Emergent brain-like representations in a goal-directed neural network model of visual search

Motahareh Pourrahimi[1,2,3], Pouya Bashivan[1,2,3,4]

[1] Integrated Program in Neuroscience, McGill University, Montréal, Canada
[2] Mila, Université de Montréal, Montréal, Canada
[3] Department of Physiology, McGill University, Montréal, Canada
[4] Department of Computer Science, McGill University, Montréal, Canada

June 7, 2025

Abstract

Visual search, the act of locating a target among distractors, is a fundamental cognitive behavior and a core paradigm for studying visual attention. While its behavioral properties are well characterized in humans and non-human primates, the underlying neural mechanisms remain largely unspecified. To address this gap, we developed a biologically aligned neural network model trained to perform visual search directly from pixels in natural scenes. This model exhibits strong generalization to novel scenes and objects, produces human-like scanpaths, and replicates previously known behavioral biases in humans. By analyzing the model's internal representations, we found that it naturally develops a retinocentric cue-similarity map and prospective fixation signals–features that closely resemble neural activity in the primate fronto-parietal network. Beyond reproducing known behavior and neural signatures, the model makes testable predictions about the geometry and dynamics of internal representations underpinning cue-driven prioritization, fixation preferences, their perspective memories, and prospective plans. These findings offer a computational framework for understanding visual search and a roadmap for future neurophysiological and behavioral studies.

Introduction

Visual search, the process of locating a target option among distractors in a visual scene Duncan and Humphreys (1989); Wolfe (2018); Chan and Hayward (2013), underpins countless daily tasks; Whether we are rummaging through a messy desk to find our car keys, scanning a crowded cafe for a friend, or simply picking out the salt shaker at the dinner table Eckstein (2011); Wolfe (2021), we constantly rely on this ability to guide our actions. Because of its ubiquity and importance, visual search has been extensively studied in psychology, cognitive science, machine learning, and neuroscience, both as a core cognitive paradigm and as a building block in numerous primate behaviors Treisman and Gelade (1980); Gelade (2001); Wolfe (1994); Duncan and Humphreys (1989); Eckstein (2011); Mitroff and Biggs (2014); Chan and Hayward (2013); Deco and Heinke (2007); Malcolm and Henderson (2009); Wolfe and Horowitz (2004); Egeth and Yantis (1997); Yang et al. (2017); Li et al. (2020); Yuan and Li (2020); Itti and Koch (2001); Mazer and Gallant (2003); Ischebeck et al. (2024); Desimone et al. (1995); Chelazzi et al. (1993); Wolfe and Horowitz (2017); Kristjánsson and Draschkow (2021); Bella-Fernández et al. (2022).

We search for things by moving our eyes because of the inherent limitations of our visual system. First, our eyes can only see a limited part of our surroundings at every moment, our visual field Findlay and Gilchrist (2003). Second, we cannot see everything in detail within our visual field. The visual acuity decreases drastically with increasing eccentricity from the center of the gaze Carrasco (2011); Findlay and Gilchrist (2003). This happens because the center of gaze is projected onto the fovea, a

small region on the retina with a high density of photoreceptors. However, as we move away from the fovea and into the periphery, the density of photoreceptors decreases rapidly, resulting in a decline in the resolution of the projected image on the retina Rossi and Roorda (2010); Watson (2014); Curcio et al. (1990); Wandell (1995). This sampling process serves as the crucial interface between the brain and the external world, enabling the efficient allocation of limited neural resources for perception and action within the complex and high-dimensional surrounding world. In turn, these properties of the visual system shape the strategy of active vision.

Over the years, several computational models have been proposed to replicate the human search scanpaths (the spatiotemporal sequence of fixations), a central challenge for models of overt visual search. Broadly, these models could be grouped into data-driven Yang et al. (2024); Mondal et al. (2023); Yang et al. (2022); Travi et al. (2022); Yang et al. (2020); Zelinsky et al. (2019) and mechanistic models Zhang et al. (2018); Rashidi et al. (2023); Ding et al. (2022); Zhang et al. (2022); Bujia et al. (2022); Gupta et al. (2021); Adeli and Zelinsky (2018); Adeli et al. (2017); Wei et al. (2016); Zelinsky et al. (2013); Pomplun and Gray (2007); Rao et al. (2002); Zelinsky (2008); Navalpakkam and Itti (2006, 2005). Data-driven models typically consist of parametrized functions that are fitted to replicate the human scanpaths from eye tracking data, without an explicit intention to reveal the underlying neural mechanisms of visual search Yang et al. (2024); Mondal et al. (2023); Yang et al. (2022); Travi et al. (2022); Yang et al. (2020); Zelinsky et al. (2019). Despite their strong performance in predicting human behavior, it remains unknown whether and how their components and learned representations correspond to brain networks involved in visual search.

On the other hand, mechanistic models typically consist of rule-based or parameterized models driven by specific hypotheses about how the brain performs visual search. A prominent class of these models relies on a cue-similarity map, a map that represents the similarity of each location within the scene to the cue object Zhang et al. (2018); Ding et al. (2022); Zhang et al. (2022); Gupta et al. (2021); Adeli et al. (2017); Pomplun and Gray (2007); Zelinsky (2008); Rao et al. (2002); Navalpakkam and Itti (2006, 2005). To generate the cue-similarity map, the visual representation of the cue is compared to different locations in the visual scene using feature-based comparisons or template matching using convolution or correlation. Each location in the resulting map reflects its match to the cue. Fixations are typically selected by a winner-take-all mechanism that chooses the point with the highest similarity. However, alternative methods such as population averaging or spatial optimization have also been employed Rao et al. (2002).

Earlier mechanistic models utilized manually-defined feature dimensions or relied on simple measures of similarities such as template matching Rao et al. (2002); Pomplun and Gray (2007); Zelinsky (2008); Navalpakkam and Itti (2006, 2005); Wolfe (2021), and as a result, were unable to perform more complex search tasks such as categorical search that requires invariance to category-orthogonal object properties. More recent models addressed these limitations by leveraging pre-trained deep vision models, including convolutional neural networks (CNNs) Zhang et al. (2018); Gupta et al. (2021) and vision transformers Ding et al. (2022). These models provide rich, generalizable visual representations, enabling them to predict human scanpaths across diverse visual search tasks and to zero-shot transfer to novel objects.

While these models have successfully replicated certain aspects of human search behavior, they suffer from multiple shortcomings. First, most of them consist of rule-based mechanisms that are not grounded in biologically-plausible neural architectures. As a result, these models are unable to explicitly inform the role of different brain areas in executing visual search behavior as they lack distributed representations that can be compared with the function and response properties of neurons in the brain. Notably, none of these prior models are image-computable. Although Zhang et al. (2018); Ding et al. (2022) incorporate image-computable vision modules, they replace the neural computations beyond sensory visual areas with a symbolic rule-based algorithm. Consequently, they do not offer any concrete predictions about the neural responses in the higher-order areas such as the frontal eye fields (FEF), lateral intraparietal area (LIP), ventral prearcuate area (VPA), and the superior colliculus (SC) during visual search. Moreover, as these models are engineered specifically for solving the visual search problem, they lack a general architecture that can extend to other cognitive tasks.

Second, to prevent the model from getting trapped on the same location, these models often incorporate a mechanism known as inhibition of return (IOR), which discourages revisiting previously attended locations and encourages the exploration of new areas during successive saccades Zhang et al. (2018);

Gupta et al. (2021); Zelinsky (2008); Pomplun and Gray (2007). While incorporating IOR helps prevent repeated fixations, the technical need for such a mechanism highlights a limitation of this modeling framework: the absence of a dynamic internal representation that integrates information from partial observations across successive fixations. Engineering an IOR mechanism to avoid getting stuck at the same location is not limited to mechanistic models and is a widely adopted practice in data-driven models of search as well. However, growing evidence suggests that primates do revisit previously fixated locations, albeit with lower probability Smith and Henderson (2009); Zhang et al. (2022); Wilming et al. (2013). Therefore, a question remains: whether a model solely trained to perform the task with no such constraints would learn human-like inhibition of return.

Third, most mechanistic and data-driven models assume uniform access to the entire visual field, an unrealistic and biologically implausible simplification. Eccentricity-dependent acuity is a key factor shaping fixation selection Lleras et al. (2022) and, by extension, eye movement behavior. Ignoring this aspect can therefore impair a model's ability to replicate primate visual search. Recognizing this limitation, some models attempted to approximate the effects of nonuniform eccentricity-dependent visual acuity, for example, by applying Gaussian blurring centered at the fixation location Zelinsky (2008); Pomplun and Gray (2007); Yuan and Li (2020) or by implementing eccentricity-dependent average pooling on the intermediate layers of the visual network Gupta et al. (2021). While these techniques represent a step toward greater biological realism, they fall short in capturing a critical function of retinal transformation: reducing the high-dimensional visual input from the external world into a compressed, lower-dimensional representation centered on the current fixation Curcio et al. (1990); Iyer and Burge (2019).

These limitations underscore the need for models that are both biologically grounded and capable of performing the task from first principles. To this end, we 1) used a network architecture capable of integrating information across fixations through a dynamically updated internal state; 2) trained the model solely on the task of performing visual search without assuming any symbolic rules and computations; 3) implemented eccentricity-dependent visual acuity similar to that in primates via eccentricity-dependent downsampling of the input to simulate the sensory limitations of peripheral vision. Importantly, the model was not hand-engineered for visual search, making it flexible to extend to other cognitive tasks. Its distributed representation and image-computable nature also enable direct comparisons between model activations and neural activity in the brain. Moreover, as the model is trained to directly perform the visual search task, it does not require large-scale human behavioral data.

We show that after training to categorically search for objects in natural scenes, the model learns a general mechanism for search that generalizes to novel scenes and object categories, exhibiting human-like behavior while relying on neural representations analogous to those observed in the primate brain. Having a dynamic internal neural representation, the latent space of the model, provides us with the opportunity to not only study the brain-similarity of the learned representations but also to make testable predictions about the geometry and dynamics of these latent representations, the computations they support, and possible mechanisms that implement visual search behavior in the brain's neural circuitry.

Results

Goal-driven modeling of visual search in natural scenes

Training large neural networks requires large datasets. In the absence of such datasets for natural visual search, we constructed a large-scale dataset of natural scenes using images from the Places365 dataset Zhou et al. (2017). We applied a state-of-the-art object detector He et al. (2017) to identify instances of 80 object categories, extracting corresponding segmentation masks that served as ground-truth target locations (Fig. 1A). These masks also formed a cue bank from which example images were drawn. Each trial began with the presentation of a cue image, an object exemplar sampled from the target category, followed by a visual scene depicting a natural scene. The visual scene was presented at every time step until the end of the trial. To discretize spatial information, we overlaid a 10×10 grid on both the visual scene and target mask (Fig. 1B). The task design mirrored the psychophysical protocol of the COCO-Search18 dataset Chen et al. (2021), which we used to assess generalization to

novel scenes and alignment with human behavior (Fig. 1C). The COCO-Search18 dataset is derived from the MS COCO dataset Lin et al. (2014) and includes 18 target categories, all contained within our training set.

Our proposed Goal-directed Visual Search Model (GVSM) consists of the following four components: First, a CNN for processing visual information. The CNN was pretrained on object recognition using the ImageNet dataset Deng et al. (2009). Imagenet-trained CNNs have been shown to closely replicate neural activity across the ventral visual cortex Yamins et al. (2014); Ratan Murty et al. (2021); Bashivan et al. (2019); Cadieu et al. (2014); Khaligh-Razavi and Kriegeskorte (2014). Second, a Glimpse Network that combines the perceived information with an efference copy of the fixation action generated by the model Mnih et al. (2014). Third, a sequence processing module (transformer Vaswani et al. (2017)) for integrating information across fixations that assumes the role of the fronto-parietal attentional control network in guiding fixations, along with the fourth component, the Policy network, a multilayer perceptron (MLP) for generating fixation probability maps from the transformer latent space (Fig. 1D). At each time step, the CNN receives a retinal input image, produced by a multi-resolution cropping module that approximates eccentricity-dependent visual acuity, yielding a foveated representation Mnih et al. (2014) (Fig. 1E). The CNN output is then combined with the fixation location through a glimpse network (Methods), forming the glimpse representation that is passed to the Information Integrator module (a causal transformer). Then, the Fixation Selection module (MLP) generates an action probability map of the next fixation location by processing the latest output token of the transformer (the Information Integrator's latent space). The next fixation location is sampled from the generated fixation probability map. This process continues until the model fixates the target (i.e., the predicted fixation location falls within the ground-truth target mask) or the trial reaches the maximum number of time steps (Fig. 1D).

Model parameters are trained to perform visual search using a two-stage training paradigm, inspired by prior work on saccade-augmented visual categorization Elsayed et al. (2019). In the first stage, we use backpropagation to train the Information Integrator (transformer) to predict the target area from a random sequence of fixations over the image (Fig. 2A). In the second stage, we freeze the transformer parameters and train an MLP policy network with reinforcement learning (RL) to select fixation locations (Fig. 2B). During the supervised stage, the transformer learns to integrate information across partial glimpses, accumulating evidence across fixations to localize the target. In the RL stage, the policy network exploits these learned representations, embedded in the transformer's latent space, to decide where to look next to find the target efficiently. The policy network is incentivized to minimize the number of steps by incurring a penalty for each fixation taken without locating the target. Fig. 2C illustrates an example trial performed by the trained model over time. During operation, the transformer output token at each step, i.e., the model's latent state, may serve as a computational analogue of the evolving representations within the fronto-parietal attentional control network.

GVSM replicates human search behavior

We compared the model's generalization ability and its behavioral alignment with humans against three baselines including two prominent mechanistic models of visual search: IVSN and eccNET. These models extract visual representations of the cue frame and the visual scene using pretrained CNNs, and compute a cue-similarity map by convolving the cue representation across the visual scene representation. The resulting similarity map guides fixations through a winner-take-all mechanism, coupled with inhibition of return. IVSN assumes uniformly high resolution across the entire visual field, while eccNET applies eccentricity-dependent average pooling to the intermediate layers of the feature-extracting CNNs to emulate retinal transformations. In addition to these two baselines, we also included a data-driven model of visual search, an inverse reinforcement learning model (IRL), which we treated as a gold-standard baseline due to its training on human eye-tracking data. This model formulates search as an inverse RL problem, using a generator-discriminator architecture to learn to generate reward maps over the input images conditioned on the cue category, by generating fixation sequences that match those from human subjects.

GVSM generalizes to novel scenes and objects. Humans can flexibly generalize their visual search behavior to any natural scene, containing diverse objects. A general model of visual search is naturally expected to not merely memorize training examples but to generalize to novel scenes and

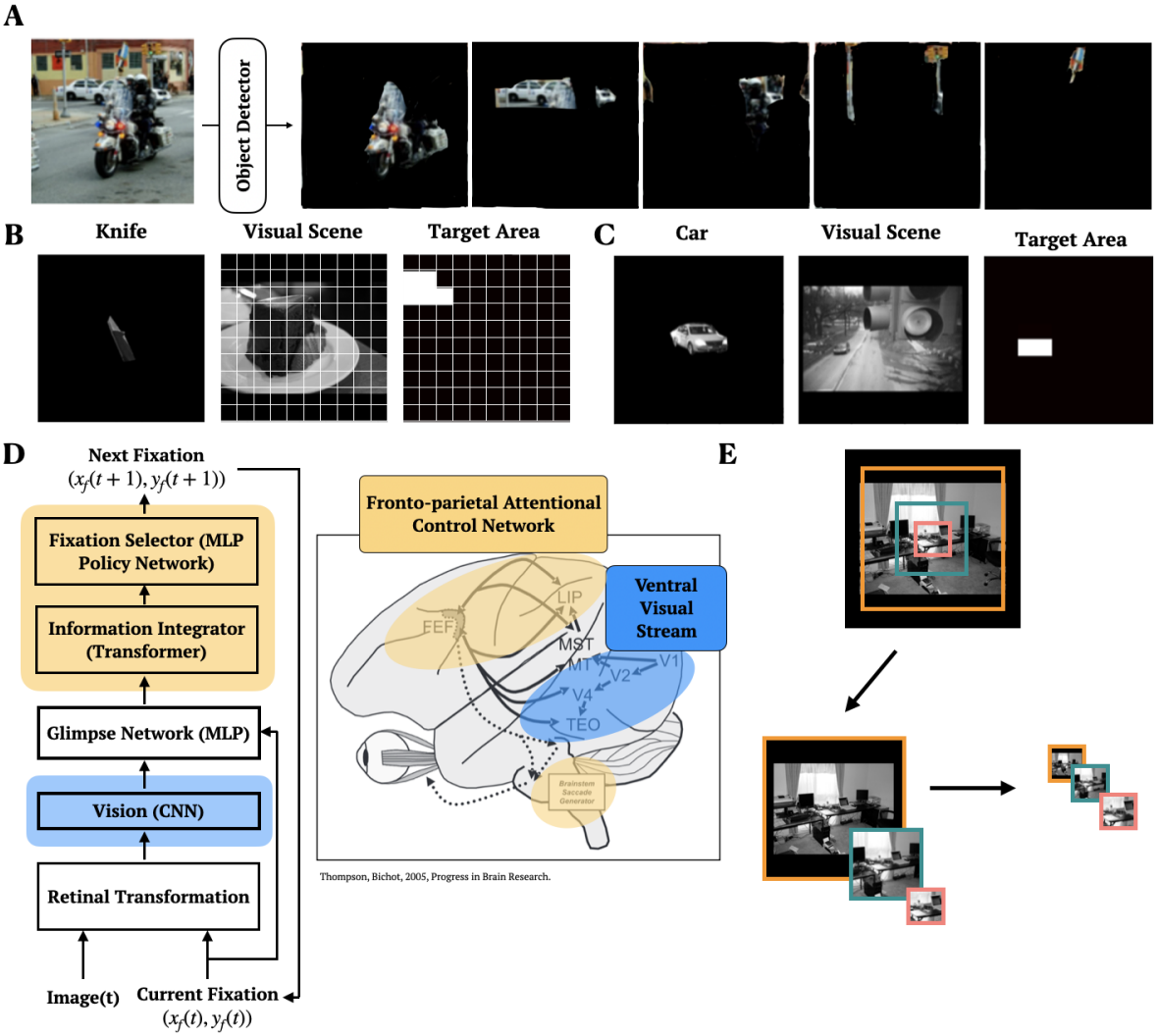

Figure 1. Behavioral task and model architecture. A. Dataset generation pipeline. We used a state-of-the-art object detector to extract object masks for 80 categories (e.g., motorcycle, human, car, traffic light, kite) from natural scene images. Extracted object instances were stored in a cue bank, serving as visual query examples for the corresponding cue categories. Each image and its associated object masks formed the basis for a visual search trial. **B. Example training trial.** Each training trial consisted of a visual scene (from Places365), a visual cue image (an object from the cued category sampled from the cue bank and placed on a black background), and the corresponding ground-truth target area map. The scene was discretized using a 10×10 grid applied to the image and the target mask. The ground-truth target map contained 1s for grid locations entirely covered by the cued object mask and 0s elsewhere. **C. Example test trial.** Test trials were sampled from the COCO-Search18 dataset. Unlike the original psychophysics protocol that used text-based cues, we employed visual cues generated using the same procedure as in A, with novel cue examples drawn from MS COCO. **D. Model architecture.** At each time step, a pretrained CNN, mimicking the primate visual system, processed a retinal image and generated a visual representation. This was combined with the fixation location via the Glimpse Network to form a glimpse representation, which was passed to the Information Integration module (a causal transformer), followed by a Fixation Selector (an MLP), taking on the role of the fronto-parietal attentional control in guiding fixations. The next fixation location was read out from the latest transformer token using the MLP. The model iterated until it fixated the target or reached the maximum time steps (10). **E. Retinal transformation.** Biological retinal transformation was approximated using a multi-resolution crop module. Three crops (128×128, 256×256, 512×512) are extracted from a 640×640 image, centered at the current fixation location, and resized to the smallest crop size. This process mimics the progressive peripheral blurring observed in the primate retina.

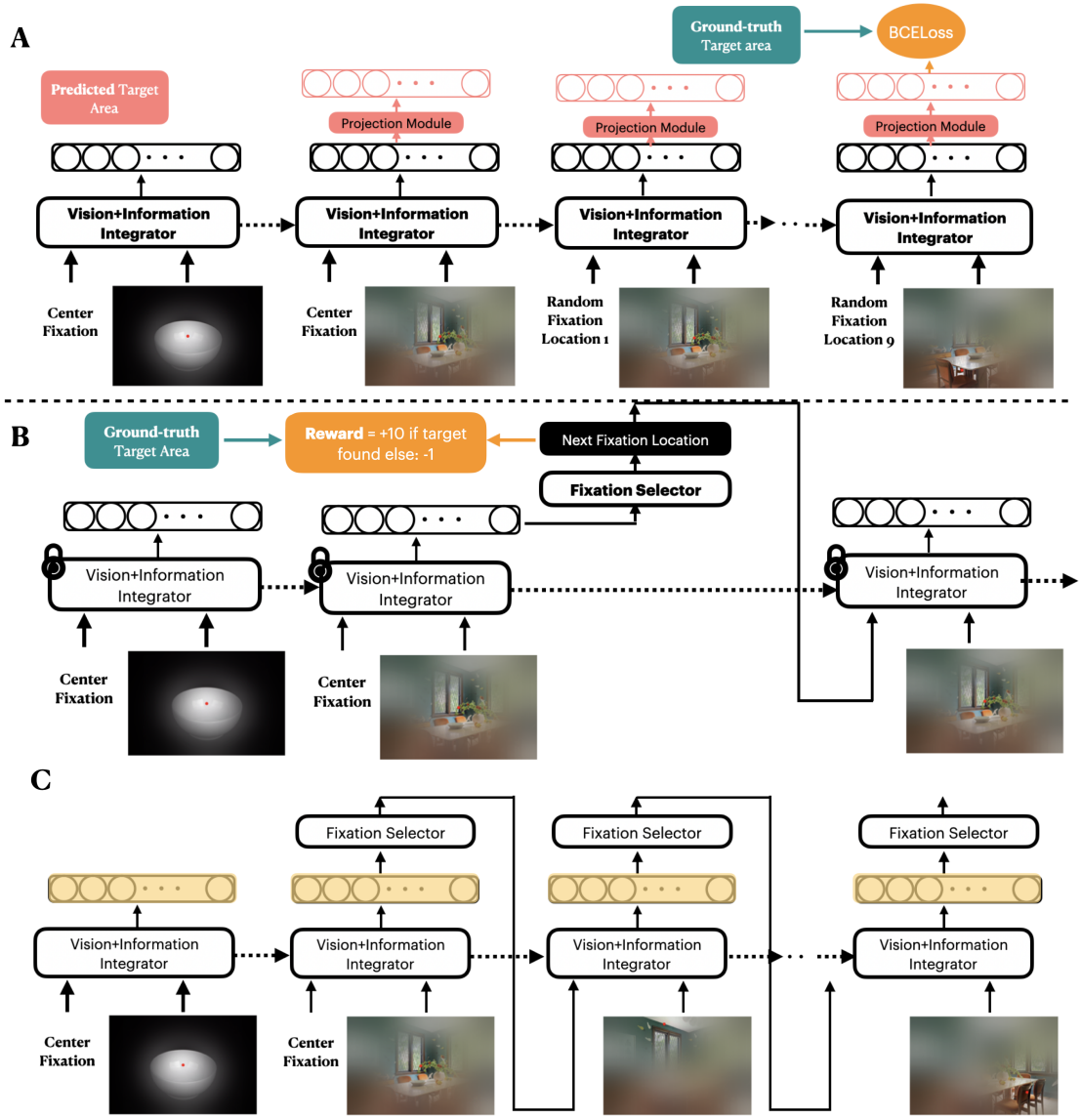

Figure 2. Model training. A. First stage of model training. We trained the parameters of the Information Integrator (the transformer) and a linear readout using backpropagation to output the target area map at each step, given the observations obtained through a random sequence of fixations over the image. The retinal transformed images depict a schematic of a progressively blurred image and do not accurately reflect the multi-resolution crop retinal transformation used in the modeling as described above. **B. Second stage of model training.** During this stage, the information integrator's parameters were fixed. We trained the policy network using RL to generate a 10×10 action probability map from which the next fixation location was sampled. The RL agent received a reward of +10 when its generated fixation location lay on the target and a penalty of -1 for each unsuccessful step. **C. Test-time behavior of the model.** At test time, given a cue and a natural scene, the model searches for the position containing an instance of that object category.

objects , reflecting the flexibility and robustness of human search behavior. We tested the generalization of GVSM on unseen visual scenes of the COCO-Search18 test set (average accuracy=84.61%, S.D.=1.30%), which contained novel cue examples and visual scenes (Fig. 3A). GVSM outperformed all baseline models of natural scene search in visual search generalization (IRL: $t = 68.75$, $p < 0.001$; IVSN: $t = 8.13$, $p < 0.001$) (Fig. 3B). This included the IRL model, which has been explicitly trained on human eye-tracking trajectories from the COCO-Search18 train set and is regarded as a gold standard in this context.

Next, we investigated behavioral alignment of GVSM with humans while searching for objects in natural scenes addressing three critical aspects of human search behavior, starting from the coarser characteristics to more fine-grained patterns: 1) search efficiency, an important high-level characteristic of search behavior, measuring how quickly the target is located; 2) inhibition of return characteristics: humans' substantially low tendency to revisit previously fixated locations Posner et al. (1985); Klein (1988, 2000); 3) trial-by-trial spatiotemporal sequence of fixations, search scanpaths.

Search efficiency alignment. Overall search efficiency was visualized using cumulative performance curves showing the proportion of successful trials as a function of fixation number (Fig. 3B). While a significant performance gap remains between models and humans, GVSM captured the general shape of the human cumulative performance curve and achieved the highest search performance.

We then assessed whether GVSM captured second-order structure in search efficiency, namely, systematic biases linked to trial features such as target size. Humans required a greater number of fixations to locate the target in trials with smaller target sizes (Spearman's $\rho = 0.17$, $p < 0.001$; Fig. 3C). GVSM closely replicated this dependency, showing a comparable correlation between the number of fixations and target size (Spearman's $\rho = 0.19$, $p < 0.001$), best matching the degree of the human dependency compared to the baselines, whereas other models either overestimated the influence of target size on search efficiency (IRL: Spearman's $\rho = 0.37$, $p < 0.001$) or failed to capture it altogether (Fig. 3C).

Emergence of inhibition of return. Previous studies have shown that fixation history plays a key role in guiding eye movements. Specifically, previously fixated locations are revisited at a much lower probability Posner et al. (1985); Klein (1988, 2000). We aimed to assess whether GVSM too captures this behavioral effect. To do so, we examined how the probability of fixating on each location evolved across successive fixations. To this end, we tracked the probability of fixating at each location at time steps prior to being fixated, at the fixation time step, and at the following time steps using the fixation probability maps generated by the Fixation Selector at each time step (Fig. 3D; inset). The probability of fixating on a location gradually increased until it peaked at the time of fixation, indicating progressive integration of information across fixations by the model (Fig. 3D). GVSM successfully replicated the post-fixation decline in probability, which decreased to a nonzero level, indicating that GVSM successfully learned an inhibition of return mechanism mimicking those documented in human behavior, reducing the probability of previously visited locations (Fig. 3D). This suggests that inhibition of return may emerge naturally in neural network models trained to perform visual search from raw sensory inputs. These findings reaffirm that fixation history plays a significant role in fixation selection Posner et al. (1985); Klein (1988, 2000); Smith and Henderson (2009), and GVSM captures this effect.

Trial-by-trial consistency between GVSM and human search scanpaths. Beyond replicating high-level characteristics of human search behavior, a strong model of visual search should also align with humans at a fine-grained level, capturing the sequence of eye movements used to find the target, the scanpath, on each trial. We compared model and human scanpaths using the fixation-by-fixation scanpath prediction method, one of the most recent methods of scanpath alignment evaluation Kümmerer and Bethge (2021). In this approach, we obtain the model's fixation probability map at each step (the 10×10 output of the Fixation Selector module), conditioned on the subject's saccadic history thus far (Fig. 3E), and evaluate it against the subject's ground-truth fixation at that step using standard metrics from the saliency prediction literature (Fig. 3F). We found that GVSM's scanpaths best matched human behavior, outperforming all baselines(aggregate scores comparison: GVSM vs. eccNET: $t = 18.83$, $p < 0.001$, GVSM vs. IVSN: $t = 9.5$, $p < 0.001$) (Fig. 3F). Together, these results demonstrate that GVSM not only captures high-level aspects of human search behavior but also achieves the strongest alignment so far with human spatiotemporal fixation sequences despite not being trained on eye-tracking data.

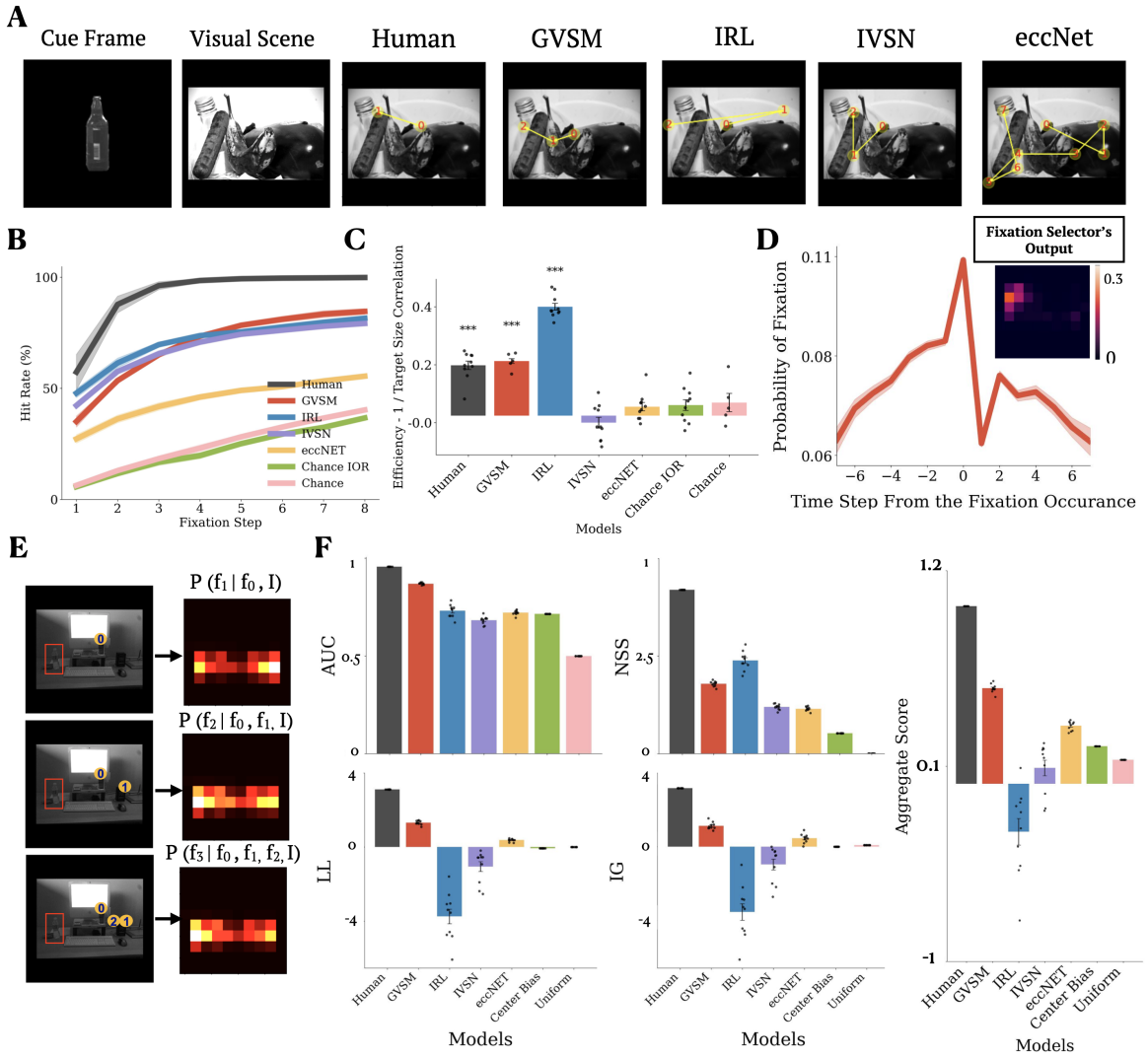

Figure 3. GVSM replicates human search behavior. A. Example test set search scanpaths. Search scanpaths of a human subject, GVSM, and the baseline models on an example trial from the COCO-Search18 test set. **B. Cumulative performance curves.** The proportion of trials in which the target was found within a given number of fixations is shown. **C. Comparing the effect of target size on search efficiency across models.** Spearman correlation of the number of fixations required to find the target with the $\frac{1}{\text{target size}}$ averaged across participants or model instances. Each point represents a human participant, a run of the GVSM model, or a bootstraped instance of the deterministic baselines. We bootstrapped the trials to create multiple instances of the deterministic baselines. Error bars represent the standard error magnitude (SEM) across the instances. The stars indicate the significance of the correlation (three stars for $p < 0.001$). **D. Fixation probability evolution curves.** The probability of each eventually fixated location is plotted across time relative to when it is fixated. The x-axis denotes fixation offset (in time steps), and the y-axis indicates its probability in the fixation map. The solid line represents the average evolution curve across all the fixated locations in all the test trials, and the shaded areas represent their SEM. GVSM gradually integrates information across time steps and demonstrates inhibition of return behavior observed in humans. **E. Illustration of the fixation-by-fixation scanpath prediction method.** At each time step, the model receives the sequence of human fixations up to that point and outputs a conditional fixation probability map (10×10 grid from the fixation selector module), P conditioned on the human fixation selection history thus far on the image I. **F. Evaluation of fixation predictions.** Each conditional probability map is compared to the participant's ground-truth fixation at that step using four standard metrics from the saliency prediction literature (Methods). GVSM outperforms all baselines on three of the four metrics and ranks second only to IRL, a gold-standard baseline trained directly on eye-tracking data, on normalized similarity. Each dot represents a comparison of the model scanpaths with one subject, and error bars represent the SEM across those comparison pairs. Overall alignment was quantified by normalizing each metric by the corresponding average human score and averaging across the four metrics. GVSM achieves the highest overall alignment with human search scanpaths.

GVSM captures the horizontal-vertical biases in human search behavior

Human visual behavior is asymmetric along different polar directions at isoeccentric locations. In particular, visual performance in various tasks, including visual search, is higher along the horizontal axis compared to the vertical axis Barbot et al. (2021); Kupers et al. (2022); Chaikin et al. (1962); Carrasco et al. (2004); Corbett and Carrasco (2011); Mackeben (1999); Altpeter et al. (2000). This horizontal-vertical asymmetry has also been characterized by a stronger preference for horizontal fixations in human fixation selection behavior Tatler and Vincent (2009); Najemnik and Geisler (2005); Koevoet et al. (2025); Bahill et al. (1975a). Consistent with previous literature, participants in the COCO-Search18 task also exhibited a preference for horizontal eye movements with the highest frequency of fixations occurring along the X axis (Fig. 4A). GVSM best captures this horizontal bias, closely replicating the distribution of fixation directions indicated by a nonsignificant difference between the distributional distance between humans and that between the model and humans (GVSM-Human vs. Human-between: $t = 0.35$, $p = 0.73$; GVSM-Human vs. IVSN-Human: $t = -6.55$, $p < 0.001$, and GVSM-Human vs. IRL-Human: $t = -7.15$, $p < 0.001$; Fig. 4B).

Another facet of this asymmetry is the higher variability of fixated X coordinates versus Y coordinates (Fig. 4C). All the models show this higher variability across the X axis, although the difference between the variability along the X axis and Y axis for GVSM is closest to humans after IRL, the gold standard baseline trained on eye-tracking data (Fig. 4C). To further examine this asymmetry, we analyzed the temporal correlation for each of the X and Y coordinates of fixations in both human participants and models. We found that the Y coordinates of fixated locations are highly correlated across time whereas the X coordinates show no such correlation. This reflects the preference towards horizontal fixations and particularly the higher variability across the X axis. Again, using this measure, GVSM best captured the humans' difference between the correlation across time along the X and Y axis (Fig. 4D).

To investigate whether such a bias appears in the model's latent space, we fitted two sets of linear regression models to decode the X and Y coordinates of the upcoming fixation locations from the immediate leading Information Integrator's latent space ($X : 0.27 \pm 0.06$, $Y : 0.70 \pm 0.02$; diagonals in Fig. 4E). X coordinates were consistently less linearly decodable from the model's latent space, reflecting the higher variability along the X axis. Moreover, we found that the encoding subspace of the upcoming fixation locations is consistent across time as demonstrated by high generalization accuracy of fixation coordinate decoders trained on one time step to the others (Fig. 4E).

Next, we fitted linear regression models decoding the previously visited X and Y coordinates from the model's latent space at each time step to see if the bias existed at the memory storage stage as well as the encoding of upcoming fixations. We found that although upcoming X cooridnates were weakly decodable from the model's latent space, once they were realized by the model, i.e. the fixations were sampled from the fixation probability map and were used at the following time step to fixate the input, they were strongly encoded in the model's latent space as well as the memory of the visited Y coordinates (memory decoding accuracy for X: 0.78 ± 0.01; Y: 0.81 ± 0.02) (Fig. 4F). Notably, because the correlation across time is high along the Y axis, much of this high decodability of the memory of the Y coordinates can be explained by the higher correlation across time whereas the memory of the fixated X coordinates are as strongly decodable despite no correlation across time (Y: mean Spearman's $\rho = 0.50 \pm 0.01$, $t = 38.6$, $p < 0.001$ and X: mean Spearman's $\rho = -0.04 \pm 0.01$, $t = -1.76$, $0.05 < p$, difference shown in Fig. 4D). These results suggest that although GVSM exhibits greater variability in fixation choices along the horizontal versus vertical axis, it does not display stronger *memory* for horizontal than vertical ones.

Remarkably, GVSM captured this horizontal–vertical asymmetry despite lacking any built-in oculomotor constraints, previously hypothesized as a key factor in shaping this bias in contrast to the view that such biases may arise from task demands and scene structure alone Tatler and Vincent (2009); Torralba and Oliva (2003). Supporting this interpretation, GVSM also reproduces other canonical properties of human saccades, for instance, the typical skew toward shorter saccades and the tendency for horizontal saccades to be longer than vertical ones Tatler and Vincent (2009); Bahill et al. (1975b); Gajewski et al. (2004); Pelz and Canosa (2001); Tatler et al. (2006).

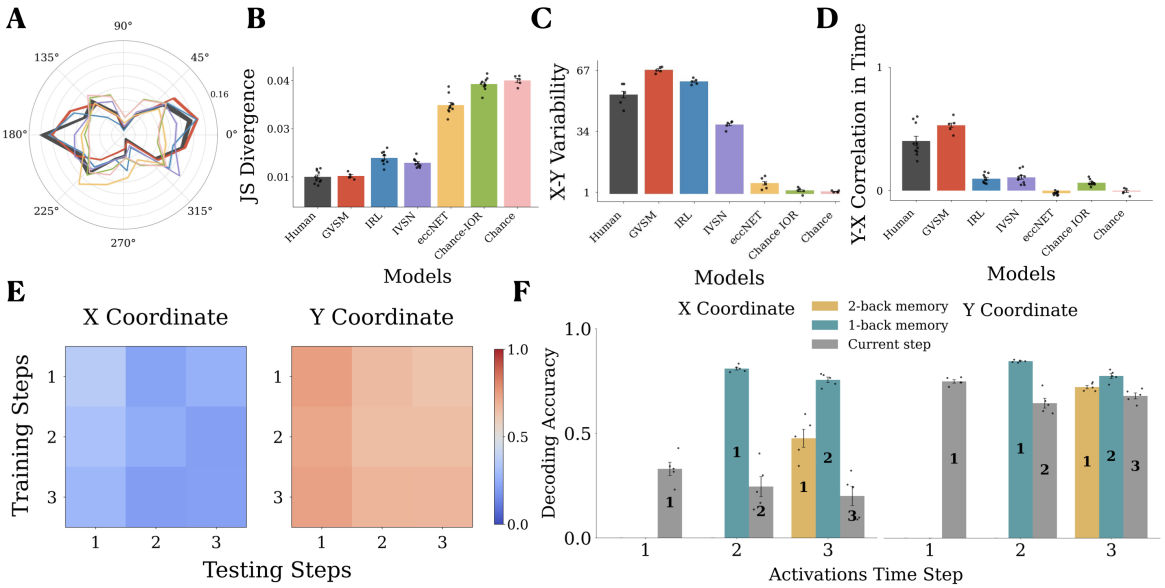

Figure 4. Human-like horizontal-vertical biases captured by GVSM. A. Histogram of fixation directions. Histograms of fixation directions show that both humans and GVSM exhibit a preference for horizontal over vertical fixations on the COCO-Search18 test set. Each curve represents the average probability distribution function of the fixation directions for humans. Shaded bands represent the SEM across observers (human participants, GVSM runs, or the bootstrapped version of the deterministic baselines). Colors follow the same conventions as in panels B-D. Fixation directions are shown in degrees (0-360), and the range was divided into 15 bins. Each concentrated circle represents a frequency level (outermost circle = 0.16). **B. Distance from the human distribution of fixation directions.** The distribution of GVSM's fixation directions most closely matches that of human observers. The y-axis shows the distance between each model's and humans' directional distributions (Methods) averaged over instances (human participants, GVSM runs, or bootstrapped versions of the deterministic baselines). Each dot represents an instance, and the error bars represent the SEM across them. Human-between is calculated by leaving one participant out and assessing the distance between its distribution of fixation directions from the average distribution of the other participants. **C. Variability difference along the X and Y axes.** Standard deviation of fixated X coordinates minus that of the fixated Y coordinates is shown for all the models and humans, averaged across instances. Each point represents an instance and error bars represent the SEM across them. **D. Difference in correlation across time along the X and Y axes.** Average correlation of the consecutive fixated X and Y coordinates across time averaged over the pairs is computed for all the instances of models and human participants. Each dot represents the difference between the average correlation across time along X axis minus that along the Y axis for an instance (a human participant, or a GVSM run, or a bootstrapped version of the baselines) and error bars represent the standard deviation across them. **E. Encoding of fixation coordinates in the model's latent space, capturing the horizontal bias.** The (x, y) coordinates of immediately upcoming fixations are linearly decodable from the latent space, with decoders generalizing across time steps. Notably, Y coordinates are decoded more reliably than X coordinates, consistent with the human tendency to favor horizontal over vertical fixations, resulting in higher variability of fixated X coordinates. **F. Fixation selection history encoding in the model's latent space.** Previously fixated (x, y) locations are decodable from the unit activations of the model's latent space at later time steps. Unlike the selection stage, X coordinates are reliably decoded, indicating specificity of the horizontal bias to selection rather than memory encoding.

Brain-like cue-similarity representations in latent space

GVSM replicates the role of cue-similarity in guiding fixation selection during visual search. Cue-similarity has long been considered as one of the main guiding factors of fixation selection in visual search Wolfe (2021); Wolfe and Horowitz (2017) as indicated by a tendency to fixate on locations that are visually or semantically similar to the target object Radulescu et al. (2022) and success of cue-similarity based models in predicting search scanpaths Pomplun and Gray (2007); Zelinsky (2008); Zhang et al. (2018); Ding et al. (2022). Most studies, however, assess similarity in an allocentric reference frame, independent of the viewer's gaze Radulescu et al. (2022); Pomplun and Gray (2007); Zelinsky (2008); Zhang et al. (2018); Ding et al. (2022). In contrast, human vision is inherently retinocentric, with acuity highest at the center of gaze and declining toward the periphery.

To highlight this distinction, we measured between allocentric and retinocentric frames by measuring two types of cue-similarity values at each image location. In the allocentric frame, similarity is computed globally; each location's similarity to the cue remains constant regardless of where the observer is fixating (Fig. 5A). In contrast, the retinocentric frame accounts for how the retinal image transforms with gaze shifts: similarity at each location is computed relative to the foveated view centered on the current fixation location. Thus, by computing cue-similarity at every location, we derive two distinct maps: a static allocentric cue-similarity map and a dynamic retinocentric cue-similarity map that updates with each fixation (Fig. 5A; see Methods). Hereafter, only the first three time steps are included in the analyses since the majority of trials ($65.09 \pm 1.06\%$) are solved by humans within the first 3 fixation steps. Consistent with previous studies, humans fixations were, on average, more similar to the cue object than non-fixated locations, highlighting cue-similarity as a guiding factor in visual search behavior (Fig. 5B). GVSM successfully captured this bias toward high cue-similarity locations.

Emergence of a cue-similarity map in the model's latent space. The behavioral findings indicated accordance of the GVSM behavior with the well-known bias of human fixation selection towards locations with higher similarity to the cue object, but important questions remain regarding the latent representations that drive this behavior. For example, which form of cue-similarity map, allocentric or retinocentric, is encoded in the model's latent space? What mechanisms the model employs to leverage these representations toward successful visual search in natural scenes? To what extent these internal representations and computations align with those observed in the primate brain? And what new predictions they offer about the underlying neural mechanisms? To approach these questions, we turned to the analysis of the model's latent space by first testing two hypotheses about how the cue-similarity map and fixation history guide fixation selection in GVSM, and how each compares with existing evidence from neuroscience (Fig. 5C):

- **H1: Retinocentric-based**: The Information Integrator encodes the current and prior states of the dynamic retinocentric cue-similarity map directly within its latent space. These representations, together with fixation history, are then used by the Fixation Selector to guide subsequent fixations.
- **H2: Allocentric-based**: The Information Integrator progressively constructs an internal allocentric cue-similarity map by projecting each retinocentric input into an allocentric reference frame and integrating them over time. The Fixation Selector then relies on this evolving allocentric representation, along with fixation history, to determine the next fixation.

To test these hypotheses, we first examined whether the model's latent space exhibits cue-similarity map representations similar to those observed in the primate brain during visual search. One of the most prominent neural signatures of visual search in primates is a distributed encoding of the retinocentric cue-similarity map across a fronto-parietal network, including areas such as LIP, FEF, and VPA, collectively known as the fronto-parietal attention control network Bichot et al. (2015); Bisley and Mirpour (2019); Machner et al. (2020); Powell et al. (1999); Sapountzis et al. (2018); Bichot and Schall (1999). A subset of neurons in these regions encode the similarity of a stimulus to the search cue relative to the current gaze position, responding more strongly when a target appears within their receptive field compared to distractors. In contrast, although allocentric cue-similarity maps have served as the foundation for many computational models of visual search, the neural encoding of such maps in the primate brain remains largely unknown. Here, we sought to investigate whether either or

both of these representations emerge in the model's latent space during visual search.

We reasoned that a cue-similarity map can be instantiated in the model's latent space through distributed patterns of neural activity, where different combinations of neurons encode cue similarity at different spatial locations. Geometrically, these encodings correspond to vectors in the latent space, with distinct directions representing different locations. Some of these directions may involve overlapping subsets of units, suggesting that individual neurons can contribute to the encoding of multiple locations.

To search for such a code, we fitted linear regression models to decode cue-similarity at each location from the model's unit activations at each time step. We found that a retinocentric cue-similarity map, similar to those observed in the fronto-parietal attentional control network, emerges in the model's latent space, supporting the retinocentric-based hypothesis (Fig. 5D). This was supported by the robust linear decodability of retinocentric cue-similarity values across the visual field from the model's unit activations over time (decoding accuracy=0.46 ± 0.02). The allocentric cue-similarity map was also linearly decodable from the model's latent space (decoding accuracy=0.54 ± 0.02), albeit this encoding did not improve over time and the decodability even slightly declined, suggesting the absence of an evolving internal representation of an allocentric map. This finding contradicts the allocentric-based hypothesis (H2), which posits that the model should progressively construct and refine such a map over time.

A natural question to ask is, why would the model encode both allocentric and retinocentric cue-similarity maps in its latent space? We hypothesized that the existing allocentric information in the model's latent space may be the remnant of retinocentric information collected from the first center fixation. If true, the allocentric cue-similarity map encoding found in the model latent space would be identical to the first center-fixating retinocentric observation of the search image, rather than being a dynamically evolving allocentric representation of the entire field of view.

To test this hypothesis, we assessed the generalization accuracy of retinocentric cue-similarity map decoders on decoding the allocentric cue-similarity map. We expected high generalization accuracy if this hypothesis were true, as it would reveal a shared encoding space for the two. Generalization accuracy at the first step (0.36 ± 0.03) was comparable to the decoding accuracy of the retinocentric cue-similarity map itself (0.39 ± 0.05), supporting this hypothesis. Taken together, these findings were compatible with the retinocentric-based hypothesis (H1) as a plausible mechanism for visual search in the primate brain–one that aligns with empirical evidence indicating the presence of retinocentric cue-similarity map representations.

Brain-like activity predictive of future fixations

We found that GVSM's internal representations reflect a retinocentric cue-similarity map resembling the spatial priority signals observed across the fronto-parietal attentional control network, particularly in regions such as FEF, LIP, VPA, and the superior colliculus (SC), which lies closer to motor execution of eye movements. This emergent structure in the model's latent space provides a plausible computational substrate for the core mechanism proposed in the retinocentric-based hypothesis (H1) of fixation selection during visual search. According to this hypothesis (Fig. 5C), the history of retinocentric cue-similarity maps, combined with the current retinocentric map derived from ongoing partial observations, guides the selection of subsequent fixation locations. Previous work has shown that neural activity in the FEF can predict not only the next, but also subsequent fixations Phillips and Mark (2006). Likewise, recordings from the SC show that target selection signals can emerge before visual input reaches the neuron's receptive field, indicating that future fixation locations may be preselected in parallel Shen and Paré (2014).

To assess whether GVSM similarly engages in multi-step fixation selection based on early visual input, we first conducted a behavioral ablation analysis. In this experiment, we corrupted the model's visual input at all time steps following the first, such that after receiving the cue and an initial central glimpse of the scene, subsequent observations were replaced with black frames. Despite the absence of new visual information beyond the first step, the model retained a substantial degree of its original performance (Fig. 6A). This indicates that it had preplanned significant aspects of its search behavior. The performance decline induced by this manipulation grew with increasing trial length (i.e., number

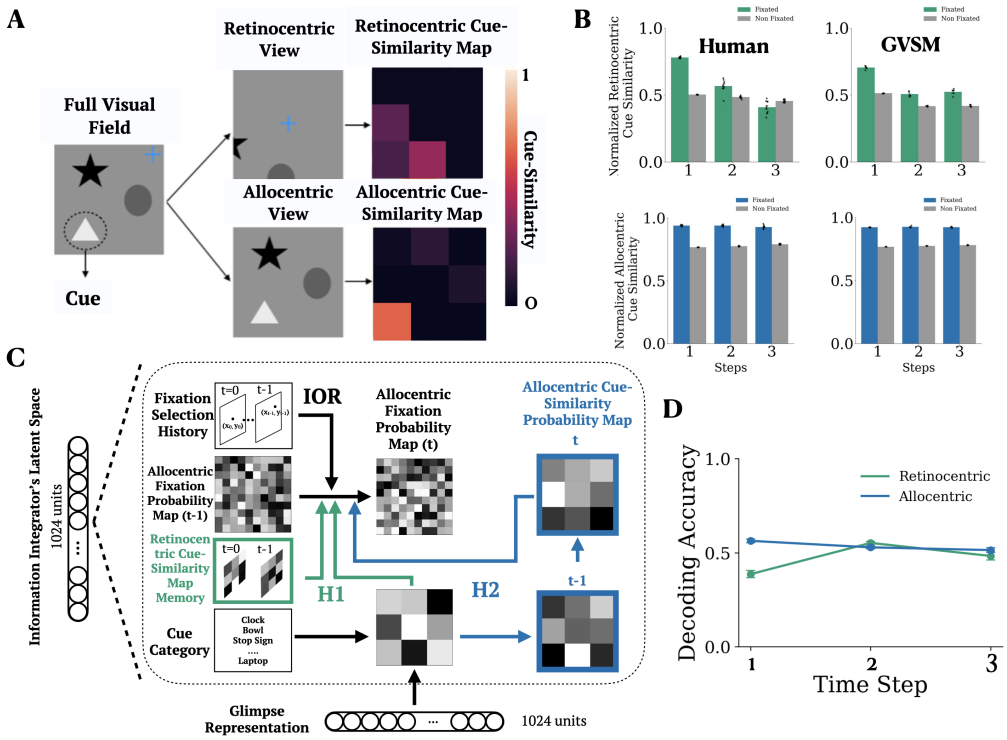

Figure 5. Brain-like cue-similarity representations in GVSM latent space A. Schematic of retinocentric vs. allocentric cue-similarity maps. Cue-similarity was computed allocentrically (fixation-independent) or retinocentrically (fixation-dependent). **B. Cue-similarity at fixated vs. non-fixated locations.** Cue-similarity normalized by per-step maximum at fixated locations is higher than at non-fixated locations. Each dot represents an observer instance, and the error bars show the SEM across them. Only the first three time steps are included. **C. Schematic of two hypotheses about representations underlying fixation selection in GVSM**: H1) Model encodes current and past retinocentric cue-similarity maps, along with fixation selection history. These elements are combined to form a fixation probability map that guides future fixations (green). H2) The model incrementally constructs an allocentric cue-similarity map of the full field of view by integrating the sequence of incoming retinocentric observations (blue). In both cases, fixation selection history contributes to the updating of the fixation probability map by inhibition of return (IOR), and the retinocentric cue-similarity map is computed in the model's latent space from the stored cue category information and the incoming retinocentric observations. **D. Cue-similarity map encoding in the model's latent space.** Decoding accuracy of the linear regression models on predicting the cue-similarity at each location on the cue-similarity map from the model's unit activations is shown across time steps. Decoding accuracies were averaged across the observer instances for each location, then averaged across the 9 locations. Error bars represent SEM across 9 locations.

of fixation steps), suggesting a gradual shift from reliance on preplanned trajectories to dependence on new sensory input during more complex searches (Fig. 6B).

The ablation experiment suggested that the model relies on the intial observation, especially during the first few saccades. To more directly probe the presence of activity predictive of future fixations in the model's internal representations, we fitted linear regression decoders to predict future fixation locations from the model's latent state at each time step. Indeed, the upcoming fixation coordinates were significantly decodable from GVSM's latent representations (Fig. 6C), paralleling the predictive activity found in both the FEF and SC during visual search Shen and Paré (2014); Phillips and Mark (2006).

Together, these results demonstrate that GVSM learns brain-like strategies for visual search, where latent representations encode both short-term goals and longer-horizon plans, much like the predictive dynamics observed in primate oculomotor systems.

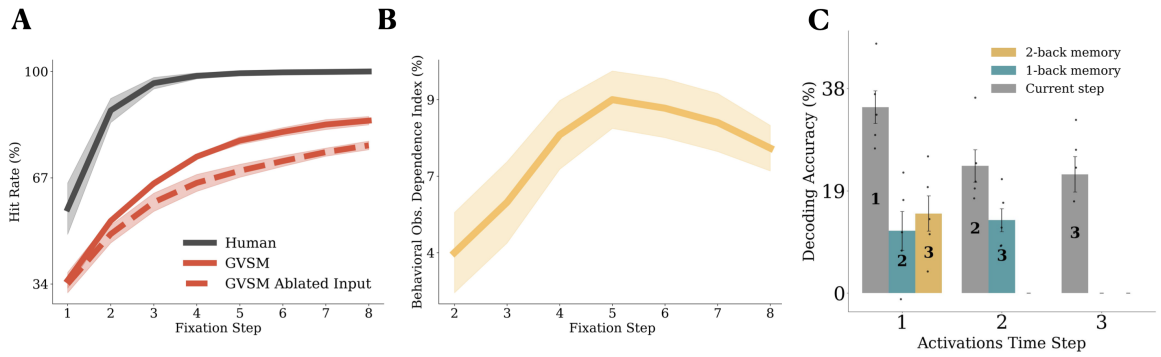

Figure 6. GVSM activity predicts future fixations. A. Cumulative performance curves of GVSM with normal and ablated inputs. In the ablated condition, the model receives black inputs after seeing the curve frame, and a center-fixated visual scene once. Error bars represent the SEM across searcher instances. **B. Behavioral measure of dependence on the incoming observations.** Dependence on the current observation is quantified by the performance gap between the normal and ablated-input conditions. For trials requiring a higher number of fixations, and further into the search process, the reliance of the model on the initial observation decreases, as shown by an increase in dependence on the incoming visual input. **C. Activity predictive of future fixations.** Linear decodability of the future fixation locations from the model's latent space indicates a latent representation supporting the multi-step fixation selection akin to the predictive activity observed in the primate brain. Grey represents the decoding accuracy of the immediate fixation locations. Numbers on each bar denote the time step of the decoded coordinate. Each dot represents a searcher instance, and the bars show the average across them, and the error bars represent their SEM. Only results for the fixated X coordinates are shown here because Y coordinates are highly decodable from previous time step unit activations due to their high correlation across time, thus not providing nontrivial information in this regard.

Predicting the geometry of the latent cue-similarity map representations

Geometry of the latent cue-similarity map representation. We showed that GVSM's latent space encodes cue-similarity of different spatial locations in retinocentric and allocentric reference frames. Next we asked how different units in the model participated to encode this information. We devised two competing hypotheses about the geometry of the cue-similarity map representation in the model's latent space. Under the first hypothesis, the cue-similarity map is encoded in a discontinuous configuration where the similarity values at different visual field locations are arbitrarily represented by distinct neural subspaces irrespective of their spatial proximity (Fig. 7A, left). Under the second hypothesis, the map is encoded in a continuous topographic configuration, such that cue-similarity values at nearby locations in the visual field are represented by more aligned neural subspaces (Fig. 7. A, right).

To investigate this, we first visualized the learned cue-similarity decoders in the three-dimensional principal component (PC) space derived from the unit activations. This revealed a continuous topographic structure consistent with the second hypothesis, where encoding directions corresponding to

spatially proximal locations were more aligned than those corresponding to distant locations (Fig. 7. B). Furthermore, we reasoned that if two locations share overlapping representations, a decoder trained at one should generalize better to the other. Consistent with this idea, we observed a systematic decline in decoder generalization accuracy as the distance between training and test locations increased, supporting a continuous and topographically organized encoding of cue-similarity map in the model (Fig. 7C). Altogether these results suggest that cue-similarity at proximal parts of the visual field are encoded by partially overlapping populations and distant parts are encoded in partially opposing directions (Fig. 7C; negative generalization).

Transformation of the cue-similarity map encoding subspace over time. To gain a better understanding of how the model tracks cue-similarity across time, we additionally examined whether the cue-similarity maps are consistently encoded within the same subspace at different time points during the task (Fig. 7. D). To investigate this, we evaluated the generalization accuracy of cue-similarity decoders by fitting them on one time step and testing on subsequent time steps (Fig. 7E). The results showed low generalization accuracy between the first time step and later steps, while there was high generalization among the later time steps. This suggests that the initial encoding of the retinocentric cue-similarity map occurs in a distinct subspace, while later encodings share a common subspace. This distinct encoding of the first time step retinocentric cue-similarity map is consistent with the previous finding that the retinocentric observation at the first time step plays an important role in prospective planning of future fixations, suggesting the criticality of this information for the search performance (Fig. 6).

We next asked whether the model reuses the same geometric structure to encode retinocentric cue-similarity across time, despite their seemingly distinct representations at different steps. To test this, we applied orthogonal Procrustes analysis Schönemann (1966) to determine whether a simple rotation could align the subspace from the first-step with the common subspace observed in later steps. The successful alignment suggests that the model maintains a shared geometry for encoding cue-similarity across time, even if the representation appears distinct (Fig. 7. F-G). This finding indicates that while the model may instantiate separate subspaces for encoding cue-similarity at different time steps, these subspaces are not arbitrary but are instead geometrically related through a consistent transformation—highlighting a stable, reusable representational structure.

Neural mechanisms of maintenance and retrieval of memory during visual search

Memory of the cue category information in the latent space. To accurately perform the visual search task, GVSM must retain crucial information about its prior observations and actions. We first investigated how GVSM retains the cue information during visual search. Similar to prior analyses, we used a linear classifier (support vector machine) to decode the cue category from the model's latent space. We found that the cue category information could be reliably decoded from the latent space after observing the cue frame during the task (Fig. 8. A). Furthermore, the cue-similarity was maintained within a stable subspace of the model's latent representation throughout the search sequence. This stability is evidenced by the successful generalization of cue category decoders trained on unit activations from one time step to others (Fig. 8A). Such consistency enables the model to generate retinocentric cue-similarity maps by integrating the sustained cue representation with incoming visual observations at each fixation.

Memory of the retinocentric cue-similarity map in latent space. Our previous results showed that retinocentric cue-similarity maps underlie the ability of the model to perform visual search (Fig. 5). Based on this hypothesis, the current retinocentric cue-similarity map and the memory of previous retinocentric cue-similarity maps contribute to the fixation selection (Fig. 5C). To test whether such memory is preserved, we fitted linear regression models to decode cue-similarity maps from earlier time steps using the unit activations of the Information Integrator. We found that the cue-similarity maps from the first two time steps could be reliably decoded from later time steps, although decoding was less accurate for the second time step (Fig. 8B).

Memory maintenance and retrieval mechanisms. Results in Fig. 8B confirmed that GVSM encodes memories of prior cue-similarity maps in its latent space. However, it is unclear how the

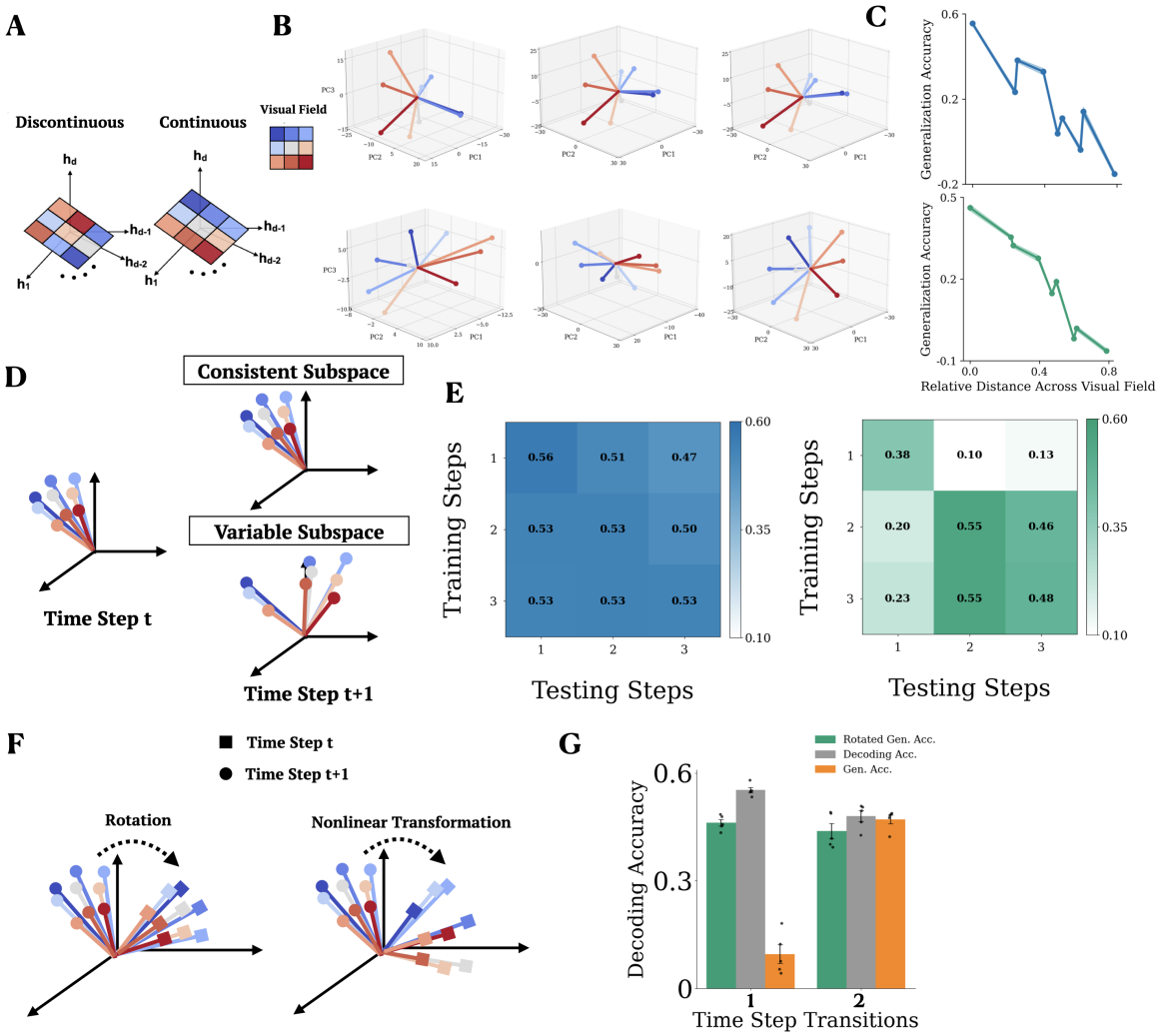

Figure 7. Geometry of the latent cue-similarity map representations. A. Schematic of the geometry of the latent cue-similarity map representation. The cue-similarity map can be represented in a discontinuous, nontopographic configuration in the model's latent space, where cue-similarities across different locations in the visual field are encoded in separate subspaces, irrespective of their physical proximity (left). Alternatively, it can be represented along a continuous topographic latent representation, where cue-similarities of locations that are closer in the visual field are encoded in more aligned subspaces (right). **B.** Geometry of the latent cue-similarity map representation in the PC space. Cue-similarity encoding axes are color-coded according to their location in the visual field. Axes corresponding to nearby locations in the visual field show greater alignment in the latent space, demonstrating a continuous topographic configuration. The top and the bottom row show the allocentric and retinocentric cue-similarity decoders, respectively. **C.** Cross-location generalization accuracy of cue-similarity decoding. The generalization accuracy of linear regression models, trained to decode cue-similarity at one location, was tested on decoding cue-similarity at other locations. Generalization accuracy decreased as the distance between locations in the visual field increased, confirming a continuous topographic representation. **D.** Schematic of the cue-similarity map representation over time. The cue-similarity map could either be encoded consistently in the same subspace over time or in separate subspaces at different time points. **E.** Cross-time generalization accuracy of cue-similarity decoding. We tested the generalization of cue-similarity decoders (allocentric depicted in blue, and retinocentric in green) trained on one time step and applied to others. Generalization accuracy was low between the first time step and subsequent ones, but high thereafter, suggesting that a distinct cue-similarity map encoding subspace is used at the first time step, with a shared subspace for later time steps. **F.** Schematic of possible geometrical relationship between the retinocentric cue-similarity map representations across time. **G.** Transformation of the retinocentric cue-similarity map encoding subspace over time. Orthogonal Procrustes analysis revealed that the distinct retinocentric cue-similarity map encoding subspace at the first time step could be rotated to align with the shared subspace encoding the retinocentric cue-similarity map at later fixations. Grey bars represent the baseline (average decoding accuracy at the encoding time step), orange bars indicate the generalization accuracy across time steps, and green bars show the accuracy of decoders rotated by the identified transformation. Error bars represent SEM across multiple model instances.

model retains and reuses the cue-similarity maps across time. We reasoned that the Information Integrator component of GVSM may implement one of two distinct mechanisms to store memory representations of retinocentric cue-similarity maps or fixation selection history (Fig. 8C).

- **H1: Slot-based memory subspaces Whittington et al. (2025); Wang et al. (2025); Luck and Vogel (1997)**: In this framework, the model's latent space is partitioned into distinct, time-indexed subspaces, or "slots", with each slot dedicated to storing a single item, such as a computed retinocentric cue-similarity map or a visited fixation location. Once an item is encoded, it remains in its designated slot until retrieval. These slots occupy unique and temporally stable regions of the latent space, supporting sustained, interference-resistant memory representations over time.

- **H2: Chronologically indexed memory subspaces Lei et al. (2024)**: In this alternative framework, memory is organized by the relative age of encoded content. The latent space is divided into subspaces based on how long ago the information, such as a cue-similarity map or fixation location, was stored. Each subspace is dynamically updated as the task progresses, maintaining information according to its temporal distance from the current time step.

To test these hypotheses, we performed decoding analyses using linear regression models on both the retinocentric cue-similarity maps computed at each time step from new sensory input and the maintained cue category memory, and the generated fixation locations (fixation selection history). For fixation history, we focused on X-coordinate decoding only, as the Y-coordinates exhibited strong correlation across time, resulting in a shared encoding subspace across all time steps (Fig. 5G). This prevented a meaningful analysis of y-coordinate memory mechanisms

We found that the retinocentric cue-similarity map is encoded via a time-dependent slot-based mechanism, with the memory slots functioning differently depending on the encoding time step of the incoming information. We earlier showed that the first-step retinocentric cue-similarity map is encoded in a subspace distinct from other time steps (low generalization accuracy of the first-step cue-similarity decoders to the next time steps in Fig.7E). To test whether the memory is encoded in slot-based subspaces, we measured the generalization accuracy of the decoders trained on the encoding time step on decoding the same information from the later time steps. The first step retinocentric cue-similarity map decoder largely generalized to later time steps (T=1: 0.38±0.03; T=2: 0.35±0.01; T=3: 0.31±0.005; Fig. 8D) and its accuracy largely matched the memory decodability at those time steps (accuracy of linear regression models trained and tested on decoding the first-step retinocentric cue-similarity maps from the unit activations at $T = 2$, and $T = 3$; average: 0.36±0.02; Fig.8D). This suggests a slot-based mechanism in retaining the first retinocentric cue-similarity map (Fig. 8E, Slot 1).

However, when repeating the same analysis on the second time step, we observed a sharp drop in decoder generalization between steps 2 and 3 (T=2: 0.54±0.02; T=3: 0.40±0.02; Fig.8D), suggesting that the memory of this time step is not fully maintained in later time steps. From our earlier analysis, we already knew that the retinocentric cue-similarity maps in steps 2 and 3 are encoded in a shared subspace (Fig. 7E). This suggests that in the steps following step 1, retinocentric cue-similarity maps are encoded into a shared slot (Fig. 8E, Slot 2^+), and inevitably overlap with each other. Why should a model choose to do this? We believe that this shared space underpins the implementation of a dynamic representation of cue-similarity in the model that is updated by every fixation and is the key in guiding the model's fixation selection mechanism.

Furthermore, to test H2, we fitted linear regression models to decode n-steps-back information from unit activations at each time step and evaluated their generalization to decoding the same n-steps-back information at other time points. The relative chronologically indexed memory subspaces hypothesis (H2) was ruled out as a candidate for the retinocentric cue-similarity map memory encoding and retrieval mechanism, indicated by low generalization accuracy of the 1-back decoders trained on unit activations of step 2 to 3 (0.14±0.02) and vice versa (0.10±0.03) substantially lower than the corresponding memory decoding accuracies (0.35±0.02 and 0.39±0.02, respectively). Conversely, the 1-step-back fixation history decoders exhibited high generalization from one step pair to the other (Fig. 8F), indicating storage of these memories in chronologically indexed memory subspaces. Low slot retrieval accuracies compared to the memory decoding accuracies confirmed the lack of a slot-based memory storage and retrieval mechanism for the fixation selection history (Fig. 8G). This is

consistent with the previous finding that each of the fixation coordinates X and Y are encoded in a shared stable subspace across time (Fig. 4E).

Overall, our results revealed a dissociation between the two mechanisms: memories of the retinocentric cue-similarity maps were encoded in the time-dependent slot-based subspaces (a variation of H1), while fixation selection history was maintained within chronologically indexed subspaces (H2) (Fig. 8D-E).

This observed difference may be attributed to the distinct functional roles of the fixation selection history and the retinocentric cue-similarity maps in guiding the model's virtual eye movements. Fixation selection history supports inhibition of return and exploration by tracking previously visited locations, which inherently requires maintaining a temporal sequence. Encoding this information in chronologically indexed subspaces may enable the model to flexibly update the fixation probability map, inhibiting recently visited locations. In contrast, retinocentric cue-similarity maps act as relatively stable, spatially organized templates for guiding search based on similarity to the cue. Consistent with this, we previously found that when the model's input was ablated such that it received the scene only once at the beginning of the trial, its performance remained high, indicating a strong reliance on early visual input and the initial cue-similarity map. The current finding that these maps are stored in stable, slot-based memory subspaces suggests a mechanism for preserving their fidelity across time and fixations. Importantly, this mechanism is found to be time-dependent; more stably encoding the first time step retinocentric cue-similarity map in a distinct subspace and sharing a common subspace encoding partial memories of the cue-similarity map at later time steps. This allows the model to repeatedly retrieve and apply the first-step map, which was behaviorally found to be critical for the search performance, without interference. Together, these results imply that GVSM prioritizes durable storage of early scene representations while dynamically updating its memory of fixation history to support ongoing exploration.

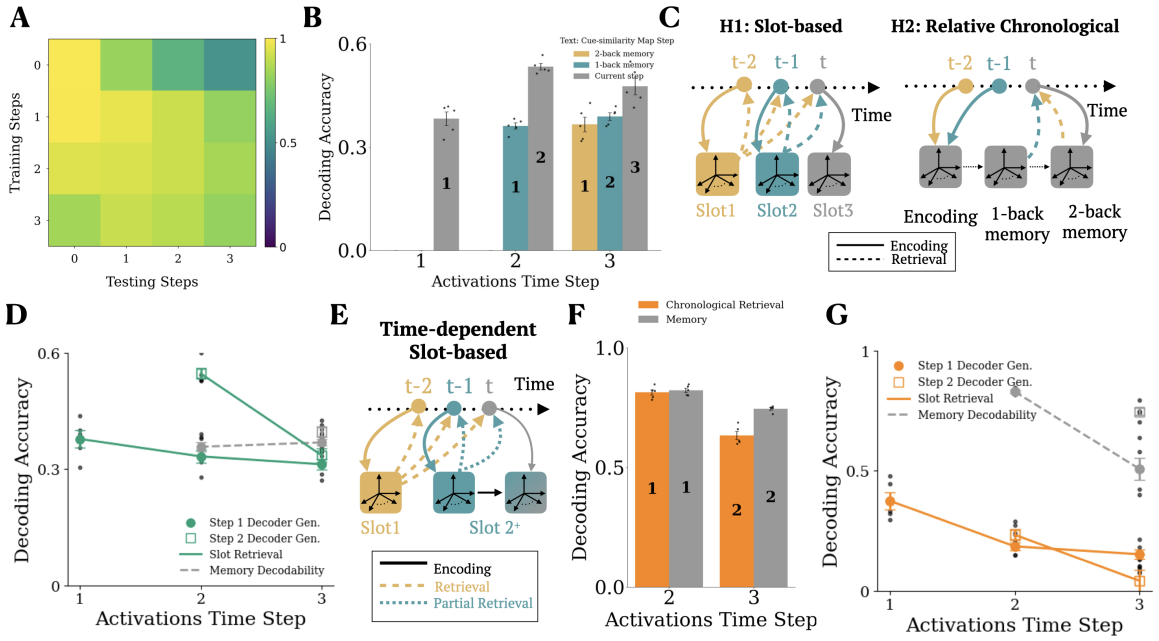

Figure 8. Neural mechanisms of maintenance and retrieval of memory. A.Cue category memory representations. High generalization accuracy of linear regression models trained to decode cue category information from unit activations across time steps indicates that cue information is encoded in a stable subspace of the model's latent space. **B. Memory of the retinocentric cue-similarity map.** Retinocentric cue-similarity maps computed at earlier time steps could be accurately decoded from unit activations at later time steps, demonstrating successful memory storage within the latent space. Green bars show memory decoding accuracy, i.e., the average ability to decode previously computed cue-similarity maps from later time steps. Grey bars show baseline decoding accuracy at the encoding time step when the map was first available. Error bars indicate SEM across model instances. **C. Schematic of two candidate memory encoding mechanisms.** H1: Slot-based memory: Each item is stored in a distinct, temporally stable subspace of the latent space. H2: Chronologically indexed memory: Items are stored in subspaces indexed by relative temporal distance, with memory contents dynamically updated as time progresses. **D. Evidence for time-dependent slot-based memory encoding of retinocentric cue-similarity maps.** First step retinocentric cue-similarity map decoder generalizes well to later time steps (filled green circles) and matches memory decodability in those steps (filled grey circles). Second step retinocentric cue-similarity map decoder's generalization is substantially lower at step 3. **E. Schematic of the time-dependent slot-based memory storage and retrieval mechanism.** Memory of the first time-step retinocentric cue-similarity map is stably preserved in its encoding subspace, while the following retinocentric maps update their shared encoding subspace, maintaining a partially combined memory of the maps beyond the first time step. **F. Evidence for chronologically indexed memory encoding of fixation selection history.** To test H2, linear regression models were trained to decode the X-coordinates of previously fixated locations. Generalization across time steps revealed consistent subspace structure indexed by temporal distance, supporting the chronologically indexed memory hypothesis for fixation history. **G. Lack of slot-based memory encoding subspaces for the fixation selection history.** Linear regression models trained to decode the fixation X-coordinates at either steps 1 or 2 do not generalize to later time steps (orange points) and are significantly lower than the memory decodability in those steps (grey points), indicating a lack of a slot-based memory code for the fixation selection history.

Discussion

We developed a neural network model of visual search, called GVSM, to address key limitations of prior models, namely, the lack of a dynamic internal state for integrating information across fixations, reliance on symbolic rules rather than learned latent representations, and unrealistic assumptions about uniform visual acuity or biologically implausible retinal transformations. GVSM features a biologically inspired architecture that searches for targets directly from pixel inputs, making it both image-computable and neurally plausible. It incorporates eccentricity-dependent visual acuity, maintains a dynamic internal representation to integrate information across fixations, and generates artificial neu-

ral activations that are testable against brain data. Trained solely on the visual search task objective, GVSM exhibits strong generalization to unseen stimuli and captures both high-level and fine-grained aspects of human search behavior, producing human-like search scanpaths. Furthermore, the model adopts human-like search strategies and encodes information using brain-like neural representations. We found that its fixations are predicted by cue similarity and fixation history in ways that closely mirror human behavior. Importantly, because GVSM is not explicitly hand-engineered with mechanisms such as inhibition of return, it naturally integrates fixation history into its fixation selection process, resulting in fixation dynamics that are more aligned with those observed in humans.

Brain-like activity predictive of future fixations. The results from our ablation and decoding analyses provide evidence that GVSM's internal representations support prospective fixation selection during visual search. The ability to predict future fixation locations from the model's latent state prior to their execution suggests that GVSM maintains information about upcoming saccades over multiple steps, consistent with prior findings of anticipatory activity in regions such as the frontal eye fields and superior colliculus Phillips and Mark (2006); Shen and Paré (2014). The model's preserved performance when visual input was restricted after the first glimpse further indicates that early observations can inform subsequent behavior, allowing search to proceed in the absence of new sensory input. At the same time, the decline in performance with longer trial durations suggests a growing reliance on incoming visual information as search complexity increases. Together, these results indicate that GVSM combines elements of preplanned and reactive control, and may offer a computational account of how short- and longer-term goals are jointly represented in primate oculomotor systems.

Emergence of brain-like cue-similarity representations in the latent space. We found evidence of the emergence of Brain-like neural representations in GVSM's latent space. One of the most prominently studied representations in the neuroscience of visual search is the distributed encoding of a retinocentric cue-similarity map in the fronto-parietal attentional control network, including FEF, VPA, and LIP Bichot et al. (2015); Bisley and Mirpour (2019); Machner et al. (2020); Powell et al. (1999); Sapountzis et al. (2018); Bichot and Schall (1999). We found that a similar cue-similarity map emerges in the model's latent representations, despite the model being trained solely to perform the task. These learned representations support a retinocentric-based mechanism of visual search, in which current and previous retinocentric cue-similarity maps are encoded and guide fixation selection, in conjunction with memory representations of previously visited locations.

This stands in contrast to allocentric-based hypotheses that underlie prior cue-similarity-based computational models of visual search Zhang et al. (2018); Ding et al. (2022); Zhang et al. (2022); Adeli et al. (2017); Pomplun and Gray (2007); Zelinsky (2008); Rao et al. (2002); Navalpakkam and Itti (2006, 2005), where fixation selection is guided by winner-take-all mechanisms applied to allocentric cue-similarity maps. In our model, however, we found no evidence of a strongly encoded allocentric cue-similarity map in the latent space, nor any indication of improving decodability over time that would suggest an evolving allocentric representation. This lack of evidence provides direct evidence against the allocentric-based hypothesis. While earlier models based on this framework have been successful in capturing aspects of search behavior, our results demonstrate that task-driven training alone is sufficient to produce a model that better aligns with both human behavior and brain-like neural representations.

Overall, we demonstrated that the model relies on neural representations closely resembling those observed in the primate brain during visual search. Specifically, a retinocentric cue-similarity map emerges in the model's latent space, akin to those found in the primate fronto-parietal attentional control network. This finding supports a retinocentric-dependent mechanism of visual search, in contrast to the allocentric framework typically assumed in computational models.

While GVSM establishes a new state-of-the-art in replicating human search behavior and, unlike prior models, exhibits brain-like neural representations, a significant gap between the model and humans remains in both task performance and scanpath prediction. Future improvements may be achieved by scaling model capacity and training data, or by incorporating additional brain-inspired inductive biases, for instance, top-down modulatory connections from the information integration module to the vision module, analogous to feedback from prefrontal to visual areas in the primate brain. Further alignment with biological systems may also be attained by training the model to generate retinocentric actions.

Horizontal bias due to statistics of natural scenes. Humans are known to have a bias for making horizontal saccades, characterized by greater variability, lack of autocorrelation, and higher frequency of fixations along the x-axis, which was captured by the GVSM model at both behavioral and representational levels. In particular, the model's latent space asymmetrically encoded upcoming fixation coordinates: the y-coordinate was highly linearly decodable, whereas the x-coordinate showed low linear decodability. This asymmetry in human preference for horizontal eye movements have been hypothesized to arise from either the oculomotor constraints or ecological relevance and scene statistics Tatler and Vincent (2009). Here, we showed that these biases emerge in GVSM in the absence of any oculomotor or biomechanical constraints. This challenges the view that such biases arise primarily from movement-related effort or physical limitations. Instead, it supports the hypothesis that directional saccade preferences can emerge from the structure of the task and the statistics of the visual environment since they appeared in the model simply through task optimization, without any built-in constraints. Everyday behaviors such as reading and exploration, which unfold predominantly along the horizontal axis, may have shaped both the human visual system and the visual input used to train our model, contributing to the emergence of horizontal bias. Consistent with this interpretation, the model also captures other human-like saccadic properties such as saccade size, direction distributions, changes in saccade direction, and fixation eccentricities, despite lacking an oculomotor system. This further supports the view that these behavioral signatures can result from task optimization rather than from embodied constraints.

A neurally-plausible model of visual search. Having a neurally plausible model, i.e., one that generates distributed activations akin to neural population responses in the brain, offers a unique opportunity to study the learned neural representations not only for their alignment with the brain but also to generate testable predictions about the neural mechanisms underlying visual search. Our findings enabled several concrete predictions: (1) The retinocentric cue-similarity map is encoded in a continuous topographic format, such that locations in close proximity within the visual field are represented by more closely aligned subspaces in the neural state space. (2) This cue-similarity representation rotates across time: it begins in a distinct neural subspace that dynamically reorganizes, preserving its internal geometry while rotating into a new, stable representation across subsequent time steps. (3) Cue category information is consistently encoded in a stable subspace over time. (4) Memory representations follow two distinct formats: the retinocentric cue-similarity map is stored in time-dependent slot-based subspaces that remain uninterrupted by the incoming sensory information or get updated across time, depending on the encoding time step, while previously visited fixation locations are encoded in chronologically indexed subspaces, reflecting their temporal order. Further investigation of the model's latent space through ablation studies and mechanistic interpretability tools may yield additional insights into the potential circuitry and computational principles supporting visual search in the primate brain.

A goal-directed neural network model replicates complex behavioral phenomena. Our findings contribute to a growing body of work that shows task-optimized neural networks can serve as predictive models of brain function, capturing both behavioral patterns and neural representational dynamics. While much of this work has focused on sensory cortices, for instance, CNN models of visual object recognition predict neural responses along the ventral visual stream in nonhuman primates Yamins et al. (2014); Ratan Murty et al. (2021); Bashivan et al. (2019); Cadieu et al. (2014); Khaligh-Razavi and Kriegeskorte (2014), models of the auditory cortex have successfully predicted population-level activity in response to natural sounds Kell et al. (2018), and deep networks trained on odor recognition tasks have recapitulated representational structures observed in the olfactory system Wang et al. (2021); Taleb et al. (2024), more recent studies have begun to extend this approach to higher-level cognition. Notably, large language models have been shown to predict activation patterns across the cortical language network, including Broca's area, the posterior superior temporal gyrus, and angular gyrus Caucheteux and King (2022); Schrimpf et al. (2021). However, these models have rarely been extended to predict neural responses in higher-order cortical regions such as those in the prefrontal or parietal cortex, areas involved in flexible, task-dependent control, likely due to the broader behavioral repertoire and complexity of the computations performed there.

Our model extends these efforts to the domain of goal-directed vision, providing a neurally grounded framework for understanding visual search behavior in naturalistic settings, taking a first step toward building neurally predictive models of the fronto-parietal attentional control network, including regions

such as the FEF, VPA, and LIP, which play a key role in visual search behavior. We show that the representations learned by our model during natural scene search closely align with those observed in these areas, suggesting that task-optimized models can capture internal representations beyond early sensory cortices. By linking high-level task objectives, such as object search, to the emergent representational geometry of a deep neural system, we offer a computational account of how dynamic visual memory, cue guidance, and fixation selection might be jointly implemented in the primate brain. This framework not only accounts for empirical patterns in eye movements but also generates concrete predictions for future neurophysiological investigations. When large-scale neural recordings during naturalistic visual search become available, our model can be directly tested for its predictive power in these areas. Furthermore, the general architecture of our model makes it a promising foundation for extension to other cognitively demanding tasks, offering a path toward unified models of high-level cognitive function that can predict neural responses across the prefrontal and parietal cortices, moving beyond the sensory and language domains into the realm of attention, memory, and decision-making. As predictive models continue to advance across domains, our work highlights the value of using goal-driven architectures not merely as engineering tools, but as scientific instruments to probe the computational principles of perception and action in the brain.

Methods

The model was implemented and trained in Python (v3.12; Python Software Foundation, https://www.python.org/) using the PyTorch library (v1.14; https://www.pytorch.org/) on a GPU-accelerated machine equipped with 1–4 NVIDIA RTX A5000 graphics cards. All analyses of human and model data were conducted in Python v.3.12, using the numpy v.1.26.4 and/or scikit-learn v.1.5.2 packages.

Dataset

Training Dataset: Large-Scale Natural Scene Visual Search Dataset Generation. To train GVSM, we constructed a large-scale natural scene visual search dataset based on the Places365 image dataset, which contains over 10 million images of diverse real-world environments Zhou et al. (2017). Each sample in the dataset consists of an arbitrary instance of a cue category (e.g. car), a visual scene that contains at least one instance of the cue category, and a ground truth mask that specifies the position of the cue category in the given scene. To generate the ground truth target masks, we processed each image with a state-of-the-art object detection model (Mask R-CNN) He et al. (2017) and extracted object masks for all detected instances of 80 cue categories (e.g., car, table, kite, person). The resulting cue bank comprised an average of 62,138±257,067.33 exemplars per category, with "person" having the largest set (2,261,195 exemplars) and "hairdryer" the smallest (230 exemplars). To simplify training and avoid multiple simultaneous targets, we merged all object masks of the same category within an image into a single composite mask. Consequently, any region containing an instance of the target cue category was considered part of the target area.

Each (image, cue category) pair served as a distinct trial. The order of presenting the trials was randomized on each pass through the dataset. Task frames were generated on the fly: the cue frame was created by placing a randomly selected exemplar from the cue bank at the center of a black 640×640 pixels canvas, while the search image was resized such that its longer side matched 640 pixels, with the shorter side padded symmetrically to yield a 640 × 640 pixels search frame. Ground-truth target maps were generated by overlaying a 10×10 grid on the cue object mask, thereby discretizing the visual field into 100 spatial bins and reducing the model's action space from pixel-level to grid-level selection. During the initial training phase (supervised training), we used strict ground-truth maps, marking a grid location as a target only if it was fully covered by the object mask. In a subsequent training phase, we adopted a more lenient labeling approach, similar to other visual search models, where any grid cell with partial overlap with the target area was marked as a hit. The final dataset comprised 6,754,405 total trials, drawn from 4,472,151 unique images and 80 cue categories.

Testing Dataset: COCO-Search18. To evaluate the behavioral alignment between GVSM and human visual search behavior, we used the COCO-Search18 dataset Chen et al. (2021). This large-scale benchmark comprises over 6,000 natural images (target-present only) from the MS COCO dataset,

each annotated with eye movement trajectories recorded during visual search. The dataset includes data from 10 human participants who were instructed to search for one of 18 object categories (e.g., microwave, stop sign, backpack) per trial. Each trial contains the sequence of fixations, their durations, and the identity of the search target, offering a detailed account of goal-directed attention and saccadic decision-making in complex real-world scenes.

To generate corresponding search trials for our model, we adapted the same pipeline used for training data generation. Instead of text queries (as used in the human experiments), we constructed cue frames by presenting an exemplar image of the target category at the center of a black 640×640 pixels background. We also derived lenient ground-truth target maps for the model using the object bounding boxes provided in the dataset, assigning a grid location as a target if it overlapped with the ground-truth target area. Our evaluation focused exclusively on the target-present subset, and we assessed both spatial and temporal alignment between the model's predicted fixation sequences and human scanpaths.

GVSM

Retinal transformation. To approximate the progressive loss of visual acuity in the periphery caused by the retinal transformation, we implemented a multi-resolution cropping module adopted from prior work Mnih et al. (2014). For each 640×640 task frame, the module extracts three square crops centered at the model's current fixation location with sizes of 128, 256, and 512 pixels. All three crops are then resized to 128×128 pixels. The original 128×128 pixels crop serves as the high-resolution foveal input, while the resized 256×256pixels and 512×512 pixels crops provide increasingly blurred peripheral representations. This setup mimics the eccentricity-dependent degradation of visual detail in the human retina.

Model Architecture. The model consists of three main components: a visual processing module, an information integration module, and a fixation selection module. In the visual processing module, the visual inputs are processed by an EfficientNet-B0 CNN Koonce (2021), which simulates the visual processing along the ventral visual pathway Schrimpf et al. (2018). From each input image, we extract the final convolutional feature maps before the last fully connected layer, yielding a $1280 \times 4 \times 4$ representation of the scene.

This visual representation is combined with the current fixation location through a Glimpse Network. Fixation coordinates (X,Y) are defined in a normalized retinocentric reference frame, with the image center as the origin. The x-axis increases rightward and the y-axis upward, with both coordinates scaled to lie in the range [-1,1]. The fixation location and the CNN-derived visual features are separately projected to half the embedding dimension (512) using linear layers followed by ReLU nonlinearities. These two projected vectors are then independently transformed by additional linear layers to the full embedding dimension (1024), summed, and passed through another ReLU, yielding the glimpse representation.

The glimpse representation is processed by the Information Integrator module, an 8-layer causal transformer with 8 attention heads and an embedding dimension of 1024. The transformer's final output token, representing the model's latent state, feeds into the Fixation Selector, a multi-layer perceptron (MLP) policy network. This network, composed of N fully connected layers with X nonlinearity, outputs a 100-dimensional probability distribution over the 10×10 spatial grid. The next fixation location is sampled from this distribution. This new fixation location is then passed to the retinal transformation module to generate a foveated input for the next time step. The foveated input re-enters the CNN, initiating another pass through the model. This loop continues until the model either fixates on the target area or reaches the maximum number of time steps (10). The first two steps are fixed: the model first fixates on the cue frame, then at the center of the search image. Thus, the model is allowed to make up to 8 fixation decisions per trial.

Model Training

We first pretrained the CNN to perform object recognition on images from the ImageNet dataset, each subjected to a simulated retinal transformation centered at randomly sampled fixation locations. The task was to classify each image into one of 1,000 categories. Training hyperparameters followed

standard practice in the CNN literature, and in particular, the specifications introduced in the original EfficientNet-B0 paper Koonce (2021) adjusted to the available resources (1 GPU: NVidia V100SXM2) (maximum learning rate = 0.36, weight decay = 10^{-5}, learning rate scheduler = 5 epochs warmup followed by waterfall scheduling, multiplying the learning rate by 0.97 every 2.4 epochs, optimizer = RMSProp Tieleman (2012), batch size = 512). We then adopted a two-phase training paradigm inspired by prior work on saccade-augmented visual categorization Elsayed et al. (2019). This approach mirrors the two-stage training framework used in recent foundation models, such as large language models, in which an initial self-supervised learning phase (here, supervised) is followed by an RL stage Christiano et al. (2017); OpenAI Achiam et al. (2023); Touvron et al. (2023); Baker et al. (2022).

Phase 1: Supervised training. In the first phase, we optimized the parameters of the transformer (with CNN parameters kept frozen) to predict the target area given a random sequence of fixations over the search image. A linear readout layer was applied to the final output token of the transformer to produce a 100-dimensional vector, representing predicted probabilities over a 10×10 grid. This output was trained using binary cross-entropy loss against the strict ground-truth target maps. To address the class imbalance, since most grid locations in a trial do not contain the target, we introduced per-sample, per-class weighting to balance the contribution of positive (on-target) and negative (off-target) locations. The weighted binary cross-entropy loss is defined as:

$$\ell_c(x, y) = -\frac{1}{N}\sum_{n=1}^{N} w_{n,c}\left[y_{n,c} \cdot \log\sigma(x_{n,c}) + (1 - y_{n,c}) \cdot \log(1 - \sigma(x_{n,c}))\right] \tag{1}$$

where $\sigma(.)$ is the sigmoid function, N is the batch size, $C = 100$ is the number of classes (grid locations), $x \in R^{N\times C}$ are the model logits, $y \in \{0, 1\}, N \times C$ are the binary labels (1 if on the target, 0 otherwise), $w_{n,c}$ is the weight for sample n and class c defined as:

$$w_{n,c} = \begin{cases} \frac{C}{\sum_{c'=1}^{C} y_{n,c'}} & \text{if } y_{n,c} = 1 \\ \frac{C}{C-\sum_{c'=1}^{C} y_{n,c'}} & \text{if } y_{n,c} = 0 \end{cases} \tag{2}$$

Here, $\sum_{c'=1}^{C} y_{n,c'}$ is the number of positive labels (on-target) for sample n. This formulation ensures comparable loss magnitude across images with large differences in target sizes. for each trial, the total loss contribution from on-target and off-target locations is balanced, regardless of their relative frequency.

Phase 2: reinforcement learning. In the second phase, we trained the Fixation Selector, a multi-layer perceptron (MLP) policy network, to sequentially choose fixation locations based on the transformer's final output token. During this stage, the parameters of the CNN and transformer modules were frozen. The MLP received the 1024-dimensional transformer output and produced a 100-dimensional probability distribution over the grid locations. The fixations were sampled from this distribution over grid locations using softmax sampling. We used the Proximal Policy Optimization (PPO) algorithm Schulman et al. (2017) to optimize the fixation policy. A reward of +10 was assigned if the selected fixation landed on the target area (as defined by the lenient target map), and a penalty of -1 was given for each step taken without finding the target. This reward structure encouraged the model to locate the target as efficiently as possible. To maximize training efficiency and avoid uninformative episodes, we curated a subset of the trials from Phase 1 for RL. We excluded trials in which the target was excessively large or small, as oversized targets can inflate cumulative rewards without improving performance in realistic conditions. Target size thresholds were matched to the minimum and maximum target sizes observed in COCO-Search18. We also limited cue categories to the 18 used in COCO-Search18 Despite being trained on this curated subset, the RL agent successfully generalized to the full dataset used in Phase 1, owing to its exposure to the most informative and representative trials during training.

Behavioral and Latent Space Analysis.

Baselines. To ensure a fair comparison between our model and baseline models, we closely matched the target-found criteria and other evaluation conditions to those used in GVSM. First, we aligned

the fixation box size across models. The fixation box refers to the area around the fixation point that is considered "seen." If this area overlaps with the target, the search is deemed successful; otherwise, the area is inhibited, and its probability of being fixated again is set to zero. While our model does not implement explicit inhibition of return, its action space is restricted to a 10×10 grid, making each 64×64-pixel location analogous to fixation boxes in IVSN and eccNET, and to grid cells in IRL. To standardize evaluation, we matched the fixation-box-to-image ratio across all baselines to that of our model ($64/640 = 0.1$). Using the original fixation sizes of the baselines did not meaningfully alter results. We used the publicly available implementations of the baseline models and the fixation-by-fixation scanpath prediction method provided by Travi et al. (2022). Note that IRL received cue categories in textual form, while IVSN, eccNET, and our model used visual exemplars of the cue. To further ensure consistency, we mapped all scanpaths, human and model, to a common reference frame: a 640×640 pixel image space discretized into a 10×10 grid. We reduced consecutive fixations falling within the same grid cell to a single fixation at that location. The same target-found criterion was applied: if the fixated grid location overlapped with the target region, the target was considered found and the trial ended. This alignment of evaluation conditions did not significantly alter the performance metrics or scanpath prediction scores for human participants or baseline models.

Fixation-by-fixation evaluation of search scanpath predictions. To evaluate how well the models predicted human visual search behavior at a fine-grained level, we used a fixation-by-fixation scanpath prediction method Kümmerer and Bethge (2021). This approach represents the state-of-the-art in scanpath comparison and addresses several critical limitations of traditional sequence-based metrics, which typically treat scanpaths as holistic sequences and rely on distance-based similarity measures. In contrast, the fixation-by-fixation approach (1) better reflects the decision-making process underlying eye movements by treating each fixation as a choice conditioned on previous ones, (2) aligns with standard evaluation methods used in saliency map prediction, (3) enables detailed behavioral alignment analyses by scoring individual fixations rather than entire scanpaths (as shown in Fig. 5), and (4) is free of hyperparameters. Notably, sequence-based metrics failed to distinguish between models in our setting.

In this method, human fixations are fed into the model sequentially; at each step, the model generates a fixation probability map conditioned on the history of prior human fixations. This conditional output is then compared to the actual human fixation at that time step using four standard metrics borrowed from the saliency prediction literature: Area Under the Curve (AUC), Normalized Scanpath Similarity (NSS), Log-Likelihood (LL), and Information Gain (IG). AUC treats the predicted fixation map as a classifier and evaluates its ability to discriminate the ground-truth human fixation from other locations. NSS computes the value of the predicted map at the true fixation location, normalized by the map's mean and standard deviation. IG quantifies the added information provided by the model relative to a center-bias baseline by comparing the log-probabilities assigned to the fixation location. Positive IG values indicate better model alignment with human gaze behavior. Finally, the LL metric is defined as the information gain over a uniform baseline.

Comparing distributions of fixation directions. To compare the distribution of fixation directions between models and humans, we computed the Jensen–Shannon (JS) divergence between their respective direction histograms Lin (2002). Given two normalized distributions P (model) and Q (human), the JS divergence is defined as:

$$\mathrm{JS}(P \parallel Q) = \frac{1}{2}\mathrm{KL}(P \parallel M) + \frac{1}{2}\mathrm{KL}(Q \parallel M), \quad \text{where } M = \frac{1}{2}(P + Q) \tag{3}$$

where the Kullback–Leibler (KL) divergence is given by Kullback and Leibler (1951):

$$\mathrm{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{4}$$

Unlike KL divergence, which is asymmetric and can be undefined when Q(i)=0, JS divergence is symmetric, smooth, and always finite. These properties make it a robust and interpretable measure of dissimilarity between distributions. A lower JS divergence indicates that the model's fixation direction distribution more closely resembles that of humans.

Computing retinocentric and allocentric cue-similarity maps. To assess the similarity between the cue object and various locations on the search image, we convolve the visual representation of the cue frame (i.e., the final feature maps of the CNN before the global pooling and the classifier fully connected head) with the image Zhang et al. (2018). The images, retinally transformed in the retinocentric case so 3-channel 128×128 and intact in the allocentric case so 1-channel 640×640 (grayscale) are processed by the EfficientNet-B0 CNN. In the retinocentric case, the CNN is the same as the one used as the vision module in the model, pretrained on retinal-transformed ImageNet to do object recognition. In the allocentric case, it is a pretrained CNN on object recognition on ImageNet with no retinal transformation. The resulting retinocentric and allocentric cue-similarity maps are 3×3 and 11×11, respectively. Thus, the allocentric maps are downsampled using bilinear interpolation to match the lower resolution of the retinocentric maps to equalize the conditions for comparing the decoding accuracies.

Latent decoding analysis. To examine the representational content of the model's latent state, we conducted linear decoding analyses using ridge regression. We fitted linear decoders to predict various target variables, retinocentric and allocentric cue-similarity maps, fixation probability maps, and (X,Y) coordinates of visited locations from the model's latent activations at each time step. Specifically, for each model instance, a separate decoder was trained for each spatial location (i.e., each cell in a 10×10 grid over the image), using the output token of the Information Integrator module as input features. Fitting was performed with 5-fold cross-validation, and decoding accuracy was quantified as the Spearman correlation between predicted and ground truth values on held-out trials. To evaluate subspace generalization, such as from cue-similarity maps to fixation probability maps, across time steps, or across spatial locations, we tested trained decoders on mismatched targets and compared their performance to within-target decoding baselines.

Orthogonal Procrustes analysis. To examine the geometrical structure of neural representations in the model's latent space, we used orthogonal Procrustes analysis to quantify how well one set of decoder weight vectors could be aligned with another via rotation. Each set of weights, corresponding to either cue-similarity maps or fixation probability maps, comprised 100 vectors (one per 10×10 grid cell), trained to predict the value of the respective spatial map at each location based on the latent activation at a given time step. We used this framework in two distinct ways: (1) to assess whether the representation of cue-similarity maps rotated across time, i.e., whether the encoding subspace at one time point could be aligned to that of another via a rigid rotation; and (2) to test whether the cue-similarity encoding subspace could be rotated to align with the fixation probability map encoding subspace, potentially reflecting a transformation underlying fixation selection.

For each case, we first centered and normalized the source and target sets of weight vectors. We then applied orthogonal Procrustes analysis to compute the optimal rotation matrix and global scaling factor, aligning the source to the target. The transformed source weights were then inverse-transformed to recover their original scale, and alignment quality was quantified by evaluating the decoding performance of the reconstructed decoders on the target prediction task. This approach allowed us to test both the temporal stability of cue-similarity representations and their geometric relationship to the representations driving fixation behavior.

Acknowledgments

This research was supported by the Healthy-Brains-Healthy-Lives startup supplement grant, the NSERC Discovery grant RGPIN-2021-03035, and CIHR Project Grant PJT-191957. P.B. was supported by FRQ-S Research Scholars Junior 1 grant 310924, and the William Dawson Scholar award. M.P. was supported by the UNIQUE Masters and PhD Fellowships, and the Stichting Formation Award. All analyses were executed using resources provided by the Digital Research Alliance of Canada (Compute Canada) and funding from Canada Foundation for Innovation project number 42730.

References

Adeli, H., Vitu, F., and Zelinsky, G. J. (2017). A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *Journal of Neuroscience*, 37(6):1453–1467.

Adeli, H. and Zelinsky, G. (2018). Deep-bcn: Deep networks meet biased competition to create a brain-inspired model of attention control. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1932–1942.

Altpeter, E., Mackeben, M., and Trauzettel-Klosinski, S. (2000). The importance of sustained attention for patients with maculopathies. *Vision Research*, 40(10-12):1539–1547.

Bahill, A., Adler, D., and Stark, L. (1975a). Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology & Visual Science*, 14(6):468–469.

Bahill, A. T., Adler, D., and Stark, L. (1975b). Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology & Visual Science*, 14(6):468–469.

Baker, B., Akkaya, I., Zhokov, P., Huizinga, J., Tang, J., Ecoffet, A., Houghton, B., Sampedro, R., and Clune, J. (2022). Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654.

Barbot, A., Xue, S., and Carrasco, M. (2021). Asymmetries in visual acuity around the visual field. *Journal of vision*, 21(1):2–2.

Bashivan, P., Kar, K., and DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436.

Bella-Fernández, M., Suero Suñé, M., and Gil-Gómez de Liaño, B. (2022). Foraging behavior in visual search: A review of theoretical and mathematical models in humans and animals. *Psychological research*, 86(2):331–349.

Bichot, N. P., Heard, M. T., DeGennaro, E. M., and Desimone, R. (2015). A source for feature-based attention in the prefrontal cortex. *Neuron*, 88(4):832–844.

Bichot, N. P. and Schall, J. D. (1999). Effects of similarity and history on neural mechanisms of visual selection. *Nature neuroscience*, 2(6):549–554.

Bisley, J. W. and Mirpour, K. (2019). The neural instantiation of a priority map. *Current opinion in psychology*, 29:108–112.

Bujia, G., Sclar, M., Vita, S., Solovey, G., and Kamienkowski, J. E. (2022). Modeling human visual search in natural scenes: a combined bayesian searcher and saliency map approach. *Frontiers in Systems Neuroscience*, 16:882315.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., and DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision research*, 51(13):1484–1525.

Carrasco, M., Giordano, A. M., and McElree, B. (2004). Temporal performance fields: Visual and attentional factors. *Vision research*, 44(12):1351–1365.

Caucheteux, C. and King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134.

Chaikin, J. D., Corbin, H. H., and Volkmann, J. (1962). Mapping a field of short-time visual search. *Science*, 138(3547):1327–1328.

Chan, L. K. and Hayward, W. G. (2013). Visual search. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(4):415–429.

Chelazzi, L., Miller, E. K., Duncan, J., and Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature*, 363(6427):345–347.

Chen, Y., Yang, Z., Ahn, S., Samaras, D., Hoai, M., and Zelinsky, G. (2021). Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1):8776.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Corbett, J. E. and Carrasco, M. (2011). Visual performance fields: Frames of reference. *PLoS One*, 6(9):e24470.

Curcio, C. A., Sloan, K. R., Kalina, R. E., and Hendrickson, A. E. (1990). Human photoreceptor topography. *Journal of comparative neurology*, 292(4):497–523.

Deco, G. and Heinke, D. (2007). Attention and spatial resolution: A theoretical and experimental study of visual search in hierarchical patterns. *Perception*, 36(3):335–354.

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

Desimone, R., Duncan, J., et al. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222.

Ding, Z., Ren, X., David, E., Vo, M., Kreiman, G., and Zhang, M. (2022). Efficient zero-shot visual search via target and context-aware transformer. *arXiv preprint arXiv:2211.13470*.

Duncan, J. and Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological review*, 96(3):433.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of vision*, 11(5):14–14.

Egeth, H. E. and Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual review of psychology*, 48(1):269–297.

Elsayed, G., Kornblith, S., and Le, Q. V. (2019). Saccader: Improving accuracy of hard attention models for vision. *Advances in neural information processing systems*, 32.

Findlay, J. M. and Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Number 37. Oxford University Press.

Gajewski, D. A., Pearson, A. M., Mack, M. L., Bartlett III, F. N., and Henderson, J. M. (2004). Human gaze control in real world search. In *International Workshop on Attention and Performance in Computational Vision*, pages 83–99. Springer.

Gelade, G. (2001). A feature-integration theory of attention. *Visual perception: Essential readings*, page 347.

Gupta, S. K., Zhang, M., Wu, C.-C., Wolfe, J., and Kreiman, G. (2021). Visual search asymmetry: Deep nets and humans share similar inherent biases. *Advances in neural information processing systems*, 34:6946–6959.

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

Ischebeck, A., Kreilinger, H., Peiris Miller, J., Höfler, M., Gilchrist, I. D., and Körner, C. (2024). Fixating targets in visual search: The role of dorsal and ventral attention networks in the processing of relevance and rarity. *Imaging Neuroscience*, 2:1–16.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203.

Iyer, A. and Burge, J. (2019). The statistics of how natural images drive the responses of neurons. *Journal of Vision*, 19(13):4–4.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915.

Klein, R. (1988). Inhibitory tagging system facilitates visual search. *Nature*, 334(6181):430–431.

Klein, R. M. (2000). Inhibition of return. *Trends in cognitive sciences*, 4(4):138–147.

Koevoet, D., Van Zantwijk, L., Naber, M., Mathôt, S., Van der Stigchel, S., and Strauch, C. (2025). Effort drives saccade selection. *ELife*, 13:RP97760.

Koonce, B. (2021). Efficientnet. In *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization*, pages 109–123. Springer.

Kristjánsson, Á. and Draschkow, D. (2021). Keeping it real: Looking beyond capacity limits in visual cognition. *Attention, Perception, & Psychophysics*, 83:1375–1390.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Kümmerer, M. and Bethge, M. (2021). State-of-the-art in human scanpath prediction. *arXiv preprint arXiv:2102.12239*.

Kupers, E. R., Benson, N. C., Carrasco, M., and Winawer, J. (2022). Asymmetries around the visual field: From retina to cortex to behavior. *PLoS computational biology*, 18(1):e1009771.

Lei, X., Ito, T., and Bashivan, P. (2024). Geometry of naturalistic object representations in recurrent neural network models of working memory. *Advances in Neural Information Processing Systems*, 37:100604–100629.

Li, F., Kant, S., Araki, S., Bangera, S., and Shukla, S. S. (2020). Neural networks for fashion image classification and visual search. *arXiv preprint arXiv:2005.08170*.

Lin, J. (2002). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Lleras, A., Buetti, S., and Xu, Z. J. (2022). Incorporating the properties of peripheral vision into theories of visual search. *Nature Reviews Psychology*, 1(10):590–604.

Luck, S. J. and Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–281.

Machner, B., Lencer, M. C., Möller, L., Von der Gablentz, J., Heide, W., Helmchen, C., and Sprenger, A. (2020). Unbalancing the attentional priority map via gaze-contingent displays induces neglect-like visual exploration. *Frontiers in Human Neuroscience*, 14:41.

Mackeben, M. (1999). Sustained focal attention and peripheral letter recognition. *Spatial vision*, 12(1):51–72.

Malcolm, G. L. and Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision*, 9(11):8–8.

Mazer, J. A. and Gallant, J. L. (2003). Goal-related activity in v4 during free viewing visual search: Evidence for a ventral stream visual salience map. *Neuron*, 40(6):1241–1250.

Mitroff, S. R. and Biggs, A. T. (2014). The ultra-rare-item effect: Visual search for exceedingly rare items is highly susceptible to error. *Psychological science*, 25(1):284–289.

Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. (2014). Recurrent models of visual attention. *Advances in neural information processing systems*, 27.

Mondal, S., Yang, Z., Ahn, S., Samaras, D., Zelinsky, G., and Hoai, M. (2023). Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1441–1450.

Najemnik, J. and Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391.

Navalpakkam, V. and Itti, L. (2005). Modeling the influence of task on attention. *Vision research*, 45(2):205–231.

Navalpakkam, V. and Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimizing detection speed. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2049–2056. IEEE.

OpenAI Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. arxiv. *arXiv preprint arXiv:2303.08774*.

Pelz, J. B. and Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision research*, 41(25-26):3587–3596.

Phillips, A. N. and Mark, A. (2006). Predictive activity in macaque frontal eye field neurons during 5. *Neural Netw*, 19:1467–1474.

Pomplun, M. and Gray, W. (2007). Advancing area activation toward a general model of eye movements in visual search.

Posner, M. I., Rafal, R. D., Choate, L. S., and Vaughan, J. (1985). Inhibition of return: Neural basis and function. *Cognitive neuropsychology*, 2(3):211–228.

Powell, K. D., Colby, C. L., Gottlieb, J., Kusunoki, M., and Goldberg, M. E. (1999). Space and salience in parietal cortex. *Current Oculomotor Research: Physiological and Psychological Aspects*, pages 25–35.

Radulescu, A., van Opheusden, B., Callaway, F., Griffiths, T. L., and Hillis, J. M. (2022). Modeling human eye movements during immersive visual search. *bioRxiv*, pages 2022–12.

Rao, R. P., Zelinsky, G. J., Hayhoe, M. M., and Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision research*, 42(11):1447–1463.

Rashidi, S., Xu, W., Lin, D., Turpin, A., Kulik, L., and Ehinger, K. (2023). An active foveated gaze prediction algorithm based on a bayesian ideal observer. *Pattern Recognition*, 143:109694.

Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J., and Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications*, 12(1):5540.

Rossi, E. A. and Roorda, A. (2010). The relationship between visual resolution and cone spacing in the human fovea. *Nature neuroscience*, 13(2):156–157.

Sapountzis, P., Paneri, S., and Gregoriou, G. G. (2018). Distinct roles of prefrontal and parietal areas in the encoding of attentional priority. *Proceedings of the National Academy of Sciences*, 115(37):E8755–E8764.

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007.

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

Shen, K. and Paré, M. (2014). Predictive saccade target selection in superior colliculus during visual search. *Journal of Neuroscience*, 34(16):5640–5648.

Smith, T. J. and Henderson, J. M. (2009). Facilitation of return during scene viewing. *Visual Cognition*, 17(6-7):1083–1108.

Taleb, F., Vasco, M., Ribeiro, A., Björkman, M., and Kragic, D. (2024). Can transformers smell like humans? *Advances in Neural Information Processing Systems*, 37:72032–72060.

Tatler, B. W., Baddeley, R. J., and Vincent, B. T. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision research*, 46(12):1857–1862.

Tatler, B. W. and Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17(6-7):1029–1054.

Tieleman, T. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26.

Torralba, A. and Oliva, A. (2003). Statistics of natural image categories. *Network: computation in neural systems*, 14(3):391.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Travi, F., Ruarte, G., Bujia, G., and Kamienkowski, J. E. (2022). Visions: visual search in natural scenes benchmark. *Advances in Neural Information Processing Systems*, 35:11987–12000.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wandell, B. A. (1995). *Foundations of vision*. sinauer Associates.

Wang, M., Fusi, S., and Stachenfeld, K. (2025). Brain-like slot representation for sequence working memory in recurrent neural networks. In *Second Workshop on Representational Alignment at ICLR 2025*.

Wang, P. Y., Sun, Y., Axel, R., Abbott, L., and Yang, G. R. (2021). Evolving the olfactory system with machine learning. *Neuron*, 109(23):3879–3892.

Watson, A. B. (2014). A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of vision*, 14(7):15–15.

Wei, Z., Adeli, H., Nguyen, M. H., Zelinsky, G., and Samaras, D. (2016). Learned region sparsity and diversity also predicts visual attention. *Advances in Neural Information Processing Systems*, 29.

Whittington, J. C., Dorrell, W., Behrens, T. E., Ganguli, S., and El-Gaby, M. (2025). A tale of two algorithms: Structured slots explain prefrontal sequence memory and are unified with hippocampal cognitive maps. *Neuron*, 113(2):321–333.

Wilming, N., Harst, S., Schmidt, N., and König, P. (2013). Saccadic momentum and facilitation of return saccades contribute to an optimal foraging strategy. *PLoS computational biology*, 9(1):e1002871.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1:202–238.

Wolfe, J. M. (2018). Visual search. *Stevens' handbook of experimental psychology and cognitive neuroscience*, 2:1–55.

Wolfe, J. M. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic bulletin & review*, 28(4):1060–1092.

Wolfe, J. M. and Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature reviews neuroscience*, 5(6):495–501.

Wolfe, J. M. and Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature human behaviour*, 1(3):0058.

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

Yang, F., Kale, A., Bubnov, Y., Stein, L., Wang, Q., Kiapour, H., and Piramuthu, R. (2017). Visual search at ebay. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2101–2110.

Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., Samaras, D., and Hoai, M. (2020). Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 193–202.

Yang, Z., Mondal, S., Ahn, S., Xue, R., Zelinsky, G., Hoai, M., and Samaras, D. (2024). Unifying top-down and bottom-up scanpath prediction using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1683–1693.

Yang, Z., Mondal, S., Ahn, S., Zelinsky, G., Hoai, M., and Samaras, D. (2022). Target-absent human attention. In *European Conference on Computer Vision*, pages 52–68. Springer.

Yuan, A. and Li, Y. (2020). Modeling human visual search performance on realistic webpages using analytical and deep learning methods. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12.

Zelinsky, G., Yang, Z., Huang, L., Chen, Y., Ahn, S., Wei, Z., Adeli, H., Samaras, D., and Hoai, M. (2019). Benchmarking gaze prediction for categorical visual search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological review*, 115(4):787.

Zelinsky, G. J., Peng, Y., Berg, A. C., and Samaras, D. (2013). Modeling guidance and recognition in categorical search: Bridging human and computer object detection. *Journal of Vision*, 13(3):30–30.

Zhang, M., Armendariz, M., Xiao, W., Rose, O., Bendtz, K., Livingstone, M., Ponce, C., and Kreiman, G. (2022). Look twice: A generalist computational model predicts return fixations across tasks and species. *PLoS computational biology*, 18(11):e1010654.

Zhang, M., Feng, J., Ma, K. T., Lim, J. H., Zhao, Q., and Kreiman, G. (2018). Finding any waldo with zero-shot invariant and efficient visual search. *Nature communications*, 9(1):3730.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.