Look Twice: A Computational Model of Return Fixations across Tasks and Species

Mengmi Zhang[1,2], Will Xiao[3], Olivia Rose[4], Katarina Bendtz[1,2], Margaret Livingstone[3], Carlos Ponce[4], and Gabriel Kreiman[1,2]

Address correspondence to gabriel.kreiman@tch.harvard.edu

[1]Children's Hospital, Harvard Medical School

[2]Center for Brains, Minds and Machines

[3]Department of Neurobiology, Harvard Medical School, Boston, MA, USA

[4]Department of Neuroscience, Washington University School of Medicine, MO, USA

Abstract

Saccadic eye movements allow animals to bring different parts of an image into high-resolution. During free viewing, inhibition of return incentivizes exploration by discouraging previously visited locations. Despite this inhibition, here we show that subjects make frequent return fixations. We systematically studied a total of 44,328 return fixations out of 217,440 fixations across different tasks, in monkeys and humans, and in static images or egocentric videos. The ubiquitous return fixations were consistent across subjects, tended to occur within short offsets, and were characterized by longer duration than non-return fixations. The locations of return fixations corresponded to image areas of higher saliency and higher similarity to the sought target during visual search tasks. We propose a biologically-inspired computational model that capitalizes on a deep convolutional neural network for object recognition to predict a sequence of fixations. Given an input image, the model computes four maps that constrain the location of the next saccade: a saliency map, a target similarity map, a saccade size map, and a memory map. The model exhibits frequent return fixations and approximates the properties of return fixations across tasks and species. The model provides initial steps towards capturing the trade-off between exploitation of informative image locations combined with exploration of novel image locations during scene viewing.

Introduction

Foveal vision is marked by small receptive field sizes compared to the larger receptive fields in the periphery. Primates and other animals move their eyes several times a second through ballistic excursions called saccades, bringing different parts of a scene into high resolution at the center of fixation. Saccades are critical to visual processing and are orchestrated by multiple brain areas involved in determining the location of the next fixation and implementing the corresponding eye movements (*1–6*). Certain locations in an image are more salient in the sense that they draw fixations more frequently than others; for example, subjects rarely make saccades to the middle of a white wall, while a moving yellow car tends to attract saccades.

Models that aim to predict eye movements generally postulate an attention map that specifies how saliency differs across an image (*7–10*). A winner-take-all mechanism selects the maximum of the attention map as the location for the next fixation. Thereafter, some change must occur in the attention map to allow the eyes to explore other locations and prevent the system from repeatedly selecting the same maximum. An inhibition-of-return (IOR) mechanism is typically imposed to ensure that the model can choose the next maximum (*11*). A balance must be struck between an IOR mechanism that is too strong, which would prevent the system from scrutinizing the areas of maximum interest in the attention map, and a weak IOR, which would prevent image exploration. This trade-off can be thought of as a variant of the exploration-exploitation balance.

A finite IOR mechanism would allow subjects to return to previously visited locations. Indeed, behavioral studies have shown that *return fixations* often take place during normal gaze behaviors including reading (*12*), pattern copying or block sorting (*13*, *14*), portrait painting (*15*), solving arithmetic and geometry problems (*16*), visual search (*11*, *17–22*), and free viewing (*18*, *23–25*). Return fixations have been used in neurophysiological studies of target detection (*26*), to study working memory during visual search in change detection tasks (*11*, *21*, *22*, *27*), in object and location recall tasks (*25*), to study the effects of memory load (*19*, *25*, *28–32*), in rejection of distractors (*18*), and the comparison of IOR versus memory-less models (*33*).

The mechanisms driving return fixations remain poorly understood and likely involve multiple factors including image contents, contextual relations among objects, goal-relevance, object familiarity, visual working memory, and eye muscle constraints (*33*, *34*). In this study, we set out to quantify and model the properties of return fixations across a wide variety of naturalistic tasks. We assessed the general principles underlying return fixations, as opposed to only capturing how locations are revisited under a single experimental condition. Therefore, we studied eight experiments that included different species (humans and monkeys), different types of images (isolated object arrays, natural images, Waldo images), different tasks (free viewing and visual search), and different stimulus dynamics (static images and egocentric videos). We studied the frequency of return fixations, their temporal distribution, their spatial distribution and their dependence on image contents and task demands. We show that both monkeys and humans make frequent return fixations with few intervening saccades, especially but not exclusively during visual search; these return fixations show longer durations, and are more prevalent in areas of high saliency and similarity to the target during visual search. To gain insights into the mechanisms that drive return fixations, we propose a biologically-inspired computational model that provides a first-order approximation to the spatio-temporal dynamics of revisiting previous fixation locations by humans and monkeys.

Results

A typical eye movement sequence is shown in **Figure 1A**. The subject fixated at location 3 (red triangle), then made a saccade to location 4 (yellow circle), and quickly returned to location 5 (red circle), which overlaps with location 3. We refer to location 3 as a *to-be-revisited* fixation, location 5 as a *return fixation*, and to all other locations as *non-return fixations*. We used a threshold of one degree of visual angle (dva, approximately the resolution of our eye tracking system, **Methods**) to determine whether two fixations overlapped. Other examples of return fixations are shown in **Figure 1B-D**. We evaluated the pattern of return fixations in eight experiments schematically illustrated in **Figure 2** (see **Methods** for experiment details). These eight experiments encompassed different primate species (humans, **Figure 2A-D,G-H**, and non-human primates, **Figure 2E-F**), different tasks (free viewing, **Figure 2D-G**, and visual search, **Figure 2A-C,H**), and different stimulus presentation formats (static images, **Figure 2A-F**, and free-moving recorded in egocentric videos, **Figure 2G-H**). We characterized the prevalence of return fixations and their properties, and propose a biologically-inspired computational model (**Figure 6**) that captures the main properties of return fixations.

Return fixations are ubiquitous

We started by re-examining the fixation patterns of human subjects during three progressively more challenging visual search tasks in a dataset that we had studied previously (*7*), which included object array images (**Figure 2A**), natural images (**Figure 2B**), and "Where is Waldo?" images (**Figure 2C**). Even though many computational models of visual search assume infinite inhibition of return (IOR), that is, no possibility of returning to a previously visited location, we observed that human subjects made a large number of return fixations in all three cases: $11.9 \pm 0.07\%$ (**Figure 3A**, here and throughout, mean $\pm$ SEM), $18.8 \pm 0.06\%$ (**Figure 3B**), and $15.6 \pm 0.09\%$ (**Figure 3C**), respectively. In all cases, the proportion of return fixations was higher than expected by a null model implementing random eye movements while respecting the distribution of saccade sizes ($p < 10^{-15}$, two-tailed t-test, $t < -16, df = 1,598$, **Methods**). Moreover, subjects returned to the same location not just once, but often multiple times. **Figure S1** shows the proportion of cases where subjects made two return fixations.

We reasoned that return fixations could constitute a useful strategy during visual search, especially in difficult tasks, where it is easy to miss the target and it may therefore be advantageous to revisit previous locations. To assess whether return fixations constitute a unique property of visual search tasks, we conducted a free-viewing experiment where there was no obvious incentive to revisit previous locations (**Figure 2D**). Under free-viewing conditions, subjects still made multiple return fixations (**Figure 3D**, $9.9 \pm 0.08\%$), above the proportion expected by chance ($p < 10^{-15}$, two-tailed t-test, $t = -22, df = 1,098$). The free-viewing experiment in **Figure 2D** used the same images as the visual search experiment in **Figure 2B** and we can therefore directly compare the fraction of return fixations. There were nearly twice as many return fixations during visual search compared to free-viewing conditions.

We analyzed data from two additional free-viewing experiments in macaque monkeys (**Figure 2E-F**). Monkeys also demonstrated extensive return fixations: $20.1 \pm 0.03\%$ and $8.2 \pm 0.01\%$ (**Figure 3E-F**). In all the free-viewing experiments, the proportion of return fixations was higher than expected by chance ($p < 10^{-15}$, two-tailed t-test, $t < -498, df = 198$). The proportion of return fixations made by free-viewing monkeys was comparable or higher than the proportion by humans during visual search. It should be noted that image content, image sizes, and stimulus presentation times differed between

the monkey and human experiments. Therefore, the different proportions of return fixations between humans and monkeys may reflect differences in stimulus selection, rather than a true difference across species.

A large fraction of eye movements studies have focused on behavioral responses to flashed static images. Intrigued by the consistency of return fixations across visual search and free viewing of static images, we asked whether subjects also revisit fixation locations during more naturalistic conditions. To address this question, we extended the analyses to two egocentric video datasets where eye movements were tracked in free-moving subjects during a cooking task (**Figure 2G**, reference(*35*)), or a real-world visual search task (**Figure 2H**, reference(*36*)). In the cooking egocentric video dataset, subjects were asked to follow a sequence of steps on recipes to prepare a meal (**Methods**). In the visual search egocentric video dataset, subjects were asked to navigate an indoor home, search for a list of commonly used items, such as a thumb drive, and put those items on a designated table (**Methods**). To avoid the complexities of head movements during free movement and also to account for fixation locations that may disappear from the field of view, we focused on stable 5-second segments (**Figure S3**, **Methods**). Under these conditions, subjects still made repeated return fixations: $33.4 \pm 0.02\%$ (cooking) and $8.6 \pm 0.02\%$ (visual search task) (**Figure 3G-H**). In both egocentric video datasets, the proportion of return fixations was higher than expected by chance ($p < 10^{-15}$, two-tailed t-test, $t < -79, df = 198$). During the cooking task, subjects manipulated kitchenware, foods, and the recipe in front of them and tended to make a large number of return fixations. In sum, return fixations were ubiquitous across tasks, species, and static images or free-moving conditions.

The total number of fixations varied across tasks and the proportion of return fixations depends on the total number of fixations in a non-linear way. We therefore evaluated the prevalence of return fixations during the first 6 fixations **FigureS2**. We chose to examine the first 6 fixations because this number allowed us to incorporate most of the data, including those experiments that had few fixations per trial. Except for the visual search 3 experiment, the proportion of return fixations was above chance in all the experiments, even when considering exclusively the first 6 fixations in each trial.

In contrast with our initial conjecture, return fixations do not constitute a behavior that is unique to visual search tasks; return fixations are also prevalent during free viewing of scenes. Consistent with our expectations, the proportion of return fixations was higher during visual search than free-viewing when comparing the same experimental conditions. In sum, subjects revisit certain locations within an image, even multiple times, across a large variety of experimental conditions, tasks, and species.

Return fixations are consistent across subjects

We asked whether the locations of return fixations were consistent across subjects in the first six experiments with static images (**Figure S4**). We omitted this analysis in the egocentric video datasets because the field of view in each frame could be different across subjects, making comparisons between subjects difficult to interpret. **Figure S17** shows examples that illustrate consistent return fixation locations across subjects for the same image. For example, seven out of ten subjects made a return fixation to the location at "9 o'clock" in **Figure S17A** and five out of seven subjects made return fixations to the framed picture on the upper left in **Figure S17D**. On the same image, to quantify the degree of consistency across subjects, we divided the image into a grid and calculated the probability of observing return fixations at each location (**Methods**, **Figure S4B**). We summarized these probability distributions of return fixation locations by computing their entropy. An

extreme case of perfect consistency would lead to a probability of 1 at a given location and 0 elsewhere, resulting in minimal entropy. In contrast, a complete lack of consistency would lead to an approximately uniform probability distribution, except for random overlaps, resulting in high entropy. The chance level was computed by considering the same total number of return fixations and distributing them at random locations in the image (**Methods**). The entropy was lower than expected by chance in four of the six experiments, the exceptions being the visual search 3 experiment and the Free Viewing 2 experiment in monkeys, both of which showed the same trend but did not reach statistical significance (**Figure S4E,H**). In sum, in most experiments different subjects tended to revisit the same locations.

Subjects promptly revisit return fixation locations and tend to linger at those locations

We divided all fixation locations into the following three non-overlapping categories: to-be-revisited fixations, non-return fixations, and return fixations (**Figure 4A**). We calculated the offset between to-be-revisited and return fixations. In the example in **Figure 1A**, the return offset is 1 (intervening location 4 between fixation 3 and 5). A strong inhibition-of-return would imply that there should be a large return offset between to-be-revisited and return fixations. In stark contrast, in all the experiments, the distribution of return offsets showed a rapid decay (**Figure 4B**). The percentage of all return fixations with an offset of 1 ranged from 28.3% (humans, Free Viewing) to 80.5% (monkeys, Free Viewing 2) and the percentage of return fixations with a return offset less than or equal to 3 ranged from 47.3% (Visual Search 2) to 100% (monkeys, Free Viewing 2). Regardless of the species, experimental task or stimulus mode, subjects held recently visited locations in memory and tended to move their eyes back to those locations after a very short delay, often the minimum possible delay of one intervening fixation.

After returning their eyes to a given location, subjects tended to fixate longer at the return location (**Figure 4C**). The difference between the return fixation durations and non-return fixation durations ranged from 35.4 ± 5 ms (Human Visual Search 1) to 77.8 ± 5.3 ms (Human Visual Search 2). The duration of return fixations was significantly longer than non-return fixations for all the visual search experiments and two of the free-viewing experiments. Intriguingly, in the egocentric video visual search experiment, fixation durations were longer for the non-return fixations (see **Discussion**).

Return fixation durations were also longer than to-be-revisited fixation durations. The difference between return fixation durations and to-be-revisited fixation durations ranged from 22.2 ± 8.5 ms (Monkeys Free Viewing 1) to 68.8 ± 6.6 ms (Human Visual Search 2). The duration of return fixations was significantly longer than to-be-revisited fixations for all the visual search experiments and the first two free-viewing experiments. There was no consistent relationship between the duration of to-be-revisited fixations and non-return fixations. Therefore, the increased lingering at return locations cannot be simply ascribed to a specific content property of the image (by definition, the image content at to-be-revisited and return locations is very similar).

Saccade sizes preceding return fixations were generally smaller than those preceding non-return fixations. The difference in saccade sizes ranged from 0.7 ± 0.2 dva (Monkey Free Viewing 1) to 4.3 ± 0.5 dva (Human Videos 2) (**Figure 4D**). The visual search on object arrays experiment did not show this effect, perhaps because subjects tended to fixate on the objects, which were isolated with no background, and thus there was only a limited repertoire of possible saccade sizes given the geometry of the display. There was also no difference in saccade sizes in the Monkey Free Viewing 2 experiment.

In sum, return fixations were distinct from non-return fixations and also distinct from to-be-revisited fixations. Subjects tended to remember previously visited locations and revisit them shortly after their first encounter, typically after making a smaller saccade, and generally spending an additional ~ 50 milliseconds the second time around.

Subjects return more often to salient locations and locations more similar to the target during visual search

Beyond the proximity to recent previously explored locations, we asked whether features in the image had an impact on the locations of return fixations. The consistency between subjects described in **Figure S4** suggested that there are special locations in the image that tend to be revisited more often. In addition, the distribution of all return fixation locations was not uniform, further suggesting that there are spatial biases that impact the return fixation locations (**Figure S9**). For example, there was a center bias, especially during free-viewing tasks for both humans and monkeys (**Figure S9D-F**), and the return fixation locations were skewed towards the bottom part of the image in egocentric videos (**Figure S9G-H**). To test whether these location biases are specific to return fixations or general to all fixations, we plotted the spatial biases for both return fixations (**Figure S9A1-H1**) and non-return fixations separately (**Figure S9A2-H2**). We compared the spatial distribution of return versus non-return fixations using the Kullback–Leibler divergence (KLD, which a value of 0 for two identical distributions and a large value when the two distributions are very dissimilar). To avoid the sparsity in the fixation distribution and to account for the resolution of the eye tracker, we quantized the fixation locations with a 2D grid of resolution 1 dva, and computed the KLD between return and non-return fixations. KLD was greater than 38 for all datasets except for Visual Search 1 (where there were only 6 object locations), and Monkey Free Viewing 2. The KLD values suggest that return fixation distribution tended to be more spatially biased than non-return fixations.

Previous studies have shown that fixations tend to gravitate towards locations with high bottom-up saliency, such as regions of high contrast changes (*9*, *10*, *37*). Therefore, we asked whether return fixations were distinct in terms of their bottom-up saliency. We used low-level image features, including edges, contrast, intensity, and color, defined in reference(*38*) to calculate the bottom-up saliency at each location in each image and compared the bottom-up saliency at return fixation locations versus non-return locations. In all experiments except for Visual Search 1, saliency at return fixation locations was higher than at non-return locations (**Figure 5A**).

Although saliency was higher at return fixation locations, the difference in saliency seemed to be too small to fully explain the pattern of return fixations, especially during visual search conditions. In particular, in the visual search experiment 1, return fixations could not be distinguished from non-return fixations in terms of their bottom-up saliency. We hypothesized that the decision making process driving return fixations during visual search might also incorporate top-down information, leading us to investigate how the task demands impacted the locations of return fixations. First, we separately considered the sought target location versus non-target locations. In the three visual search experiments (**Figure 2A-C**), there were more frequent return fixations to target locations than to non-target locations (**Figure 5B**). In the visual search experiments in **Figure 2B-C** (but not in the object array experiment in **Figure 2A**), subjects had to use the computer mouse to click on the target location. Therefore, return fixations to target locations are most likely to imply that subjects fixated on the target but were unaware that they had found it, moved their eyes to other locations, and then returned to the target location, became aware that they had finished the search, and clicked the target location with the mouse.

Even though returning to the target location makes sense in terms of the task goals, subjects also returned to non-target locations. In all three visual search experiments, the proportion of return fixations was higher than expected by chance both for target and non-target locations (**Figure 5B**). In particular, we compared the Visual Search 2 and Human Free Viewing tasks, which used the same images. The proportion of non-target return fixations in the Visual Search 2 task was larger than the proportion of return fixations in the Human Free Viewing task ($p < 10^{-15}$, two-tailed t-test, $t < 70, df = 2,498$). Subjects revisit more locations during visual search compared to free-viewing conditions, even when those locations do not contain the target.

A simple hypothesis of why subjects may return to non-target locations is that those locations may share some degree of visual similarity with the target. To test this hypothesis, we designed an experiment to assess the degree of visual similarity between different fixation locations with the sought target (**Figure 5C**). Subjects were presented with the target image plus two options and were asked to choose the image that was most visually similar to the target (**Methods**). The subjects participating in these two psychophysics experiments on target similarity were different from the ones in the two original visual search tasks. To ensure that subjects performed the task as directed, we included two controls where one of the options was identical to the target (control 1), or one of the options was a different exemplar from the same category (control 2). As expected, subjects chose the control images over non-return fixations (**Figure 5D**). In the test condition, one of the images was a non-target return fixation and the other one was a non-return fixation. Subjects indicated that the return fixations were slightly more similar to the target than non-return fixations $55.8 \pm 1.25\%$ of the time in the visual search experiments 1 ($p = 0.002$, one-sample t-test comparing to chance, $t = 4.45, df = 9$) and $56.3 \pm 1.57\%$ of the time in the Visual Search 2 experiment ($p = 0.003$, one-sample t-test comparing to chance, $t = 3.99, df = 9$). In sum, subjects returned more frequently to salient locations, to locations containing the target, and to locations resembling the target in visual search experiments.

A computational model captures key properties of return fixations

To further understand the mechanisms that give rise to return fixations, we sought to develop an image-computable model capturing the basic observations in **Figures 3-5**. A schematic diagram of the model is shown in **Figure 6**. The starting point was the neurophysiologically-inspired invariant visual search network (IVSN) (*7*). IVSN consists of a "ventral visual cortex" module, implemented by a pre-trained deep convolutional neural network (the VGG-16 network, (*39*)), and a "pre-frontal cortex" module. The visual features from the target image are temporarily stored in pre-frontal cortex and modulate the features of the search image in a top-down fashion, creating a target feature similarity map, M_{sim} (**Figure S5A, Figure S7D**). The IVSN model uses this map to generate a sequence of fixations. The model does not have any mechanism to process motion information or integrate temporal information across video frames; therefore, we focus here on modeling the results of the first six experiments on static images (**Methods**).

Several modifications were introduced into the IVSN architecture. First, to produce a sequence of eye movements during free-viewing conditions, we incorporated the possibility of having uniform top-down modulation and introduced a bottom-up saliency map, M_{sal} (**Figure S5A, Figure S7C**, **Methods**). The saliency map (M_{sal}) depended exclusively on the image contents, while the similarity map (M_{sim}) additionally depended on the target during visual search. Of note, the weight parameters used for extracting visual features for M_{sim} and M_{sal} were neither trained with any of the images used in this

study, nor were they trained to match human performance: all the weights in the VGG-16 architecture were pre-trained using the ImageNet dataset in a visual recognition task (*39*).

Second, we incorporated a constraint on the saccade sizes. The distribution of saccade sizes in humans and monkeys is not uniform (**Figure S10**): occulomotor constraints imply that there are few large saccades and the saccade sizes follow an approximately gamma distribution. Therefore, for each fixation t, we included a saccade prior map, $M_{sac,t}$, computed from the current fixation location and an empirical saccade size distribution for each task (**Figure S6B-G, Figure S7E**). Thus, in contrast to the previous two maps, $M_{sac,t}$ does not depend on the image content.

A third and critical modification is the introduction of a memory decay function for previous fixation locations. Many visual search models, including the initial implementation of IVSN include an infinite inhibition-of-return mechanism that prohibits fixations to revisit previous locations. Instead, we introduced a memory decay map $M_{mem,t}$ that contained information about previously visited locations (**Figure S6A, Figure S7F**). The map $M_{mem,t}$ does not depend on the image contents but rather it is calculated from the eye positions at all the previous fixations $1, ..., t$.

The model linearly combines the four maps, M_{sim}, M_{sal}, $M_{sac,t}$ and $M_{mem,t}$, producing a final attention map, $M_{f,t}$. (**Methods, Figure S7G**). The linear combination involved two scalar weights, w_{sac} and w_{mem}, which control the relative importance of the $M_{sac,t}$ and $M_{mem,t}$ maps. We heuristically searched for adequate values for w_{sac} and w_{mem} based on the Visual Search 2 experiment. These two parameters were then fixed and used to test all the datasets. The location of the next fixation for the model is dictated by the maximum of $M_{f,t}$ (**Figure S7G**). The saccade and memory maps are updated after each fixation (the similarity and saliency maps are fixed), and a new fixation is generated. During visual search, the model decides whether the current fixation contains the target or not by using the ventral visual cortex to extract visual features in the current fixation and comparing those features to the stored target features (**Figure S5B, Methods**). The process is iterated until the target is found in visual search tasks or for a fixed number of steps in the free viewing tasks.

An example sequence of fixations produced by the model in the free-viewing experiment is shown in **Figure 7A2**. The model makes one return fixation denoted by the red triangle, which matches the location of a return fixation made by the monkey in the same experiment. In this example, both the monkey and the model revisited a location within the face (note that the model has no special bias towards faces other than the features extracted from the image by the ventral visual cortex module). Further visualization examples for each of the six datasets are shown in **Figure S8**.

Similar to the results shown for humans and monkeys in **Figure 3**, the model made more return fixations than expected by chance in all six experiments (**Figure 7B, Figure S12**, $p < 10^{-15}$, two-tailed t-test, $t < -75, df = 198$), with a proportion of return fixations ranging from $5 \pm 0.02\%$ in the human Free-Viewing experiment to $25.2 \pm 0.07\%$ in the human Visual Search 2 experiment. The proportion of return fixations was higher in the visual search tasks (**Figure S12A-C**) compared to the free-viewing tasks (**Figure S12D-F**, $p < 10^{-15}$, two-tailed t-test, $t = 49, df = 598$). Consistent with the results shown in **Figure 5B**, even though the model does not explicitly incorporate any target location information, the model tended to produce a higher proportion of return fixations at the target locations than at non-target locations in the Visual Search 2 and 3 experiments (**Figure S13**), but not in the Visual Search 1 experiment.

Consistent with the results in humans and monkeys (**Figure 4B**), most of the return offsets for the model tended to be small and the return offset distribution showed an approximately exponential decay (see example in **Figure 7C** and results

for all experiments in **Figure S14**). The model does not have any notion of fixation duration and therefore we cannot plot the equivalent to **Figure 4C**. In the Visual Search 2 and 3 tasks, the saccade sizes preceding a return fixation for the model tended to be larger than those preceding non-return fixations (**Figure S15A-C**), which is different from the trend observed in humans and monkeys in **Figure 4D**. In the human and monkey free-viewing tasks, the saccade sizes preceding return fixations for the model tended to be smaller (**Figure S15D-F**), consistent with the results in humans and monkeys. Image properties also influenced the probability of making a return fixation to a particular location in the model, as shown by the increased bottom-up saliency at return fixations compared to non-return fixations (**Figure S16**). In sum, without any task-specific training, the model approximates most of the basic properties of return fixations in humans and monkeys.

Discussion

We examined 44,328 return fixations out of a total of 217,440 fixations (20.4%) in 8 experiments monitoring eye movements (**Figure 2**). Return fixations are ubiquitous across visual search tasks of different complexity levels, during free-viewing conditions, in humans and monkeys, and also under naturalistic free-moving conditions (**Figure 3**). Return fixations tend to occur shortly after the first exploration of a given location, in many cases with the minimum possible offset of one intervening fixation (**Figure 4B**). Return fixations last $\sim$ 50 ms longer (**Figure 4C**) and often follow small saccades (**Figure 4D**). The locations of return fixations are consistent across subjects (**Figure S4**) and tend to gravitate towards regions of higher bottom-up saliency (**Figure 5A**) and also towards regions that resemble the target during visual search tasks (**Figure 5B-D**).

We developed an image-computable model that mimics the basic properties of return fixations (**Figure 6**, **Figure 7**). The proposed model has four main components: (i) a saliency map, (ii) a target similarity map, (iii) a constraint on saccade sizes, and (iv) a finite inhibition-of-return with an approximately exponential memory decay function (**Figure 6B**). The first two components depend exclusively on the image contents. Salient locations include changes in color, orientation, and texture, among other properties. There is extensive literature documenting the role of salient locations in attracting eye movements (*40*, *41*). The target similarity map, which is only relevant during visual search tasks, makes image locations that resemble the target especially attractive for fixations (*7*).

The third model component is based on the observed distribution of saccade sizes (**Figure S10**), which is far from uniform and leads to a low probability of making large saccades. The avoidance of large saccades is likely to be imposed by the eye movement musculature itself (*42*, *43*). Additionally, saccadic suppression might also be more difficult to implement with large saccades. This constraint also makes it more likely to revisit recent locations (**Figure 4B**), but previous work has shown that the saccade size distribution is not sufficient to account for the frequency of return fixations (*33*).

The fourth model component is memory decay, which approximates the finite inhibition-of-return (IOR). The strength of IOR plays a central role in balancing exploration (stronger inhibition of prior fixations enhances foraging of novel image locations) versus exploitation (weaker memory for inhibiting previous locations facilitates return fixations). The proportion of return fixations is lower than what would be expected by a memoryless system (*33*), which can be explained by a finite IOR. As an initial approximation, the computational model assumes that the IOR function is fixed and independent of the task, species, or experimental conditions. Differences in the return fixation properties across tasks are thus accounted for in the model by the integration of the four components.

Task demands can play an important role in determining the frequency and duration of return fixations. For example, during the cooking task, say that the subject is cutting carrots, the eyes are constantly drawn to the knife and carrot, increasing the number of return fixations. Under these conditions, there is no need to process extensive information at each fixation and the fixation durations are shorter. In contrast, during the Waldo search task, there is a stronger incentive for exploring other locations, thus reducing the frequency of return fixations, yet each return fixation lasts longer, as the large amount of clutter makes the target recognition decision harder.

It is interesting to speculate that return fixations may be especially linked to an imperfect visual recognition machinery. In an extreme case where the visual recognition machinery achieves perfect performance in interpreting the contents at the fovea in every fixation, there would be little incentive to revisit locations to gain further insights. Multiple studies have praised the virtues of fast recognition in approximately 150 ms after flashing a stimulus (*44–46*). However, many of those studies have focused either on isolated objects or large objects with minimal clutter. Under more natural conditions and especially for smaller objects embedded in clutter, subjects make many recognition mistakes (*47*, *48*). Indeed, a strong example of recognition errors is the case of return fixations to the target during visual search (**Figure 5B**, see also discussion in (*7*)). Consistent with the link between return fixations and imperfect recognition during a single fixation, several studies have argued that return fixations allow re-inspection of incomplete or dynamic regions in scenes (*33*), recovery of lost information (*29–32*), and rehearsals of visual working memory (*25*, *28*).

The proposed model is deliberately founded on using pre-trained neural networks and has very few free parameters. The weights to extract image features in the ventral visual cortex module are trained on a visual recognition task using the ImageNet dataset and are *not* fine-tuned for any of the tasks in the current study. The saccade size constrain, the IOR memory decay function and the relative weight of those two components are derived from experimental data. Specifically, they were tuned using the Visual Search 2 experiment. All of those parameters were fixed thereafter, and the results for all the other datasets shown throughout the text do not use any type of data fitting. The model does not always quantitatively match the observations in humans and monkeys (e.g., compare **Figure 3** versus **Figure S12**, **Figure 4** versus **Figure S14** and **Figure S15**, **Figure 5** versus **Figure S16**). While fine-tuning the model for each dataset would improve the fitting, the purpose of the model was to provide a conceptual proof-of-principle demonstration of the key ingredients underlying return fixations rather than an attempt at fine-tuning multiple parameters to fit the experiments.

Most eye movement studies have focused on flashing two-dimensional images on a computer screen. This paradigm as a surrogate for natural vision has been criticized for lacking depth information, natural spatiotemporal statistics, and natural head and body movements. Egocentric videos provide a more naturalistic venue to study first-person viewing behaviors where subjects interact with physical objects while freely moving their eyes, heads and bodies. Despite the notable experimental differences to flashing images, in terms of return fixations, subjects revisit previously fixated locations during naturalistic behaviors like visual search and cooking, in a similar fashion to the behavior observed in static images. Task demands and natural behaviors may impose additional constraints. Several studies have shown that the egocentric gaze in a natural environment requires the combination of gaze direction (the line of sight in a head-centered coordinate system), head orientation, and body pose (*49–51*). For example, during the cooking task, the gaze point tends to fall on the object that is currently being manipulated. This constraint results in a higher proportion of return fixations compared with the other tasks.

The amount of time devoted to visual processing during a saccade has been used as a proxy for the computational demands of given tasks(*52*, *53*). Consistent with this notion, the average fixation duration during free viewing (259.7 ± 0.9 ms) was shorter than during visual search (297.9 ± 0.9 ms). Interestingly, return fixations showed longer durations (**Figure 4C**). A possible interpretation of this observation is that the brain tags these locations as return fixations, perhaps acknowledging the difficulties in visual recognition during the first pass, and devotes additional computational time to improving recognition the second time around.

Even though the model captures essential properties of return fixations, the model does not behave exactly the way humans and monkeys do. First, the model shows constant acuity over the entire visual field, which is clearly not the case for primate vision where acuity drops rapidly from the fovea to the periphery. Second, humans and monkeys have a better recognition system to decide whether the target is present or not at the current fixation (**Figure S11**). Third, humans and monkeys may capitalize on contextual information integrated over multiple saccades to decide whether a return fixation is warranted or not (*48*). For example, the object relations in the environment might attract primates to check back for relevant objects. Fourth, there is no learning in the current model, but humans and monkeys can adapt and change their exploration and exploitation strategies in a task-dependent manner. Despite these limitations, the model suggests a biologically plausible mechanism to describe the balance of exploration and exploitation during visual scene understanding across species, tasks and viewing conditions.

Methods

Datasets

We evaluated return fixations on eight datasets (**Figure 2**), individually described below.

Visual search on static images

We evaluated eye movements during three visual search tasks with increasing level of difficulty, reported in reference (*7*): object arrays (Visual Search 1, **Figure 2A**), natural images (Visual Search 2, **Figure 2B**), and Waldo images (Visual Search 3, **Figure 2C**). Forty-five naive observers (19–37 years old, 15 subjects per experiment) participated in these tasks. Subjects had to fixate on a cross shown in the middle of the screen for 500 ms, a target object was presented followed by another fixation delay (object arrays and natural images), a search image was presented, and subjects had to move their eyes to find the target. In the natural images (Visual Search 2) and Waldo images (Visual Search 3), subjects had to indicate the target location via a mouse click. If the clicked location fell within the target, subjects went on to the next trial; otherwise, subjects stayed on the same search image until the target was found. If the subjects could not find the target within 20 seconds, the trial was aborted and the next trial was presented. For further details about the images and the tasks, see reference (*7*).

Human free viewing

To compare human eye movements during visual search versus free viewing, we conducted an experiment with 10 subjects (18-37 years old, 5 female). We used the same 240 natural images from the Visual Search 2 task, described above. Subjects

had to first fixate on the center cross for 500 ms and then freely move their eyes to explore the image for 4500 ms (**Figure 2D**). The stimulus presentation duration of 4500 ms was chosen because subjects were able to find the target in 90% of the trials within this time during the visual search task (Visual Search 2) (*7*), therefore providing us with an approximately comparable number of fixations. Subjects were instructed to look at the images, without any other task demands. The stimuli were presented in grayscale on a 19-inch CRT monitor (Sony Multiscan G520) occupying full screen (1024 $\times$ 1280 pixels, subtending 25 $\times$ 30 degrees of visual angle (dva)). Observers were seated at a viewing distance of approximately 66.4 cm. The participants' eye movements were recorded using the EyeLink 1000 plus system (SR Research, Canada). All participants had normal or corrected-to-normal vision. Participants were compensated for participation in the experiments. All the human psychophysics experiments were conducted with the subjects' informed consent and according to the protocols approved by the Institutional Review Board at Children's Hospital.

Monkey free viewing

To compare eye movements under free viewing conditions across species, we analyzed eyetracking data from free-viewing monkeys (**Figure 2E-F**). These two datasets were initially collected for other experiments with the goal of studying neuronal responses during free viewing; the neuronal responses are not discussed here. These two experiments were not designed to specifically match the conditions in the previous experiments. The purpose of introducing these experiments in the current study is not to quantitatively assess whether humans make more or less return fixations than monkeys, which would require matching the experimental conditions. Rather, the purpose here is to qualitatively assess the properties of return fixations in different species and to evaluate whether the model can capture those properties. Six monkeys (5-13 years old, all male) from one lab were tested in the Monkey Free Viewing 1 experiment. Two monkeys (both 7 years old males) from a second lab were tested in the monkey Free Viewing 2 experiment. Procedures were approved by the Harvard Medical School Institutional Animal Care and Use Committee (Monkey Free Viewing 1), and the Washington University School of Medicine Institutional Animal Care and Use Committee (Monkey Free Viewing 2), and conformed to NIH guidelines provided in the Guide for the Care and Use of Laboratory Animals.

In the Monkey Free Viewing 1 experiment (**Figure 2E**), each trial of the experiment did not start with a center cross. Instead, the initial fixation of a trial could start anywhere on the screen. The presentation duration of each trial vary from 1,000 ms to 2,000 ms. There were 121 images in total with repeated presentations in random order. In our analysis, we focused on fixation sequences longer than 1,500 ms during *only* first stimulus presentations. We discarded other fixation sequences in repeated presentations due to concerns about the impact of memory across trials on return fixations. The trial sequence intermixed both natural images and images containing salient visual features, such as other monkeys or body parts. There were 36 images out of 121 containing faces. To ameliorate the center fixation bias in each trial, the image location was randomly jittered relative to the original image in a [-3:1:3] dva grid. For example, in **Figure 2E**, the stimulus was shifted to the left; with the vacant space shown as grey background. Monkeys were seated at a viewing distance of approximately 57-58 cm away. All images were presented in color on a monitor screen (635 $\times$ 635 pixels, subtending 16 $\times$ 16 dva).

In the Monkey Free Viewing 2 experiment (**Figure 2F**), monkeys were trained to first look at the center fixation for 500 ms,

followed by the stimulus presentation for 1000 - 1500 ms. There were 1,761 images in total with 1,380 natural images from the MSCOCO dataset (*54*), around 240 images that contain either monkey faces or their body parts, and around 140 pictures of local laboratory staff and animal shelters. As in the Monkey Free Viewing 1 experiment, to eliminate the center bias, all the images were shifted randomly in a 2-degree radius circle. Monkeys were seated at a viewing distance of approximately 57-58 cm away. All images are presented in color on a monitor screen (596 $\times$ 596 pixels, subtending 15 $\times$ 15 dva).

Human egocentric videos

While the majority of eye movement studies have focused on static images, here we also considered a more naturalistic setting where subjects could move freely and interact with physical objects while eye positions and first person videos were recorded (Egocentric videos, **Figure 2G-H**). We used two existing egocentric video datasets (*35*, *36*). In both cases, subjects wore an SMI mobile eyetracker (iMotions, Denmark) with a sampling rate of 30 Hz and precision of $\approx$ 0.5 dva.

The egocentric video dataset 1 (Humans, Videos 1) consisted of eye positions in 86 videos showing the field of view captured from a fist-person perspective (*35*). In these videos, 32 subjects performed cooking activities. Each video clip lasted 20 minutes and was captured at 24 frames per second and 960 $\times$ 1280 pixels (corresponding to $\approx$ 46 $\times$ 60 dva). In the beginning of each cooking task, subjects were instructed to follow the steps on a recipe. There were 10 recipes, such as northern american breakfast, pizza, and turkey sandwich. Each recipe entailed a sequence of meal preparation steps. **Figure S3B** shows example video frames and their corresponding fixations overlaid on the last frame of each video clip (yellow circles).

The egocentric video dataset 2 (Humans, Videos 2) consisted of eye positions in 57 videos showing the field of view captured from a first-person perspective (*36*). In these videos, 55 subjects performed a visual search task . Each video clip lasted around 15 minutes and was captured at 24 frames per second and a resolution of 960 $\times$ 1280 pixels (corresponding to $\approx$ 46 $\times$ 60 dva). The experiment site was a fully furnished and functional model home including a master bedroom, children's room, living room, open kitchen, dining area, study room, recreational room, bathroom and exercise area. Each subject was asked to search for a list of 22 items commonly used in daily life (including thumb drive, shampoo, etc.) and move them to the designated packing location (dining table). **Figure S3C** shows example video frames and their corresponding fixations overlaid on the last frame of each video clip (yellow circles).

Psychophysics experiments on target feature similarity for return fixations

In **Figure 5C-D**, we asked whether the return fixation locations were visually similar to the sought target during the visual search tasks. To answer this question, we conducted two psychophysics experiments on Amazon Mechanical Turk (Mturk). We recruited 20 subjects (10 subjects for the object arrays task and 10 subjects for the natural images task). Subjects were compensated for participation. Participants provided informed consent in electronic format. We followed all the protocols approved by the Institutional Review Board at Children's Hospital.

Subjects were presented with a target object and two alternative images. Subjects performed a two-alternative forced choice task indicating which of the two options was more similar to the target object. The images remained on the screen until the subjects made a choice. The two image options were randomly mapped onto choice A or B. The images were fixation

patches obtained by cropping the search image. In the main test condition, one of the options always corresponded to a return fixation patch and the other option corresponded to a non-return fixation patch, where return and non-return were defined based on the eye movement data independently obtained from different subjects in reference (*7*). For each trial, the return and non-return fixation image patches were extracted from the same image and trial. In the object array experiment (**Figure 2A**), the patch encompassed the entire object (subtending about 3.6 dva). In the natural images experiment (**Figure 2B**), the patch encompassed a square box of size 156 $\times$156 pixels, subtending 3.6 dva and centered at each fixation. We collected 412 pairs on object arrays and 1,041 pairs on natural images.

As a sanity check to evaluate the quality of the online results from Mturk, we introduced two control conditions that were randomly intermixed with the test trials. In control 1 (3% of the trials), the two options were a non-return fixation patch versus the actual identical target. We (obviously) expected subjects to indicate that the identical target was more similar to the target than the non-return fixation locations. In control 2 (3 % of the trials), the two options were a non-return fixation patch versus an object belonging to the same semantic category as the target object but showing a different exemplar, different rotation and scaling. We expected subjects to indicate that the exemplar from the same category was more similar to the target than the non-return fixation locations. We set an exclusion criteria for subjects that made more than 3 errors in the control trials, but all 20 subjects satisfied these two controls and none were excluded from the analyses.

Computational model to predict return fixations

We first provide a high-level intuitive outline of the proposed computational model, followed by a full description of the implementation details (**Figure 6**). The model is based on the previously published architecture for invariant visual search (IVSN(*7*)). The current model incorporates many modifications, most notably the prediction of eye movements during free viewing conditions when there is no target to search for, the incorporation of multiple maps discussed below, and finite inhibition-of-return.

The output of the model is a sequence of fixations. During visual search tasks, there are two inputs to the model: the target image (I_T) and the search image (I_S). During free viewing tasks, there is only a single stimulus input (I_S). The model posits an attention map $M_{f,t}$ at each fixation time t by integrating four components: a bottom-up saliency map M_{sal}, a target visual feature similarity map M_{sim}, a saccade prior map $M_{sac,t}$ dependent on the previous fixation location, and a visual working memory map $M_{mem,t}$ dependent on the location of previous fixations (**Figure 6**).

In the visual search tasks, both the target image (I_T) and the search image (I_S) are processed through the same deep convolutional neural network, which aims to mimic the transformation of pixel-like inputs through the ventral visual cortex (*55–57*). The target feature similarity map (M_{sim}) indicates the similarity between the target image and each location of the search image. M_{sim} is computed using the same procedure described previously (*7*). Briefly, feature information from the top level of the visual hierarchy provides top-down modulation, based on the target high-level features, on the activation responses to the search image. The target feature similarity map depends exclusively on I_T and I_S and does not change with each fixation. During the free viewing tasks, there is no target image, and we therefore remove the top-down modulation, and use the same deep convolutional neural network to extract the high-level features of the stimulus I_S, aggregating all the feature maps into one saliency map M_{sal}.

Inhibition-of-return refers to the observation that previously fixated locations tend to be inhibited (*11*). Many models of visual search, including the initial version of IVSN (*7*), assume infinite inhibition-of-return. Under these conditions, there cannot be any return fixations since these models remember perfectly the previously visited locations and never look back. Modeling return fixations requires a finite memory. Many studies (*25*, *28–32*) have capitalized on return fixations to study visual working memory. Here we introduced a memory decay function to keep track of previous visited locations and constantly update the visual working memory map $M_{mem,t}$ over all past fixations from 1 to $t-1$ (**Figure S6A**). $M_{mem,t}$ depends only on the previous fixations and is independent of the content of I_T or I_S.

Researchers have also shown that occulomotor biases constrain the saccade sizes (e.g., subjects are more likely to make two 10 dva saccades than one 20 dva saccade). This occulomotor constraint can also impact the frequency of return fixations (*33*). To take saccade size constraints into account, the model incorporated a saccade prior distribution map $M_{sac,t}$. $M_{sac,t}$ depends only on the previous fixation $t-1$ and is independent of the content of I_T or I_S.

The final attention map $M_{f,t}$ integrates M_{sim}, M_{sal}, $M_{mem,t}$ and $M_{sac,t}$. A winner-take-all mechanism selects the maximum local activity as the location for the next fixation at $t+1$. During visual search tasks, if the model recognizes the target at the current fixation location, the search stops. Otherwise, the maps are updated and the model produces a new fixation. During the free viewing tasks, the model stops when it reaches the average number of fixations made by humans or monkeys in the corresponding datasets.

The model was always presented with the exact same images that were shown to the subjects in all the tasks. We focus here on modeling the results for only the first six experiments on static images for several reasons. First, the model does not have any mechanism to process motion information or integrate temporal information across video frames. Second, the model does not have any mechanism to incorporate specific task information such as following a recipe in the cooking task. Third, a model of the egocentric videos would require constructing a memory map in 3D.

Target feature similarity map

The computation of the target feature similarity map M_{sim} follows the IVSN model in reference (*7*). Of note, this is a zero-shot model which does not require training on any eye movement data. We describe the computation of M_{sim} briefly here and refer the reader to reference (*7*) for further details. The "ventral visual cortex" module builds upon the basic bottom-up architecture for visual recognition (*39*, *55*, *56*, *58–60*). We used a state-of-the-art deep feed-forward network, implemented in VGG16 (*39*), pre-trained for image classification on the 2012 version of the ImageNet dataset (*61*). The same set of weights, that is, the same network, is used to process the target image I_T and the search image I_S. The output of the ventral visual cortex module is given by the activations at the top-level (Layer 31 in VGG16, $\phi_{31}(I_T, W)$, and the layer before that (Layer 30 in VGG16), $\phi_{30}(I_S, W)$, in response to the target image and search image, respectively. The top level activation is stored in a "pre-frontal cortex" module. We use the activations in layer 31 in response to the target image to provide top-down modulation to layer 30's response to the search image (**Figure S5A**). This modulation is achieved by convolving the representation of the target with the representation of the search image before max-pooling:

$$M_{sim} = f(\phi_{31}(I_T, W), \phi_{30}(I_S, W)) \tag{1}$$

where $f(\cdot)$ is the target modulation function defined as a 2D convolution operation with kernel $\phi_{31}(I_T, W)$ on the search feature map $\phi_{30}(I_S, W)$.

Saliency map

The saliency map M_{sal} is computed using the same ventral visual cortex module (without any weight changes or retraining). We obtained the activations of size $C_{30} \times W_{30} \times H_{30}$ at the top-level (Layer 30 in VGG16) where C_{30} is the number of channels, W_{30} and H_{30} are the width and height respectively. We take the average over all channels. In other words, $\phi_{30}(I_S, W)$ gets uniformly modulated by an all-ones matrix $J_{C_l \times 1 \times 1}$ of size $C_l \times 1 \times 1$.

$$M_{sal} = f(J_{C_l \times 1 \times 1}, \phi_{30}(I_S, W)) \tag{2}$$

Saccade prior map

Humans and monkeys make relatively small saccades due to occulumotor constraints (*7*, *33*). We used the empirical distribution of saccade sizes to constrain the saccade sizes for the model. Specifically, we plotted the saccade size distribution of the subjects on the corresponding datasets and interpolated to create a 2D map $M_{sac,t}$ centered at the tth fixation. **Figure S6B** plots the empirical saccade size distributions of all fixations over all trials and subjects for each dataset and their corresponding 2D saccade maps when the fixation is at the center. The saccade prior map is updated after each fixation. **Figure S7E** shows example visualizations of saccade priors over fixations.

Memory decay map

Humans and monkeys have limited memory capacity and finite inhibition of return (*11*, *62*, *63*). We added a finite memory module to the model where the 2D memory map $M_{mem,\widetilde{t}}$ at the $\widetilde{t}$th fixation keeps track of memories at all the past fixation locations $\{(x_1, y_1), (x_2, y_2), ..., (x_t, y_t), ..., (x_{\widetilde{t}}, y_{\widetilde{t}})\}$. From the $\widetilde{t}$th back to the 1st fixation locations, the memory value a_t at the tth fixation location gets degraded using the following memory decay function:

$$a_t = \begin{cases} \alpha^{\widetilde{t}-t}, & \text{if } \alpha^{\widetilde{t}-t} \geq \beta \\ \beta, & \text{otherwise} \end{cases} \tag{3}$$

where we set memory decay parameter $\alpha = 0.92$ and clipping threshold $\beta = 0.5$. **Figure S6A** shows the plot of memory value a_t as a function of fixation number t when $\widetilde{t} = 15$. The model's memory decays for the most recent fixations and maintains a low memory level for the rest of past fixations. To avoid sparseness of the 2D memory map $A_{mem,t}$ for the tth fixation at (x_t, y_t), we applied Gaussian filtering centered at that fixation location:

$$A_{mem,t}(x_t, y_t) = a_t \exp \frac{(x-x_t)^2 + (y-y_t)^2}{2\sigma^2} \tag{4}$$

where σ is the standard deviation of the Gaussian. The value of σ controls how much the model remembers adjacent pixels centered around fixation location (x_t, y_t). We set $\sigma = 0.08$ on object arrays (Visual Search 1) and $\sigma = 0.02$ in the other datasets. The different σ is because in object arrays, each object stands alone, and the choice of the Gaussian memory mask

is large enough to cover the complete object on the arrays. To avoid overfitting, we performed a heuristic search for suitable parameters α, β and σ *only* on the Visual Search 2 task, and fixed those parameters for the rest of the tasks.

After updating $A_{mem,t}(x_t, y_t)$ for each fixation t, the model predicts the final memory map $M_{mem,\widetilde{t}}$ by taking the largest memory value across $A_{mem,t}(x_t, y_t)$ for all previous fixation locations $t \in \{1, 2, ..., t, ..., \widetilde{t}\}$. **Figure S7F** shows visualization examples of the memory map $M_{mem,\widetilde{t}}$. When there is a return fixation, by taking the largest memory value across $A_{mem,t}(x_t, y_t)$, the memory value at the revisited location overwrites the decayed memory value at the "to-be-revisited" location.

Integration of feature maps

The model predicts the final attention map $M_{f,t}$ by taking the weighted linear combination of M_{sim}, M_{sal}, $M_{mem,t}$ and $Msac, t$, after normalizing them to [0,1] (**Figure S7**).

$$M_{f,t} = w_{mem} M_{mem,t} + w_{sac} M_{sac,t} + w_{sim} M_{sim} + w_{sal} M_{sal} \tag{5}$$

In the visual search tasks, $w_{sim} = 1$ and $w_{sal} = 0$ and in the free viewing tasks $w_{sim} = 0$ and $w_{sal} = 1$. We fit the 2 weights w_{mem} and w_{sac} to match the return fixation properties *only* on the Visual Search 2 experiment with natural images. We used the following weights: $w_{mem} = -0.93, w_{sac} = 0.2346$. These weights are fixed throughout the experiments and do not depend on t. The model takes the maximum in the attention map $M_{f,t}$ as the location of the $t + 1$-th fixation.

Object recognition

In the visual search tasks, given a fixation location, the model needs to decide whether the target was found or not (in a similar way that humans need to decide whether they found the target after moving their eyes to a new location). The model performs visual recognition to decide whether the target is present at the fixated location. We used a simplified visual recognition mechanism consisting of four steps: (1) we cropped a patch of 1 dva centered at the current fixation; (2) we used the same deep feed-forward architecture described above to extract the activations in the last classification layer of VGG16 in response to the cropped patch; (3) we similarly extracted the activation in the last classification layer of VGG16 in response to the target image I_T, and (4) we computed the cosine similarity distance between the activations for the image patch and the target image.

We computed M_{recog} for each location in the image. At each fixation location, the model retrieved the cosine similarity distance in M_{recog}. We empirically set a hard threshold for cosine similarity distance (0.5 for object arrays and Waldo images; 0.3 for natural images). If the distance between the current fixation patch and the target image is below the threshold, the model decides that the target is found and the search trial stops; otherwise, the model continues the visual search process by updating the four maps and the overall attention map (**Equation 5**).

In the free viewing tasks, there is no target image to recognize. Instead, we stop the model after it generated N_c fixations, where N_c is the average number of fixations by humans or monkeys in the corresponding dataset.

Null model

We compared the model performance against a null model that made random eye movements. Similar to reference (*33*), the null model is memoryless: it does not have any history dependency during prediction of the next fixation location. The only constraint that the null model has is the saccade size, ensuring that the null model can also make return fixations by selecting random locations in the vicinity of the current fixation. Thus, the null model randomly samples the next fixation location from the saccade amplitude distribution $M_{sac,t}$. We used the same stopping criteria for the null model as the one described in the previous paragraph. We ran simulations generating at least 25,000 random sequences of fixations and reported their proportion of return fixations for all datasets in **Figure 3** and **Figure S12**. These random sequences of fixations have the same length as the average number of fixations per trial for each dataset. Similarly, the number of return fixations also impacts the entropy value (**Figure S4**). In order to calculate the chance level for the between-subject consistency analyses, we used the number of return fixations collected from all subjects in each trial (that is, each image) to randomly generate an equal number of random return fixations and computed the entropy for these random return fixations. We repeated this process 100 times for every trial per dataset.

Data analyses

Fixation extraction and calibration

In the visual search tasks, we used the fixations from the previous work (*7*). In the human free viewing task, we used the fixation clustering function from reference (*64*), implemented in MATLAB. During all the human eyetracking experiments on static images, if a fixation was not detected during the initial fixation window in each trial, the experimenter re-calibrated the eye tracker.

In the Monkey Free Viewing 1 task, there were 3-5 re-calibrations. We minimized the number of times for checking re-calibrations in order to maintain the monkeys' attention. The checking process was only activated if the monkeys failed 3 times or more to complete a trial. We used two eyetrackers (four monkeys using ISCAN and two monkeys using Eyelink 1000 Plus) to record monkeys' eye positions. In the Monkey Free Viewing 2 task, calibration was conducted in the beginning of each session. We used the ISCAN eyetracker to record the monkeys' eye positions. We used the built-in Monkeylogic 2 graphics library (*65*) to monitor eye movements on MATLAB.

In the two egocentric datasets (*35*, *36*), calibration was only performed once before each experiment. In a natural environment the resulting egocentric videos represent a combination of head pose and gaze position. Different from static images, both foreground and background objects move with respect to the egocentric coordinate. This implies that the coordinates of a fixated object on the current video frame might be shifted with respect to the next frame or even disappear in the next frame due to abrupt large head motions. To avoid the complexities of using optic flow to track coordinates across frames, we simplified the analyses by considering short segments with minimal head motion. We uniformly split the long egocentric videos into short 5-second video clips. To approximate head motion, we calculated the Euclidean distance between the first frame and the last frame of the video clip at the pixel level and normalized the Euclidean distance by the total number of pixels on each frame (**Figure S3**). We only considered video clips if the normalized Euclidian distance between the first

frame and the last frame was ≤ 0.4.

The field of view of the camera capturing the egocentric videos was limited to 46×60 dva while the field of view for human eyes is at least 120 dva. Thus, we could have cases when the eye tracking data goes beyond the size of the video frame, leading to missing fixations in some video frames. Missing eye tracking data could also arise as a consequence of large head rotations. The coordination between head and gaze movements results in an early movement of eye gazes to the anticipated direction of what subjects intend to look at before the head rotates such that subjects can re-position the object of interest in the center of the field of view (*51*). Therefore, in addition to the constraint based on the normalized Euclidian distance, we also discarded from analyses those video clips with more than 14 consecutive frames with missing fixations. Fourteen consecutive frames at 24 frames per second corresponds to about 3 fixations. For the rest of the video clips with fewer missing fixations, we performed linear interpolation to estimate eye positions in frames with missing data. After all these pre-filtering steps, we ended up with 8,468 video clips in the cooking egocentric videos (**Figure 2G**) and 1,186 clips in the visual search egocentric videos (**Figure 2H**). Despite these efforts to remove large head motion, we could still have video clips with small head movements, which would lead to inaccurate analyses of return fixations. Therefore, for the egocentric videos, we relaxed the threshold in the definition of return fixations to 1.5 dva overlap, instead of 1 dva overlap as used in all the other datasets.

Evaluation of object recognition during visual search

During visual search, subjects can fixate on the target object without realizing that they have found the target and continue searching (there are multiple examples of this phenomenon and discussion in reference (*7*)). We refer to these cases of missing the targets as "false negatives" in visual recognition (**Figure S11**). To compute the false negative rate for humans, we counted the total number of fixations on the targets without mouse clicks and divided it by the total number of fixations for all the trials (**Figure S11**).

Conversely, there could also be cases when humans are not looking at the target but rather they are fixating on a distractor; yet, they misclassify the distractor as the target by clicking the mouse at the wrong location. We refer to these failure cases as "false positives". We computed the false positive rate as the number of trials when false clicks happen within that trial, normalized by the total number of trials (**Figure S11**).

In the object array experiments, subjects were not asked to click the target location with the mouse and therefore we cannot compute false positives or false negatives. Thus, we only reported false positive and false negative rates in the Visual Search 2 and 3 datasets (natural images and Waldo). The model similarly makes false positives and false negatives, also reported in (**Figure S11**).

Evaluation of return fixation properties

Definition of return fixations: A fixation location was considered to be a *return fixation* if it was within one degree of visual angle (dva) of a previous fixation location (**Figure 1**, **Figure 4A**). One dva is approximately the resolution of the eye tracking data in the experiments reported here. The original fixation is referred to as a *to-be-revisited* location. All other fixations

are referred to as *non-return* locations. This definition is consistent with criteria previously used in the literature (e.g., (*20*)). In the egocentric video datasets (**Figure 1G-H**), we relaxed the degree of overlap to 1.5 dva to account for the alignment imprecision introduced by potential head movements.

Proportion of return fixations: The proportion of return fixations is defined as the number of return fixations normalized by the total number of fixations in each trial. **Figure 3** reports the proportion of return fixations averaged over all subjects and **Figure S12** reports the proportion of return fixations for the model. In the visual search tasks, we further divided the return fixations based on whether they landed on the target objects or non-target locations (**Figure 5B**). The proportion of return fixations at target locations was calculated as the number of on-target return fixations divided by the total number of on-target fixations per trial. Similarly, we calculated the proportion of non-target return fixations. We omitted the division of on-target and non-target return fixations on the egocentric visual search dataset because of the lack of annotations of target object locations in each video frame.

Return offset: The *return offset* was defined as the number of intervening fixations in between a to-be-revisited location and a return fixation. For example, in **Figure 4A** the return offset between to-be-revisited location 3 and return fixation 6 is 2. **Figure 4B** shows the distributions of return offsets for all experiments and **Figure S14** shows the corresponding distributions for the model.

Saliency: To study whether return fixations correlate with saliency, we evaluated saliency maps using Graph-based Visual Saliency (GBVS), which is a bottom-up saliency prediction algorithm in computer vision using low-level visual features, such as color, orientation, and contrast (*38*). We computed the average of all saliency values for each fixation patch, defined as a squared region covering one dva and centered at the fixation location. **Figure 5A** shows saliency for all the experiments and **Figure S14** shows saliency for the model.

In the case of egocentric videos, it is not justifiable to establish a one-to-one mapping from an extracted fixation (lasting approximately 250 ms) to a single video frame (about 42 ms, 24Hz video frame rate). The fact that one fixaiton involves many frames implies the need for further assumptions to calculate the saliency value for a fixation. Assuming that we have little head motion for the selected video clips, we approximated the saliency value for a fixation by projecting all the fixations within a video clip back to the last frame, and computed saliency using GBVS on the last frame.

Mouse clicks: In the Visual Search 2 and Visual Search 3 experiments, subjects were asked to use the mouse to click on the location of the target. We asked whether there were additional return fixations at the end of each trial which were related to this testing procedure. For example, subjects may find the target, then look for the mouse position, and then make a second saccade to the target. We conducted two additional analyses to evaluate this possibility. First, if subjects look back for the mouse pointer, the fixation preceding the return fixations should overlap with the mouse pointer location. Subjects were instructed not to move the mouse during visual search, except when the target was found at the end of the trial. We computed the proportion of fixations preceding return fixations that overlapped with the mouse pointer location with respect to the total number of return fixations in both experiments. There were only 1.5% and 1.3% of preceding return fixations overlapping with mouse pointer locations in the Visual Search 2 and 3 experiments, respectively. Second, to further quantify the effect of looking for mouse pointers, we examined the first 6 fixations in each trial (**FigureS2**). In both experiments, it took subjects much more than 6 fixations to find the target(*7*). During the initial 6 fixations, it is unlikely that the subjects were looking

for the mouse pointer. During the first 6 fixations, the proportion of return fixations was significantly above chance in the Visual Search 2 experiment, but not in the Visual Search 3 experiment. In sum, the presence of return fixations in the visual search experiment 1 (where subjects did not use the mouse), combined with the presence of return fixations during the first 6 fixations in the Visual Search 2 experiment, and together with the low fraction of return fixations where subjects look for the mouse pointer before making a return fixation, suggest that the majority of return fixations cannot be ascribed to subjects searching for the mouse pointer.

Between-subject consistency in return fixation locations

To evaluate between-subject consistency (**Figure S4, Figure S17**), we performed the following steps: (1) we mapped all return fixations from all subjects on each image to a uniform 2D grid of size 32 by 40 (**Figure S4B**); (2) we computed the proportion of subjects that showed a return fixation at location l, with $l = 1, ..., 32 \times 40$; (3) computed the entropy H for this distribution over L locations using the following equation:

$$H(L) = -\sum_{l=1}^{32\times 40} p_l \log(p_l) \tag{6}$$

An extreme case of perfect consistency would lead a probability of 1 at a given location and 0 elsewhere, leading to minimal entropy, whereas a complete lack of consistency would lead to an approximately uniform probability distribution except for random overlaps, resulting in high entropy. We omitted this analysis in the egocentric video datasets because the field of view in each frame could be different across subjects, making comparisons difficult to interpret.

Statistical analyses

We used two-tailed t-tests when comparing two distributions and considered results to be statistically significant when $p < 0.05$. Because calculations of p values tend to be inaccurate when the probabilities are extremely low, we reported all p values less than 10^{-15} as $p < 10^{-15}$ (as opposed to reporting, for example, $p = 10^{-40}$); clearly none of the conclusions depend on this.

Code and Data availability

All the source code and raw data is publicly available through the lab's GitHub repository: https://github.com/kreimanlab/RefixationModel.

Acknowledgments

This work was supported by NIH grant R01EY026025 and by the Center for Brains, Minds and Machines, funded by NSF Science and Technology Centers Award CCF-1231216. MZ is supported by postdoctoral fellowship of Agency for Science, Technology and Research (Institute of Infocomm Research). We thank Jeremy Wolfe for helpful discussions and advice, Pranav Misra for help with the datasets, Yin Li, Miao Liu and Keng Teck Ma for helping with the egocentric video datasets.

Competing interests

The authors declare that they have no competing interests.

References

1. J. L. Orquin, S. M. Loose, *Acta psychologica* **144**, 190–206 (2013).
2. A. C. Schütz, D. I. Braun, K. R. Gegenfurtner, *Journal of vision* **11**, 9–9 (2011).
3. S. K. Ungerleider, L. G, *Annual review of neuroscience* **23**, 315–341 (2000).
4. J. W. Bisley, *The Journal of physiology* **589**, 49–57 (2011).
5. M.-H. Grosbras, A. R. Laird, T. Paus, *Human brain mapping* **25**, 140–154 (2005).
6. S. Paneri, G. G. Gregoriou, *Frontiers in neuroscience* **11**, 545 (2017).
7. M. Zhang *et al.*, *Nature communications* **9**, 1–15 (2018).
8. T. Miconi, L. Groomes, G. Kreiman, *Cerebral cortex* **26**, 3064–3082 (2015).
9. L. Itti, C. Koch, *Vision research* **40**, 1489–1506 (2000).
10. S. S. Kruthiventi, K. Ayush, R. V. Babu, *IEEE Transactions on Image Processing* **26**, 4446–4456 (2017).
11. R. M. Klein, *Trends in cognitive sciences* **4**, 138–147 (2000).
12. K. Rayner, *Psychological bulletin* **124**, 372 (1998).
13. D. H. Ballard, M. M. Hayhoe, J. B. Pelz, *Journal of cognitive neuroscience* **7**, 66–80 (1995).
14. M. Hayhoe, D. Bensinger, D. Ballard (1997).
15. I. Neath, *Human memory: An introduction to research, data, and theory.* (Thomson Brooks/Cole Publishing Co, 1998).
16. M. Hegarty, R. E. Mayer, C. E. Green, *Journal of educational psychology* **84**, 76 (1992).
17. M. R. Beck, M. S. Peterson, M. Vomela, *Journal of Experimental Psychology: Human Perception and Performance* **32**, 235 (2006).
18. C. A. Dickinson, G. J. Zelinsky, *Psychonomic Bulletin & Review* **12**, 1120–1126 (2005).
19. G. J. Solman, J. A. Cheyne, D. Smilek, *Vision research* **51**, 1185–1191 (2011).
20. N. C. Anderson, W. F. Bischof, K. E. Laidlaw, E. F. Risko, A. Kingstone, *Behavior research methods* **45**, 842–856 (2013).
21. Z. Wang, R. M. Klein, *Vision research* **50**, 220–228 (2010).
22. W. J. MacInnes, A. R. Hunt, M. D. Hilchey, R. M. Klein, *Attention, Perception, & Psychophysics* **76**, 280–295 (2014).
23. K. Ruddock, D. Wooding, S. Mannan, *Spatial vision* **10**, 165–188 (1996).
24. S. Mannan, K. Ruddock, D. Wooding, *Spatial vision* **11**, 157–178 (1997).

25. G. J. Zelinsky, L. C. Loschky, C. A. Dickinson, *Memory & Cognition* **39**, 600–613 (2011).

26. D. L. Sheinberg, N. K. Logothetis, *J Neurosci* **21**, 1340–50. (2001).

27. T. J. Smith, J. M. Henderson, *Journal of Vision* **11**, 3–3 (2011).

28. L. E. Thomas *et al.*, *Psychonomic bulletin & review* **13**, 891–895 (2006).

29. I. D. Gilchrist, M. Harvey, *Current Biology* **10**, 1209–1212 (2000).

30. C. Körner, I. D. Gilchrist, *Psychological Research* **72**, 99–105 (2008).

31. K. Shen, A. R. McIntosh, J. D. Ryan, *Journal of Vision* **14**, 11–11 (2014).

32. B. W. Tatler, I. D. Gilchrist, M. F. Land, *The Quarterly Journal of Experimental Psychology Section A* **58**, 931–960 (2005).

33. P. M. Bays, M. Husain, *Journal of vision* **12**, 8–8 (2012).

34. A. R. Nikolaev, R. N. Meghanathan, C. van Leeuwen, *Journal of neurophysiology* **120**, 2311–2324 (2018).

35. Y. Li, M. Liu, J. M. Rehg, presented at the Proceedings of the European Conference on Computer Vision (ECCV), pp. 619–635.

36. M. Zhang, K. Teck Ma, J. Hwee Lim, Q. Zhao, J. Feng, presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4372–4381.

37. T. Foulsham, G. Underwood, *Journal of vision* **8**, 6–6 (2008).

38. J. Harel, C. Koch, P. Perona, presented at the Advances in neural information processing systems, pp. 545–552.

39. K. Simonyan, A. Zisserman, *arXiv preprint arXiv:1409.1556* (2014).

40. L. Itti, C. Koch, E. Niebur, *IEEE Transactions on pattern analysis and machine intelligence* **20**, 1254–1259 (1998).

41. A. Borji, D. N. Sihite, L. Itti, *Vision research* **91**, 62–77 (2013).

42. S. Shaunak, E. O'sullivan, C. Kennard, *Journal of neurology, neurosurgery, and psychiatry* **59**, 115 (1995).

43. P. Glimcher, in *Fundamental neuroscience* (Academic Press, 1998).

44. M. C. Potter, *J Exp Psychol Hum Learn* **2**, 509–22, ISSN: 0096-1515 (Print) 0096-1515 (Linking) (1976).

45. S. Thorpe, D. Fize, C. Marlot, *Nature* **381**, 520–522 (1996).

46. H. Liu, Y. Agam, J. Madsen, G. Kreiman, *Neuron* **62**, 281–290 (2009).

47. T. Serre, A. Oliva, T. Poggio, *PNAS* **104**, 6424–6429 (2007).

48. M. Zhang, C. Tseng, G. Kreiman, *CVPR* (2019).

49. Y. Li, A. Fathi, J. M. Rehg, presented at the Proceedings of the IEEE International Conference on Computer Vision, pp. 3216–3223.

50. J. Pelz, M. Hayhoe, R. Loeber, *Experimental brain research* **139**, 266–277 (2001).

51. M. F. Land, *Experimental brain research* **159**, 151–160 (2004).

52. H. Tang *et al.*, *Neuron* **83**, 736–748 (2014).

53. H. Tang *et al.*, *PNAS* **115**, 8835–8840 (2018).

54. T.-Y. Lin *et al.*, presented at the European conference on computer vision, pp. 740–755.

55. M. Riesenhuber, T. Poggio, *Nature neuroscience* **2**, 1019–1025 (1999).

56. T. Serre *et al.*, *Progress in brain research* **165**, 33–56 (2007).

57. G. Kreiman, T. Serre, *Annals of the New York Academy of Sciences* **1464**, 222–241 (2020).

58. G. Wallis, E. T. Rolls, *Progress in neurobiology* **51**, 167–194 (1997).

59. A. Krizhevsky, I. Sutskever, G. E. Hinton, presented at the Advances in neural information processing systems, pp. 1097–1105.

60. K. Fukushima, S. Miyake, in *Competition and cooperation in neural nets* (Springer, 1982), pp. 267–285.

61. O. Russakovsky *et al.*, *International journal of computer vision* **115**, 211–252 (2015).

62. T. S. Horowitz, *Visual Cognition* **14**, 668–684 (2006).

63. T. S. Horowitz, J. M. Wolfe, *Nature* **394**, 575–577 (1998).

64. S. D. Koenig, PhD thesis, 2017.

65. J. Hwang, A. R. Mitz, E. A. Murray, *Journal of neuroscience methods* **323**, 13–21 (2019).

Figures and figure captions

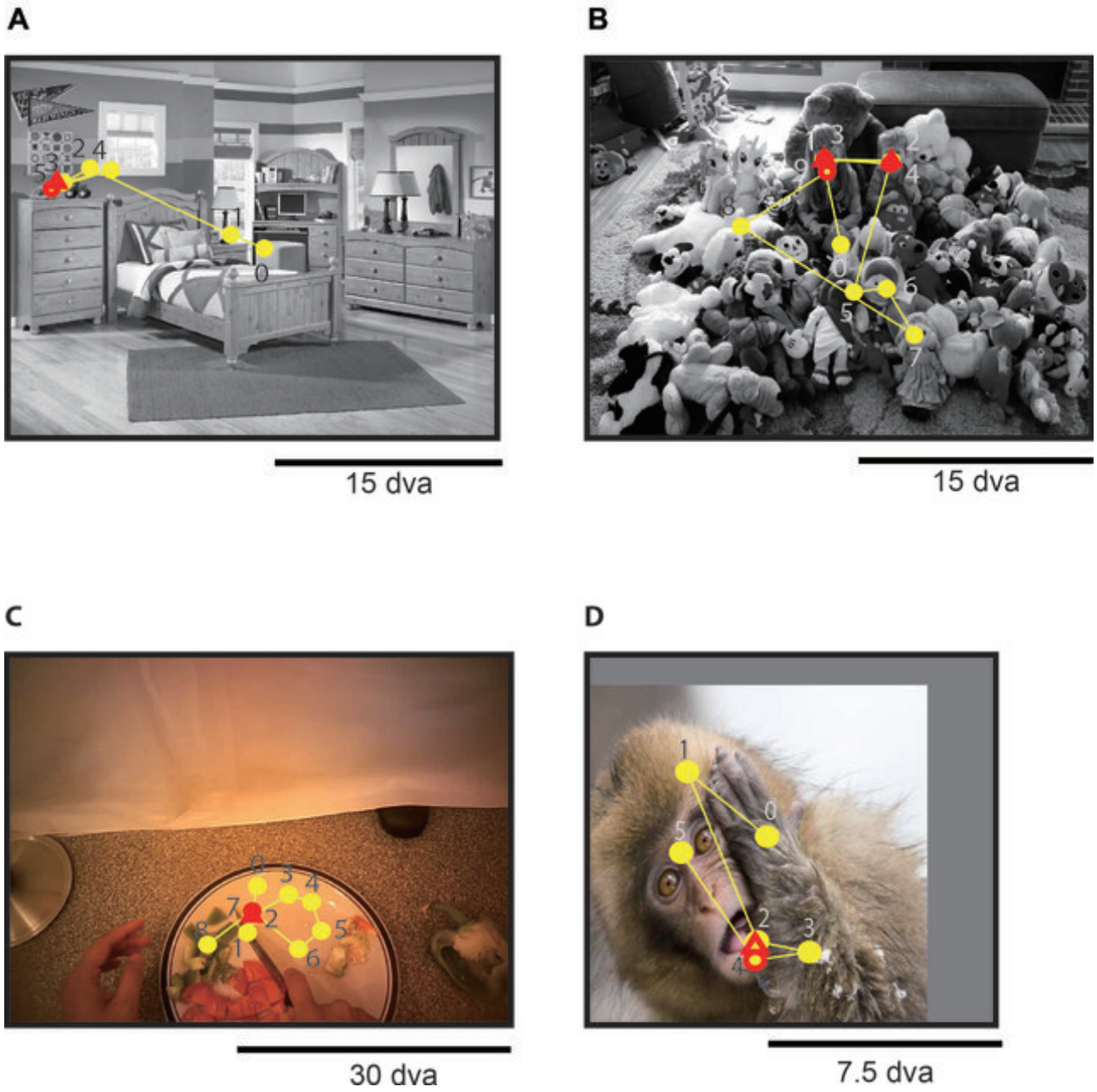

Figure 1. **Primates make return fixations during natural vision.** Example fixation sequences (yellow circles) of humans (**A-C**) and monkeys (**D**) during visual search (**A**), free viewing of static images (**B, D**), and a cooking task with a head-mounted eye tracker (**C**) (see **Figure 2** for task definitions). The numbers denote the fixation order. A fixation (yellow circles) is referred to as a "return fixation" (red triangle) if the Euclidean distance to any of the previous fixations is less than 1 degree of visual angle (dva). The previous fixation overlapping with the return fixation is referred to as "fixation-to-be-revisited" (red circle). There are two return fixations in **B** and one in **A**, **C**, and **D**.

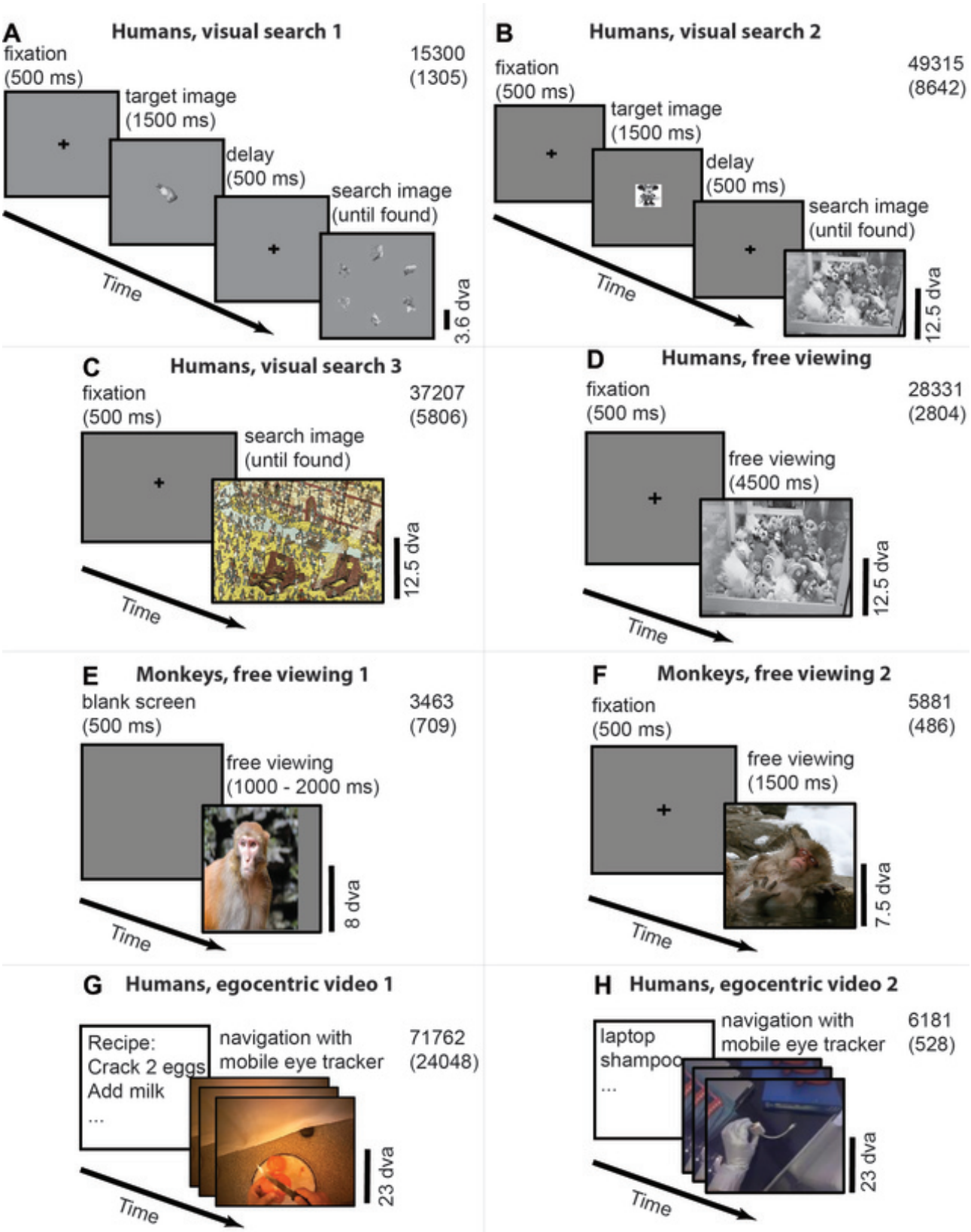

Figure 2. **Schematic description of the eight experimental paradigms.** (**A-C**) Visual search tasks with object arrays (**A**), natural images (**B**), and "Waldo" images (**C**), (see reference(*7*) for details). (**D-F**) free viewing experiments with static natural images in humans (**D**) and monkeys (**E, F**). (**G**) Egocentric video dataset where subjects had to follow various recipes to make breakfast (see (*35*) for details). (**H**) Egocentric video dataset where subjects had to search for 22 items (see reference(*36*) for details). The numbers in the top right corner of each subplot denote the total number of fixations (top) and the total number of return fixations (bottom).

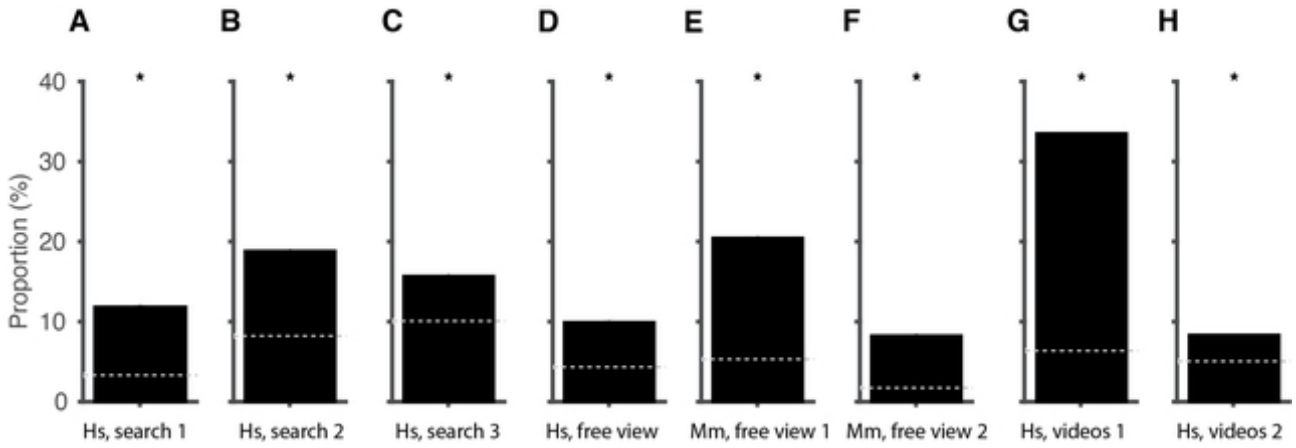

Figure 3. **Human and non-human primates make frequent return fixations.** The proportion of return fixations is defined as the total number of return fixations normalized by the total number of fixations. **A-H** refer to the eight experiments in **Figure 2**. Error bars indicate SEM across all subjects The chance level (dashed lines) was computed by generating sequences of random fixations (**Methods**). The proportion of return fixations was higher than chance in all experiments ($*$ denotes $p < 0.05$, one-sample t-test).

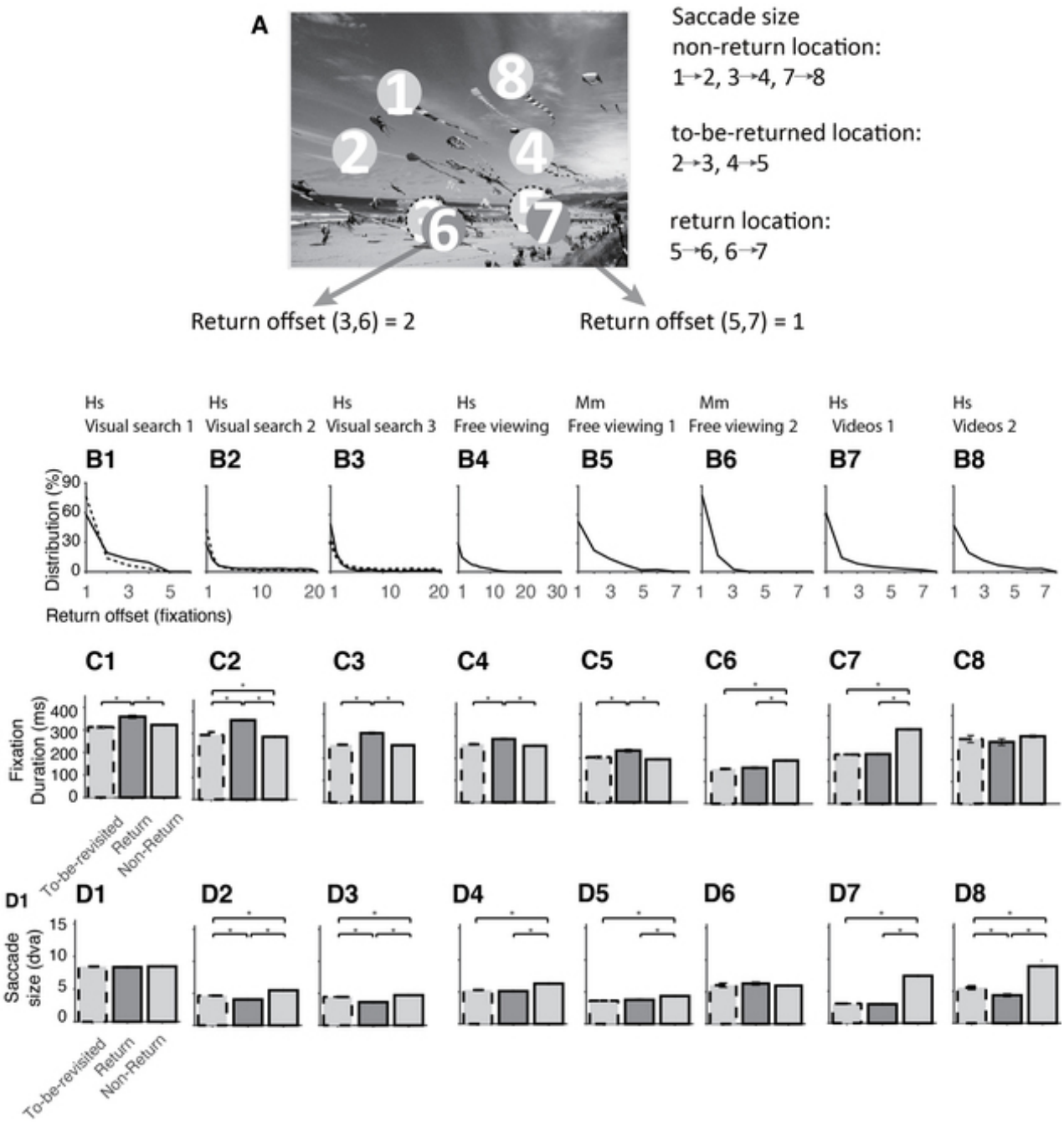

Figure 4. **Return fixations showed short return offsets, tended to last longer and to follow smaller saccades**. (**A**) Example sequence of eight fixations on an image, including return fixations (6 and 7), to-be-revisited fixations (3 and 5), and non-return fixations (1, 2, 4 and 8). The return offset is defined as the number of intervening fixations for a given return location. (**B**) Distribution of the return offset for the 8 experiments. In the visual search experiments (**B1, B2, B3**), the solid line shows return fixations to target locations and the dashed line shows return fixations to non-target locations. (**C**) The durations of return fixations tended to be longer for return locations. Asterisks ($*$) denote statistically significant differences between two distributions. (**D**) The sizes of saccades to return locations tended to be smaller ($*$ denotes $p < 0.05$, two-tailed t-test).

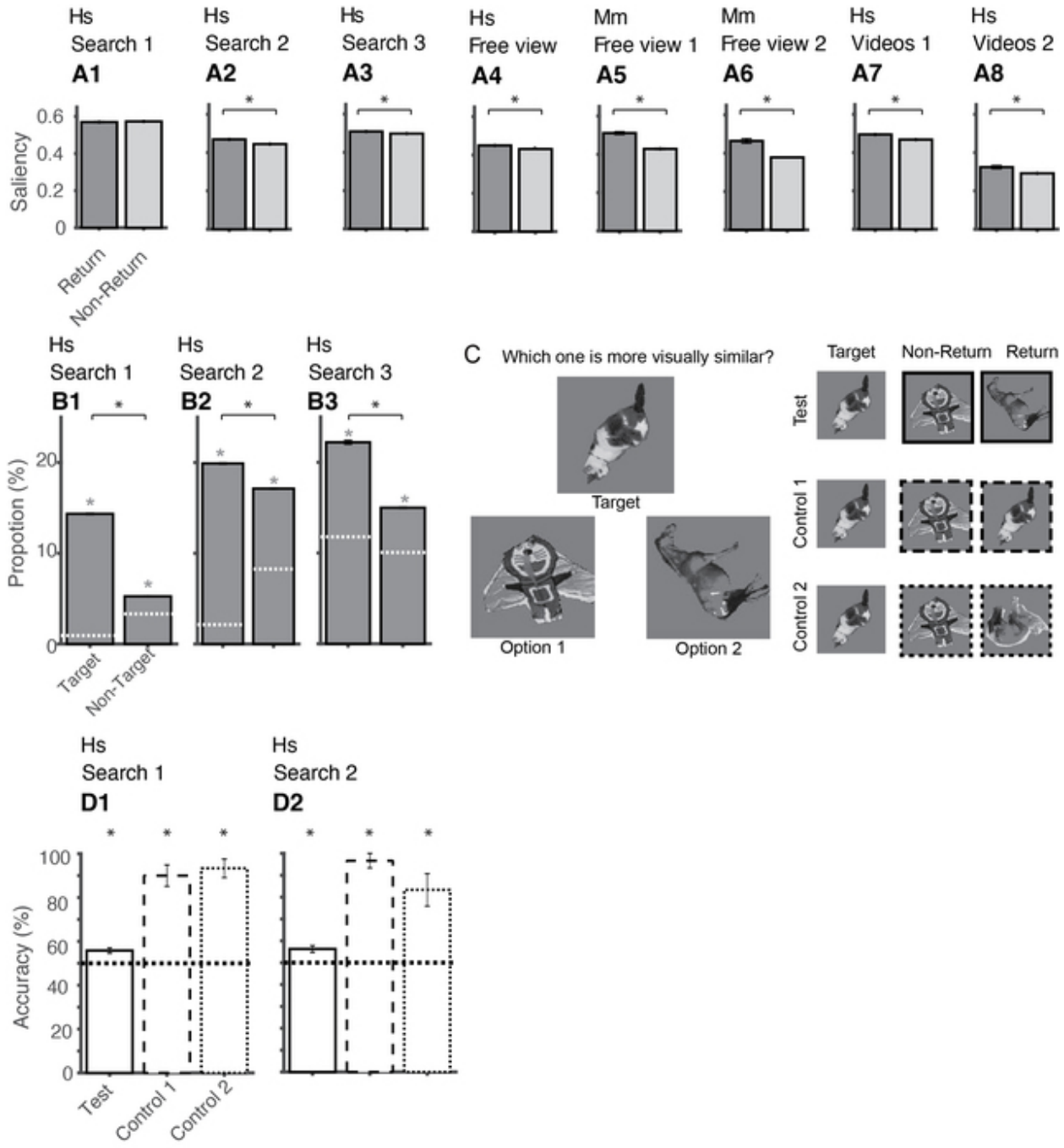

Figure 5. **Return fixation locations depended on the image and task**. (**A**) Return locations tended to show higher saliency. Asterisks denote statistical significance ($*$ denotes $p < 0.05$, two-tailed t-test). Because to-be-revisited locations and return locations overlapped by definition, saliency at to-be-revisited locations was similar to that at return locations and is not shown here. (**B**) During visual search tasks, the proportion of return fixations was larger for target than non-target locations. Gray asterisks denote a statistically significant difference in the proprtion of return fixations with respect to the chance levels. Black asterisks denote a statistically significant difference betwee the proportion of return fixations in target versus non-target locations. (**C**) Schematic of an experiment to assess whether return fixations shared similarity to the target. In each trial, subjects were presented with an image (Target) and had to choose the more similar image between two alternatives. There were three conditions: (1) Test, where the two alternatives were return fixations versus non-return fixations, (2) Control 1, where the two alternatives were an identical copy of the target versus non-return fixations, and (3) Control 2, where the two alternatives were a different object from the same category as the target versus non-return fixations. (**D**) Accuracy in distinguishing images from each of the conditions in **C** versus non-return fixations. Asterisks denotes statistically significant difference from chance levels (horizontal dashed line at 50%, ($*$ denotes $p < 0.05$, one sample t-test)).

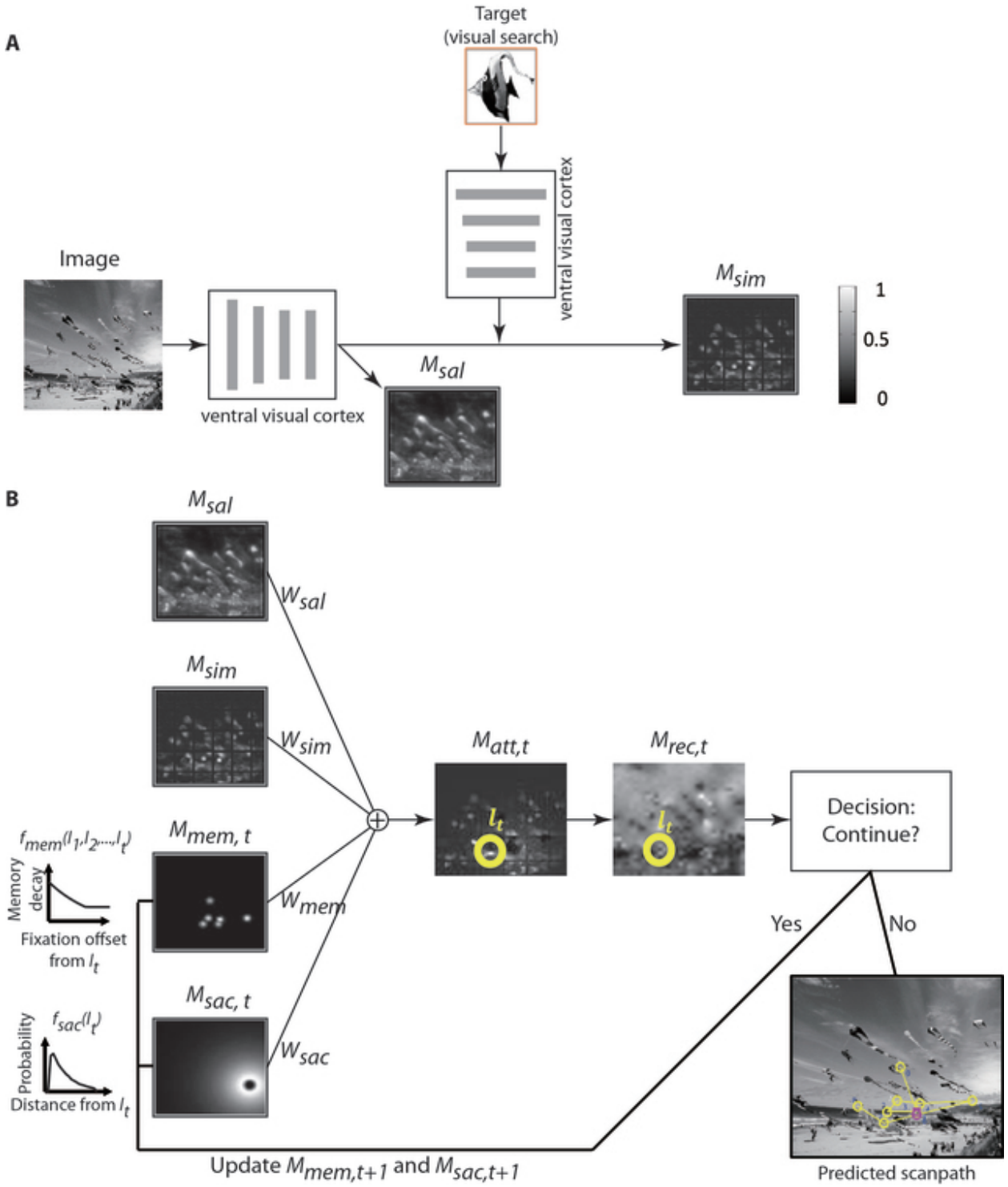

Figure 6. **Architecture of the computational model**. (**A**) The model has a ventral visual cortex module (pre-trained VGG-16) that extracts features from the image. These features constitute the saliency map (M_{sal}). In visual search tasks, the same ventral visual cortex module processes the target image and modulates the features in the search image via top-down modulation, generating a target similarity map (M_{sim}) (*7*). See **Figure S5** and **Methods** for detailed architecture. (**B**) The generation of eye movements is governed by a weighted combination of 4 maps: M_{sal} (part **A**), M_{sim} (part **A**, only in visual search tasks), a time-dependent memory map ($M_{mem,t}$, and a time-dependent saccade size constrain map ($M_{sac,t}$). At each time point t, the 2D spatial memory decay map $M_{mem,t}$ is updated based on all previous visited locations $\{l_1, l_2, \ldots, l_t\}$. The brighter the color on the memory decay map, the stronger the effect of memory inhibition. The saccade size constrain map is also updated at each time point according to the current fixation location, l_t. A winner-take-all chooses the maximum in the combined attention map $M_{att,t}$ (yellow circle) as the location for the next fixation. In visual search tasks, the model computes a recognition map M_{recog} indicating the confidence that the current fixation contains the sought target (**Figure S5B**). If the target is found, the search stops; otherwise, it continues. During free viewing tasks, the recognition map is not used and the model keeps generating eye movements for a fixed amount of time.

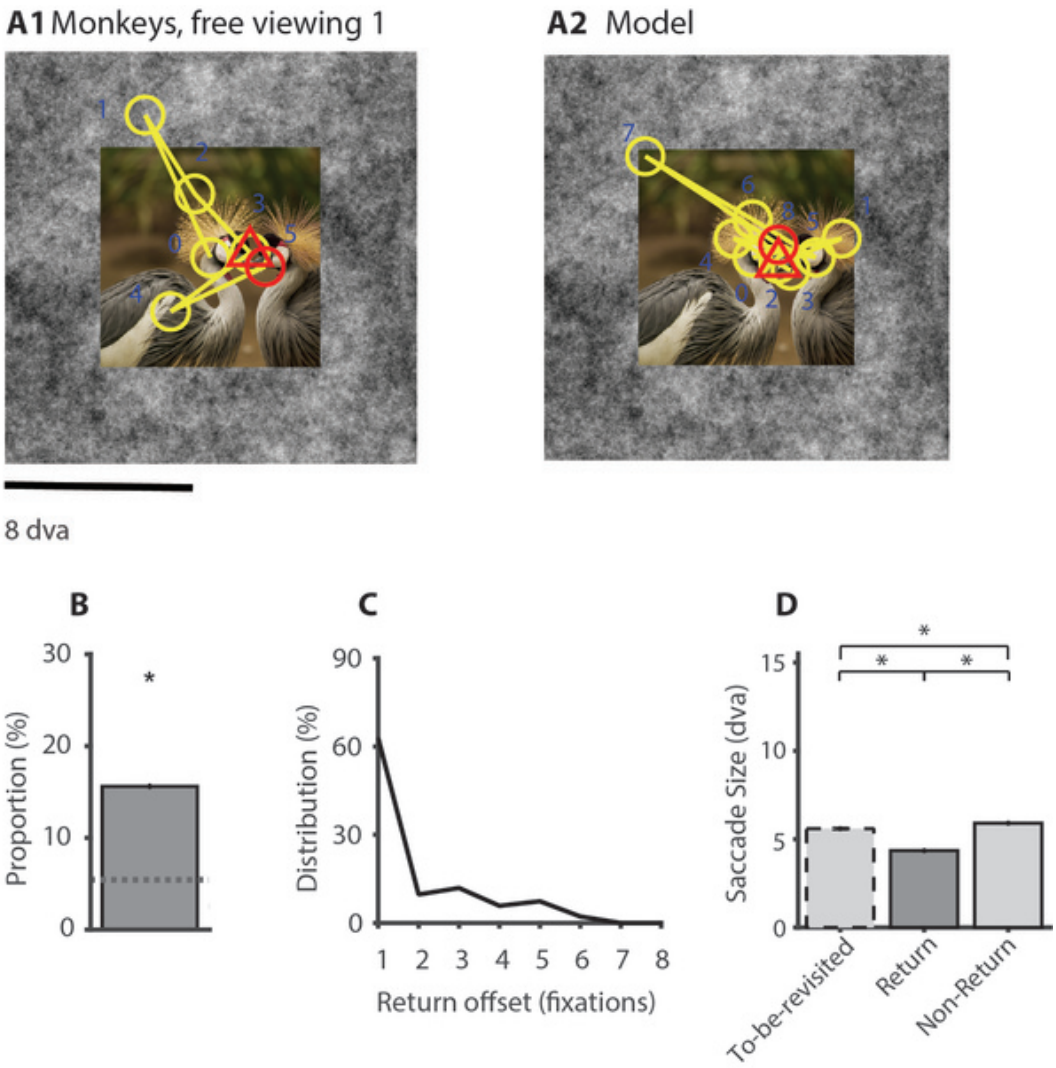

Figure 7. **The computational model generates return fixations.** (**A**). Example scanpaths by monkeys (**A1**) and by the model (**A2**) while free viewing natural images (conventions follow those in **Figure 1**.) (**B**) Proportion of return fixations. (**C**) Distribution of the offset between two fixations on the same location. (**D**) Saccade sizes. Panels (**B, C, D**) follow the same conventions as those in **Figure 3E** and **Figure 4B, D**. See **Figure S5** for model implementation and a full comparison of all return fixation properties predicted by the model for all static image datasets in **Figure S12, Figure S14, Figure S15**, and **Figure S16**.

Supplementary Materials

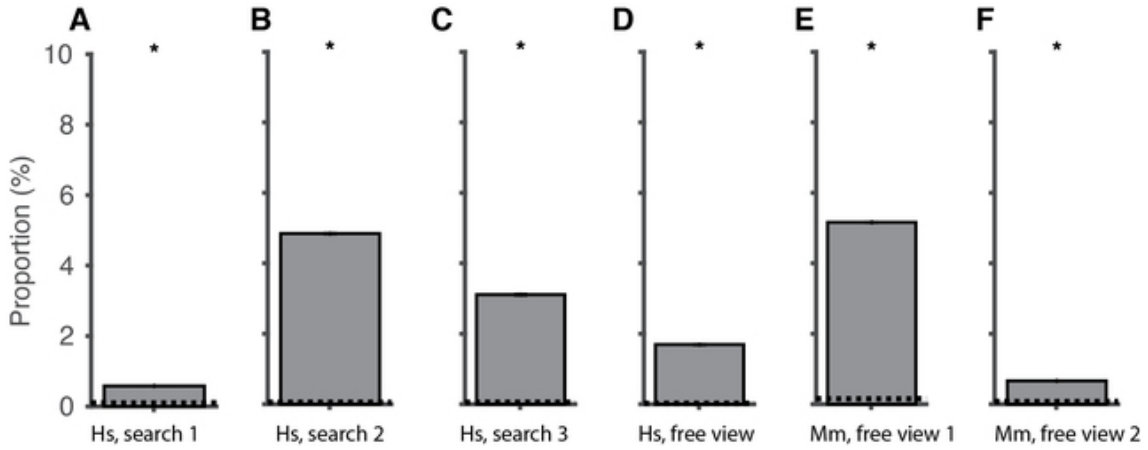

Figure S1. **Proportion of fixations that revisit the same location twice**. Following the format in **Figure 3**, here we show the proportion of fixations that return to the same location twice (as in the sequence L_1-L_2-L_1-L_3-L_4-L_1 where L indicates different image locations).

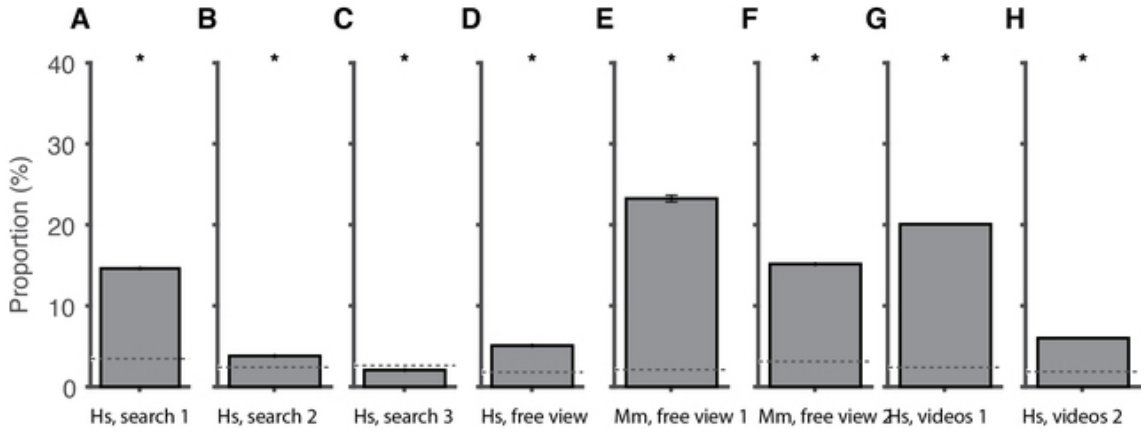

Figure S2. **Proportion of return fixations among the first six fixations**. Following the format in **Figure 3**, here we show the proportion of return fixations among the first six fixations of all scanpaths over all eight datasets. The proportion of return fixations among the first six fixations was higher than chance ($*$ denotes $p < 0.05$, one-sample t-test) in all experiments except Visual Search 3.

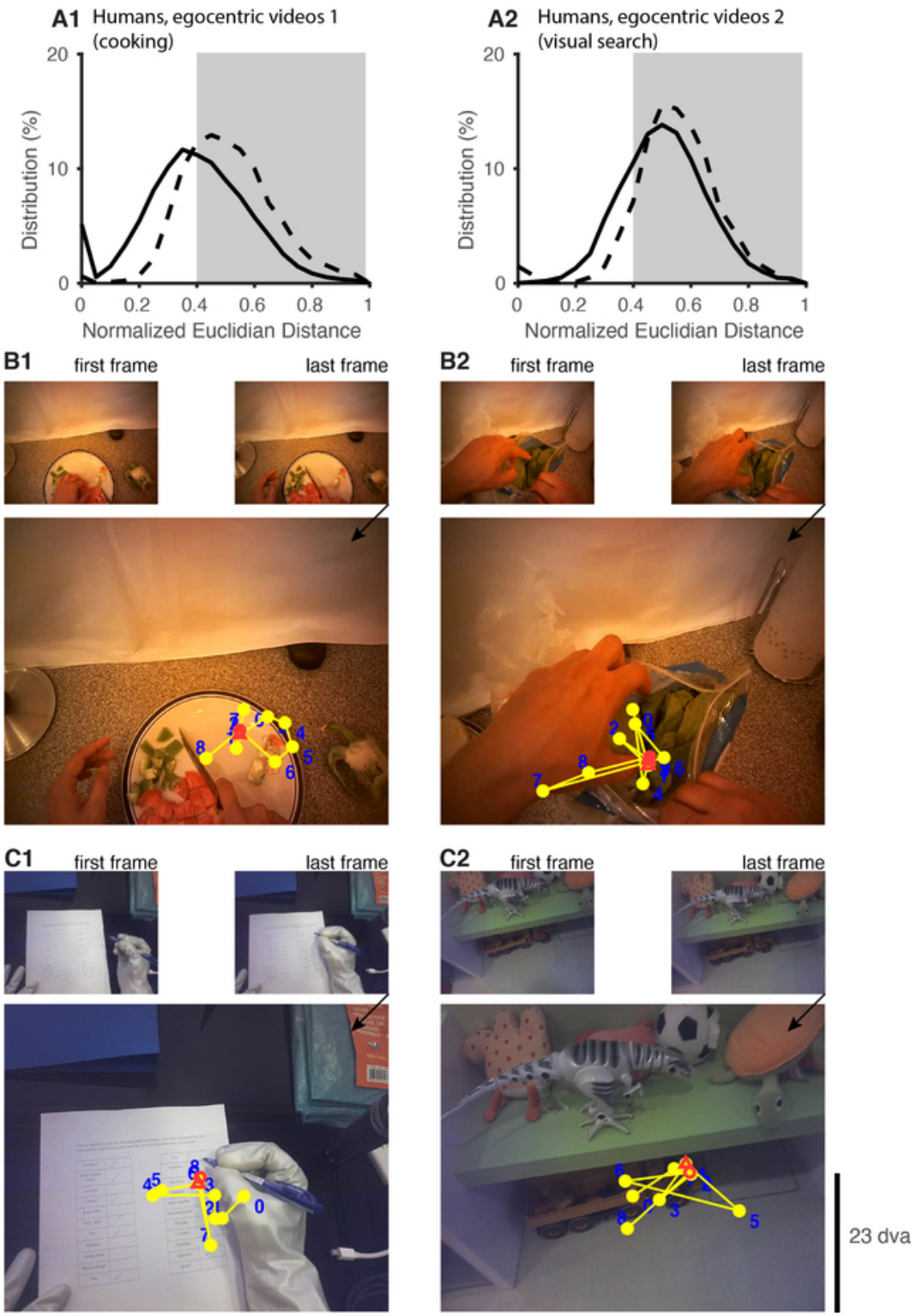

Figure S3. **Extraction of fixations on egocentric videos**. Egocentric videos (**Figure 2G-H**) were analyzed in 5-second segments. (**A**). Distribution of normalized Euclidian distance between the first frame and the last frame (solid lines). The chance distribution of normalized Euclidian distance values was computed by considering random pairs of frames from 100 different video sequences (dashed line). Segments with large head motion (normalized Euclidian distance > 0.4, gray rectangle) were excluded from analyses. (**B-C**). Examples of the first and last frame in 5-second clips (top) and extracted fixations (yellow circles) mapped to the last frame of example video clips (bottom) during the cooking task (**B**, **Figure 2G**) and visual search (**C**, **Figure 2H**). The figure conventions follow **Figure 1**.

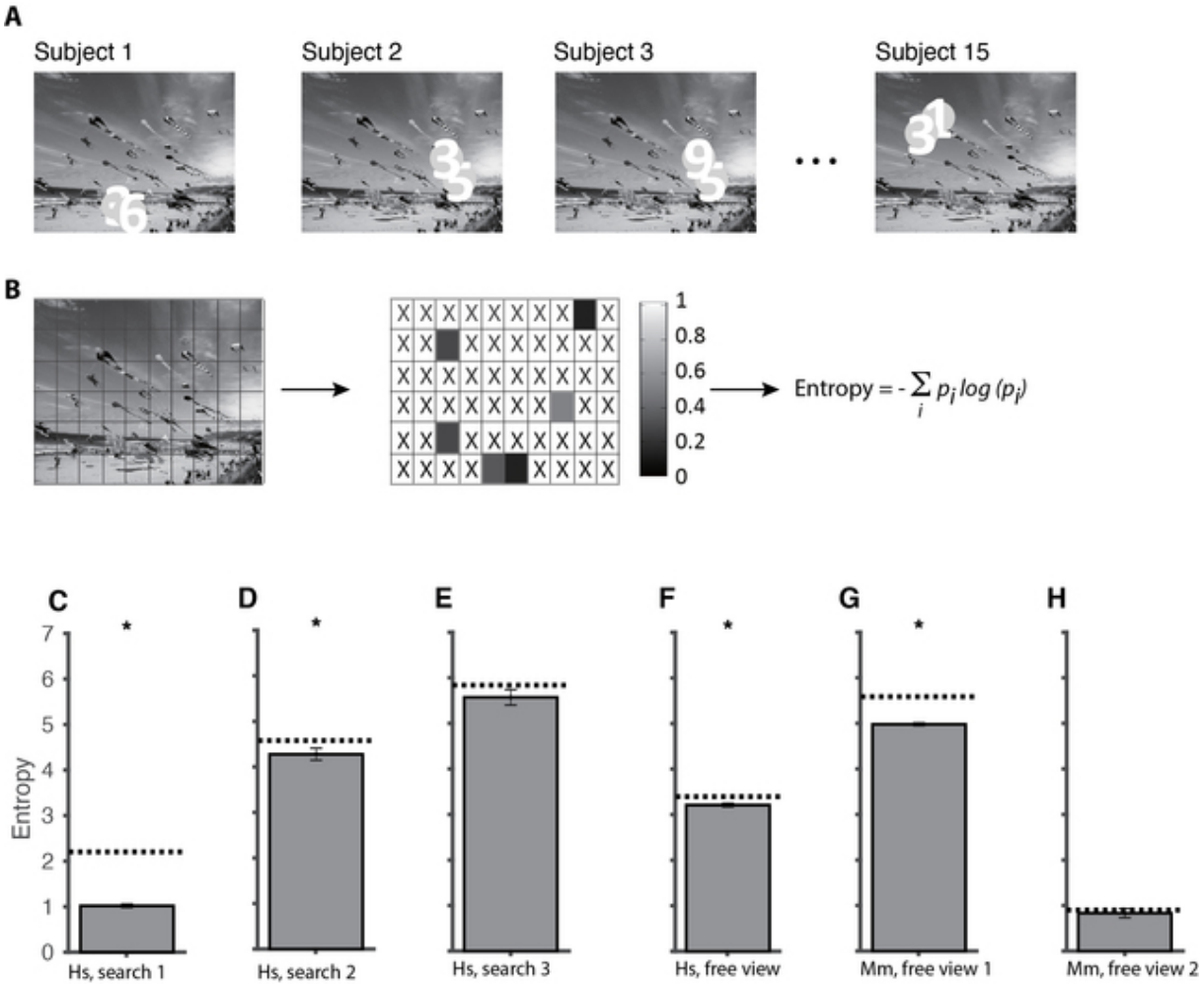

Figure S4. **Return fixations are consistent across subjects**. (**A**). Example return fixations over multiple subjects. The numbers indicate the fixation number in the sequence for each subject. (**B**). The image is divided into a 32×40 grid (only a few of the lines are shown in the image). For those locations where there is at least one return fixation, we compute the proportion of the total number of return fixations that land on that location (shown here by the grayscale colormap). The entropy of the distribution is then computed from those probability values. The lower the entropy, the higher the consistency between subejcts. (**C**). Entropy for each experiment. Horizontal dashed lines show the expected chance value obtained by assigning the total number of return fixations to random locations for each image.

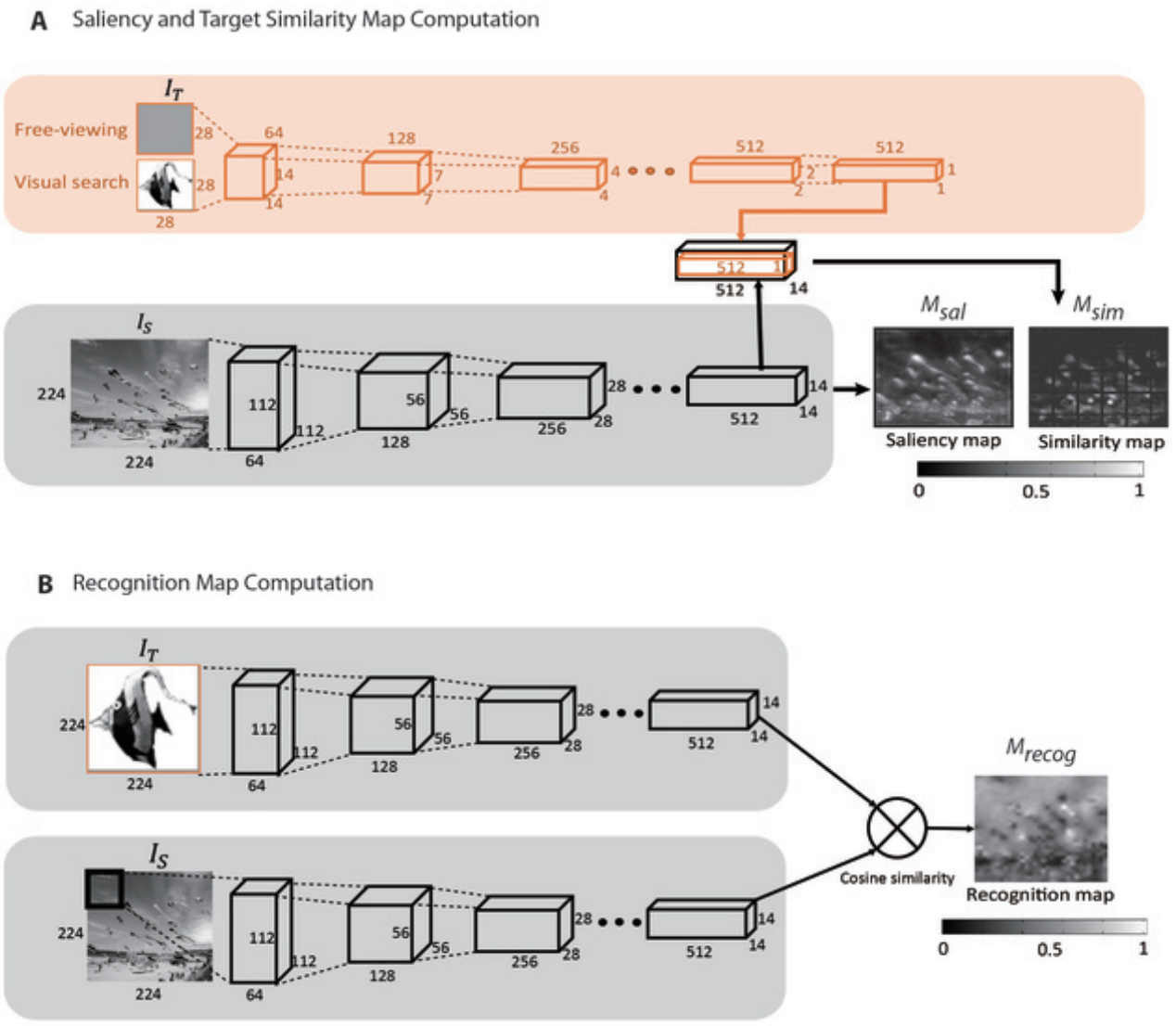

Figure S5. **Calculation of bottom-up saliency map, target similarity map, and recognition map in the computational model**. (**A**). At the heart of the model is a pre-trained deep convolutional network that mimics image processing along the ventral visual cortex. Here we used the VGG-16 architecture (*39*). The features extracted from the top convolutional layer in the model yield the saliency map, M_{sal}. During visual search tasks, we following the work of Zhang et al (*7*), to create a target similarity map, M_{sim}, based on comparing the target features (orange box) and the search image features (gray box). Only some of the layers are shown here for simplicity, the dimensions of the feature maps are indicated for each layer. (**B**). For each location on the search image, we cropped a fixation patch of size 224×224 and used it as input to the same pre-trained recognition network as (**A**) to extract the feature maps. Similarly, we obtained the feature maps of the target image. We computed the cosine similarity between the target and cropped features, resulting in a 2D spatial recognition map M_{recog} where each location on the map denotes a confidence value of recognizing the target at each location on I_S.

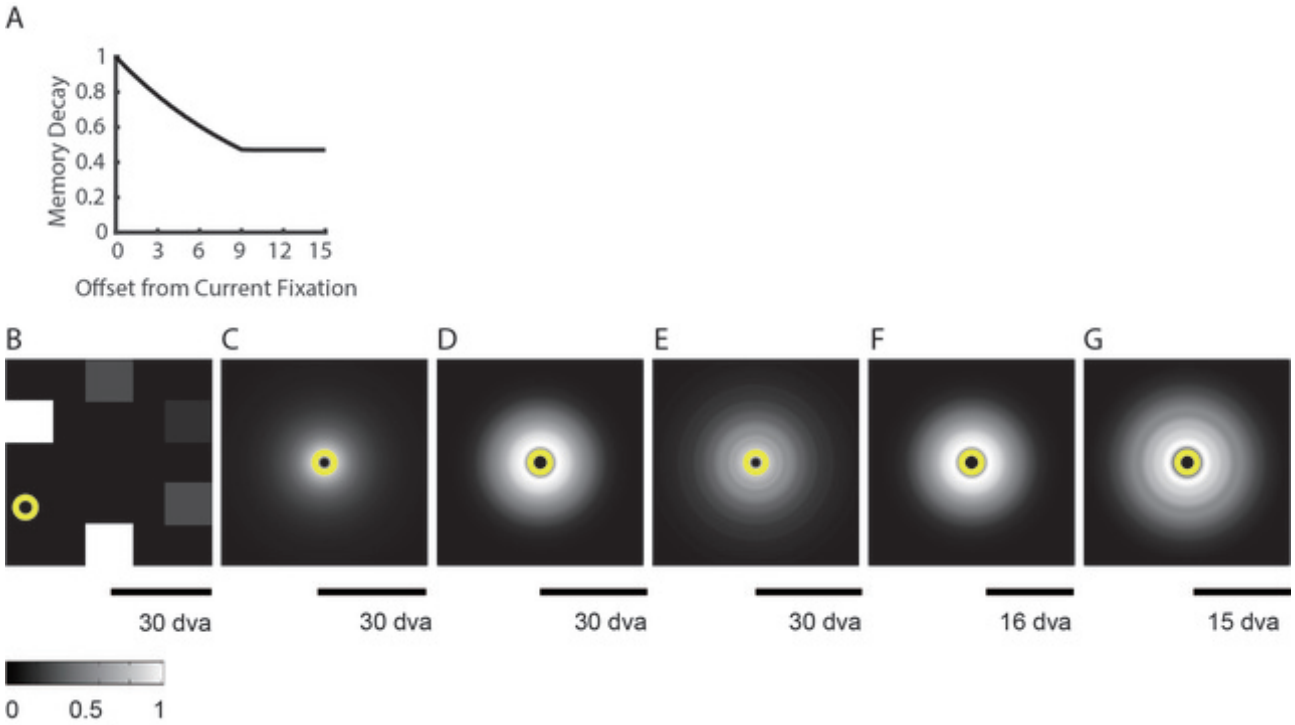

Figure S6. **Memory Decay Function and 2D Empirical Distribution of saccade sizes**. **A**. The memory decay is a function of the offset from the current fixation (see **Methods**). (**B-G**). Saccade size prior maps (M_{sac}) in Visual Search 1 on object arrays (**B**), Visual Search 2 in natural images (**C**), Visual Search 3 in Waldo images (**D**), Free Viewing in humans (**E**), and Free Viewing 1 and 2 in monkeys (**F-G**). The map activation color scale is shown on the bottom left. In all cases, we show the map assuming fixation in the center (yellow circle), except in **B** where the map is shown assuming fixation in the lower left (yellow circle).

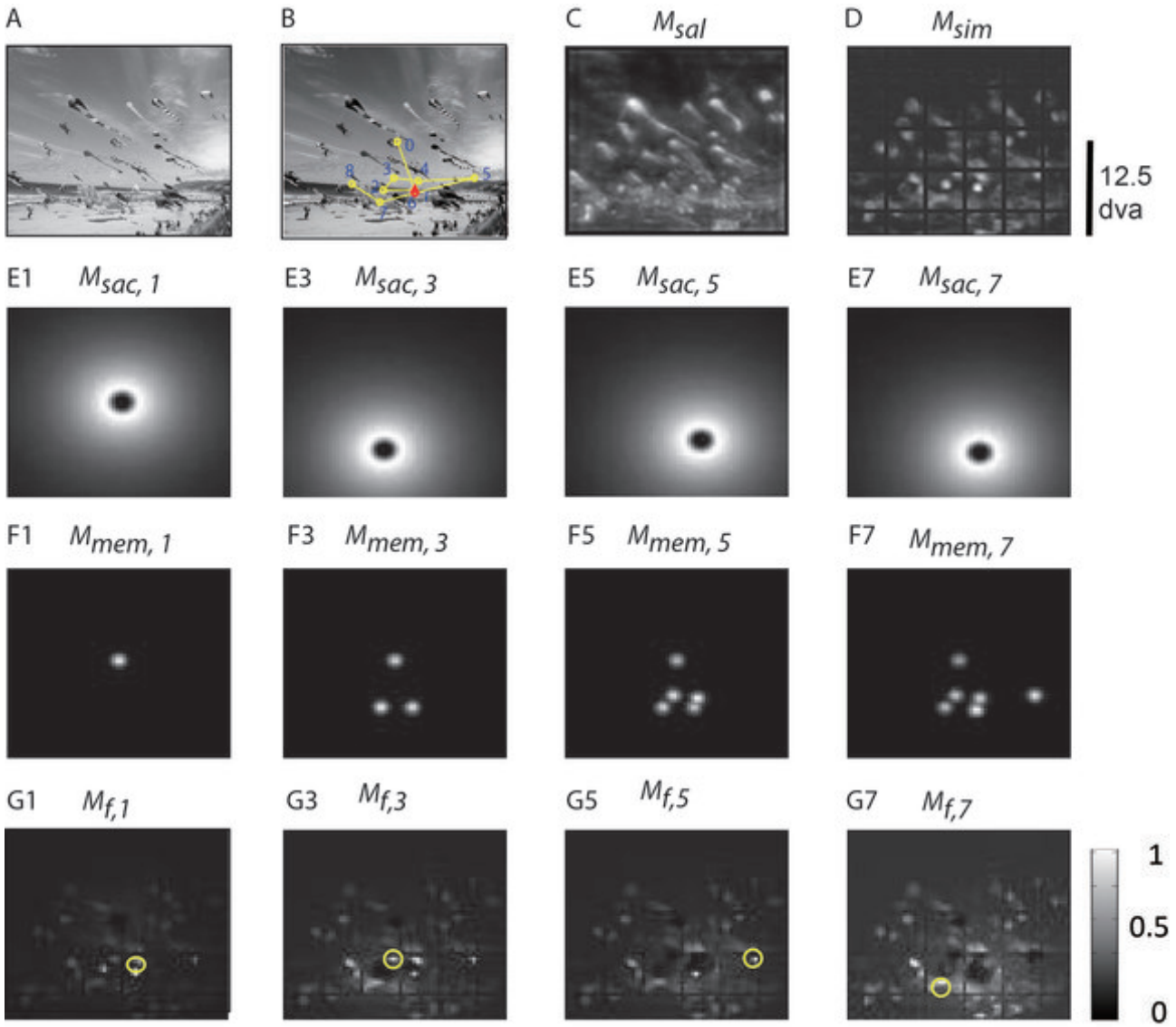

Figure S7. **Visualization examples of attention maps for the model.** (**A**) Example image. (**B**) Pattern of fixations predicted by the model. The plot follows the conventions in Figure 1. (**C**) Saliency map (M_{sal}). (**D**) Similarity map (M_{sim}). ((**E**)) Saccade map (M_{sac}) for fixations 1, 3, 5 and 7. (**F**) Memory map (M_{map}) for fixations 1, 3, 5 and 7. (**G**) Attention map (M_f) for fixations 1, 3, 5 and 7. The yellow circle indicates the maximum, which corresponds to the fixation location.

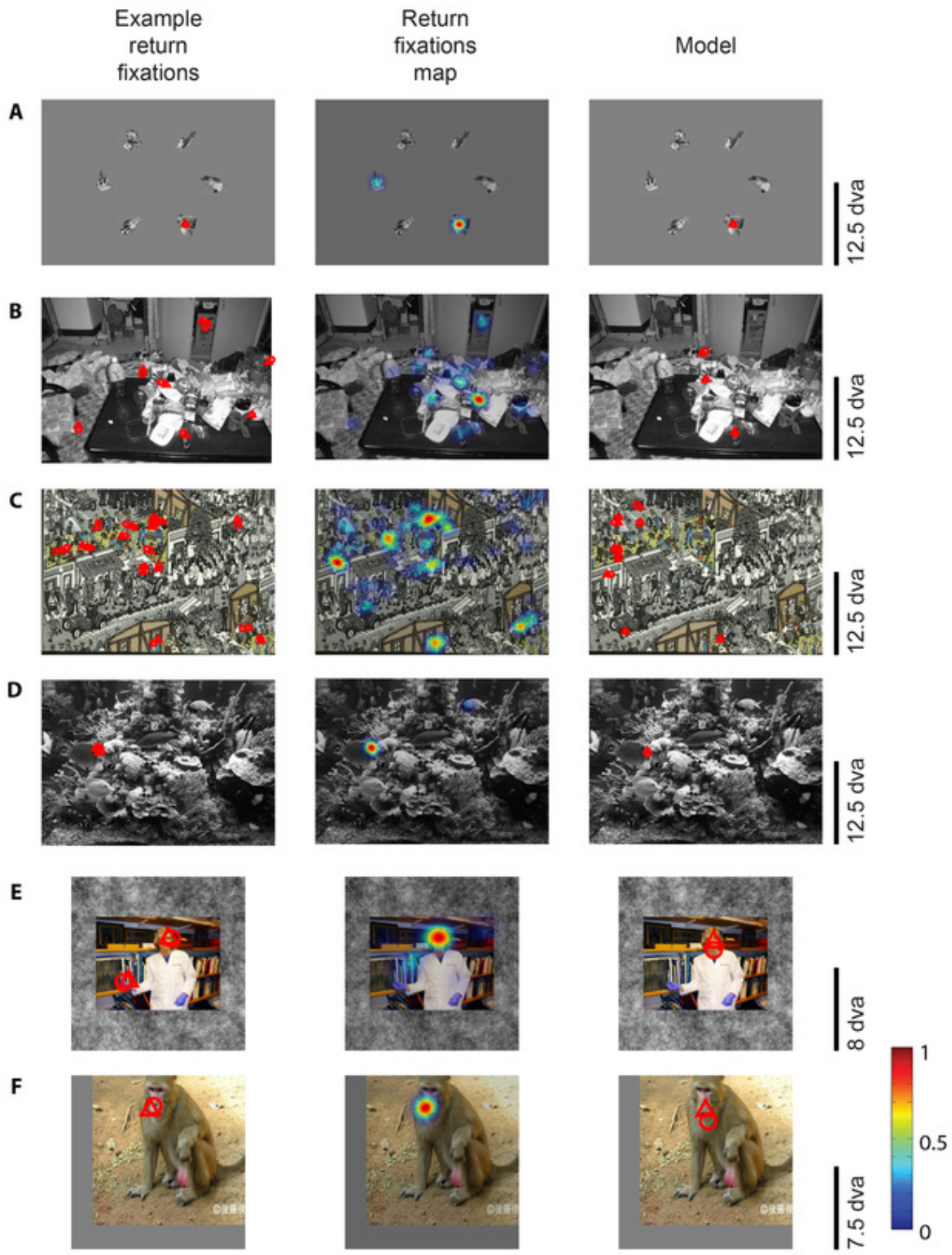

Figure S8. **Example return fixations and model predictions**. Red circles = to-be-revisited locations; Red triangles = return fixations. Middle column shows return fixation probability map across all subjects (see color scale on bottom right).

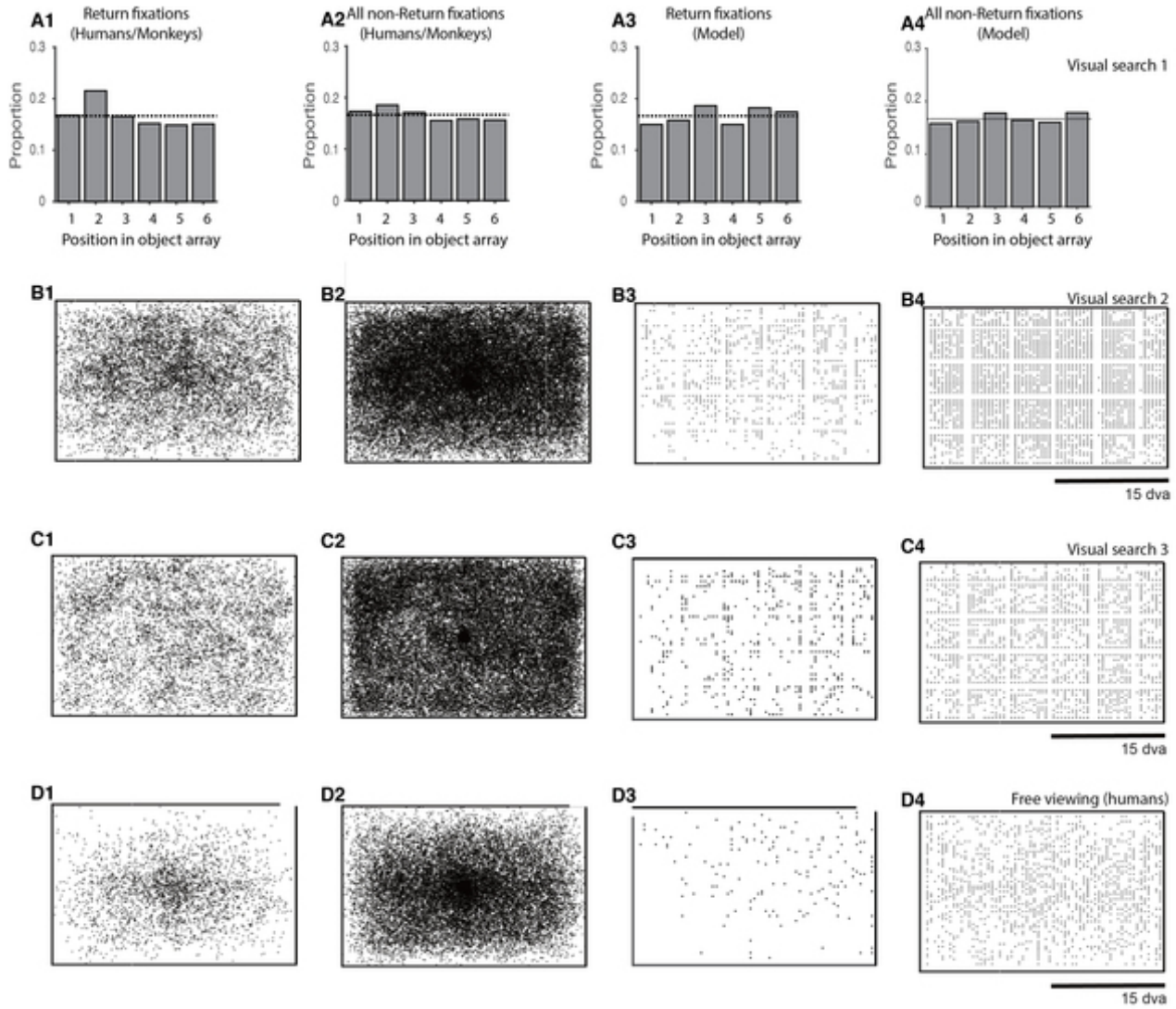

Figure S9. **A. Return and non-return fixation locations**. Locations of return fixations (Column 1) and non-return fixations (Column 2) for each of the experiments for humans/monkeys (**A12-H12**) and the model (**A34-F34**). Each dot indicates the location of a return/non-return fixation.

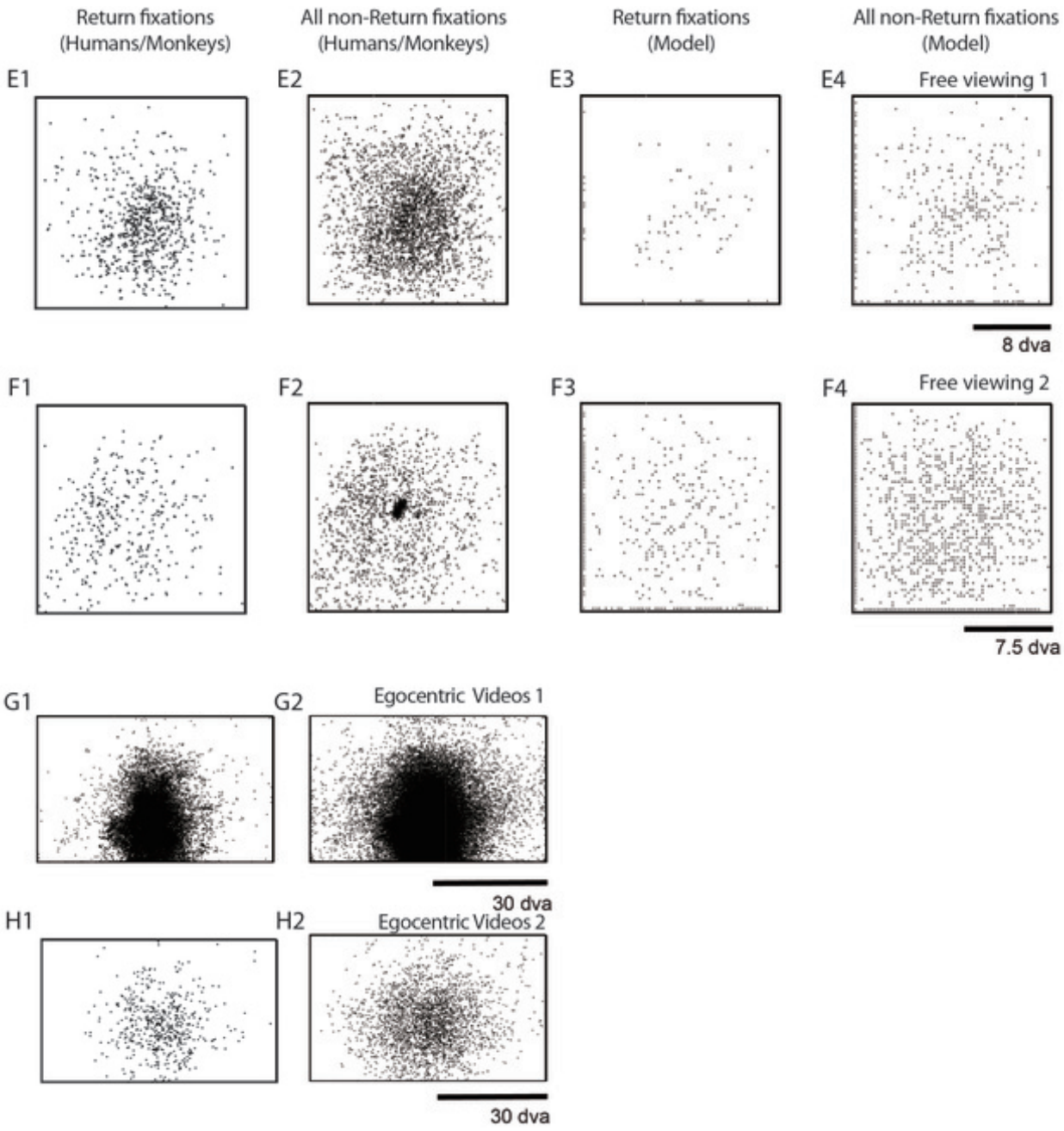

Figure S9. **B. Return and non-return fixation locations**. Locations of return fixations (Column 1) and non-return fixations (Column 2) for each of the experiments for humans/monkeys (**A12-H12**) and the model (**A34-F34**). Each dot indicates the location of a return/non-return fixation.

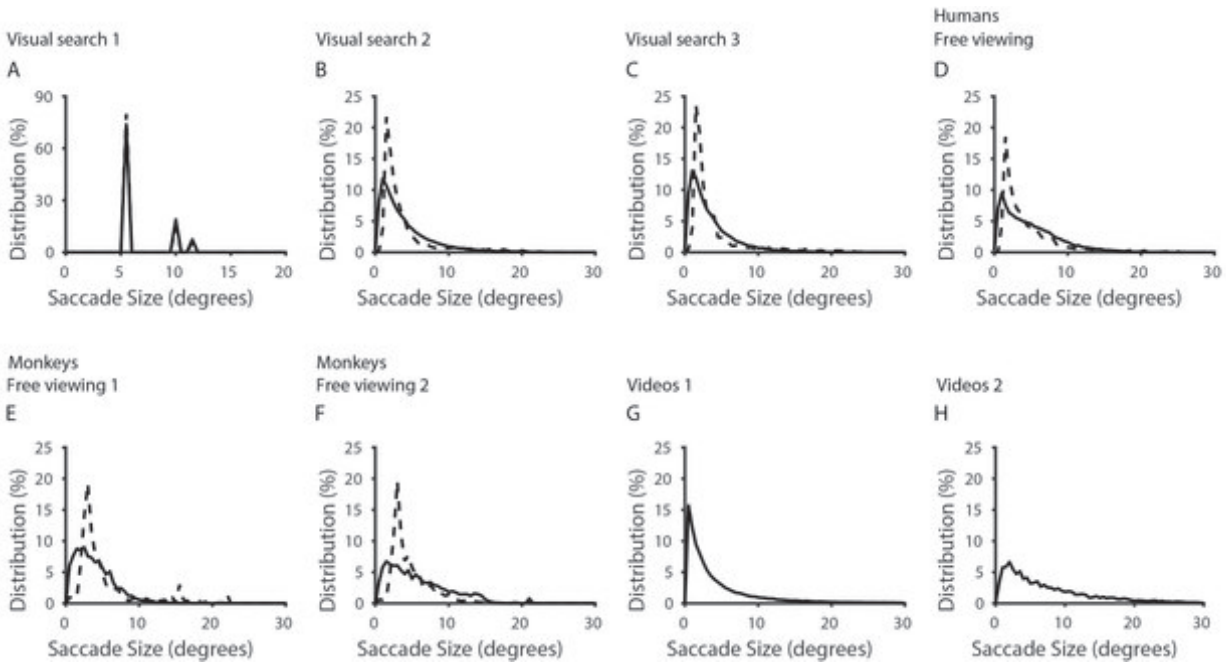

Figure S10. **The distribution of saccade sizes in the model matches the one from humans and monkeys**. Distribution of saccade sizes for humans/monkeys (solid lines) and the model (dashed lines).

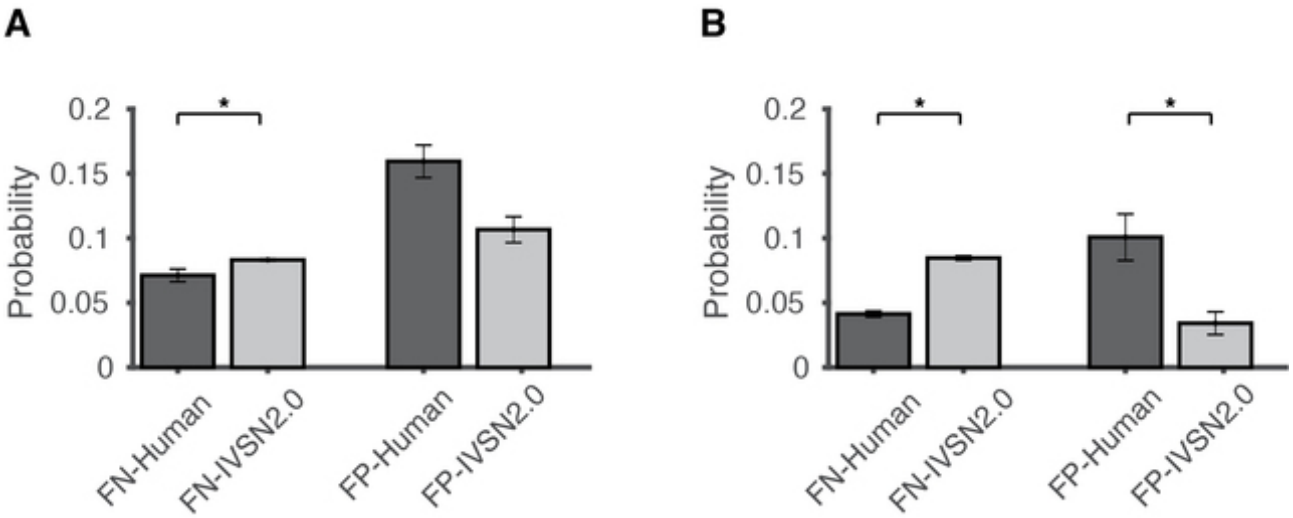

Figure S11. **False negative and positive rates for humans and the model (IVSN2.0) in Visual Search 2 on natural images (A) and in Visual Search 3 on Waldo images (B)**. In visual search tasks, IVSN2.0 has to decide whether the current fixation patch is the target. We introduced the recognition map indicating the confidence value of recognizing that fixation patch belongs to the target. See **Figure S5B** for recognition map M_{recog} calculation. We empirically set a threshold for deciding whether the target is found based on the corresponding confidence value obtained from M_{recog} at current location l_t. The false positive rates for humans are defined as the number of mouse clicks at the wrong locations normalized by total number of mouse clicks over all the trials. The false negative rates for humans are defined as the proportion of number of fixations falling on the targets and yet, humans do not recognize the targets and continue to move their eyes to other locations out of total number of fixations over all the trials. For IVSN2.0, since the fixations and the mouse clicks are always consistent, we calculated both false positive rates and false negative rates based on the empirical thresholds and normalized them based on total number of fixations.

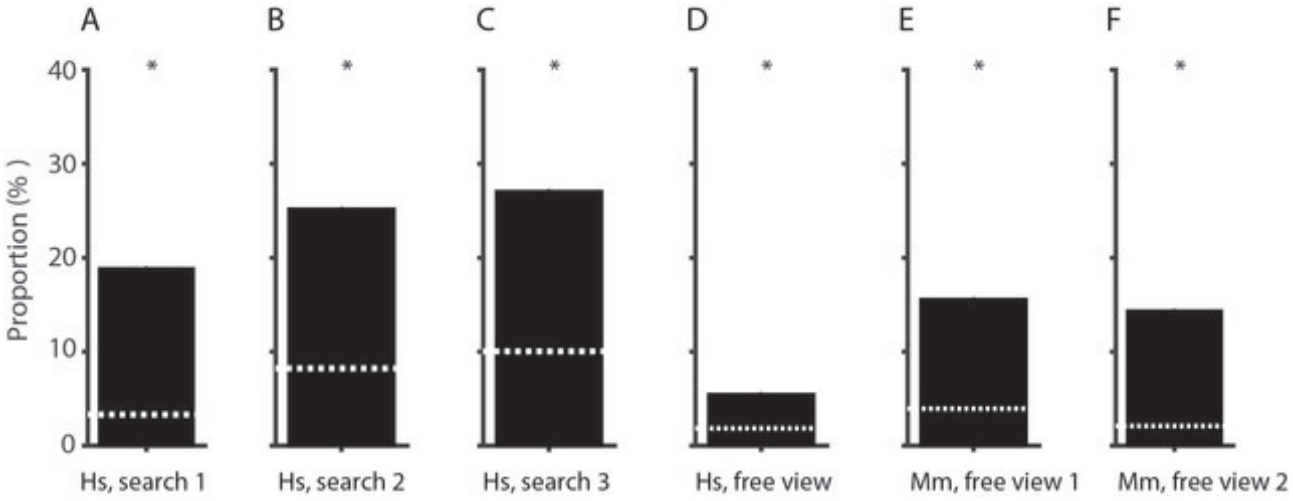

Figure S12. **The model makes frequent return fixations**. Following the format in **Figure 3**, this figure shows the proportion of return fixations in each static image dataset for the model.

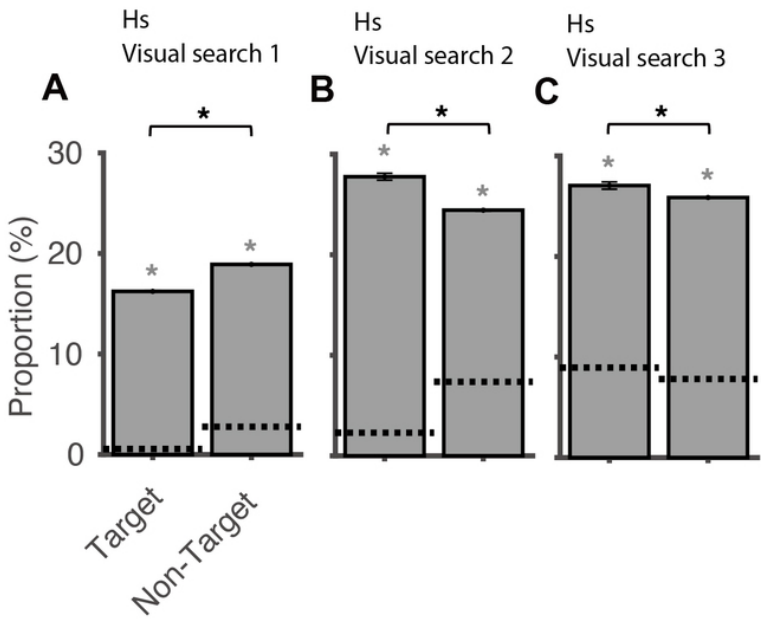

Figure S13. **The model makes more return fixations at target locations than non-target locations in visual search**. Following the format in **Figure 5B**, this figure shows the proportion of return fixations in each visual search dataset at target and non-target locations for the model. Asterisks denote statistical significance above chance ($*$ denotes $p < 0.05$, two-tailed t-test).

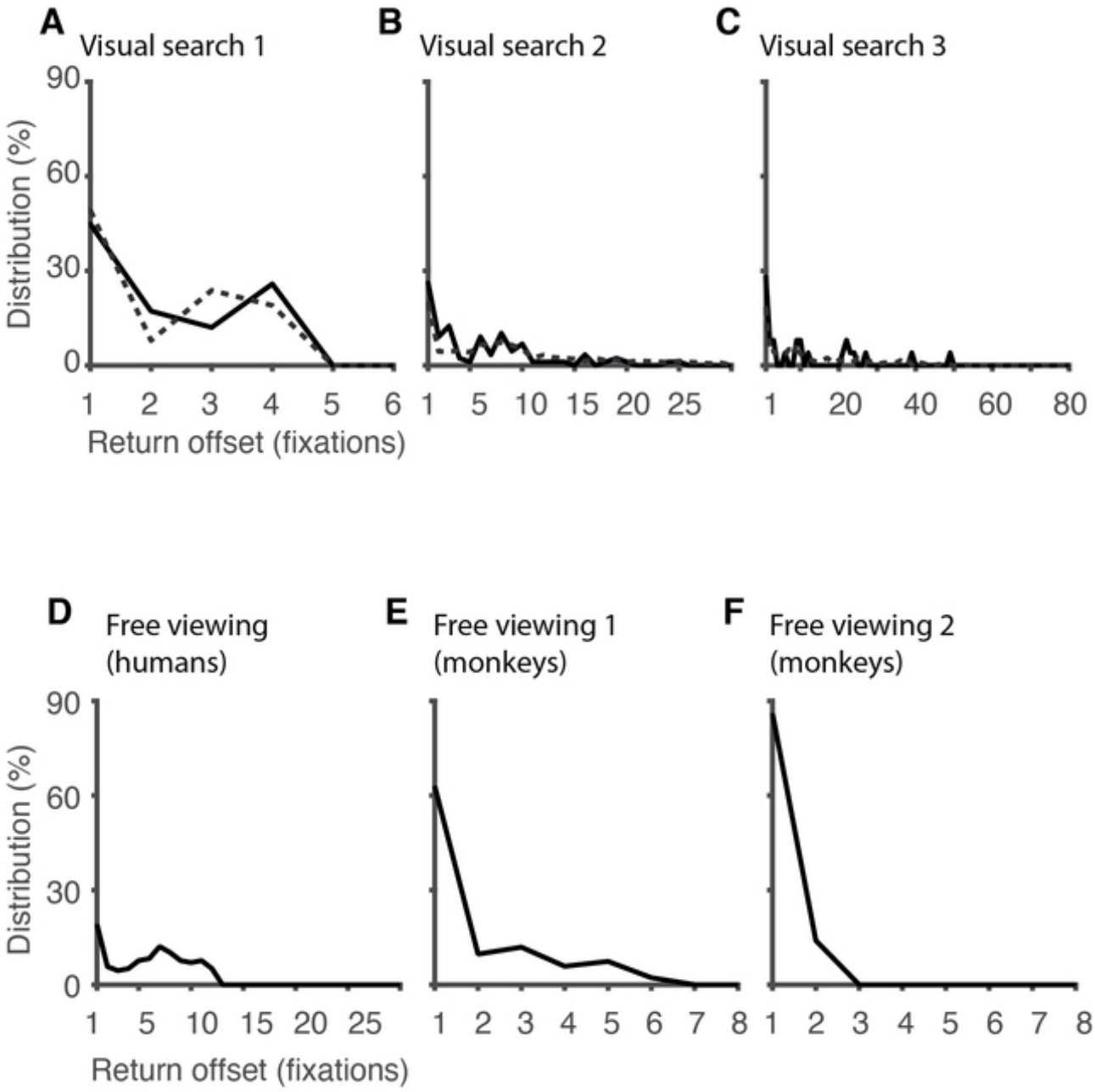

Figure S14. **Return fixations show small return offsets for the model**. Following the format of **Figure 4B**, we show distribution of return fixation offsets for the computational model. In the visual search experiments (**A, B, C**), the solid line shows return fixations to target locations and the dashed line shows return fixations to non-target locations.

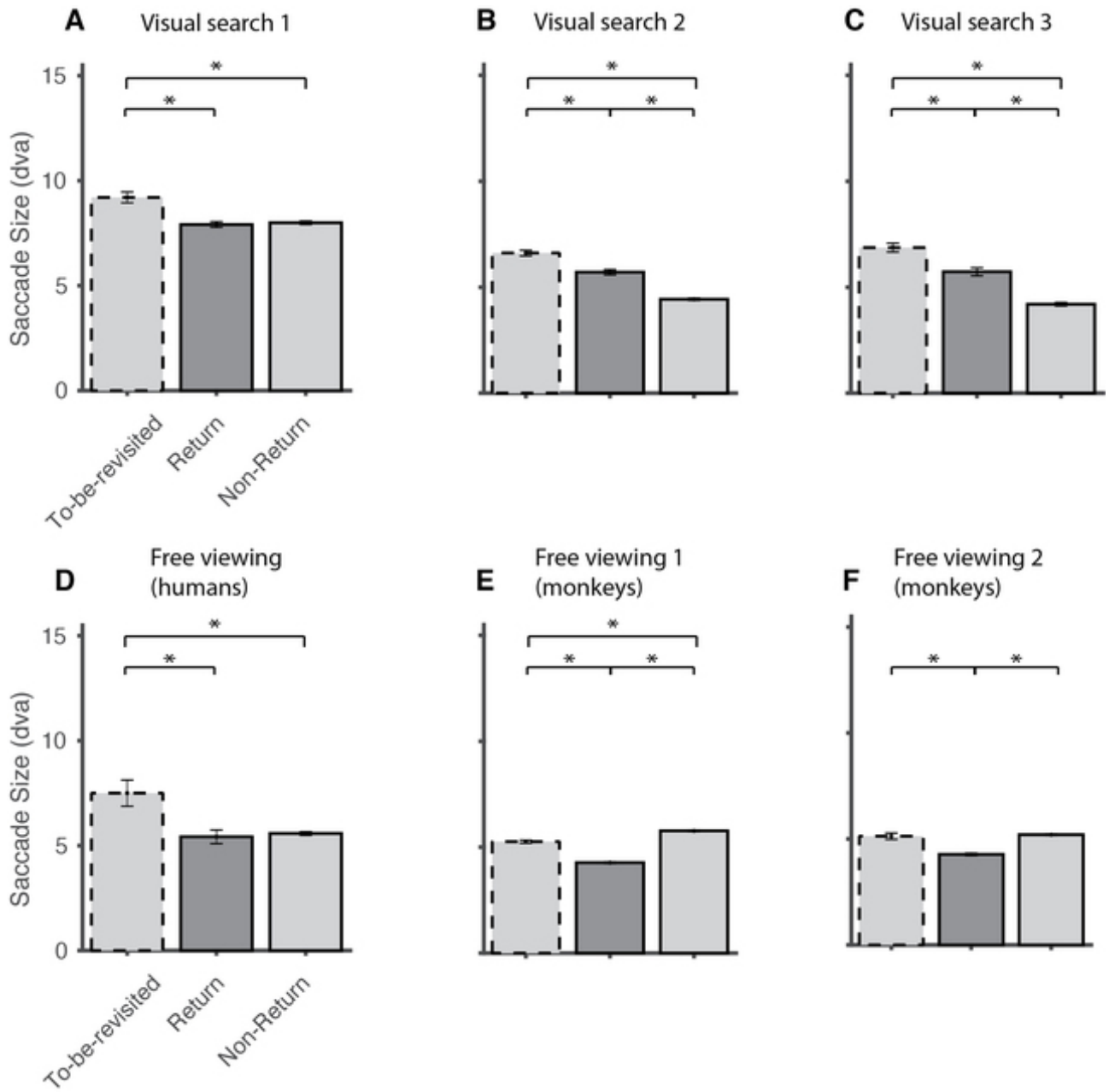

Figure S15. **The model makes shorter saccades at return fixations than non-return fixations**. Following the format of **Figure 4D**, bar plots show saccade sizes of to-be-reivisted, return, and non-return fixations for the computational model over all static image datasets. Asterisks denote statistical significance ($*$ denotes $p < 0.05$, two-tailed t-test).

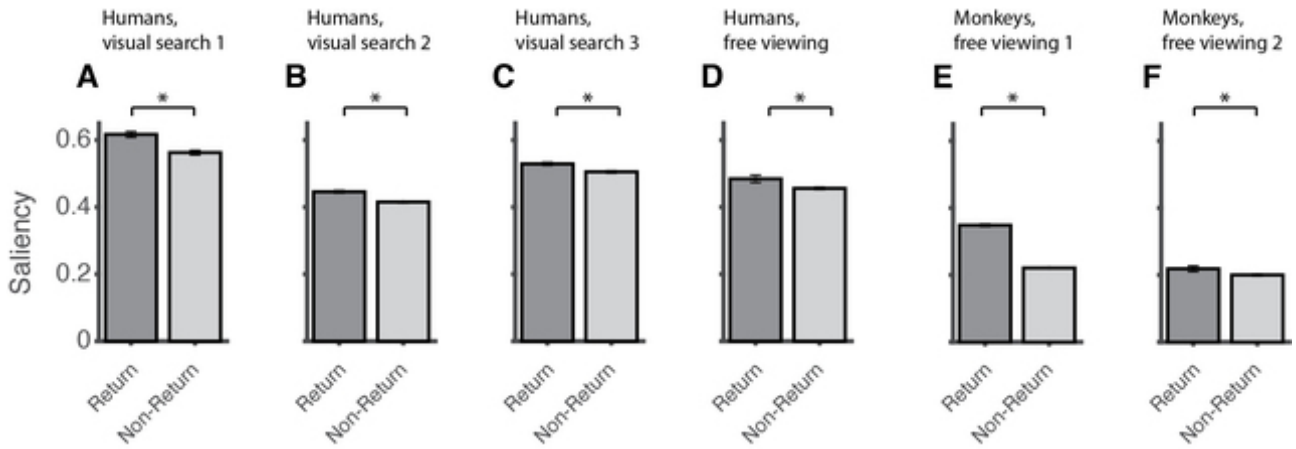

Figure S16. **Return fixation locations in the model showed higher saliency**. Following the format in **Figure 5A**, saliency for return and non-return fixations for the model in each of the static image datasets. Asterisks denote statistical significance ($*$ denotes $p < 0.05$, two-tailed t-test).

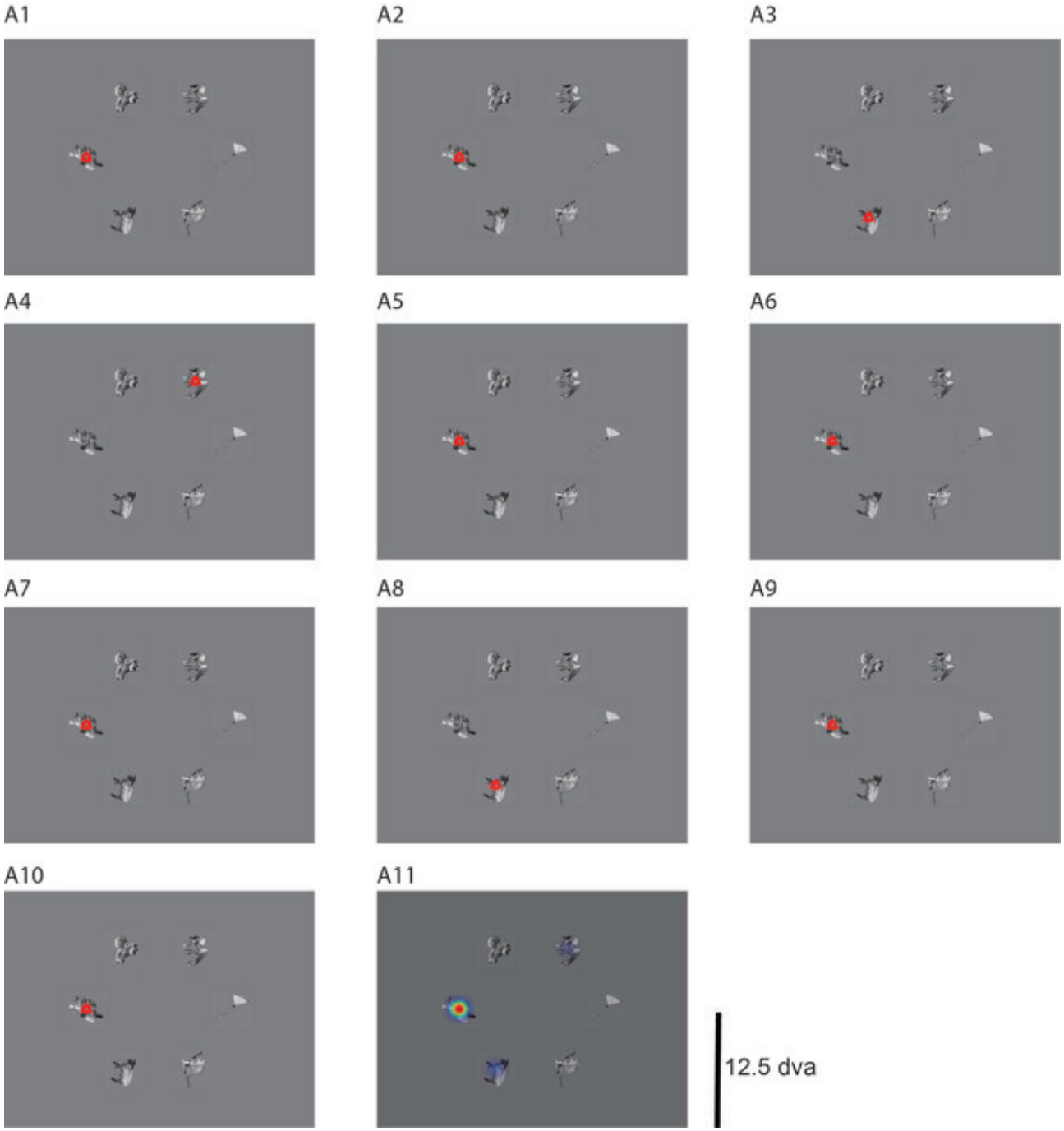

Figure S17. **A. Example of consistent return fixations across subjects, Visual Search 1.**

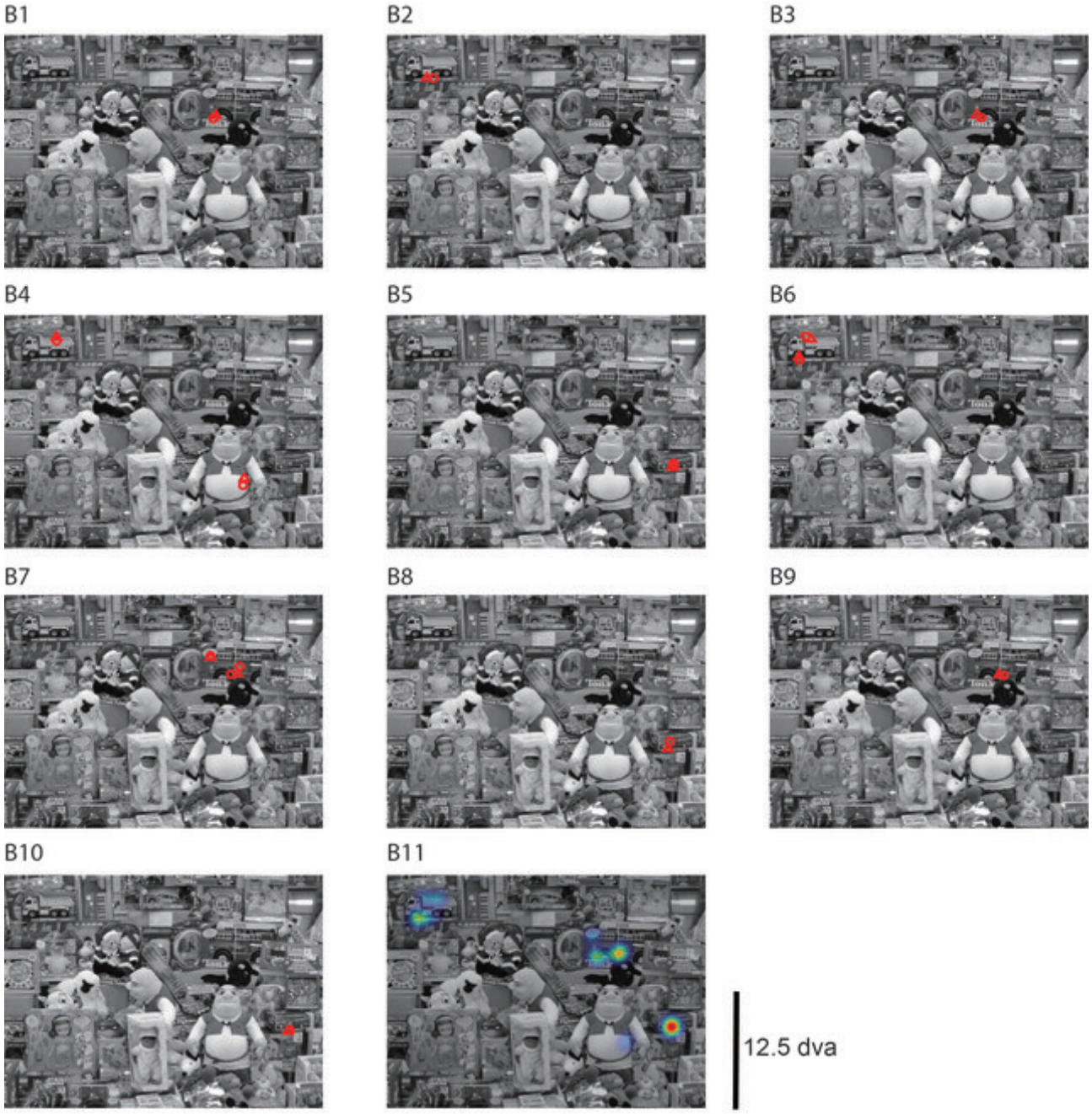

Figure S17. **B. Example of consistent return fixations across subjects, Visual Search 2.**

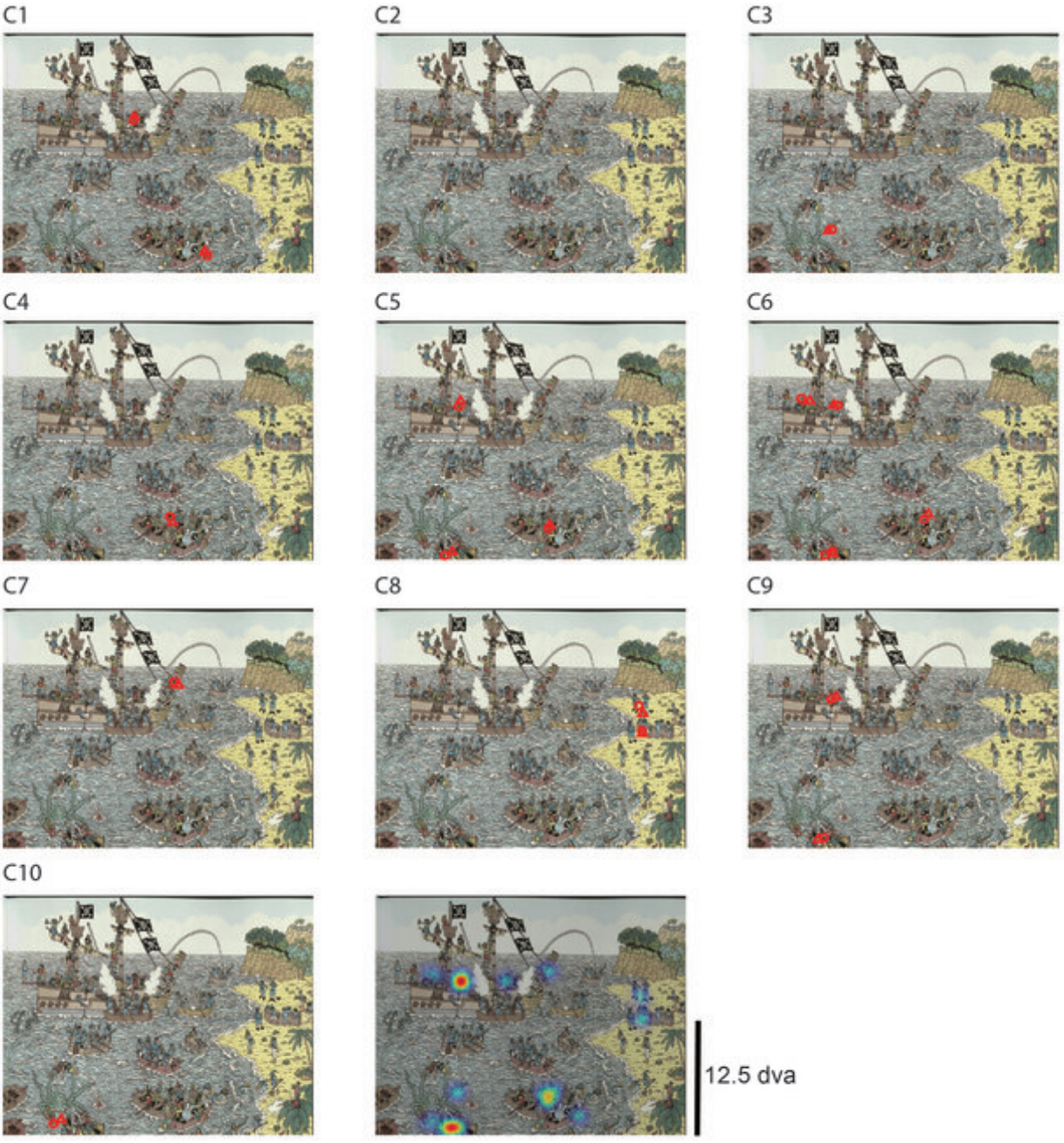

Figure S17. **C. Example of consistent return fixations across subjects, visual search 3.**

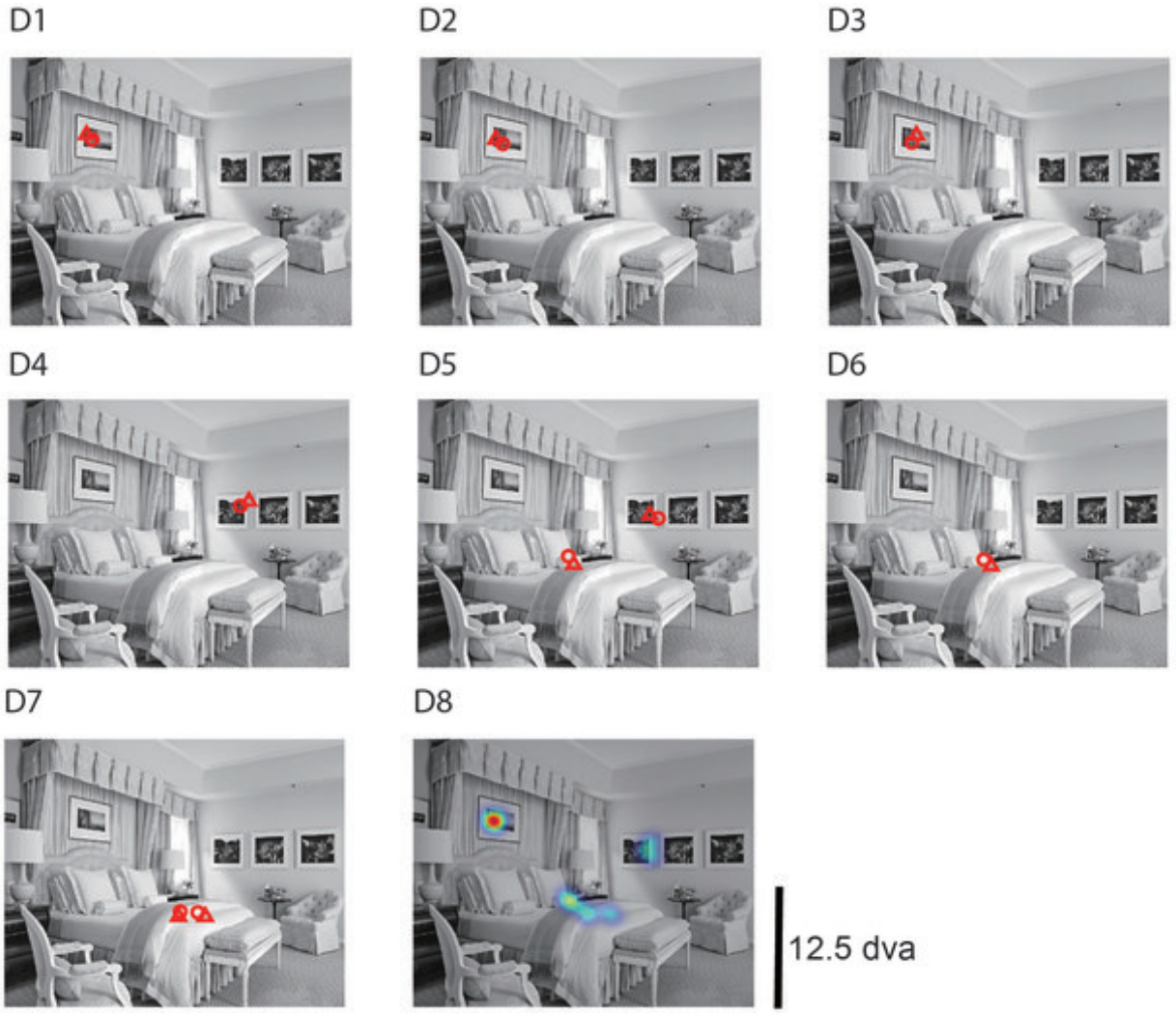

Figure S17. **D. Example of consistent return fixations across subjects, free viewing (humans).**

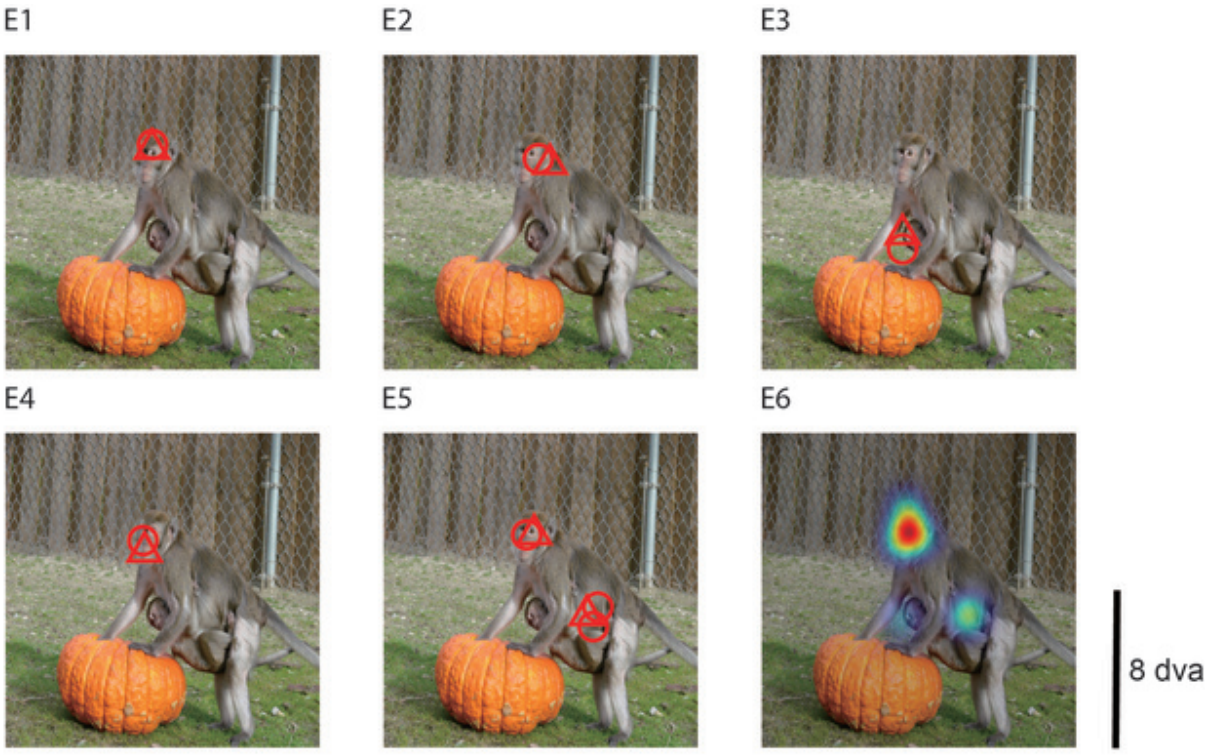

Figure S17. **E. Example of consistent return fixations across subjects, free viewing 1 (monkeys).**

Figure S17. **F. Example of consistent return fixations across subjects, free viewing 2 (monkeys).**