (POET[72]), these tasks are not visual search and therefore these datasets cannot be used to train models of search behavior. These collective inadequacies demanded the creation of a newer, larger, and higher-quality dataset of search fixations, enabling deep network models to be trained on people's movements of attention as they pursue target-object goals.

Selection of target categories and search images

Here we more fully describe how we selected from COCO's trainval dataset [28] the 18 target categories and the 6,202 images included in COCO-Search18. A goal in implementing our selection criteria was to elicit the behavior that we are trying to measure, namely, the guidance of search by a target category. We also put care into excluding images that might elicit other gaze patterns that would introduce noise with respect to identifying the target-control signal. This sort of attention to detail is uncommon in datasets created for the training of deep network models, where the approach seems to be that more images is always better because this leads to better-trained models. However, in creating a dataset of human behavior, this more-is-better impulse should be tempered with some quality control to be confident that the behavior serving the critical role of labels is of the purported type, which in the current context are search fixations. Because a standard search paradigm collects behavioral responses for both TP and TA images, separate selection criteria were needed. All image selection was based on object labels and/or bounding boxes provided by COCO. On this point, while inspecting the images that were selected we noticed that exemplars in some categories were mislabeled, probably due to poor rater agreement on that category. For instance, several chair exemplars were mislabeled as couches, and vice versa. Rather than attempting to correct these mislabels, which would be altering COCO, we decided to keep them and tolerate a higher-than-normal error rate for the affected categories. This action seemed best, given our plan to discard error trials from the search performance analyses in our study, but researchers interested in interpreting button press errors in COCO-Search18 should be aware of this labeling issue.

Target-present image selection. Six criteria were imposed on the selection of images to be used for target-present search trials (positive samples).

(1) Images were excluded if they depicted people or animals. We did this to avoid the known biases to fixate on these objects when they appear in a scene [21,73]. Such biases would compete with guidance from target-category features, thereby distorting study of the target-bias that is more central to search.

(2) Images were excluded if they depicted multiple instances of the target. A scene showing a classroom with many chairs would therefore be excluded from the "chair" target category because one, and only one, instance of a chair would be allowed in an image.

(3) Images were excluded if the size of the target, measured by the area of its bounding box, was smaller than 1% or larger than 10% of the total image area. This was done to create searches that were not too hard or too easy.

(4) Images were excluded if the target appeared at the image center, based on a 5×5 grid. We did this because the participant's gaze was pre-positioned at this central location at the start of each search trial.

(5) Images were excluded if their width/height ratio fell outside the range of 1.2-2.0 (based on a screen ratio of 1.6). This criterion excluded very elongated images, which we thought might distort normal viewing behavior.

(6) Images, and entire image categories, were excluded if the above criteria left fewer than 100 images per object category. We did this because fewer than 100 images would likely be insufficient for training a deep network model.

Applying these exclusion criteria left 32 object categories from COCO's original 80. Given that this number of images was still far too large images for people to practically annotate with search fixations, we decided to attempt exclusion of images where targets were highly occluded or otherwise difficult to recognize. We did this out of concern that such images would largely introduce noise into the search behavior. To do this, we trained object detectors on cropped views of these

32 categories, and excluded images if the object bounding boxes had a classification confidence < .99. Specifically, for these 32 categories we thresholded the classification confidence of the object in all COCO2014 trainval images, and created a validation set consisting of images meeting this criterion and a training set consisting of the excluded images. The bounding box of the object, for each of the 32 object classes, was then cropped in the image to obtain the positive training samples. Negative samples were same-sized image patches that had 25% intersection with the target (area of intersection divided by area of target), meaning that they were class-specific hard negatives. All cropped patches (over 1 million) were resized to 224×224 pixels while maintaining the aspect ratio using padding. The classifier was a ResNet50 pre-trained on ImageNet, which we fine-tuned by dilating the last fully-connected layer and re-training on 33 outputs (32+"Negative"). Images were excluded if the cropped object patch had a classification score of less than .99. This procedure resulted in 18 categories with at least 100 images in each category, totaling 3,131 TP images.

Two final exclusion criteria were implemented by manual selection. First, for the clock target category we included only images of analog clocks, meaning that we excluded digital clocks from being clock targets. We did this because the features of analog and digital clocks are highly distinct and very different, and we were concerned that this would introduce variability in the search behavior and reduce data quality. Five images depicting only digital clocks were excluded for this reason. Lastly, images from all 18 of the target categories were screened for objectionable content, which we defined as offensive content or content evoking discomfort or disgust. The "toilet" category had the most images (17) excluded for objectionable content, with a total of 25 images excluded across all target categories. After implementing all exclusion criteria discussed in this section, we obtained 3,101 TP images from 18 categories: bottle, bowl, car, chair, analog clock, cup, fork, keyboard, knife, laptop, microwave, computer mouse, oven, potted plant, sink, stop sign, toilet, and tv. See Figure 3 for the specific number of images in each category.

Target-absent image selection. To balance the selection of the 3,101 TP images, we selected an equal number of TA images from COCO. To do this, we kept the criteria excluding images depicting people or animals, extreme width/height image ratios, and images with objectionable content, all as described for the TP image selection, but added two more exclusion criteria that were specific to each of the 18 target-object categories.

(1) Images were excluded if they depicted an instance of the target, a prerequisite for a TA image.

(2) Images were excluded if they depicted less than two instances of the target category's siblings, a criterion introduced to discourage searchers from making TA responses purely on the basis of scene type. For example, a person might be biased to make a TA response if they are searching for a toilet target and the image is a street scene. Because COCO has a hierarchical organization, parent, child, and sibling relationships can be used for image selection. For example, COCO defines the siblings of a microwave to be an oven, toaster, refrigerator, and sink, all under the parent category of appliance. By requiring that the TA scenes for a target category have at least two of that category's siblings, we impose a sort of scene constraint that minimizes target-scene inconsistency and make a scene appropriate to use as a TA image. A scene that has an oven and a refrigerator is very likely to be a kitchen, thereby making it difficult to answer on the basis of scene type alone whether a microwave target is present or absent.

These exclusion criteria still left us with many thousands more TA images than we needed, so we sampled randomly within each of the 18 target categories to match the 3,101 TP images.

Order of target-category presentation

Collecting the search behavior for 6,202 images required dividing each participant's effort into six days of testing. Each testing session was conducted on a different day, lasted about 2 hours, and consisted of about 1000 search trials, evenly divided between TP and TA. Because images from different categories can overlap (e.g., images depicting a microwave may also depict an oven), the presentation order of the target-category blocks was constrained to minimize the repetition of images in consecutive categories and consecutive sessions. For example, because 49 images satisfied the selection criteria for both the sink and microwave target categories, we prevented the microwave and sink categories from appearing in, not only the same session, but the sessions preceding and following. We did this to minimize possible biases resulting from seeing the same scene in different search contexts. A heuristic for maximizing this distance between repeating images resulted in the following fixed target category presentation order across the six sessions:

(1) tv + sink;
(2) fork + chair;
(3) car + bowl + potted plant + mouse;
(4) knife + keyboard + oven + clock;
(5) cup + laptop + toilet; and
(6) bottle + stop sign + microwave.

Each participant viewed from Session 1 to Session 6, or from Session 6 to Session 1, with this order counterbalanced across participants.

Data-collection procedure

Participants were 10 Stony Brook University undergraduate and graduate students, 6 males and 4 females, with ages ranging from 18-30 years. All had normal or corrected to normal vision, by self report, were naive with respect to task design and paradigm when recruited, and were compensated with course credit or money for their participation. Informed consent was obtained from each participant at the beginning of

testing, in accordance with the Institutional Review Board responsible for overseeing human-subjects research at Stony Brook University.

The target category was designated to participants at the start of each block. This was done using the type of display shown in Figure S2 for the potted-plant and analog clock categories. The name of the target category was shown in text at the top, with examples of objects that would, or would not, qualify as exemplars of the named category. In selecting exemplars to illustrate as positive target-category members, we attempted to capture key categorical distinctions at a level immediately subordinate to the target category. When needed, we also gave negative examples by placing a red X through the object. We did this to minimize potential confusions and to enable the participant to better define the target category's boundary.

The procedure (Figure S3) on each trial began with a fixation dot appearing at the center of the screen. To start a trial, the participant would press the "X" button on a game-pad controller while carefully looking at the fixation dot. An image of a scene would then be displayed and the participant's task would be to answer, "yes" or "no", whether an exemplar of the target category appears in the displayed scene by pressing the right or left triggers of the game-pad, respectively. The search scene remained visible until the manual response. Participants were told that there were an equal number of TP and TA trials, and that they should make their responses as fast as possible while maintaining high accuracy. No accuracy or response time feedback was provided.

The presentation of images during the experiment was controlled by Experiment Builder (SR research Ltd., Ottawa, Ontario, Canada). Stimuli were presented on a 22-inch LCD monitor (1680×1050 pixel resolution). Participants viewed stimuli at a distance of 47cm from the monitor, enforced by chin and head rests, resulting in horizontal and vertical visual angles of $54^\circ \times 35^\circ$, respectively. They were asked to keep their gaze on the fixation point at the start of each trial, but were told that they should feel free to move their eyes as they searched. Eye movements were recorded throughout the experiment using an EyeLink 1000 eye-tracker in tower-mount configuration (SR research Ltd., Ottawa, Ontario, Canada). Eye-tracker calibrations occurred before every block or whenever necessary, and these 9-point calibrations were not accepted unless the average calibration error was $\leq .51^\circ$ and the maximal error was $\leq .94^\circ$. The experiment was conducted in a quiet laboratory room under dim lighting conditions.

SM2: Behavioral evaluation of COCO-Search18

Effects of set size and target eccentricity

The visual search literature has done excellent work in identifying many of the factors that increase search difficulty (for reviews, see:[12,18,59,74]). Larger set sizes (number of items in the search display), smaller target size, larger target eccentricity, and greater target-distractor similarity are all known to make search more difficult. However, most of this work was done in the context of simple stimuli, and generalization to realistic images is challenging. For example, what to consider an object in a scene is often unclear, making it difficult to define a set size[75]. Objects in images also do not usually come annotated with labels and bounding boxes. These problems of object segmentation and identification, which largely do not exist for search studies using object arrays, become significant obstacles to research when scaled up to images of scenes.

With COCO-Search18, we can begin to ask how the search for targets in images is affected by set size and target eccentricity. Set size is determined based on the COCO object and stuff labels, which collectively map every pixel in an image to an object or stuff category. Set size is the count of the number of these object labels for a given image. Figure S4 shows the relationship between the number of fixations made on an image, averaged over participants, and the set size of that image, grouped by target category. Some target categories, such as laptop, oven, microwave, and potted-plant, have positive set size effects, indicating a less efficient search with more objects. A similar pattern is shown in Figure S5 for the relationship between the number of fixations on a search image and the initial visual eccentricity of the target (distance between the image center and the target bounding-box center), where for these same objects there was a decrease in search efficiency with increasing target eccentricity. For other target object categories, such as: stop sign, fork, and keyboard, search efficiency was unaffected by either set size or target eccentricity, possibly because these objects are either highly salient (stop sign) or highly constrained by scene context (keyboard).

Distance from fixation to the target

How much closer does each search fixation bring gaze to the target? This is another measure of search efficiency that has been used in the overt attention control literature, and we report this analysis in Figure S6. Plotted is the Euclidean distance between the target and the starting fixation (0) and the gaze landing positions after the first six eye movements (1-6). The most salient pattern is the rapid decrease in fixation-target distance in the first two new fixations, which dovetails perfectly with the steep increase in the cumulative probability of target fixation over these same eye movements reported in Figure 2B. From a starting location near the center of the image, these eye movements brought gaze steadily closer to the target. Note that because this fixation-target distance is averaged over images and participants, the roughly 5 degrees of visual angle at the bottom of these functions should not be misinterpreted as gaze being this distance from the target on a given trial. More interpretable are the overall trends, where a steep drop in distance is followed by a plateau, or even a smaller increase in distance with the 5th and 6th new fixations. This small increase is likely an artifact of these 5 and 6-fixation trials being the most difficult, with more idiosyncratic search behavior, or perhaps even eye movements leaving the target to idiosyncratic destinations while the button press was programmed and executed.

Target-absent search fixations

In the main text we focused on the TP data, where the guidance signal is clearer and the modeling goals are better defined, but we conducted largely parallel analyses of the TA data. Figure S7A shows representative TA images with fixation data from one participant, and Figure S7B shows FDMs from all participants for the same images. Comparing these data with the TP data from Figure 1, it is clear that people made many more fixations in the absence of a target. This was expected from the search literature, but it should also be noted that the FDMs are still much sparser than what would be hypothesized by an exhaustive search (xxx). Paralleling Figure 3, in Figure S8 we also report analyses of the search behavior that are applicable to the TA data. These are again grouped by manual accuracy and response time, and the mean number of fixations made before the target-absent button press terminating a trial. As in Figure **??**, Figure S9 visualizes the consistency among these measures. The columns show ranked performance, with dark red indicating more difficult (or least efficient) search and dark blue indicating relatively easy or efficient search. The rows in Figure S9A group the measures by target category. Similar to the TP data, there was again good consistency among the measures. Also consistent is the fact that bottles and cups were among the most difficult target categories, whereas toilets was the easiest. There was also evidence in the TA data for a speed-accuracy trade-off for some target categories. For example, microwaves and stop signs had relatively low error rates, but these categories were searched with relatively high effort, as measured by ranked response time and number of fixations. Figure S9B visualizes the measures by participant instead of category, where we again found individual differences between participants in search efficiency.

Practice effects

Each of the participants contributing to COCO-Search18 searched more than 6000 images, making it possible to analyze how their search efficiency improved with practice. Figure S10 shows practice effects for both response time (top) and the number of fixations before the button press (bottom), were we define practice effects as performance on the first 1/3 of the trials compared to performance on the last 1/3 of the trials for each target category. Note that practice effects were larger for TA trials (right) than for TP trials (left), and that considerable differences existed across categories. Some categories, such as bottles, showed large practice effects, while other categories, such as analog clocks, showed none at all. We speculate that this difference is due to some categories requiring more exemplars to fully learn compared to others. For example, analog clock was perhaps the most well defined of COCO-Search18's categories, and bottle certainly one of the least well defined, creating greater opportunity to better learn the bottle category over trials.

Search fixation durations

Figures S11 and S12 show density histograms of the search fixation durations for the TP and TA data, respectively, plotted for each of the target categories. x-axis represents the fixation durations with bin size = 50 and y-axis shows the normalized probability density at each fixation. Of note in the TP data is that the mode initial fixation durations (blue lines) were a bit longer than the mode duration of the rest (averaged difference = 63ms), consistent with the very strong guidance observed in the initial eye movements, and they tended to have more bi-modal distritutions. The main peak was at $\sim$250 msec, with a smaller and very short-latency peak at $\sim$50 msec that is likely a truncation artifact of fixation duration being measured relative to the onset of the search display. In contrast, the distribution of second fixations (orange lines) were consistently shorter, even relative to the subsequent fixations. Speculatively, this may be due to a greater proportion of the first new fixations being "off object" [76], which are often followed by short-latency corrective saccades that bring gaze accurately to an object. This interpretation is consistent with the high probability of the target being fixated by the second eye movement (Figure 2B). As for the subsequent fixations, they tended to be short ($\sim$200ms) and not highly variable in their durations. The TA fixations showed similar trends, except for the durations of the first new fixations no longer differing from the rest.

Saccade amplitudes

We also analyzed the distribution of saccade amplitudes during visual search, defined here as the Euclidean distance between consecutive fixations in visual angle. Figure S13 and Figure S14 show the distributions of saccade amplitudes in the TP and TA data, respectively. In the TP data, saccade amplitudes were larger in some categories (toilet and stop sign) than others (bottle and potted plant), likely because easier target categories could be identified from farther in the visual periphery. Some categories also had a bimodal distribution of amplitudes, most clearly shown for clocks, forks, stop signs, and tvs. We speculate that this may be due to the exemplars of these categories sometimes appearing in atypical views, such as a stop sign from the back, and in these cases search is likely less efficient and saccade amplitudes smaller. Saccade amplitudes in the TA data were clearly larger than for TP, and this difference was consistent across target categories. We attribute this to the relatively large viewing angle of the search displays (54 $\times$ 35 degress) creating a greater need for exploration, but this is again speculation. The distributions of saccade amplitudes were also more consistent across categories in the TA data, and there was very little evidence of a bi-modality. Both of these findings likely reflect a generally weaker TA guidance signal.

SM3: Model Methods

Training and testing datasets

Model success depends on the training dataset being an accurate reflection of the test dataset. When the training dataset includes a behavioral annotation, as does COCO-Search18, it

is therefore important to know that similar patterns exist in the training and testing search behavior. The analyses shown in Figure 4A included images from all of COCO-Search18, which recall were randomly split into 70% for training, 10% for validation, and 20% for testing. Figure S15 replots the data from Figure 4A, but divides it into the training/validation (left) and testing (right) sets. Note the high agreement between the test and train/val datasets across this battery of behavioral performance measures.

Inverse Reinforcement Learning

The specific inverse-reinforcement learning (IRL) method that we used was generative adversarial imitation learning (GAIL [48]) with proximal policy optimization (PPO) [77]. The model policy is a generator that aims to create state-action pairs that are similar to human behavior. The reward function (the logarithm of the discriminator output) maps a state-action pair to a numeric value. The generator and discriminator are trained within an adversarial optimization framework to obtain the policy and reward functions. The discriminator's task is to distinguish whether a state-action pair was generated by a human (real) or by the generator (fake), with the generator aiming to fool the discriminator by maximizing the similarity between its state-action pairs and those from people. The reward function and policy that are from the fixation-annotated images images during training are then used to predict new search fixations in the unseen test images.

SM4: Model evaluation and prediction metrics

Metrics for comparing search efficiency and scanpaths

We considered five metrics for comparing search scanpath efficiency (Table 1). Two of these metrics follow directly from Figure 2B and the group target-fixation probability (TFP) function. The first metric computes the area under the TFP curve, a metric we refer to as TFP-auc. Second, and relatedly, we compute the sum of the absolute differences between the human and model target-fixation-probability in a metric that we refer to as Probability Mismatch. These metrics emphasize target-fixation efficiency by penalizing the number of fixations needed to achieve this goal. We also report Scanpath Ratio, which is an older and widely used metric for overt search efficiency. It is the ratio of the Euclidean distance between the initial fixation location and the target to the summed Euclidean distances between the fixation locations in the search scanpath [41,61–63]. An initial saccade that lands directly on the target would therefore result in a scanpath ratio of 1 because the start-target distance and the summed saccade distance would be the same. The scanpath ratio metric emphasizes the distance travelled by gaze in its computation of search efficiency, and not the number of fixations. Finally, we compute two metrics that focus on quantifying the similarity in the search fixation scanpaths, and particularly in comparing search scanpaths between people and the models. The first of these computes a Sequence Score by first converting a scanpath into a string of fixation cluster IDs, and then using a string matching algorithm[64] to measure the similarity between the two strings. Figure S16 shows examples of behavioral and model scanpaths and their sequence scores to help develop an intuition for this metric. Lastly, we use MultiMatch [65,66] to measure the scanpath similarity at the pixel level. MultiMatch measures five aspects of scanpath similarity: shape, direction, length, position, and duration. We exclude the duration metric because the models considered here do not predict fixation duration.

Metrics for comparing search fixation-density maps and models of their prediction

Although scanpath prediction is the higher bar, the prediction of FDMs is also important in that it opens the potential for model comparison across a broader literture. We used three widely accepted metrics for evaluating model predictions against behavioral FDMs on the test images. The AUC metric (Area under the Receiver Operating Characteristic Curve) uses a predicted priority map as a binary classifier to discriminate behavioral fixation locations from non-fixated locations. NSS (Normalized Scanpath Saliency) finds the model predictions at each of the behavioral fixation locations, then averages and normalizes these values. Lastly we computed a Pearson's Correlation Coefficient (CC) between the predicted and behavioral FDMs, although this metric reflects only the degree of linear relationship between predicted and behavioral FDMs.

For both scanpath and FDM prediction, each model generated 10 sequences of fixations (scanpaths), each of maximum length 6 (excluding the first fixation). The 10 scanpaths were obtained by sampling 6 new fixations from the predicted priority map for 10 times. This was done for each testing image. Table S1 reports the purely spatial evaluation of FDM predictions for each model$\times$state combination in our comparison. The results that we report in the main text in the context of scanpath prediction also hold in the case of FDM prediction. Specifically, the IRL-Hi-Low-C model outperformed the others, and did so for every metric. Additionally, the Detector-Hi model performed the second best in all the metrics, supporting our conclusion that a simple detector does a relatively good job in predicting fixations in visual search.

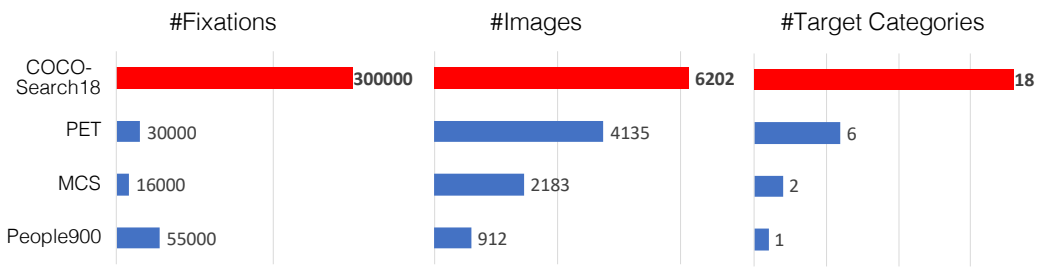


Figure S1. Comparisons between COCO-Search18 and other large-scale datasets of search behavior. COCO-Search18 is the largest in terms of number of fixations (∼300,000), number of target categories (18), and number of images (6,202).

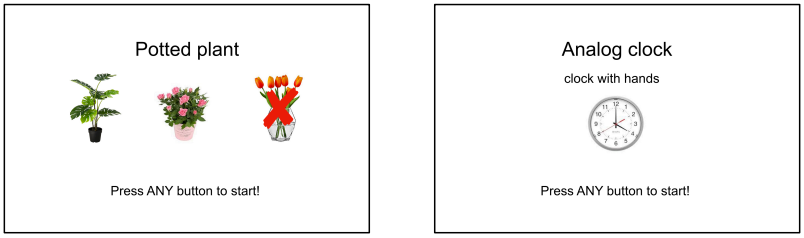


Figure S2. Examples of target-designation displays, shown for the potted-plant and analog clock targets, that preceded the block of trials for a given target category.

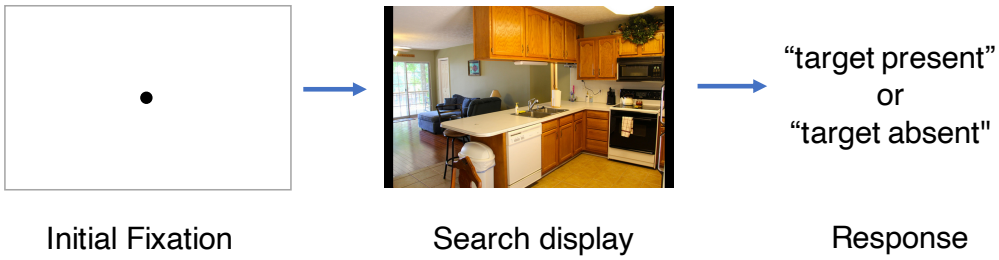


Figure S3. Example of the search procedure. Each trial began with a fixation dot appearing at the center of the screen. Participants would start a trial by pressing a button on a game-pad controller while carefully looking at the fixation dot. An image of a scene would then be displayed and the participant's task was to make a speeded "yes" or "no" target-presence judgment by pressing the right or left triggers of the game-pad, respectively.

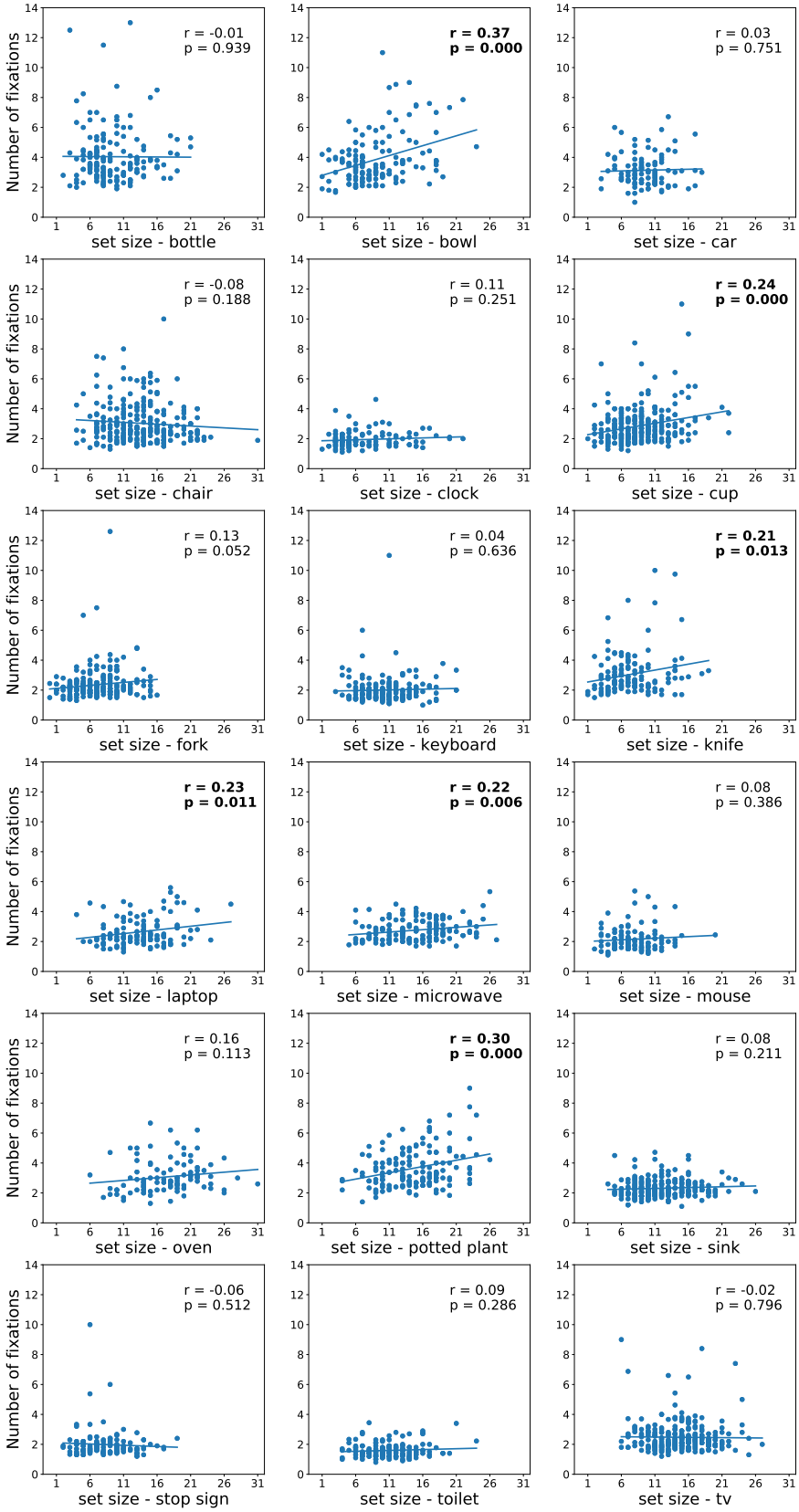


Figure S4. Number of fixations made on an image plotted as a function of the image set size (using COCO object and stuff labels), averaged over participants and grouped by target category.

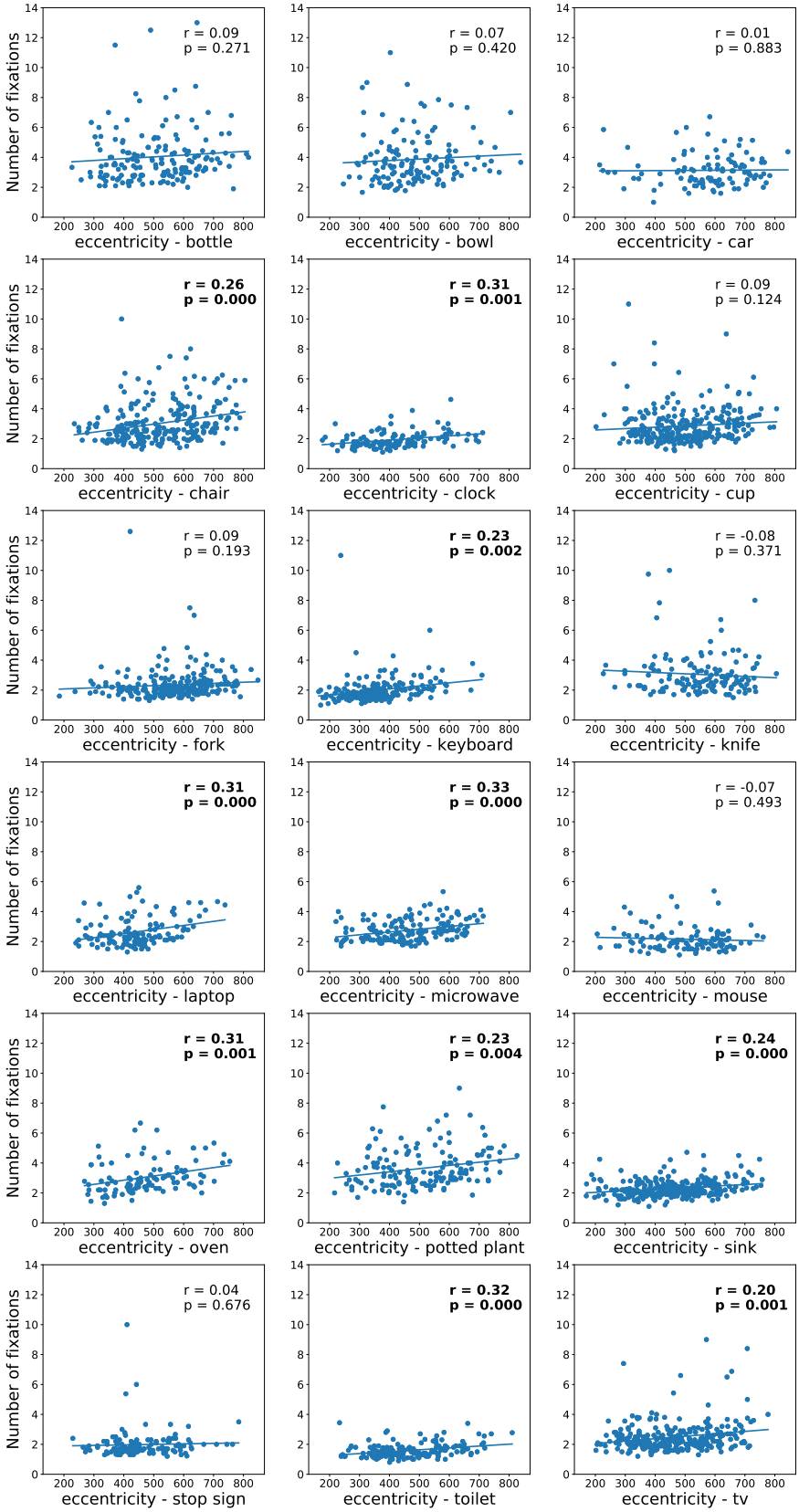


Figure S5. Number of fixations made on an image plotted as a function of initial target eccentricity (using the center of the COCO bounding-box), averaged over participants and grouped by target category.

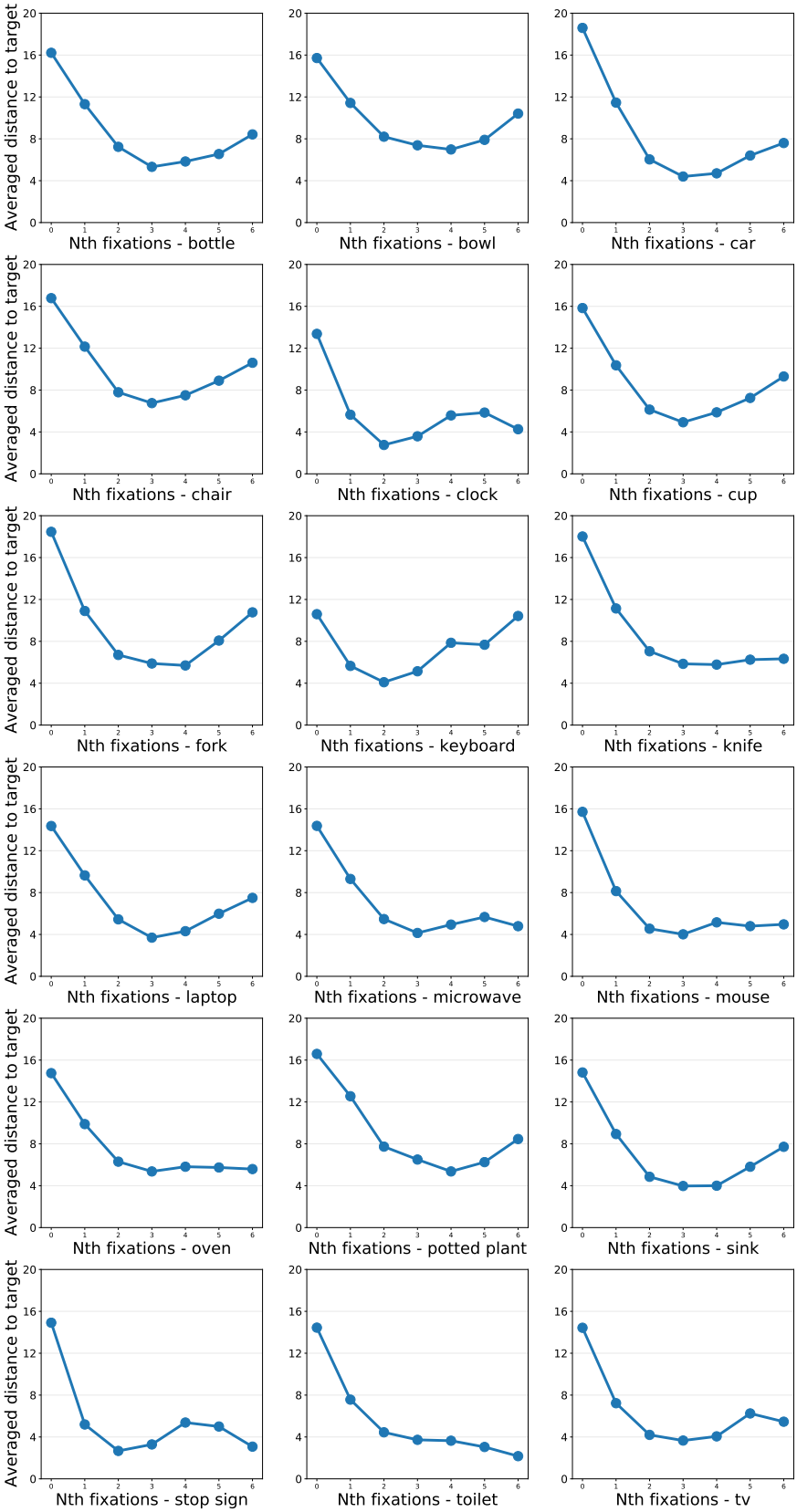
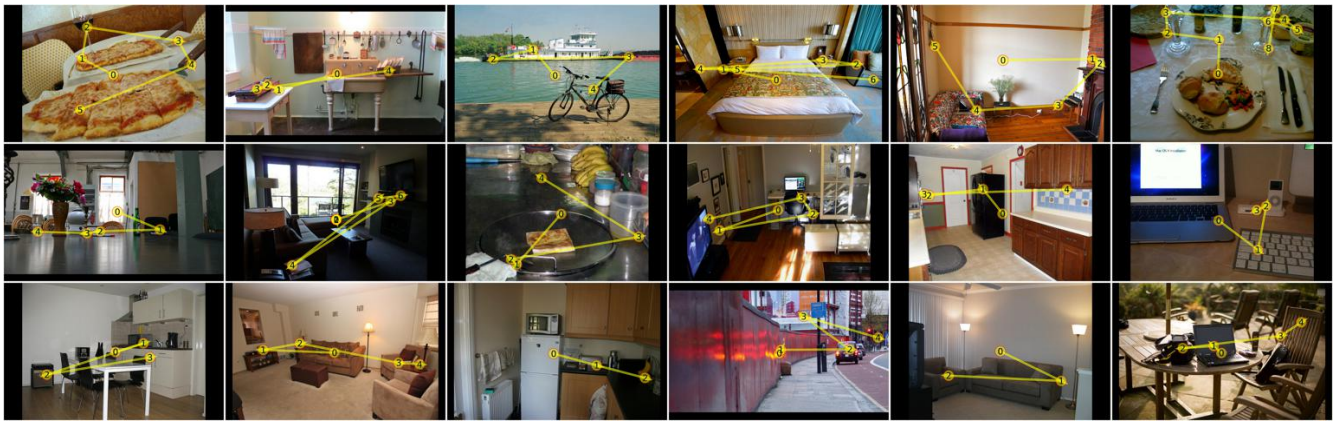


Figure S6. Averaged Euclidean distance (in visual angle) between gaze and the target's center (using COCO bounding-box labels) over the first 6 saccades, grouped by target category.

A



B

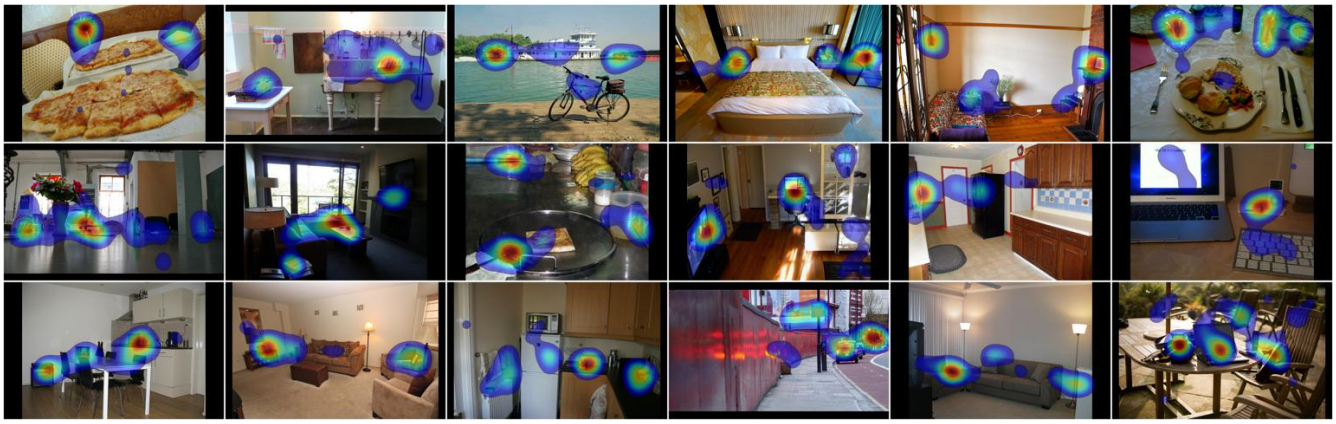


Figure S7. (A). Examples of a target-absent image for each of the 18 target categories. Yellow lines and discs indicate a representative search scanpath from a single participant. From left to right, top to bottom: bottle, bowl, car, chair, analog clock, cup, fork, keyboard, knife, laptop, microwave, mouse, oven, potted plant, sink, stop sign, toilet, tv. (B). Examples of a fixation density map of a target-absent image for each of the 18 target categories. Categories from left to right, top to bottom are the same as in A.

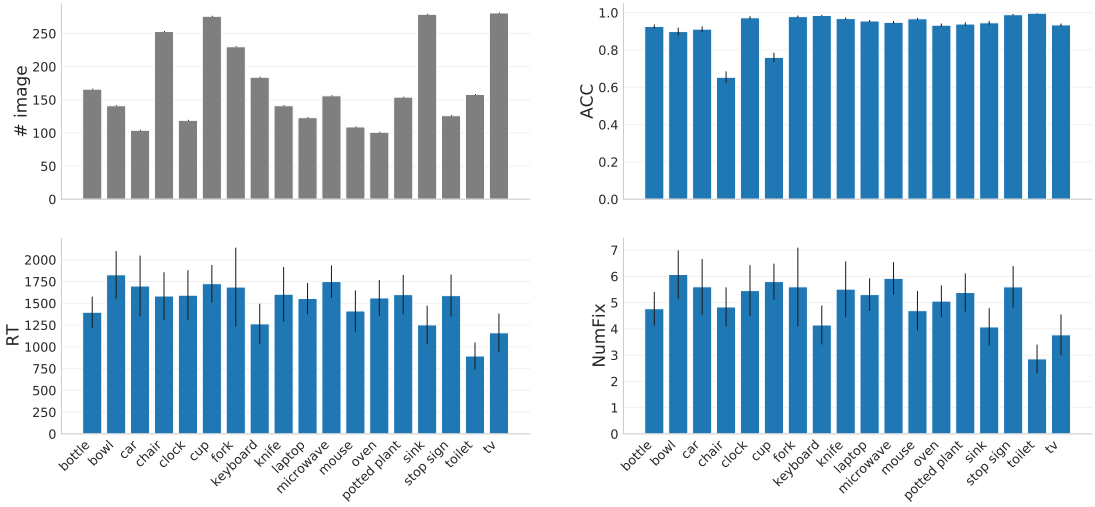


Figure S8. COCO-Search18 analyses for all 18 target categories in target-absent trials. Top: number of images in each category (gray), and response accuracy (ACC). Bottom: reaction time (RT) and number of fixations made before the button press (NumFix). Values are means over 10 participants, and error bars represent standard errors.

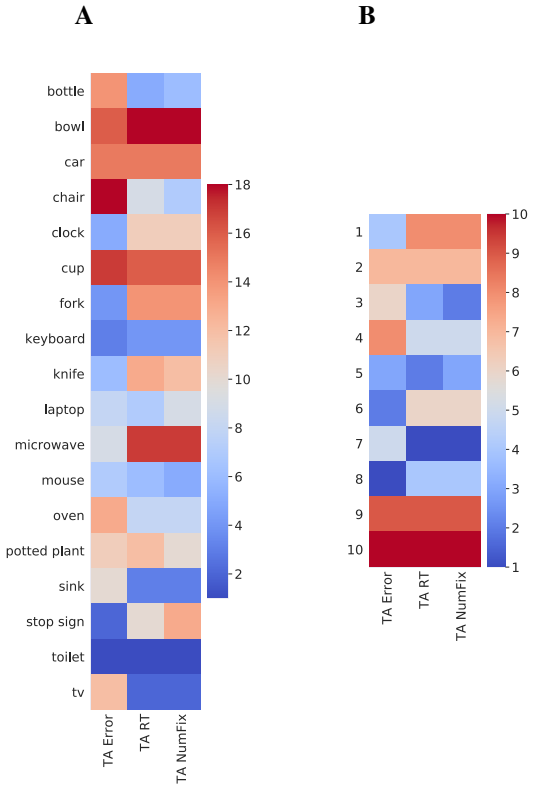


Figure S9. (A). Target-absent data, ranked [1-18] by target category (rows) and averaged over participants, shown for multiple performance measures (columns). These include: response Error, reaction time (RT), and number of fixations (NumFix). Redder color indicates higher rank and harder search targets, bluer color indicates lower rank and easier search. (B) Target-absent data, now ranked by participant [1-10] and averaged over target category (rows). Performance measures and color coding are the same as in (A).

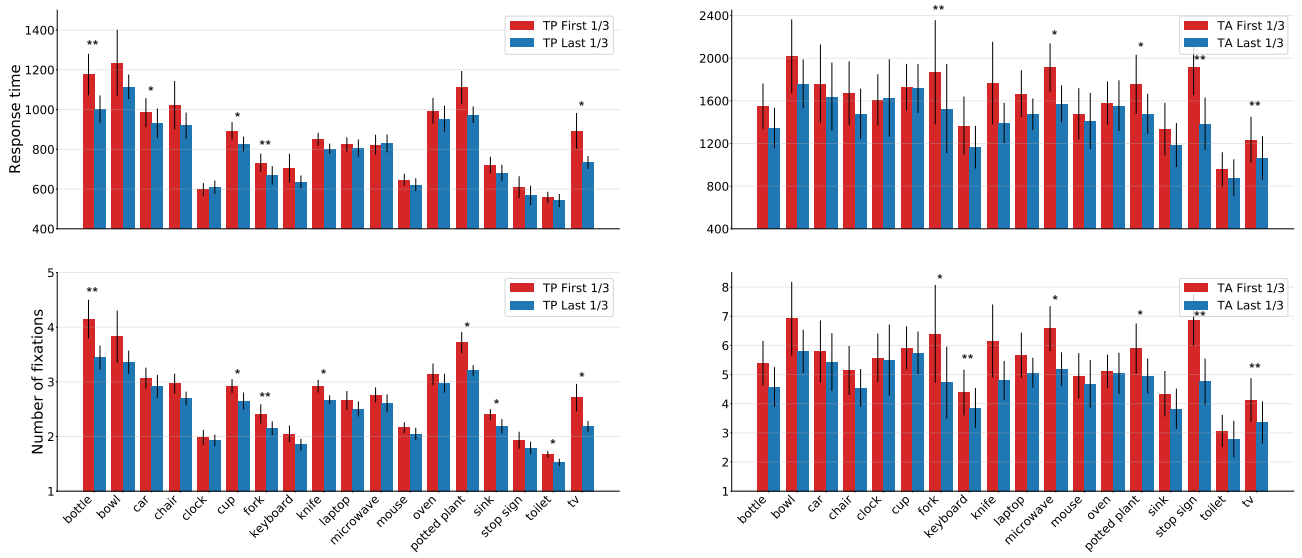


Figure S10. Practice effects, visualized as the difference in search performance between the red (first 1/3 of the trials) and the blue (last 1/3 of the trials) bars, grouped by the 18 target categories. The top row shows response time, and the bottom row shows the number of fixations before the button press. Target-present data are shown on the left, target-absent data are shown on the right. Only correct trials were included. *: $p < .05$, **: $p < .01$

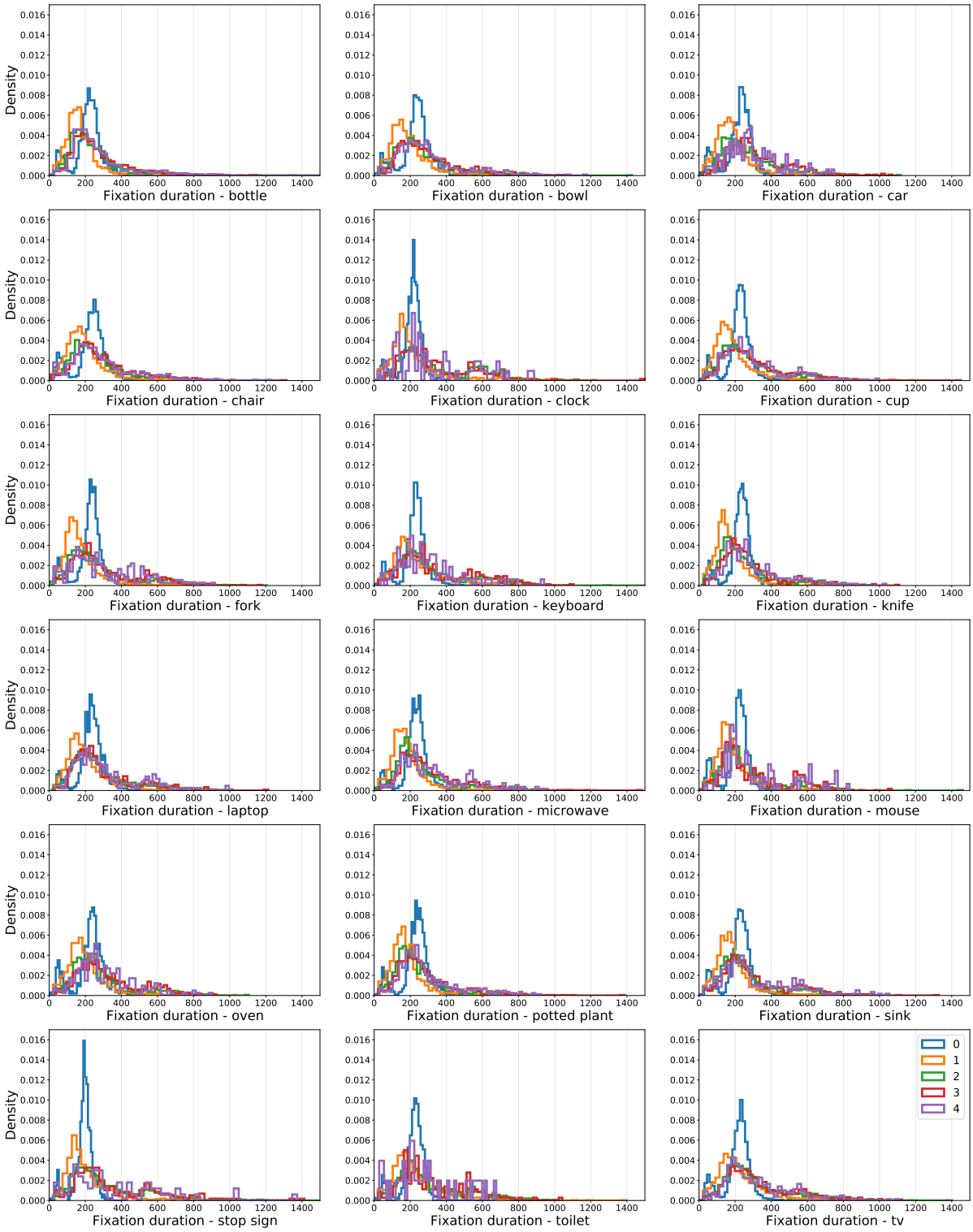


Figure S11. Density distributions of target-present fixation durations, plotted for each of the target categories (bin size = 50). The color lines refer to the initial fixation durations (0, blue), followed by the first four new fixations (1-4).

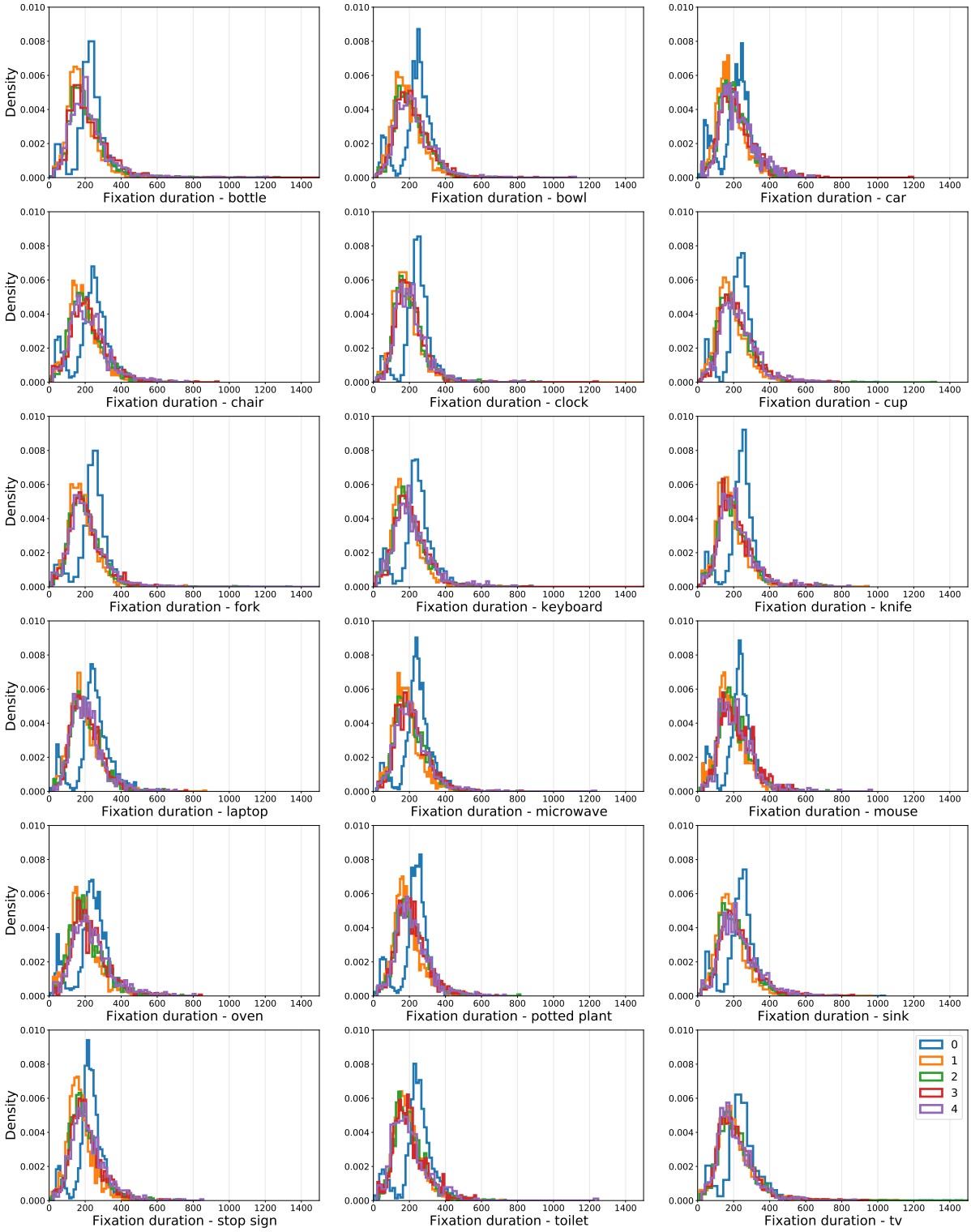


Figure S12. Density distributions of target-absent fixation durations, plotted for each of the target categories (bin size = 50). The color lines refer to the initial fixation durations (0, blue), followed by the first four new fixations (1-4).

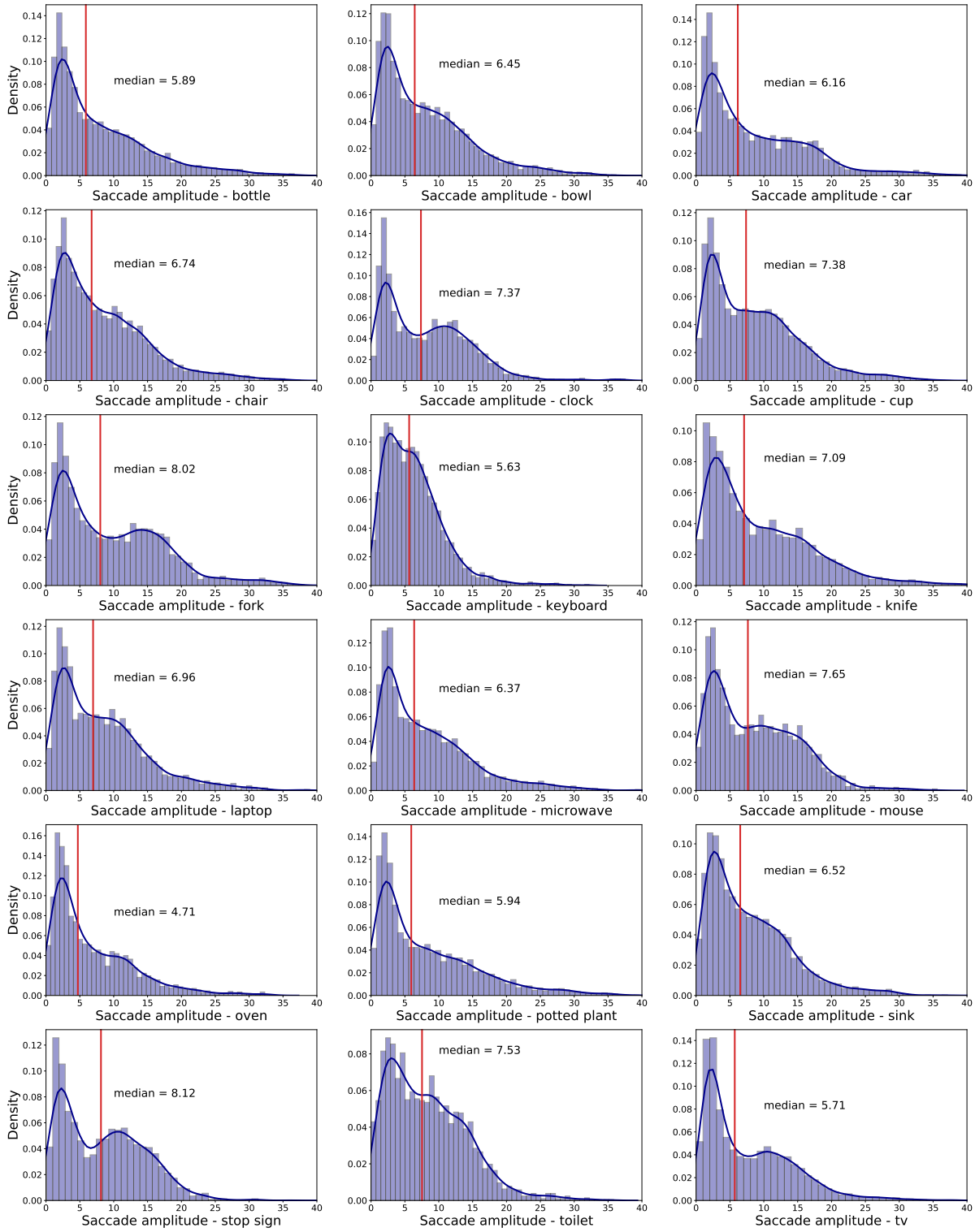


Figure S13. Density distributions of target-present saccade amplitudes, plotted by target category. Red vertical lines indicate median amplitudes. Dark blue lines represent Gaussian kernel density estimates.

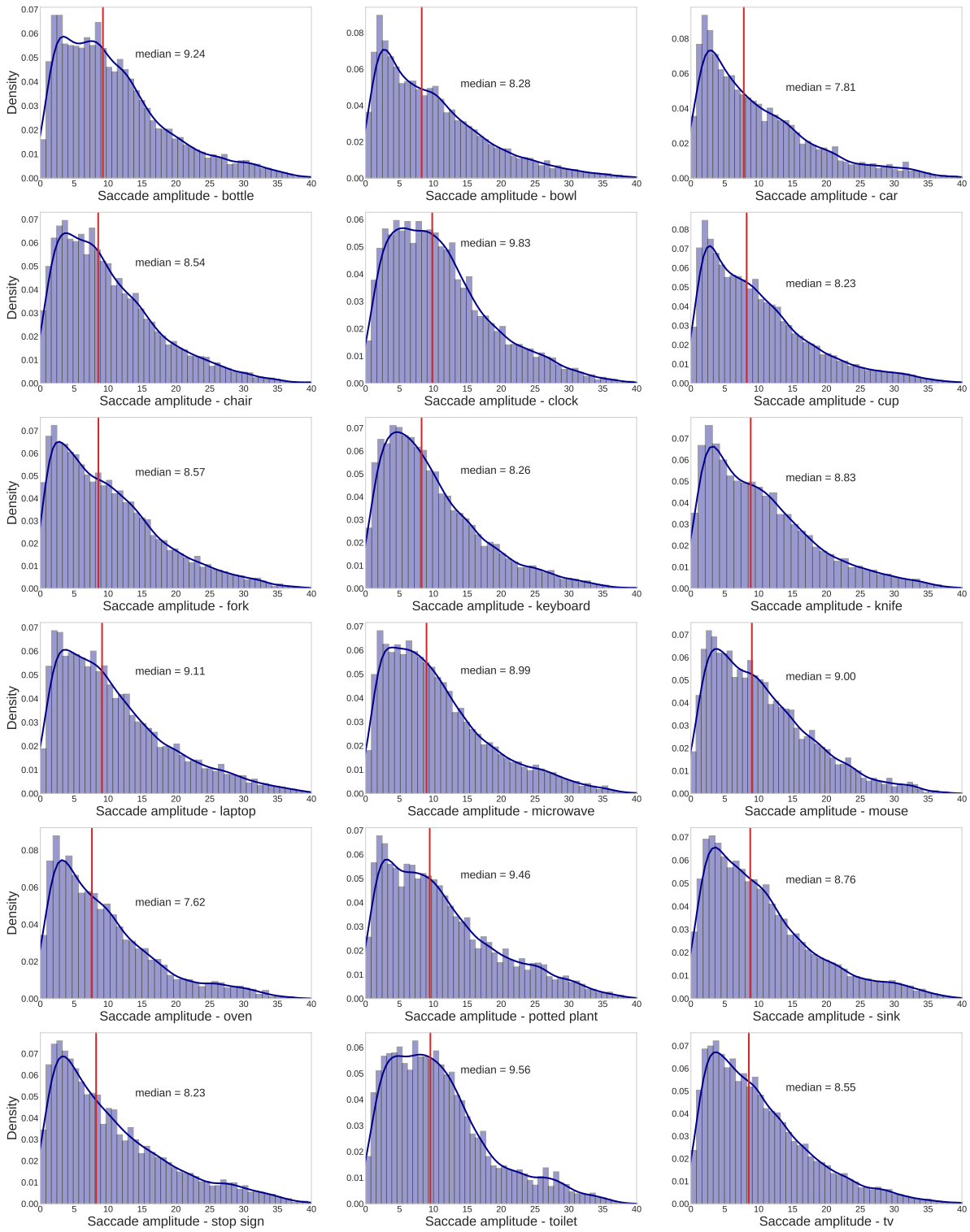


Figure S14. Density distributions of target-absent saccade amplitudes, plotted by target category. Red vertical lines indicate median amplitudes. Dark blue lines represent Gaussian kernel density estimates.

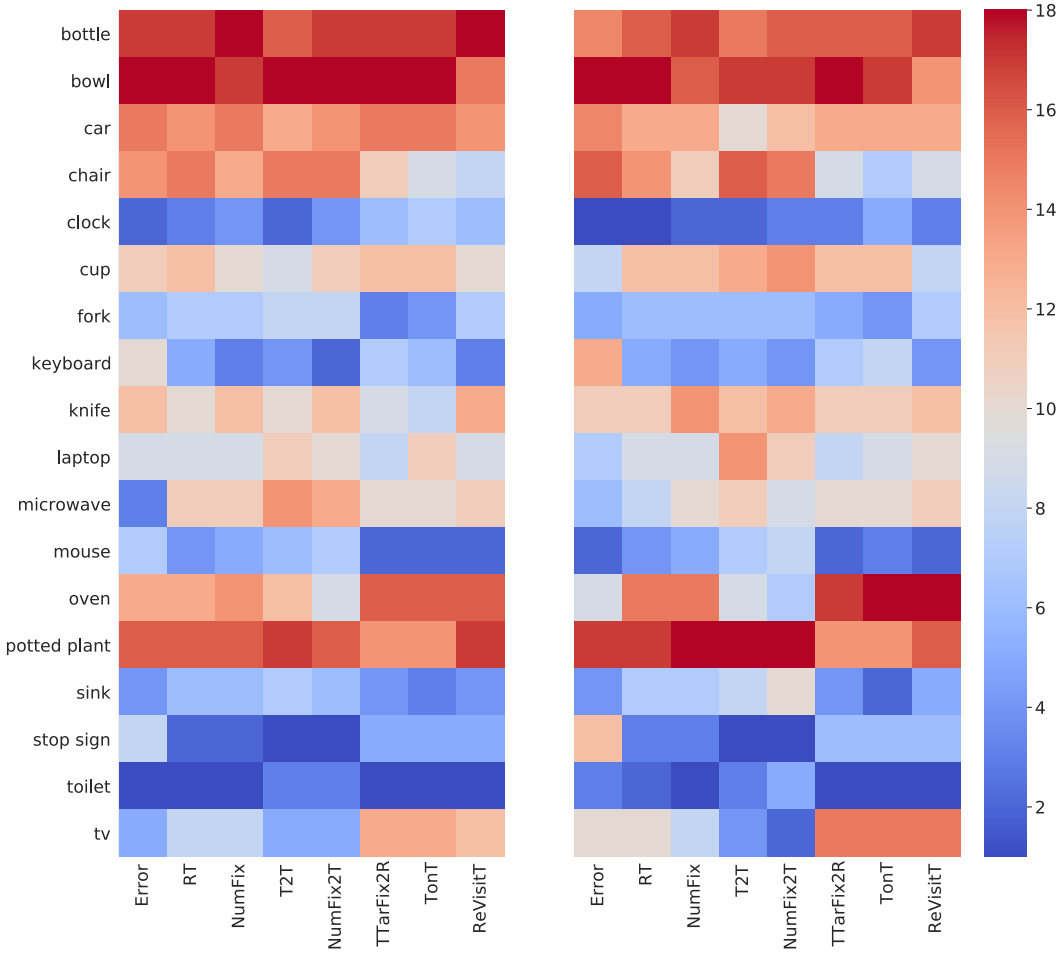


Figure S15. Target-present data, ranked by target category (1-18, rows) and shown for multiple performance measures (columns) in the trainval (left) and test (right) COCO-Search18 datasets. Redder color indicates higher rank and harder search targets, bluer color indicates lower rank and easier search. Measuers include: response error, reaction time (RT), number of fixations (NumFix), time to target (T2T), number of fixations to target (NumFix2T), time from first target fixation until response (TTarFix2R), time spent fixating the target (TonT), and the number of target re-fixations (ReVisitT).

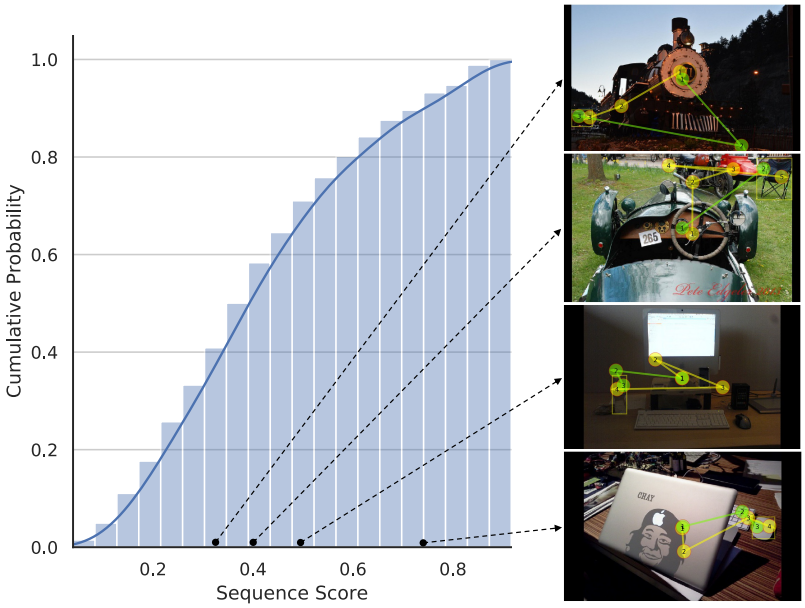


Figure S16. Left: cumulative distribution of sequence scores computed between the IRL model and participant behavior in the test images of COCO-Search18. Right: Examples illustrating the scanpaths producing four different sequence scores. Behavioral scanpaths are colored in yellow, and the IRL-generated scanpaths are in green. Sequence scores for the generated scanpaths are 0.33, 0.40, 0.50, and 0.75, from top to bottom. Note that the predictive results in this illustration are from a slightly different version of the IRL model than the one reported here.

	AUC ↑	NSS ↑	CC ↑
Human	0.675	3.396	0.356
Random	0.536	0.320	0.044
Detector-Hi	0.677	1.383	0.189
Detector-Low	0.614	0.867	0.116
Deep Search-Hi	0.655	1.137	0.156
Deep Search-Low	0.625	0.902	0.125
IRL-ReT-C	0.634	1.123	0.155
IRL-Hi-Low-C	**0.685**	**1.510**	**0.210**
IRL-Hi-Low	0.670	1.350	0.189

Table S1. Results from models (rows) predicting behavioral fixation-density maps using multiple spatial comparison metrics (columns), applied to the COCO-Search18 test images. "Human" refers to an oracle method whereby the FDM from half of the searchers are used to predict the the FDM from the other half of the searchers. See the main and supplemental text for additional details about the spatial fixation comparison metrics.