A robotics-inspired scanpath model reveals the importance of uncertainty and semantic object cues for gaze guidance in dynamic scenes

Vito Mengers[*] Technische Universität Berlin, Berlin, Germany
Science of Intelligence, Research Cluster of Excellence, Berlin, Germany

Nicolas Roth[*] Technische Universität Berlin, Berlin, Germany
Science of Intelligence, Research Cluster of Excellence, Berlin, Germany

Oliver Brock[#] Technische Universität Berlin, Berlin, Germany
Science of Intelligence, Research Cluster of Excellence, Berlin, Germany

Klaus Obermayer[#] Technische Universität Berlin, Berlin, Germany
Science of Intelligence, Research Cluster of Excellence, Berlin, Germany

Martin Rolfs[#] Humboldt-Universtät zu Berlin, Berlin, Germany
Science of Intelligence, Research Cluster of Excellence, Berlin, Germany

The objects we perceive guide our eye movements when observing real-world dynamic scenes. Yet, gaze shifts and selective attention are critical for perceiving details and refining object boundaries. Object segmentation and gaze behavior are, however, typically treated as two independent processes. Here, we present a computational model that simulates these processes in an interconnected manner and allows for hypothesis-driven investigations of distinct attentional mechanisms. Drawing on an information processing pattern from robotics, we use a Bayesian filter to recursively segment the scene, which also provides an uncertainty estimate for the object boundaries that we use to guide active scene exploration. We demonstrate that this model closely resembles observers' free viewing behavior on a dataset of dynamic real-world scenes, measured by scanpath statistics, including foveation duration and saccade amplitude distributions used for parameter fitting and higher-level statistics not used for fitting. These include how object detections, inspections, and returns are balanced and a delay of returning saccades without an explicit implementation of such temporal inhibition of return. Extensive simulations and ablation studies show that uncertainty promotes balanced exploration and that semantic object cues are crucial to forming the perceptual units used in object-based attention. Moreover, we show how our model's modular design allows for extensions, such as incorporating saccadic momentum or presaccadic attention, to further align its output with human scanpaths.

Introduction

Humans actively move their eyes to pay attention to individual parts of their environment. Several seminal studies have explored eye movements in natural contexts (Land & Lee, 1994; Land, Mennie, & Rusted, 1999; Pelz, Hayhoe, & Loeber, 2001; Triesch, Ballard, Hayhoe, & Sullivan, 2003; Rothkopf, Ballard, & Hayhoe, 2007; Tatler, Hayhoe, Land, & Ballard, 2011; Mital, Smith, Hill, & Henderson, 2011; Matthis, Yates, & Hayhoe, 2018), yet we lack a mechanistic understanding of

Citation: Mengers, V., Roth, N., Brock, O., Obermayer, K., & Rolfs, M. (2025). A robotics-inspired scanpath model reveals the importance of uncertainty and semantic object cues for gaze guidance in dynamic scenes. *Journal of Vision*, *25*(2):6, 1–33, https://doi.org/10.1167/jov.25.2.6.

https://doi.org/10.1167/jov.25.2.6 Received July 31, 2024; published February 10, 2025 ISSN 1534-7362

gaze control in such natural conditions. Computational models of visual attention provide an invaluable tool to analyze the contributions of distinct mechanisms and link them to observable behavior (Itti & Koch, 2001; Borji & Itti, 2012; Roth, Rolfs, Hellwich, & Obermayer, 2023; Kümmerer & Bethge, 2023). In this work, we present an object-based computational model that reproduces human free-viewing eye-tracking data (with a stationary head position) when observing natural dynamic scenes. In addition to the saccadic decision-making process, we also model how the basic building blocks—on which object-based attention can act—can be formed. Our model is mechanistic in the sense that it implements algorithmic principles behind attentional mechanisms. Specifically, we aim to capture how information is integrated to determine the next saccade target and how different object cues contribute to the formation of perceptual units for object-based attention. However, we do not prioritize the plausibility of how the inputs to these mechanisms are computed in the first place, nor do we make claims about the neural implementation of these mechanisms in the brain. The model's modularity then allows us to systematically test the effect and contribution of different attentional mechanisms on the simulated gaze behavior, which can be directly compared with human eye-tracking results.

Visual attention sequentially selects objects for perceptual processing and provides the information to generate a motor plan for eye movements (Deubel & Schneider, 1996). Different psychophysical experiments have, depending on the task and presented visual stimulus, uncovered different aspects of attention (for an overview, see Carrasco, 2011; Nobre & Kastner, 2014). The most prominent theories of visual attention describe it as space, feature, or object based. Space-based attention is classically characterized as a spotlight (Posner, 1980) or zoom lens (Eriksen & Yeh, 1985) that enhances processing at the attended location. The attended location is typically selected based on maxima in a priority or saliency map (Koch & Ullman, 1985; Itti & Koch, 2001). Independent of a specific location, feature-based attention can be deployed covertly to objects that share a specific attribute, like color or motion direction (Treue & Trujillo, 1999; Saenz, Buracas, & Boynton, 2002; White & Carrasco, 2011). Evidence for object-based attention was, for example, found in experiments where attention was allocated to one of two objects that share the same location (Duncan, 1984; O'Craven, Downing, & Kanwisher, 1999; Blaser, Pylyshyn, & Holcombe, 2000) and where attention was directed faster to locations within an attended object than to locations outside the object (Egly, Driver, & Rafal, 1994; Malcolm & Shomstein, 2015). The object specificity of attention suggests that, at least in some cases, the underlying units of attentional processing and selection are discrete visual objects (for reviews, see Scholl, 2001; Peters & Kriegeskorte, 2021). Cavanagh et al. (2023) presented a compelling framework for how experimental findings attributed to space- or feature-based attention can be conceptualized as forms of object-based attention. We have previously demonstrated using a computational modeling approach that objects are particularly important for gaze guidance during free viewing of dynamic natural scenes (Roth et al., 2023).

When simulating human eye movements in natural scenes, models are typically limited in at least one of two ways: modeling only the average spatial gaze density instead of individual scanpaths, or being only applicable to static images instead of videos. Classic saliency models have been extended to include motion (e.g., Molin, Etienne-Cummings, & Niebur, 2015), and deep learning models have been used successfully for video saliency prediction (e.g., Wang, Shen, Guo, Cheng, & Borji, 2018; Droste, Jiao, & Noble, 2020). However, saliency models are restricted to modeling the average spatial distribution of gaze positions. Models capable of describing the attentional dynamics of individual saccadic decisions usually assume the scene to be static and are not applicable to dynamic scenes (e.g., Itti, Koch, & Niebur, 1998; Tatler, Brockmole, & Carpenter, 2017; Wloka, Kotseruba, & Tsotsos, 2018; Schwetlick, Rothkegel, Trukenbrod, & Engbert, 2020; Kümmerer, Bethge, & Wallis, 2022). Rather than relying on these simplifications of common models (for reviews, see Borji & Itti, 2012; Bylinskii et al., 2015; Kümmerer & Bethge, 2023), our approach predicts full scanpaths, including the order and timing of fixation and smooth pursuit events, for dynamic videos. Our previous scanpath model (Roth et al., 2023) describes the saccadic decision-making processes during the free-viewing of dynamic scenes but requires explicitly provided object segmentations for modeling object-based attention. How the building blocks of object-based attention arise before being actively attended and what mechanisms contribute to the formation of these perceptual units are, however, open questions (Wagemans et al., 2012).

Classic theories of the visual system propose that visual processing involves organizing elements of the scene into coherent units through structured operations (Ullman, 1984; Bundesen, 1990). To describe what object-based attention can act on, "proto-objects" were introduced as pre-attentive volatile units that can be accessed and further shaped by selective attention (Rensink, 2000). Walther and Koch (2006) proposed a model that generates such proto-objects for static scenes based on salient regions defined based on color, edges, and luminance. In contrast, psychophysical studies showed that saliency-based proto-objects are less predictive of where people look in real-world scenes than semantically defined objects (Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013). This suggests that pre-attentive objects can also be formed based on semantics and do not rely solely on low-level saliency. In the same vein, human reconstruction of

local image regions is controlled by semantic object boundaries, which are constructed within 100 ms of scene viewing (Liu, Agam, Madsen, & Kreiman, 2009; Neri, 2017), whereas rapid serial visual representation tasks show that scene identification can be even faster (Potter, Wyble, Hagmann, & McCourt, 2014). Although object boundaries are formed globally, the recognition of individual objects and perceiving their visual details still requires selective attention (Wolfe, 1994; Henderson, 2003; Underwood, Templeman, Lamming, & Foulsham, 2008; Wolfe, 2021) and the confidence in information about the foveated objects increases (Stewart, Ludwig, & Schütz, 2022).

Because perceived objects guide eye movements while gaze shifts influence object perception, the modeling of object-based saccadic decisions requires linking the two interdependent processes. Such interdependences pose a challenge for many modeling approaches that tend to treat model components as almost independent. A similar challenge exists in robotics, where a robot usually needs to decide on actions given the highly interdependent information from its different sensors (Eppner et al., 2016). Therefore, we model the interdependent segmentation and saccadic decision-making by using an information processing pattern from robotics, called Active InterCONnect (AICON; see Battaje, Godinez, Hanning, Rolfs, & Brock, 2024), which has been applied to robustly solve such problems for real-world robotic systems (Martín-Martín & Brock, 2022). It is centered around building bidirectional connections between components that allow for the interpretation of sensory cues while taking into account the extracted information from other components. In a recent example of this approach, we combined motion and appearance segmentation of objects to disambiguate each cue (Mengers, Battaje, Baum, & Brock, 2023). By additionally extracting kinematic object motion constraints from their observed motion, predicting their future motion becomes easier (Martín-Martín & Brock, 2022), which in turn simplifies segmenting them (Mengers et al., 2023). Although these bidirectional interactions of components are similar to the top-down influence of higher abstractions on low-level visual processing in reverse hierarchy theory (Ahissar & Hochstein, 2004) or interpretation-guided segmentation (Tenenbaum & Barrow, 1977), they become more informative by estimating the uncertainty of each component's extracted information. This way, the information of different components can be weighted in their connections, and the robot can act according to the current uncertainty, for example, by moving more carefully or actively obtaining more information (Bohg et al., 2017).

We transfer this idea to the modeling of interdependent segmentation and visual exploration in dynamic real-world scenes: The components for visual target selection and segmentation of the scene are in an active interconnection regulated by uncertainty. Segmented objects can act as uncertain perceptual units for target selection, while moving the gaze toward a particular object can resolve the uncertainty over its segmentation. The initial segmentation of a presented scene is estimated globally, meaning that objects that have not yet been foveated are also segmented throughout the visual field (cf., Neisser, 1967). We build on psychophysical evidence, showing that an initial global scene segmentation can be obtained already within the first fixation based on low-level appearance (Schyns & Oliva, 1994), motion (Reppas, Niyogi, Dale, Sereno, & Tootell, 1997), and semantic (Neri, 2017) object cues. These pre-attentive object boundaries are sequentially refined through high-quality segmentation masks of the actively attended (i.e., foveated) objects (Henderson, 2003). We treat these different sources of object information as inherently uncertain cues, which we combine in a Bayes filter, a recurrent mechanism that optimally combines the different input sources and updates its compressed representation based on new measurements over time (Särkkä, 2013). Similar to the related Bayes filter for object segmentation in robotics (Mengers et al., 2023), the tracked uncertainty over the segmentation is estimated based on the agreement of its measurements over time. Thus, this uncertainty describes where the existence or location of boundaries between objects is ambiguous (for more details, see Appendix A). Combined with other scene features, like visual saliency, this uncertainty about the object segmentation drives the active exploration of the scene and contributes to the saccadic decision-making process. The high-resolution semantic segmentation of the object at the current gaze position, in turn, provides a high-confidence measurement and updates the object representation in the Bayes filter. This reduces the uncertainty at the current gaze position and encourages further exploration of other parts of the scene.

The automatic generation of an uncertainty map as a result of our object segmentation hence provides us with an advantage over existing mechanistic computational models of visual attention. They typically rely on an explicitly implemented mechanism, called inhibition of return (IOR), to propel exploration (cf., Itti and Koch, 2001; Zelinsky, 2008; Schwetlick et al., 2020; Roth et al., 2023). IOR as an attentional effect was first described by Posner and Cohen (1984) as the temporary inhibition of the visual processing of recently attended scene parts. Although the initial experiment did not involve eye movements, subsequent studies have found a temporal delay of return saccades (*temporal* IOR; cf., Luke, Smith, Schmidt, & Henderson, 2014) and that saccades are spatially biased away from previously attended locations (*spatial* IOR; cf., Klein & MacInnes, 1999). These effects were interpreted as a foraging factor to encourage attentional orientation to previously

unexplored parts of the scene (Klein & MacInnes, 1999; Klein, 2000). Itti, Koch, and Niebur (1998) hence used IOR as a convenient mechanism to inhibit locations in the saliency map to prevent their model from repeatedly selecting the same most salient location. Including this inhibition subsequently became the de facto standard for mechanistic scanpath models. However, mounting evidence suggests that IOR effects observed in cueing tasks (Posner & Cohen, 1984; Tipper, Driver, & Weaver, 1991) do not play a significant role in gaze behavior under most conditions: Fixation distributions in scene viewing and visual search actually find an increased probability of returns and an absence of spatial IOR (Hooge, Over, van Wezel, & Frens, 2005; Smith & Henderson, 2009; Smith & Henderson, 2011). The effect of temporal IOR in scene viewing has been explained by Wilming, Harst, Schmidt, and König (2013) through "saccadic momentum," a general dependency of fixation durations and subsequent relative saccade angles tendency for saccades to continue the trajectory of the last saccade (Anderson, Yadav, & Carpenter, 2008; Smith & Henderson, 2009).

In the present work, we propose a mechanistic computational scanpath model that does not rely on active IOR as a mechanism to drive scene exploration. Instead, we used a close interaction between the object segmentation and the saccadic decision-making processes to leverage uncertainty over the object boundaries in the scene to encourage exploration. We show that these interconnected processes lead to human-like gaze behavior for dynamic real-world scenes. The modular implementation of our model allows for principled hypothesis testing by analyzing the influence of different implementations on the simulated gaze behavior. We systematically explore the influence of the object uncertainty on the model scanpaths and find that it leads to an exploration behavior that closely resembles the human data. It even reproduces the temporal IOR effect without the need for an explicit IOR implementation. Moreover, we show that access to high-level object information leads to more realistic scanpaths, suggesting that perceptual units of human attention are shaped by semantic knowledge. Finally, we demonstrate how the model can easily be extended to include additional mechanisms like saccadic momentum and presaccadic attention.

Materials and methods

A model for interdependent saccadic decisions and object segmentation

We propose a model for the two processes of saccadic decision-making and object segmentation in natural scenes. To establish an active interconnection between them, we use a design principle from robotics (Martín-Martín & Brock, 2022) that focuses on bidirectional interactions between components. For our model, this means that we implement both saccade target selection and object segmentation as components that require the other's current state as input, as shown in Figure 1. Critically, we consider the uncertainty of the current segmentation to weigh different segmentation measurements. This segmentation uncertainty is also an input to our saccade target selection, as studies of eye movements in natural environments have shown that uncertainty about the state of the visual environment is important to understand and predict gaze behavior (Gottlieb, Oudeyer, Lopes, & Baranes, 2013; Hayhoe & Matthis, 2018).

We explain how each component models the respective process based on the visual input and the other component's current state. We start with the component for object segmentation, which we adapted from our previous work in robotic perception (Mengers et al., 2023) to account for object information at the current gaze position and top-down semantic information. Then we explain how we modified our previous model for the saccadic decision-making process (Roth et al., 2023) to take advantage of both the segmentation and its estimated uncertainty.

Estimating object segmentation and its uncertainty

In real-world scenes, object segmentations based on semantics, motion, and appearance will typically not wholly agree (Hackett & Shah, 1990; Pantofaru, Schmid, & Hebert, 2008). This leads to ambiguity and uncertainty when combining different object cues (Chen & Pavlidis, 1980; Hackett & Shah, 1990; Pantofaru et al., 2008; Mengers et al., 2023). For example, an object of similar color to the environment that does not move might be counted toward the background (in Figure 2, the shirt of the person on the right disappears in appearance segmentation), whereas an object made up of multiple similarly colored parts that can move relative to another might be segmented into these parts according to their motion (in Figure 2, the lower half of the person on the left is not moving together with the upper and hence disappears in motion segmentation). Therefore, we aim not only to estimate the object segmentation, but also to explicitly estimate the current uncertainty over it. To do so, we combine multiple cues for object segmentation as measurements in a recursive Bayesian filter (Särkkä, 2013). This filter updates the object segmentation with each new measurement while also estimating its uncertainty, similar to the segmentation filter in previous work on object segmentation for robotics (Mengers et al., 2023). As shown in Figure 2 on the left, we consider three measurements of pre-attentive global segmentation based on motion, appearance, and semantics, as well as segmentation of only the locally attended object.

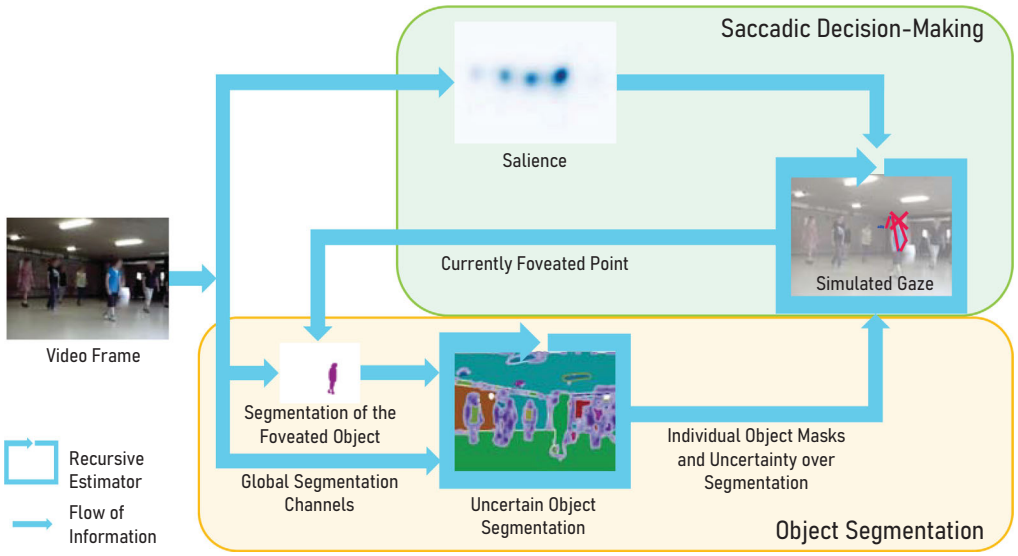

Figure 1. Saccadic decisions and object perception influence each other, as reflected by their interconnection in our model. We illustrate the information flow in our model during the processing of a single frame from a dynamic video. Object segmentation is informed by multiple global object cues and a high confidence prompted segmentation of the foveated object. The segmented objects act as perceptual units for the saccade target selection. The uncertainty over object segmentation plays a key role in driving exploration while being resolved through high-confidence measurements at the current gaze position. Because both the dynamic scene and gaze change over time, the recursive estimator continuously updates the segmentation and its uncertainty.

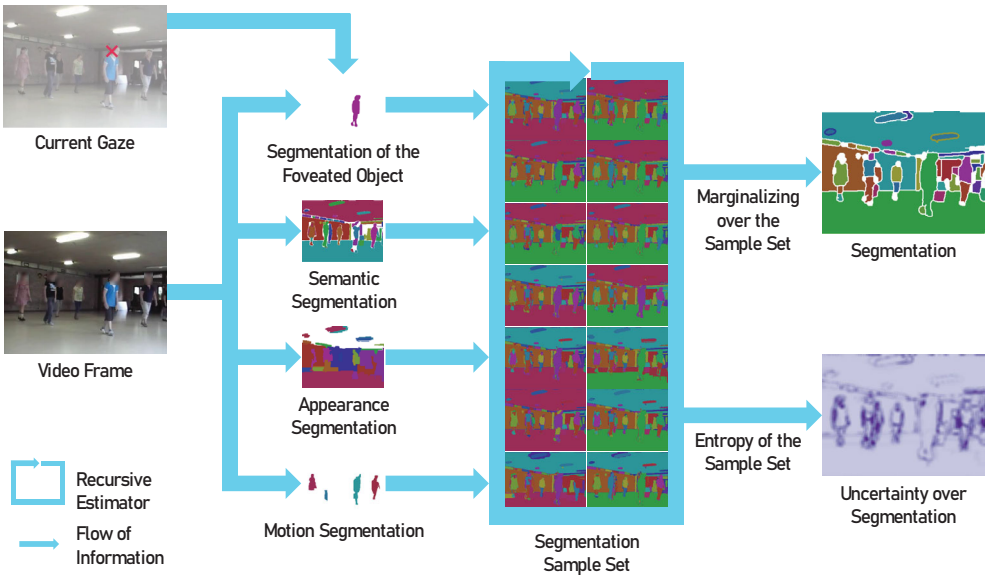

Figure 2. Our model combines multiple object cues to estimate both object segmentation and its uncertainty recursively. We integrate foveated and global segmentations of the scene (left) in a Bayesian filter (middle), which maintains a belief over the current state, represented by a weighted set of multiple possible segmentation samples (14 example samples from the full set of 50 are shown). We then compute the currently most likely segmentation and its uncertainty (right), which we use to inform saccadic decisions.

This attentive segmentation is particularly important because it has greater confidence (Stewart et al., 2022), thereby reducing segmentation uncertainty dependent on the current gaze. This is one direction of the strong interaction between object segmentation and saccadic decision-making in our model. We now describe how we obtain the different measurements of object segmentation, before explaining how we combine them in a Bayesian way using a particle filter to estimate both segmentation and its uncertainty.

Cues for the current object segmentation

We aim to design a directly image-computable model and thus rely only on the RGB video as input for pre-attentive global segmentation. We extract three object segmentation cues from it: low-level appearance, higher-level semantic features, and common motion. For the appearance segmentation, we use the simple graph-based method by Felzenszwalb and Huttenlocher (2004) because it already provides reliable regions of common appearance. For the semantic segmentation, we face a more complex problem, for which we leverage recent advances in large, data-driven segmentation models (Kirillov et al., 2023). Concretely, we obtain a semantic segmentation using the method by Ke et al. (2023). To find common motion in the scene, we first quantify motion as optical flow between subsequent frames using a state-of-the-art, data-driven method (Shi et al., 2023). We then find parts that move together by applying the same graph-based method (Felzenszwalb & Huttenlocher, 2004) as for appearance, because it proves to be sufficiently reliable.

Moreover, we use the current gaze location to inform object segmentation because gazing at an object should afford higher-confidence measurements of its boundaries (Henderson, 2003). To model such precise measurements around the currently attended object, we use a large data-driven segmentation model (Zhao et al., 2023) that can develop a prompted segmentation around a provided point (for more details, see Appendix B). If we provide it with the current gaze location, we obtain its highest confidence object that contains this point. To further increase the quality of this prompted high-confidence segmentation, we perform it on the highest available resolution of the input image, which we downsample for other cues (see Table D1). We use the prompted segmentation as an additional input to our filter for object segmentation. Because the current gaze location is a result of the previous saccadic decision process, this represents the connection of the two components in one direction. We explain the other, richer direction in Scanpath simulation, but now continue to explain how we combine all the described inputs to obtain one object segmentation with uncertainty.

Combining different object segmentation measurements in a particle filter

Our aim is to represent object segmentation and its uncertainty, which means a *belief* over object segmentation, and update this belief with new measurements over time. Representing such a belief is hard, because the space of possible segmentations is complex, high-dimensional, and can have multiple modes. Consequently, we cannot simply represent this belief with a Gaussian over object segmentation. We have shown previously that, instead, we can use a Monte Carlo approach for such representations, where each set of particles corresponds with the likely segmentation of the scene (Mengers et al., 2023, Sec. III-A). These particles together represent a belief over the segmentation, which we can recursively update with a *particle filter* (Thrun, Burgard, & Fox, 2005). To give an intuition for this particle filtering approach, let us consider the general problem of estimating a belief over a state s_t that dynamically changes over time and for which we obtain measurements z_t. When using a particle filter, we represent the belief over the state s_t by a set of different particles, each a hypothesis $s^{[i]}$ for the current state. If the state was not dynamic, we could now use the measurements over time to determine the true state by weighting each hypothesis with a weight $w_t^{[i]}$ (Equation 1, where η is a normalizing factor and i is the index of the particle). Unlikely states are removed using weighted resampling, that is, redetermining the particle set by randomly drawing with replacement particles from the current set according to their weights. To account for dynamism, we can add a prediction step (Equation 2), where we adapt each hypothesis $s_t^{[i]}$ according to available information a_t on the current development of the state s_t. For a more detailed introduction and derivation of the particle filter, please see (Thrun et al., 2005).

$$\forall_i : w_t^{[i]} = \frac{1}{\eta} \cdot p(z_t | s_t^{[i]}) \quad (1)$$

$$\forall_i : s_t^{[i]} \sim p(s_t | s_{t-\Delta t}^{[i]}, a_t) \quad (2)$$

When using such a particle filter to update a belief over the segmentation of a scene, each particle $s_t^{[i]}$ is one possible segmentation of the scene into objects (see Figure 2). Together, these particles represent different hypotheses for the object segmentation of the scene and—in their (dis-)similarities for different parts of the scene—varying levels of uncertainty. We recursively filter this set to account for the dynamism of the scene and integrate new measurements of the real segmentation by implementing Equations 1 and 2: To perform predictions (Equation 2) of these particles, we use the current optical flow as a_t to shift the boundaries between objects in each particle's segmentation according to the estimated motion between frames. Then, we weigh the resulting segmentation particles according to their distance to each of our measurements (Equation 1), resampling the set according to the product of the resulting weights. To determine this distance between a particle's segmentation and a measured segmentation, we compute the average distance of their object boundaries, as described in more detail in Appendix C1. In addition to weighting and resampling the particles based on current segmentation measurements, we also adjust the belief by directly

incorporating measured segments into some of the particle segmentations. This is not strictly necessary because these measurements are, in principle, already incorporated in the particles. Still, modifying some of the particles to more closely resemble the current measurements is computationally favorable because it allows for a higher quality of the belief approximation around the most likely areas and makes the approach more robust for a smaller number of particles, as explained in (Mengers et al., 2023, Sec. III-C). The resulting resampled set then represents the currently most likely segmentation hypotheses according to the measurement history.

Obtaining object segmentation and its uncertainty

Although the set of segmentation samples is useful to maintain a belief over the segmentation into objects, it is challenging to utilize in saccadic decision-making. Therefore, we marginalize across the sample set at each time step to obtain one object segmentation and uncertainty estimate, as illustrated on the right in Figure 2. We first determine the likelihood $p_b(x, y)$ that each image pixel (x, y) is part of a boundary between two segments by comparing the weights of all particles with a boundary at a given pixel (the particle set $\mathcal{B}(x, y)$) against the weights of those without (the particle set $\bar{\mathcal{B}}(x, y)$) as shown in Equation 3. Based on these boundary likelihoods, we can then obtain the currently most likely segmentation by thresholding and closing contours. Compared with the full set of segmentation samples, this is, of course, some loss of information, but we preserve the information on the agreement between particles by explicitly deriving the current uncertainty. To do so, we evaluate the entropy $H(x, y)$ of the previously thresholded boundary likelihood (Equation 4), resulting in high values where some samples have boundaries, although others do not.

$$p_b(x, y) = \frac{\sum_{i \in \mathcal{B}(x,y)} w_t^{[i]}}{\sum_{i \in \mathcal{B}(x,y) \cup \bar{\mathcal{B}}(x,y)} w_t^{[i]}} \quad (3)$$

$$\begin{aligned} H(x, y) = & -p_b(x, y) \cdot \log(p_b(x, y)) \\ & -(1 - p_b(x, y)) \cdot \log(1 - p_b(x, y)) \end{aligned} \quad (4)$$

We use the obtained object segmentation and uncertainty to select saccade targets in a drift-diffusion model (DDM) over the objects. To do so, we need to ensure that the same object keeps the same ID within the segmentations over time. Therefore, we use a variation of the Hungarian algorithm (Hopcroft & Karp, 1973) to match object IDs between object segmentations. Specifically, we determine the similarity of the segments in the current object segmentation to those in the past 10 time steps by determining their intersection over union, discounted for older segmentations to favor keeping the currently used object IDs. This results in a weighted bipartite graph from old segment IDs to new segments, in which we find the matching where each new segment is matched with an ID such that the sum of all weights is maximized (see Jonker & Volgenant, 1987). If no existing ID can be matched, a new ID is assigned. For further details on this matching procedure, see Appendix C2. The segmentation map then forms the basis for the object-based attention mechanism in the scanpath simulation, which we describe in detail in the next section.

Scanpath simulation

We model the saccadic decision-making process by adapting the object-based *Scan*path simulation in *Dy*namic scenes (*ScanDy*) framework (Roth et al., 2023). The scanpath simulation updates its internal state, which includes a decision variable for all potential target objects in the scene, as segmented through the particle filter (Figure 2). We model the target selection process of where and when to move the gaze position with a DDM, in which each object represents a potential saccade target (cf. Figure 2). The decision variable for each object depends on its eccentricity given the current gaze position, how relevant the visual scene features are, as measured by salience, and the uncertainty of the local object boundaries, as provided by the segmentation particle filter. Notably, the model does not rely on an explicit implementation of the IOR mechanism.

Scene relevance based on salient features

We quantify the relevance of the scene content for gaze behavior by computing frame-wise feature maps F. Because we model free-viewing gaze behavior, where the observers have no explicit task, we approximate the relevance of different parts of the scene through visual saliency. We used the video saliency model UNISAL (Droste et al., 2020), which was jointly trained on both image and video visual saliency datasets, since it is lightweight and produces state-of-the-art results on the DHF1K Benchmark (Wang et al., 2018). We inferred the video saliency maps using the model with the domain adaptation for the DHF1K video dataset, which is most similar to the videos used in this study. The resulting video saliency predictions used as frame-wise feature maps $F(x, y)$ are normalized to [0,1]. F is typically strongly localized around the most salient object (cf. Droste et al., 2020). To allow the model to rely less on this strongly focused map, we introduce a model parameter $f_{\min}$ that linearly scales F to $F' \in [f_{\min}, 1]$. By reducing the effective value range, a higher $f_{\min}$ parameter decreases the influence of the salience on the saccadic decision-making process.

Gaze-dependent visual sensitivity

The foveation of the human visual system leads to a decrease in visual sensitivity with eccentricity from the current gaze position. As in Roth et al. (2023), we model the visual sensitivity S across the scene with an isotropic Gaussian G_S. We account for the well-documented object-based attentional benefit (Egly et al., 1994; Scholl, 2001; Malcolm & Shomstein, 2015) by approximating the covert spread of attention across the currently foveated object O_f (1 if pixel is part of the object, 0 if not) with a uniform sensitivity, replacing the part of G_S that falls within O_f.

$$G_S(x, y) = \frac{1}{2\pi\sigma_S^2} \exp\left(-\frac{(x - x_0)^2 + (y - y_0)^2}{2\sigma_S^2}\right) \quad (5)$$

$$S(x, y) = \begin{cases} 1, & \text{if } O_f(x, y) = 1, \\ G_S, & \text{else,} \end{cases} \quad (6)$$

where the standard deviation $\sigma_S = 7$ dva is set according to similar models (cf. Schwetlick et al., 2020; Roth et al., 2023) and based on preliminary model explorations.

In addition, we implemented two possible model extensions, which are not part of the base model but can be incorporated into the visual sensitivity S, namely *saccadic momentum* and *presaccadic attention*. In our explicit implementation of saccadic momentum, we increase the visual sensitivity in the direction of the previous saccade by generating an angle preference map based on the current gaze position and the angle of the previous saccade. We set a maximal value, which will be the sensitivity value in the direction of the previous saccade angle, that decreases linearly with the angle within a specified angle range to a minimum value. The resulting map (see Figure F1a) is then multiplied with S. In our implementation of presaccadic attention, we assume a uniform spread of visual sensitivity across not only the currently foveated object but also objects that are likely to be the next saccade target (see Figure F1b).

Uncertainty over object segmentation

The visual system integrates different sources of information into a coherent visual representation of the environment (Milner & Goodale, 2006). If an object moves or input sources differ, for example, when the appearance and the motion-based segmentation find different object boundaries, this leads to a disagreement between instances in the segmentation particle filter. We include the resulting uncertainty over the object segmentation as the third contributing factor for the saccadic decision-making process, in addition to the relevance of the scene features and the gaze-dependent visual sensitivity. The uncertainty measure is directly obtained from the entropy $H(x, y)$ of the previously obtained boundary likelihood threshold in the object segmentation particle filter (see Equation 4). We smooth the resulting map with a Gaussian blur, so uncertainties at the object boundaries are attributed to both objects. The values in the uncertainty map are, by construction, in the range $U(x, y) \in [0, 1]$. Analogous to the scaling of the scene feature map, we introduce the model parameter u_{min} that linearly scales U to $U' \in [u_{min}, 1]$. Higher values for u_{min} hence effectively downscale the influence of the object uncertainty on U'. The uncertainty at the current gaze position is typically low since the prompted segmentation of the currently foveated object provides a refined object mask, which is incorporated in the particle filter with high confidence. Through this interaction, the uncertainty contribution encourages the exploration of other objects in the scene.

Saccadic decision-making process

We describe gaze behavior as a sequential decision-making process where objects in the scene accumulate evidence for becoming the next saccade target over time. As in the *ScanDy* framework (Roth et al., 2023), we model this latent cognitive process using a modified DDM (Ratcliff, Smith, Brown, & McKoon, 2016) with multiple options. The DDM accumulates evidence for each object over time (drift), while random fluctuations perturb each decision variable (diffusion). Unlike a classic DDM model, which includes only one decision variable and two thresholds for alternative choices, our model assigns each potential target object i a decision variable V_i that accumulates toward a shared decision threshold θ (see *Drift Diffusion Process* in Figure 3). As soon as the accumulated evidence for one object exceeds θ, a saccade to this target is initiated. Hence, the DDM by design does not only model where but also when the eyes move. The DDM drift rate μ_i for an object at a given time depends on the task relevance based on scene features $F'(x, y)$, the visual sensitivity depending on the current gaze position $S(x, y)$, and the uncertainty of the object segmentation $U'(x, y)$. We multiply these maps in every frame to an evidence map $E(x, y, t) = S \cdot F' \cdot U'$, as shown in Figure 3. Next, we calculate μ_i for each object mask O_i (1 if pixel is part of the object, 0 if not) in the resulting object segmentation of the particle filter (see Figure 2) as the average evidence across the mask $\bar{E}(O_i, t)$, scaled by the area A_i of the object, with

$$\bar{E}(O_i, t) = \frac{\sum_{x,y} E(x, y, t)}{\sum_{x,y} O_i(x, y, t)}, \quad (7)$$

$$A_i(t) = \sum_{x,y} O_i(x, y, t) \cdot (1 \text{ dva}/1 \text{ px})^2, \quad (8)$$

$$\mu_i(t) = \bar{E}(O_i, t) \cdot \max\left(1, \log_2 A_i(t)\right). \quad (9)$$

We convert the area from px^2 to dva^2 to ensure that videos with different resolutions are treated appropriately and scale the object's perceptual size

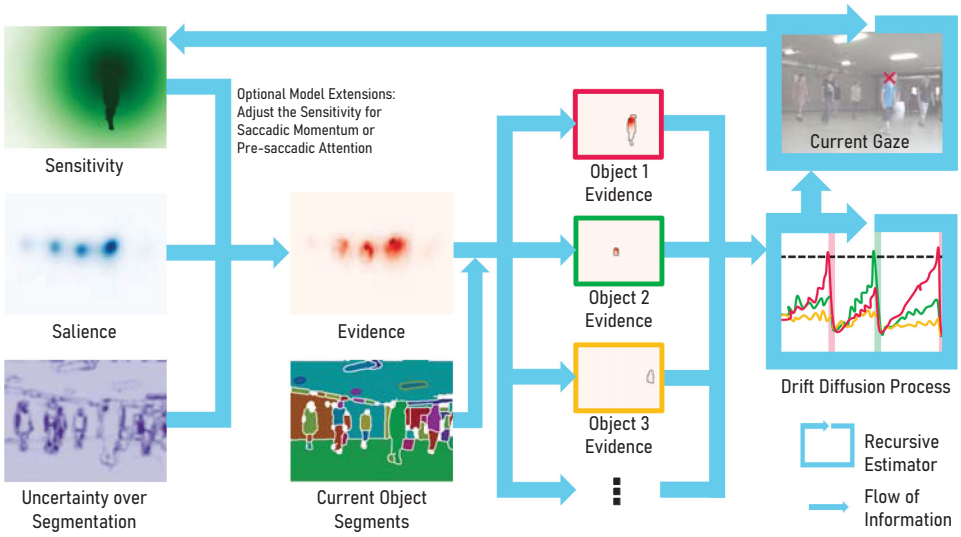

Figure 3. Our model makes saccadic decisions based on objects and is driven by uncertainty. It combines the uncertainty over object segmentation with salience and gaze-dependent sensitivity (left) into evidence for individual objects (middle). This evidence is then accumulated for each object in a drift-diffusion process (right). As soon as its threshold is passed, a saccade to this object is executed, otherwise the gaze smoothly pursues the currently foveated object.

logarithmically (cf., Nuthmann, Einhäuser, & Schütz, 2017) to account for the difference in object sizes.

The decision variable V_i for each object is then updated based on μ_i and random fluctuations in the diffusion term $\epsilon \sim \mathcal{N}(0,\ 1)$, with

$$V_i(t + \Delta t) = V_i(t) + \nu \cdot (\mu_i(t)\Delta t + s\epsilon\sqrt{\Delta t}), \qquad (10)$$

where the noise level s is a free parameter, and ν is the fraction of time within Δt spent on foveation since no evidence is accumulated during saccades. We set the update time resolution $\Delta t = 1$, measured in frames. We assume a linear update in V_i and can hence calculate the exact time when the decision threshold θ is crossed. As soon as θ is reached, we reset all decision variables $V_i = 0\,\forall i$, and a saccade is executed to the corresponding object. The saccade duration τ_s scales linearly with the saccade amplitude a_s (Collewijn, Erkelens, & Steinman, 1988; Roth et al., 2023) with

$$\tau_s = 2.7\,\text{ms/dva} \cdot a_s + 23\,\text{ms}. \qquad (11)$$

Gaze update

We update the simulated gaze position at each time step (i.e., video frame). If the DDM threshold θ is not reached, the gaze point moves with the optic flow at its current position. This results, depending on the object and camera motion in the video, in either fixation or smooth pursuit behavior where the gaze moves with the object. If an i exists with $V_i > \theta$, a saccade is triggered to O_i. The exact landing position within O_i is determined probabilistically, with the probability $p_i(x, y)$ of each pixel being proportional to the scene features F and gaze-dependent visual sensitivity S:

$$p_i(x, y) \sim O_i(x, y, t_0) \cdot F'(x, y, t_0) \cdot S(x, y, t_0). \qquad (12)$$

Dataset

We compared the simulated scanpaths with human eye-tracking data recorded under free-viewing conditions on videos of natural scenes. We collected eye-tracking data from 10 participants (8 female; mean age, 34.4 years; range, 23–69 years) on 43 video clips from the unidentified video objects (Wang, Feiszli, Wang, & Tran, 2021) dataset (10 used for parameter tuning, 33 used for testing the model; randomly split). The videos were selected to show diverse content and to have temporally consistent, densely annotated object masks for the first 90 frames (cf. Wang et al., 2021).

We recorded eye-tracking data for the here-used videos with an Eyelink 1000+ tabletop system (SR Research, Osgoode, ON, Canada) with a 1,000 Hz sampling rate, as part of an ongoing collaborative large-scale eye movement database (publication of full dataset in preparation). We presented the videos in a dark room on a wall-mounted 16:9 video-projection screen (size: 150 × 84 cm, Stewart Luxus Series "GrayHawk G4," Stewart Filmscreen, Torrance, CA) at a distance of 180 cm from the study participants. We used a PROPixx projector (Vpixx Technologies, Saint-Bruno, QC, Canada) operating with 1,920 × 1,080 pixels resolution on its native vertical refresh rate of 120 Hz. All videos were shown with a 30

fps framerate and (depending on their aspect ratio) scaled to a size of maximally 38.2 dva horizontally or 21.5 dva vertically (1,536 × 864 pixels) to avoid high eccentricities. Participants started each trial with a fixational control (red dot on a black background) at a random location within the area where the scene was shown. The video was presented as soon as the participant fixated the target location (tolerance radius of 2 dva), ensuring high data quality and variation in the initial gaze position. All participants provided informed consent according to the World Medical Association (2013) before data collection.

Event detection algorithm

Identifying saccades in gaze data in dynamic scenes with object and camera movement in the scene can be a challenging task due to the presence of smooth pursuit eye movements. Potentially large pursuit velocities lead to a high number of false positive saccade detections in classic velocity-based algorithms such as the Engbert-Mergenthaler (EM) algorithm (Engbert & Mergenthaler, 2006). We, therefore, used the state-of-the-art U'n'Eye neural network architecture (Bellet, Bellet, Nienborg, Hafed, & Berens, 2019) and fine-tuned the network to our dataset. We labeled saccades, foveations (combining fixation and smooth pursuit events), and post-saccadic oscillation (PSO) events for one randomly selected second per video from different subjects. Detecting PSOs is important to reliably define the endpoint of a saccade and hence precisely determine the duration of a foveation event (Schweitzer & Rolfs, 2022). The U'n'Eye network, with the training data we provided, was not able to reliably detect PSOs. Hence, we used the PSO detection based on saccade direction inversion, as described by Schweitzer and Rolfs (2022). This algorithm expects saccades in the format provided by the EM algorithm. We, therefore, ran both the EM and U'n'Eye saccade detection algorithms, determined the saccades that were detected with both algorithms and then specified the exact saccade endpoint using the direction inversion criterion for PSO detection.

Metrics and parameter fitting

We determined the model parameters by comparing the foveation duration and saccade amplitude distributions of the simulated scanpath with the human ground truth. We measured the similarity between a simulated distribution N to the ground truth M using the two-sample Kolmogorov-Smirnov (KS) statistic $D = \sup_x |N(x) - M(x)|$. We systematically varied the DDM noise level s, the decision threshold θ, and the relative importance of the feature map F' and uncertainty map U', quantified by the rescaling parameters $f_{\min}$ and $u_{\min}$. We performed a coarse grid search in this four-dimensional parameter space on the 10 videos in the training set. We simulated five different scanpaths (different random seeds) for each parameter configuration and compared them. Because we were particularly interested in the effect of uncertainty on the simulated scanpaths, we refined the grid search for each $u_{\min}$ value around the parameter sets leading to the lowest mean of the KS statistics for the foveation duration D_{FD} and the saccade amplitude D_{SA}. We present an overview of all fixed and fitted model parameters, the parameter grid, and details of the fitting procedure in Appendix D.

With the model parameters chosen such that the basic scanpath summary statistics of foveation durations and saccade amplitudes matched the human data, we evaluated the simulated scanpaths out-of-domain on the test set, that is, on 33 previously unseen videos and on different metrics than what the parameters on the training set were selected for. For each parameter set, we simulated 10 scanpaths and compared them with the data from the 10 human observers. We focused on the analysis of how gaze behavior balances the exploration of the background of a scene (*Background*), uncover an object for the first time for foveal processing (*Detection*), explore further details of the currently foveated object by making a within-object saccade (*Inspection*), or return to a previously uncovered object (*Return*) (Linka & de Haas, 2021; Roth et al., 2023). Comparing the foveation durations in each category provides an insightful metric of the exploration behavior, which is particularly suited for dynamic scenes (see Roth et al., 2023). In addition to evaluating models on the test set, we also chose the later described *base model* among different uncertainty values $u_{\min}$ based on this metric on the training set (see Model ablation 1: Uncertainty drives exploration and Appendix D for more details).

Because our model does not have an explicit IOR mechanism, we were particularly interested in whether it could reproduce typical IOR effects. IOR describes the inhibition of recently attended stimuli and the resulting delayed response to them (Posner & Cohen, 1984; Klein, 2000). In a free-viewing condition, as in the data used for this study, we therefore expect that saccades that return to a previous gaze position require more time to prepare. Hence, we analyzed the distribution of relative saccade angles of the human and simulated scanpaths. We divided all foveation events into 30 bins depending on the relative angle of the previous saccade (i.e., bin size of 12°). With the expectation that foveations preceding a return saccade (±180°) would be longer and foveations preceding forward saccades (±0°) would be shorter, we calculated the median foveation duration for each relative saccade angle bin. We used the median foveation duration instead of the mean to avoid a few very long events (in particular, smooth pursuit events can last multiple seconds), distorting the statistics for individual bins.

Results

Our aim was to build an image-computable and mechanistic computational model that closely resembles human gaze behavior in dynamic real-world scenes. In this section, we compare our model with human scanpaths, first qualitatively in Qualitative scanpath analysis and then quantitatively by reviewing aggregated statistics in Aggregated scanpath statistics. We systematically explore the influence of uncertainty on visual exploration behavior in Model ablation 1: Uncertainty drives exploration. In an additional ablation study, we probed the impact of individual object representations as input sources and the importance of the interaction between object perception and saccadic decision-making for the simulated scanpaths, as described in Model ablation 2: Semantic object cues and component interconnections form suitable perceptual units. Last, we show how our model can be easily extended with additional modules, such as saccadic momentum or presaccadic attention, leading to more human-like saccade angle statistics and slight improvements in early object detections (Model extensions: Saccadic momentum improves saccade angle statistic and presaccadic attention benefits early object detections).

Qualitative scanpath analysis

Our model produces scanpaths that qualitatively closely resemble human visual exploration behavior; one example scanpath is shown in Figure 4 (see videos in Appendix E for more examples and direct comparisons with human scanpaths). The access to individual mechanistic components of our model makes the individual saccadic decisions transparent and interpretable: Initially, all unexplored salient objects have relatively high uncertainty, which is resolved through large saccades towards them (Figure 4a; for a more detailed analysis of the uncertainty development, see Appendix A). Objects with particularly high saliency are likely to be revisited (Figure 4b) or are further inspected (Figures 4c and 4d). Return saccades to previously foveated objects also become more likely with time, as uncertainty over object boundaries can rise again, for example, through object motion. This qualitatively similar behavior of our model can also be seen in Appendix E, where we show the exact scanpath and all intermediate computational steps as videos for 10 simulations of our model as well as for 10 human participants. We now further quantify these qualitative similarities between the human and modeled gaze behavior by comparing summary statistics of human scanpaths with the model predictions across the whole dataset.

Aggregated scanpath statistics

We compare our base model predictions to human scanpaths on a set of videos not used for parameter search. As described before, we selected the model parameters to resemble the statistics of human foveation duration and saccade amplitude on 10 videos. The model generalizes well to the previously unseen set of 33 videos, as shown based on the aggregated scanpath statistics in Figure 5. Similar to human eye-tracking data, the foveation durations (Figure 5a) of the simulated scanpaths follow a log-normal distribution with a mean of 390 ms and a median of 332 ms (humans: mean of 433 ms and median of 316 ms). The distribution of the model is more narrow compared to humans, which—if other metrics would not be considered—could be corrected by increasing both the decision threshold and the noise level in the DDM, as described in Scanpath simulation. The saccade amplitudes in the simulated scanpaths (Figure 5b) follow the gamma distribution of the human data with a mean of 3.70 dva and a median of 2.81 dva (humans, mean of 3.40 dva and median

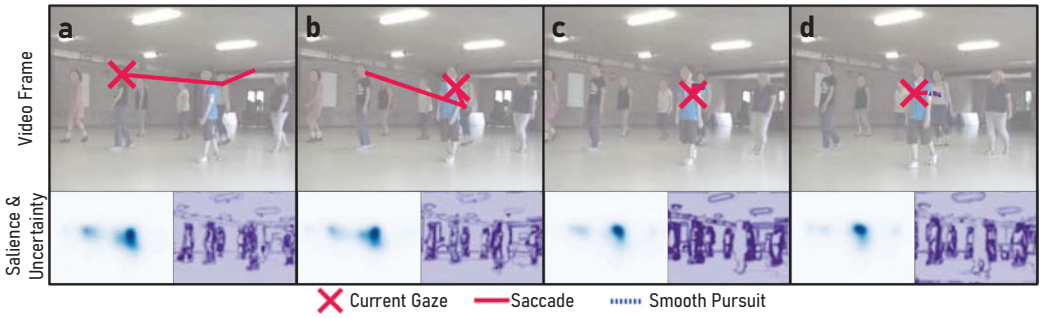

Figure 4. The predicted scanpaths of our model show human-like exploration in dynamic scenes. In this video of the test dataset, the model first follows uncertainty and detects two novel objects (dancers) (a), then returns to the first before detecting another one (b), which is then further inspected primarily due to its high visual saliency (c and d). For a video version, see Appendix E.

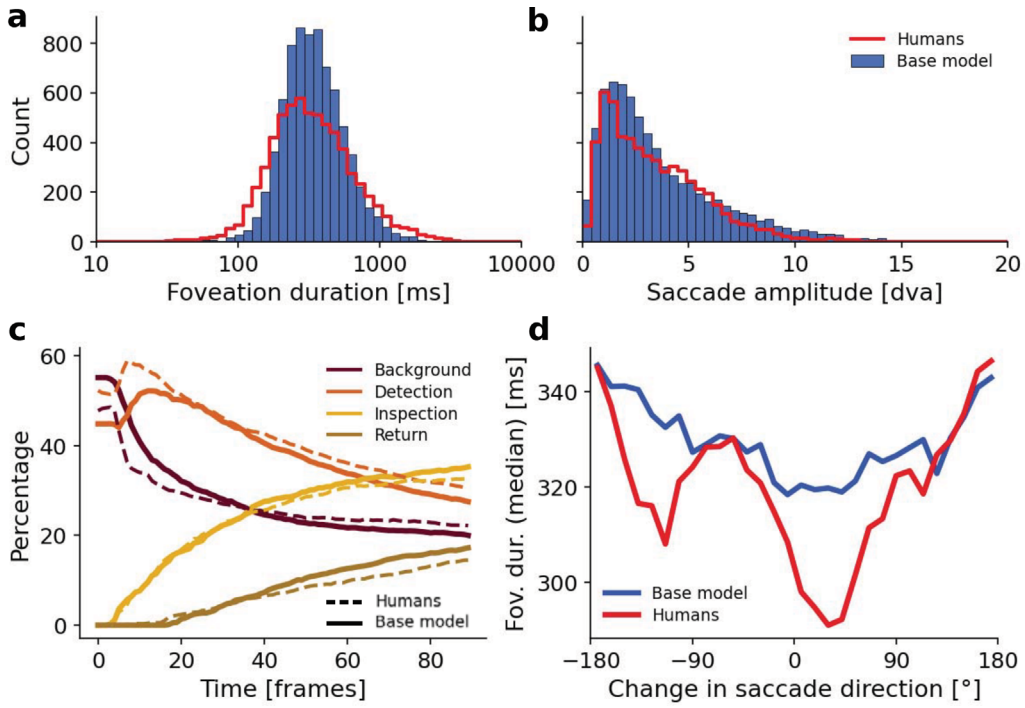

Figure 5. Aggregated statistics of the simulated scanpaths of the base model resemble the human eye-tracking data. (a) Histogram of the duration of all foveations in the human ground truth data (red) and the base model (blue). (b) Histogram of the saccade amplitude distributions. (c) Percentage of foveation events in the categories "Background" (maroon), "Detection" (orange), "Inspection" (yellow), and "Return" (khaki) across all human (solid) and model (dashed) scanpaths as a function of time. (d) Median duration of the preceding foveation durations for each saccade. We applied a centered circular moving average across five bins (12° bin size) to reduce fluctuations in the median.

of 2.90 dva). These well-described statistics are not explicitly implemented in the model, but emerge from model constraints: Foveation durations are a consequence of the way evidence is accumulated in the decision-making process. Saccade amplitudes result from the balance of local exploration, as encouraged by the visual sensitivity, and global exploration, as driven by uncertainty and noise in the DDM.

Beside replicating these basic summary statistics, we are interested in how the exploration behavior of our model compares to that of humans. Our model, like the participants in our dataset, starts at a random initial location on the scene. Hence, about half of the scanpaths start on the background. In Figure 5c, we can observe that the model—similar to humans—quickly starts to favor the exploration of novel objects (detections) rather than further exploring the background. We additionally confirmed that, in a large proportion of the scenes, the object that was first detected by the majority of human observers was also first detected in the majority of simulated scanpaths (24 of 33 scenes in the test set, 72.7% agreement; base rate, 23.6%, estimated as the average of $\frac{1}{N_{0,s}}$ across scenes, where $N_{0,s}$ is the number of objects in the first frame of each scene). After an initial peak in detections in both models and humans, the amount of detection decreases in both cases in favor of further saccades within the currently foveated object (inspections) or revisits of previously foveated objects (returns). Overall, both in relative amounts and trends over time, the balance between the exploration of the background, new objects, and already seen objects of the model resembles the human behavior well. Typically, such a balance in exploration can be achieved through a suitable parametrization of an explicitly implemented IOR mechanism (cf., Itti & Koch, 2001; Roth et al., 2023). We find that in our model, the relative influence of the uncertainty map plays a crucial role in achieving this balance, which we describe in detail in Model ablation 1: Uncertainty drives exploration.

The model even shows the expected *temporal* IOR effect (Klein & MacInnes, 1999; Luke et al., 2014), as shown in Figure 5d, without explicitly implementing it and without adjusting any parameters to reproduce this statistic. We find a characteristic dip in foveation durations before a saccade is executed in the same direction as the previous saccade (forward saccades), as observed in the human scanpaths. The preparation of saccades with larger turning angles is slower. This is a result of the uncertainty at the previous gaze position

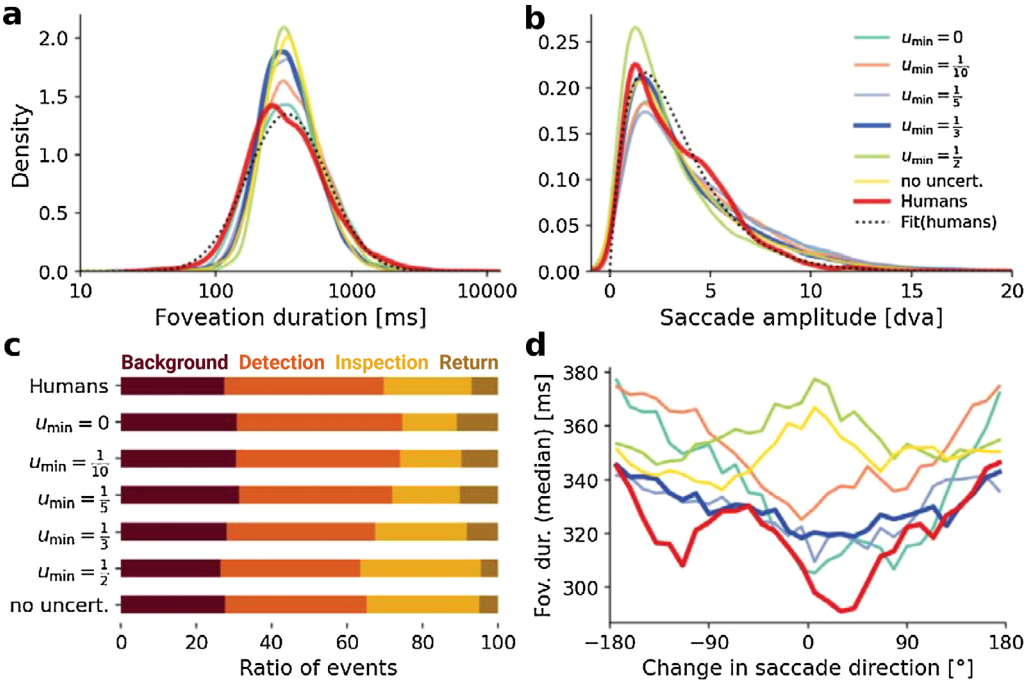

Figure 6. The uncertainty contribution in the model determines the exploration behavior. (a) Kernel density estimation (KDE) of the distribution of foveation durations for the human data and simulated scanpaths with different uncertainty contributions. The dotted line indicates a log-normal fit to the human data with $\mu = 5.815$ and $\sigma = 0.681$ (equiv. to an expected value of $e^{\mu+\frac{\sigma^2}{2}} = 422.8$ ms). (b) KDE for the saccade amplitude distributions with a fitted Gamma distribution to the human data with shape $\alpha = 2.01$ and rate $\beta = 0.59$ (equiv. to an expected value of $\frac{\alpha}{\beta} = 3.40$ dva). (c) Ratio of time spent in the different foveation categories, as shown in Figure 5c, averaged across time. (d) Temporal IOR effect for the different uncertainty contributions, as in Figure 5d. The model with $u_{min} = \frac{1}{3}$ corresponds to the base model in Figure 5. Further information about the individual model parameters can be found in Appendix D.

being reduced through the foveated object cue, such that the accumulation of evidence will take longer for a return saccade (for a more detailed analysis, see Model ablation 1: Uncertainty drives exploration).

In summary, our model also quantitatively resembles human scanpaths in dynamic scenes, both in its basic statistics and its exploration behavior. In the next two sections, we further show that the similarities between human and modeled scanpaths, particularly the exploration behavior balances between the different functions of foveations, can be attributed to two features of our model: The consideration of uncertainty and bidirectional interaction between object perception, and saccadic decision-making to generate appropriate perceptual units to operate on.

Model ablation 1: Uncertainty drives exploration

Our model uses the uncertainty measure of the object segmentation module as an estimate for the perceptual uncertainty that influences the gaze behavior depending on the scanpath history. Here, we evaluate the effect of this uncertainty mechanism, comparing the simulated model scanpaths with a varying influence of the uncertainty on the saccadic decision-making process. Specifically, we vary the u_{min} parameter of the model where a *higher* value *decreases* the importance of uncertainty. We also compare results from this model with those of a version that does not consider uncertainty at all.

We select the threshold θ and the noise level s of the DDM for each value of u_{min} anew to fit the foveation duration and saccadic amplitude statistics (see Metrics and parameter fitting). Hence, varying the importance of uncertainty does not strongly influence the basic scanpath statistics, as shown in Figures 6a and 6b. Althogh the specific densities change, the general shape of the log-normal (foveation duration) and gamma (saccade amplitude) distributions remain stable. However, how the model balances exploration behavior changes considerably (Figure 6c): For high importance of uncertainty ($u_{min} < \frac{1}{3}$), the model focuses on the exploration of previously unvisited parts of the scene (background, detection) or returns to previously detected objects while only rarely further inspecting the currently attended object. For low importance ($u_{min} > \frac{1}{3}$ or "no uncert."), in contrast, we observe the

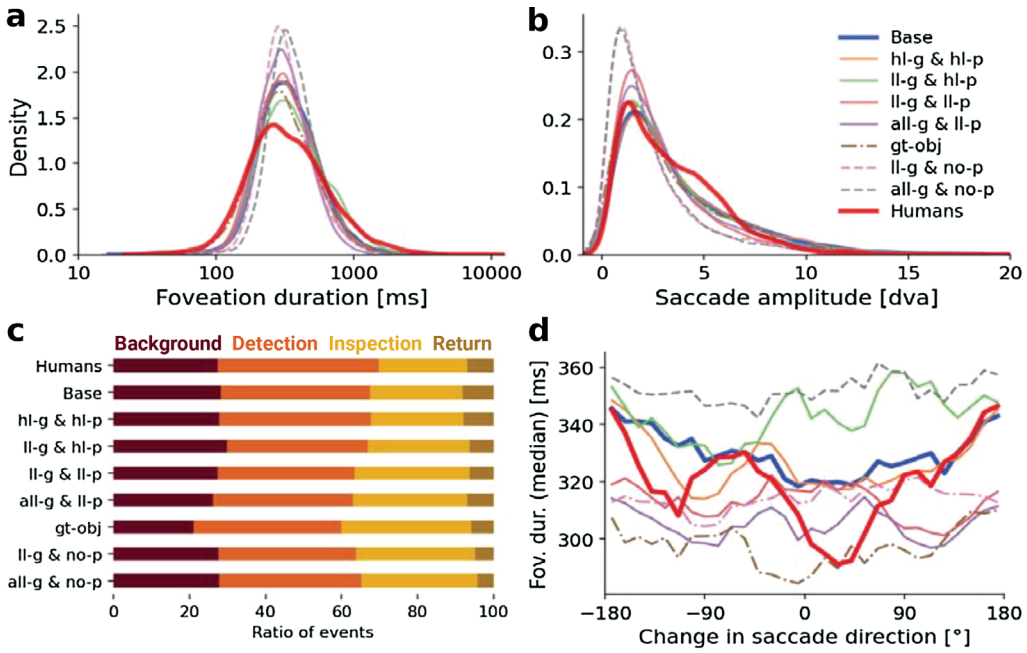

Figure 7. Semantic object cues and the interconnection through the gaze-dependent prompt are crucial for human-like simulated scanpaths. (a–d) Analogous to Figure 6 for models that use different object cues in the segmentation module. We compared the human data and the base model with models that use only the high-level/semantic object cues for the global and the prompted segmentation (*hl-g & hl-p*), only the low-level/appearance & motion-based global segmentation and the high-level prompt (*ll-g & hl-p*), a low-level/appearance-based prompt either combined with only low-level or with all global cues (*ll-g & ll-p*, *all-g & ll-p*), a model that uses ground truth objects together with the base model uncertainty (*gt-obj*), and models that use either only low-level or all global object cues without any prompted object (*ll-g & no-p*, *all-g & no-p*).

opposite trend, where the currently attended object is inspected further because the uncertainty over other parts of the scene does not drive the exploration there. For this reason, we find the right trade-off between the exploration of novel parts and the further inspection of attended parts at a medium importance of $u_{\min} = \frac{1}{3}$.

If we further analyze the temporal IOR effect under the different variations of uncertainty importance (Figure 6d), we observe that the model requires a certain level of uncertainty importance (at least $u_{\min} = \frac{1}{3}$) to reach the same effect. With too low importance of uncertainty ($u_{\min} > \frac{1}{3}$ or "no uncert."), the effect is instead inverted, such that saccades in the same direction are preceded by longer foveation durations. Hence, incorporating uncertainty into the model increases the probability of return events while simultaneously giving rise to the temporal IOR effect. The capability of accounting for this effect highlights how uncertainty can replace an IOR mechanism that is explicitly built into the model to drive human-like exploration behavior. The uncertainty of our model is computed based on the different object cues the model receives to build the object segmentation of the scene. In the following, we analyze the influence of these different inputs and how the resulting object representations affect the simulated scanpaths.

Model ablation 2: Semantic object cues and component interconnections form suitable perceptual units

Our model updates both the current object segmentation and perceptual uncertainty from the current image of the scene using different object cues. The segmentation then defines the perceptual units in saccadic decision-making, and the uncertainty influences the likelihood of selecting these perceptual units. In this ablation, we investigated the extent to which different object cues in the segmentation algorithm affect the predicted scanpaths. We compared different combinations of low-level (basic appearance and motion) and high-level cues (with semantic/top-down influence) for both the global and gaze-dependent object segmentation. The primary scanpath statistics did not change by much if we replaced the object sources (Figures 7a and 7b). For simplicity (high computational cost of the parameter exploration) and since we are primarily interested in the overall trends, we used the same model parameters as in the

base model. Only for the models without a prompted object segmentation (i.e., in which uncertainty is not lowered at the current gaze position), we found a new parameter set for the comparison (see Appendix D for more details). We compared the resulting scanpaths with our base model (low- and high-level global segmentation combined with a high-level prompted mask, i.e., *all-g & hl-p*), a model that uses a ground truth object segmentation (provided as labels in the dataset, cf. Dataset), and the human scanpaths.

We found that a model that does not use any low-level object cues but instead relies only on the high-level, semantic global segmentation and the semantic prompt (*hl-g & hl-p*) explored the scene very similarly to the base model and the human observers in terms of both the functional categories (Figure 7c) and the temporal IOR effect (Figure 7d). If, instead, only appearance- and motion-based segmentations were used as global object cues (*ll-g & hl-p*), the exploration behavior of the model remained close to the human data as long as foveated segmentations take advantage of high-level cues. The lack of a global semantic segmentation, however, led to more exploration of the background due to the uncertain low-level segmentation and, thus, made the characteristic dip in the temporal IOR effect disappear. We also implemented a model that used exclusively low-level object cues by replacing the high-confidence prompted object segmentation with an appearance-based low-level object prompt (*ll-g & ll-p*). In this case, the model segmented individual pieces of clothing based on color when foveating a person. If, instead, semantics were used, the person, including their clothes, would be considered a single object. This, in turn, would lead to a higher number of inspections as there is more uncertainty within the remaining ground truth objects. Adding the global semantic segmentation to the model with low-level prompts had almost no effect on the scanpath statistics (*all-g & ll-p*).

We next investigated the influence of the interaction between saccadic decisions and segmentation. We removed one of the two directions of these interactions. First, we replaced the perceptual units generated by our segmentation component with the ground truth objects provided by the dataset, while still computing and using the uncertainty map as in the base model. As a result, the few labeled objects were often and reliably foveated, leading to a high amount of inspections, while the background was explored much less. In an additional ablation, we removed the foveated segmentations from our model (*all-g & no-p* and *ll-g & no-p*), using the particle filter for the global segmentations but making the segmentation into perceptual units independent of the gaze. Hence, we removed the ability of the model to actively resolve uncertainty through saccades. This changed its exploration behavior considerably: Inspections became much more frequent, while detection times decreased. Moreover, we no longer observed any temporal IOR effect.

In summary, we found that removing low-level object cues from the segmentation filter does not lead to big changes in the resulting scanpaths. High-level semantic segmentation cues, however, were needed to simulate human-like gaze behavior. In particular, high-level prompted object cues entailed a temporal IOR effect. When we removed the ability of the model to reduce the object uncertainty through saccadic decisions through the prompted object cue, we observed an even larger effect on the simulated scanpaths. Even if the model included a global semantic segmentation, the uncertainty-driven interaction between the two components was crucial.

Model extensions: Saccadic momentum improves saccade angle statistic and presaccadic attention benefits early object detections

We have shown so far that our model reproduces important hallmarks of scanpaths in dynamic real-world scenes. One instructive metric we have not yet investigated is the distribution of relative saccade angles. Importantly, this distribution shows how many forward and return saccades were made and is therefore also interesting in the context of *spatial* IOR, that is, the reduced probability of returning to a previously visited location. The human scanpaths in our data show a strong bias for saccades in the opposite direction relative to the previous saccades, as shown in Figure 8a. This is in line with work that showed that return saccades are much more frequent in complex scenes than expected from the IOR literature (Smith & Henderson, 2009; Burlingham, Sendhilnathan, Komogortsev, Murdison, & Proulx, 2024). Because our model does not explicitly inhibit return saccades, this behavior is replicated well. Yet, the base model did not reproduce the human bias to make saccades in the same direction as the previous saccades, called *saccadic momentum* (Anderson et al., 2008; Smith & Henderson, 2009). Different mechanisms have been discussed to explain saccadic momentum, including a continuation of the motor plan and a visual bias in V4 neurons (Motter, 2018). Although no such mechanism is implemented in the base model, its modular implementation makes it easy to account for the saccadic momentum effect.

We thus extended our base model by introducing a bias towards forward saccades into the gaze-dependent visual sensitivity S (see Scanpath simulation), while keeping all other model parameters the same. Unsurprisingly, the model with saccadic momentum reproduced the relative saccade angle distribution (Figure 8a). Importantly, the previously investigated

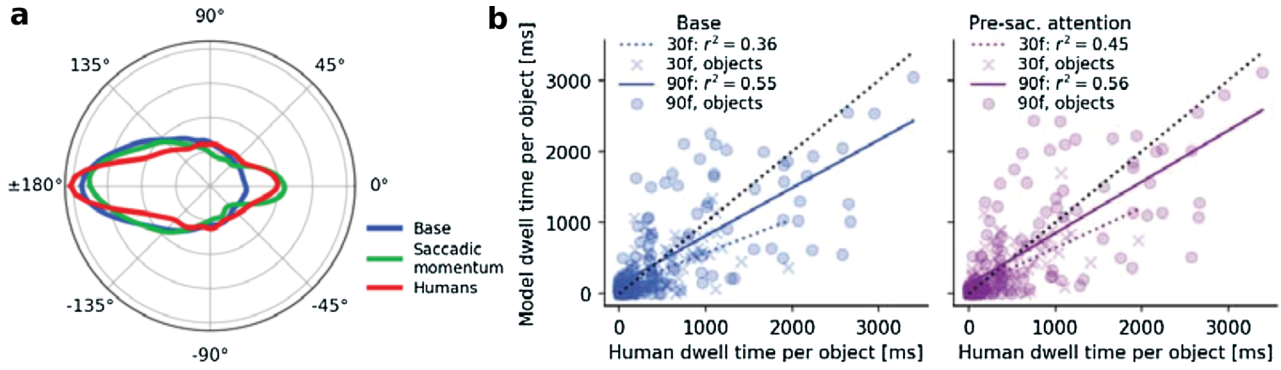

Figure 8. Extending the model through saccadic momentum or presaccadic attention leads to improvements in certain statistics. (a) Histograms of the change in saccade direction for scanpaths simulated with the base model (blue) and the model with saccadic momentum (green) compared to the human data (red). Forward saccades with ±0° go in the same direction, while return saccades with ±180° go in the opposite direction compared to the previous saccade. (b) Dwell time for each individual object averaged across human observers compared to simulated model scanpaths of the base model (left, blue) or the model with presaccadic attention (right, purple). We distinguish between the time objects were foveated in the first 30 frames (marked with *x*, dotted regression line) and in the first 90 frames (maximum number of frames with objects; marked with *o*, solid regression line). A perfect prediction would correspond to the data points for all objects lying on the dotted line with slope $m = 1$ and intercept $y_0 = 0$. See Figure F2 in Appendix F for the aggregated scanpath statistics analogous to Figures 6 and 7 of the extended models.

statistics of human exploration behavior remained largely unaffected (see Figure F2).

In a second extension, we included the well-established finding of *presaccadic attention* shifts (Deubel & Schneider, 1996; Rolfs, Jonikaitis, Deubel, & Cavanagh, 2011) into the model. We implemented this by prompting objects whose evidence exceeded 30% of the DDM threshold θ and setting the sensitivity map S for these objects to 1, just as if they were foveated (see Figure F1b). Again, we added this component to the base model while keeping all other model parameters unchanged. Effectively, this provides the model with additional saliency information at the most likely saccade targets, which should help to better prioritize between them. Therefore, we expected this presaccadic attention model to be more consistent in exploring the same objects as the human observers than the base model. We did not see a considerable change in the correlation of the overall object-specific dwell time when considering the whole duration for which object masks are available (90 frames; $m = 0.67$, $y_0 = 139.6$, $r^2 = 0.55$ for the base and $m = 0.72$, $y_0 = 126.6$, $r^2 = 0.56$ for the presaccadic attention model; Figure 8). In an exploratory analysis where we only considered the objects foveated in the first second, which in human scanpaths primarily corresponds with detections of the most salient objects (cf., Parkhurst, Law, & Niebur, 2002; Donk & Van Zoest, 2008), we did see an improvement in the correlation through this model extension (30 frames; $m = 0.47$, $y_0 = 101.1$, $r^2 = 0.36$ for the base and $m = 0.56$, $y_0 = 76.9$, $r^2 = 0.45$ for the presaccadic attention model). We predict that this attentional benefit would become more pronounced if we were to fit all free model parameters again for the presaccadic attention model and/or specifically fit the models to reproduce the object-specific dwell times.

Finally, we accounted for the simplified assumption that a saccade is executed immediately after the decision threshold is reached. It has been shown that new visual information does not influence the movement plan anymore in the final 50 to 70 ms of the preceding fixation (Hooge & Erkelens, 1999; Ludwig, Gilchrist, & McSorley, 2005). We implemented such a saccadic dead time in our model by prolonging each foveation by 50 ms, during which no evidence is accumulated. After this dead time, the saccade is executed as in the base model. Without fitting the parameters again, we only lowered the decision threshold $\theta = 4.0$ (base model) to $\theta' = 3.5$ to account for the otherwise 50 ms longer foveation durations, and keep all other parameters as in the base model. We find that the inclusion of this dead time does not make a qualitative difference on any of the investigated metrics (see Figure F2).

Discussion

We presented a model for object-based attention and gaze behavior in complex dynamic scenes that builds on a previous model for saccadic decision-making (Roth et al., 2023) and an object segmentation model for interactive perception in robotics (Mengers et al., 2023). The active interconnection between the two model components resembles an algorithmic information processing pattern from robotics, AICON (see Battaje

et al., 2024), which we further examine in Using an information processing pattern from robotics. Prior to this, we discuss the results of our study (Summary and evaluation of the results) as well as the limitations and advantages of our approach (Advantages and limitations of our model). In particular, we elaborate on the conclusions we can draw about uncertainty as a driving factor for visual exploration (Uncertainty drives exploration), and what we can learn from the model about the perceptual units of visual attention (Perceptual units for object-based attention).

Summary and evaluation of the results

Our scanpath model successfully replicates key aspects of human visual exploration in dynamic real-world scenes. Qualitative (Qualitative scanpath analysis) and quantitative (Aggregated scanpath statistics) comparisons between simulated and human gaze behavior demonstrate that the model closely resembles human behavior and accurately reproduces scanpath statistics. We selected the model parameters such that the simulated scanpaths fit the foveation duration and saccade amplitude statistics of human eye-tracking data. Without further fitting, the model captures meaningful exploration patterns on unseen videos, including the temporal balance between detecting new objects, inspecting currently foveated objects, and returning to previously viewed areas. This balance is primarily driven by the influence of uncertainty on saccadic decisions, which also leads to a temporal IOR effect without the need for an explicit implementation of an IOR mechanism (see Model ablation 1: Uncertainty drives exploration). We further investigated how different object sources, such as low-level and high-level cues, influence scanpaths and found that semantic object cues played a crucial role in obtaining human-like exploration (see Model ablation 2: Semantic object cues and component interconnections form suitable perceptual units). Additionally, model extensions incorporating psychophysically uncovered mechanisms like saccadic momentum and presaccadic attention have the potential to further align the model's resemblance to human behavior in terms of saccade angle distributions and object dwell time (see Model extensions: Saccadic momentum improves saccade angle statistic and presaccadic attention benefits early object detections).

Combined, the scanpath evaluation metrics in this work offer a comprehensive view of how well the model mimics human gaze behavior by assessing both temporal and spatial dynamics in visual exploration. Ideally, a single metric would capture all aspects of the simulated behavior, but currently, no established evaluation metric exists for scanpaths in dynamic scenes. For models with a readily computable sequential likelihood function, data assimilation has shown promise as an approach for both parameter fitting and model evaluation (Schütt et al., 2017; Schwetlick et al., 2020; Seelig et al., 2020; Engbert et al., 2022). Although it might be conceivable to approximate the spatiotemporal likelihood function for our model's scanpaths and update them frame by frame, this approach would be computationally infeasible. In addition to recomputing the likelihood for every frame, it is unclear how to extend the point processes used in the sequential likelihood approach to include smooth pursuit events (for a detailed discussion on additional scanpath evaluation metrics in dynamic scenes, see Roth et al., 2023).

Advantages and limitations of our model

The here presented model still has many of the simplifications of our previous framework for *Scan*path simulation in *Dy*namic scenes (*ScanDy*) (Roth et al., 2023). Importantly, we assume that attention spreads instantaneously and uniformly across objects and that saccades are always precisely executed without attempting to model the saccade programming and oculomotor control. Although we focus on scene segmentation and scanpath simulation in the current work, our modular implementation should make it easy to further extend the model in that direction. The current extensions of saccadic momentum and presaccadic attention both only required the addition of a few lines of code.

So far, we have only modeled scanpaths during free viewing, that is, observers had no task instructions. In the future, we plan to apply the same modeling approach to simulate scanpaths in complex dynamic scenes during goal-directed tasks, such as visual search and scene memorization. We expect that additional top-down attentional control during these tasks can be incorporated into the modeling by adapting the feature map F (see Scanpath simulation, F currently represents only visual saliency) and tuning the model parameters. For example, we would anticipate that our model could already reasonably simulate scanpaths for scene memorization through a down-scaling of the importance of F through $f_{\min}$ and visual search through the inclusion of a target similarity map in F. In both cases, the threshold of the DDM θ should be lowered to account for typically shorter foveation durations under such task conditions (Rayner, 2009).

The important improvement over the existing *ScanDy* framework is the active interconnection with object segmentation. Through this interaction, the model becomes image computable, that is, we do not have to define what constitutes an object a priori, but the object representations change based on the scanpath. The implementation of the object

segmentation as a recursive Bayesian filter leads to a serial dependence of the segmentation, using both prior and present object information to represent the scene (Fischer & Whitney, 2014). Furthermore, the segmentation module automatically provides us with an uncertainty map, which depends on the prior and present gaze position. We show that, through the automatic reduction of uncertainty as a consequence of saccadic decisions, this uncertainty map is well-suited to drive saccadic exploration behavior during dynamic free-viewing scenes.

Importantly, when we say we have a mechanistic model, we refer to attentional mechanisms in the sense of algorithmic principles and do not make claims on the biological or implementational level (cf., Marr, 1982). Although there is evidence for Bayesian updating in the brain (Knill & Pouget, 2004; Ma, Beck, Latham, & Pouget, 2006), even in the form of a neural particle filter (Kutschireiter, Surace, Sprekeler, & Pfister, 2017), we want to argue more conceptually for principled ways of information processing, independently of neural implementation. For example, there is evidence of bidirectional information exchange between different components of perceptual processing, similar to the exchanges between our components for object segmentation and saccadic decision-making. Such exchanges have been observed not only between different hierarchical levels of processing (Ahissar & Hochstein, 2004), but also laterally between the processing of different cues (Livingstone & Hubel, 1988) or even between separate sensory modalities (McGurk & MacDonald, 1976).

In our model, we recursively update the segmentation in the object component and the evidence in the saccadic decision component. Hence, the model makes use of the temporal consistency of the visual environment, which has also been observed and described in human behavior during visual search (e.g., Niemi & Näätänen, 1981; Kristjánsson, Sigurjónsdóttir, & Driver, 2010) and object perception (Blake & Yang, 1997; Liu, 2008). For this segmentation, we aimed to combine and compare object cues based on low-level appearance (Schyns & Oliva, 1994), motion (Reppas et al., 1997), and semantics (Neri, 2017), which have been shown to play a role in the human visual system. Although the recursive Bayesian integration of these object cues is mechanistically plausible, the way our model computes these inputs is certainly different from how the visual system might infer them. The computer vision algorithms used to obtain these cues, as described in Estimating object segmentation and its uncertainty, and particularly the semantic segmentation, on which we provide further details in Appendix B, were not chosen based on their biological plausibility but rather for how well their results represent the respective object cues as uncovered in psychophysical experiments. Similarly, the prompted semantic segmentation of the currently foveated object does not use a more plausible foveated input frame, since this would be outside the training distribution of the algorithm. Instead, we use a higher resolution of the input frame compared to the pre-attentive global segmentations, prompt the model at the current gaze position (see Appendix B for details), and include the resulting mask with higher confidence into the particle filter. An additional foveal benefit plays a role in the subsequent saccadic decision-making process, where the combination of global scene features F, and the gaze-dependent visual sensitivity S approximates the incoming information at any point in time. Our model is hence plausible on the level of attentional mechanisms and used object cues, but not on the level of how these are currently implemented.

The modular and mechanistic design of the model allows us to explore essential hypotheses about attention and gaze behavior in dynamic scenes—which can be challenging to test experimentally. By studying the model's behavior, we can generate hypotheses that can later be tested in eye-tracking experiments specifically designed for this purpose. The model offers complete control over its internal processes, allowing us to perform various ablation studies, including those on latent variables, which are usually difficult to assess in behavioral experiments. In the interpretation of our model ablation results, we assume that the other parts of our model are mechanistically similar to the human visual system. This strategy allows us to deduce how the investigated mechanism (i.e., the inclusion of uncertainty for gaze guidance or the formation of perceptual units for object-based attention) best interacts with the other model components to produce human-like gaze behavior in dynamic scenes.

In our implementation of attentional mechanisms, we focused on what we consider the core components of the vast literature on attentional guidance. In theory, including other mechanisms may change the interplay between model components and, as a result, the interpretation of our ablations. In practice, however, we find that— although our extensions of the model improve certain statistics of the simulated scanpaths—they do not qualitatively change the model's overall behavior. While this is not a guarantee that it will be the same for future model extensions, it increases our confidence in the robustness of our model and its predictive power for mechanisms of visual attention. Therefore, we can develop hypotheses about the inner workings of the human visual system by systematically examining how our model produces certain behaviors. These hypotheses can then be tested in psychophysical experiments guided by the model. In the following sections, we discuss two insights from the model and how they may inspire psychophysical experiments.

Uncertainty drives exploration

The connection between active exploration behavior and the reduction of perceived uncertainty of the environment is well established in the literature (Renninger, Verghese, & Coughlan, 2007; Sullivan, Johnson, Rothkopf, Ballard, & Hayhoe, 2012; Friston, Adams, Perrinet, & Breakspear, 2012). Gottlieb et al. (2013) summarized that "information-seeking obeys the imperative to reduce uncertainty and can be extrinsically or intrinsically motivated" (p. 586) and that "the key questions we have to address when studying exploration and information-seeking pertain to the ways in which observers handle their own epistemic states, and specifically, how observers estimate their own uncertainty and find strategies that reduce that uncertainty" (p. 586). It is, however, not obvious how uncertainty should be measured and quantified in an image-computable model of visual attention.

In this context, it is important to clarify again what we mean by uncertainty since the term can refer to many things. Our model specifically considers the uncertainty of the boundary between potential objects, both about their existence and exact location, but not about the object's identity or other possible features (for more details, see Appendix A). For example, if an object in the periphery moves, this typically would increase the uncertainty estimate in our model. One could argue that the additional motion cue should reduce the uncertainty about the shape of the object. Indeed, this intuition is reflected in our model since the input from the motion segmentation will clearly show the object. However, the overall uncertainty of the object might still increase because the exact position, shape, or state of the moving object might change, which would be reflected in conflicting object measurements from different sources or in a strong deviation from the prior belief. This prior belief is calculated as the segmentation of the previous frame, shifted by the optical flow.

Our results show that including the uncertainty map of the object segmentation module as a driving factor in the saccadic decision-making process leads to human-like simulated scanpaths. The weight of the uncertainty map for the decision-making process, parameterized through $u_{\min}$, strongly influences the ratio between foveation categories, in particular, the frequency at which objects are inspected. The prompted high-confidence object segmentation typically leads to a low uncertainty at the current position, encouraging further exploration of the scene and more return events for a strong influence of uncertainty (low $u_{\min}$). If the influence is weak (high $u_{\min}$), the gaze-dependent spread of attention leads to a strong tendency to further inspect objects with high salience. Interestingly, the $u_{\min}$ parameter also influences the strength of the temporal IOR effect. Despite returns occurring more often with a lower $u_{\min}$, the uncertainty of recently foveated objects is typically reduced, thereby slowing down the evidence accumulation process. Although IOR is generally conceived as a viewing bias that both reduces (spatial) and delays (temporal) return events, our uncertainty-guided model captures not only the temporal IOR but also the spatial "facilitation of return" (Smith & Henderson, 2009) observed in the human scanpaths.

Most mechanistic scanpath models require an explicit implementation of IOR (cf. Itti, Koch, & Niebur, 1998; Zelinsky, 2008; Schwetlick et al., 2020; Roth et al., 2023) to avoid being bound to the objects or locations with the highest salience (Itti & Koch, 2001). Our model takes a different approach, similar to previous computational models that have incorporated uncertainty-based strategies, where exploration is driven by high variance or entropy (Cohn, Ghahramani, & Jordan, 1996; Rothkopf & Ballard, 2010). It is closely related to the principle of information maximization, which has been applied before to simulate eye movements in static scenes (Lee & Yu, 1999; Renninger, Coughlan, Verghese, & Malik, 2004; Wang, Chen, Wang, Jiang, Fang, & Yao, 2011). Where our model is uncertain is also closely related to "Bayesian surprise," which was introduced by Itti and Baldi (2009) in the context of scanpaths as a measure for how eye movement data affects differences between posterior and prior beliefs of an observer about the world. These models also do not require an explicit IOR implementation, since there is little information to be gained by revisiting already foveated parts of the scene. However, when observing dynamic real-world scenes, further inspections and returns are frequent, and defining an information maximization or uncertainty-driven approach that can account for this behavior is not trivial. In our model, we do not need a separate estimation of the uncertainty, since it is a natural by product of the AICON-ic way in which we obtain the object segmentation.

Perceptual units for object-based attention

Object-based attention is a well-established concept that has been thoroughly investigated in a large variety of experimental paradigms (Scholl, 2001; Peters & Kriegeskorte, 2021; Cavanagh et al., 2023). However, it remains unclear what constitutes a visual object in this context (Spelke, 1990; Scholl, Pylyshyn, & Feldman, 2001; Feldman, 2003; Palmeri & Gauthier, 2004; Cavanagh et al., 2023). Our model allows us to systematically vary the input sources (e.g., semantic, motion-based, or appearance-based object cues) used for the formation of the scene segmentation, which defines the perceptual units on which the object-based attentional selection process operates. Under the assumption that our implementation of saccadic decision-making mechanisms is similar to the human

visual system, we expect that the object cues that lead to more human-like scanpaths are also the cues primarily used for saccadic decision-making in humans.

Our results suggest that attentional guidance primarily relies on semantic object cues in dynamic scenes. Only models that used the semantic cues both for the global and prompted scene segmentation showed the temporal IOR effect and could reproduce the balance between foveation categories seen in humans (cf. Figure 7). This result is consistent with evidence for global semantic understanding of natural scenes (Neri, 2017; Cavanagh et al., 2023). As expected, the model scanpaths also became less human-like if we replaced the prompted semantic segmentation at the gaze position with an appearance-based, low-level object cue prompted at the fixation position (*all-g & ll-p* in Figure 7, overestimating the amount of inspection and not showing the temporal IOR effect). This model corresponds to the assumption that a foveated object would get more finely segmented (e.g., a t-shirt, which was previously part of a person, becomes its own object when foveated). However, we do not see support for this assumption since the simulated scanpaths based on it were less plausible compared to the base model. Removing the global low-level object cues (*hl-g & hl-p* in Figure 7) did not impact the simulated scanpath statistics in any major way. There is ample evidence for the brain using appearance- and motion-based object cues to segment complex dynamic scenes (Schyns & Oliva, 1994; Reppas et al., 1997; Von der Heydt, 2015). Based on our results, however, we would argue that low-level object cues do not play an important role in the formation of the perceptual units on which object-based attention is operating.

These results could be tested experimentally by probing the visual sensitivity within or outside the currently foveated object as predicted by the model. A promising method to study this would be to test the response to gaze-contingent narrow-band contrast increments during free viewing (Dorr & Bex, 2013). Under the assumption of a delayed response to probes outside an attended object (Egly et al., 1994; Scholl, 2001) and in combination with the predictions from our model, this would allow us to disentangle the object cues used in the visual system to construct perceptual units for object-based attention.

Using an information processing pattern from robotics

Our model is based on the robotics-inspired information processing pattern AICON, which structures information processing at a mechanistic level to generate adaptive behavior. Our results and recent studies show that AICON is not limited to robotics, but is applicable to domains like human perception of visual illusions (Battaje et al., 2024) or even collective behavior (Mengers, Raoufi, Brock, Hamann, & Romanczuk, 2024), where systems must integrate uncertain, interdependent inputs to make perceptual decisions. Here, we present our evidence for how AICON's algorithmic patterns address the specific challenges of human vision, which show strong parallels to those in robotics. Based on this evidence, we then provide a "recipe" for building AICON-ic models of other perceptual processes.

Building a model with AICON means constructing a system of recursive components that interact through actively modulated bidirectional connections (*active interconnections*). As discussed for our model in Advantages and limitations of our model, there is ample evidence for recursive updating in human perception. These recursions within perceptual processes—often implemented in a Bayesian way—are critical for resolving ambiguous inputs, whether from sensory neurons or a robot's camera. For example, recursive processing turns depth perception, a nearly impossible task when attempted with a single image, into a trivial one by incorporating motion parallax. Active interconnections between components further refine perception, because they can share relevant information extracted through other means with each other. In integrating cues in this way, perception becomes more robust, as seen in robotics (Martín-Martín & Brock, 2022; Mengers et al., 2023). But this goes beyond simple one-directional cue integration: Because each recursive component remains uncertain, it should use all available information from its related components to reduce its uncertainty. Active interconnections are *bidirectional* and the conveyed information between components needs to adapt to changing uncertainties to ensure an information flow from more to less certain components at all times (*active modulation*). In our model, the active interconnection between the object segmentation and saccadic decision-making module leads to human-like visual exploration behavior. Likewise, Battaje et al. (2024) have shown that active interconnections between color and shape perception and between luminance and motion perception enable models to replicate the human perception of visual illusions while accounting for individual variability.

We believe AICON will be transferable to other vision processes. For those interested in building AICON-inspired models, we offer both our code (on GitHub: https://github.com/rederoth/AICONic_ScanDy) and suggest a general three-step recipe: (1) Identify key perceptual processes or representations likely to contribute to the high-level process of interest. Build a recursive model for each, ideally one that estimates uncertainty over its representation, and verify each component's behavior in isolation using controlled inputs. (2) Define and implement active interconnections between components based on possible

interdependencies, modulating these connections based on component states and uncertainties. Add connections incrementally, observing and tuning system behavior to align with expected outcomes. (3) Once fully connected, observe how behaviors emerge from component interactions. Experiment with ablating connections or adjusting parameters to refine alignment with experimental data or to generate new predictions, such as individual variability in perceptual processes.

By applying this recipe, AICON offers a versatile framework that fosters knowledge exchange across disciplines studying behavior, like robotics and vision science. Although behaviors are very different on a lower level (how computation is exactly performed) and a higher level (the exact ecological niche and its constraints), the common mechanistic challenges—integrating uncertain information and adapting across contexts—often result in convergent solutions. Therefore, we believe that studying mechanistic information processing patterns like AICON across disciplines offers a promising path toward a more unified and deeper understanding of the fundamental drivers of behavior.

Conclusion

We developed and evaluated a model for object-based attention and gaze behavior in real-world dynamic scenes. By integrating saccadic decision-making mechanisms with an object segmentation framework, our model successfully simulates human-like scanpaths. This integration, an implementation of the AICON information processing pattern from robotics, enables the model to progressively refine its object segmentation through active exploration, while uncertainty over that segmentation guides the scanpath.

The modular design of our model allows for systematic hypothesis testing and ablation studies, providing a valuable tool for exploring the mechanisms of visual attention. We found that the uncertainty in object segmentation plays a crucial role in guiding human-like visual exploration. Instead of relying on an explicit IOR mechanism, we propose the active reduction of uncertainty through saccadic decisions as the driving mechanism of scene exploration. Furthermore, our results suggest that attentional guidance primarily relies on semantic object cues, highlighting the importance of high-level scene understanding in active vision. By capturing the interplay of segmentation and saccadic decision-making, our model highlights the power of mechanistic information processing patterns like AICON, encouraging future research to explore information processing patterns that transcend disciplinary boundaries.

Keywords: scanpath simulation, object-based attention, probabilistic image segmentation, active interconnect, eye movement, active vision

Acknowledgments

The authors thank Richard Schweitzer, Madeleine Gross, and Olga Shurygina for their permission to use the UVO eye-tracking dataset and Julie Ouerfelli-Ethier for her helpful comments on the manuscript. We gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2002/1 "Science of Intelligence"—project number 390523135. We acknowledge support by the Open Access Publication Fund of TU Berlin.

Commercial relationships: none.
Corresponding authors: Vito Mengers; Nicolas Roth
Emails: v.mengers@tu-berlin.de; roth@tu-berlin.de.
Address: Technische Universität Berlin, MAR 5-1, Marchstr. 23, Berlin 10587, Germany.

[*]VM and NR contributed equally to this work.
[#]OB, KO, and MR equal supervision.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8*(10), 457–464.

Anderson, A. J., Yadav, H., & Carpenter, R. H. S. (2008). Directional prediction by the saccadic system. *Current Biology, 18*(8), 614–618.

Battaje, A., Godinez, A., Hanning, N. M., Rolfs, M., & Brock, O. (2024). An information processing pattern from robotics predicts properties of the human visual system. *bioRxiv:2024.06.20.599814*.

Bellet, M. E., Bellet, J., Nienborg, H., Hafed, Z. M., & Berens, P. (2019). Human-level saccade detection performance using deep neural networks. *Journal of Neurophysiology, 121*(2), 646–661.

Blake, R., & Yang, Y. (1997). Spatial and temporal coherence in perceptual binding. *Proceedings of the National Academy of Sciences of the USA, 94*(13), 7115–7119.

Blaser, E., Pylyshyn, Z. W., & Holcombe, A. O. (2000). Tracking an object through feature space. *Nature, 408*(6809), 196–199.

Bohg, J., Hausman, K., Sankaran, B., Brock, O., Kragic, D., Schaal, S., . . . Sukhatme, G. S. (2017). Interactive perception: Leveraging action in perception and perception in action. *IEEE Transactions on Robotics, 33*(6), 1273–1291.

Borji, A., & Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review, 97*(4), 523.

Burlingham, C. S., Sendhilnathan, N., Komogortsev, O., Scott Murdison, T., & Proulx, M. J. (2024). Motor "laziness" constrains fixation selection in real-world tasks. *Proceedings of the National Academy of Sciences of the USA, 121*(12), e2302239121.

Bylinskii, Z., DeGennaro, E. M., Rajalingham, R., Ruda, H., Zhang, J., & Tsotsos, J. K. (2015). Towards the quantitative evaluation of visual attention models. *Vision Research, 116*, 258–268.

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

Cavanagh, P., Caplovitz, G. P., Lytchenko, T. K., Maechler, M. R., Tse, P. U., & Sheinberg, D. L. (2023). The architecture of objectbased attention. *Psychonomic Bulletin & Review, 30*, 1643–1667.

Chen, P. C., & Pavlidis, T. (1980). Image segmentation as an estimation problem. *Computer Graphics and Image Processing, 12*(2), 153–172.

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research, 4*, 129–145.

Collewijn, H., Erkelens, C. J., & Steinman, R. M. (1988). Binocular coordination of human horizontal saccadic eye movements. *Journal of Physiology, 404*(1), 157–182.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*(12), 1827–1837.

Donk, M., & Zoest, W. V. (2008). Effects of salience are short-lived. *Psychological Science, 19*(7), 733–739.

Dorr, M., & Bex, P. J. (2013). Peri-saccadic natural vision. *Journal of Neuroscience, 33*(3), 1211–1217.

Droste, R., Jiao, J., & Alison Noble, J. (2020). Unified image and video saliency modeling. In *European Conference on Computer Vision,* (pp. 419–435).

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General, 113*(4), 501.

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General, 123*(2), 161.

Engbert, R., & Mergenthaler, K. (2006). Microsaccades are triggered by low retinal image slip. *Proceedings of the National Academy of Sciences of the USA, 103*(18), 7192–7197.

Engbert, R., Rabe, M. M., Schwetlick, L., Seelig, S. A., Reich, S., & Vasishth, S. (2022). Data assimilation in dynamical cognitive science. *Trends in Cognitive Sciences, 26*(2), 99–102.

Eppner, C., Höfer, S., Jonschkowski, R., Martín-Martín, R., Sieverling, A., Wall, V., . . . Brock, O. (2016). Lessons from the amazon picking challenge: Four aspects of building robotic systems. In *Robotics: Science and Systems* (pp. 4831–4835). Ann Arbor, Michigan, USA: Robotics Science and Systems Foundation.

Eriksen, C. W., & Yeh, Y-Yu. (1985). Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance*,*11*(5), 583.

Feldman, J. (2003). What is a visual object? *Trends in Cognitive Sciences, 7*(6), 252–256.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision, 59*, 167–181.

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience, 17*(5), 738–743.

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3*, 151.

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, & attention: Computational and neural mechanisms. *Trends in Cognitive Sciences, 17*(11), 585–593.

Hackett, J. K., & Shah, M. (1990). Multi-sensor fusion: A perspective. In *IEEE International Conference on Robotics and Automation (ICRA),* (pp. 1324–1330). IEEE.

Hayhoe, M. M., & Matthis, J. S. (2018). Control of gaze in natural environments: Effects of rewards and costs, uncertainty and memory in target selection. *Interface Focus, 8*(4), 20180009.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7*(11), 498–504.

Hooge, I Th C, & Erkelens, C. J. (1999). Peripheral vision and oculomotor control during visual search. *Vision Research, 39*(8), 1567–1575.

Hooge, I Th C, Over, E. A.B., Wezel, R. J.A., & Frens, M. A. (2005). Inhibition of return is not a foraging facilitator in saccadic search and free viewing. *Vision Research, 45*(14), 1901–1908.

Hopcroft, J E., & Karp, R M. (1973). An $n5/2$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing, 2*(4), 225–231.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49*(10), 1295–1306.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194–203.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.

Jonker, R., & Volgenant, T. (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing, 38*, 325–340.

Ke, L., Ye, M., Danelljan, M., Liu, Y., Tai, Yu-W, Tang, C.-K., . . . Yu, F. (2023). Segment anything in high quality. In *Proceedings of the 37th International Conference on Neural Information Processing Systems,* (pp. 29914–29934).

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., . . . Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* (pp. 4015–4026).

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences, 4*(4), 138–147.

Klein, R. M., & MacInnes, W. J. (1999). Inhibition of return is a foraging facilitator in visual search. *Psychological Science, 10*(4), 346–352.

Knill, D. C., & Pouget, A. (2004). The bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*(12), 712–719.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*(4), 219–227.

Kristjánsson, Á., Sigurjónsdóttir, Ó., & Driver, J. (2010). Fortune and reversals of fortune in visual search: Reward contingencies for pop-out targets affect search efficiency and target repetition effects. *Attention, Perception, & Psychophysics, 72*, 1229–1236.

Kümmerer, M., & Bethge, M. (2023). Predicting visual fixations. *Annual Review of Vision Science, 9*(1), 269–291.

Kümmerer, M., Bethge, M., & Wallis, T. S.A. (2022). DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision, 22*(5), https://doi.org/10.1167/jov.22.5.7.

Kutschireiter, A., Surace, S. C., Sprekeler, H., & Pfister, J.-P. (2017). Nonlinear bayesian filtering and learning: A neuronal dynamics for perception. *Scientific Reports, 7*(1), 8722.

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception, 28*(11), 1311–1328.

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature, 369*(6483), 742–744.

Lee, T. S., & Yu, S. (1999). An information-theoretic framework for understanding saccadic eye movements. *Advances in Neural Information Processing Systems, 12*.

Linka, M., & de Haas, B. (2021). Detection, inspection and re-inspection: A functional approach to gaze behavior towards complex scenes. *Journal of Vision, 21*(9), https://doi.org/10.1167/jov.21.9.1971.

Liu, H., Agam, Y., Madsen, J. R., & Kreiman, G. (2009). Timing, timing, timing: Fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron, 62*(2), 281–290.

Liu, T. (2008). Learning sequence of views of three-dimensional objects: The effect of temporal coherence on object memory. *Journal of Vision, 8*(6), https://doi.org/10.1167/8.6.516.

Livingstone, M., & Hubel, D. (1988). Segregation of form, color, movement, & depth: Anatomy, physiology, and perception. *Science, 240*(4853), 740–749.

Ludwig, C. J.H., Gilchrist, I. D., & McSorley, E. (2005). The remote distractor effect in saccade programming: Channel interactions and lateral inhibition. *Vision Research, 45*(9), 1177–1190.

Luke, S. G., Smith, T. J., Schmidt, J., & Henderson, J. M. (2014). Dissociating temporal inhibition of return and saccadic momentum across multiple eye-movement tasks. *Journal of Vision, 14*(14), https://doi.org/10.1167/14.14.9.

Ma, W Ji, Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9*(11), 1432–1438.

Malcolm, G. L., & Shomstein, S. (2015). Object-based attention in realworld scenes. *Journal of Experimental Psychology: General, 144*(2), 257–263.

Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company.

Martín-Martín, R., & Brock, O. (2022). Coupled recursive estimation for online interactive perception of articulated objects. *International Journal of Robotics Research, 41*(8), 741–777.

Samir Matthis, J., Yates, J. L., & Hayhoe, M. M. (2018). Gaze and the control of foot placement when walking in natural terrain. *Current Biology, 28*(8), 1224–1233.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748.

Mengers, V., Battaje, A., Baum, M., & Brock, O. (2023). Combining motion and appearance for robust probabilistic object segmentation in real time. In: *2023 IEEE International Conference on Robotics and Automation (ICRA),* (pp. 683–689).

Mengers, V., Raoufi, M., Brock, O., Hamann, H., & Romanczuk, P. (2024). Leveraging uncertainty in collective opinion dynamics with heterogeneity. *Scientific Reports, 14*(1), 27314.

Milner, D., & Goodale, M. (2006) *The visual brain in action,* volume *27*. Oxford, UK: Oxford University Press.

Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation, 3*(1), 5–24.

Lottier Molin, J., Etienne-Cummings, R., & Niebur, E. (2015). How is motion integrated into a proto-object based visual saliency model? In: *2015 49th Annual Conference on Information Sciences and Systems (CISS)*.

Motter, B. C. (2018). Saccadic momentum and attentive control in v4 neurons during visual search. *Journal of Vision, 18*(11), https://doi.org/10.1167/18.11.16.

Neisser, U. (1967) *Cognitive psychology*. New York: Appleton-Century-Crofts.

Neri, P. (2017). Object segmentation controls image reconstruction from natural scenes. *PLoS Biology, 15*(8), e1002611.

Niemi, P., & Näätänen, R. (1981). Foreperiod and simple reaction time. *Psychological Bulletin, 89*(1), 133–162.

Nobre, K., & Kastner, S. (2014) *The Oxford handbook of attention*. Oxford, UK: Oxford Library of Psychology.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*(8), https://doi.org/10.1167/10.8.20.

Nuthmann, A., Einhäuser, W., & Schütz, I. (2017). How well can saliency models predict fixation selection in scenes beyond central bias? A new approach to model evaluation using generalized linear mixed models. *Frontiers in Human Neuroscience, 11* (491).

O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fmri evidence for objects as the units of attentional selection. *Nature, 401*(6753), 584–587.

Pajak, M., & Nuthmann, A. (2013). Object-based saccadic selection during scene perception: Evidence from viewing position effects. *Journal of Vision, 13*(5), https://doi.org/10.1167/13.5.2.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience, 5*(4), 291–303.

Pantofaru, C., Schmid, C., & Hebert, M. (2008). Object recognition by integrating multiple image segmentations. In: *European Conference on Computer Vision (ECCV)* (pp. 481–494). Springer.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*(1), 107–123.

Pelz, J., Hayhoe, M., & Loeber, R. (2001). The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research, 139*, 266–277.

Peters, B., & Kriegeskorte, N. (2021). Capturing the objects of vision with neural networks. *Nature Human Behaviour, 5*(9), 1127–1144.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*(1), 3–25.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. *Attention and Performance X: Control of Language Processes, 32*, 531–556.

Potter, M. C., Wyble, B., Erick Hagmann, C., & McCourt, E. S. (2014). Detecting meaning in rsvp at 13 ms per picture. *Attention, Perception, & Psychophysics, 76*, 270–279.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20*(4), 260–281.

Rayner, K. (2009). The 35th Sir Frederick Bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*(8), 1457–1506.

Renninger, L., Coughlan, J., Verghese, P., & Malik, J. (2004). An information maximization model of eye movements. *Advances in Neural Information Processing Systems, 17*.

Renninger, W. L., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision, 7*(3), https://doi.org/10.1167/7.3.6.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition, 7*(1–3), 17–42.

Reppas, J. B., Niyogi, S., Dale, A. M., Sereno, M. I., & Tootell, R. B.H. (1997). Representation of motion boundaries in retinotopic human visual cortical areas. *Nature, 388*(6638), 175–179.

Rolfs, M., Jonikaitis, D., Deubel, H., & Cavanagh, P. (2011). Predictive remapping of attention across eye movements. *Nature Neuroscience, 14*(2), 252–256.

Roth, N., Rolfs, M., Hellwich, O., & Obermayer, K. (2023). Objects guide human gaze behavior in dynamic real-world scenes. *PLOS Computational Biology, 19*(10), e1011512.

Rothkopf, C. A., & Ballard, D. H. (2010). Credit assignment in multiple goal embodied visuomotor behavior. *Frontiers in Psychology, 1*(173).

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision, 7*(14), https://doi.org/10.1167/7.14.16.

Saenz, M., Buracas, G. T., & Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience, 5*(7), 631–632.

Särkkä, S. (2013) *Bayesian filtering and smoothing*. Cambridge, UK: Cambridge University Press.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80*(1–2), 1–46.

Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object? Evidence from target merging in multiple object tracking. *Cognition, 80*(1–2), 159–177.

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Reich, S., AWichmann, F., & Engbert, R. (2017). Likelihood-based parameter estimation and comparison of dynamical cognitive models. *Psychological Review*, *124*(4), 505.

Schweitzer, R., & Rolfs, M. (2022). Definition, modeling, and detection of saccades in the face of post-saccadic oscillations. In: *Eye Tracking: Background, Methods, and Applications,* (pp. 69–95). New York: Springer Science+Business Media.

Schwetlick, L., Rothkegel, L. O. M., Trukenbrod, H. A., & Engbert, R. (2020). Modeling the effects of perisaccadic attention on gaze statistics during scene viewing. *Communications Biology, 3*(727).

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science, 5*(4), 195–200.

Seelig, S. A., Rabe, M. M., Malem-Shinitski, N., Risse, S., Reich, S., & Engbert, R. (2020). Bayesian parameter estimation for the swift model of eye-movement control during reading. *Journal of Mathematical Psychology, 95*, 102313.

Shi, X., Huang, Z., Bian, W., Li, D., Zhang, M., Cheung, K. C., . . . Li, H. (2023). Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 12469–12480.

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

Smith, T. J., & Henderson, J. M. (2011) (1990). Looking back at waldo: Oculomotor inhibition of return does not prevent return fixations. *Journal of Vision, 11*(1), https://doi.org/10.1167/11.1.3.

Spelke, E. S. Principles of object perception. *Cognitive Science, 14*(1), 29–56.

Stewart, E. E.M., Ludwig, C. J.H., & Schütz, A. C. (2022). Humans represent the precision and utility of information acquired across fixations. *Scientific Reports, 12*(1), 2411.

Sullivan, B. T., Johnson, L., Rothkopf, C. A., Ballard, D., & Hayhoe, M. (2012). The role of uncertainty and reward on eye movements in a virtual driving task. *Journal of Vision, 12*(13), https://doi.org/10.1167/12.13.19.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11*(5), https://doi.org/10.1167/11.5.5.

Tatler, B. W., Brockmole, J. R., & Carpenter, R. H.S. (2017). Latest: A model of saccadic decisions in space and time. *Psychological Review, 124*(3), 267.

Tenenbaum, J. M., & Barrow, H G. (1977). Experiments in interpretationguided segmentation. *Artificial Intelligence, 8*(3), 241–274.

Thrun, S., Burgard, W., & Fox, D. (2005) *Probabilistic Robotics*. Cambridge, MA: MIT Press.

Tipper, S. P., Driver, J., & Weaver, B. (1991). Object-centred inhibition of return of visual attention. *Quarterly Journal of Experimental Psychology, 43*(2), 289–298.

Treue, S., & Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature, 399*(6736), 575–579.

Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. *Journal of Vision, 3*(1), https://doi.org/10.1167/3.1.9.

Ullman, S. (1984). Visual routines. *Cognition, 18*(1-3), 97–159, https://doi.org/10.1016/0010-0277(84)90023-4.

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition, 17*(1), 159–170.

Von der Heydt, R. (2015). Figure–ground organization and the emergence of proto-objects in the visual cortex. *Frontiers in Psychology, 6*(1695).

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., ... Von der Heydt, R. (2012). A century of gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin, 138*(6), 1172.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*(9), 1395–1407.

Wang, W., Chen, C., Wang, Y., Jiang, T., Fang, F., & Yao, Y. (2011). Simulating human saccadic scanpaths on natural images. In: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition,* (pp. 441–448).

Wang, W., Feiszli, M., Wang, H., & Tran, Du. (2021). Unidentified video objects: A benchmark for dense, open-world segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* (pp. 10776–10785).

Wang, W., Shen, J., Guo, F., Cheng, M.-M., & Borji, A. (2018). Revisiting video saliency: A large-scale benchmark and a new model. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* (pp. 4894–4903).

White, A. L., & Carrasco, M. (2011). Feature-based attention involuntarily and simultaneously improves visual performance across locations. *Journal of Vision, 11*(6), https://doi.org/10.1167/11.6.15.

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

Wloka, C., Kotseruba, I., & Tsotsos, J. K. (2018). Active fixation control to predict saccade sequences. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* (pp. 3184–3193).

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

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

World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *Journal of the American Medical Association, 310*(20), 2191–2194, ISSN 0098-7484, https://doi.org/10.1001/jama.2013.281053.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review, 115*(4), 787–835.

Zhao, Xu, Ding, W., An, Y., Du, Y., Yu, T., Li, M., ... Wang, J. (2023) Fast segment anything. *arXiv preprint arXiv:2306.12156*.

Appendix A: Uncertainty over object segmentation

We estimate the uncertainty over object segmentation using a particle filter, as described in Estimating object segmentation and its uncertainty. This uncertainty reflects the ambiguity in the existence and location of boundaries between objects and not the identity or category of individual objects. Given its key role in our model, we provide an intuitive explanation of this uncertainty and how it is updated on a frame-by-frame basis, illustrated with an example sequence in Figure 4. We visualize the model's uncertainty U' both as an average across the entire scene (Figure A1b) and as an average over the ground truth objects in the video (Figure A1a, with the ground truth objects shown in Figure A1c). For individual ground truth objects, we show how gaze position affects object uncertainty for a single simulated scanpath over the first 90 frames with available ground truth. For the global uncertainty across the scene, we average over 10 stochastic scanpath realizations to observe the general relationship between uncertainty and scene content, independent of the specific scanpath.

After the initial pre-attentive global segmentation, the uncertainty for different objects varies based on visual ambiguity in appearance and semantic cues (cf., Figure A1a). Uncertainty for non-foveated objects depends on factors like motion or occlusion. However, if an object is foveated (e.g., the red object at frame 50), its uncertainty rapidly decreases to nearly the minimum ($u_{\min} = 1/3$). When gaze shifts away from this object (e.g., in frame 58), its uncertainty rises again. This dependency on various factors, combined with diverse uncertainties across objects, means that average uncertainty over the scene remains relatively steady throughout the sequence (cf. Figure A1b). This stability arises because we are not estimating uncertainty over object identity, which would remain low once identified, but over object boundaries, which can quickly become ambiguous again as objects move, their visible parts change, or occlusions occur. If there is no camera movement and the objects in the scene remain mostly static, the overall uncertainty decreases (cf., the green intervals in Figure A1b, where dancers either slowly rotate or have a stop-step and thus move less).

As a result, the global uncertainty remains on a similar level over time as long as there are ongoing scene changes in the video. Hence, we did not expect a strong effect on the temporal development of uncertainty-related effects. To test this, we compared the temporal IOR effect for the first and second half of the videos (frames 0–149 and frames 150–299) for both the base model and the human data, as shown in Figure A2. The general trend of a temporal IOR effect remains

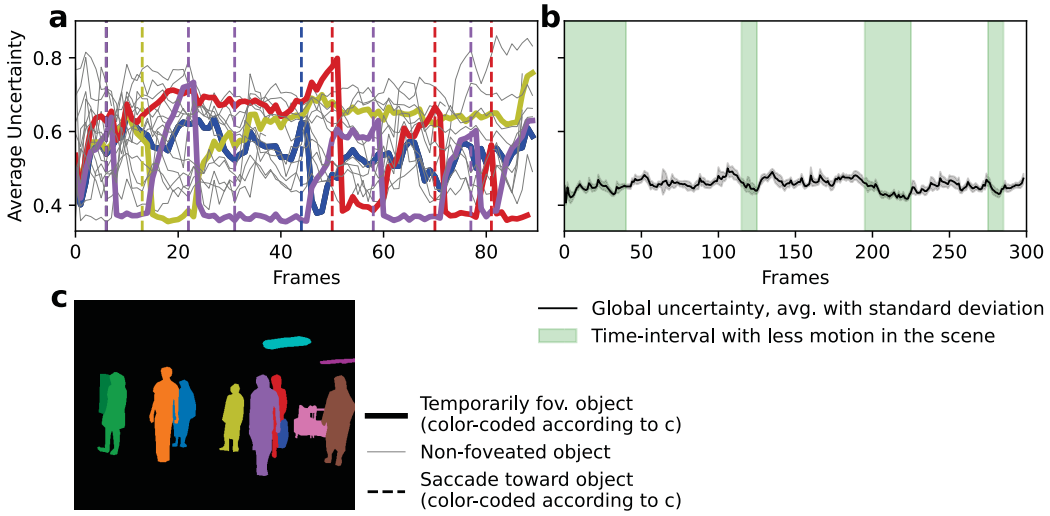

Figure A1. Uncertainty in our model represents where the boundary of objects is currently ambiguous. We visualize the uncertainty U' of our model for the same scene as in Figure 4. We show the uncertainty for individual ground truth objects in (a) with the ground truth objects of that scene shown for the initial frame in (c). The uncertainty of non-foveated objects (thin gray lines) varies for different objects and over time, but after a saccade towards an object (dashed colored line indicates a saccade towards the same color object) uncertainty for that object (thick colored line) rapidly reduces and remains low until the gaze moves away. (b) Global uncertainty (averaged over 10 scanpaths with standard deviation) remains in the same regime but reduces if there is less motion in the scene (green time intervals), since ambiguity can be resolved while less new ambiguity arises from motion.

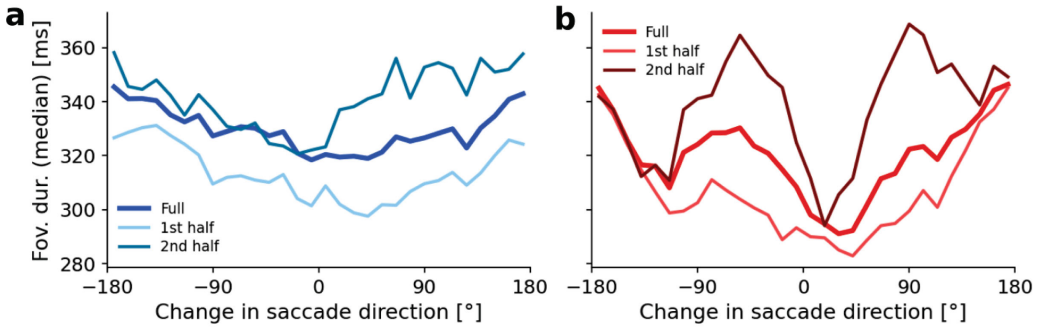

Figure A2. Temporal IOR effect during the first and second half of presentation time. (a) Evaluation of the base model with the result of the full duration, as shown in Figure 5d, compared to an evaluation of foveations only during the first (frames 0–149) or second half (frames 150–299). (b) Same as (a) for human data.

visible for both halves, only with higher variability due to the smaller number of samples. However, the average foveation durations are shorter for the first half of frames than for the second for both our model and the human data. In the model, this must be the result of changes in uncertainty, since the salient feature map F is always normalized and the visual sensitivity S does not change in magnitude. However, because the global uncertainty also does not change drastically, this is most likely a result of the distribution of uncertainty compared with the distribution of saliency. The first saccades favor objects with both high saliency and uncertainty, leading to a reduction of uncertainty for these objects. Due to this reduced uncertainty, evidence for these objects that initially drive a fast sequence of saccades is then accumulated more slowly, leading to overall longer foveation durations.

In summary, uncertainty in our model represents where the boundary of objects is currently ambiguous. This can fluctuate depending on the scene content, but by foveating on an object, the model ensures a lower uncertainty for that object. However, uncertainty for objects can rise after gaze shifts away, leading to a relatively stable amount of global uncertainty over the scene that distributes according to scanpath history and scene content over time.

Appendix B: Semantic and foveated segmentation from prompt-based models

We use a state-of-the-art data-driven model (Kirillov et al., 2023) to generate both the pre-attentive global segmentations and the segmentations of the currently foveated object, as described in Estimating object segmentation and its uncertainty. This family of models is fundamentally based on the formulation of segmentation as a prompt-based task. Here, we give a high-level explanation of this task formulation and how such models are trained with vast amounts of data to provide an intuition on how the semantic object cues were obtained. For further details on the general concept, refer to Kirillov et al. (2023), and for details on the specific derived models used here, see Ke et al. (2023) and Zhao et al. (2023).

Traditionally in computer vision, segmentation has been formulated as the task of generating a dense map for a given image that separates it into regions, often based on known object classes. Kirillov et al. (2023) introduced an alternative formulation in which, given a prompt relating to one object in the image (a point, a bounding box, text, or a dense mask), the mask covering that object needs to be identified. For this formulation of the segmentation task, a learnable model consists of encoders for the image as well as all possible prompts and a decoder that, given the encoded image and prompt, generates a mask. These encoders and decoders are usually transformer-based and can then be trained in unison based on a vast amount of data of labeled segmentations collected from various sources (Kirillov et al., 2023). The new task formulation, moreover, allows leveraging labeled segmentations in multiple ways, since one image can be used for all types of prompts, as well as variations of the same prompt, e.g., by shifting the prompt point within the labeled mask of that object. Thus, the amount of training samples increases and a model that can solve a variety of segmentation tasks can be learned.

We use such learned models with prompt-points to generate both pre-attentive global semantic segmentations and the segmentations of the currently foveated object. Let us give an example based on Figure 1 starting with the more intuitive segmentation of the foveated object: the current gaze acts as prompt-point that currently lies on the face of the person in the foreground, and thus an appropriate mask (the segmentation of the foveated object) contains the entire person. Note that alternative appropriate masks might only be the person's face or maybe even only the specific part of the face the prompt is on, for example, the left eye. Hence, models for this task do not output just one mask, but a weighted set of masks from which the user can select (we simply use the highest weighted mask throughout this work). This segmentation procedure is in itself not foveated, because it leverages the entire image in a spatially invariant manner during encoding and only leverages the prompt-point during the final mask construction. To generate a global semantic segmentation of the scene with such a model, instead of passing one point, we pass a grid of points to obtain a set of masks that can be combined into one segmentation (cf., the semantic segmentation in Figure 2).

Appendix C: Further details on the particle filter implementation

We track a belief over scene segmentation by combining different measurements over time within a particle filter, as we describe in Estimating object segmentation and its uncertainty. Here, we want to give further details on its implementation, especially regarding the computation of each particle's weight and the matching process for segmentation IDs when marginalizing the particle set into a single object segmentation.

Appendix C1: Particle weighting

When computing the weight of each particle (Equation 1), we weigh it according to each segmentation cue. To do so, we first compute the unnormalized weights $\widetilde{w}_t^{[i]}(z_t)$ according to each cue z_t, using a distance function between two segmentations (Equation 13). We can determine this distance between two segmentations as the sum of the distances of each boundary pixel in one segmentation to the closest boundary pixel in the other, which is easily computable using the distance transform disttransform(s) of the boundary image s of a segmentation. Because this distance is non-symmetric, we, however, need to use it in both directions (Equation 14 where W and H are the width and height of the image frame).

$$\widetilde{w}_t^{[i]}(z_t) = \frac{1}{d(s_t^{[i]}, z_t)} \quad (13)$$

$$d(s_1, s_2) = \sum_{x=1}^{W} \sum_{y=1}^{H} \Big((s_1)_{xy} \cdot (\text{disttransform}(s_2))_{xy} + (s_2)_{xy} \cdot (\text{disttransform}(s_1))_{xy} \Big) \quad (14)$$

To determine the overall weight of a particle according to the set of all cues $\mathcal{Z}_t$ at time t, we combine

the unnormalized weights $\widetilde{w}_t^{[i]}(z_t)$ for each cue z_t as in a product, as shown in Equation 15, where η is a normalizing factor between particles. However, as the cues have different amounts of information and thus confidence, we combine with an additional exponential importance factor α_z. These importance factors were set during some initial explorations on the dataset to produce satisfactory segmentations as shown in Table D1.

$$w_t^{[i]}(\mathcal{Z}_t) = \frac{1}{\eta} \prod_{z_t \in \mathcal{Z}_t} \left(\widetilde{w}_t^{[i]}(z_t)\right)^{\alpha_z} \qquad (15)$$

Appendix C2: Matching segmentation IDs for consistency over time

We obtain a single segmentation from the particle set during each iteration to inform saccadic decision-making, as we have described in Estimating object segmentation and its uncertainty. To keep the IDs of this segmentation consistent, we use a variation of the Hungarian algorithm (Hopcroft & Karp, 1973) to match object IDs between object segmentations. To do so, we must determine the matching weights $w_m(m_1, m_2)$ between the mask m_1 of an object in one segmentation and the mask m_2 in another. We use the well-established intersection over union $\mathrm{IOU}(m_1, m_2)$ metric to measure their overlap:

$$\mathrm{IOU}(m_1, m_2) = \frac{m_1 \cap m_2}{m_1 \cup m_2}. \qquad (16)$$

However, if we only consider these overlaps between the current and last segmentation, some object IDs will get lost due to perceptual uncertainty. Hence, we consider the last 10 segmentations, but discount their importance with the factor β. We compute resulting matching weights $w_m(m_1, m_2)$ that the mask m_1 in the current segmentation should have the same ID as the mask m_2 in each of the last $T = 10$ segmentations following

$$w_m(m_1, m_2) = \sum_{t=0}^{T} \mathrm{IOU}(m_1, m_{2,(T-t)}) \cdot \beta^{(T-t)}, \qquad (17)$$

and then match the IDs using maximum weight full matching in bipartite graphs (Jonker & Volgenant, 1987), allowing for new IDs if no existing ID can be matched.

$\boldsymbol{\theta} = 4.0$	Decision threshold of the drift-diffusion model.
$\mathbf{s} = 0.4$	Decision noise of the drift-diffusion model.
$\mathbf{u_{min}} = 1/3$	Parameter to rescale the uncertainty map U to $U' \in [u_{\mathrm{min}}, 1]$. Increasing u_{min}, reduces the importance of uncertainty in our model.
$\mathbf{f_{min}} = 0$	Parameter to rescale the salience map F to $F' \in [f_{\mathrm{min}}, 1]$. Increasing f_{min}, reduces the importance of salience in our model.
$\sigma_S = 7$ dva	Standard deviation of the isotropic Gaussian $G_S = \frac{1}{2\pi\sigma_S^2} \exp\left(-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma_S^2}\right)$ that we use to model visual sensitivity.
$N = 50$	Number of segmentation particles. Higher number improves estimation of both segmentation and uncertainty, but heavily increases computational load. We find that $N = 50$ is already sufficient, given the direct insertion of regions to prevent divergence.
$\alpha_{\mathrm{appearance}} = 0.4$	Importance factor of the appearance segmentation to determine the weight of each particle.
$\alpha_{\mathrm{motion}} = 0.05$	Importance factor of the motion segmentation to determine the weight of each particle. It is set much lower than others since the motion segmentation only provides information about some parts of the scene, while also being noisy.
$\alpha_{\mathrm{semantic}} = 1.0$	Importance factor of the semantic segmentation to determine the weight of each particle.
$\alpha_{\mathrm{foveated}} = 0.6$	Importance factor of the foveated segmentation to determine the weight of each particle. It is lower than for the semantic segmentation, since it only provides information on part of the scene, while having in principle the highest confidence.
$r_{\mathrm{scale,foveated}} = 1.0$	Scale factor of the input resolution for the foveated object segmentation.
$r_{\mathrm{scale,other}} = 0.35$	Scale factor of the input resolution for all other segmentation cues. They are downsampled to model lower-confidence information.

Table D1. Parameters of our model. We show their settings for our base model, with the parameters that are fitted for different versions of the model in **bold**.

Appendix D: Parameter exploration

We found appropriate parameter values through extensive grid searches in a four-dimensional parameter space, as described in Metrics and parameter fitting. To make the computational cost of the grid search feasible, we fixed all parameters except for the decision threshold θ for the DDM, the DDM noise level s, and the scaling parameters for the importance of the uncertainty u_{min} and salient scene features f_{min}. All free and fixed model parameters are described in Table D1.

We used 10 videos from our dataset as a training set (33 for testing) and generated 5 scanpaths stochastically for each video and parameter configuration. To confirm that five scanpath realizations were sufficient to estimate the model's free parameters reliably, we assessed the variability of the KS statistic over different numbers of realizations. For this, we simulated for each video in the training set 30 scanpath realizations using the base model parameters. For each number $N \in [2, 29]$, we then randomly drew 25 sets of size N out of the 30 realizations and calculated the KS statistic for each set. The standard deviation of the KS statistic across the 25 sets for each number of stochastic realizations N is shown in Figure D1. We used the difference in KS statistic between the two best-fitting parameter sets as a reference value and found that with five realizations, the KS statistic's standard deviation was already below that difference. Increasing the number of realizations beyond five only slightly reduced variability. Therefore, we used five stochastic scanpath realizations to reduce the computational cost of the grid search. The model's strong generalization from training to test set throughout our results validated this approach.

We first ran a coarse parameter grid exploration for the parameters $\theta \in [2, 3, 4, 5, 6]$, $s \in [0.1, 0.2, 0.3, 0.4]$, $u_{\text{min}} \in [0, \frac{1}{10}, \frac{1}{5}, \frac{1}{3}, \frac{1}{2}]$, and $f_{\text{min}} \in [0, \frac{1}{10}, \frac{1}{5}, \frac{1}{3}]$. Around the best-performing parameters, we performed a finer grid search in θ and s, as shown in Figures D2 and D3. We did not consider parameter sets with $s > 0.4$ because previous model explorations have shown that the simulated scanpaths for higher noise levels are more likely to explore the background or objects that are not often foveated by human observers. As the main indication for noise-driven scanpaths, we took a lower correlation of the object dwell time between simulated and human scanpaths, as it is shown in Figure 8b, which in fact decreases for models with $s > 0.4$.

To ensure a fair comparison between models in our ablation studies, we run additional parameter explorations for the models where the foveation duration and saccade amplitude change considerably compared to the base model with the parameters in Table D1. For the model without uncertainty contribution (*no uncert.* in Figure 6) we set $U' = u_{\text{min}}$, resulting in a model with $\theta = 3.0$, $s = 0.3$, $f_{\text{min}} = 0$, $u_{\text{min}} = \frac{1}{3}$ having the lowest mean of KS statistics D_{FD} and D_{SA} across the four-dimensional grid of free parameters. When investigating the influence of different object cues, we explored a fine parameter grid $f_{\text{min}} = 0$, $u_{\text{min}} = \frac{1}{3}$ for better comparability with the other models. This

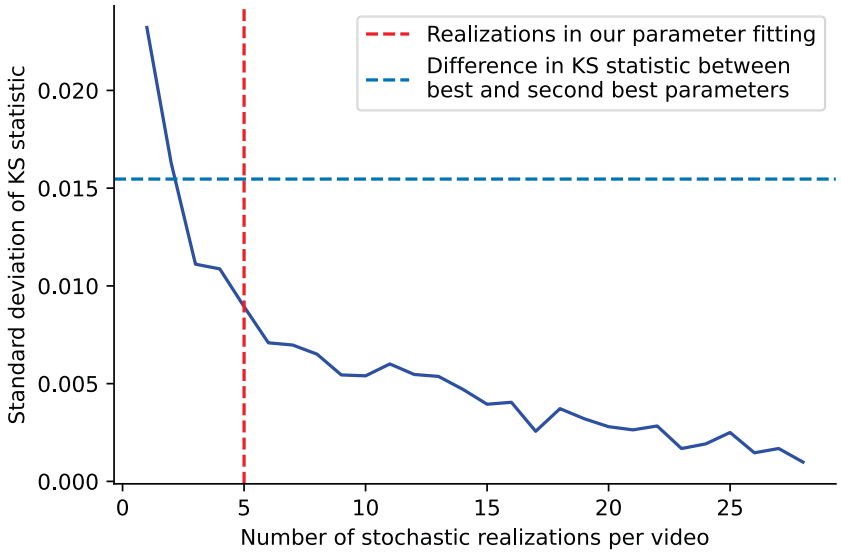

Figure D1. To ensure a reliable estimation of the model's free parameters on the training set, we compared the variability of the KS statistic for the base model across different numbers of stochastic scanpath realizations to the difference between the two best-fitting parameter sets. Initially, adding more realizations significantly reduced the standard deviation, bringing it below the difference in KS statistic between the best-fitting parameter sets. Beyond this point, additional realizations only gradually reduced variability. Therefore, we selected 5 stochastic scanpath realizations per video to fit our model parameters.

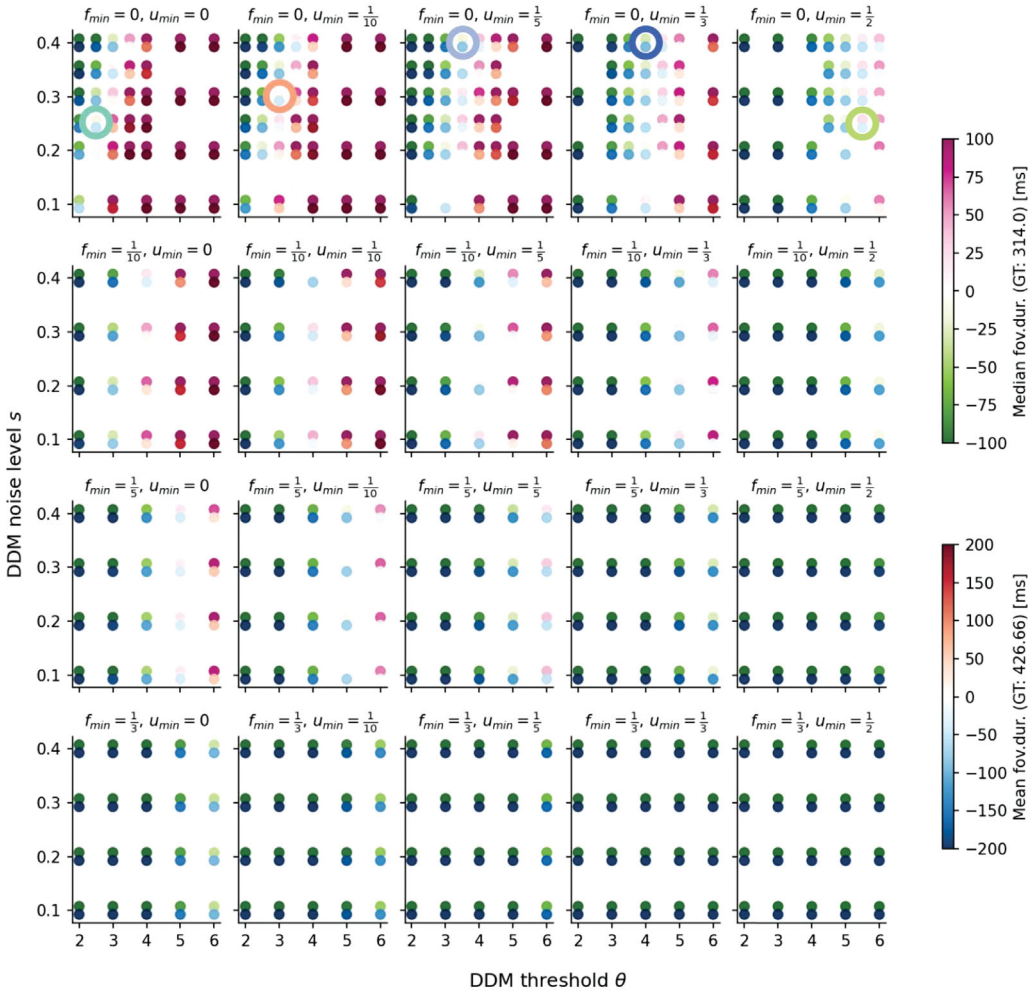

Figure D2. The distribution of foveation durations is one criterion to determine model parameters within the four-dimensional grid of free parameters. Each dot-tuple characterizes the deviation of the median (higher) and mean (lower dot) foveation duration of simulated scanpaths compared to the human ground truth (GT) in the training set (10 videos; 5 random seeds each). Brighter dots indicate more suitable parameters. Circles mark the chosen parameter sets for each value of u_{min}, which we subsequently analyzed in detail as shown in Figure 6.

resulted in parameter values of $\theta = 4.0$, $s = 0.4$ for the model using ground truth objects (*gt-obj* in Figure 7), $\theta = 5.5$, $s = 0.4$ for the model with all global object cues, but without a prompted object (*all-g & no-p*), and $\theta = 5.5$, $s = 0.4$ for the model with global appearance and motion-based segmentation only (*ll-g & no-p*).

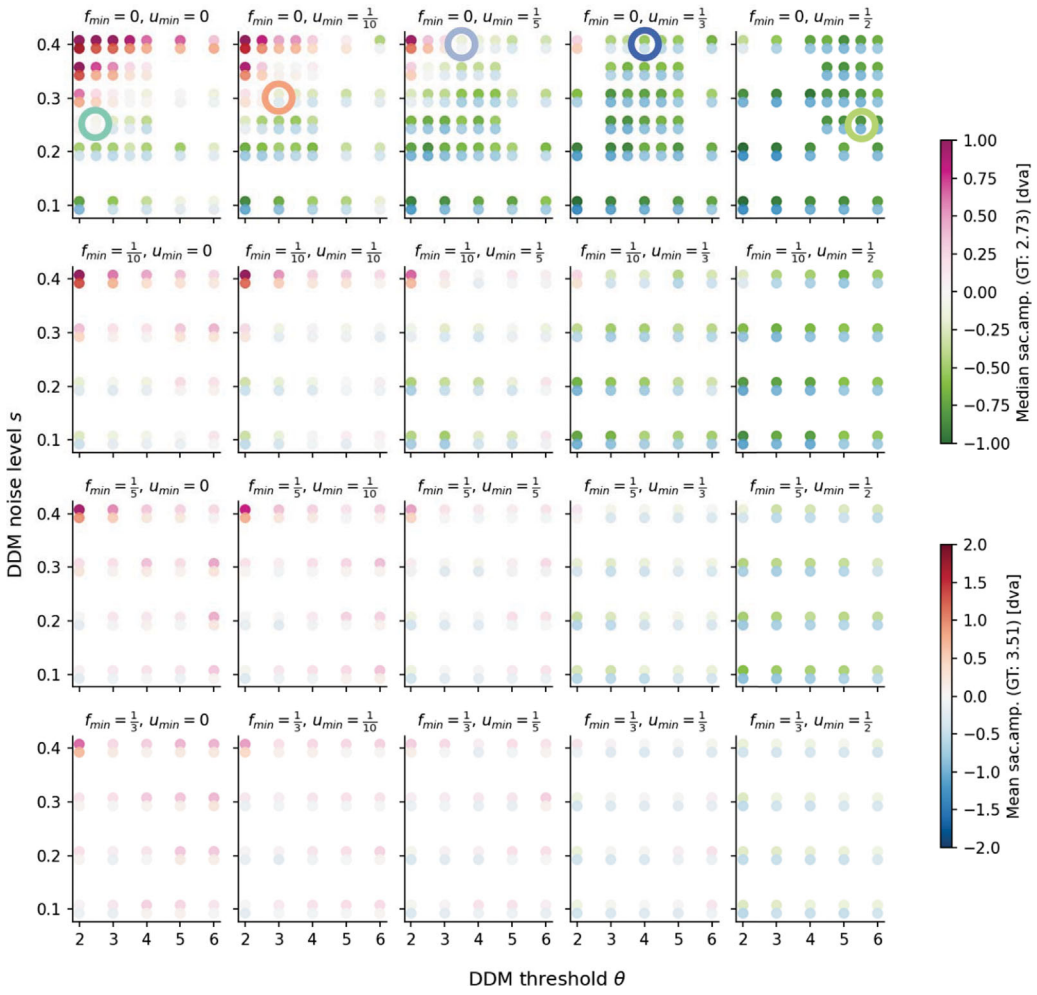

Figure D3. The saccade amplitude distribution of the simulated scanpaths is the second criterion, plotted analogously to Figure D2.

Appendix E: Videos of human and model scanpaths

The visualizations of our model parts shown in Figures 1 to 4 can be seen as downloadable (https://doi.org/10.14279/depositonce-22812) videos for 10 different simulated scanpaths on that input sequence.

We show 10 simulated scanpaths for 10 additional videos from the test set to illustrate the variability of our dataset. For comparison, we also show the scanpaths of 10 human participants on the respective input sequences. All videos are shown with a playback speed of 0.5 (i.e., 15 fps instead of 30 fps) to make it easier to compare the scanpaths.

Appendix F: Extended models: Details and statistics

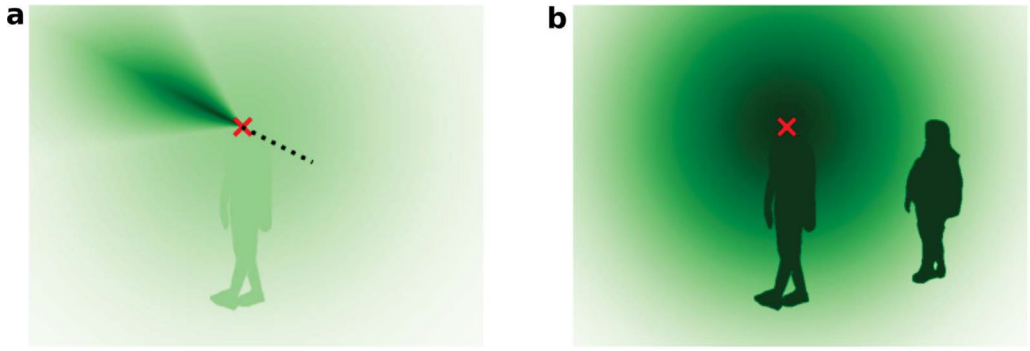

Figure F1. Illustration of the modified sensitivity maps S' for the two extended models. (a) Saccadic momentum: We set the maximal value in the direction of the previous saccade (indicated with the dotted line) to 2.5, which decreases linearly to 0.85 within an angle of 35°, and multiply the resulting map with S. (b) Presaccadic attention: If the evidence of an object crossed 30% of the decision threshold θ, we obtain a prompted object mask at its location and set the sensitivity of this object to 1.

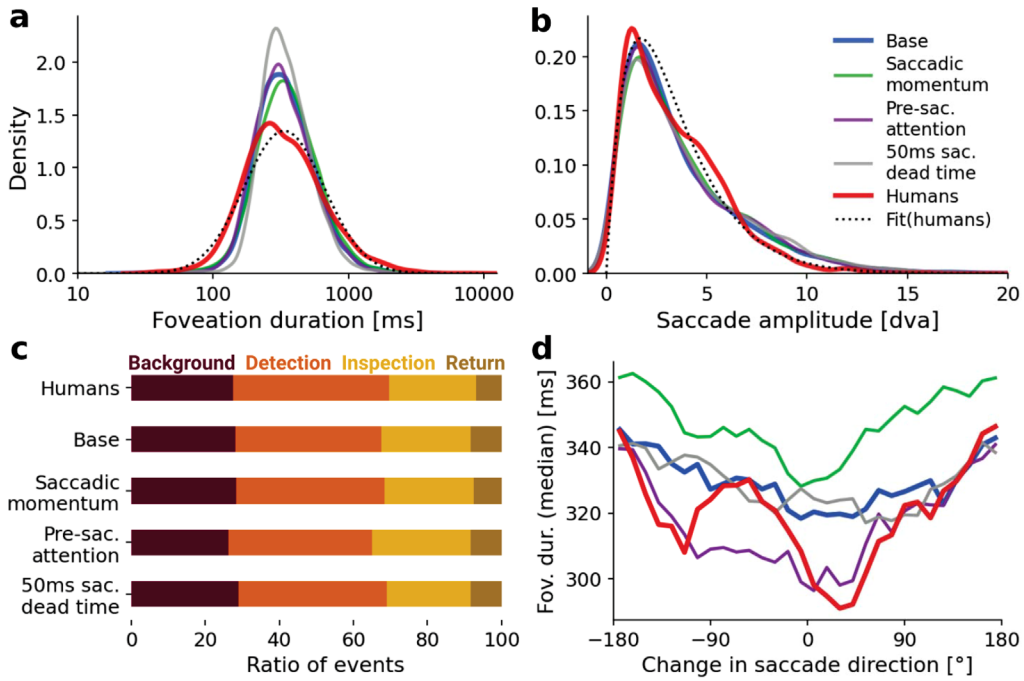

Figure F2. Model extensions do not qualitatively change the aggregated scanpath statistics. (a–d) Analogous to Figure 6 but for different model extensions. Plotted are the base model (*blue*, cf. Figure 5), its extension with saccadic momentum (*green*), pre-saccadic attention (*purple*), and a saccadic dead time of 50 ms (*gray*) compared to the human data (*red*).