Behavioral/Cognitive

Spatiotemporal Evidence Accumulation Through Saccadic Sampling for Object Recognition

Zhihao Zheng,[1] Jiaqi Hu,[1,2] and Gouki Okazawa[1,2]

[1]Institute of Neuroscience, Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Center for Excellence in Brain Science and Intelligence Technology, International Center for Primate Brain Research, Chinese Academy of Sciences, Shanghai 200031, China and [2]University of Chinese Academy of Sciences, Beijing 101408, China

Visual object recognition has been extensively studied under fixation conditions, but our natural viewing involves frequent saccadic eye movements that scan multiple local informative features within an object (e.g., eyes and mouth in a face image). These saccades would contribute to object recognition by subserving the integration of sensory information across local features, but mechanistic models underlying this process have yet to be established due to the presumed complexity of the interactions between the visual and oculomotor systems. Here, we employ a framework of perceptual decision making and show that human object categorization behavior with saccades can be quantitatively explained by a model that simply accumulates the sensory evidence available at each moment. Human participants of both sexes performed face and object categorization while they were allowed to freely make saccades to scan local features. Our model could successfully fit the data even during such a free viewing condition, departing from past studies that required controlled eye movements to test trans-saccadic integration. Moreover, further experimental results confirmed that active saccade commands (efference copy) do not substantially contribute to evidence accumulation. Therefore, we propose that object recognition with saccades can be approximated by a parsimonious decision-making model without assuming complex interactions between the visual and oculomotor systems.

Key words: decision making; evidence accumulation; face recognition; object recognition; saccadic eye movement

Significance Statement

When we view an object to judge its identity or properties, we move our eyes to inspect multiple local features, gathering dynamic information. How does object recognition unfold during this complex sequence of events? To explain object recognition with saccades, should we model precisely how the visual and oculomotor systems exchange information in the brain? Instead, we demonstrate that human object recognition can be quantitatively explained by a decision-making model that processes each snapshot of an image sequence and simply integrates information over the course of multiple eye movements. This model approximates human behavior without additional mechanisms, even under experimental conditions in which people freely move their eyes to scan local features without constraint during face and object recognition.

Introduction

Object recognition is often studied as the process of extracting object information from a static image, but during a natural visual experience, our visual system is bombarded with frequent changes in the retinal image due to our own eye movements. Saccadic eye movements, which occur on average 1–3 times per second (Otero-Millan et al., 2013), help us search for and focus on an important object in a visual scene (Eckstein, 2011). However, even when looking at a single object, we often make saccades within the object to scan multiple local features. For example, in classic demonstrations by Yarbus (1967), people viewing a face made frequent saccades across features such as eyes and mouth.

A number of studies have investigated eye movement patterns during object viewing (Schurgin et al., 2014; Peterson et al., 2016;

Received Dec. 29, 2024; revised Aug. 16, 2025; accepted Aug. 21, 2025.

Author Contributions: Z.Z., J.H., and G.O. designed research; Z.Z. performed research; Z.Z. and J.H. analyzed data; Z.Z. and G.O. wrote the first draft of the paper; Z.Z. and G.O. edited the paper; Z.Z. and G.O. wrote the paper.

We thank Tianlin Luo for comments on earlier versions of the manuscript. This work was supported by the National Science and Technology Innovation 2030 Major Program (Grant No. 2021ZD0203703, G.O.), Strategic Priority Research Program of the Chinese Academy of Sciences (XDB1010202, G.O.), Shanghai Municipal Science and Technology Major Project (Grant No. 2019SHZDZX02), National Natural Science Foundation of China (Grant No. 32371077, G.O.), and the National Natural Science Fund for Excellent Young Scientists Fund Program (overseas).

The authors declare no competing financial interests.

Z.Z.'s present address: CERVO Brain Research Center, Department of Psychiatry and Neuroscience, Faculty of Medicine, Université Laval, Quebec City, Canada.

Correspondence should be addressed to Gouki Okazawa at okazawa@ion.ac.cn.

This paper contains supplemental material available at: https://doi.org/10.1523/JNEUROSCI.2453-24.2025

https://doi.org/10.1523/JNEUROSCI.2453-24.2025

Hessels, 2020) and have suggested the importance of saccades (Heisz and Shore, 2008; Schurgin et al., 2014; Hessels, 2020). People fixate on the most informative region of an image during face recognition (Buchan et al., 2007; Peterson and Eckstein, 2012) and adopt different saccade patterns depending on task demands (Schurgin et al., 2014; Kanan et al., 2015; Hessels, 2020). Limiting saccades impairs object recognition and learning performance (Henderson et al., 2005; Hsiao and Cottrell, 2008; Hsiao and Liu, 2012). Integration of visual information across saccades (i.e., trans-saccadic integration) has been well demonstrated in studies using simplified stimuli such as Gabor orientations (Ganmor et al., 2015; Wolf and Schütz, 2015), color patches (Wijdenes et al., 2015), or 2D shapes (Demeyer et al., 2009, 2010; Herwig et al., 2015; Poth et al., 2015).

However, important questions remain to be addressed regarding how eye movements contribute to object vision. First, most existing work on trans-saccadic integration has used simple visual features (Ludwig et al., 2014; Ganmor et al., 2015; Wijdenes et al., 2015; Wolf and Schütz, 2015), whereas the recognition of complex object images (e.g., faces) may pose different challenges because saccades bring different complex features (e.g., eyes and mouth) to both the fovea and periphery. Existing models suggest that sensory evidence is integrated even in such cases (Renninger et al., 2004, 2007; Akbas and Eckstein, 2017), but an empirical test with natural images would be crucial to verify this. Second, the integration during saccades may require additional neural processes of combining visual inputs with efference copy from the oculomotor system (Melcher, 2007; Binda and Morrone, 2018); whether such processes play a major role in object recognition should also be tested empirically. Finally, trans-saccadic integration is usually studied under conditions in which participants are explicitly asked to make a saccade (Demeyer et al., 2009, 2010; Ganmor et al., 2015; Herwig et al., 2015; Poth et al., 2015; Wijdenes et al., 2015; Wolf and Schütz, 2015). It remains to be seen whether people integrate evidence even when they make free saccades during object viewing.

Here, we develop a framework for measuring and modeling object recognition behavior with saccades to address these questions. Our key innovation is to employ a theory of perceptual decision making (Shadlen and Kiani, 2013) and study object recognition behavior as a process of accumulating sensory evidence (Mack and Palmeri, 2011; Okazawa et al., 2018, 2021; Heidari-Gorji et al., 2021; Luo et al., 2025). Using behavioral paradigms with parametrically controlled object and face stimuli, we demonstrate that a simple model that accumulates evidence of local visual features across saccades is sufficient to quantitatively account for participants' behavior even when they view an image freely. Moreover, participants' behavioral performance was minimally affected by the efference copy. We conclude that object recognition with saccades can be approximated by a parsimonious model that accumulates available sensory evidence from dynamic retinal images without assuming complex interactions between the visual and oculomotor systems.

Methods

Participants and experimental setup

We recruited 22 participants (age 20–40, 4 males and 18 females, students or employees of the Chinese Academy of Sciences). Our participant sampling did not consider sex, as it was deemed unlikely to influence the outcomes of this study. All participants had normal or corrected-to-normal vision and were naive to the purpose of the experiment. Written informed consent was obtained from all participants. Six of them were dropped before the main data collection, either because of scheduling issues or poor eye-tracking quality. All experimental procedures were approved by the Institutional Review Board of the Center for Excellence in Brain Science and Intelligence Technology (Institute of Neuroscience), Chinese Academy of Sciences.

Nine participants performed the free saccade task (Fig. 1), while nine performed the guided saccade task (Fig. 5). Two of them performed both tasks. In each task, we assigned participants to perform categorization of face identity stimuli, face expression stimuli, or car stimuli (Fig. 1*A*; three participants each), but this study focused on the results consistently observed across these stimulus conditions. Sample sizes were chosen following the convention of studies using psychophysical reverse correlations and modeling of decision-making behavior (Okazawa et al., 2021). We had to set our sample sizes small because we sought to collect a large number of trials from each participant (∼3,500 trials per participant; 63,984 total trials in this study) after extensive practice sessions (∼2,000 training trials per participant prior to data collection) in order to obtain as much reliable behavioral data as possible from individual participants (Smith and Little, 2018). We performed a sensitivity analysis for a *t*-test using MATLAB function *sampsizepwr* function. With our sample size (9), our study had 80% power to detect a statistically significant effect ($p < 0.05$) assuming a true mean difference of 0.9 and a standard deviation of 1 (Lakens, 2022).

During the experiments, the participants sat in a height-adjustable chair in a semi-dark room. Their chins and foreheads were supported by a chin-rest mounted on a table, which was fixed at a specific position to ensure a stable viewing distance of 57 cm from a cathode-ray-tube monitor (17-inch IBM P77; 75 Hz refresh rate; 1,024 × 768 pixel screen resolution). The Psychophysics Toolbox (Brainard and Vision, 1997) and MATLAB (MathWorks) were used to control the stimulus presentation. The eye movements were monitored using a high-speed infrared camera (Eyelink 1,000 Plus; SR Research). The gaze position was recorded at 1 kHz.

Task designs

Face and object categorization task with free eye movement. To examine the contribution of saccades to object recognition, we designed three versions of the object categorization task: face identity, face expression, and car categorization (Fig. 1*A*). In each version of the task, participants classified an image into one of two categories while they were allowed to freely make saccades inside the image. The stimuli were chosen from a morph continuum created by interpolating two prototype images, and the participants were required to report which prototype the given stimulus looked similar to. The two prototypes were male and female faces in the identity task, happy and sad faces of the same individual in the expression task, and two types of cars in the car task (Fig. 1*A*). We used face stimuli because they allow easy definition of informative features (i.e., eyes and mouth) and previous studies have successfully explained face categorization during fixation conditions using an evidence accumulation model (Okazawa et al., 2018, 2021). We further designed a car categorization task to ensure that the results were not specific to face recognition.

Participants began each trial by fixating on a fixation point (0.3° diameter), which appeared randomly at one of six peripheral locations (11.5° away from the screen center for the face tasks, 8° for the car task; Fig. 1*B*). After a short delay (400–700 ms, truncated exponential distribution), a stimulus appeared at the center of the screen. The randomized fixation point locations were intended to minimize bias in the location that participants initially looked at in the image (Arizpe et al., 2012; Peterson and Eckstein, 2012). Participants then had to make a saccade in the stimulus within 500 ms of its appearance. The stimulus was kept ambiguous (halfway between the two prototypes on the morph continuum) and thus uninformative until the participants made a saccade. After fixation, the stimulus was replaced with a face image of the morph level chosen for the trial. The participants were then allowed to look at any place in the image, but if their fixation left the image, the trial was aborted. Participants reported the category of the stimulus by pressing one of two keyboard buttons whenever they were ready [reaction time (RT) task]. The stimulus was extinguished immediately when the button was pressed. If the participants did not make a decision within 5 s, the trial was aborted. In total, 1.75% of trials were aborted either due to fixation breaks during stimulus presentation or time out. Auditory feedback was provided for correct and incorrect decisions. If the stimulus was ambiguous (halfway between the two prototypes on the morph continuum), the correct feedback was provided in a random half of the trials. Following feedback, the next trial began after a 1 s inter-trial interval.

We created each stimulus set by continuously morphing two prototype images using a custom-made program (Okazawa et al., 2021). The prototypes for the expression task were obtained from the Nim Face set (Tottenham et al., 2009), and the prototypes for the identity task were obtained from the Tsinghua Facial Expression Database (Yang et al., 2020). The face images shown in the figures of this paper were from these databases and presented with permission. The prototypes for the car task were generated by authors using Midjourney (https://www.midjourney.com) with prompts such as "Clean and minimalist product photography of a white SUV with soft edges, highlighting its sleek and modern design." We then used Photoshop (Adobe) to edit the image parts that were difficult to morph, such as the steering wheels, to create a naturalistic morph continuum. Our program generated the morphed intermediates of the two prototypes by linearly interpolating the positions of manually defined anchor points on the prototypes and textures within the tessellated triangles defined by the anchor points. The linear weights for the two prototypes determined the morph level of an image (ranging from −100 to 100%, where the two extremes corresponded to the two prototypes). In each trial, we chose an average morph level from −96, −48, −24, −12, −6, 0, 6, 12, 24, 48, and 96%. For participants with higher performance, we also added −3 and 3% morph levels.

Our algorithm could morph local stimulus features independently. For face images, we manually circumscribed the regions containing the eyes and mouth and morphed only the inside of the regions (Fig. 1*A*). Similarly, for the car images, we manually defined the front and rear regions (Fig. 1*A*). The regions outside these features were maintained halfway between the two prototypes and thus remained uninformative. For face images, because the regions outside the eyes and mouth show a limited contribution to judgments (Schyns et al., 2002; Okazawa et al., 2021), this segmentation of informative and uninformative regions was unlikely to influence participants' behavior. For the car images, the front and rear parts were split at the midline (Fig. 1*A*), but the two prototypes were mostly different around the hood (bonnet) and the rear window. While the choice of prototype images would affect which parts become informative, we do not consider that this choice affected the main conclusions of this study. We could also independently adjust the full dynamic range of the morph line for each feature. This adjustment was made when we realized that the participant relied heavily on (kept looking at) one feature during training. We gradually decreased the range of the over-sampled features up to 50% while confirming that the participant's overall performance was maintained. Once the main data collection began, we did not make any adjustments.

To examine how each object feature contributed to the participants' decisions, we added random temporal fluctuations of the morph level to the individual features in each trial. The mean morph level was fixed within a trial and matched between the two features, but the morph level of a feature was randomly updated every 106.7 ms (eight monitor frames of the 75 Hz display) drawn from a Gaussian distribution with an SD of 20%. When the drawn number exceeded ±100%, it was resampled. The 106.7 ms fluctuation duration provided us with sufficiently precise measurements of the participants' temporal weighting in their ∼1 s decision time, while the duration was sufficiently long to ensure a subliminal transition of the morph levels from one image to another. Between two morphed face images, we interleaved a noise mask (phase randomization of the 0% morph face) with a smooth, half-cosine transition function during the eight monitor frame (Okazawa et al., 2021). More specifically, during each of the eight-frame cycle, a face image was first shown without a mask for one monitor frame (13.3 ms). Then, it gradually faded out over the next seven frames as a mask image faded in. For these frames, the mask and the face images were linearly combined, pixel by pixel, according to a half-cosine weighting function, so that in the last frame, the weight of the mask was 1 and the weight of the face image was 0. Immediately afterward, a new face frame with new morph levels was shown, followed by another cycle of masking, and so forth. This masking procedure minimized the chance that participants noticed fluctuations in morph levels during stimulus presentation. Movie 1 shows an example of our dynamic face stimuli.

In each trial, the stimulus fluctuations started only after the participants made a saccade into the stimulus. Prior to the saccade, participants fixated on a fixation point placed peripherally (8–11.5° away from the image). During this period, the stimulus remained uninformative (0% morph). Thus, the participants could start judging the stimulus category only after fixating on it. Exactly at the moment of saccade to a stimulus, the stimulus underwent a sudden change in the morph level, but no participants noticed this change. Since the stimulus fluctuations started only after fixation on the image, psychophysical kernels (Figs. 3*E* and 4) were aligned to the timing of the participant's fixation on the stimulus rather than the timing of the actual stimulus onset.

We determined the stimulus sizes in our tasks to characterize human object recognition behavior under natural conditions. We thus set the distance between the two informative features (eyes and mouth for face stimuli, and front and rear parts of car stimuli) to be five visual degrees apart. This is approximately the size that we experience when seeing objects and faces at natural distances (McKone, 2009). Under this constraint, the full stimulus size was ∼9.3° × 11° ($W \times H$) for the identity task, ∼9.4° × 13.6° for the expression task, and ∼7° × 2.8° for the car task. We expect that participants' saccade patterns would be greatly affected by stimulus size (von Wartburg et al., 2007; Otero-Millan et al., 2013). For example, participants would cease making saccades when the stimulus becomes too small. However, our primary goal was to study the effect of saccades under conditions with naturalistic object sizes, where saccades would spontaneously occur, and we believe that our conclusions hold as long as they are tested under such naturalistic ranges (see Discussion). We also confirmed that, according to the human contrast sensitivity function (CSF), the unfixated feature retained the information about the stimulus category while participants fixated on the other feature (Fig. S1*C–D*, Peterson and Eckstein, 2012; Or et al., 2015).

We recruited nine participants for the free saccade tasks, of whom three each were randomly assigned to perform each of the three categorization tasks (31,128 trials in total; 3,459 ± 106 trials per participant). Prior to the main data collection, the participants underwent extensive training (on average 2,000 trials) to ensure stable behavioral accuracy. During training, we informed the participants that the images contained

Movie 1. An example image sequence similar to those used in the experiments. The sequence consists of face images interleaved by masks. For each face image, the morph levels of two facial features (eyes and mouth) fluctuated around the mean morph level for the trial (Fig. 1*B* inset). The masks made these stimulus fluctuations subliminal. Note that the size and frame rate of the video do not accurately represent the stimuli used in the experiments.

multiple informative features for categorization and encouraged them to use multiple features to solve the task. However, we did not directly ask them to make eye movements or to look at particular parts of an image.

Guided saccade task. To examine whether oculomotor commands are necessary for feature integration, we designed a guided saccade task in which participants categorized objects with or without a saccade (Fig. 5*A*). In the saccade condition (Fig. 5*B*), we instructed participants to make a saccade during the stimulus presentation by moving the fixation point from one region to another. In the no-saccade condition (Fig. 5*E*), participants maintained fixation while the stimulus position suddenly moved, mimicking the change in retinal input resulting from a saccade. Similar to the free saccade task, we used face identity, expression, and car categorization conditions with the same stimuli and categorization rules.

In the saccade condition, participants initially viewed a fixation point that appeared at one of two locations near the center of the screen. The two locations corresponded to the location of the two informative features of a stimulus shown shortly afterward (eyes or mouth in the face tasks, front or rear in the car task). There was a variable delay between the participant's fixation onset and stimulus onset (400–700 ms, truncated exponential distribution). Immediately after the stimulus onset, the fixation point shifted to the location of the other informative feature, and participants had to make a saccade to this location in between 100 and 400 ms. After sufficient training, participants could consistently make a saccade with ∼200 ms latency (timeout: 5.1% of trials). After the saccade, the stimulus continued for another 213.4 ms; thereafter, it disappeared together with the fixation point, and the participants had to report their decision by pressing a keyboard button within 1 s (timeout: 1.5% of trials). These brief stimulus presentations replicated previous studies that investigated trans-saccadic integration (Ganmor et al., 2015) and were ideal for testing the temporal integration of evidence because behavioral performance is likely to saturate with longer stimulus duration (Kiani et al., 2008).

In separate blocks, we performed the no-saccade condition, which mimicked the change in retinal input during the saccade condition without asking participants to make an actual saccade (Fig. 5*E*). In this condition, the fixation point remained in the same place, but a stimulus briefly appeared with one feature centered at the fixation point, and after a brief blank period, it reappeared with the other feature centered at the fixation point. Thus, the condition approximated what the participants would have seen during the saccade condition. The duration of the first stimulus presentation and the blank were randomly sampled from the distribution of the saccade latency (170.0 ms ± 5.3 ms) and saccade duration (53.5 ms ± 1.3 ms) obtained in the main condition for each participant. To obtain these numbers, we first collected half of the data for the saccade condition. Subsequently, we collected the remaining half together with the no-saccade condition in the same sessions to ensure that each participant's training level was similar between the two conditions.

In the saccade and no-saccade blocks, we also included trials in which a stimulus was shown only before or after a saccade/stimulus jump (Fig. 5*B, E*). In these trials, we removed/displayed a stimulus contingent on the timing of the participant's saccade (either "Pre only" or "Post only" trials in Fig. 5*B, E*, right). These trials were randomly interleaved with the main condition in which a stimulus was shown in both periods ("Both" trials; Fig. 5*B, E*, left). This allowed us to test whether participants' performance improved when a stimulus was present both before and after a saccade, indicating the integration of evidence across saccades. While the full stimulus duration in "Both" trials was longer than that in "Pre only" or "Post only" trials, participants could only make use of this longer presentation time by integrating evidence across saccades. Thus, this form of comparison has been often used for the test of trans-saccadic integration (Ganmor et al., 2015; Wolf and Schütz, 2015).

The stimulus morph levels fluctuated in this task, similar to those in the free saccade task. Because the morph levels were updated every 106.7 ms (eight monitor frames, during which a stimulus image made a smooth transition to a noise mask; see above), approximately two cycles of fluctuations occurred before a saccade, as the saccade latency was, on average, ∼170 ms. After the saccade, we reset the fluctuation cycle such that one cycle starts immediately after the saccade landing, ensuring consistency in the pattern of stimulus-mask cycles before and after the saccade. The post-saccade stimulus fluctuations continued for two cycles (213.4 ms) before the stimulus was terminated. As in the free saccade task, fluctuations occurred independently for the two informative features, whereas the average morph level was the same for the two features and was constant during the trial. The average morph levels were chosen from −96, −48, −24, −12, −6, 0, 6, 12, 24, 48, and 96%. For participants with higher performance, we added −3 and 3% morph levels.

Nine participants performed this guided saccade task, of whom three each was randomly assigned to the facial identity, expression, or car categorization conditions (32,856 trials in total; 3, 651 ± 68 trials per participant). Prior to the main data collection, the participants underwent extensive training (on average 2,000 trials) to ensure stable saccade latency and behavioral accuracy.

Data analysis

Detection of cross-feature saccades and quantification of saccade patterns. Saccades were detected from the 1 kHz eye-tracking data using the Eyelink 1,000 Plus' default saccade detection parameters with additional criteria (Larsson et al., 2013) to ensure accuracy. We first applied a small smoothing (a Gaussian filter with 3 ms standard deviation) to the tracking data to remove high-frequency noise and then detected the timings of eye traces with a velocity exceeding 30°/s for 4 ms and acceleration exceeding 8,000°/s^2 for 2 ms. These timings were considered potential saccade onsets. We then estimated the end time of these potential saccades by looking for the time when the velocity fell below 20°/s for 2 ms. Finally, we classified them as saccades if their duration was longer than 6 ms and they were observed at least 20 ms after the last saccade (Larsson et al., 2013). Through manual inspection, we confirmed that these parameters accurately detected saccades. In rare cases, noise in the eye traces led to the false detection of saccades, which were removed during manual inspection.

Our key objective was to examine how large saccades spanning multiple features in an image contribute to the integration of evidence across features. We thus focused on these cross-feature saccades in our main analyses. We considered a saccade a cross-feature if it satisfied the following criteria. (1) The saccade start point was inside or near the region of one feature (<1.5° for the face tasks and <0.5° for the car task) and its end point was inside or near the region of the other feature. These numbers were chosen based on manual inspection of eye movement patterns. The region for each feature was manually circumscribed (Fig. S1*B*), and the average gaze positions were calculated 50 ms before and after the saccade to determine whether they were near or inside the regions. The distance to a feature was defined as the minimum Euclidean distance between the gaze position and any point on the manually drawn contour line of the feature. (2) The amplitude of the saccade was greater than 2°. This second criterion ensured that the saccade was not small enough to occur right around the boundary of the two features but was large enough to cause a considerable change in the retinal input. The number was determined through the manual inspection of saccade patterns and the distribution of saccade amplitudes (Fig. 2*E*). (3) The saccade started at least 50 ms after the participant fixated on the stimulus and at least 50 ms before the participant's response. This condition ensured that it occurred during decision formation.

To examine how stimulus fluctuations influenced the participants' decisions depending on their gaze positions, we defined the fixated and unfixated features in each cycle of stimulus fluctuations in several analyses (Figs. 3, 4, and S3*C–D*). We first averaged eye positions within each of the 106.7 ms fluctuation cycles and then checked whether the averaged position was inside or near (<1.5° for the face tasks and <0.5° for the car task) the region of a feature circumscribed manually. If so, the feature was defined as fixated, whereas the other feature was defined as unfixated. This definition follows the criteria used to define cross-feature saccades above. If the average eye position was outside the range of both features, the fixated feature was not defined, and the corresponding cycle of stimulus fluctuations was excluded from the analysis. A fluctuation cycle was also excluded if there was a cross-feature saccade within this period.

During quantitative analyses and model fitting (Figs. 3*G*, 4, S3*E*, S4, and S5), we also calculated the distance between the participant's gaze position and each object feature each time in a trial. This distance followed the definition described above as the minimum Euclidean distance between the gaze position and any point on the contour line circumscribing the region of the informative feature (Fig. S1*B*). If the gaze position was within the circumscribed region, the distance was set to zero. Thus, this definition is agnostic of where the exact center of the informative features is.

Psychometric and chronometric functions. To quantify behavioral performance in the free saccade task, we fitted the following logistic function to the choice data of each participant for each stimulus condition (Fig. 1*C*, top):

$$\mathrm{logit}[P(\mathrm{choice}\ 2)] = \alpha_0 + \alpha_1 s, \tag{1}$$

where $\mathrm{logit}(p) = \log(p/1-p)$, s is the nominal stimulus strength of a trial ranging from −1 (−100% morph level) to +1 (+100% morph level), and α_i are regression coefficients. α_0 quantifies the choice bias and α_1 quantifies the slope of the psychometric function.

The relationship between stimulus strength and the participants' mean RTs was assessed using a hyperbolic tangent function (Fig. 1*C*, bottom):

$$T = \frac{\beta_0}{s}\tanh(\beta_1 s) + \beta_2, \tag{2}$$

where T is the mean RTs in seconds and β_i indicates the model parameters. β_0 and β_1 determine the stimulus-dependent changes in RTs, whereas β_2 quantifies the portion of RTs independent of the stimulus strength.

Behavioral performance in the guided saccade task (Fig. 5) was quantified using the following logistic function:

$$\mathrm{logit}[P(\mathrm{correct})] = \alpha_1 s + \alpha_2(s \cdot I), \tag{3}$$

where an indicator variable, I, was used to quantify the difference in the slope of the psychometric functions between two conditions. For example, when we compared behavioral performance between the "Both" and "Pre only/Post only" conditions in the guided saccade task (Fig. 5*B, E*), I was set to 1 in the former condition and 0 in the latter condition. We fitted the above function to individual participants' data and examined the performance difference between conditions by testing if α_2 was significantly different from 0 using t-test across participants. The function did not have a bias term because it was fit to the probability of correct, which is 0.5 at zero stimulus strength by definition.

Joint psychometric functions of features across saccades. To directly test whether participants used fixated features before and after cross-feature saccades, we plotted their choice performance as a function of the morph levels of these features (Fig. 3*A–C*). For example, if a participant first fixated on the eye region and then made a saccade to the mouth region before committing to a choice, we computed the average morph fluctuations in the eye region before the saccade as well as the average in the mouth region after the saccade (Fig. 3*A*). We then projected each trial in a 2D space defined by the morph levels before and after a saccade. In this space, we computed the probability of choice of the trials in a Gaussian window with a standard deviation of 5% and visualized the probability of choice by drawing iso-probability contours at 10% intervals (Fig. 3*B*). Similarly, if a participant made two cross-feature saccades, we plotted their performance as a function of the first, second, and third fixation features (Fig. 3*C*). Since the participants made fewer than three cross-feature saccades in most trials (Fig. 2*F*), we did not consider trials with more saccades.

To test the significance of the influence of each fixated feature on participants' choices, we performed the following logistic regression for trials with one cross-feature saccade:

$$\mathrm{logit}[P(\mathrm{choice}\ 2)] = w_0 + w_1 s_1 + w_2 s_2 + w_{1,2} s_1 s_2, \tag{4}$$

where s_1 and s_2 correspond to the morph levels of the features fixated the first and second times in a trial, respectively. w_0 quantifies choice bias, w_1 and w_2 are linear coefficients, whereas $w_{1,2}$ is a coefficient of the multiplicative interaction term. For trials with two cross-feature saccades, we used:

$$\begin{aligned}\mathrm{logit}[P(\mathrm{choice}\ 2)] = {} & w_0 + w_1 s_1 + w_2 s_2 + w_3 s_3 \\ & + w_{1,2} s_1 s_2 + w_{1,3} s_1 s_3 + w_{2,3} s_2 s_3 \\ & + w_{1,2,3} s_1 s_2 s_3,\end{aligned} \tag{5}$$

where s_1, s_2, and s_3 corresponded to the morph levels of the features fixated first, second, and third times in a trial. We performed regression using all trials within individual participants and performed a two-tailed t-test across participants to test the significance of the contribution of each feature.

Psychophysical reverse correlation. To test whether features at different points in time across saccades affected the participants' choices, we performed psychophysical reverse correlations (Ahumada, 1996; Okazawa et al., 2018; Figs. 2*H*, 3, 4, 5, and S3–S6). Psychophysical kernels [$K_f(t)$] were calculated as the difference in the average fluctuations of morph levels conditional on the participant's choices:

$$K_f(t) = E[s_f(t) \mid \mathrm{choice}\ 1] - E[s_f(t) \mid \mathrm{choice}\ 2], \tag{6}$$

where $s_f(t)$ represents the morph level of feature f at time t. This reverse correlation analysis was originally proposed to reveal perceptual templates of linear observers seeing images containing pixel-level white noise (Murray, 2011). By contrast, our reverse correlation was performed using random fluctuations of the morph levels of object images, which were highly non-linear with respect to pixel-level inputs. However, we have previously confirmed that the morph levels created using our method almost linearly mapped onto estimated subjective evidence (Okazawa et al., 2018); thus, as long as we are concerned with the relationship between the morph levels and participants' performances, we consider that the analysis largely satisfies the original assumption of linear observers. This analysis only used trials with low stimulus strength (nominal morph level, 0–12%). For the non-zero strength trials, the mean strength was subtracted from the fluctuations, while the residuals were used for the reverse correlation. In Figure 2*H*, we averaged the kernels for each feature over time when the participants viewed the feature. For the guided saccade tasks (Fig. 5*D, G*), we averaged the two cycles of stimulus fluctuations both before and after the saccade to calculate the kernels (but see Fig. S6*G, H* for unaveraged kernels).

When plotting the time course of psychophysical kernels (Figs. 3*E*, 4, and S3–S5), we sorted individual stimulus fluctuations depending on which feature the participant fixated on during that cycle of fluctuations, and then generated kernels for the fixated and unfixated features. When a participant's gaze position could not be classified into a feature or a cross-feature saccade occurred during a cycle of fluctuation, it was excluded from the analysis. Since the stimulus fluctuation started when participants fixated on an image (see above), the kernels aligned to stimulus onset started from the moment of fixation and were calculated up to the first cross-feature saccade or up to 1 s in the trials without saccades. The kernels aligned to the participants' responses were calculated using stimulus fluctuations after the last cross-feature saccade or using the fluctuations for 1 s from the response when there was no saccade in a trial. For the kernels aligned to saccades, we used five stimulus cycles before and after the saccade onset. Figure 4*D* shows an example trial with only one saccade, but if there were more than one cross-feature saccade, all were used when computing the kernels. For the kernels shown in Figures 3*E* and 4, we averaged the kernels across all participants. The kernels of the individual participants can be found in Figure S4. Three-point boxcar smoothing was applied to the temporal kernels for denoising. However, we did not perform smoothing when evaluating the fitting quality (R^2).

To further quantify how gaze position modulated the contribution of local features to decisions, we realigned the same stimulus fluctuations according to the distance between the participants' gaze position and each feature location (Fig. 3*F, G*). As explained in the "Detection of cross-feature saccades and quantification of saccade patterns" section above, we averaged the gaze positions during each cycle of stimulus fluctuation and computed its distance from any point on the counter line circumscribing the region of each feature (Fig. S1*B*). If the gaze position was within the circumscribed region, the distance was set to zero. If a saccade occurred during one cycle of stimulus fluctuation, the fluctuation was excluded from the analysis. We then sorted the fluctuations according to the calculated distance and generated psychophysical kernels at each distance d as:

$$K_f(d) = E[s_f(d) \mid \text{choice } 1] - E[s_f(d) \mid \text{choice } 2], \tag{7}$$

where $s_f(d)$ is the morph fluctuation of feature f at distance d. This was calculated using stimulus cycles concatenated across the trials with low stimulus strength (nominal morph level, 0–12%) within individual participants.

Calculation of saccade frequency and probability. We calculated the frequency of participants' saccades for each stimulus strength to test any potential dependence on stimulus difficulty (Fig. 6*B*). Frequencies could not be simply estimated by dividing saccade counts by trial duration (i.e., RTs) because saccades tended to be periodic. Suppose that saccades occur every 400 ms regardless of stimulus strength. If the average RT was 500ms for one stimulus and 700 ms for another, saccade counts are expected to be one per trial, and thus, saccade counts per time tend to be underestimated for stimuli with longer RTs. Therefore, we had to match the RT distributions across stimulus strengths for a proper comparison. We generated RT histograms with 100 ms intervals and performed histogram matching by randomly subsampling trials from each stimulus strength. We then counted the total number of saccades in these subsampled trials and divided this number by the total stimulus duration across the trials. The frequencies were calculated in this manner for individual participants and then averaged (Fig. 6*B*).

We further took an alternative approach to estimate saccade frequency without matching RT distributions (Fig. 6*C*). In this method, we calculated the number of saccades that occurred at each time point (100 ms bins) and divided it by the number of trials whose RT was longer than that time point. This saccade probability is not affected by the interaction of RTs and saccade timing outlined above because this metric does not depend on the duration of trials after each time point to compute the probability. However, the results can be noisy, particularly for later time points, because fewer trials contribute to the calculation. We therefore classified trials into two groups (easy: >20% morph level, difficult: <20%) to perform this analysis.

Model fit and evaluation

To quantitatively examine whether the participants' choice behavior during the free saccade task could be explained by the integration of sensory evidence over saccades, we constructed a simple extension of evidence accumulation models widely used to explain behavioral data in a variety of perceptual decision-making tasks (Shadlen and Kiani, 2013; Okazawa et al., 2021). We also developed multiple alternative models to confirm that no alternative mechanisms account for the behavioral data. In what follows, we first describe the expression and fitting procedure for the main model and then extend them to the alternative models.

Main model. Our main model is an extension of the drift-diffusion model, which considers multiple informative image features and their distances from the participants' gaze positions (Fig. 4*A*). The model was extended from that of a previous study that was demonstrated to accurately explain human face categorization behavior measured under stable fixation conditions (Okazawa et al., 2021). This previous model first linearly integrates the fluctuations of the local features [$s_i(t)$ for feature i at time t]:

$$\mu(t) = \sum_{i=1}^{N} k_i \cdot s_i(t), \tag{8}$$

where μ is momentary evidence for the model, N is the number of features, and k_i is the sensitivity parameter for each feature i. Momentary evidence is then accumulated over time to form the decision variable (v) at each time t as:

$$v(t) = \int_0^t \mu(\tau) + \eta(\tau)\, d\tau, \tag{9}$$

where $\eta(\tau)$ represents internal (neural) noise in the sensory, inference, or integration processes, assumed to follow a Gaussian distribution with mean 0 and SD $\sigma(t)$. When the decision variable [$v(t)$] reaches an upper or lower bound (+B or −B), the model commits to a decision associated with the bound. RT was defined as the time required to reach a bound plus a non-decision time including sensory and motor delays. The non-decision time was drawn from a Gaussian distribution with a mean of T_0 and an SD of σ_{T_0}.

Our present model extends the above formalism by incorporating participants' gaze positions as a factor that influences the informativeness of local features (Krajbich et al., 2010; Tavares et al., 2017; Yang and Krajbich, 2023). We added one free parameter (λ) that quantified the degree to which the informativeness of each feature decays as a function of the distance from the gaze position (i.e., visual eccentricity). The decay was expressed as an exponential function based on the previous studies that successfully modeled visual acuity as a function of visual eccentricity (Peli et al., 1991; Peterson and Eckstein, 2012). We also tested a linear decay function and confirmed that it yielded similar results (Fig. S5*B*). In the exponential model, Equation 8 was modified as:

$$\mu(t) = \sum_{i=1}^{N} k_i \cdot e^{-\lambda d_i(t)} \cdot s_i(t), \tag{10}$$

where $d_i(t)$ is the Euclidean distance (in units of visual angle) between the gaze position and object feature i at time t. As mentioned above, the distance was defined as the minimum length between the gaze position and any point on the counter line manually drawn to circumscribe each feature (Fig. S1*B*). If the gaze position was inside the circumscribed region, the distance was set to zero. During the saccades, both momentary evidence and diffusion noise were set to zero to simulate the absence of visual input.

Once the momentary and accumulated evidence is defined as above, we can numerically derive the probability that the decision variable has value v at time t by solving the Fokker–Planck equation:

$$\frac{\delta p(v, t)}{\delta t} = \left[-\frac{\delta}{\delta v}\mu(t) + 0.5\frac{\delta^2}{\delta v^2}\sigma^2(t) \right] p(v, t), \tag{11}$$

where $p(v, t)$ denotes the probability density. The accumulation process started from zero evidence and continued until the decision variable reached one of the two bounds (±B), indicating two choices. Thus, the partial differential equation above has the following initial and boundary conditions:

$$\begin{aligned} p(v, 0) &= \delta(v), \\ p(\pm B, t) &= 0, \end{aligned} \tag{12}$$

where $\delta(v)$ denotes the Dirac delta function. The diffusion noise [$\sigma(t)$] was set to 1, and the bound and drift rate were defined in a unit of diffusion noise. The RT distribution for each choice was obtained by convolving the distribution of bound crossing times with the distribution of non-decision time (a Gaussian distribution with a mean of T_0 and an SD of σ_{T_0}). The SD, σ_{T_0}, was always set to one-third of T_0 to reduce the number of free parameters.

Overall, our main model had five degrees of freedom: decision bound height (B), sensitivity parameters for two features (k_1, k_2), mean non-decision time T_0, and the decay rate of visual sensitivity λ. We fit the model parameters by maximizing the likelihood of the joint distribution of the observed choices and RT distributions of individual participants in each stimulus condition (Okazawa et al., 2018). Given a set of parameters, the stimulus fluctuations and participant's gaze points in each trial were used to calculate the RT distributions of the two choices according to the model formulation above. These distributions were used to calculate the log-likelihood of the observed choice and RT for individual trials. These log-likelihoods were summed across the trials to calculate the likelihood function for the dataset. We used a simplex search method (*fminsearch* in Matlab) to determine the parameter set that maximized the summed likelihood. To avoid local maxima, we repeated the fitting process from multiple initial parameter sets and selected the set that converged to the largest likelihood as the final result. Because maximum likelihood estimation is sensitive to outliers, we excluded trials with reaction times greater than three SDs from the mean for each stimulus strength during model fitting. Fitting was performed for each participant and included the trials with all stimulus strengths. The fitting performance was quantified using the coefficient of determination (R^2) for the joint distributions of choices and RTs. For each morph level, we generated the RT distribution for each choice (bin size, 100 ms) and computed the R^2 between the data and model outputs after concatenating the bins for all morph levels and choices. The fitting curves shown in Figure 4*B–D* are the averages across participants.

Alternative models. To examine whether different mechanisms accounted for the behavioral data, we developed multiple alternative models. They included the "gaze-independent" model, which had constant sensitivity to each local feature regardless of the participant's gaze position, the "evidence-reset" model, which resets the accumulated evidence every time the participant makes a cross-feature saccade, and the "independent-accumulator" model, which does not integrate evidence across saccades but accumulates the evidence from two features independently. These models were fitted to the behavioral data using the aforementioned procedure.

The gaze-independent model was designed to test whether the information of participants' gaze positions was necessary to account for their behavior. In our main model, the sensitivity to each feature was modulated according to the distance between the gaze position and the feature (Eq. 10), whereas the gaze-independent model removed this term and computed momentary evidence assuming that the sensitivity to each feature is constant regardless of gaze position (thus using Eq. 8) to determine the drift rate. The other components of the model were the same as those used in our main model. This model has one fewer parameter (four) than our main model.

The evidence-reset model was created to test the possibility that, when sensory evidence from one feature was insufficient to form a decision, people would make a saccade to the other feature and restart their decision-making process. To simulate this, the model resets the accumulated evidence to zero after a cross-feature saccade. Thus, the choices and RTs of the model were based solely on the feature fixated on after the last cross-feature saccade in a trial. To fit this model, we extracted the timing of the last cross-feature saccade and the feature fixated afterward from each trial and simulated the bounded evidence accumulation using them to predict the choice and RT of that trial. The model was equivalent to our main model if a trial did not contain a cross-feature saccade. The model becomes unrealistic when the last cross-feature saccade was too close to the RT of a trial; we thus did not count saccades that occurred within the mean non-decision time (i.e., T_0) from the RT. Note that we only excluded these saccades close to RTs but did not exclude any trials, thus the comparison with the main model was performed using all trials. Besides this resetting mechanism, all components of our main model, including the sensitivity of each feature and the dependency of sensitivity on the gaze positions, were preserved in this model. The number of parameters in this model is the same as that in our main model.

The independent-accumulator model tested the possibility that the evidence was not integrated across saccades. Instead, it independently accumulated evidence for the two features and committed to a choice based on evidence from one of them that reached a bound earlier. The evidence for a feature was accumulated when the feature was fixated, whereas the accumulated evidence stayed frozen when the feature was not fixated. To implement this, we classified each time of each trial into one of the two features based on the participants' eye positions and then simulated the evidence accumulation process for each feature. The model had five parameters.

To test fit performance, we computed the difference in the Bayesian information criterion (ΔBIC) between the main model and each of the alternative models (Figs. 4 and S5; positive values indicate poorer fits of the alternative models). For each model, we summed the log-likelihood of all trials and averaged the sum across participants to derive the BIC. When fitting and evaluating the evidence-reset and independent-accumulator models, the likelihood involved both the probability that they reached a decision at the observed RT after the last saccade and the probability that they did not reach a decision before the last saccade. This is consistent with the definition of the likelihood for the main model, which was based on the probability of reaching a decision at the observed RT without reaching a decision bound at any other point from stimulus onset.

Generation of model psychophysical kernels and RT distributions. The models above were fit to the choices and RTs, but the model formulation does not prescribe its psychophysical kernel. Therefore, we relied on simulations to estimate the model kernels. We created 10^5 simulated trials with stimulus strengths ranging from 0 to 12% using the same stimulus distributions as in the main task (i.e., Gaussian distribution with 20% SD). The model responses for these trials were simulated using the same parameters fitted for each participant. We then used the simulated choices and RTs to calculate model psychophysical kernels, following the same procedure used for the human data (Fig. 4*D, F, H, J*, S4, and S5). Thus, the model kernels were not directly fitted to the participants' kernels but were generated from an independent set of stimulus fluctuations, making the comparison of data and models informative. Similarly, the RT distributions of the models (Fig. 4*C*) were generated using simulations with an independent set of morph fluctuations to ensure an accurate comparison of the data and models.

To generate the model predictions, it was necessary to simulate eye movement data because the models needed to compute the distance between gaze positions and feature locations to calculate the strength of momentary evidence (Eq. 10). To generate realistic eye data, we used the participants' actual eye movement data from randomly selected trials; however, when the duration was shorter than the duration required for model simulations, we extended the eye data in two different ways. For the main model, we stitched a chunk of the eye trace obtained from another trial such that it could be smoothly connected to the end of the eye data. To do so, we looked for a chunk that started at a position <0.3° distance from the endpoint of the eye data. We repeated this stitching procedure until the eye trace reached the desired length. For the evidence-reset and independent-accumulator models, we simply extended the last eye position to become the desired length of the eye trace because the model had to assume that no-saccade occurred during this extended period.

Ideal observer analysis for the guided saccade task

In the guided saccade task (Fig. 5), we examined whether the participants' performance could be accounted for by the optimal integration of the evidence before and after a saccade (Fig. S6). The task had "Pre only" and "Post only" conditions where a stimulus was only shown before or after a saccade, and "Both" condition where a stimulus was shown in both epochs (Fig. 5*B*). To build an ideal observer model, we first estimated the precision of the participant's judgment of the stimulus in the "Pre only," "Post only," and "Both" conditions ($\hat{M}_{\text{pre}}$, $\hat{M}_{\text{post}}$, and $\hat{M}_{\text{both}}$) assuming Gaussian judgment noise:

$$\begin{aligned}\hat{M}_{\text{pre}} &= M + N(\mu_{\text{pre}}, \sigma^2_{\text{pre}}),\\ \hat{M}_{\text{post}} &= M + N(\mu_{\text{post}}, \sigma^2_{\text{post}}),\\ \hat{M}_{\text{both}} &= M + N(\mu_{\text{both}}, \sigma^2_{\text{both}}),\end{aligned} \tag{13}$$

where M is the actual stimulus value (morph level), μ_{pre}, μ_{post}, and μ_{both} are biases in the judgment, σ^2_{pre}, σ^2_{post}, and σ^2_{both} are variances in the judgment, and N represents the Gaussian distribution. These biases and variances could be estimated by fitting a cumulative Gaussian distribution to the psychometric function of the "Pre only," "Post only," and "Both" conditions, respectively.

An ideal observer model that optimally combines evidence from "Pre" and "Post" epochs makes the following maximum a posteriori estimate (Ernst and Banks, 2002; Oruç et al., 2003; Ganmor et al., 2015):

$$\begin{aligned} \hat{M}_{\text{ideal}} &= N(\mu_{\text{ideal}}, \sigma^2_{\text{ideal}}), \\ \mu_{\text{ideal}} &= w_{\text{pre}}(M + \mu_{\text{pre}}) + w_{\text{post}}(M + \mu_{\text{post}}), \\ \sigma^2_{\text{ideal}} &= \left(\frac{1}{\sigma^2_{\text{pre}}} + \frac{1}{\sigma^2_{\text{post}}}\right)^{-1}, \end{aligned} \tag{14}$$

where:

$$\begin{aligned} w_{\text{pre}} &= \frac{1}{\sigma^2_{\text{pre}}}\left(\frac{1}{\sigma^2_{\text{pre}}} + \frac{1}{\sigma^2_{\text{post}}}\right)^{-1}, \\ w_{\text{post}} &= \frac{1}{\sigma^2_{\text{post}}}\left(\frac{1}{\sigma^2_{\text{pre}}} + \frac{1}{\sigma^2_{\text{post}}}\right)^{-1}. \end{aligned} \tag{15}$$

The obtained σ^2_{ideal} corresponds to the variance of the judgment by the ideal observer. We compared this value against σ^2_{both} calculated above to test the optimality of the integration (Fig. S6*A, B*). We further performed the same analysis for the no-saccade condition (Fig. S6*C, D*).

Modeling feature discriminability based on human contrast sensitivity

One potential reason that we did not observe the effects of saccades on the integration of the to-be-fixated feature (Fig. 5) is that the contrast sensitivity in the periphery may have been too low to perceive the feature. To exclude this possibility, we applied an ideal observer model (Peterson and Eckstein, 2012; Or et al., 2015) based on the spatial variation in human contrast sensitivity to our stimuli. We assumed that the ideal observer foveated at one of the two informative features of the images in each task and tested whether the unfixated feature was still informative by comparing the model's psychometric function between the fixated and unfixated features (Fig. S1*C, D*).

We first convolved the stimuli in the spatial domain using a previously reported CSF:

$$\text{CSF}(f, r) = c_0 f^{a_0} \exp(-b_0 f - d_0 r^{n_0} f), \tag{16}$$

where *f* is the spatial frequency (in cycles per degree) and *r* is the distance from the fixation position (in degrees). The constants a_0, b_0, and c_0 describe the peak contrast and the shape of the CSF, d_0 and n_0 define how fast the contrast declines with the distance from the fixation. All the values of these constants were chosen from previous studies [$a_0 = 1.2$, $b_0 = 0.3$, $c_0 = 0.625$, $d_0(\frac{\pi}{2}) = 0.0001$, $d_0(-\frac{\pi}{2}) = 0.00024$, $d_0(0) = 0.00005$ and $n_0 = 5$; Peterson and Eckstein, 2012; Or et al., 2015]. For a given fixation point, the input image (the stimuli blended with a phase-scrambled noise mask used in the experiment) was divided into small spatial bins without overlap in polar coordinates, and each bin was assigned a CSF according to the distance *r* (step size 0.25°) and the angle (divided into three regions: horizontal 0, up $\frac{\pi}{2}$ and down $-\frac{\pi}{2}$, which corresponded to the three d_0 parameters above). We then convolved the whole image with the CSF defined by this spatial bin and extracted the corresponding area as the filtered result of each bin (Fig. S1*C*).

Using the filtered images, we performed an ideal observer analysis and obtained the model performances. The images were compared with two templates (−100 and 100% morph images in each task) to calculate the likelihoods of choosing one of the two categories:

$$l_{f,k} = \exp\left(-\frac{(g - s_{f,k})^T (g - s_{f,k})}{2\sigma^2}\right), \tag{17}$$

where *g* is the filtered input image, the $s_{f,k}$ is the template (−100 or 100% morph image) filtered in the same way, and σ is a free parameter corresponding to the noise level for decisions. We set the fixation point to be the center of one of the two features (eye and mouth for the face tasks, and front and back for the object task) and used the pixels within the contour of either fixated or unfixated feature (the contours shown in Fig. S1*B*). For each morph level, we calculated the likelihoods with 1,000 different phase-scrambled noise masks used in the experiment and derived the model's correct rate by averaging them (Fig. S1*D*). The noise level, σ, was set such that the model yielded similar performances with human participants for the fixated feature (expression task: 2.5, identity and car tasks: 1.5). This choice does not affect our conclusion as our goal was to compare the relative model performances between the fixated and unfixated features.

Results

Saccadic sampling of local features during object recognition

We designed an object categorization task in which participants freely made eye movements inside a stimulus to report its category. We assigned participants to perform either facial identity categorization, facial expression categorization, or car categorization (Fig. 1*A*). A stimulus in each trial was sampled from a morph continuum of two prototype images (e.g., two facial identities in the identity categorization, which corresponded to −100 and 100% morph; Fig. 1*A* and S1*A*), and the participants were asked to report which prototype category the stimulus was closer to. Before stimulus onset, participants were required to look at a fixation point whose position was randomly selected from six possible peripheral locations (8–11.5° away from the monitor center; Fig. 1*B*). Following stimulus onset, the participants had to immediately make a saccade to the stimulus and could then look at any part inside it. They subsequently reported their decisions by pressing a button as soon as they were ready (RT task; Fig. 1*B*). RTs were defined as the time between the fixation on the stimulus and the button press. We recruited nine participants and assigned three of them to each categorization task (Fig. 1*A*). These sample numbers were determined based on our needs for collecting a large number of trials (∼3,500 trials per participant) to perform psychophysical reverse correlation and model fits (Smith and Little, 2018). Since we had only three participants for each task, the present study focused on the behavioral patterns common to all three tasks. Furthermore, we provided individual participants' results in supplementary figures, ensuring that we report the results consistently observed across participants.

To assess how participants sampled the local features during decision-making, we defined two informative features for each stimulus set (eyes and mouth in the face sets; front and rear parts in the car set; Fig. 1*A*) and added random fluctuations to their morph levels every 106.7 ms (eight monitor frames) during stimulus presentation (Fig. 1*B*, inset; fluctuation SD, 20% morph). The mean morph levels of the two features were maintained identical and constant within each trial. The morph level outside of the two features was always set to 0% and remained uninformative. This design allowed us to use psychophysical reverse correlation (Ahumada, 1996; Okazawa et al., 2018, 2021) and test how each feature at each moment during stimulus viewing influenced the participants' decisions. Two features were sufficiently separated for saccades to be made between them (inter-feature distance, 5°), and at the same time, the image sizes were within the range of naturalistic viewing conditions (McKone, 2009).

We first confirmed that the participants showed stereotypical choice accuracy and RTs under the three stimulus conditions (Fig. 1*C*). Hereafter, we focus on the results qualitatively consistent across the three conditions and present the results either

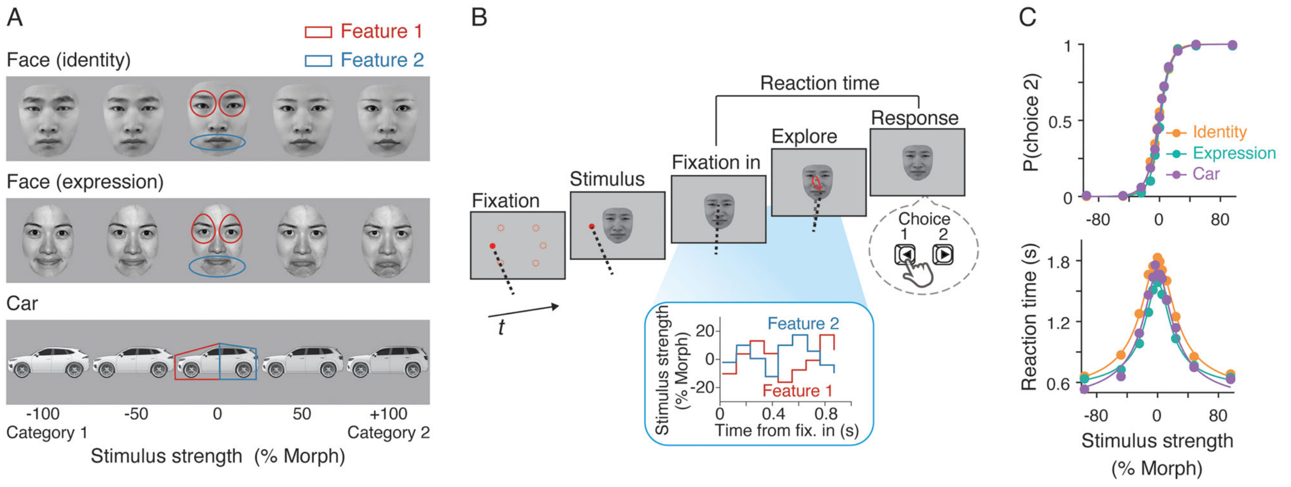

Figure 1. Object categorization task with free saccades. ***A***, Participants were assigned to perform either facial identity, expression, or car categorization ($n = 9$). In each trial, participants viewed a stimulus chosen from a morph continuum ranging from −100 to 100% morph levels between two prototype images. Participants were then required to report which prototype the stimulus was closer to. We defined two informative features (red and blue contour lines) and morphed images inside these two regions. The face images were from the Nim Face set (Tottenham et al., 2009) and the Tsinghua Facial Expression Database (Yang et al., 2020) and presented with permission. The same face images were used in the subsequent figures. ***B***, Participants began each trial by fixating on a red dot that appeared at one of six possible locations on the screen. Shortly afterward, a stimulus appeared, and participants were required to make a saccade to the stimulus. Thereafter, participants could view any part of the image until they reported the stimulus category by pressing one of two buttons as soon as they were ready (reaction time task). During stimulus viewing, the morph levels of the two informative features fluctuated randomly every 106.7 ms, while their mean was maintained constant within a trial (SD: 20%). This fluctuation allowed us to examine the weighting of each feature during decision making. An example movie of the dynamic face stimuli can be found in Movie 1. ***C***, Participants showed stereotypical psychometric and chronometric curves as a function of the mean morph levels. Lines represent logistic and hyperbolic tangent fits for psychometric and chronometric functions, respectively (Eqs. 1 and 2).

individually or averaged across conditions depending on the purpose of visualization (where averaged results are presented, individual results are shown in the supplementary figures). In all conditions, choice accuracy was monotonically modulated by the morph level [logistic regression slope $\alpha_1 = 13.7 \pm 1.2$, mean ± SEM across participants, Eq. 1; $t_{(8)} = 11$, $p = 4.0 \times 10^{-6}$, two-tailed t-test]. RTs were systematically longer for lower morph levels [$\beta_1 = 7.14 \pm 0.52$ fitted to a hyperbolic tangent function, Eq. 2; $t_{(8)} = 14$, $p = 7.5 \times 10^{-7}$, two-tailed t-test]. These patterns are consistent with many previous behavioral results of perceptual tasks (Shadlen and Kiani, 2013) and thus suggest that decision-making models similar to those previously proposed, such as bounded evidence accumulation (Okazawa et al., 2021; Luo et al., 2025), can explain our results.

While performing the task, participants often made multiple saccades between informative features (Fig. 2). Immediately following stimulus onset, their fixations tended to land just below the eyes in the face categorization tasks (Fig. 2*A*, left and middle), which is consistent with previous studies that investigated fixation patterns during face recognition (Peterson and Eckstein, 2012). For expression categorization, landing positions appeared slightly closer to the nose (Fig. 2*A*, middle), which also agrees with previous reports (Peterson and Eckstein, 2012). For car categorization, the participants' initial fixation landed on the rear region in most trials. Following this initial fixation, participants often made multiple saccades (Fig. 2*D*), and their fixation positions were dispersed during decision formation. The density of fixation positions during the full stimulus duration revealed a concentration around the two informative features (Fig. 2*B*). For identity and car categorization, the density plots exhibited two distinct peaks corresponding to the two features. For expression categorization, the two peaks were less distinct but still covered the two features. These patterns were qualitatively similar across the participants (Fig. S2*A*, *B*).

The distinct peaks in the density plots resulted from frequent saccades between the informative features. We plotted the distribution of saccade amplitudes and identified two peaks (Fig. 2*E*), one corresponding to small saccades within local features and the other to larger saccades spanning across features. We were particularly interested in larger saccades as they lead to large changes in the retinal image and can contribute to the integration of sensory information across distant features. We therefore extracted these "cross-feature" saccades and focused our analysis on their effects on decision-making behavior in the following sections. In brief, we defined saccades as cross-features if their start and end points were near the two features and their amplitudes were greater than two degrees (see Methods for more details). Examples of the start and end points of these saccades are shown in Figure 2*C*. These cross-feature saccades occurred on average 1.02 ± 0.12 times per trial (Fig. 2*F*) and appeared to be periodic with an average interval of ∼400 ms (Fig. 2*G*). We did not consider saccades spanning the left and right eyes as cross-features because the two eyes had the same morph level and did not provide distinct information in our stimuli. The influence of smaller saccades on decision-making is shown in Fig. S3*A*.

As expected from the frequent saccades between the two features, we observed that participants relied on both of them to judge the stimulus categories. We performed psychophysical reverse correlation (Ahumada, 1996; Okazawa et al., 2018, 2021), in which we averaged the fluctuations of morph levels of each feature during the participants' viewing of the feature in each trial and computed the difference in the average fluctuations between the trials in which participants chose Category 1 and Category 2 (Eq. 6 in Methods). The amplitudes of the resulting psychophysical kernels quantified the extent to which the fluctuations influenced the participants' choices (Fig. 2*H*). The kernels were positive for both features in all the tasks; even when we selected the feature weighted less for each participant and averaged them across the participants (Fig. 2*H*, right), they

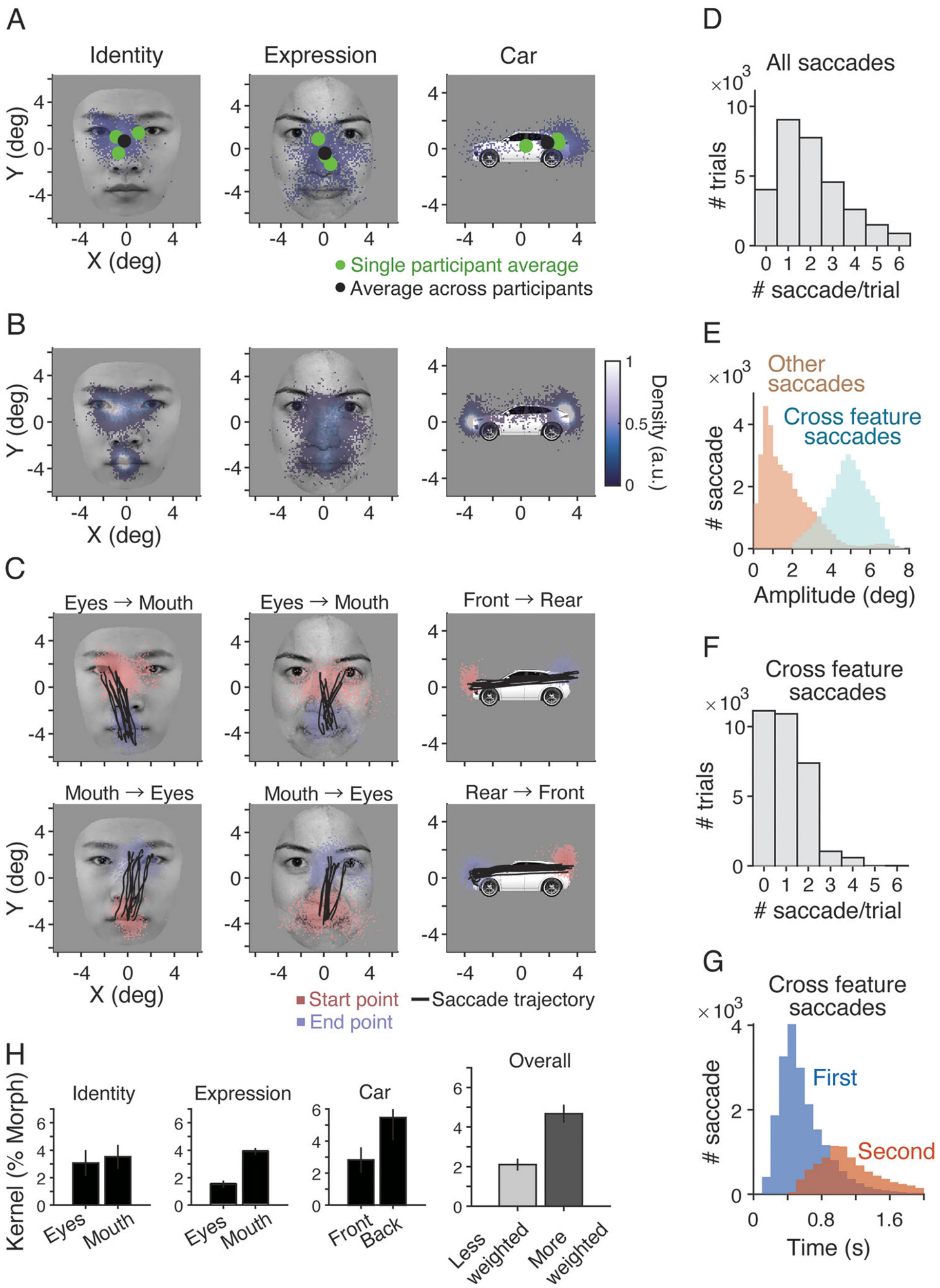

Figure 2. Participants made frequent saccades to sample two informative features. ***A***, Scatter plots of the first fixation positions after stimulus onset show a concentration just below the eyes in the face tasks and the rear part in the car task. The blue dots represent individual trials from a representative participant. Plots for each participant are shown in Fig. S2*A*. ***B***, Density plots of the fixated positions during the entire stimulus viewing period show that the participants primarily fixated around the two informative features. The plots shown are from representative participants; plots for all the participants can be found in Fig. S2*B*. ***C***, Example saccades spanning the two informative features. Plots for each participant are shown in Fig. S2*C*. ***D***, Distribution of saccade counts per trial. The trials were aggregated across participants. ***E***, The distribution of saccade amplitudes revealed two peaks. The peak with larger amplitudes corresponded to the saccades spanning the two features (cross-feature saccades; see Methods for its definition). ***F***, Participants made at most two cross-feature saccades in most of the trials. ***G***, Distribution of the timings of cross-feature saccades. ***H***, Consistent with the frequent fixations on the two features, psychophysical reverse correlation (Eq. 6) revealed positive influences of both features on participants' decisions. Error bars indicate SEM across participants.

were still significantly positive [$t_{(8)} = 7.1$, $p = 9.9 \times 10^{-5}$ for the feature weighted less; two-tailed paired t-test]. The kernel amplitudes seemed to differ between the features, such as higher values for mouth than for eyes in the expression task, but such a difference in feature weighting may depend on our choice of prototype stimuli and is not the focus of the current study. Rather, the key observation was that both features were used to solve the task, providing a basis for investigating whether and how eye movements were involved in gathering evidence from distant features.

Both features fixated before and after saccades contribute to decisions

Having extracted the cross-feature saccades, we now address one of our main questions: whether and how the information of local object features is integrated across these saccades. Many previous studies have demonstrated the trans-saccadic integration of simple visual stimuli (e.g., grating or color patch) seen at the fovea and periphery (Ganmor et al., 2015; Herwig et al., 2015; Wijdenes et al., 2015; Wolf and Schütz, 2015; Paeye et al., 2017; Shafer-Skelton et al., 2017; Stewart et al., 2020), but object recognition poses a different challenge because participants see different features across saccades.

To address this question, we leveraged our stochastic stimuli and first tested whether the morph fluctuations of the features fixated on before and after each saccade correlated with participants' choices regardless of the features being fixated on. We averaged the morph fluctuations of the fixated features between saccades (Fig. 3*A*) and plotted the participants' choice performance as a function of these morph levels (Fig. 3*B*, *C*). In the trials with one cross-feature saccade, this became a two-dimensional (2D) psychometric function (Fig. 3*B*). The plot displayed prominent diagonal iso-performance contours, indicating that both features before and after a saccade influenced the participants' decisions. Fitting a logistic function (Eq. 4) to this pattern revealed significant weights for both pre-saccade [$t_{(8)} = 7.2$, $p = 9.1 \times 10^{-5}$, two-tailed *t*-test across participants] and post-saccade features [$t_{(8)} = 6.3$, $p = 2.4 \times 10^{-4}$, two-tailed *t*-test] without a strong interaction term [$t_{(8)} = 0.62$, $p = 0.55$, two-tailed *t*-test]. Likewise, we analyzed the trials with two cross-feature saccades and confirmed that the morph levels of the features fixated during the first, second, and third fixation periods all contributed to the participants' choices [Fig. 3*C*; first period: $t_{(8)} = 3.7$, $p = 0.0063$, second period: $t_{(8)} = 8.5$, $p = 2.9 \times 10^{-5}$, third period: $t_{(8)} = 4.7$, $p = 0.0016$, interactions: $p = 0.082$, two-tailed *t*-test; Eq. 5]. Thus, the participants relied on information both before and after saccades to make their decisions. In the next section, we show that these results indicate the integration of evidence rather than random reliance on features before or after saccades.

We then used psychophysical reverse correlation to quantify the temporal dynamics of feature weighting and found a persistent contribution of fixated features across saccades. We computed the psychophysical kernels over time by calculating the difference in stimulus fluctuations at each time point between the trials in which the participants chose Category 1 and Category 2 (Eq. 6). The resulting kernels revealed rich temporal dynamics (Fig. 3*E*) and, importantly, had positive weights throughout the stimulus presentation when the feature was fixated (Fig. 3*E*; "pre" and "post" features indicate the features fixated before and after a saccade, Fig. 3*D*). When aligned to stimulus onset, the kernels tended to gradually decrease over time. Around the time of cross-feature saccades, the amplitudes of the pre- and post-saccade features were swapped. Because the temporal resolution of our stimulus fluctuations was ∼100 ms, we did not analyze further details of temporal dynamics around saccades (cf. Wolf and Schütz, 2015; but see Fig. S3*B* for kernels plotted with a higher temporal resolution). When aligned to the time of the participants' choice, we observed a characteristic peak around 400–500 ms before the choice (Fig. 3*E*, right). As demonstrated in the next section, these complex kernel dynamics can be explained quantitatively using a simple evidence accumulation model.

Psychophysical reverse correlations could also be used to quantify the spatial integration of sensory evidence. Instead of sorting stimulus fluctuations over time, we sorted the same data according to the visual distance between the participant's gaze position and each feature at each time point (Fig. 3*F*; see

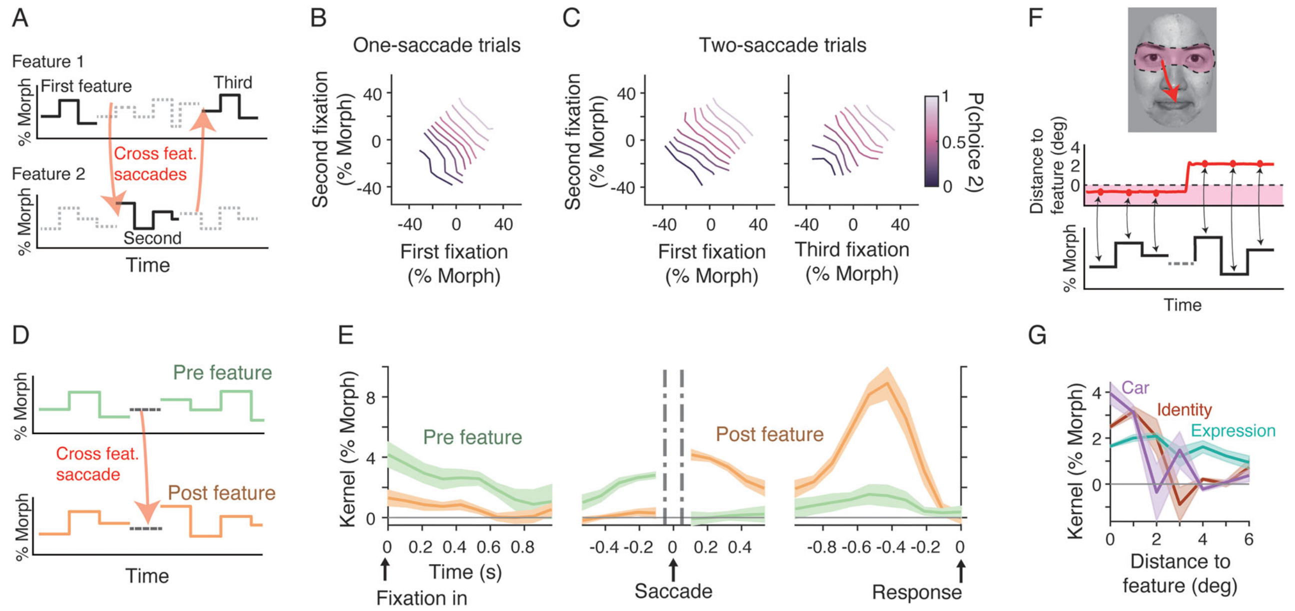

Figure 3. Both features fixated before and after saccades contribute to decisions. ***A***, We extracted the morph levels of the fixated feature (solid lines) at different fixation epochs split by cross-feature saccades and averaged them in each trial for the analysis in ***B*** and ***C***. The panel shows an example trial, in which a participant first fixated on feature 1, made a saccade to feature 2, and then fixated back on feature 1. ***B***, ***C***, 2D psychometric functions based on the average morph level of each fixation epoch. The lines are diagonal, indicating that both features fixated across saccades influenced participants' choices in the trials with one cross-feature saccade (***B***) and with two cross-feature saccades (***C***). ***D***, To quantify the temporal weighting of features across saccades, we performed psychophysical reverse correlation (Eq. 6) using the morph fluctuations of the features fixated before and after cross-feature saccades. The schematic shows an example of one saccade trials. The trials with more saccades are also included in the analysis (see Methods). ***E***, Psychophysical kernels indicate continuous influences of fixated features on participants' decisions. Their rich temporal dynamics can be explained by a simple evidence accumulation model (Fig. 4). Shading indicates SEM across participants. ***F***, Schematic for the analysis in ***G***. To examine spatial integration, we sorted the morph fluctuations based on the distance to each feature from participants' gaze position. The distance was calculated from a contour line manually circumscribing each feature (Fig. S1*B*; see Methods). ***G***, The amplitudes of psychophysical kernels decreased largely monotonically as a function of the distance to the features from the gaze position.

Methods for the definition of distance). The resulting kernels revealed a largely monotonic reduction as a function of distance (Fig. 3*G*) with some variability across the stimulus conditions. In the identity and car tasks, kernel amplitudes decreased sharply with distance, whereas they were much flatter in the expression task. Consistent with this finding, another line of analysis (Fig. S3*C, D*) confirmed that the influence of unfixated features was markedly greater on the expression task. Thus, the extent of the spatial window for integration can be either stimulus- or task-dependent (see Discussion). Nonetheless, the influence of the features can still be described as a monotonic function of the visual distance from the features in all tasks.

Across-saccade evidence accumulation accounts for behavior

Encouraged by our observation of the positive influence of features fixated on before and after saccades, we formally tested the integration of evidence across saccades by fitting an evidence accumulation model to the behavioral data. Previous studies have shown that face categorization behavior during fixation conditions could be explained using a simple model that computes the weighted sum of evidence from each facial feature and accumulates this sum over time (Okazawa et al., 2018, 2021). We extended this model by incorporating participants' eye movements such that the model kept accumulating evidence across saccades, but the informativeness of each feature depended on the gaze position (Tavares et al., 2017).

Our model integrates fluctuating sensory evidence from object features (e.g., eyes and mouth in face tasks) and accumulates evidence over time across saccades to form a decision variable (Fig. 4*A*). Each feature has a different strength of evidence (sensitivity parameter k_i in Eq. 10), which decays as a function of the distance between the feature and the participant's gaze position at each time point (decay rate λ in Eq. 10). Decay was modeled using an exponential function based on previous studies (Peli et al., 1991; Peterson and Eckstein, 2012), but other monotonic functions could similarly fit the data (Fig. S5*B*). Aside from

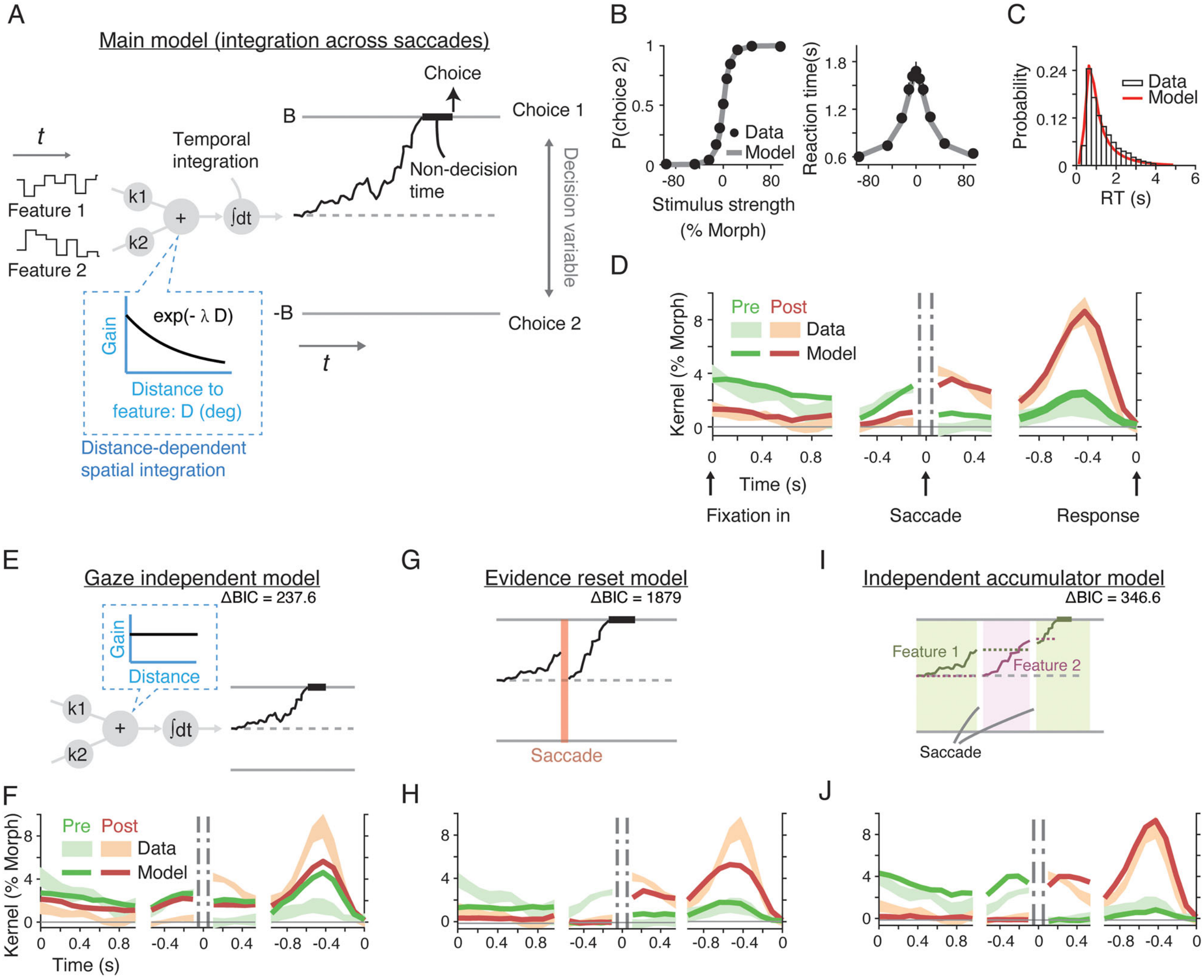

Figure 4. Across-saccade evidence accumulation model explains the behavioral results. ***A***, Our main model accumulates evidence throughout stimulus presentation across saccades until the accumulated evidence (decision variable) reaches a bound. The choice associated with the crossed bound is made after a non-decision time. Momentary evidence is computed as a linear sum of the morph levels of the two features with their weights as free parameters (k_1 and k_2), which are modulated by the distances to the features from the gaze position at each moment (blue inset). ***B–D***, The model quantitatively accounts for choices, mean reaction times (RTs), RT distributions (the panel includes the trials of all morph levels), and psychophysical kernels. Plots for individual participants are shown in Fig. S4. ***E, F***, If the weights for the features in the model are fixed regardless of the gaze positions (***E***), the amplitudes of fixated and unfixated kernels become similar, deviating from the data (***F***). ΔBIC indicates the difference in fit performance relative to the main model (positive values indicate poorer fits). ***G, H***, A model that resets evidence accumulation after saccades (***G***) fails to account for the amplitudes of data kernels before saccades (***H***). ***I, J***, Alternatively, if a model accumulates evidence independently for each feature and makes a decision based on one of the features that reached a bound first (***I***), it shows deviation of the kernel amplitudes from the data (***J***).

this decay, we did not assume any component in the model that depended on eye positions and saccades. When the decision variable reaches an upper or lower bound, the model makes a choice associated with the bound after a non-decision time that accounts for sensory and motor delays.

This simple extension of an evidence accumulation model accounted for all aspects of the behavioral data examined. The model accurately fitted the choices, mean RTs, and the distributions of RTs (Fig. 4*B, C*; $R^2 = 0.83 \pm 0.028$). Furthermore, it quantitatively accounted for the patterns of psychophysical kernels aligned to all of the time epochs (Fig. 4*D*; $R^2 = 0.51 \pm 0.074$). Note that the model kernels were not directly fitted to the participants' kernels but were simulated from an independent set of stimulus fluctuations (see Methods), making the comparison of data and models informative.

The dynamics of kernels observed in the data can be accounted for by the mechanistic components of evidence accumulation (Okazawa et al., 2018, 2021). The model explains the decreasing kernels aligned to stimulus onset because there is a temporal gap between the bound crossing and the report of a decision (i.e., the non-decision time), making a later portion of the stimulus fluctuations irrelevant to the decision. Because the timing of the bound crossing varies across trials, the model predicts a gradual reduction in the effect of stimulus fluctuations over time. The peak of the kernels aligned to the behavioral responses corresponded to the moment of crossing a decision bound in the model. At that moment, even tiny stimulus fluctuations bring the decision variable beyond a bound and dictate the decision, leading to the large kernel amplitudes. Subsequently, the kernels sharply drop to zero because of the non-decision time. The small peak for the kernel of the unfixated feature was observed because the evidence from the unfixated feature was also accumulated (weighted by the exponential decay function) and contributed to crossing the bound. When aligned to the time of saccades, the amplitudes of the kernels swapped between the pre- and post-saccade features because of the change in sensory sensitivity caused by the distance between the features and the gaze location.

Note that the peak of the response-aligned kernels indicates that the non-decision time was ∼400 ms, substantially longer than the expected minimum necessary time for sensory and motor processing (Bompas et al., 2024). In our model formalism, any time that did not depend on stimulus morph levels was explained as non-decision time (see T_0 in Methods), thus it could include the time needed to process complex object information and map object categories to the associated action plans.

We further confirmed that no other model accounted for the behavioral data without assuming gaze-dependent evidence accumulation. First, we tested a model that did not consider gaze position but accumulated evidence from two informative features with constant sensory sensitivity over time (Fig. 4*E*). This gaze-independent model could fit choices and RTs (Fig. S5*C*) but clearly failed to explain the differences in the amplitudes of psychophysical kernels between fixated and unfixated features (Fig. 4*F*). The model kernels showed slight differences between fixated and unfixated features because participants tended to fixate on features with higher sensitivity more frequently, but the differences were far smaller than those observed in the actual data.

Second, we considered a model that did not integrate evidence across saccades but restarted evidence accumulation after each saccade (Fig. 4*G*). According to this evidence-reset model, if participants cannot make a decision based on one feature, they switch their focus to the other feature and make a decision based on it. This model could also fit choices and RTs (Fig. S5*D*), but, as expected, failed to explain the amplitude of psychophysical kernels before saccades (Fig. 4*H*). The model kernels before saccades were positive owing to the inclusion of trials without saccades, but they were much smaller than the data.

Finally, we ruled out a hypothesis that participants relied on either of the two features to render a decision but did not integrate the evidence from both features across saccades (independent accumulator model; Fig. 4*I*). This model had two accumulators, one for each feature, which accumulated evidence only when the feature was fixated and was frozen when it was not fixated. A decision was made when either accumulator reached a bound. This model could fit choices and average RTs well (Fig. S5*E*) and also generated psychophysical kernels that resembled the data overall (Fig. 4*J*). However, the kernel amplitude for the unfixated feature was near zero, since the evidence from the two features was not integrated. The response-aligned kernel for the unfixated feature was weakly positive because the last saccade could have occurred shortly before the response. By contrast, the kernel amplitude for the fixated feature was higher than the amplitude of the data throughout the trials because a decision had to be based solely on the fixated feature, requiring greater sensitivity to it. Model comparison revealed that our main model provided a better fit to the data (ΔBIC = 346.6).

Overall, we found that a simple mechanism that accumulates sensory evidence across saccades is sufficient to account for the participants' object categorization behavior. Eye position information was required in the model to explain the decrease in sensitivity to features as a function of visual eccentricity; however, aside from that, we did not need to model the complex interactions between eye movements and the object recognition process. While these results cannot prove the absence of such interactions, our results in the next section also support the idea that the interactions of the visual and oculomotor systems do not play a substantial role in the types of object recognition behavior we examined.

Active saccade commands are unnecessary for feature integration

The observed integration of evidence across saccades could have depended on neural processes that combine visual signals and active saccade commands (i.e., efference copy), such as the predictive coding of visual features prior to saccades. To examine the extent of the contribution of efference copy, we next designed a "guided saccade" task that could compare object recognition performances between conditions with and without saccades (Fig. 5*A*). This task used the same object stimuli as in the free saccade task, but we strictly controlled the participants' eye movement and stimulus presentation. In the saccade condition, participants were explicitly instructed to make cross-feature saccades during object categorization (Fig. 5*B*). In the no-saccade condition, participants maintained fixation while there was a sudden change in the visual display, mimicking the change caused by saccades (Fig. 5*E*).

The saccade and no-saccade conditions had almost identical trial structures and stimulus durations. In each trial, a fixation point was initially placed at the center of one of the informative features of the stimulus. In the saccade condition, the fixation point moved to the location of the other informative feature immediately following the stimulus onset (Fig. 5*B*). The participants were required to make a saccade following this jump of the fixation point. The saccade was followed by another stimulus period (∼200 ms), and the participants reported their decisions

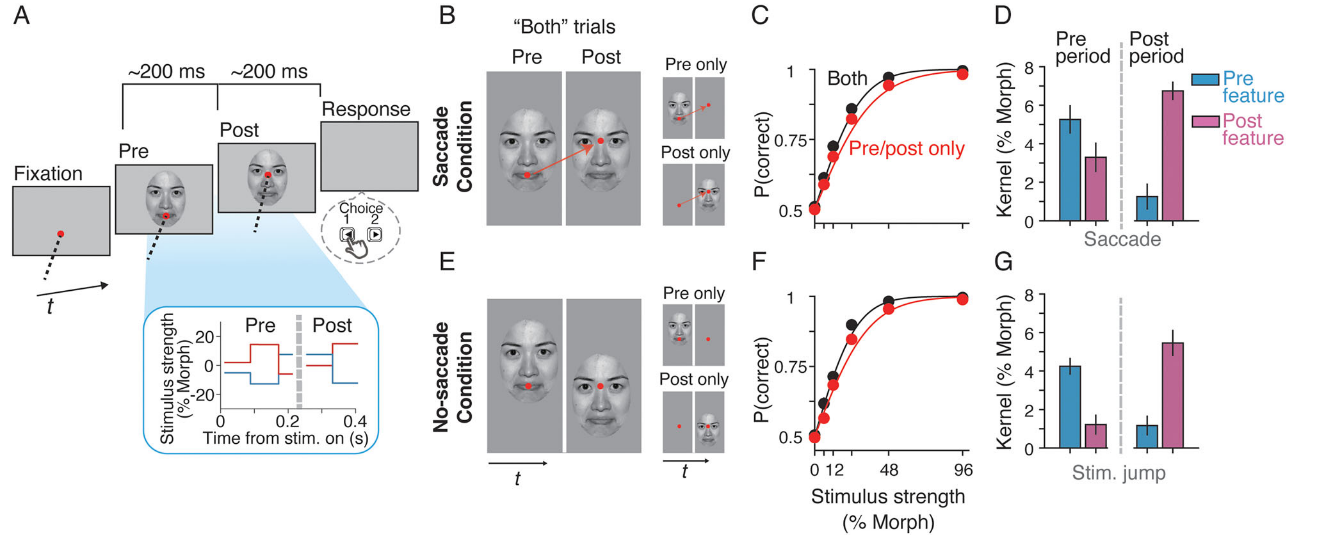

Figure 5. Guided saccade task revealed the lack of influence of efference copy on feature integration. ***A***, In this task, participants were asked to look at the red fixation dot, which moved from the position of one feature to the other in the saccade condition (***B***). After the saccade, the stimulus was extinguished in ∼200 ms, and participants reported their choice by pressing a button. As in the free saccade task, morph levels fluctuated every ∼100 ms (inset). Each participant was assigned to perform either identity, expression, or car categorization ($n = 9$). ***B, E***, We compared the saccade condition (***B***) with the non-saccade condition (***E***), in which the fixation point stayed at the same position, but the stimulus jumped, mimicking the visual display in the saccade condition. The stimulus duration was matched between the conditions within each participant (see Methods). To quantify the integration of evidence across saccades, we also had trials in which a stimulus was shown only before ("Pre only") or after ("Post only") the saccade or stimulus jump event. ***C, F***, Whether or not participants made a saccade, their performance improved when a stimulus was shown in both epochs. Error bars denoting SEM across participants were smaller than the data points. ***D, G***, Psychophysical kernels in the "both" condition showed positive influences of the fixated features before and after a saccade or stimulus jump, consistent with evidence integration. The data of individual participants are shown in Fig. S6.

by pressing a button after the stimulus was extinguished. In the no-saccade condition, the stimulus position suddenly shifted from one region to another, while the fixation point stayed the same, and the participants had to maintain fixation (Fig. 5*E*). The duration of the first stimulus and the blank between the two displays were set according to each participant's saccade latency (170.0 ms ± 5.3 ms) and duration (53.5 ms ± 1.3 ms) in the saccade condition; thus both the spatial and temporal profiles of stimuli were approximately matched between the two conditions. Each of the two conditions had trials where a stimulus was shown both before and after a saccade/stimulus jump ("Both" trials; Fig. 5*B*, *E*, left) and trials where a stimulus was shown only before or after ("Pre only" and "Post only" trials; Fig. 5*B*, *E*, right).

In the saccade condition, we confirmed that participants integrated evidence across a saccade. Their behavioral accuracy was significantly higher when an image was present across a saccade ("Both" trials) than when it was present only before or after a saccade [Fig. 5*C*; the difference in logistic slope $\alpha_2 = 1.69 \pm 0.36$, Eq. 3; $t_{(8)} = 4.75$, $p = 0.0014$, two-tailed *t*-test]. The improvement in performance was subtle but consistent with the near-optimal integration of evidence (Fig. S6), in line with prior studies that examined the saccadic integration of simpler features (Ganmor et al., 2015). Similar to the free saccade task, we also identified positive kernels for the fixated feature both before and after a saccade [Fig. 5*D*; before saccade $t_{(8)} = 6.41$, $p = 2.1 \times 10^{-4}$, after saccade $t_{(8)} = 10.8$, $p = 4.8 \times 10^{-6}$, two-tailed *t*-test].

Critically, higher performance for the "Both" trials was also observed in the no-saccade condition, supporting evidence integration [Fig. 5*F*; the difference in logistic slope $\alpha_2 = 2.01 \pm 0.39$, Eq. 3; $t_{(8)} = 5.12$, $p = 9.1 \times 10^{-4}$, two-tailed *t*-test]. In support of this, we also identified the positive kernels for the fixated features both before and after a saccade in the no-saccade condition [Fig. 5*G*; before saccade $t_{(8)} = 9.69$, $p = 1.1 \times 10^{-5}$, after saccade $t_{(8)} = 7.97$, $p = 4.5 \times 10^{-5}$, two-tailed *t*-test]. The size of the performance improvement was statistically indistinguishable from the saccade condition [$\alpha_2 = 0.32 \pm 0.50$, Eq. 3; $t_{(8)} = -0.64$, $p = 0.54$, two-tailed *t*-test]. One potential difference we noted was that the kernel for the unfixated feature before a saccade looked slightly higher in the saccade condition than in the no-saccade condition, which may indicate pre-saccadic enhancement in visual sensitivity (Li et al., 2021), but this difference was also statistically indistinguishable [Fig. 5*D*, *G*; $t_{(8)} = -1.71$, $p = 0.13$, two-tailed *t*-test]. Thus, the results indicate that saccade commands are not a prerequisite for feature integration and do not substantially improve behavioral performance even if they are effective (see Discussion for the interpretation).

Saccade frequency was minimally influenced by ongoing decision formation

Thus far, we have focused on the mechanisms of perceptual decision-making during object recognition, but our free saccade task (Fig. 1) also allowed us to examine whether and how saccade patterns are modulated by the ongoing decision-making process. For example, when the currently fixated feature is uninformative, people may make frequent saccades for the other feature to seek for more evidence (Li et al., 2023), or people may spend more time fixating on each feature (Horstmann et al., 2017; Einhäuser et al., 2020). If so, there could be a correlation between stimulus difficulty and saccadic frequency.

Contrary to these expectations, we did not find any significant relationship between stimulus difficulty and saccade frequency in our tasks. The average number of cross-feature and other saccades in each trial was clearly higher in more difficult trials (Fig. 6*A*), but this did not indicate more frequent saccades because the trial duration (i.e., RT) was longer in difficult trials (Fig. 1*C*, bottom). Therefore, the number of saccades per time had to be estimated, but the calculation requires greater complexity than a simple division of the number of saccades by the trial duration, as saccades tend to be periodic (Fig. 2*G*), and the

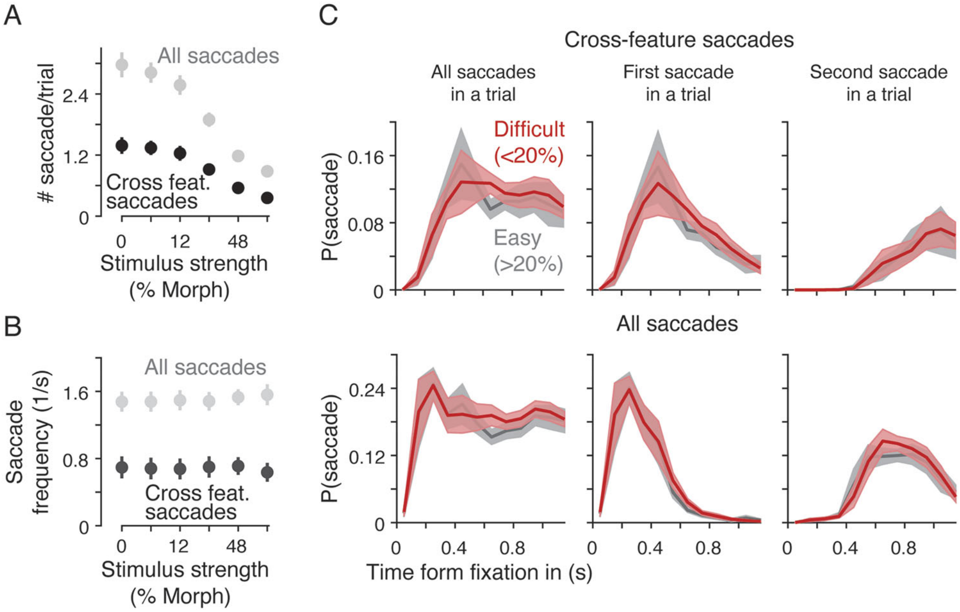

Figure 6. The frequency of saccades did not significantly depend on stimulus difficulty. ***A***, The average number of saccades per trial was higher for more difficult (low morph level) trials. Error bars indicate the SEM across participants. But reaction times were also longer for these trials (Fig. 1*C*). ***B***, When saccade frequency per time was calculated, it was not correlated with stimulus difficulty. To allow the unbiased estimation of saccade frequency, we matched RT distributions across stimulus strength when calculating the frequency (see main text and Methods). ***C***, Without matching RT distributions, we compared the probability of making a saccade at each time point by dividing the count of trials with a saccade at that time point by the count of trials whose RTs were longer than that time. We compared this probability between easy and difficult trials and found that they were statistically indistinguishable.

calculation strongly depends on the relative distributions of RTs and saccade timing. Therefore, we matched the RT histograms of different stimulus strengths by randomly subsampling trials. After the RT matching, saccade frequency was not correlated with stimulus difficulty [Fig. 6*B*; cross-feature saccades: $F_{(5,48)} = 0.08$, $p = 0.995$; all saccades: $F_{(5,48)} = 0.1$, $p = 0.992$, repeated-measures one-way ANOVA].

We further corroborated this conclusion through an analysis that does not rely on matching RTs (Fig. 6*C*). This was important as RTs could be highly correlated with participants' subjective uncertainty (Kiani et al., 2014) and thus matching RTs could also align subjective uncertainty across stimulus strengths. Instead, we calculated the probability of making a saccade at each time point using only trials with RTs longer than the time point. As long as there were a sufficient number of trials at each time point, this probability should not be biased by the duration of the RTs following each time point. We plotted this probability using either all cross-feature saccades, the first cross-feature saccade in a trial, or the second saccade in a trial, and did not find differences between easy and difficult trials (Fig. 6*C*, top; split by 20% morph boundary; $p = 0.747$, repeated-measures two-way ANOVA). Including within-feature saccades did not affect this result (Fig. 6*C* bottom; $p = 0.507$, repeated-measures two-way ANOVA). In summary, saccades appeared to occur in a stochastic manner without clear influence from the ongoing decision-making process in our tasks.

Discussion

Humans often make saccades to sample local informative features when viewing object images, but the mechanisms by which saccades contribute to object recognition have yet to be established. Studies on eye movements have proposed complex interactions between the visual and oculomotor systems, such as predictive visual processing based on saccadic commands (Binda and Morrone, 2018), whereas studies on object recognition lean toward eschewing this complexity and favor a briefly flashed static image under fixation conditions (DiCarlo et al., 2012). Here, we applied a decision-making theory to reformulate the problem as the accumulation of sensory evidence from multiple local features across saccades. Our results indicate a parsimonious relationship between eye movements and object recognition; humans integrated evidence across saccades (Figs. 3 and 4), but behavioral performance did not strongly depend on active saccade signals (Fig. 5). As such, a simple evidence accumulation model that does not assume complex interactions between the visual and oculomotor systems can approximate decision-making behaviors.

Many prior studies have documented the integration of visual information across saccades (Ganmor et al., 2015; Herwig et al., 2015; Wijdenes et al., 2015; Wolf and Schütz, 2015; Paeye et al., 2017; Shafer-Skelton et al., 2017; Stewart et al., 2020), but to our knowledge, ours is the first attempt to apply a mechanistic model grounded on evidence accumulation to account for object recognition involving saccades. The accumulation of evidence is common in many perceptual tasks (Shadlen and Kiani, 2013), and existing models of form and object vision suggest that information is integrated across saccades (Renninger et al., 2004, 2007; Akbas and Eckstein, 2017). Thus, one might argue that our results were largely expected. However, we believe that we have provided important empirical tests for the following three points. First, we were able to demonstrate the integration of evidence across saccades using naturalistic stimuli (faces and objects) through a model fitting approach that quantitatively explains various aspects of the participants' choice behavior. Second, this modeling approach allowed us to test the integration process during a free viewing condition in which participants' gaze locations and times were unrestricted. Finally, we demonstrated a limited role for efference copy in this process, confirming that a simple evidence accumulation model (Fig. 4) is sufficient to explain behavior in our tasks.

The apparent lack of necessity for efference copy (Fig. 5) seemingly contradicts a large body of existing studies underscoring its role in visual perception (Ross et al., 2001; Melcher, 2005, 2007; Binda and Morrone, 2018), but this could stem from the fact that the demands for our tasks were different. In conventional trans-saccadic perceptual tasks, participants are required to judge a simple stimulus (e.g., the orientation of a Gabor patch) initially viewed in the periphery and subsequently foveated through a saccade (Ganmor et al., 2015; Herwig et al., 2015; Wijdenes et al., 2015; Wolf and Schütz, 2015). Here, the remapping of receptive fields induced by oculomotor commands (Binda and Morrone, 2018) could serve as a vehicle to fuse information across saccades (Melcher, 2005, 2007), although the necessity of an efference copy is still debated even under this condition (Henderson and Anes, 1994; Bays and Husain, 2007). In contrast, object images comprise multiple distinct features. We believe that the integration of foveated features across saccades is more important for improving behavioral accuracy in this case, which may not require oculomotor information. It is also possible that our behavioral data did not have sufficient statistical power to detect the effects of oculomotor signals on feature integration, but even if there is an effect, its effect size must be limited according to our results.

Regarding the frequency of saccades, we found minimal influence of uncertainty in ongoing decision making (Fig. 6). Previous studies have found that the timing of saccades is modulated by internal states of visual recognition (Henderson and Pierce, 2008; Castelhano et al., 2009; Nuthmann et al., 2010; Laubrock et al., 2013; Trukenbrod and Engbert, 2014; Nuthmann, 2017; Einhäuser et al., 2020). We did not observe such effects in our tasks, possibly because our participants were extensively trained to perform saccadic sampling on the same visual images and had established stereotyped saccadic patterns based on the learned statistics of the stimuli. Also, the average stimulus strength was always matched between two informative features in our tasks, potentially providing less incentive for participants to modulate their saccade patterns depending on decision uncertainty. Nevertheless, the lack of effects suggests that evidence accumulation during decision making and eye movement planning can be independent. To explain periodic saccades, some previous theories have considered the drift-diffusion process that triggers a saccade when it reaches a threshold, effectively functioning like a stochastic clock counter (Nuthmann et al., 2010; Trukenbrod and Engbert, 2014; Mengers et al., 2024), but it is likely that such a process and the evidence accumulation we modeled here are separate processes in the brain.

The spatial distribution of saccades we observed were consistent with previous studies (Fig. 2). Participants tended to look at informative regions or nearby areas in all tasks (Fig. 2*B*) and made frequent saccades across features (Fig. 2*C*). A seminal study by Peterson and Eckstein (2012) showed that people look just below the eyes during face recognition, which is consistent with the prediction of an ideal observer model. Our observations of the initial fixation positions replicated these findings (Fig. 2*A*), whereas fixations diverged to other locations during prolonged viewing, which is also consistent with their results (Or et al., 2015). A tantalizing question is whether this behavior is optimal in terms of information sampling (Najemnik and Geisler, 2005; Vandormael et al., 2017; Callaway et al., 2021), but defining optimality in our tasks is not trivial. An ideal observer model that learned the statistical regularity of our stimuli may just continue to fixate on the most informative part of an image throughout stimulus viewing, but sampling multiple informative features might also become optimal depending on its definition (Renninger et al., 2004, 2007; Hoppe and Rothkopf, 2019).

It would also be worthwhile to discuss how our findings extend to other natural object stimuli. We designed our stimuli to have two distant informative features to manipulate informativeness and detect the participants' saccades across the features. Although we designed our stimuli to be naturalistic—eyes and mouth are indeed informative features for face recognition (Schyns et al., 2002; Okazawa et al., 2021)—other object images would have more than two diagnostic features that could also overlap with each other. In such cases, people may not directly look at each feature but rather look at a place between the features to sample evidence. Indeed, in the expression task, the participants tended to fixate between the eyes and mouth (Fig. 2*B*) and had a broader spatial window (Fig. 3*G*) sampling evidence from both fixated and unfixated features (Fig. S3*C–E*). This might indicate that the participants adjusted their spatial sampling windows depending on the task context. Stimulus size would also influence sampling strategies (von Wartburg et al., 2007; Otero-Millan et al., 2013). In our experiments, we used a naturalistic range of object sizes (McKone, 2009), but if the viewing distance increases, the image would become too small to make saccades inside, in which case features could be spatially integrated without saccades (Okazawa et al., 2021). Finally, we covered our stimuli with dynamic noise, which could have made it difficult to glean evidence from the unfixated feature before saccades. Had we removed the noise, the participants' spatial accumulation window might have been broader than observed (Fig. 3*G*). Although we were unable to test all possible conditions, our model—evidence accumulation that depends only on the eccentricity of features—can readily offer quantitative predictions of object recognition behaviors in any of these settings.

We hope that the simplicity of the proposed framework encourages further investigation of object recognition under naturalistic viewing. Our findings suggest that, despite the apparent complexity of oculomotor events, there could be stable representations of momentary and accumulated sensory evidence across saccades in the neural circuitry for object processing (Bonnen et al., 2023; Xiao et al., 2024) and decision-making (So and Shadlen, 2022). Such a solution for active object vision can also be easily implemented in image-computable models, since it only requires accumulating evidence from the model responses to each snapshot of image sequences.

Data and Code Availability

Data and code used in this study are available at https://osf.io/ckap7/.

References

Ahumada Jr A (1996) Perceptual classification images from Vernier acuity masked by noise. Perception 25:2–2.

Akbas E, Eckstein MP (2017) Object detection through search with a foveated visual system. PLoS Comput Biol 13:e1005743.

Arizpe J, Kravitz DJ, Yovel G, Baker CI (2012) Start position strongly influences fixation patterns during face processing: difficulties with eye movements as a measure of information use. PLoS One 7:e31106.

Bays PM, Husain M (2007) Spatial remapping of the visual world across saccades. Neuroreport 18:1207–1213.

Binda P, Morrone MC (2018) Vision during saccadic eye movements. Annu Rev Vis Sci 4:193–213.

Bompas A, Sumner P, Hedge C (2024) Non-decision time: the Higgs boson of decision. Psychol Rev 132:330–363.

Bonnen T, Wagner AD, Yamins DL (2023) Medial temporal cortex supports compositional visual inferences. bioRxiv, pp 2023–09.

Brainard DH, Vision S (1997) The psychophysics toolbox. Spat Vis 10:433–436.

Buchan JN, Paré M, Munhall KG (2007) Spatial statistics of gaze fixations during dynamic face processing. Soc Neurosci 2:1–13.

Callaway F, Rangel A, Griffiths TL (2021) Fixation patterns in simple choice reflect optimal information sampling. PLoS Comput Biol 17:e1008863.

Castelhano MS, Mack ML, Henderson JM (2009) Viewing task influences eye movement control during active scene perception. J Vis 9:6–6.

Demeyer M, De Graef P, Wagemans J, Verfaillie K (2009) Transsaccadic identification of highly similar artificial shapes. J Vis 9:28–28.

Demeyer M, De Graef P, Wagemans J, Verfaillie K (2010) Parametric integration of visual form across saccades. Vision Res 50:1225–1234.

DiCarlo JJ, Zoccolan D, Rust NC (2012) How does the brain solve visual object recognition? Neuron 73:415–434.

Eckstein MP (2011) Visual search: a retrospective. J Vis 11:14–14.

Einhäuser W, Atzert C, Nuthmann A (2020) Fixation durations in natural scene viewing are guided by peripheral scene content. J Vis 20:15–15.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429–433.

Ganmor E, Landy MS, Simoncelli EP (2015) Near-optimal integration of orientation information across saccades. J Vis 15:8.

Heidari-Gorji H, Ebrahimpour R, Zabbah S (2021) A temporal hierarchical feedforward model explains both the time and the accuracy of object recognition. Sci Rep 11:5640.

Heisz JJ, Shore DI (2008) More efficient scanning for familiar faces. J Vis 8:9–9.

Henderson JM, Anes MD (1994) Roles of object-file review and type priming in visual identification within and across eye fixations. J Exp Psychol Hum Percept Perform 20:826–839.

Henderson JM, Pierce GL (2008) Eye movements during scene viewing: evidence for mixed control of fixation durations. Psychon Bull Rev 15:566–573.

Henderson JM, Williams CC, Falk RJ (2005) Eye movements are functional during face learning. Mem Cognit 33:98–106.

Herwig A, Weiß K, Schneider WX (2015) When circles become triangular: how transsaccadic predictions shape the perception of shape. Ann N Y Acad Sci 1339:97–105.

Hessels RS (2020) How does gaze to faces support face-to-face interaction? A review and perspective. Psychon Bull Rev 27:856–881.

Hoppe D, Rothkopf CA (2019) Multi-step planning of eye movements in visual search. Sci Rep 9:144.

Horstmann G, Becker S, Ernst D (2017) Dwelling, rescanning, and skipping of distractors explain search efficiency in difficult search better than guidance by the target. Vis cogn 25:291–305.

Hsiao JH-w., Cottrell G (2008) Two fixations suffice in face recognition. Psychol Sci 19:998–1006.

Hsiao JH, Liu TT (2012) The optimal viewing position in face recognition. J Vis 12:22–22.

Kanan C, Bseiso DN, Ray NA, Hsiao JH, Cottrell GW (2015) Humans have idiosyncratic and task-specific scanpaths for judging faces. Vision Res 108:67–76.

Kiani R, Hanks TD, Shadlen MN (2008) Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. J Neurosci 28:3017–3029.

Kiani R, Corthell L, Shadlen MN (2014) Choice certainty is informed by both evidence and decision time. Neuron 84:1329–1342.

Krajbich I, Armel C, Rangel A (2010) Visual fixations and the computation and comparison of value in simple choice. Nat Neurosci 13:1292–1298.

Lakens D (2022) Sample size justification. Collabra Psychol 8:33267.

Larsson L, Nyström M, Stridh M (2013) Detection of saccades and postsaccadic oscillations in the presence of smooth pursuit. IEEE Trans Biomed Eng 60:2484–2493.

Laubrock J, Cajar A, Engbert R (2013) Control of fixation duration during scene viewing by interaction of foveal and peripheral processing. J Vis 13:11–11.

Li H-H, Pan J, Carrasco M (2021) Different computations underlie overt presaccadic and covert spatial attention. Nat Hum Behav 5:1418–1431.

Li X, Su R, Chen Y, Yang T (2023) Optimal policy for uncertainty estimation concurrent with decision making. Cell Rep 42:112232.

Ludwig CJ, Davies JR, Eckstein MP (2014) Foveal analysis and peripheral selection during active visual sampling. Proc Natl Acad Sci 111:E291–E299.

Luo T, Xu M, Zheng Z, Okazawa G (2025) Limitation of switching sensory information flow in flexible perceptual decision making. Nat Commun 16:172.

Mack ML, Palmeri TJ (2011) The timing of visual object categorization. Front Psychol 2:165.

McKone E (2009) Holistic processing for faces operates over a wide range of sizes but is strongest at identification rather than conversational distances. Vision Res 49:268–283.

Melcher D (2005) Spatiotopic transfer of visual-form adaptation across saccadic eye movements. Curr Biol 15:1745–1748.

Melcher D (2007) Predictive remapping of visual features precedes saccadic eye movements. Nat Neurosci 10:903–907.

Mengers V, Roth N, Brock O, Obermayer K, Rolfs M (2024) A robotics-inspired scanpath model reveals the importance of uncertainty and semantic object cues for gaze guidance in dynamic scenes. Preprint, arXiv:2408.01322.

Murray RF (2011) Classification images: a review. J Vis 11:2–2.

Najemnik J, Geisler WS (2005) Optimal eye movement strategies in visual search. Nature 434:387–391.

Nuthmann A, Smith TJ, Engbert R, Henderson JM (2010) CRISP: a computational model of fixation durations in scene viewing. Psychol Rev 117: 382–405.

Nuthmann A (2017) Fixation durations in scene viewing: modeling the effects of local image features, oculomotor parameters, and task. Psychon Bull Rev 24:370–392.

Okazawa G, Sha L, Purcell BA, Kiani R (2018) Psychophysical reverse correlation reflects both sensory and decision-making processes. Nat Commun 9:3479.

Okazawa G, Sha L, Kiani R (2021) Linear integration of sensory evidence over space and time underlies face categorization. J Neurosci 41:7876–7893.

Or CC-F, Peterson MF, Eckstein MP (2015) Initial eye movements during face identification are optimal and similar across cultures. J Vis 15:12–12.

Oruç I, Maloney LT, Landy MS (2003) Weighted linear cue combination with possibly correlated error. Vision Res 43:2451–2468.

Otero-Millan J, Macknik SL, Langston RE, Martinez-Conde S (2013) An oculomotor continuum from exploration to fixation. Proc Natl Acad Sci U S A 110:6175–6180.

Paeye C, Collins T, Cavanagh P (2017) Transsaccadic perceptual fusion. J Vis 17:14–14.

Peli E, Yang J, Goldstein RB (1991) Image invariance with changes in size: the role of peripheral contrast thresholds. J Opt Soc Am A Opt Image Sci Vis 8:1762–1774.

Peterson MF, Lin J, Zaun I, Kanwisher N (2016) Individual differences in face-looking behavior generalize from the lab to the world. J Vis 16: 12–12.

Peterson MF, Eckstein MP (2012) Looking just below the eyes is optimal across face recognition tasks. Proc Natl Acad Sci U S A 109:E3314–E3323.

Poth CH, Herwig A, Schneider WX (2015) Breaking object correspondence across saccadic eye movements deteriorates object recognition. Front Syst Neurosci 9:176.

Renninger L, Coughlan J, Verghese P, Malik J (2004) An information maximization model of eye movements. Adv Neural Inf Process Syst 17:1121–1128.

Renninger LW, Verghese P, Coughlan J (2007) Where to look next? Eye movements reduce local uncertainty. J Vis 7:6.

Ross J, Morrone MC, Goldberg ME, Burr DC (2001) Changes in visual perception at the time of saccades. Trends Neurosci 24:113–121.

Schurgin M, Nelson J, Iida S, Ohira H, Chiao J, Franconeri S (2014) Eye movements during emotion recognition in faces. J Vis 14:14–14.

Schyns PG, Bonnar L, Gosselin F (2002) Show me the features! understanding recognition from the use of visual information. Psychol Sci 13:402–409.

Shadlen MN, Kiani R (2013) Decision making as a window on cognition. Neuron 80:791–806.

Shafer-Skelton A, Kupitz CN, Golomb JD (2017) Object-location binding across a saccade: a retinotopic spatial congruency bias. Atten Percept Psycho 79:765–781.

Smith PL, Little DR (2018) Small is beautiful: in defense of the small-N design. Psychon Bull Rev 25:2083–2101.

So N, Shadlen MN (2022) Decision formation in parietal cortex transcends a fixed frame of reference. Neuron 110:3206–3215.

Stewart EE, Valsecchi M, Schütz AC (2020) A review of interactions between peripheral and foveal vision. J Vis 20:2–2.

Tavares G, Perona P, Rangel A (2017) The attentional drift diffusion model of simple perceptual decision-making. Front Neurosci 11:268904.

Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, Marcus DJ, Westerlund A, Casey B, Nelson C (2009) The NimStim set of facial expressions: judgments from untrained research participants. Psychiatry Res 168:242–249.

Trukenbrod HA, Engbert R (2014) ICAT: a computational model for the adaptive control of fixation durations. Psychon Bull Rev 21:907–934.

Vandormael H, Herce Castañón S, Balaguer J, Li V, Summerfield C (2017) Robust sampling of decision information during perceptual choice. Proc Natl Acad Sci U S A 114:2771–2776.

von Wartburg R, Wurtz P, Pflugshaupt T, Nyffeler T, Lüthi M, Müri RM (2007) Size matters: saccades during scene perception. Perception 36: 355–365.

Wijdenes LO, Marshall L, Bays PM (2015) Evidence for optimal integration of visual feature representations across saccades. J Neurosci 35:10146–10153.

Wolf C, Schütz AC (2015) Trans-saccadic integration of peripheral and foveal feature information is close to optimal. J Vis 15:1–1.

Xiao W, Sharma S, Kreiman G, Livingstone MS (2024) Feature-selective responses in macaque visual cortex follow eye movements during natural vision. Nat Neurosci 27:1157–1166.

Yang T, Yang Z, Xu G, Gao D, Zhang Z, Wang H, Liu S, Han L, Zhu Z, Tian Y, et al. (2020) Tsinghua facial expression database–a database of facial expressions in Chinese young and older women and men: development and validation. PLoS One 15:e0231304.

Yang X, Krajbich I (2023) A dynamic computational model of gaze and choice in multi-attribute decisions. Psychol Rev 130:52–70.

Yarbus AL (1967) *Eye movements and vision*. New York, NY: Springer.