

Reviewed Preprint

v1 • June 29, 2026

Not revised

✉ For correspondence:

xmch@ucdavis.edu

Competing interests: No competing interests declared

Funding: See [page 17](#)

Reviewing editor: Emilio Salinas, Wake Forest University School of Medicine, United States

© 2026, Soyuhos et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

Meaning-based guidance of attention in rhesus monkeys during naturalistic scene viewing

Orhan Soyuhos^{1,2}, Taylor R Hayes³, Wenqing Hu^{1,2}, Taylor P Hamel^{1,4}, Brinda Sevak^{1,5}, John M Henderson^{2,3}, Xiaomo Chen^{1,4} ✉

¹Center for Neuroscience, University of California, Davis, Davis, United States • ²Department of Psychology, University of California, Davis, Davis, United States • ³Center for Mind and Brain, University of California, Davis, Davis, United States •

⁴Department of Neurobiology, Physiology and Behavior, University of California, Davis, Davis, United States •

⁵Department of Biomedical Engineering, University of California, Davis, Davis, United States

eLife Assessment

This **valuable** study shows that macaque monkeys preferentially fixate regions in natural scenes that are classified as "meaningful" by a computational model - an earlier model that was developed to identify locations that are semantically informative to humans - suggesting that overt attention to structured visual content is shared across primates. However, support is **incomplete** for the stronger claim that macaques are guided by semantic meaning, which is confounded by lower-level visual features that co-vary with it and by methodological limitations that complicate interpretation. If the semantic interpretation were more reliably established, the significance of the findings would increase, as they would connect the human cognitive process of scene understanding to neural circuit mechanisms accessible in non-human primates.

<https://doi.org/10.7554/eLife.111311.1.sa3>

Abstract

In humans and other primates, high-acuity vision is restricted to the fovea, requiring frequent saccadic eye movements to sample visual information, a process known as overt attention. Classical visual salience theory explains how low-level image features guide these movements, but recent human studies have shown that overt attention is also strongly guided by scene meaning—the spatial distribution of semantic informativeness. Whether this form of attentional guidance is uniquely human or shared across primate vision remains unknown. Here, we addressed this question by recording eye movements from two rhesus macaques freely viewing naturalistic indoor scenes. Fixation selection was modeled using meaning maps alongside image-based salience maps and center proximity. In both monkeys, meaning robustly predicted fixation selection after controlling for visual salience and center bias. Moreover, high-meaning regions captured attention independently of visual salience, whereas salience played an increasingly important role as meaning decreased. While this prioritization of meaningful regions remained robust across environments, familiarity broadened visual exploration by increasing the likelihood of fixating less meaningful areas. Finally, the influence of meaning on fixation selection strengthened with attentional engagement. These findings suggest that meaning-based attention is an evolutionarily conserved component of primate vision and establish a behavioral foundation for investigating its neural mechanisms.

Introduction

Real-world visual scenes contain far more information than can be processed at once (Fei-Fei et al., 2007; Henderson, 2003). In humans and other primates, high-acuity vision is limited to the fovea, requiring frequent saccadic eye movements to sample the visual environment—a process known as overt attention (Findlay & Gilchrist, 2003; Henderson, 2013). Classical models of attentional selection emphasize the role of visual salience in guiding attention during naturalistic scene viewing, proposing that gaze is drawn toward locations with low-level visual features such as luminance contrast, color, and orientation, which are combined into a salience map that guides stimulus-driven attention (Itti & Koch, 2001; Parkhurst et al., 2002; Wolfe, 1994). Although these models successfully explain salience-driven aspects of fixation behavior, they do not fully account for gaze patterns during free-viewing, in which both visual salience and scene meaning contribute to attentional selection. For example, during everyday navigation, observers reliably fixate traffic signals, road boundaries, and signage even when these elements are small, low in contrast, or visually inconspicuous.

A major obstacle to evaluating the role of meaning in attentional guidance has been the lack of methods for quantifying semantic information in visual scenes (Henderson, 2017). The meaning map approach was first developed to systematically quantify the spatial distribution of semantic informativeness across real-world scenes (Henderson & Hayes, 2017; Henderson et al., 2019). However, this method relied on labor-intensive human rating procedures. More recently, the DeepMeaning model provided an image-computable extension of this framework by leveraging a pretrained vision–language model to estimate local semantic informativeness (Hayes & Henderson, 2025). Using these approaches, scene meaning has been shown to predict fixation selection beyond low-level salience in human adults (Hayes & Henderson, 2022a, 2022b; Henderson & Hayes, 2017, 2018; Zhang et al., 2022) and infants (Oakes et al., 2024), even when meaning is task-irrelevant (Peacock et al., 2019). Notably, meaning-based guidance emerges rapidly during attentional deployment, influencing the earliest fixations (Henderson & Hayes, 2017, 2018), and neural signatures associated with the processing of scene meaning appear shortly after visual onset during EEG recording (Kiat et al., 2022). Together, these findings suggest that attentional selection during naturalistic viewing is robustly shaped by high-level semantic information in humans.

Whether meaning-based guidance of attention reflects a uniquely human strategy or a conserved feature of the primate visual system remains unknown. Non-human primates provide a critical model for investigating the circuit mechanisms underlying visual attention, given their close anatomical, functional, and behavioral similarities to humans (Bichot et al., 2015; Buschman & Kastner, 2015; Buschman & Miller, 2007). Extensive work in macaques has identified attentional control signals distributed across interconnected cortical and subcortical networks, including frontoparietal, temporal, and midbrain structures, that support both visual salience-driven and goal-directed attention (Boshra & Kastner, 2022; Moore & Zirnsak, 2017; Xia et al., 2024). In parallel, neurons throughout the visual hierarchy exhibit selectivity for complex object features and higher-order visual content, indicating that primate visual representations extend well beyond low-level image statistics (Peelen & Kastner, 2014; Ponce et al., 2019; Taubert et al., 2017; Vinken et al., 2025). However, despite this extensive knowledge of attentional circuitry and visual representation, it remains unclear whether scene meaning itself guides attentional behavior in non-human primates during naturalistic viewing.

Here, we examined the extent to which overt attention in macaques is guided by scene meaning and whether this guidance depends on behavioral context. We recorded eye movements from two rhesus macaques freely viewing naturalistic indoor scenes and modeled fixation selection using meaning maps, low-level salience maps, and spatial center bias. To assess contextual modulation of meaning-based guidance, we compared gaze behavior during viewing of unfamiliar and familiar scenes and across different levels of attentional engagement. We found that scene meaning robustly predicts fixation selection in macaques and exerts an influence comparable to that of low-level salience after accounting for center bias. In addition, meaning significantly

interacted with visual salience in guiding attention, such that high-meaning regions captured attention independently of visual salience, whereas salience played an increasing role as meaning decreased. Moreover, attentional guidance was modulated by behavioral context: while attention remained consistently directed toward high-meaning regions in both unfamiliar and familiar scenes, familiarity increased the likelihood of fixations on low-meaning areas. Additionally, the influence of meaning strengthened when scenes were more engaging, as quantified by viewing time. Overall, these findings demonstrate that meaning-based guidance of attention is an evolutionarily conserved feature of the primate visual system and establish a behavioral foundation for future mechanistic investigations into how scene meaning is represented and utilized by neural circuits.

Results

We recorded eye movements from two male rhesus macaques (Monkey V and Monkey I) freely viewing naturalistic indoor scenes. The stimulus set consisted of 200 images evenly split between unfamiliar and familiar scenes. Unfamiliar scenes depicted indoor environments the monkeys had never experienced, such as restaurants and residential interiors, whereas familiar scenes were photographs of their housing and laboratory surroundings. Each trial began with central fixation followed by 5 s of free-viewing (Fig. 1A). In total, we analyzed 3,140 fixations from Monkey V and 2,943 fixations from Monkey I across 193 and 176 images, respectively.

Scene meaning guides overt attention during free-viewing

To determine whether overt attention in macaques is guided by scene meaning, we quantified fixation selection as a function of three spatial predictors and their interactions: scene meaning, image salience, and center proximity (Fig. 1B). Scene meaning indexed the spatial distribution of semantic informativeness (Hayes & Henderson, 2025; Henderson & Hayes, 2017). Image salience captured local contrast in lowlevel features such as luminance and edge orientation (Itti et al., 1998). Center proximity measured distance from the screen center to control explicitly for the known bias toward central fixation, independent of image content (Hayes & Henderson, 2022b; Tatler, 2007). Importantly, we verified that these spatial predictors captured distinct sources of variance with negligible multicollinearity prior to modeling (Fig. S2). For each predictor, values at fixated locations were compared against those at randomly sampled non-fixated locations (see Methods; Fig. 1C–D).

If scene meaning guides attention in macaques, then locations with higher meaning values should be more likely to be fixated, even after accounting for salience and center bias. To test this, we fit a Bayesian Bernoulli generalized linear mixed model (GLMM) predicting whether a sampled location was fixated based on meaning, salience, and center proximity, with random effects for scene (Table 1; Fixation model). For Monkey V, meaning showed a robust positive effect ($\beta = 0.73$, 95% HDI [0.64, 0.83], $P(\beta > 0) > 0.999$), after controlling for the positive effects of salience ($\beta = 0.71$, 95% HDI [0.64, 0.79], $P(\beta > 0) > 0.999$) and center proximity ($\beta = 0.30$, 95% HDI [0.23, 0.36], $P(\beta > 0) > 0.999$). Monkey I showed a similar pattern, with meaning acting as a strong predictor of fixation ($\beta = 0.33$, 95% HDI [0.27, 0.39], $P(\beta > 0) > 0.999$) alongside salience ($\beta = 0.23$, 95% HDI [0.17, 0.29], $P(\beta > 0) > 0.999$) and center proximity ($\beta = 0.45$, 95% HDI [0.38, 0.51], $P(\beta > 0) > 0.999$). To interpret the magnitude of these effects, we converted the model coefficients into predicted fixation probabilities. For a region with average features, the baseline probability of fixation was approximately 51% for both animals. Increasing the meaning by 1 standard deviation (SD) raised this probability to 68% in Monkey V and 59% in Monkey I. In comparison, increasing salience by 1 SD yielded probabilities of 67% in Monkey V and 57% in Monkey I (Fig. 2A–D). Thus, meaning exerted an influence comparable to that of visual salience.

Relative contributions of meaning, salience, and center bias

We next quantified the overall explanatory power of the predictors (Table S3). For Monkey V, the Fixation model achieved a Bayesian R^2 of 0.26 (95% HDI [0.25, 0.28]), corresponding to a point-biserial correlation of $r = 0.48$ (95% HDI [0.479, 0.484]) between the model's predicted probabilities

Fig. 1. Experimental task and spatial predictors of fixation selection.

(A) Behavioral paradigm. Monkeys initiated each trial by fixating a central point (500 ms) before freely viewing a naturalistic scene for 5 s to receive a juice reward. (B) Example scene and corresponding spatial predictor maps. The top-left panel shows an example scene with fixation locations (green dots) and randomly sampled non-fixated regions (black dots). Together, these locations provide an account of regions that did and did not capture the subject’s attention. Each location was used to compute a mean value from the meaning, salience, and center proximity maps across an approximately 3-degree window, illustrated by the circles around example fixated (green) and non-fixated (black) locations. These spatial maps respectively visualize the spatial distribution of semantic informativeness, low-level physical features, and the central fixation bias used to model fixation selection. (C) Predictor distributions for Monkey V. Density plots comparing the normalized (z-scored) values of meaning, salience, and center proximity at actual fixated locations (green) versus randomly sampled non-fixated locations (black). (D) Predictor distributions for Monkey I. Same as in (C).

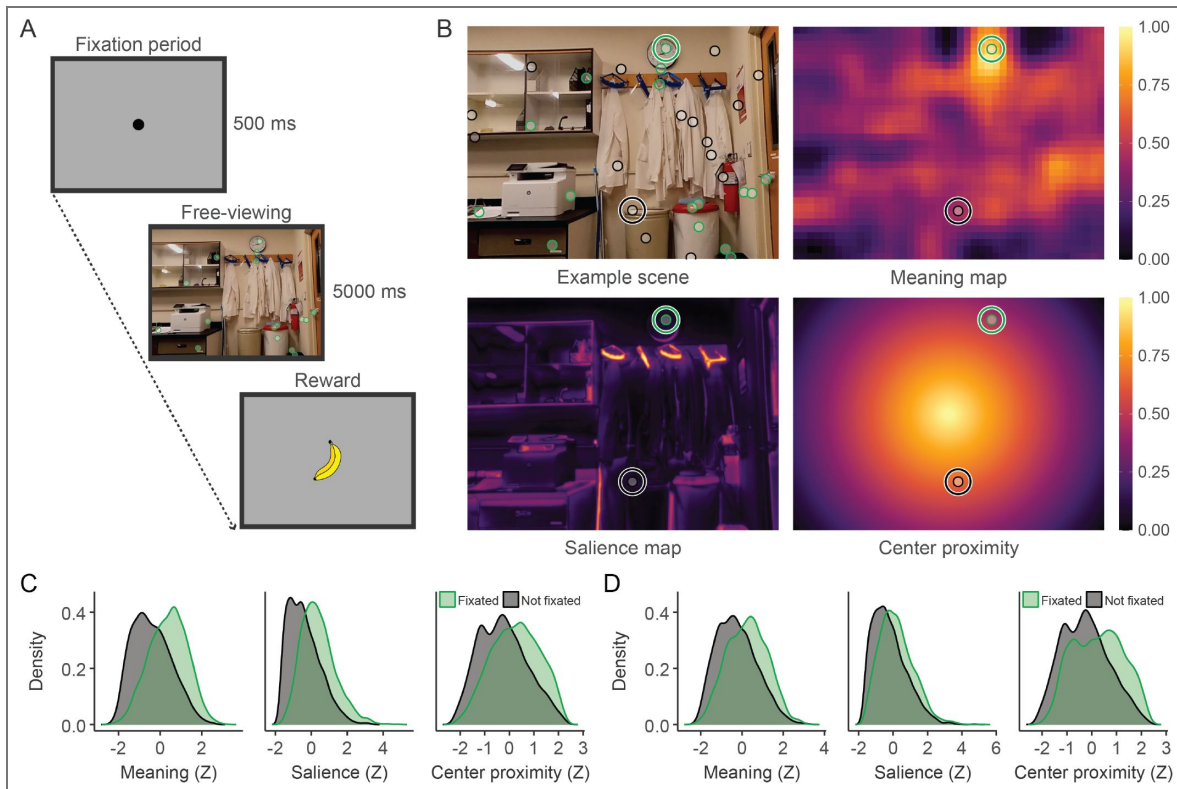


Table 1. Posterior estimates for the fixation model.

The fixed effects (main effects and interactions) and the scene-level random effect are presented for the fixation-likelihood GLMM. Columns display the posterior mean and standard deviation (SD) of β coefficients, the 95% highest density interval (HDI), and the probability of a positive effect ($P(\beta > 0 \mid \text{data})$). Probabilities close to 1 indicate strong evidence for a positive effect, whereas probabilities close to 0 indicate strong evidence for a negative effect. Values that would round to 1.00 or 0.00 are reported as $P(\beta > 0 \mid \text{data}) > 0.999$ or $P(\beta > 0 \mid \text{data}) < 0.001$ to avoid implying exact certainty.

	Monkey V				Monkey I			
	β Mean	β SD	95% HDI	$P(\beta > 0 \mid \text{data})$	β Mean	β SD	95% HDI	$P(\beta > 0 \mid \text{data})$
<i>Fixed effects</i>								
Intercept	0.02	0.04	[-0.06, 0.11]	0.71	0.05	0.03	[-0.01, 0.11]	0.95
Meaning	0.73	0.05	[0.64, 0.83]	> 0.999	0.33	0.03	[0.27, 0.39]	> 0.999
Saliency	0.71	0.04	[0.64, 0.79]	> 0.999	0.23	0.03	[0.17, 0.29]	> 0.999
Center proximity	0.30	0.03	[0.23, 0.36]	> 0.999	0.45	0.03	[0.38, 0.51]	> 0.999
Meaning \times saliency	-0.17	0.04	[-0.25, -0.09]	< 0.001	-0.08	0.03	[-0.14, -0.02]	0.01
Meaning \times center proximity	0.11	0.04	[0.04, 0.19]	> 0.999	-0.09	0.03	[-0.15, -0.02]	< 0.001
Saliency \times center proximity	0.08	0.04	[0.00, 0.16]	0.98	0.00	0.03	[-0.07, 0.06]	0.49
<i>Random effects</i>								
Scene sigma	0.43	0.06	[0.33, 0.55]	> 0.999	0.03	0.02	[0.00, 0.08]	> 0.999

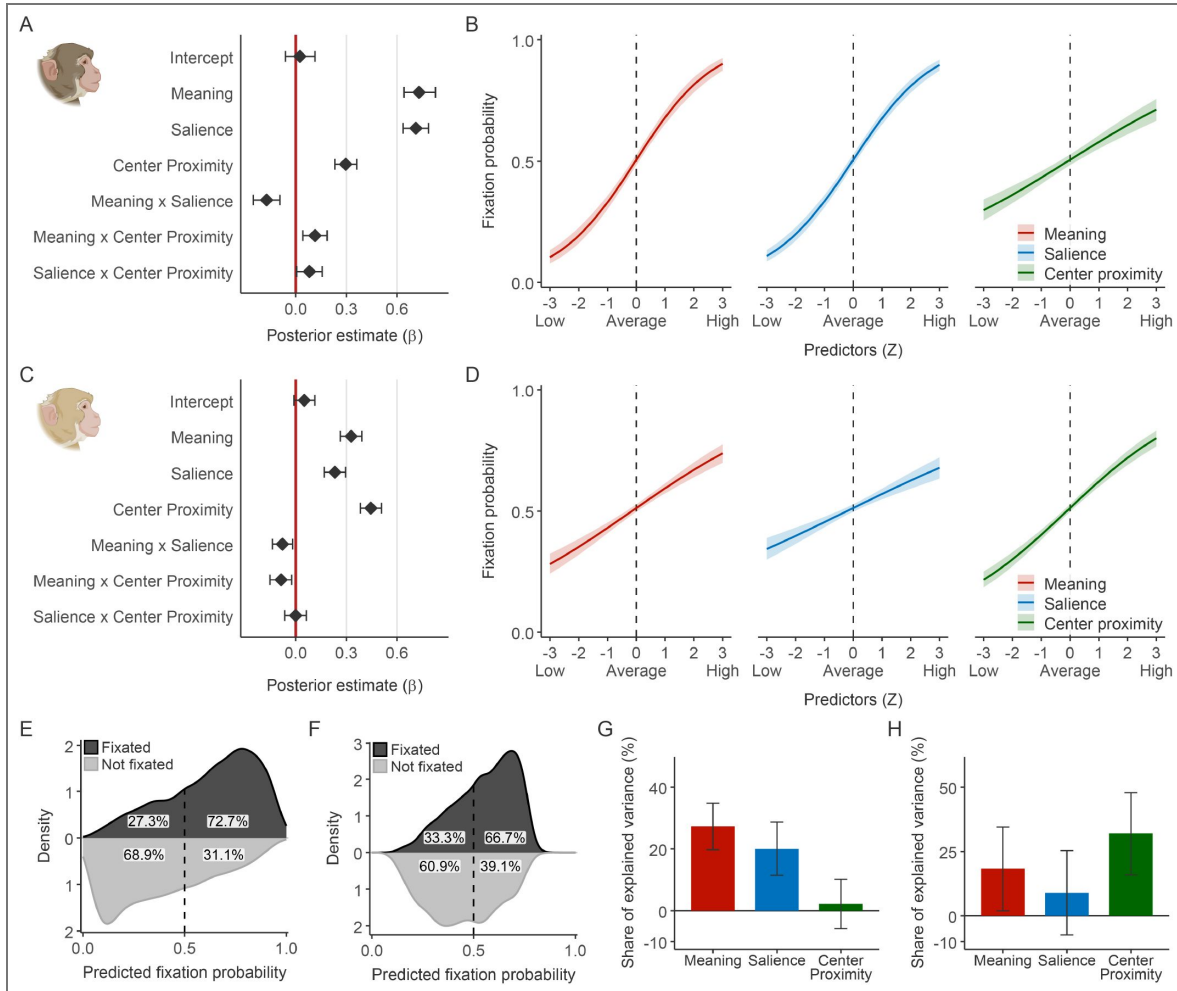


Fig. 2. Scene meaning is a robust predictor of fixation selection.

(A, C) Posterior estimates of the fixed effects for Monkey V (A) and Monkey I (C). Points represent the posterior mean of β coefficients, and error bars indicate the 95% highest density interval (HDI). The vertical red line marks 0 (no effect). Predictors are considered reliable if their 95% HDI excludes 0. (B, D) Predicted fixation probability as a function of the three main spatial predictors for Monkey V (B) and Monkey I (D). Curves visualize the marginal effects of meaning (red), saliency (blue), and center proximity (green) as they vary from -3 to $+3$ standard deviations (z-scores), while holding other predictors constant at their means (0). Solid lines indicate the posterior mean prediction, and shaded areas represent the 95% HDI. (E, F) Model classification performance for Monkey V (E) and Monkey I (F). Density distributions show the model's predicted probabilities for observed fixations (dark gray) and non-fixated locations (light gray). The vertical dashed line represents the classification threshold (0.5), and the percentages indicate the proportion of observations in each class that fall above or below the threshold. (G, H) Relative importance of spatial predictors for Monkey V (G) and Monkey I (H). Bars represent the unique share of explained variance (ΔR^2) attributable to meaning (red), saliency (blue), and center proximity (green). Error bars indicate the 95% highest density interval (HDI).

and the observed fixation behavior. In terms of classification accuracy, the model successfully identified 73% of fixated locations and 69% of non-fixated locations (Fig. 2E). For Monkey I, the model provided moderate predictive power, with a Bayesian R^2 of 0.11 (95% HDI [0.10, 0.13]) and a point-biserial correlation of $r = 0.33$ (95% HDI [0.329, 0.332]), correctly identifying 67% of fixated locations and 61% of non-fixated locations (Fig. 2F). To place these model results in a behavioral context, we estimated how consistently each monkey fixated the same locations when viewing the same scenes a second time (159 scenes for Monkey V; 146 scenes for Monkey I). For Monkey V, the mean scene-wise correlation between the two viewings was 0.43, corresponding to a mean scene-wise R^2 of 0.23. For Monkey I, the mean scene-wise correlation was 0.33, corresponding to a mean scene-wise R^2 of 0.15. These values provide a behavioral benchmark for the amount of fixation structure that is reproducible within subjects across viewings. The predictive performance of the GLMM was comparable to this level of behavioral reproducibility for both monkeys, indicating that the model captured a substantial portion of the fixation structure that is reliably expressed across repeated viewings.

To assess the relative importance of each map, we computed the incremental explained variance (ΔR^2) by refitting reduced models that excluded each predictor and all associated interaction terms, and calculated each predictor's relative percentage share of the total explained variance (% ΔR^2) (Fig. 2G-H). In Monkey V, meaning accounted for the largest unique share of variance ($\Delta R^2 = 0.07$, 95% HDI [0.05, 0.09]), representing 27.3% (95% HDI [19.7%, 34.8%]) of the model's total explanatory power. Saliency also provided a substantial unique contribution ($\Delta R^2 = 0.05$, 95% HDI [0.03, 0.07]; % $\Delta R^2 = 20.0\%$, 95% HDI [11.4%, 28.7%]), while center proximity contributed negligible unique variance ($\Delta R^2 = 0.006$, 95% HDI [-0.01, 0.03]; % $\Delta R^2 = 2.2\%$, 95% HDI [-5.8%, 10.2%]) (Fig. 2G). Monkey I exhibited a different weighting of predictors. Center proximity was the strongest unique predictor ($\Delta R^2 = 0.04$, 95% HDI [0.02, 0.05]), explaining 32.0% (95% HDI [15.9%, 47.8%]) of the total explained variance. Nevertheless, meaning remained a key factor relative to saliency, contributing 18.4% (95% HDI [2.0%, 34.5%]) of the model's uniquely explained variance ($\Delta R^2 = 0.02$, 95% HDI [0.002, 0.04]). In contrast, the unique contribution of saliency was statistically uncertain ($\Delta R^2 = 0.01$, 95% HDI [-0.01, 0.03]; % $\Delta R^2 = 9.0\%$, 95% HDI [-7.4%, 25.3%]) (Fig. 2H), suggesting that, for this subject, saliency offered little predictive gain beyond what was already captured by meaning and center bias. In summary, these variance-partitioning analyses indicate that, across both monkeys, meaning-based guidance exerts a robust influence on overt attention while accounting for visual saliency and oculomotor bias.

Attention prioritizes high-meaning regions, with saliency's role increasing as meaning decreases

We next examined whether meaning and saliency combine additively to guide fixation selection or instead compete with each other. An additive account predicts independent contributions of meaning and saliency to fixation selection. A competitive account, in contrast, suggests integration within a shared neural representation, such as a priority map (Bisley & Goldberg, 2010; Fecteau & Munoz, 2006), in which stronger input from one signal suppresses the contribution of the other. We observed a consistent negative interaction between meaning and saliency in both monkeys (Monkey V: $\beta = -0.17$, 95% HDI [-0.25, -0.09], $P(\beta > 0) < 0.001$; Monkey I: $\beta = -0.08$, 95% HDI [-0.14, -0.02], $P(\beta > 0) = 0.01$; Table 1). Thus, when scene meaning was high, fixation probability remained consistently high regardless of visual saliency. In contrast, when scene meaning was low, visual saliency strongly modulated fixation probability.

Model-based predictions further quantified this pattern (Fig. 3). We calculated the change in fixation probability caused by shifting image saliency from low (-2 SD) to high (+2 SD) at different levels of scene meaning. At low meaning (-2 SD), increasing saliency from low to high raised fixation probability by 63% in Monkey V (95% HDI [53%, 73%]) and 34% in Monkey I (95% HDI [23%, 46%]). In contrast, at high meaning (+2 SD), this change was reduced to 22% (95% HDI [11%, 34%]) for Monkey V and 7% (95% HDI [-5%, 18%]) for Monkey I. These results are inconsistent with purely additive integration. Instead, meaning and saliency appear to interact competitively: attention is directed to high-meaning regions regardless of visual saliency, with saliency playing

an increasing role as meaning decreases. Overall, these findings suggest a model in which meaning and salience are integrated within a shared selection mechanism rather than contributing independently to attentional guidance.

Scene familiarity increases the exploration of low-meaning regions

We further asked whether meaning-based attentional guidance is context-dependent or remains stable across conditions. To test this, we examined whether long-term familiarity with an environment modulates meaning-based guidance. Three competing hypotheses motivated this analysis. First, familiarity might sharpen meaning-based guidance, such that prior knowledge of a scene's layout drives an even stronger prioritization of high-meaning regions. Second, familiarity might broaden visual exploration, such that knowing the environment reduces the need to exclusively target high-meaning regions, allowing attention to be distributed more frequently to less meaningful areas. Third, the strength of meaning-based attentional guidance might be a fixed trait, operating identically regardless of the observer's prior experience with the environment.

To distinguish among these hypotheses, we extended the Fixation model to include familiarity (familiar vs. unfamiliar environment) and its interactions with meaning and salience (Table S1; Familiarity model). In both monkeys, we observed a significant increase in fixation probability when viewing scenes from familiar environments (Monkey V: $\beta = 0.47$, 95% HDI [0.31, 0.64], $P(\beta > 0) > 0.999$; Monkey I: $\beta = 0.13$, 95% HDI [0.01, 0.26], $P(\beta > 0) > 0.98$) (Fig. 4B–C). In addition, we observed a significant negative interaction between meaning and familiarity (Monkey V: $\beta = -0.19$, 95% HDI [-0.34, -0.02], $P(\beta > 0) = 0.01$; Monkey I: $\beta = -0.19$, 95% HDI [-0.32, -0.06], $P(\beta > 0) < 0.001$) (Fig. 4B–C). Crucially, the probability of fixating high-meaning regions remained consistently high regardless of scene familiarity. However, familiar scenes increased the likelihood of fixations on less meaningful areas, supporting the hypothesis that familiarity broadens visual exploration (Fig. 4D–E). In contrast, low-level salience remained stable across scene contexts. Neither monkey showed a reliable interaction between salience and familiarity (Monkey V: $\beta = -0.09$, 95% HDI [-0.24, 0.07]; Monkey I: $\beta = 0.09$, 95% HDI [0.23, 0.03]) (Fig. 4B, C, F, and G). Evidence for an interaction between center proximity and familiarity was present in one monkey but not the other (Monkey V: $\beta = 0.34$, 95% HDI [0.20, 0.47], $P(\beta > 0) > 0.999$; Monkey I: $\beta = 0.10$, 95% HDI [-0.03, 0.24], $P(\beta > 0) = 0.94$) (Fig. 4B, C, H, and I). These results indicate that familiarity selectively modulated meaning-based attentional guidance while leaving bottom-up feature guidance largely unchanged.

We further quantified the unique contribution of each feature by calculating the drop in incremental Bayesian R^2 (ΔR^2) when removing meaning or salience from the model. In Monkey V, meaning accounted for 38% more unique variance in unfamiliar relative to familiar environments (unfamiliar $\Delta R^2 = 0.08$, 95% HDI [0.06, 0.11]; familiar $\Delta R^2 = 0.06$, 95% HDI [0.03, 0.09]). Monkey I showed the same directional effect, with the explanatory power of meaning nearly tripling in unfamiliar scenes (unfamiliar $\Delta R^2 = 0.04$, 95% HDI [-0.01, -0.06]; familiar $\Delta R^2 = 0.01$, 95% HDI [-0.01, 0.04]). In contrast, the contribution of salience showed little evidence of contextual modulation. For Monkey V, salience explained comparable variance in unfamiliar and familiar scenes (unfamiliar $\Delta R^2 = 0.05$, 95% HDI [0.03, 0.08]; familiar $\Delta R^2 = 0.05$, 95% HDI [0.02, 0.08]). In Monkey I, the point estimate was slightly higher in unfamiliar scenes, but both HDIs overlapped zero (unfamiliar $\Delta R^2 = 0.02$, 95% HDI [-0.01, 0.04]; familiar $\Delta R^2 = 0.01$, 95% HDI [-0.02, 0.03]), indicating weak evidence for reliable differences. In addition, familiarity modulated center bias in one subject. Monkey V exhibited increased center bias in familiar environments ($\beta = 0.34$, 95% HDI [0.20, 0.47]), whereas this effect was not significant in Monkey I ($\beta = 0.10$, 95% HDI [0.03, 0.24], $P(\beta > 0) = 0.94$). Together, these results show a common pattern across subjects: high-meaning regions consistently capture attention, familiarity broadens exploration by increasing fixations on less meaningful areas, and salience-driven selection remains largely stable across contexts.

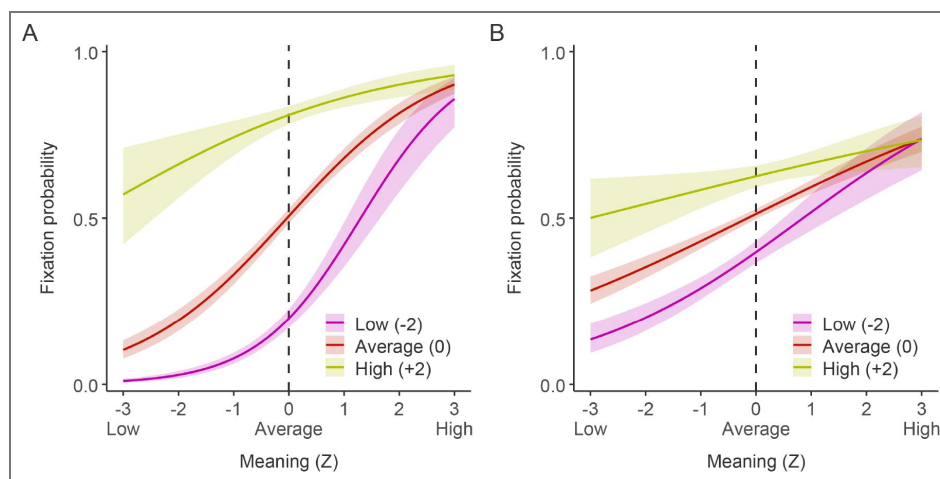


Fig. 3. Attention is directed to high-meaning regions, with salience playing an increasing role as meaning decreases.

(A, B) Visualization of the meaning \times salience interaction for Monkey V (A) and Monkey I (B). Curves show the predicted fixation probability as a function of meaning at three representative levels of image salience: low (-2 SD, purple), average (0 SD, red), and high ($+2$ SD, yellow). Solid lines represent the posterior mean prediction, and shaded areas indicate the 95% highest density interval (HDI). The vertical dashed line marks the average meaning value (0). In both animals, when scene meaning is high, fixation probability remains consistently high with little influence of visual salience. In contrast, when scene meaning is low, visual salience strongly modulates fixation probability.

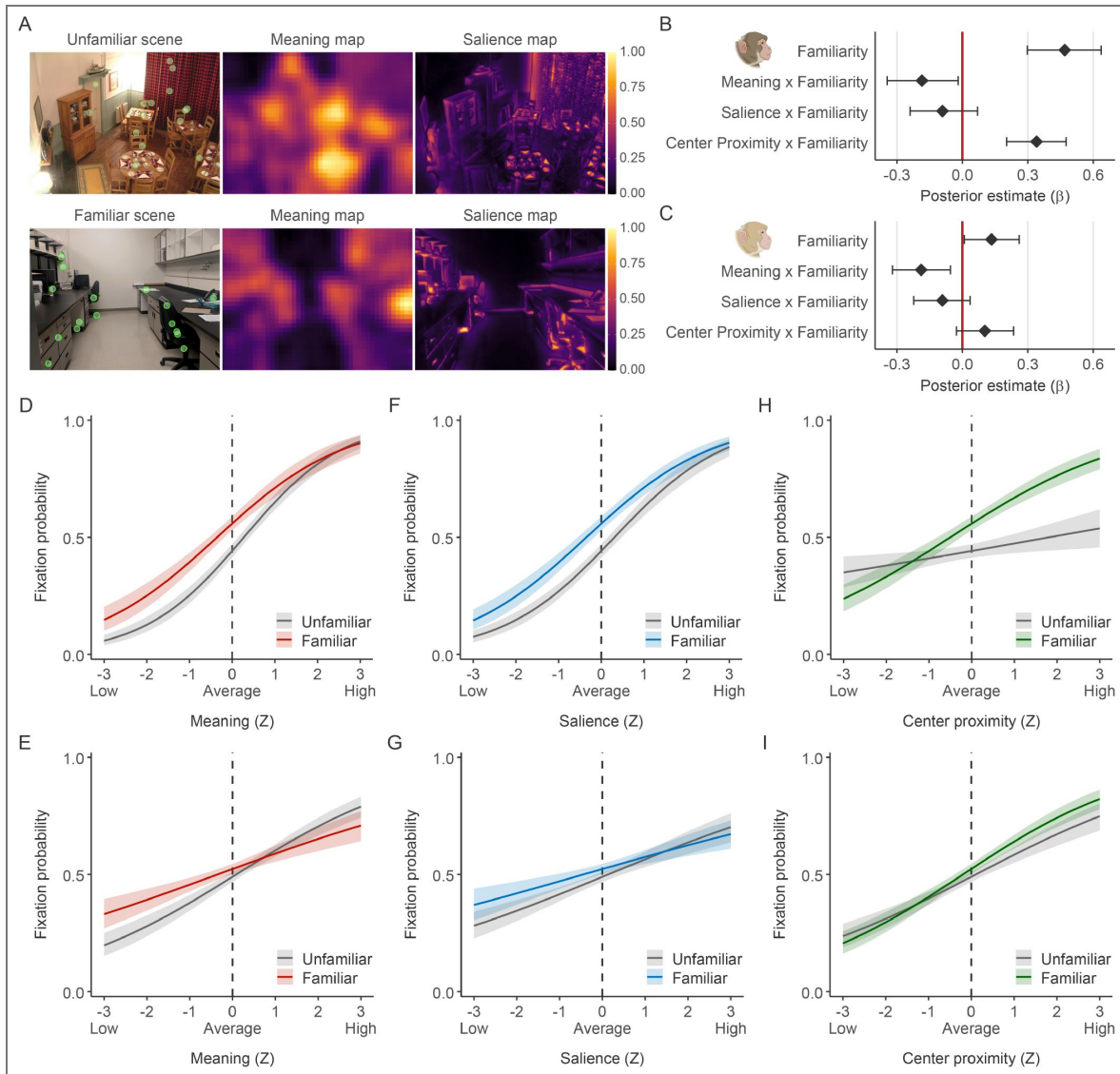


Fig. 4. Attention is directed to high-meaning regions, with familiarity playing an increasing role as meaning decreases.

(A) Example unfamiliar and familiar scenes. The top row displays an unfamiliar scene with fixations (green dots) alongside its meaning and salience maps. The bottom row displays a familiar scene with its corresponding maps. (B, C) Posterior estimates of familiarity effects for Monkey V (B) and Monkey I (C). Points represent the posterior mean of the interaction coefficients between familiarity and the spatial predictors; error bars indicate the 95% HDI. The vertical red line marks 0. Note the negative interaction between meaning and familiarity in both animals. (D–I) Predicted fixation probability as a function of spatial predictors for unfamiliar and familiar scenes. Top row (D, F, H) shows results for Monkey V; bottom row (E, G, I) shows results for Monkey I. Curves visualize fixation probability as a function of meaning (D, E), salience (F, G), and center proximity (H, I) for familiar (colored lines) versus unfamiliar (gray lines) scenes. Solid lines represent the posterior mean prediction, and shaded areas represent the 95% HDI. In both animals, the probability of fixating high-meaning regions remains consistently high regardless of scene familiarity. However, as scene regions become less meaningful, they receive increasingly more fixations in familiar compared to unfamiliar scenes.

Attentional engagement selectively strengthens meaning-based guidance

Attention not only determines where to look (spatial selection) and what to prioritize (object- or feature-based selection), but also regulates the magnitude of engagement, that is, the intensity and duration of visual sampling itself. Here, we asked whether scene-by-scene variations in engagement modulate how meaning and salience guide fixation selection. We quantified engagement as the total fixation duration accumulated within a scene during each trial. Engagement varied substantially across scenes for both Monkey V (mean = 3586 ms, SD = 547 ms; range = 2015–4463 ms) and Monkey I (mean = 3434 ms, SD = 474 ms; range = 2087–4205 ms) (Fig. 5A–B), indicating variability in sampling duration. Importantly, this trial-by-trial variance in engagement was independent of scene familiarity. Total engagement duration did not differ significantly between unfamiliar and familiar scenes for either subject (Wilcoxon rank-sum test: $p = 0.78$ for Monkey V and $p = 0.70$ for Monkey I; Fig. S3). We extended the Fixation model to include interactions between engagement and the spatial predictors (Table S2; Engagement model).

Engagement selectively scaled the influence of scene meaning on attentional guidance while having no consistent effect on overall fixation likelihood (Monkey V: $\beta = -0.25$, 95% HDI [-0.33, -0.17], $P(\beta > 0) < 0.001$; Monkey I: $\beta = 0.05$, 95% HDI [-0.01, -0.11], $P(\beta > 0) = 0.94$) (Fig. 5C–D). Specifically, in both monkeys, the effect of meaning on fixation probability increased with higher engagement (Fig. 5C–F). For Monkey V, this interaction was substantial ($\beta = 0.30$, 95% HDI [0.22, 0.38], $P(\beta > 0) > 0.999$), indicating that meaning became a progressively stronger determinant of gaze during longer exploration (Fig. 5C and E). Monkey I showed a similar pattern, but the interaction was more modest in magnitude ($\beta = 0.08$, 95% HDI [0.01, 0.15], $P(\beta > 0) = 0.99$) (Fig. 5D and F). In contrast, salience showed no consistent modulation across animals. For Monkey I, salience increased modestly with engagement ($\beta = 0.07$, 95% HDI [0.00, 0.13], $P(\beta > 0) = 0.98$), whereas for Monkey V, salience was unaffected ($\beta = -0.02$, 95% HDI [-0.09, 0.07], $P(\beta > 0) = 0.35$) (Fig. 5D). These results indicate that increased attentional engagement selectively strengthens meaning-based guidance rather than uniformly enhancing all visual signals, suggesting that the extent of visual sampling modulates the influence of meaning on overt attention.

Discussion

In this study, we asked whether meaning-based guidance of overt attention is uniquely human or instead reflects a conserved property of the primate visual system. We found that scene meaning robustly guided fixation selection in rhesus macaques, with an influence comparable to that of low-level visual salience after accounting for spatial bias. Moreover, meaning and salience interacted competitively: attention prioritized meaningful regions irrespective of their salience, while visual salience emerged as a stronger guide for eye movements as meaning decreased. This pattern suggests that meaning and salience interact within a unified attentional selection mechanism rather than contributing as independent additive signals. Importantly, meaning-based guidance was not determined solely by image structure, but was also modulated by behavioral context. Monkeys fixated high-meaning regions at similarly high rates in both unfamiliar and familiar scenes, but familiarity broadened visual exploration by increasing the likelihood of fixating less meaningful areas as meaning decreased. In addition, attentional engagement selectively modulated meaning-based guidance: the overall influence of meaning strengthened during trials with greater attentional engagement. Together, these findings provide the first evidence that scene meaning guides overt attention during free-viewing in non-human primates. They suggest that meaning-based guidance operates alongside visual salience and center bias, thereby expanding current models of primate attention and providing a behavioral foundation for investigating its underlying neural substrates.

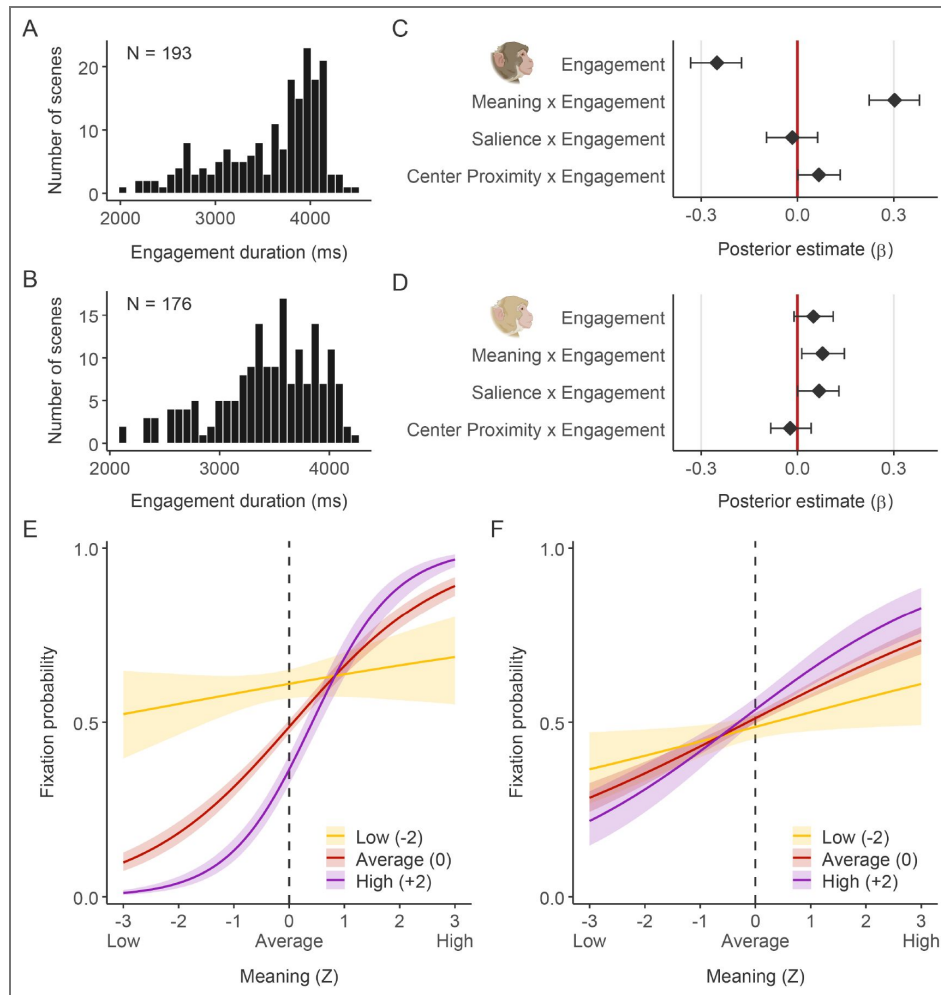


Fig. 5. Attentional guidance by meaning is enhanced during high engagement.

(A, B) Distribution of engagement duration (total fixation time per scene) for Monkey V (A) and Monkey I (B). Histograms show the variability in scene viewing time across trials. (C, D) Posterior estimates of engagement interactions for Monkey V (C) and Monkey I (D). Points represent the posterior mean of the interaction coefficients between engagement and the spatial predictors; error bars indicate the 95% HDI. The vertical red line marks 0. Note the positive interaction between meaning and engagement in both animals. (E, F) Visualization of the meaning \times engagement interaction for Monkey V (E) and Monkey I (F). Curves show the predicted fixation probability as a function of meaning at three representative levels of engagement: low (-2 SD, yellow), average (0 SD, red), and high ($+2$ SD, purple). Solid lines represent the posterior mean prediction, and shaded areas indicate the 95% HDI. In both animals, the slope of the meaning function is steepest when engagement is high (purple curve), indicating that the influence of meaning on gaze selection scales with the subject's level of scene engagement.

Meaning-based attentional guidance is a conserved property of the primate visual system

The present findings indicate that meaning-based guidance of attention reflects a conserved property of the primate visual system rather than a uniquely human strategy. To quantify local scene meaning, we used DeepMeaning, an image-computable model built on a pretrained vision–language transformer (Hayes & Henderson, 2025 [↗](#)). In this model, the transformer weights are frozen, and its multimodal image embeddings are used to train a cross-validated linear model that predicts dense human ratings of local scene meaning. Rather than relying on low-level image statistics, this framework captures structured visual representations of scene regions containing identifiable objects and informative relationships. Although trained on human ratings, the model's ability to predict macaque gaze behavior suggests that the spatial distribution of semantic informativeness it captures reflects representational mechanisms shared across primates.

Neurophysiological evidence supports this interpretation. The macaque inferotemporal (IT) cortex contains selective patches for ethologically relevant categories such as faces and bodies (Bao et al., 2020 [↗](#); Pinsk et al., 2005 [↗](#); Tsao et al., 2006 [↗](#); Vinken et al., 2025 [↗](#)), and neurons along the ventral visual hierarchy encode increasingly abstract object and scene features (Logothetis et al., 1995 [↗](#); Ponce et al., 2019 [↗](#); Riesenhuber & Poggio, 1999 [↗](#)). These hierarchical representations provide the neural substrate on which attentional circuits operate. Consistent with this organization, behavioral studies show that macaques preferentially orient toward socially relevant stimuli (Deaner et al., 2005 [↗](#); Machado et al., 2011 [↗](#); Solyst & Buffalo, 2014 [↗](#)), reward- or aversion-associated objects (Ghazizadeh et al., 2016 [↗](#)), and illusory faces in inanimate objects (Taubert et al., 2017 [↗](#)), demonstrating that categorical structure and behavioral relevance shape perceptual selection. Our results extend these category-specific findings to general scene viewing, demonstrating that the spatial distribution of informative visual structure guides attention even in the absence of explicit task demands.

Evidence from human development further supports this interpretation. Human infants preferentially fixate scene regions rated as high in meaning, and this effect strengthens across the first year of life (Oakes et al., 2024 [↗](#)). Because this prioritization emerges early in development, it suggests that attentional selection operates on object-level representations that develop through visual experience and cortical maturation. Demonstrating meaning-guided attention in macaques therefore provides a critical bridge between human behavioral findings and mechanistic circuit-level investigations in primates.

Meaning-based attentional guidance is modulated by familiarity and engagement

Despite its robust influence, the effect of meaning on attentional guidance varied with scene familiarity and behavioral context. Prior work has shown that gaze during scene viewing is shaped not only by the physical properties and content of images, but also by task demands (Castelhano et al., 2009 [↗](#); Navalpakkam & Itti, 2005 [↗](#); Yarbus, 1967 [↗](#)) and prior experience (Becker & Rasmussen, 2008 [↗](#); Li et al., 2018 [↗](#); Summerfield et al., 2006 [↗](#)). Similarly, our results suggest that the influence of meaning depends on prior experience, indicating that memory modulates attentional selection. Eye movements can reveal memory for previous experience and relational structure, even when explicit report is absent or incomplete (Hannula & Ranganath, 2009 [↗](#); Hannula et al., 2010 [↗](#); Schmidig et al., 2025 [↗](#)). More recently, Ramey et al. (2025) [↗](#) showed that episodic memory and semantic knowledge interact to guide eye movements, with the influence of memory depending on the availability and effectiveness of semantic guidance. In the present study, familiar scenes depicted environments with which the animals had extensive prior experience. These scenes increased overall fixation probability and broadened visual exploration toward less meaningful regions while preserving the prioritization of high-meaning regions. At a computational level, this pattern is broadly consistent with the exploration–exploitation framework from decision neuroscience, in which exploitation corresponds to continuing to prioritize known informative regions, whereas exploration corresponds to sampling a wider set of

regions within a familiar environment (Daw et al., 2006 [↗](#); Ebitz et al., 2018 [↗](#); Sutton, 1998 [↗](#)). Taken together, these findings suggest that long-term memory for an environment can modulate how attention is distributed within a scene while preserving the prioritization of high-meaning regions.

Behavioral context also shaped meaning-based guidance across trials. Attentional engagement can vary with multiple factors, including the animals' behavioral state (e.g., arousal) and the behavioral relevance of scene content. When engagement was higher, quantified by longer viewing durations across the whole images, fixation patterns more closely tracked the spatial distribution of meaning. In contrast, the influence of meaning weakened when engagement declined. This result suggests that meaning-based guidance is expressed more strongly when animals are behaviorally engaged with a scene, indicating that the influence of scene meaning on overt attention scales with the extent of visual sampling. Thus, these findings suggest that engagement modulates the extent to which meaning is expressed in overt attention, strengthening the meaning-based guidance of eye movements when animals are more engaged with the scene.

Neural substrates for meaning-based attentional guidance

Our findings have important implications for the neural circuits that guide overt visual attention. Dorsal frontoparietal and subcortical structures, including lateral intraparietal cortex (LIP), frontal eye fields (FEF), and superior colliculus, are thought to implement priority maps that determine saccadic targets by integrating both visual salience and behavioral goals (Buschman & Kastner, 2015 [↗](#); Moore & Zirnsak, 2017 [↗](#); Soyuhos et al., 2026 [↗](#); Xia et al., 2024 [↗](#)). These regions combine inputs from distributed cortical areas into a unified selection signal that governs eye movements. Visual salience signals, derived from image-based feature contrasts, are known to contribute to these priority maps. Our results indicate that scene meaning provides an additional and robust source of guidance. Critically, the negative interaction between meaning and visual salience suggests that these influences are not simply additive, but instead compete within a shared selection mechanism. High scene meaning dominated attentional priority regardless of visual salience, whereas salience increasingly guided selection primarily in regions where meaning was low. This pattern accords with normalization-based models of attentional control (Reynolds & Heeger, 2009 [↗](#)), in which distinct relevance signals are integrated within a common priority representation and compete via divisive normalization, biasing selection toward the strongest input.

A plausible source of meaning-related input lies in the ventral visual and scene-selective cortex. Regions such as parahippocampal place area (PPA), occipital place area (OPA/TOS), and retrosplenial cortex (RSC) encode spatial layout, object ensembles, and contextual structure (Epstein & Baker, 2019 [↗](#); Malcolm et al., 2016 [↗](#); Soyuhos et al., 2025 [↗](#)), while neurons along the ventral visual hierarchy represent increasingly abstract object and scene features beyond low-level image statistics (Bao et al., 2020 [↗](#); Riesenhuber & Poggio, 1999 [↗](#); Vinken et al., 2025 [↗](#)). These hierarchical representations may support the computation of scene meaning—the spatial distribution of recognizable and behaviorally informative structure across a scene (Henderson et al., 2020 [↗](#)). Within this framework, priority maps integrate visual salience and meaning-related signals within a common competitive architecture to determine overt visual selection. Meaning-based guidance may therefore arise from the incorporation of structured scene representations into oculomotor priority circuits during natural viewing.

Materials and methods

Subjects

The experiments involved two healthy male rhesus macaques (*Macaca mulatta*): Monkey V and Monkey I. At the time of data collection, Monkey V was 14 years old and weighed 13.3 kg, and Monkey I was 14 years old and weighed 13.8 kg. Both animals were sourced from the California National Primate Research Center and housed in standard primate facilities with environmental enrichment, in accordance with institutional guidelines. The monkeys were maintained on a

controlled fluid access schedule to ensure motivation for the juice reward. Daily health and hydration status were monitored. All procedures were approved by the University of California, Davis Institutional Animal Care and Use Committee (Protocol #24082) and were conducted in accordance with the NIH Guide for the Care and Use of Laboratory Animals.

Experimental setup and eye tracking

Animals were seated in a primate chair with the head stabilized using a surgically implanted head-post. The chair was positioned in a sound-attenuated, dimly lit booth facing a computer monitor. Visual stimuli were presented on a 32" LCD monitor (Display++, Cambridge Research Systems) with a resolution of 1920×1080 pixels and a refresh rate of 120 Hz. The physical width of the display was 70 cm, and the viewing distance was 80 cm from the subject's eyes. The screen background was a uniform mid-grey (RGB: 128, 128, 128). Stimulus presentation and experimental control were implemented in MATLAB R2023a (MathWorks) using Psychtoolbox-3. Juice reward was delivered through a tube positioned at the subject's mouth and controlled by a syringe pump (NE-4000-US, New Era Pump Systems Inc.).

Eye position was recorded from the right eye using a video-based infrared eye tracker (EyeLink 1000 Plus, SR Research) sampling at 1000 Hz. The tracker was configured to record gaze position, pupil area, and event markers for fixations, saccades, and blinks using the manufacturer's standard settings. At the beginning of each session, a horizontal-vertical 9-point calibration was performed on a mid-grey background using the standard EyeLink calibration routine, and the display coordinates in the EyeLink system were matched to the Psychtoolbox display window. Raw EyeLink data files (.edf) were imported into MAT-LAB R2023a using the Edf2Mat toolbox (developed by Etter and Biedermann, University of Zurich). For each trial, fixation events (onset time, duration, and x - y coordinates) were extracted from the EDF files. The first fixation of each trial was excluded because it was constrained by the central fixation requirement at trial onset.

Stimuli, design, and behavioral task

Stimuli consisted of 200 naturalistic indoor scenes (1440 × 1080 pixels) presented in a randomized design. We did not include humans or other animals in any image to keep the scenes emotionally neutral. The stimulus set included 100 unfamiliar scenes and 100 familiar scenes. Unfamiliar scenes depicted environments the monkeys had not directly experienced prior to the experiment, such as restaurants, bedrooms, kitchens, and offices. Familiar scenes were photographs taken in the monkeys' own living and experimental environment, including vivarium housing and laboratory spaces, and therefore reflected locations with which the animals had long-term experience. Trials were interleaved such that unfamiliar and familiar scenes appeared in a random order, preventing block effects or systematic changes in arousal over the session.

Each trial began with a central black fixation dot (0.5° radius) on a mid-grey background. The fixation window was a $6^{\circ} \times 6^{\circ}$ square centered on the dot ($\pm 3^{\circ}$ horizontally and vertically). In the first phase, the subject had up to 5000 ms to acquire fixation; once acquired, it had to be maintained continuously for 500 ms. If the subject failed to hold fixation, the trial was aborted and repeated after a 1000 ms inter-trial interval (ITI). In the second phase, the fixation dot disappeared and a naturalistic scene was presented at the center of the screen for 5000 ms. During this viewing period, the subject was free to move its eyes and explore the scene (Fig. 1A). In the third phase, a juice reward was delivered immediately following the viewing period, followed by a 1000 ms ITI.

Meaning, salience, and center proximity maps

For each scene, we generated three spatial predictor maps at the native image resolution of 1440 × 1080 pixels (Fig. 1B). A meaning map quantified the spatial distribution of semantic content using a transformer model called DeepMeaning (Hayes & Henderson, 2025). The DeepMeaning model combines features from a Contrastive Captioner (CoCa) vision-language transformer pretrained on billions of image-text pairs with linear weights trained to predict human meaning maps. A salience map was computed for each scene using the classic Itti-Koch model (Itti et al., 1998) as implemented in the pySaliencyMap package, which combines multiscale feature maps

for luminance contrast, color, and orientation into a single salience map. We chose this model rather than graph-based alternatives to minimize central bias that would otherwise be highly collinear with our center proximity predictor.

A center proximity map represented spatial bias toward the image center and was operationalized as the normalized inverse Euclidean distance from the image center. For a pixel located at coordinates (x, y) , it was defined as:

$$\text{CenterProximity}(x, y) = 1 - \frac{d(x, y)}{d_{\max}} \quad (1)$$

where $d(x, y)$ is the Euclidean distance from the image center and d_{\max} is the maximum distance from the center to any corner of the image.

Bayesian generalized linear mixed model

Prior to model fitting, we applied exclusion criteria to ensure that only trials reflecting active exploration were included. A trial (scene) was excluded from analysis if the subject made fewer than 10 fixations or if the total fixation duration was less than 2000 ms. This resulted in the exclusion of 7 scenes for Monkey V and 24 scenes for Monkey I, leaving a final dataset of 193 and 176 scenes, respectively. The final dataset included a total of 3,140 fixations for Monkey V and 2,943 fixations for Monkey I (Fig. S1). To model fixation probability, we constructed balanced datasets consisting of fixated locations and matched nonfixated locations (Hayes & Henderson, 2022a; Oakes et al., 2024). For each real fixation (Fixated = 1), we generated one corresponding non-fixated sample (Fixated = 0). To ensure that non-fixated samples represented truly unattended locations, they were drawn from random coordinates within the image boundaries, constrained to be at least approximately 3 degrees of visual angle away from all fixation locations in that scene (Fig. 1B). Prior to model fitting, we verified that the predictors for the GLMM were independent. Variance inflation factors (VIF) and pairwise Pearson correlations computed on the fixated and matched non-fixated locations indicated negligible multicollinearity across all predictors for both subjects (maximum VIF = 1.33; Fig. S2).

All models were fitted using `bambi` (version 0.13.0) with a Bernoulli likelihood and a logit link function. We used the default priors provided by `bambi`, which consist of weakly informative priors on regression coefficients and hierarchically structured priors on random effects. Each model was estimated using four Markov chain Monte Carlo (MCMC) chains with 6000 draws per chain (2000 warm-up, 4000 sampling), resulting in 16,000 posterior samples. All models included a random intercept for Scene to capture baseline heterogeneity in fixation likelihood across images. Posterior predictions were obtained by evaluating the fitted models on grids of predictor values and averaging over posterior samples.

To ensure that our predictors (including spatial maps, engagement, and familiarity) captured distinct sources of variance estimation, we inspected the correlations between their posterior parameter estimates. Multicollinearity was minimal across all three models. Pairwise correlations between fixed effects were consistently low for both Monkey V (maximum $r = 0.37$) and Monkey I (maximum $r = 0.38$; Table S4). Additionally, all parameters demonstrated stable convergence ($\hat{R} = 1.00$) and high effective sample sizes (ESS > 8,000) across all models (Table S5).

Model specifications and performance assessment

For each monkey, continuous predictors (meaning, salience, and center proximity) were z-scored across all samples so that each predictor had a mean of 0 and a standard deviation of 1 before model fitting. We fit three specific models to address our research questions. First, to test the independent and interactive effects of scene features, we fit a Fixation Model including meaning, salience, and center proximity, along with their interactions:

$$\text{Fixated or not} \sim \text{meaning} * \text{saliency} * \text{center proximity} + (1 | \text{Scene}) \quad (2)$$

Second, to examine how long-term environmental familiarity alters guidance, we fit a Familiarity Model by adding a categorical predictor for familiarity. Familiarity encoded whether the scene was familiar (Familiarity = 1) or unfamiliar (Familiarity = 0) and was included in the interaction terms.

$$\begin{aligned} \text{Fixated or not} \sim & \text{meaning} * \text{saliency} * \text{center proximity} \\ & * \text{Familiarity} + (1 | \text{Scene}) \end{aligned} \quad (3)$$

Finally, to investigate how guidance is modulated by the viewer's state, we calculated an Engagement score for each trial. Engagement was defined as the total duration of all fixations accumulated during the 5 s free-viewing period, capturing trial-to-trial fluctuations in the intensity of visual exploration. This continuous variable was z-scored and included as an interaction term in the Engagement Model.

$$\begin{aligned} \text{Fixated or not} \sim & \text{meaning} * \text{saliency} * \text{center proximity} \\ & * \text{Engagement} + (1 | \text{Scene}) \end{aligned} \quad (4)$$

To quantify the explanatory power of the fixed effects in each model, we calculated the Bayesian R^2 using the bayes R2 function in brms (Gelman et al., 2019). We reported the Bayesian R^2 for the fixed effects, which estimates the proportion of variance in the observed outcome explained by the model's expected predictions, excluding random scene intercepts. We also calculated the predictive correlation between the model's posterior expected predicted probabilities and the observed binary outcome, computed for each posterior draw and summarized using the mean and 95% HDI. To isolate the incremental predictive contribution of each scene feature, we performed a model comparison analysis. We calculated the change in Bayesian R^2 (ΔR^2) by comparing the full Fixation Model to reduced models in which a specific predictor and all its associated interactions were removed. The distribution of ΔR^2 values was computed from the posterior samples, providing a direct estimate of the incremental explanatory contribution of meaning, saliency, and center proximity, respectively. We also calculated each predictor's unique contribution as a percentage of the full model's explanatory power (% ΔR^2). Both ΔR^2 and % ΔR^2 distributions were summarized using their posterior means and 95% HDIs.

To estimate a behavioral benchmark for model performance, we quantified within-subject fixation reliability. The same scenes were presented to the subjects a second time in a subsequent block. To ensure comparability with the GLMM analysis, we applied the same trial-exclusion criteria used in the main fixation models: scenes were excluded if total fixation duration was less than 2000 ms or if fewer than 10 fixations were recorded. The repeatability analysis was then restricted to scenes that passed these criteria in both blocks, resulting in 159 repeated scenes for Monkey V and 146 repeated scenes for Monkey I. For each repeated scene, we generated smoothed fixation density maps for the initial and repeated viewings by binning fixation locations and applying Gaussian smoothing with a full width at half maximum of 3 degrees of visual angle. We then calculated the squared Pearson correlation between the two density maps for each scene. The average squared correlation across repeated scenes provided a model-free estimate of within-subject repeatability, which we used as a benchmark for the amount of fixation structure that was reproducible across viewings. This intra-observer consistency provided a behavioral reference point against which the explanatory power of the spatial predictors was interpreted.

Data availability

All data and analysis code required to reproduce the results reported in this study will be made publicly available upon acceptance of the manuscript. Raw eye-tracking data (.edf files) and fixation data (.csv files) will be deposited on the Open Science Framework (OSF). The MATLAB

scripts used for stimulus presentation, the Python (version 3.9) pipeline used for preprocessing, and the R (version 4.4.1) code used for GLMM fitting will be hosted in a public GitHub repository.

Acknowledgements

This work was supported by the National Science Foundation under grant 2152260 (NSF NRT NeuralStorm) (to O.S., W.H., and B.S.), National Institutes of Health grants EY014924 (to X.C.) and R01EY027792 (to J.M.H.), and the Brain and Behavior Research Foundation (to X.C.).

Additional information

Author contributions statement

Study conceptualization: X.C., O.S., J.M.H., and T.R.H.; Experimental design: O.S. and X.C.; Data collection: O.S., T.P.H., B.S., and W.H.; Methodology: O.S., X.C., and T.R.H.; Exploratory analysis: O.S., T.R.H., and W.H.; Formal analysis: O.S.; Visualization: O.S.; Writing – original draft: O.S. and X.C.; Writing – review & editing: O.S., X.C., J.M.H., and T.R.H.; Funding acquisition: X.C., O.S., W.H., and J.M.H.

Funding

Funder	Grant reference number	Author
National Science Foundation (NSF)	2152260	Brinda Sevak Wenqing Hu Orhan Soyuhos
HHS National Institutes of Health (NIH)	EY014924	Xiaomo Chen
HHS National Institutes of Health (NIH)	R01EY027792	John Henderson
Brain and Behavior Research Foundation (BBRF)		Xiaomo Chen

Author ORCID iDs

Orhan Soyuhos:  <https://orcid.org/0000-0002-2858-3979>

Xiaomo Chen:  <https://orcid.org/0000-0002-6142-9981>

Additional files

[Supplementary information.](#) 

References

- Bao P**, She L, McGill M, Tsao DY (2020) A map of object space in primate inferotemporal cortex. *Nature* **583**:103-108 <https://doi.org/10.1038/s41586-020-2350-5> | [PubMed](#)
- Becker MW**, Rasmussen IP (2008) Guidance of attention to objects and locations by long-term memory of natural scenes. *Journal of Experimental Psychology* **34**:1325-1338 <https://doi.org/10.1037/a0013650> | [PubMed](#)
- Bichot NP**, Heard MT, DeGennaro EM, Desimone R (2015) A Source for Feature-Based Attention in the Prefrontal Cortex. *Neuron* **88**:832-844 <https://doi.org/10.1016/j.neuron.2015.10.001> | [PubMed](#)
- Bisley JW**, Goldberg ME (2010) Attention, Intention, and Priority in the Parietal Lobe. *Annual Review of Neuroscience* **33**:1-21 <https://doi.org/10.1146/annurev-neuro-060909-152823>
- Boshra R**, Kastner S (2022) Attention control in the primate brain. *Current Opinion in Neurobiology* **76**:102605 <https://doi.org/10.1016/j.conb.2022.102605> | [PubMed](#)
- Buschman TJ**, Kastner S (2015) From Behavior to Neural Dynamics: An Integrated Theory of Attention. *Neuron* **88**:127-144 <https://doi.org/10.1016/j.neuron.2015.09.017> | [PubMed](#)

- Buschman TJ, Miller EK (2007) Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices. *Science* **315**:1860-1862 <https://doi.org/10.1126/science.1138071> | PubMed
- Castelhano MS, Mack ML, Henderson JM (2009) Viewing task influences eye movement control during active scene perception. *Journal of Vision* **9**:6 <https://doi.org/10.1167/9.3.6> | PubMed
- Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. *Nature* **441**:876-879 <https://doi.org/10.1038/nature04766> | PubMed
- Deaner RO, Khera AV, Platt ML (2005) Monkeys Pay Per View: Adaptive Valuation of Social Images by Rhesus Macaques. *Current Biology* **15**:543-548 <https://doi.org/10.1016/j.cub.2005.01.044> | PubMed
- Ebitz RB, Albarran E, Moore T (2018) Exploration Disrupts Choice-Predictive Signals and Alters Dynamics in Prefrontal Cortex. *Neuron* **97**:450-461.e9 <https://doi.org/10.1016/j.neuron.2017.12.007> | PubMed
- Epstein RA, Baker CI (2019) Scene Perception in the Human Brain. *Annual Review of Vision Science* **5**:373-397 <https://doi.org/10.1146/annurev-vision-091718-014809> | PubMed
- Fecteau JH, Munoz DP (2006) Saliency, relevance, and firing: A priority map for target selection. *Trends in Cognitive Sciences* **10**:382-390 <https://doi.org/10.1016/j.tics.2006.06.011> | PubMed
- Fei-Fei L, Iyer A, Koch C, Perona P (2007) What do we perceive in a glance of a real-world scene?. *Journal of Vision* **7**:10 <https://doi.org/10.1167/7.1.10> | PubMed
- Findlay JM, Gilchrist ID (2003) *Active vision: The psychology of looking and seeing* Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198524793.001.0001>
- Gelman A, Goodrich B, Gabry J, Vehtari A (2019) R-squared for Bayesian Regression Models. *The American Statistician* **73**:307-309 <https://doi.org/10.1080/00031305.2018.1549100>
- Ghazizadeh A, Griggs W, Hikosaka O (2016) Ecological Origins of Object Saliency: Reward, Uncertainty, Aversiveness, and Novelty. *Frontiers in Neuroscience* **10** <https://doi.org/10.3389/fnins.2016.00378> | PubMed
- Hannula DE, Althoff RR, Warren DE, Riggs L, Cohen NJ, Ryan JD (2010) Worth a glance: Using eye movements to investigate the cognitive neuroscience of memory. *Frontiers in Human Neuroscience* **4**:166 <https://doi.org/10.3389/fnhum.2010.00166> | PubMed
- Hannula DE, Ranganath C (2009) The Eyes Have It: Hippocampal Activity Predicts Expression of Memory in Eye Movements. *Neuron* **63**:592-599 <https://doi.org/10.1016/j.neuron.2009.08.025> | PubMed
- Hayes TR, Henderson JM (2022a) Meaning maps detect the removal of local semantic scene content but deep saliency models do not. *Attention, Perception, & Psychophysics* **84**:647-654 <https://doi.org/10.3758/s13414-021-02395-x> | PubMed
- Hayes TR, Henderson JM (2022b) Scene inversion reveals distinct patterns of attention to semantically interpreted and uninterpreted features. *Cognition* **229**:105231 <https://doi.org/10.1016/j.cognition.2022.105231> | PubMed
- Hayes TR, Henderson JM (2025) DeepMeaning: Estimating and Interpreting Scene Meaning for Attention Using a Vision-Language Transformer. *Open Mind* **9**:1020-1036 <https://doi.org/10.1162/opmi.a.6>
- Henderson JM (2003) Human gaze control during real-world scene perception. *Trends in Cognitive Sciences* **7**:498-504 <https://doi.org/10.1016/j.tics.2003.09.006> | PubMed
- Henderson JM (2013) Eye movements. In: Reisberg D (Ed). *The Oxford handbook of cognitive psychology* Oxford University Press. pp. 69-82 <https://doi.org/10.1093/oxfordhb/9780195376746.013.0005>
- Henderson JM (2017) Gaze Control as Prediction. *Trends in Cognitive Sciences* **21**:15-23 <https://doi.org/10.1016/j.tics.2016.11.003> | PubMed

- Henderson JM, Goold JE, Choi W, Hayes TR (2020) Neural Correlates of Fixated Low- and High-level Scene Properties during Active Scene Viewing. *Journal of Cognitive Neuroscience* **32**:2013-2023 https://doi.org/10.1162/jocn_a_01599 | PubMed
- Henderson JM, Hayes TR (2017) Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour* **1**:743-747 <https://doi.org/10.1038/s41562-017-0208-0> | PubMed
- Henderson JM, Hayes TR (2018) Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision* **18**:10 <https://doi.org/10.1167/18.6.10> | PubMed
- Henderson JM, Hayes TR, Peacock CE, Rehrig G (2019) Meaning and Attentional Guidance in Scenes: A Review of the Meaning Map Approach. *Vision* **3**:19 <https://doi.org/10.3390/vision3020019> | PubMed
- Itti L, Koch C (2001) Computational modelling of visual attention. *Nature Reviews. Neuroscience* **2**:194-203 <https://doi.org/10.1038/35058500> | PubMed
- Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**:1254-1259 <https://doi.org/10.1109/34.730558>
- Kiat JE, Hayes TR, Henderson JM, Luck SJ (2022) Rapid Extraction of the Spatial Distribution of Physical Saliency and Semantic Informativeness from Natural Scenes in the Human Brain. *Journal of Neuroscience* **42**:97-108 <https://doi.org/10.1523/JNEUROSCI.0602-21.2021> | PubMed
- Li C.-L, Aivar MP, Tong MH, Hayhoe MM (2018) Memory shapes visual search strategies in large-scale environments. *Scientific Reports* **8**:4324 <https://doi.org/10.1038/s41598-018-22731-w> | PubMed
- Logothetis NK, Pauls J, Poggio T (1995) Shape representation in the inferior temporal cortex of monkeys. *Current Biology* **5**:552-563 [https://doi.org/10.1016/S0960-9822\(95\)00108-4](https://doi.org/10.1016/S0960-9822(95)00108-4) | PubMed
- Machado CJ, Bliss-Moreau E, Platt ML, Amaral DG (2011) Social and Nonsocial Content Differentially Modulates Visual Attention and Autonomic Arousal in Rhesus Macaques. *PLOS One* **6**:e26598 <https://doi.org/10.1371/journal.pone.0026598> | PubMed
- Malcolm GL, Groen IIA, Baker CI (2016) Making Sense of Real-World Scenes. *Trends in Cognitive Sciences* **20**:843-856 <https://doi.org/10.1016/j.tics.2016.09.003> | PubMed
- Moore T, Zirnsak M (2017) Neural Mechanisms of Selective Visual Attention. *Annual Review of Psychology* **68**:47-72 <https://doi.org/10.1146/annurev-psych-122414-033400> | PubMed
- Navalpakkam V, Itti L (2005) Modeling the influence of task on attention. *Vision Research* **45**:205-231 <https://doi.org/10.1016/j.visres.2004.07.042> | PubMed
- Oakes LM, Hayes TR, Klotz SM, Pomaranski KI, Henderson JM (2024) The role of local meaning in infants' fixations of natural scenes. *Infancy* **29**:284-298 <https://doi.org/10.1111/inf.12582> | PubMed
- Parkhurst D, Law K, Niebur E (2002) Modeling the role of salience in the allocation of overt visual attention. *Vision Research* **42**:107-123 [https://doi.org/10.1016/s0042-6989\(01\)00250-4](https://doi.org/10.1016/s0042-6989(01)00250-4) | PubMed
- Peacock CE, Hayes TR, Henderson JM (2019) Meaning guides attention during scene viewing, even when it is irrelevant. *Attention, Perception, & Psychophysics* **81**:20-34 <https://doi.org/10.3758/s13414-018-1607-7> | PubMed
- Peelen MV, Kastner S (2014) Attention in the real world: Toward understanding its neural basis. *Trends in Cognitive Sciences* **18**:242-250 <https://doi.org/10.1016/j.tics.2014.02.004> | PubMed
- Pinsk MA, DeSimone K, Moore T, Gross CG, Kastner S (2005) Representations of faces and body parts in macaque temporal cortex: A functional MRI study. *Proceedings of the National Academy of Sciences* **102**:6996-7001 <https://doi.org/10.1073/pnas.0502605102> | PubMed
- Ponce CR, Xiao W, Schade PF, Hartmann TS, Kreiman G, Livingstone MS (2019) Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell* **177**:999-1009.e10 <https://doi.org/10.1016/j.cell.2019.04.005> | PubMed

- Ramey MM, Henderson JM, Yonelinas AP (2025) Episodic memory and semantic knowledge interact to guide eye movements during visual search in scenes: Distinct effects of conscious and unconscious memory. *Psychonomic Bulletin & Review* **32**:2395-2409 <https://doi.org/10.3758/s13423-025-02686-6> | PubMed
- Reynolds JH, Heeger DJ (2009) The Normalization Model of Attention. *Neuron* **61**:168-185 <https://doi.org/10.1016/j.neuron.2009.01.002> | PubMed
- Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. *Nature Neuroscience* **2**:1019-1025 <https://doi.org/10.1038/14819> | PubMed
- Schmidig FJ, Yamin D, Sharon O, Nadu Y, Nir J, Ranganath C, Nir Y (2025) Anticipatory eye gaze as a marker of memory. *Communications Psychology* **3**:122 <https://doi.org/10.1038/s44271-025-00305-7> | PubMed
- Solyst JA, Buffalo EA (2014) Social relevance drives viewing behavior independent of low-level saliency in rhesus macaques. *Frontiers in Neuroscience* **8** <https://doi.org/10.3389/fnins.2014.00354> | PubMed
- Soyuhos O, Scarpa A, Baldauf D (2025) Distinct resting-state connectomes for face and scene perception predict individual task performance. *bioRxiv* <https://doi.org/10.1101/2025.07.09.663812>
- Soyuhos O, Zirnsak M, Chaudhuri R, Chen X (2026) Selective control of prefrontal neural timescales by parietal cortex. *Nature Communications* **17**:3687 <https://doi.org/10.1038/s41467-026-70326-1> | PubMed
- Summerfield JJ, Lepsien J, Gitelman DR, Mesulam MM, Nobre AC (2006) Orienting attention based on long-term memory experience. *Neuron* **49**:905-916 <https://doi.org/10.1016/j.neuron.2006.01.021> | PubMed
- Sutton S (1998) Predicting and Explaining Intentions and Behavior: How Well Are We Doing?. *Journal of Applied Social Psychology* **28**:1317-1338 <https://doi.org/10.1111/j.1559-1816.1998.tb01679.x>
- Tatler BW (2007) The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision* **7**:4 <https://doi.org/10.1167/7.14.4> | PubMed
- Taubert J, Wardle SG, Flessert M, Leopold DA, Ungerleider LG (2017) Face Pareidolia in the Rhesus Monkey. *Current Biology* **27**:2505-2509.e2 <https://doi.org/10.1016/j.cub.2017.06.075> | PubMed
- Tsao DY, Freiwald WA, Tootell RBH, Livingstone MS (2006) A Cortical Region Consisting Entirely of Face-Selective Cells. *Science* **311**:670-674 <https://doi.org/10.1126/science.1119983> | PubMed
- Vinken K, Sharma S, Livingstone MS (2025) Mapping macaque to human cortex with natural scene responses. *Proceedings of the National Academy of Sciences* **122**:e2512619122 <https://doi.org/10.1073/pnas.2512619122> | PubMed
- Wolfe JM (1994) Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review* **1**:202-238 <https://doi.org/10.3758/BF03200774> | PubMed
- Xia R, Chen X, Engel TA, Moore T (2024) Common and distinct neural mechanisms of attention. *Trends in Cognitive Sciences* **28**:554-567 <https://doi.org/10.1016/j.tics.2024.01.005> | PubMed
- Yarbus AL (1967) Eye Movements During Perception of Complex Objects. In: Yarbus AL (Ed). *Eye Movements and Vision* Springer US. pp. 171-211 https://doi.org/10.1007/978-1-4899-5379-7_8
- Zhang H, Anderson NC, Miller KF (2022) Scene meaningfulness guides eye movements even during mind-wandering. *Attention, Perception, & Psychophysics* **84**:1130-1150 <https://doi.org/10.3758/s13414-021-02370-6> | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

The manuscript examines whether scene meaning guides overt attention in rhesus macaques. Two monkeys freely viewed naturalistic indoor scenes, including laboratory or housing scenes described as familiar and other indoor scenes described as unfamiliar. The authors compare fixation locations with matched non-fixated control locations using predictors derived from center proximity, image salience, and a DeepMeaning model intended to capture the spatial distribution of semantic informativeness. They report that meaning predicts fixation selection beyond salience and center bias, that meaning and salience interact, that familiar scenes produce broader exploration of low-meaning regions, and that the influence of meaning increases with attentional engagement.

Strengths:

A major strength of the study is its use of natural free-viewing behavior in macaques. The experimental approach takes advantage of intrinsic gaze allocation rather than relying on a more artificial task, which makes the work a useful bridge between human scene-viewing studies and future neurophysiological studies in nonhuman primates.

The statistical analyses are extensive. The authors model fixated and matched non-fixated samples with Bayesian generalized linear mixed models, including center proximity and salience as important controls, examined interactions among predictors, and reported diagnostics for multicollinearity and model convergence. These analyses support the basic observation that the human-derived meaning maps are associated with macaque fixation allocation beyond the particular center and salience terms included in the model.

The question is interesting and timely. If meaning-like scene structure can be operationalized for macaque viewing, this would provide a useful behavioral foundation for future work on the neural mechanisms that link scene analysis, gaze allocation, and natural behavior.

Weaknesses:

The main weakness is interpretive. The manuscript often treats the DeepMeaning map as though it measures scene meaning for the monkey, but the map is ultimately human-derived. Some of the examples make this issue especially salient: regions such as clocks, phones, dining tables, or other human artifacts may be meaningful to human observers, but it is not clear that they have semantic meaning for macaques. If meaning-based guidance is argued to emerge through experience, then unfamiliar human indoor scenes that the monkeys have never encountered cannot straightforwardly be meaningful to them in the same sense that they are meaningful to humans. Predictive success for these scenes may therefore indicate sensitivity to visual or object-level structure correlated with human-rated meaning, rather than macaque semantic understanding.

A related concern is that the DeepMeaning predictor may capture forms of visual salience, objectness, or high-level image structure not captured by the particular low-level salience model. For example, a clock or phone may attract gaze because of shape, contrast, face-like configuration, object boundaries, or other mid-level features rather than because it carries semantic meaning for a macaque. The present analyses show that this model is predictive, but they do not by themselves establish that the predictive variable is semantic meaning rather than visual structure beyond Itti-Koch-style salience.

The manuscript relies heavily on fitted model parameters and derived maps, with relatively little return to the raw behavioral data. The main claims would be easier to evaluate if the authors showed more direct fixation-density maps, scene-by-scene examples, and aggregate raw relationships between fixation behavior and map values. At present, much of the argument rests on interpreting fitted coefficients, without enough behavioral visualization to show what the monkeys actually did across the stimulus set.

It is also unclear whether model performance was evaluated on held-out data. The comparison to repeated viewing of the same images is useful as a behavioral benchmark, but a second viewing may itself be affected by familiarity or memory for the image. This makes it a potentially imperfect estimate of a noise ceiling for first-pass fixation predictability. Cross-validation or held-out prediction, ideally across held-out images as well as trials, would make the predictive claims more convincing.

Although the authors describe multicollinearity as negligible, Figure S2B-C appears to show some nontrivial correlations among predictors. These correlations may matter for interpretation even if variance inflation factors fall below conventional thresholds, especially when the signs of fitted effects point in directions that may be expected from the input correlations, such as relationships involving meaning and familiarity. The manuscript would benefit from reporting these correlations quantitatively and relating them to the fitted effects.

The familiarity analysis is interesting but would benefit from further control. Familiar scenes are photographs of the monkeys' housing and laboratory environments, whereas unfamiliar scenes are other indoor environments. These categories may differ not only in familiarity but also in clutter, spatial layout, object density, color distribution, luminance, contrast, edge density, texture statistics, or the distributions of salience and meaning values. Without additional characterization of the image sets, the conclusion that familiarity itself broadens exploration should be treated cautiously.

The engagement effects also appear less consistent across the two monkeys than some of the summary language suggests. The monkey-specific results should be emphasized, and claims about engagement strengthening meaning-based guidance should be stated in proportion to the cross-animal evidence.

Finally, the manuscript sometimes uses language that sounds more mechanistic than the behavioral data can support. The negative interaction between meaning and salience is an interesting result, but terms such as competitive integration in a shared priority map go beyond what can be concluded from overt fixation selection alone. The study lacks a causal or perturbational manipulation, such as image inversion or another transformation that preserves local features while altering semantic organization. The result would be clearer if described first as a model-based association or subadditive interaction in gaze allocation, with the priority-map interpretation presented as a plausible account rather than a direct conclusion.

<https://doi.org/10.7554/eLife.111311.1.sa2>

Reviewer #2 (Public review):

Summary:

In prior work, the authors developed an ML algorithm that computes spatial maps of "meaning": image regions that are likely to be given semantic labels by human observers. They also previously showed that "meaning" predicts fixations in humans and human infants. Here, these observations were extended to macaque monkeys, testing the hypothesis that meaning is a phylogenetically preserved driver of overt attention across primates.

Strengths:

The paper reports that fixated locations had higher values of meaning compared to nearby, non-fixated locations. Specifically, it shows that meaning values - as inferred from a neural network model - are useful in differentiating these two classes of locations, beyond the

established effects of image salience and centrality on gaze. The reported results were consistent in both monkeys.

Weaknesses:

It is difficult to understand what, precisely, is meant by meaning from this paper, although the prior work from this group may offer some insight. Given that, it is not clear if "high-meaning" image locations tend to be objects, for example, or faces, or other such behaviorally relevant image features. Indeed, the utility of the meaning maps was not evaluated against other algorithms that consider more complex natural scene information. This is a particular concern as the paper does not demonstrate that meaning predicts where the viewer will look within the image; instead, it shows that meaning is one of the variables that differentiates fixated locations from nearby non-fixated locations. Because this is not a causal study by necessity, caution is also needed in interpreting the results. In our view, the most parsimonious interpretation may not be that meaning guides gaze in monkeys, but instead that people tend to name things that primate brains evolved to fixate on at the expense of neighboring locations.

<https://doi.org/10.7554/eLife.111311.1.sa1>

Reviewer #3 (Public review):

Summary:

This novel study asks whether meaning-based guidance of overt attention, well-established in humans through the "meaning map" framework, extends to non-human primates. The authors recorded eye movements from two rhesus macaques freely viewing naturalistic indoor scenes and modeled fixation selection using DeepMeaning maps, Itti-Koch salience maps, and center proximity. They report that scene meaning robustly predicts fixation selection after controlling for salience and center bias, that meaning and salience interact competitively rather than additively, and that the influence of meaning is modulated by scene familiarity and attentional engagement. The cross-species extension of the meaning map approach is a valuable contribution, and the Bayesian GLMM framework with variance partitioning is well-suited to the question.

Strengths:

- (1) The cross-species extension itself is novel and well-motivated. Nobody has applied the meaning map framework to NHP gaze behavior before. Even with the interpretive caveats I raise below, creating this methodological bridge between human scene perception research and NHP circuit neuroscience is a valuable contribution.
- (2) The statistical framework is strong. The Bayesian GLMM with posterior distributions, HDIs, and probability of direction is more informative than frequentist alternatives. The variance partitioning with ΔR^2 is the right approach for disentangling predictor contributions. Random intercepts for scene are appropriate. The convergence diagnostics ($R\text{-hat} = 1.00$, ESS > 8000 across all models) are exemplary.
- (3) Transparent individual-subject reporting. With $N = 2$, reporting each monkey separately rather than pooling or averaging is the correct choice, and the authors do this consistently. The individual differences are visible because the reporting is honest.
- (4) The experimental design is excellent. 200 scenes is a substantial stimulus set by NHP standards. The inclusion of both familiar and unfamiliar environments, the repeated-viewing design for reliability estimation, and the 5-second free viewing window that yields ~15 fixations per trial all reflect thoughtful design.

(5) The familiarity and engagement analyses go beyond the basic demonstration. Even with the limitations we identified, asking how behavioral context modulates the meaning-gaze relationship is more ambitious than simply showing that the correlation exists. These analyses generate testable predictions for future work.

(6) Data and code sharing commitment. The authors plan to release raw data, preprocessing, and analysis code on OSF and GitHub.

Weaknesses:

(1) The authors' central claim is that meaning-based attentional guidance is an "evolutionarily conserved component of primate vision." This claim rests on the finding that macaque fixation patterns correlate with DeepMeaning maps. However, DeepMeaning is trained on human ratings of local scene meaning using a vision-language transformer (CoCa) pretrained on billions of human image-text pairs. What the model captures, then, is the spatial distribution of visual structure that humans judge to be semantically informative. The authors acknowledge that DeepMeaning represents "structured visual representations of scene regions containing identifiable objects and informative relationships" (lines 261-262), but this acknowledgment actually highlights the problem: regions containing identifiable objects and informative spatial relationships would plausibly attract fixations in any visual system with object-selective neurons and a bias toward structured content, regardless of whether the observer is processing "meaning" in any semantic sense. That is, the correlation between macaque gaze and DeepMeaning maps is consistent with shared object-level visual processing, but doesn't uniquely implicate shared semantic processing. The critical adversarial test from Hayes & Henderson (2022a)-where meaning maps detected the removal of semantic content via diffeomorphic scrambling while deep saliency models did not-has not been applied to macaque viewing behavior. Importantly, such a test would require new data collection (showing monkeys scrambled scenes), which may not be feasible. A more tractable approach with the existing data would be to compare DeepMeaning against some other model that captures mid-level visual structure without semantic supervision, though this would be a weaker test. Given these constraints, I would ask the authors to (a) acknowledge this limitation explicitly and temper the evolutionary conservation claim accordingly-for example, framing the result as evidence that macaques and humans share attentional biases toward visually structured scene regions, with the semantic interpretation remaining an open question-and (b) note the diffeomorphic scrambling experiment as an important future direction for establishing whether macaque attention is guided by semantic content per se.

(2) The familiar/unfamiliar scene comparison confounds long-term familiarity with systematic differences in scene content. Familiar scenes are photographs of the vivarium and laboratory; unfamiliar scenes are restaurants, bedrooms, kitchens, and offices. These two categories almost certainly differ in visual complexity, object density, spatial layout, clutter, and the types of objects present. The familiar environments (vivarium caging, lab equipment) are likely more spatially repetitive and lower in object diversity than, say, a restaurant or residential kitchen. Any difference attributed to "familiarity" could therefore reflect these systematic content differences. The negative interaction between meaning and familiarity (Monkey V: $\beta = -0.19$; Monkey I: $\beta = -0.19$), which the authors interpret as familiarity broadening exploration, could instead reflect the fact that vivarium/lab scenes have a different distribution of meaning values or a different relationship between meaning and salience than human domestic environments. The authors should address this confound directly. At minimum, comparing the distributions of meaning and salience values across the two scene categories would help the reader evaluate whether the familiarity effect can be separated from content effects. Ideally, the authors would include a subset analysis using only scenes matched on feature distributions or include scene-level summary statistics of the meaning and salience maps as covariates in the familiarity model.

<https://doi.org/10.7554/eLife.111311.1.sa0>