

UC Santa Barbara

UC Santa Barbara Electronic Theses and Dissertations

Title

Decoding Free Viewing: Using Vision-Language Models to Reveal the Optimality of Human Eye Movements for Scene Understanding

Permalink

<https://escholarship.org/uc/item/7bx5g17w>

ISBN

9798186386704

Author

Murlidaran, Shravan

Publication Date

2026-06-16

Supplemental Material

<https://escholarship.org/uc/item/7bx5g17w#supplemental>

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Santa Barbara

**Decoding Free Viewing: Using Vision-Language Models to
Reveal the Optimality of Human Eye Movements for Scene
Understanding**

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Psychological and Brain Sciences

by

Shravan Murlidaran

Committee in charge

Professor Miguel Eckstein, Chair

Professor Mary Hegarty

Professor Barry Giesbrecht

Professor William Wang

June 2026

The dissertation of Shravan Murlidaran is approved

Mary Hegarty

Barry Giesbrecht

William Wang

Miguel Eckstein, Committee Chair

June 2026

Acknowledgements

The completion of this dissertation would not have been possible without the love, support, and encouragement of the many remarkable people who have been part of my journey. First and foremost, I am profoundly grateful to my family, whose unconditional love and unwavering belief in me have been my greatest source of strength. My mom and dad have been strong pillars of support throughout my life. My mother has been the biggest cheerleader I could have ever asked for. I am the person I am today, with varied interests in music, instruments, and sports, because of the efforts she made to join me in those activities as a kid. My dad has always been the practical one, showing me how to handle failure and move forward toward success. He encouraged my curiosity about the world from a young age, and that curiosity blossomed into my love for science. Knowing I can always fall back on them if something doesn't go my way has been the biggest pillar of my strength. During the final three months of my PhD, they sat with me in the lab late into the night, making sure I was fed and cared for as I tried to finish things up. They have always made my path easier in difficult times, and no words could ever fully capture the depth of my gratitude for their presence in my life.

The next biggest blessing in my life is my sister, Sruthi. Since childhood, she has been my role model, and it would be an understatement to say she has shaped my thoughts and opinions about life. In many ways, her love for science grew on me as well, and I ended up following in her footsteps by pursuing a PhD. She has been my guiding star, and this work is my humble token of gratitude for everything she has given me. I never had a big brother figure until my brother-in-law, Mani, came into my life. I have always admired his ability to stand his ground and ensure things are done right, even under immense pressure to do otherwise. It is a quality I have long tried to emulate, and one that gave me the resolve to withstand the many pressures of

a PhD. I am also grateful to him for pushing me toward fitness and for always checking up on me, looking out for me the way only a big brother could. I am truly fortunate to have him in my life. The newest addition to my life is my nephew Avyaan, and the joy he brings is something truly incomparable. A few minutes talking to him over the phone was all it ever took to melt away whatever stress I was carrying. His laughter and innocence were a constant reminder of what truly matters. In his own unknowing way, he has been one of my greatest sources of strength through half of this PhD, and I cannot wait to watch him grow (he is 3 years old!). I would also like to dedicate this thesis to my extended family, who have been the most supportive group of people throughout my life. The conversations I have had with my cousins about life, music, movies, and cricket are something I cherish deeply. They have always felt like coming home, no matter how far away I was, and they are memories I intend to carry close to my heart forever. My cousins, uncles, aunts, and my grandparents on both my mom's and dad's sides have always been happier for me than I have been for myself, and among the loudest cheers in my corner. I am endlessly grateful for each and every one of them. I would like to thank my advisor, Prof. Miguel P. Eckstein, who over the course of my PhD has taught me what it means to be a good scientist. His attention to detail and his ability to meticulously untangle the nuances of data have helped me hone my skills and grow into a better researcher. I came into this PhD with an engineering background and virtually no knowledge of psychology or vision science, and he took me under his wing and showed me the way, guiding me toward some truly incredible work. The fruits of that journey include a published paper in Nature Communications and a patent, achievements I could not have reached without his mentorship. Beyond the science, I have also learned a great deal from him about navigating the inevitable challenges that come with life in research. I am truly grateful to have had him as my mentor. I would like to thank my committee members, Prof. Mary Hegarty, Prof. William Wang, Prof. Barry Giesbrecht, and Prof. Michael Beyeler, for their continued support and guidance throughout my PhD. Their collective expertise and thoughtful

feedback were invaluable in shaping the direction and depth of this work. I would like to extend a special thank you to Prof. Barry Giesbrecht for graciously joining my committee after some unexpected changes, a gesture I am deeply grateful for. The suggestions offered by my committee during my thesis proposal gave me a clearer vision of the path forward and played a meaningful role in shaping what this dissertation ultimately became. Each of them brought a unique perspective to the table, and their questions and insights pushed me to think more rigorously and communicate my ideas more clearly. I am fortunate to have had such a thoughtful and supportive committee by my side. I would also like to extend my heartfelt gratitude to the professors and teachers who shaped me long before this PhD began. My undergraduate and master's professors, Prof. Dmitry Korkin, Prof. Loris Fichera, Prof. Balaraman Ravindran, and Prof. Pannirselvam, each played a meaningful role in building the foundation upon which this work stands. Their dedication to teaching and their passion for their respective fields sparked something in me that I have carried into every stage of my academic journey. I am equally grateful to all my school teachers, who were the first to nurture my curiosity and set me on this path. It is often the earliest guides in our lives whose influence runs the deepest, and I would not be the scientist or the person I am today without their patience, encouragement, and belief in me. To all of them, I am profoundly thankful. I would also like to thank my labmates for always being supportive and offering valuable input to make my projects better. Being a shy person, I was fortunate that Puneeth made sure I felt comfortable as a newcomer in the lab. He was also my music buddy, someone I could talk to about different Raags. I would like to thank Sudhanshu for his advice on my talks and posters, and Nicole for being incredibly helpful, especially in the early days when I was still finding my footing in the lab. I would like to thank Devi for sharing his eye-tracking code to help me set up my own experiments and for his advice as I neared the finish line. I would also like to thank my other labmates, Byron, Srijita, Ansh, Parsa, Ziqi, Jon, Sana, Tanish, and Anqi, for being such a warm and supportive group and for helping me

with various projects and lab activities. And of course, a big thank you to all the lab RAs who stepped up as subjects or helped me run experiments over the years. I would like to thank all my friends who have been part of this journey. Vamshi, Karthik, Goirik, Barry, and Suriya have all been subjects in my eye-tracking experiments and in the scenes I photographed as part of my project. More than that, they have been my family away from home, especially during COVID, when everything felt uncertain. Their support and friendship are something I will cherish for the rest of my life. I would also like to thank the many other friends I had the pleasure of meeting throughout my PhD. I am grateful for each and every one of them. Finally, I would like to thank the music that has been a part of every step of this PhD journey. It has been one of the strongest pillars of strength, giving me the resolve to face whatever challenges came my way. Whether it was Ravaani, UCSB's premier award-winning South Asian a cappella team, my sitar classes, or fun jamming sessions with friends, these experiences helped me grow as a person and gave me memories I will carry for life. I may have unintentionally left someone out, but please know that each and every person I have interacted with during this PhD holds a special place in my heart. And finally, I would also like to thank those I had the opportunity to meet along the way, who have the potential to be part of a wonderful future ahead. Thank you all for making this difficult journey something I will cherish forever.

Vita of Shravan Murlidaran

Ph.D. Candidate in Psychological and Brain Sciences

smurlidaran@ucsb.edu

linkedin.com/in/shravan-murlidaran

github.com/shravan1394

Education

Ph.D. Candidate, Psychological and Brain Sciences

University of California, Santa Barbara

2019 – June 2026

M.S. in Robotics

Worcester Polytechnic Institute, Massachusetts

B.Tech in Mechanical Engineering

National Institute of Technology, Trichy, India

Experience

Graduate Student Researcher

University of California, Santa Barbara

Sep 2019 – Present

Graduate Teaching Assistant

University of California, Santa Barbara

Sep 2019 – Present

Graduate Research Assistant

Worcester Polytechnic Institute

Aug 2016 – Aug 2019

Robotics Club Member

National Institute of Technology, Trichy

Jun 2013 – May 2016

Patents, Publications and Preprints

Patents

- Automated Eye Movement Response Systems and Methods." U.S. Provisional Patent Application No. 64/068,590. Filed May 2026. (Patent Pending)

Journal Papers

- Murlidaran, S., Eckstein, M.P. Eye movements during free viewing to maximize scene understanding. Nat Commun 17, 940 (2025). <https://doi.org/10.1038/s41467-025-67673-w>
- Murlidaran, S., Wen, Z., Skaza, J., Wang, W., & Eckstein, M. P. (2025). Semantic Saliency from Multi-Modal Large Language Model Scene Understanding Maps. Center for Open Science. https://doi.org/10.31234/osf.io/x3mfa_v1

- Wen, Z., Skaza, J., Murlidaran, S., Wang, W. Y., & Eckstein, M. P. (2025). Predicting Reaction Time to Comprehend Scenes with Foveated Scene Understanding Maps. arXiv preprint arXiv:2505.12660.
- Skaza, J., Murlidaran, S., Varshney, A., Wen, Z., Wang, W., Eckstein, M. P., & Beyeler, M. (2025). A Deep Learning Framework for Predicting Functional Visual Performance in Bionic Eye Users. bioRxiv, 2025-06.

Conference Papers

- Murlidaran, S., Wang, W. Y., & Eckstein, M. P. (2021). Comparing visual reasoning in humans and ai. ICLR Brain2AI workshop.

Thesis

- Murlidaran, S. (2019). A mixed reality framework for surgical navigation: approach and preliminary results (Masters dissertation, Worcester Polytechnic Institute).

Presentations

- **2025** - Scene Understanding Maps: Predicting Most Frequently Fixated Object during Free Viewing with Multi-Modal Large Language Models. Shravan Murlidaran; Miguel P. Eckstein. Journal of Vision, 25 (VSS 2025). Journal of Vision
- **2024** - Eye Movements during Free Viewing to Maximize Scene Understanding. Shravan Murlidaran; Miguel P. Eckstein. 24 (VSS 2024). Journal of Vision
- **2023** - Eye Movements during Free Viewing and Scene Description are Similarly Directed to Objects Critical to Scene Understanding. Shravan Murlidaran; Miguel P. Eckstein, Gordon Research Conference on Eye Movements (GRC 2023), Mount Holyoke College, South Hadley, MA.
- **2023** - Eye Movements during Free Viewing and Scene Description are Similarly Directed to Objects Critical to Scene Understanding. Shravan Murlidaran; Miguel P. Eckstein. Journal of Vision, 23 (VSS 2023). Journal of Vision

Abstract

Decoding Free Viewing: Using Vision-Language Models to Reveal the Optimality of Human Eye Movements for Scene Understanding

By

Shravan Murlidaran

Eye movements are an important part of human vision. We constantly make them, directing our foveal processing to different locations. Research has shown eye movements to optimize accuracy in various tasks, such as search and navigation, implying that human eye movements are an active process, constantly seeking information that could aid decision-making. But what happens when there is no task to guide eye movements (free viewing task)? Is there any purpose to those eye movements? A prevalent idea is that during free viewing, humans make eye movements to low-level visually salient regions. However, decades of research have shown that humans fixate on people in scenes, on objects, on gaze, and on text, and, more recently, on local regions judged to be meaningful. The goal of this thesis is to understand which perceptual tasks humans engage in during free viewing and to assess whether free-viewing eye movements reflect the goal of optimizing task accuracy (optimal eye movements). Based on the findings, I propose that an important, general human default task during free viewing is to comprehend scenes, and that humans plan near-optimal eye movements, actively seeking information that maximizes comprehension.

The first part of this thesis focuses on experimental evidence from eye-movement measurements suggesting that scene comprehension is an important default task for humans during free viewing. Eye movements were measured under four task instructions: free viewing, scene description, object search, and counting objects. The stimuli

were image pairs, called Winograd images, containing minor visual alterations that drastically change scene interpretation without altering low-level saliency, thereby isolating the semantic factors guiding eye movements. Results indicate that free-viewing fixations closely resemble those of observers explicitly instructed to describe a scene, differing significantly from fixations during object search or counting. Furthermore, free-viewing fixations are disproportionately directed toward people and objects most critical to understanding the scene (objects that maximally impact scene descriptions when removed), rather than solely toward low-level visual saliency or locally meaningful regions. I also show that human fixations on these critical elements improve their understanding of the scene, demonstrating a causal influence of these fixations in accurate scene comprehension.

The second part of this thesis evaluates whether human fixations during free viewing approximate the optimal strategy for maximizing scene comprehension. The challenge in this objective is to create a model that estimates optimal fixations for real-world scenes and high-level goals such as scene comprehension. Classical methods based on Bayesian ideal searchers have a strong mathematical foundation but cannot be applied to real-world scenes. I implemented a model that simulates human foveated vision and sequential eye-movement exploration of scenes. Leveraging state-of-the-art vision-language models (VLMs) capable of human-level scene comprehension, the model generated descriptions of the actively explored foveated scenes. I subsequently trained a reinforcement learning (RL) agent that uses a convolutional neural network (CNN) to optimize visual exploration (Q-network), systematically executing eye movements to maximize the semantic accuracy of the VLM’s descriptions at each step. By measuring free-viewing eye movements on images carefully curated to depict complex social interactions, actions, or implied actions, I categorized the elements present in these images and counted the human fixation frequencies for these categories (people, objects relevant and irrelevant to the understanding of the scene, text, gazed and grasped objects, salient regions). The optimized RL agent matched the fixation patterns observed in

human data without any prior training on human fixation data, significantly outperforming the same RL model optimized for search or image classification tasks, as well as saliency prediction models. Furthermore, I found that VLM descriptions generated by simulating foveation at human-fixated locations, recorded during the free-viewing task, achieved semantic accuracy comparable to that obtained via fixations from an RL agent specifically optimized for scene understanding. The agreement between the RL model and human fixation frequencies also decreased when the model was trained with very low or extreme foveation, suggesting that human free viewing eye movements are an emergent property of an interaction between the goal to optimize scene comprehension and the specific foveated properties of the human visual system. I assessed the generalization of the RL model’s results to a recently published data set on eye movements from 6720 observers aged 5-72.

Together, this research demonstrates that eye movements during free viewing are an actively optimized, task-driven behavior aimed at comprehending the visual world. Fixations to people, text, objects relevant to understanding, and gazed/grasped objects are an emergent property of an interaction between the goal of optimizing scene understanding and the foveated properties of human sensory processing.

Contents

Acknowledgements	iii
Abstract	ix
Introduction	1
1 Free-viewing eye movements aim to understand scenes.	6
1.1 Abstract	6
1.2 Introduction	7
1.3 Methods:	12
1.4 Results:	24
1.5 Discussion	51
2 Why We Look Where We Look: Emergent Human-like Fixations of a Foveated Visual Language Model Maximizing Scene Understanding	59
2.1 Abstract	59
2.2 Introduction	60
2.3 Methods	66
2.4 Results	100
2.5 Discussion	118
Conclusions and Future Work	122
Bibliography	126

Introduction

You enter a busy coffee shop, find your spot, and sit down. You suddenly notice something odd in your peripheral vision and move your eyes to see a woman in shock, her hands over her mouth. You sense trouble, immediately notice her gaze, and direct your eyes in that direction to see a man on one knee, a smile on his face, his hand outstretched towards the woman, holding something. You suspect a marriage proposal and direct your gaze towards the object the man is holding to see a ring, confirming your suspicion that it is indeed a marriage proposal. Humans constantly make eye movements to explore their environment, a long-standing notion that has stood the test of time [1, 2, 3].

Over the last three decades, researchers have sought to understand the factors that influence eye movements to gain insights into our visual capabilities. One prominent theory advanced by many researchers is that humans make eye movements to regions that are most salient and computable from low-level image features, as defined by Treisman et al. [4]. Itti et al. [5] developed the first saliency maps by combining low-level image features such as color, intensity, and orientation at a given location, providing an image computable prediction of possible fixation location. Many subsequent elaborations of the model have been developed since then [6, 7, 8, 9, 10]. Saliency maps have shown reasonable success in predicting and explaining variance in eye movements.

More recent studies from the last two decades have shown that eye movements vary significantly across tasks, and that saliency alone cannot explain these differences, whether be it searching for an object or person [11, 12, 13, 14, 15], executing a motor action [16, 17, 18], navigating in an environment [19, 20], or identifying faces [21]. Humans look at regions relevant to the task, and eye movements during such tasks have

been shown to optimize task accuracy [22, 23, 24, 25], implying that humans direct their eye movements to actively seek information that could aid task performance.

But what guides eye movements in the absence of a task? As shown in the example situation, humans in the real world do not always have a fixed task when directing their eye movements. This is called free viewing [26, 6, 27]. Over the years, researchers have used free viewing to support the theory that people make eye movements to salient regions [28, 29, 30, 31, 32, 33, 34]. More recently, studies have shown that even during free viewing, people make eye movements to objects [35] or meaningful regions measured by observers’ subjective ratings of the meaningfulness of segmented local patches of scenes (meaning maps [36, 27]) rather than solely salient. This highlights the involvement of higher-level cognitive areas in directing human eye movements even in the absence of a task.

Prior studies have established defining characteristics of human gaze patterns under free-viewing conditions. Human subjects consistently direct their eye movements toward people present in the scene [37, 38], textual elements [39, 40], and distinct objects [41, 42]. More specifically, humans look at objects judged to be meaningful [43], or objects being gazed at or grasped by individuals within the image. Furthermore, initial fixations on a novel scene are overwhelmingly concentrated in the middle of the frame ([44, 45]), a well-documented phenomenon known as the center bias.

These observations raise two fundamental questions:

1. How do these universal and distinct fixation behaviors arise if eye movements during free viewing lack an explicit underlying task? Do human subjects adopt a default task while free viewing?
2. If so, are their eye movements optimized for this default task, much like the fixations observed under specific, directed tasks?

In this thesis, I address these questions across two chapters. The first chapter investigates the fixation patterns of human subjects observed during free viewing. I hypothesize that during free viewing, subjects inherently aim to understand a scene, actively directing their gaze to regions critical for this comprehension. To test this, I created a dataset of image pairs featuring minimal visual alterations but significantly altered descriptions from human subjects. Crucially, these pairs maintain similar low-level salient features across the images, isolating the semantic factors that guide eye movements. We refer to these image pairs as Winograd Images (WI).

I recorded the eye movements of human subjects across different observer groups as they viewed the Winograd pairs under four conditions with distinct task instructions: free viewing (no assigned task), scene description (describe the scene), object search (locate a target object), and object counting (count the number of specific objects). If subjects are trying to understand the scene during free viewing, their fixation patterns should closely mirror those of subjects in the scene description condition. Since Winograd image pairs fundamentally alter scene understanding, the hypothesis would predict distinct fixation patterns between the two images in a pair during scene description and free viewing. Conversely, this variance in fixation patterns should be minimal during the object search and counting conditions.

Crucially, this framework also predicts exactly where these eye movements will be directed. I hypothesize that during both free viewing and scene description conditions, fixations will be disproportionately drawn to people present in the scene and to specific objects critical to understanding the scene, rather than merely defaulting to regions of high saliency. To definitively test whether these directed fixations serve an active functional role, I also detail a second experiment in which human subjects were required to maintain fixation on either objects critical to scene understanding or objects not critical to scene understanding. This allows us to determine if fixating on critical

objects causally improves a subject’s ability to accurately describe the scene.

Building upon these behavioral findings, the second chapter explores this theoretical framework computationally. If the fixation behaviors of human subjects during free viewing, such as center bias, looking at people, text, or objects critical to scene understanding, are indeed the result of optimizing for scene comprehension, then these patterns should emerge naturally in an artificial agent with similar biological constraints that is optimized to maximize scene comprehension. I formalize this as the task-optimization theory of free-viewing eye movements. To test this, I detail the development of a novel computational agent with simulated human-like foveated vision and tasked purely with maximizing its understanding of a given scene.

The agent requires a robust scene-comprehension model to assess how varying levels of visual exploration, driven by simulated foveation, impact its understanding. To achieve this, I integrated a compact vision-language model (VLM)—capable of human-level scene comprehension—with a simulated foveation module. This setup allows the agent to actively explore a scene and generate textual descriptions based solely on foveated visual input. To quantify the quality of this exploration, these text outputs are transformed into vector embeddings within a semantic space. By computing the cosine similarity between these foveated embeddings and the VLM’s full-resolution, ground-truth description, I established a precise metric for semantic accuracy.

The core of the agent utilizes a convolutional neural network trained via reinforcement learning (RL). The semantic accuracy derived from the VLM serves as the reward signal during this active, foveated exploration, culminating in the foveated RL for scene understanding (fRL-SU) model. Crucially, unlike traditional predictive architectures [46, 47, 48] that are trained directly on massive datasets of recorded human fixations, the fRL-SU model utilizes absolutely no human behavioral data; its spatial exploration

strategy is trained strictly to maximize scene comprehension.

To evaluate whether human-like gaze is an emergent property of this active vision strategy, I compare the simulated fixations of the fRL-SU agent directly with empirical data from human subjects’ free viewing of scenes curated to depict complex social interactions, actions, or implied actions. Specifically, I assess whether the computational agent naturally learns to look at people, text, and objects critical to scene understanding at frequencies similar to biological vision. Furthermore, to confirm that scene comprehension is the specific computational goal driving this behavior, I compare the fRL-SU model’s performance against agents optimized for alternative visual tasks, such as search and scene classification (DeepGaze [49]), as well as established saliency architectures [6, 5]. I also trained the fRL-SU model on extremely high and low foveation levels to understand the constraints that foveation imposes on scene comprehension. Furthermore, I investigate whether these human-like gaze patterns emerge naturally when the model is trained without ground-truth descriptions, guided solely by the goal of minimizing uncertainty (entropy) in the VLM’s descriptions. Finally, I assess the model’s generalization by evaluating its predictive power against a large-scale, demographically diverse external dataset [50]. Ultimately, this chapter demonstrates that optimizing scene understanding within a foveated architecture provides an explanation for the universal patterns of human visual exploration during free viewing.

Chapter 1

Free-viewing eye movements aim to understand scenes.

1.1 Abstract

Human eye movements while identifying a face, searching for targets, and executing motor actions are directed to regions contributing to task accuracy. However, what humans look at and do when free-viewing a scene without a specific task is not well understood. We show that observers' free-viewing fixations are similar to fixations of observers instructed to describe the scenes and dissimilar to fixations of observers counting objects or searching for specific objects. Small visual alterations to images that change a scene's understanding, but not the most salient or its meaning map, alter where humans most frequently fixate. Free viewing fixations are more frequently directed to people and objects critical to the understanding of a scene (elements that, when erased from the scene, maximally alter the scene's description) rather than the most salient, most meaningfully judged scene region (meaning map), or the object perceived to be grasped or gazed at. By having observers describe the scene while maintaining fixation on objects relevant or irrelevant to scene understanding, we show that eye movements during free viewing are functionally important to understand scenes accurately. The theoretical framework also explains the higher frequency of fixations on people than on objects for most scenes because, when people are erased, scene descriptions are maximally altered. Thus, we conclude that an important default task during free viewing for the human brain is to comprehend scenes, reflected by frequent eye

movements toward people and objects that maximize accurate scene understanding.

1.2 Introduction

Eye movements are a critical part of human vision. Ever since Buswell (1935) [1] and Yarbus (1967) [2] conducted their classic eye-tracking studies, researchers have tried to understand what drives and influences eye movements. One of the prominent theories proposed two decades ago is that people make eye movements to salient regions. Saliency was originally defined (Itti et al. [5]) in terms of a combination of low-level features such as color, intensity, and orientation at any given location relative to its immediate surroundings, with many subsequent elaborations of the model [5, 6, 7, 8, 9, 10]. Since then, studies have shown that when humans perform specific tasks such as searching for targets [11, 12, 13, 14, 15, 51], executing a motor action [16, 17, 18], navigating in an environment [19, 20], or identifying faces [21], they do not look at the most salient object/region. Instead, humans look at locations that contain objects or visual features relevant to the task [52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63], and/or that allow maximizing task accuracy [23, 25, 21]. When instructed to artistically evaluate or describe scenes using keywords, observers fixate on objects more than salient regions [41].

Humans are not always engaged in specific tasks. They might look through a bus window, sit on a park bench, wait for a restaurant table, and explore a scene with no particular task. This is often called free viewing [26, 6]. Fixations during free viewing have been commonly used to support the theory that people direct eye movements to salient regions [28, 29, 30, 31, 32, 33, 34]. More recently, studies have shown that even during free viewing, people do not direct their eyes to salient regions but rather to objects [35] or meaningful regions measured by observers' subjective judgments of the meaningfulness of segmented local patches of scenes (meaning maps; subsequently referred to as locally meaningful regions/patches) [64, 27].

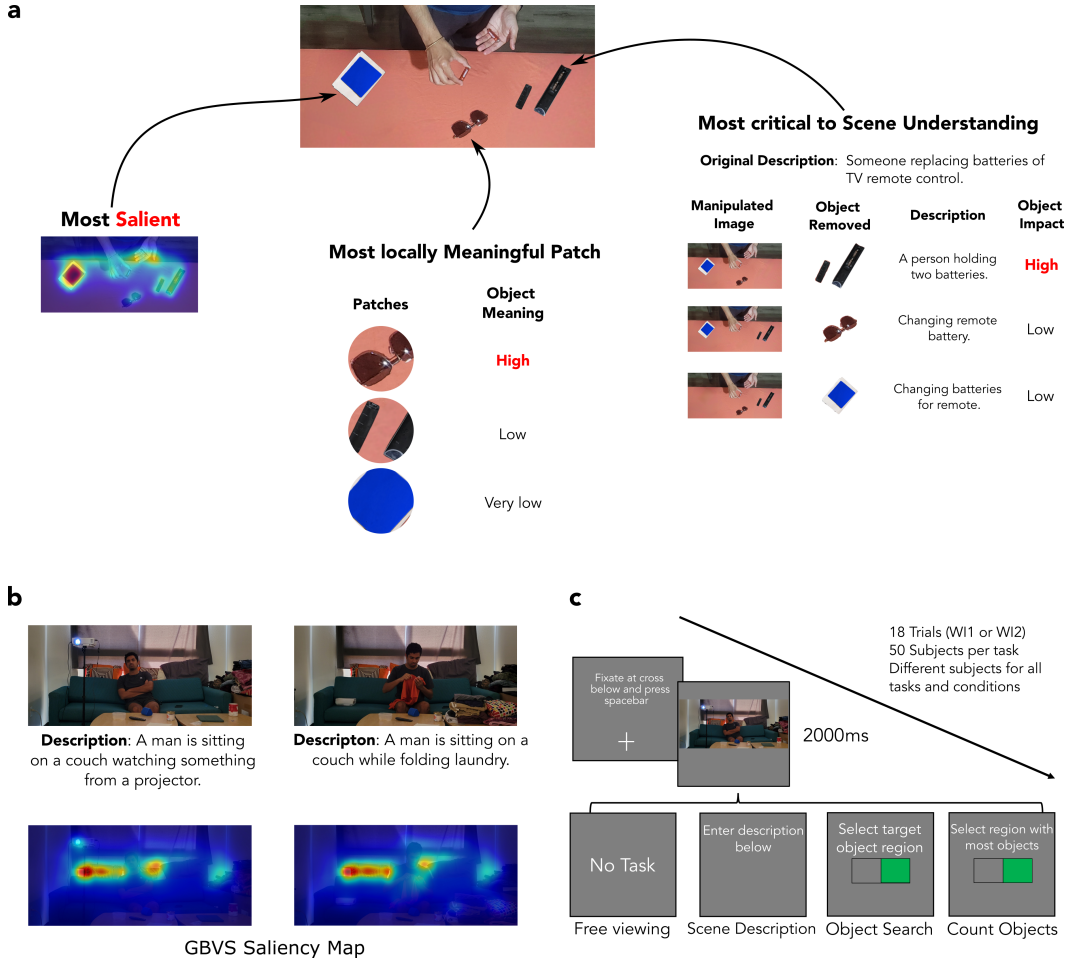


Figure 1: (a) Example of a scene where the object corresponding to the most locally meaningful patch (meaning maps), the most salient object, and the object most critical to understand the scene are different. The most salient object is the contrasting blue paper, while the most locally meaningful patch corresponds to the sunglasses. The image is described as “Someone replacing batteries of TV remote control.” The hands, the batteries, and the TV remote control are objects critical to understanding the scene, while the salient blue paper and sunglasses are irrelevant to the scene’s meaning. This image is not part of the study but included as a simple example to illustrate the concept of objects critical to scene understanding. (b) 1st row: Winograd image pair created to dissociate saliency from relevance to scene understanding with corresponding descriptions from observers. The projector is relevant to the scene understanding for the left image, while the clothes are for the right image. 2nd row: prediction of the GBVS saliency model for both Winograd images is similar, even though the description and the objects relevant to scene understanding vary. (c) Procedure for the four experimental conditions (free viewing, scene description, object search, and counting objects) of the eye movement study with 50 participants per condition (between-subjects design).

Here, we introduce a new theory. We hypothesize that, during free viewing, humans aim to understand a scene. To rapidly do so, they direct their eyes to regions critical to

scene understanding rather than looking at salient or regions judged to be meaningful. An important distinction between a locally meaningful region and one critical to scene understanding is that the latter considers how the region, people, or object contributes to understanding the whole scene. For example, Figure 1a illustrates the concept with an image described by observers as: "Someone replacing batteries of a TV remote control." The highly contrasting blue paper is the most salient object, and the sunglasses are the most locally meaningful object. However, the inverted remote with its battery cap open is the most critical object to the scene's understanding. How would one measure the contribution of an object to the scene's understanding? We propose a new method by having observers describe the scene and assess the impact of removing one object at a time on the scene descriptions. Removing the remote control from the image of Figure 1a greatly changes the description of the scene, while eliminating the sunglasses or the blue paper does not (see Figure 1a, right side). Thus, our theory hypothesizes that the remote control, critical to scene understanding, would attract more human fixations than the blue paper or the sunglasses. Furthermore, our second hypothesis is that fixating the object critical to scene understanding has a functional role in extracting the information required to understand it. The theory would predict that directing the high-resolution fovea to the remote control would be important to understand the scene (accurately describe it), while fixating on the blue paper would not. Although previous work has related eye movements to observers instructed to describe scenes [65, 66, 67], these studies have not examined how free-viewing fixations relate to the contributions of each object to scene understanding, evaluated other tasks or models, or experimentally investigated the causal functional importance of fixations on scene understanding.

Our initial methodology and analysis focused on predicting human fixations on different objects. In the latter part of the paper, we generalize the analysis to assess whether the theoretical framework can predict the well-documented frequent fixations

on people (bodies and faces) [37, 68]. For example, deleting the hand with the batteries in Figure 1a would maximally alter the description of the scene. We also assess how social cues fit the theoretical framework of eye movements to maximize scene understanding.

One of the main challenges in evaluating the hypothesis that eye movements are directed to people or objects that maximize scene understanding is that in most image data sets [69, 70, 71, 72, 73, 74] the person or object critical to scene understanding is often also the most salient object and the locally most meaningful region. This is a consequence of photographers’ practice of typically positioning the people or objects critical to the scene’s understanding as salient and central in the photograph [75]. To overcome this challenge, we developed a new dataset that decouples salient regions from regions that contribute to scene understanding.

We created pairs of real scenes that were minimally altered visually but resulted in large changes in the scene’s understanding (Figure 1b; left) while maintaining the saliency maps of the scenes and the local meaningfulness of each region. These small visual changes across image pairs could be manipulations of an object’s position, its substitution by another object, or an actor’s posture. We refer to these image pairs as Winograd Images (WI), inspired by the Winograd schema developed for sentences for which minor modifications of a word in a sentence lead to a large change in the sentence’s meaning [76]. To determine which objects were critical to the scene’s understanding, we assessed how deleting each object impacted the accuracy of scene descriptions relative to the intact images.

To test our first hypothesis, we had different groups of observers view our Winograd pair image dataset following different condition instructions (experiment 1): search for a specific object (object search; OS), describe the scene (scene description; SD), count the number of objects (counting objects; CO), or critically, free viewing (no instruc-

tions; FV). If observers execute eye movements during free viewing to understand the scene, we would expect various specific results. First, fixation patterns during free viewing should be similar to those obtained from observers during the scene description condition, which explicitly requires the observer to understand a scene to describe it. Fixation locations during free viewing should be less similar to the fixations of observers executing other specific tasks that do not require understanding the scene, such as searching for an object or counting objects. Second, Winograd image pairs, which alter scene understanding, should result in varying fixation patterns (across Winograd images) for the scene description and free viewing conditions, but less for the object search and counting conditions. Third, fixations during the free viewing and scene description conditions should be directed to objects critical to scene understanding. In contrast, fixations during the object search would be directed to the object searched (target), and fixations during object counting should be equally spread among objects. As benchmark comparisons, we also evaluated the ability of a saliency model (Graph-Based Visual Saliency (GBVS) [6]), a human eye movement-trained deep neural network (DeepGaze [49]), and predictions of the most meaningful local regions (meaning maps [64]) to predict the fixations across Winograd pairs. We hypothesize that the objects critical to scene understanding in our data would be fixated more frequently than the most salient or locally meaningful region (as measured by meaning maps) for the free viewing and scene description conditions.

To test the functional role of fixations during free viewing on scene understanding, we conducted a second experiment where observers described a scene while maintaining fixation at an object critical to scene understanding or another object irrelevant to the scene understanding. Suppose fixations have a functional role in rapidly extracting the information necessary to understand a scene. In that case, scene descriptions should be more accurate when observers maintain fixation on the objects relevant to scene understanding than when fixating on objects irrelevant to understanding the scene.

1.3 Methods:

Winograd image pairs (WI)

The experiment stimuli include 20 pairs of Winograd images photographed within the University of California, Santa Barbara premises. The images include indoor, outdoor, and table scenes. Each pair was ensured to have almost the same set of objects and positions. The small visual changes across image pairs could be manipulations of an object’s position, its substitution by another object, or an actor’s posture. The pairs were split into two sets: Winograd Set1 and Winograd Set2 [77]. The changes greatly alter how the scene was described while aiming to preserve the lower-level saliency and meaning maps. However, some changes would be introduced in the saliency and/or meaning maps by replacing an object or a variation of its location.

Each image pair was carefully curated to ensure that at least five random people described the pair differently and were consistent for a given image. Out of 180 descriptions (36 images * 5 descriptions each) in our entire dataset, 149 were consistent with their corresponding image (at least three descriptions were consistent for each image). Two pairs of Winograd images were removed from the dataset because their gold standard descriptions were inconsistent and vague. All of these images feature a person (or parts of a body) and tend to depict some future action or behavior that forms the scene’s understanding.

Eye tracking and experimental setup

Eye movements were recorded using an EyeLink 1000+ desktop-mount eye tracker (spatial resolution: 0.01°) with a sampling rate of 2,000 Hz. Participants sat 75 cm away from a 19-inch monitor, so that the screen subtended a visual angle of $26.6^\circ \times 21.8^\circ$ at $1,280 \times 1024$ pixels. The height of a stimulus image was approximately 12.7°

of visual angle. Head movements were minimized using a chin and forehead rest. A velocity threshold of $22^\circ/\text{sec}$ and an acceleration threshold of $4000^\circ/\text{sec}^2$ were used for the detection of saccades. Eye movements were recorded from the left eye. The experiment was controlled with SR Research Experiment Builder software.

Eye movement experiment with four conditions

In the first part of this study, four between-subjects eye-movement conditions were performed: free viewing, scene description, object search, and counting objects. Each condition involved 50 participants who were asked to complete 18 trials in which they viewed one set of Winograd images from all pairs (25 participants per image). Each image was shown for two seconds and was followed by condition-specific instructions. Below, we describe the instructions given to participants for the four conditions.

Free Viewing: Participants were instructed to view the scene naturally. No explicit tasks or instructions were given to them.

Scene Description: Participants were instructed to describe the presented scenes. After they viewed the scene, they typed their description with no time limit.

Object Search: Participants were instructed to search for an object within the scene. Before the start of the trial, they were given the name of the object they needed to search for, and after the trial, they reported back whether the object was on the left or right side of the image. By design, the image contained the object searched at the exact location across the Winograd image pair.

Counting Objects: Participants were instructed to count the objects on each image's left and right sides. After the trial, they reported which side of the image had more objects (see Figure 1b for the experiments' procedural flow diagram).

Fixation prediction models

The study used four fixation prediction models. The models' fixation prediction heatmaps were compared against the fixation heatmaps collectively made by participants in each of our eye-tracking experiments.

Graph Based Visual Saliency (GBVS)

A bottom-up saliency model proposed by Harel et al. [6] constructs a computational graph using Markov chains on top of the extracted image features to generate a heatmap of possible fixation locations. This study uses an implementation provided by Kümmerer [78] for the GBVS model. We chose the GBVS model because it is one of the top saliency models [79] relying purely on low-level image features to compute its saliency map, true to the original definition of saliency.

DeepGaze

DeepGaze is a neural network model trained on image features extracted from VGG-19 Convolutional Neural Network (CNN) [80] along with human fixations on those images while free viewing [49]. The model produces a heat map of possible fixation locations given an image. This study uses an implementation provided by Kümmerer [81]

Meaning maps

Meaning maps is a crowd-sourced model developed by Henderson et al. [64], which uses the subjective ratings provided by people on how meaningful they find local circular patches of an image. Each image was divided into overlapping circular patches, which were then randomized across all images. Different individuals then rated small portions of these patches to determine their meaningfulness. The procedure ensured each patch of images had three raters, and the overlap between the patches ensured that any region of the image had ratings from 27 raters. To facilitate the creation of meaning

maps for all the images in this study, 48 Amazon Mechanical Turk participants rated approximately 300 circular patches, each measuring 3 dva in size. The final result was a map of meaningful locations in a scene corresponding to possible locations people might fixate on while viewing the scenes. The procedure proposed by Henderson et al. [64] also included creating and rating 7 dva patches, which was not performed in this version of the meaning maps. Since the image height was 12.7° , using a 7 dva patch implied that every patch had multiple objects that could be recognized by the raters, making most of the image’s regions meaningful and, therefore, the 7 dva map becomes flat and averaging them with the 3 dva maps did not change the resulting meaning maps. Refer to Figure 4b for a procedural flowchart of all these models.

Scene Understanding map

The scene understanding map visualizes an object’s contributions to understanding a scene by quantifying the change measured in subjects’ descriptions of the scene after the object was removed. Creating a scene understanding map requires several steps described below: 1) Creating images with individual objects removed; 2) Collecting descriptions of images with objects removed; 3) Establishing a gold standard description for the intact image; 4) Assessing the similarity of the scene descriptions with an object removed to the scene description of the intact image; 5) Generating the scene understanding map. Below, we describe each of these steps.

Creating images with individual objects removed

Each image of the 18 Winograd pairs (36 images) was digitally manipulated to create versions with one object removed at a time. The resulting total number of images was 330. Each image had a total of 5 to 10 objects removed. Figure 4a shows an example of an image with some of its digitally manipulated versions. The photo editor app on the Samsung Galaxy S21 (version 3.4.2.43) was used to remove objects from these scenes.

Descriptions of images with objects removed

The dataset was split across the Winograd set, so participants did not see both images from the pair. To ensure that participants do not see more than one version of each image, each set was divided into 11 unequal groups, each containing exactly one copy (original or manipulated) of each image. This implied that all participants provided descriptions for at most 18 images. One hundred ten Amazon Mechanical Turk participants participated in this study, providing five descriptions for each image (or its digitally manipulated copies) in the dataset. Before this experiment, five images were used as a pre-test to ensure that observers followed the task instructions. Observers who did not accurately describe the pre-test images were not allowed to proceed to the main test with the 18 images.

Gold Standard Descriptions

To identify the impact of an object on scene understanding, we compared the descriptions of the original image to those collected for the manipulated image that had that specific object removed. We collected five descriptions for each of the original Winograd images. The gold standard description was defined as the best description among the five for each image. To establish which description was best for each image, a new group of fifty subjects (twenty-five per Winograd Image set) selected the description that best described each of the original 18 Winograd pairs. Subjects were randomly assigned to one of the Winograd pair sets (18 images). In each trial, the participant was shown an image with 7 descriptions (5 descriptions given to the image + two random descriptions from other images). Participants had to select the description that best described the presented scene. The two random descriptions were used to identify participants who did not follow the instructions. All data from participants who selected random descriptions were eliminated from the study. Importantly, the description with the highest vote (as the best description) was chosen for each image

as the gold standard description.

Description similarity ratings to determine the contribution of objects to scene description

Eighteen Amazon Mechanical Turk participants were asked to rate the similarity of the descriptions using a scale from 1 to 10, with 10 indicating highly similar and 1 indicating very low similarity. In each trial, the participant saw the gold standard description for an image, followed by descriptions corresponding to each object being removed from the scene. Since there were five such description sets for each Winograd image, each participant had to finish 180 trials (36 images * 5 sets of descriptions). Refer to Figure 4a for the procedural flowchart of this process.

Heatmap generation for scene understanding map

To generate a heat map that visualizes the contribution of each object present in the scene to the scene description, we inverted the scale to have higher scores for dissimilar descriptions. Then we computed the median rating score corresponding to the description when each object was removed. The median score was assigned to all the pixels within the object’s corresponding bounding box (defined in methods; refer to Figure 4a). To highlight relative contributions within each image, we normalized the ratings by subtracting the lowest-rated object’s rating from all the other objects’ ratings and then dividing the map by the highest score for each image (scale-inverted normalized ratings). We followed Stoll et al. [35] to generate the preferred fixation location in the images by using scale-inverted, normalized ratings corresponding to the removal of each object from the scene. Using the bounding boxes for our object, we modeled the ratings as 2D Gaussians centered at each object with its amplitude corresponding to scale-inverted normalized rating and the Gaussian horizontal and vertical standard deviation derived as a fraction of the size of the bounding box (fraction=0.29 along x-direction and fraction=0.34 along y-direction). Figure 4a shows an example of a

scene understanding map.

Defining SU-relevant and SU-irrelevant objects

Objects in each scene were categorized based on their scene understanding map score (SU-relevant and SU-irrelevant). Objects that belong to the SU-relevant category had the highest impact (most critical to scene understanding among objects in that image) on subjects’ scene descriptions when erased. If erasing them did not impact the subjects’ scene description, they would belong to the SU-irrelevant category. For each Winograd pair, the same set of objects belonged to the SU-relevant category in one image and to the SU-irrelevant category in the other image of the pairs (refer to Figure 4a). Our analyses did not consider objects that had a low impact on the subjects’ scene descriptions for both images of a Winograd pair.

Measuring objects to be grasped and/or gazed at in the dataset

Fifty Amazon Mechanical Turk participants were divided equally between the Winograd sets and were shown the images with objects in the scene covered with black boxes. This was done to remove any contextual information that could influence their judgments. Participants were asked to click on the box they perceived to be grasped or gazed at by the person in the scene. The box with the maximum clicks was considered the object perceived to be grasped or gazed at for each image. Figure 9a shows the experimental procedure and some examples. Images, where the SU-relevant object differed from the object perceived to be grasped/gazed at, constitute images with No Gaze cues.

Forced fixation scene description experiment

Sixty participants were split into four equal groups, and each group saw 18 trials. Two groups were assigned to images from Winograd set 1, while the others were assigned to set 2. A between-subjects design across the Winograd pairs ensured that observers did not use their knowledge of one image to interpret its corresponding pair. Within each set, the two groups were asked to fixate on nine SU-relevant and nine SU-irrelevant locations. Each group saw a unique combination of image and fixation location. The images were presented for 500 ms. All trials where eye movements (saccades larger than one dva) were detected within the presentation interval were discarded. On average, the discarded trials accounted for thirteen percent of all trials, resulting in thirteen descriptions per image and fixation location. Participants described each scene after it was presented. Because of the atypical nature of the task, observers were instructed to provide “their best guess of what was happening in the scene.”.

Repeating analyses including fixations/predictions on people in the scenes

Following the abovementioned procedure to generate SUMs, we recreated the scene understanding map by adding its prediction to all the people in the scenes. Similarly, we included the predictions of the other models on people. We also generated the measured fixation heat maps, including fixations on people. With these updated fixation/prediction heat maps, we recomputed the correlations of fixation heat maps across conditions and within the Winograd pairs for each condition. We also added fixations on people and the prediction of SU-relevant with the people category in the cumulative fixation distribution plots. Finally, we performed the AUROC analysis for all fixation prediction models, including fixations/predictions on people.

Dependent variables

Heatmap generation for measured fixations

Fixations on an image (a pixel white dot on a blank canvas) were convolved with a Gaussian kernel of standard deviation 0.5° visual angle. OpenCV2’s GaussianBlur function was used for this purpose. These convolved fixation points were added onto a blank map using their x and y locations to generate the heatmap.

Selecting top predictions from fixation models

Fixation prediction models can sometimes predict contrastive regions that are not objects of importance. These regions were identified by finding the maximum locations determined by the convolution between the map and a 2 dva uniform and normalized circular patch. If the model’s prediction landed on an object, the object’s bounding box was used to represent its prediction. If not, a bounding box was generated with the average dimensions of all the bounding boxes in the image.

AUROC analysis for the fixation prediction models

The shuffled AUC (sAUC) technique [82] was used for the ROC analysis to compute the performance of all fixation prediction models. For each image and observer combination, the prediction maps were thresholded at different levels, and the true positives (TPs) at each level were calculated by counting the number of fixations that fell in the areas above the thresholded level. The false positives (FPs) at each threshold level were calculated by sampling fixations (the same number of fixations as the positive set) from the same observer in other images (the negative set) in the dataset, and counting the number of these fixations that fall in areas above the threshold level. The sAUC was then calculated by finding the area under the ROC curve (TPs vs FPs) for each observer-image combination. The sAUCs were averaged across all observers and images to get a final value for each map. This analysis was conducted after removing

the fixations and model predictions for humans in the scenes. The error bars for the sAUCs were obtained from bootstrapping the images and subjects for 1000 trials.

Frequency of fixations/fixation time to object categories

Each scene had a maximum of 10 bounding boxes corresponding to different objects or regions. These bounding boxes were assigned specific identifiers or categories based on their relevance to scene understanding, as measured by the scene understanding map (SU-relevant and SU-irrelevant), the top predictions of the fixation prediction models (e.g., most salient), and condition-specific identifiers (search target from the object search condition). We counted the frequency of fixations or added the fixation times for these object/region categories across all participants. If an object category contains more than one object for a given image, the average fixations/fixation times to these objects were computed.

Inter-observer correlation of fixation maps

For each condition, the 25 subjects were divided into two random groups of 12 subjects each. The correlation of the heatmaps for each image was computed between these two groups. This process was repeated for 1000 trials. The average correlation across all images and 1000 trials was used as the inter-observer correlation for each condition. It should be noted that the correlations across Winograd images were computed using 25 subjects in each group. In contrast, within-observer correlation computations have half the number of subjects in the analysis. We computed the Winograd correlation with half the number of subjects to make a valid comparison with the inter-observer correlation. We used 1000 non-repeating combinations of 12 subjects out of the 25 and calculated the mean fixation heatmap correlation for all images for each condition. We performed a similar analysis for across-condition correlations (Figure 2c), where we take 1000 non-repeating combinations of 12 subjects for each condition across all images to compute the mean correlations.

Fixations executed by observers who did not understand the scene

The descriptions provided by participants in the scene description condition were rated for similarity against the gold standard description using the embeddings from the Large Language Model, Gemini [83]. For each image, the participants whose description rating fell below one standard deviation (SD) from the mean rating were classified as participants who did not understand the scene, and the rest were grouped as those who understood the scene. The 1st author and three other research assistants (RAs) from the lab judged the correctness of these descriptions (based on the gold standard descriptions). The mode of these judgments corresponds to the human judgment that was compared against the LLM categorization of descriptions based on SD cut-offs ranging from 0 to 2.5 below the mean similarity rating of the descriptions for each image. The fixation frequency distribution for object categories was obtained for both image groups, based on human judgments and all SD cut-offs.

Computing Similarity using Large Language Models (LLMs)

To compute the similarity of these descriptions to the gold standard (defined in the methods above), the embeddings of the large language model Gemini [83] were utilized. The model API provides access to convert text into learned feature embeddings. A cosine similarity score was used to measure the similarity between these embeddings. To ensure consistency, the same analyses were done using the embeddings of another large language model, GPT4 [84]. To ensure consistency with human judgments, we employed cosine similarity to assess the agreement between the LLM similarity metric and the human ratings obtained for all comparisons in the object erasure experiment.

Computing the lower and upper bound for similarity scores

Upper bound: Each Winograd image has five descriptions (defined in the methods above). A pairwise similarity comparison of all five descriptions was computed using

LLMs. The average similarity score across all pairwise comparisons across all images constitutes the upper bound for the similarity scores.

Lower bound: The gold standard description of each image was compared to the descriptions obtained from the force fixation condition for another random image in the dataset using LLMs. The lower bound was computed by averaging the similarity scores for 1000 such trials.

Statistical analyses

Bootstrap Resampling

Error bars for all our analyses were obtained using bootstrap resampling of subjects, images, descriptions, and ratings. Below, we describe the general procedure. We sampled 100,000 resamples of images and observers for all the fixation frequency analyses. We used 10,000 observer/image resamples for the fixation heatmap correlation analyses because of the computational cost of calculating heatmap correlations. The analysis, which quantified the inter-observer fixation heatmap correlations, used half of the subjects and 10 random samples of 12 out of the 25 subjects. For this analysis, we used only 1,000 bootstrap resamples for each of the ten different sample subjects ($n = 12$) due to computational constraints.

False discovery rate

All analyses involving more than fifteen comparisons had their significance levels corrected for the False Discovery Rate (FDR) using the Benjamini-Hochberg method [85], and the adjusted p-values were reported (q-value). For the cumulative fixation distributions, the comparison across SU-relevant and SU-irrelevant for each successive fixation depended on previous comparisons. Therefore, we added a simple, conservative modification to the FDR procedure, as detailed by the Benjamini-Yekutieli procedure [86].

All significances were determined using a one-tailed significance level.

1.4 Results:

Free viewing fixations are most similar to scene description

We created a total of 18 Winograd pairs [77]. Fifty subjects participated in each of the four conditions (a between-subjects design, with a total of 200 subjects). Within each condition, subjects were assigned randomly and equally to view one set of Winograd pairs (18 image trials). Figure 1c details the procedural flow of all four conditions.

All of our primary analyses removed fixations/model predictions on human faces or bodies in the scenes. This pre-processing enabled us to focus on fixations on objects that were critical versus those irrelevant to scene understanding, rather than on human figures, which were critical across all Winograd image pairs. Figure 2a shows an example of the fixation heatmaps (see methods for details) generated using 25 subjects for the four conditions. The gray regions indicate fixations on humans that were not considered in our primary analysis (but see Section 1.4 below for an analysis including fixations on people). Figure 2b shows the correlation of fixation heat maps across conditions (for 25 observers) computed for each of the 36 images (18 pairs) in our dataset. The fixation heat maps for free viewing were significantly more similar to the fixation heat maps from the scene description condition (FV-SD, $r = 0.54$) than those from the object search (FV-OS, $r = 0.32$, bootstrap resampling of observers and images for all analyses, $p < 0.0001$ vs. FV-SD, False Discovery Rate (FDR) adjusted for this and all subsequently reported p-values) or counting objects (FV-CO, $r = 0.42$, bootstrap, $p < 0.0001$ vs. FV-SD, see Figure 3). To compare the across-condition fixation heat map correlation to an upper bound, we estimated a within-condition (FV-FV) correlation of fixation heat maps, each using distinct groups of 12 observers (Figure 2c, see methods).

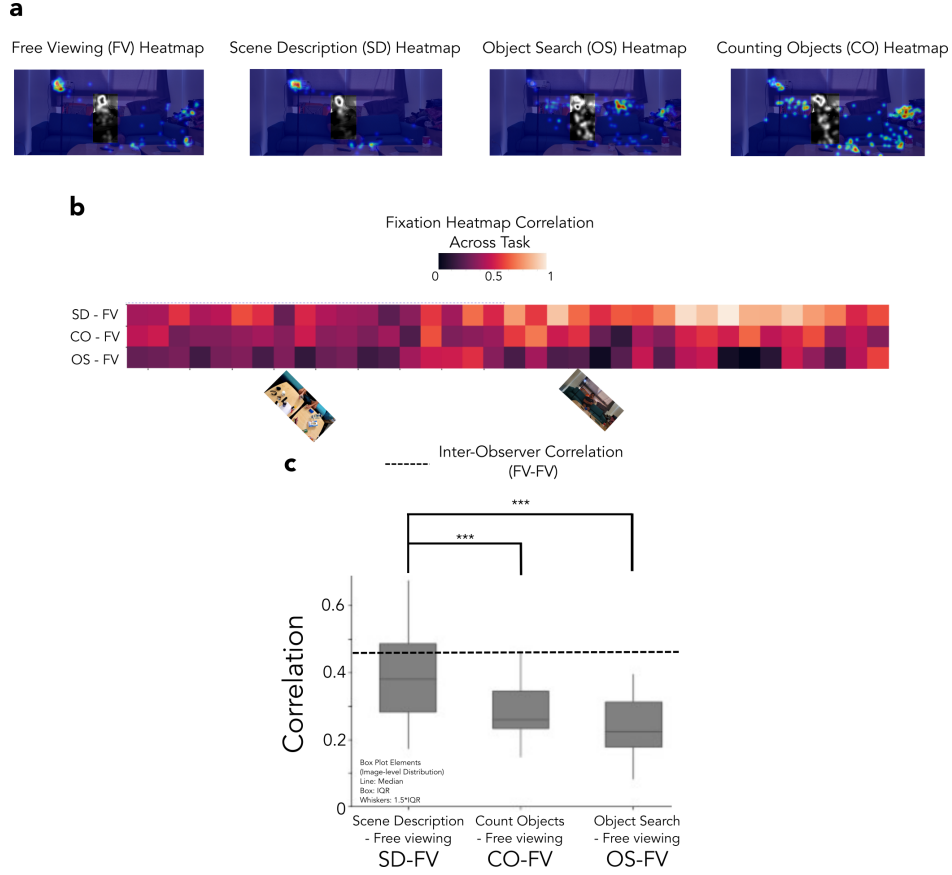


Figure 2: (a) Example of the fixation heat maps generated from 25 subjects for all four conditions. The grey heatmap regions indicate regions containing people and were excluded from the primary fixation analyses. (b) Visualize the correlation between free-viewing fixation heat maps and fixation heat maps from other conditions for each image in our dataset. The lighter the color, the higher the correlation. Three sample images are shown. There was a trend of higher correlations between the free viewing (FV) fixation heat maps and those of the scene description (SD) condition than with the heat maps of the object search (OS) and counting objects (CO) conditions. (c) The average correlation between the free-viewing fixation heat maps and scene description fixation heat maps (12 subjects) was significantly higher than the correlation with the fixation heat maps of the OS and CO conditions. (***) = $p < 0.001$). The dotted line shows the inter-observer correlation of fixation heatmaps from groups of 12 observers (FV-FV).

As with the $n=25$ analysis, we found higher FV-SD fixation heat map correlations ($r = 0.4$) than those for FV-OS ($r = 0.23$) and FV-CO ($r = 0.29$; bootstrap, $p < 0.001$ for all comparisons). Notably, the FV-SD fixation heat map correlation was the closest to the within-condition (FV-FV) inter-observer (see dotted lines in Figure 2c) heat map correlations ($\Delta r = r_{\text{FV-FV}} - r_{\text{FV-SD}} = 0.06$) relative to the other condition com-

parisons (the FV-OS vs. FV-FV correlations: $\Delta r = 0.26$; and the FV-CO vs. FV-FV correlations: $\Delta r = 0.17$).

Changes in scene understanding with small visual manipulations influence observer fixations

We compared the measured human fixation patterns across the Winograd image pairs for different conditions with other existing fixation prediction models: a purely low-level saliency model (GBVS) [6], a neural network model trained on human fixations and images to predict fixations on novel images (DeepGaze II) [49], meaning maps that quantify the meaningfulness of local regions of scenes [64], and scene understanding maps that quantify the contribution of each object to the meaning of the image. Figure 4a & b provides a flow chart of all the models we have used in this comparison.

To measure the contribution of an object to scene understanding, we developed a technique that measures the impact of an object present in the scene on the description provided by subjects. Figure 4a illustrates the procedure with an example. The original descriptions for the intact (i.e., no object removed) images were compared with the descriptions of images after the object removal to quantify the object’s impact on scene understanding. If the object is irrelevant to scene understanding, the scene descriptions with the object present or removed would be similar. In contrast, if the object is critical to scene understanding, removing it would greatly impact the scene descriptions. The similarity of the scene descriptions can be measured by relying on a separate group of subjects who view sentence pairs (without the image) and rate their similarity. Quantifying the impact of removing each object on the scene description allows us to create a map that visualizes the objects’ contribution to scene understanding (scene understanding map, SUM), which can be used to predict fixation distributions and the most fixated object (see methods and Figure 5a). Similar SUM results were obtained

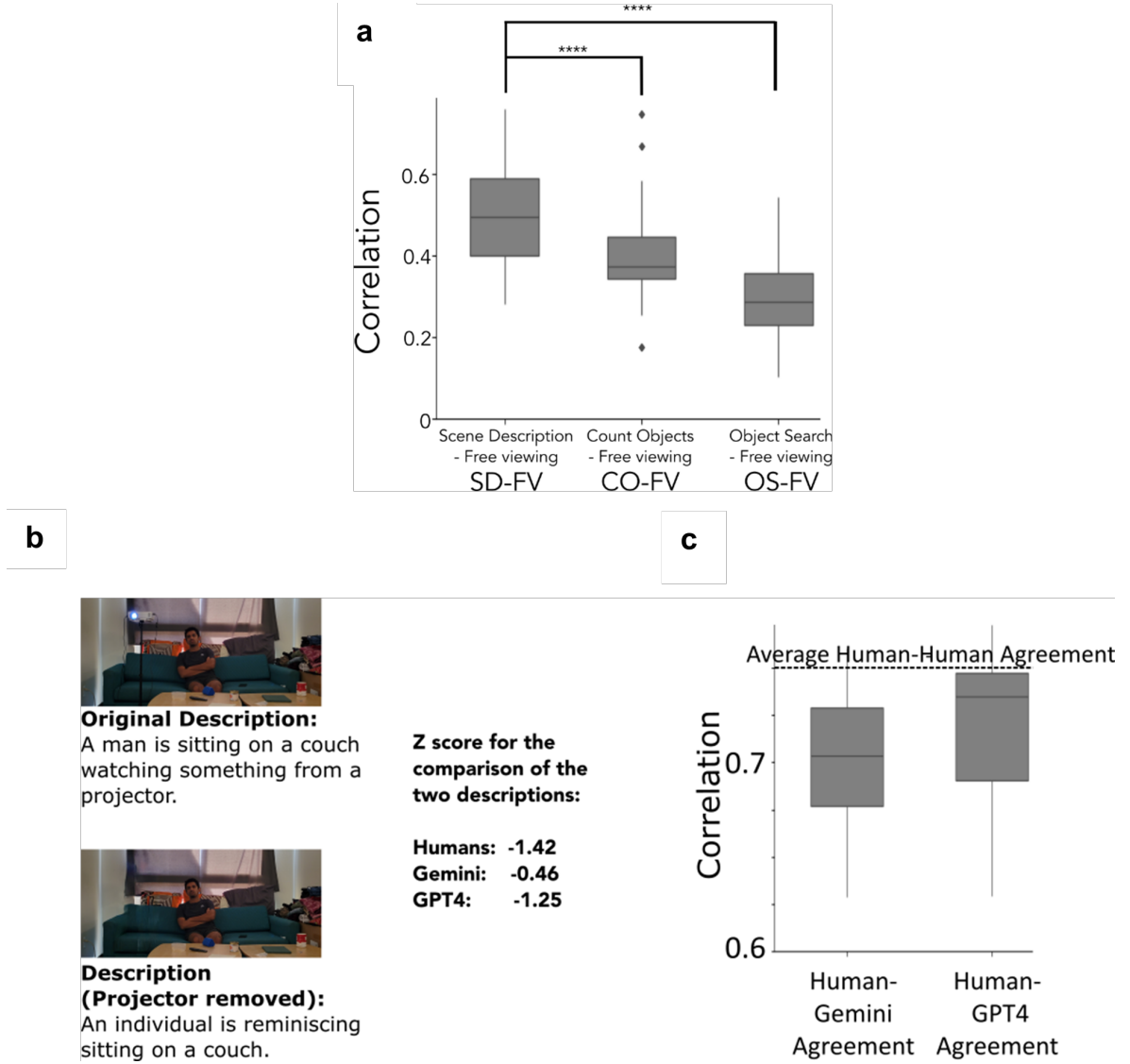


Figure 3: (a) Correlation across fixation heatmaps ($n=25$ observers) for different conditions (scene description, count objects, object search, and free viewing). We observed similar results to those of the analysis with groups of 12 observers per condition reported in the main text (Figure 2c in the main text). (b) A sentence description for the original image (top) and for the image with the projector removed (bottom). To the right, an example of the scores obtained from the LLM and human ratings for the similarity of two image descriptions. (c) The average correlation of individual human raters with semantic cosine similarity using GPT4 language embeddings and Gemini embeddings is comparable to the average inter-human rating correlation (dashed horizontal line). The above comparison used the human ratings (18 raters) and LLM embeddings obtained from the object removal experiment (averaged across all images). **** ($p < 0.0001$) A one-tailed bootstrapped analysis was conducted to test the significance of all the results. In (c), the line within the box indicates the median. The box spans the interquartile range (IQR), and the whiskers extend to the most extreme values within $1.5 \times \text{IQR}$ across participants.

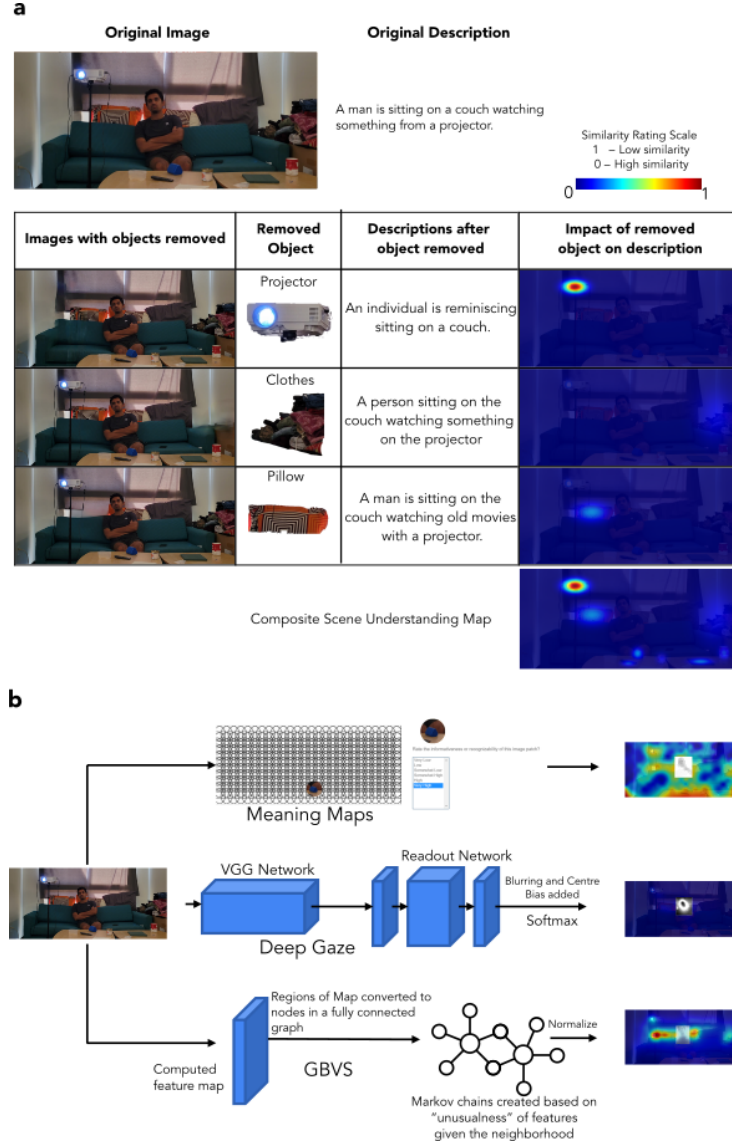


Figure 4: (a) Procedural workflow for creating scene understanding maps (SUM): The impact of an object on scene understanding was quantified by comparing the similarity (rated by humans) between descriptions that people provide when the object was present and removed. See Figure 3b & c for details on the consistency of LLMs with human raters on rating the similarity of these descriptions. The SUM was generated by placing a 2D Gaussian on each object with its amplitude determined by the impact of the object on the scene understanding and with a standard deviation determined by the size of the bounding box around the center of each object (SD range: 0.5 to 1.5 dva) ([35], see methods for details). (b) Workflow for creating meaning maps (top), DeepGaze (middle) maps, and GBVS (bottom) maps; fixation prediction models used in this study. The grey regions in the heatmaps indicate the predictions that fall on people in the scene and were excluded from the analyses (see section 1.4 below for results that include fixations on people).

using similarity ratings of descriptions based on automated Large Language Models (LLMs) based metrics (cosine similarity computed with the LLM embeddings of the sentences) rather than human ratings. The human-LLM rating agreement ($r_{\text{Gemini-Human}} = 0.7$, $r_{\text{GPT4-Human}} = 0.72$) was comparable to the average human-human ($r = 0.75$) agreement in the similarity ratings of these descriptions (see Figure 3 1b & c).

We hypothesized that altering the scene understanding with small visual manipulations would change the human fixations for the free viewing and scene description conditions. Thus, we predicted that the correlation of human fixation distributions across Winograd image pairs would be smaller than that for the saliency, local meaningful regions, and DeepGaze prediction maps. Similar to the human fixation distributions, our hypothesis predicts that the correlation across the scene understanding maps of Winograd image pairs would also be low. In addition, human fixation heat maps would be more similar (higher correlation) across the Winograd image pairs for the object search and counting objects conditions compared to the free viewing and scene description conditions.

Figure 5a shows examples of the predicted fixation heat maps of different models and measured fixation heat maps for humans for the four conditions for a Winograd image pair. The correlation between the human fixation heat maps of Winograd images for free viewing condition (Figure 5b) was significantly lower than the correlation between fixation heat maps of Winograd images for the object search (bootstrap, $p < 0.0001$) and counting objects (bootstrap, $p = 0.018$) condition. The correlation of Winograd image fixation heatmaps for the scene description condition was significantly lower than that observed in the free viewing condition (bootstrap, $p = 0.034$). The correlations of fixation prediction heatmaps between Winograd images for saliency, meaning maps, and DeepGaze were significantly higher (bootstrap, $p < 0.0001$ vs. free-viewing) than the correlations observed in free-viewing human fixation heat maps

between the Winograd pairs. In contrast, the correlation of the fixation predictions of Winograd image pairs for the SUM was the lowest and not different from that observed in human fixations for the free viewing and scene description conditions (bootstrap, $p > 0.38$).

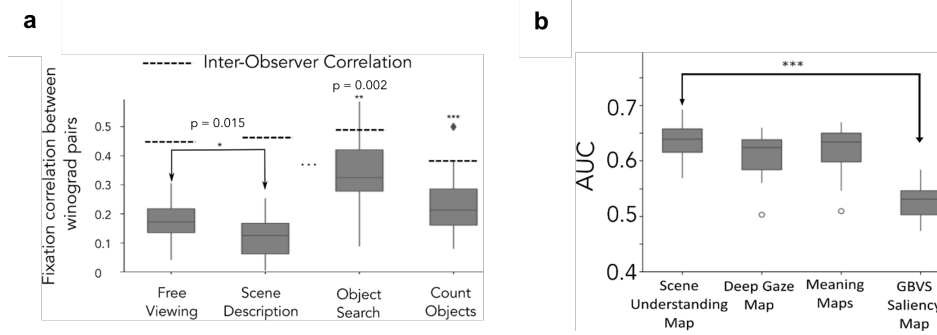


Figure 6: (a) Fixation heatmap ($n=12$ observers for each condition) correlation across Winograd pairs. The dotted lines indicate the same image inter-observer fixation heatmap correlation (12 vs. 12 observer heatmaps) averaged across the images. These results show that the difference between the same image inter-observer and the Winograd image-pair correlations is significantly higher for the scene description condition than for free viewing. The correlation difference observed for object search and counting objects is significantly lower when compared to free-viewing. (b) A shuffled AUROC analysis revealed that the SUM maps have the highest performance in predicting human fixations (25 participants per image, 36 images) during free viewing ($AUC = 0.64$) and was significantly different from GBVS maps ($AUC = 0.53$, bootstrap, $p < 0.001$) but was not significantly different from meaning maps ($AUC = 0.62$, bootstrap, $p = 0.18$) or DeepGaze ($AUC = 0.61$, bootstrap, $p = 0.94$). (***) = $p < 0.001$; ** = $p < 0.01$; * $p < 0.05$). A one-tailed bootstrapped analysis was conducted to test the significance of all the results. In both plots, the line within the box indicates the median. The box spans the interquartile range (IQR), and the whiskers extend to the most extreme values within $1.5 \times IQR$ across images for (a) and across participants for (b).

These variations in Winograd pair fixation heat map correlations across conditions cannot be accounted for by differences in inter-observer variability ($r_{\text{inter-observer}} = 0.45$ for free viewing; $r_{\text{inter-observer}} = 0.47$ for scene descriptions; $r_{\text{inter-observer}} = 0.49$ for object search; $r_{\text{inter-observer}} = 0.38$ for counting objects condition, see Figure 6a, dotted lines). In addition, as a control comparison, the correlations observed for a random pairing of images (shaded area in Figure 5b) do not result in the ordering observed in Winograd image correlations across conditions.

Even though the SUM is most helpful in predicting the most frequently fixated object, it also attains comparable accuracy at predicting all human fixations to the meaning maps and DeepGaze and is better than GBVS (shuffled AUROC analysis; see methods for implementation details and see Figure 6b for the results).

Observers look at regions/objects critical to scene understanding during free viewing

To assess where human observers fixate most frequently during free viewing, we categorized objects based on their contribution to the scene understanding map (see Figure 4a). We categorized objects as the most relevant to scene understanding (SU-relevant) if erasing them from the scene resulted in the largest change in the subjects' scene description. Similarly, we categorized objects as irrelevant to scene understanding (SU-irrelevant) if erasing them from the scene did not result in large changes in the subjects' scene description (see methods; the analysis also excluded people in scenes, see section 1.4 further below). To control for low-level visual properties of objects and object types, we designed our stimuli so that the same objects were SU-relevant for one Winograd image (example: projector in Figure 7a; left, clothes in Figure 7a; right) and SU-irrelevant to the complementary Winograd image (example: clothes in Figure 7a; left, projector in Figure 7a; right). Across the Winograd images, SU-relevant and SU-irrelevant categories were the same set of objects, providing a strong control.

We measured the frequency of human fixations on SU-relevant and SU-irrelevant objects for each condition. We also quantified the frequency of human fixations on the top predictions of the GBVS saliency model, DeepGaze, and the most locally meaningful region based on meaning maps. Figure 7b shows the fixation frequency for each object category averaged across all images and subjects. Results are shown for the four conditions. Observers fixated on objects more frequently when they were criti-

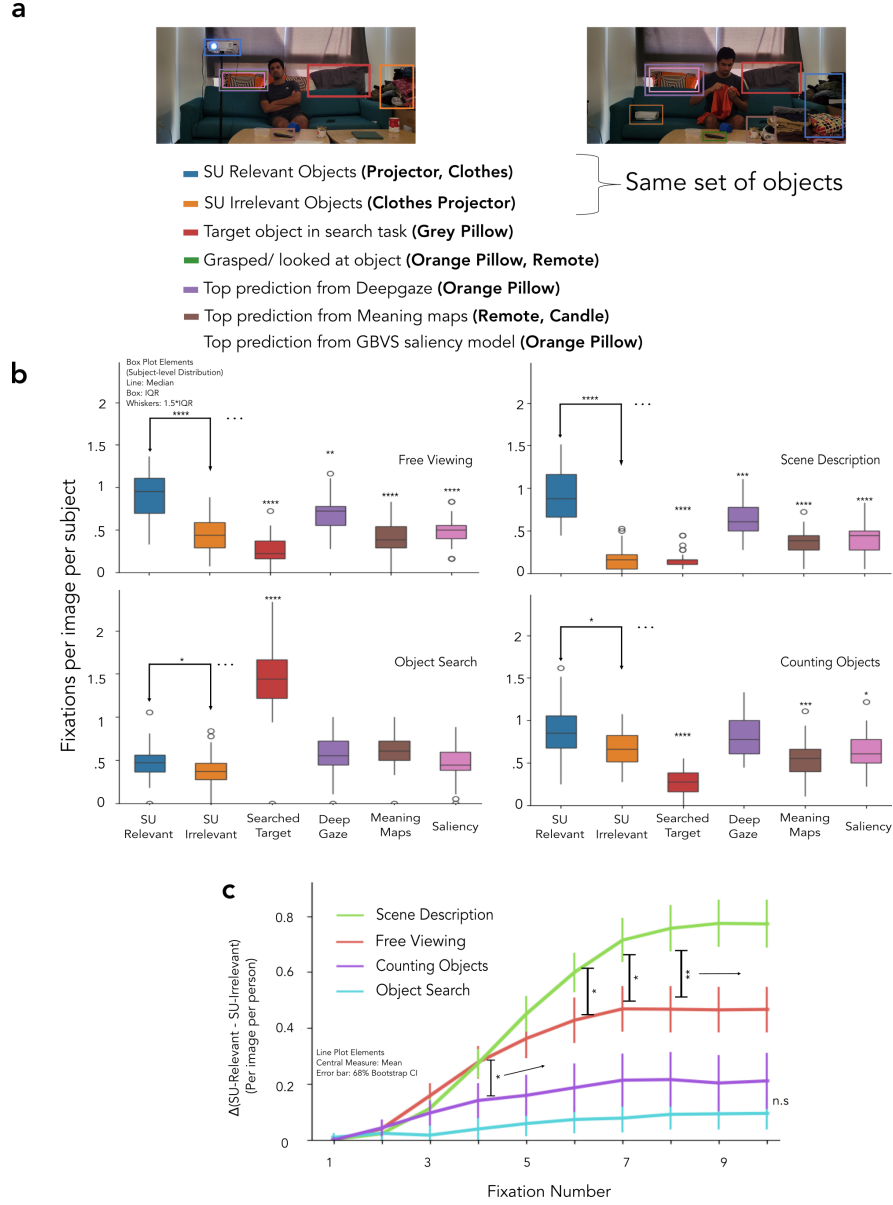
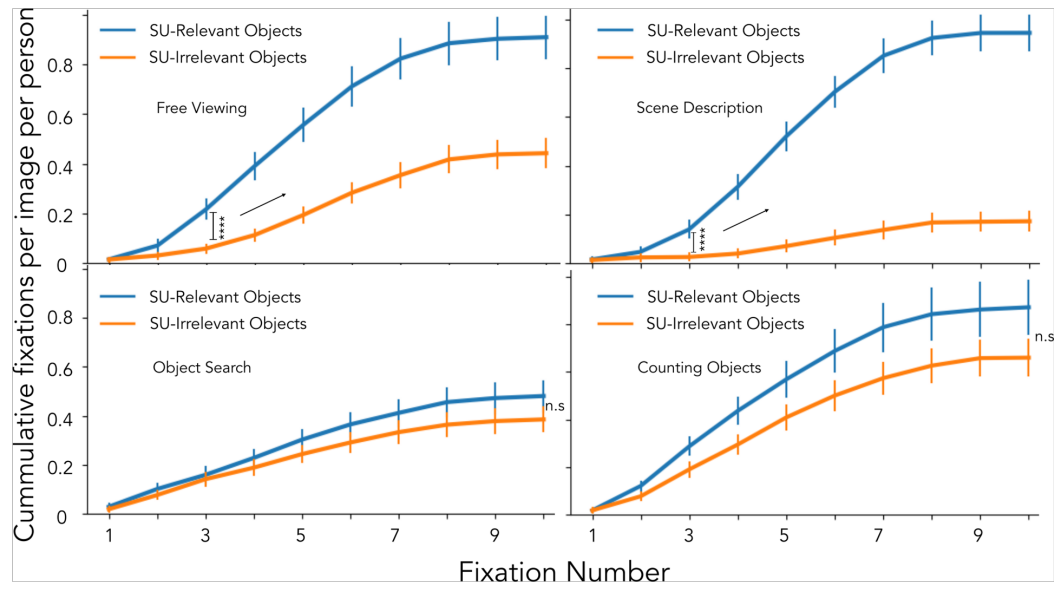
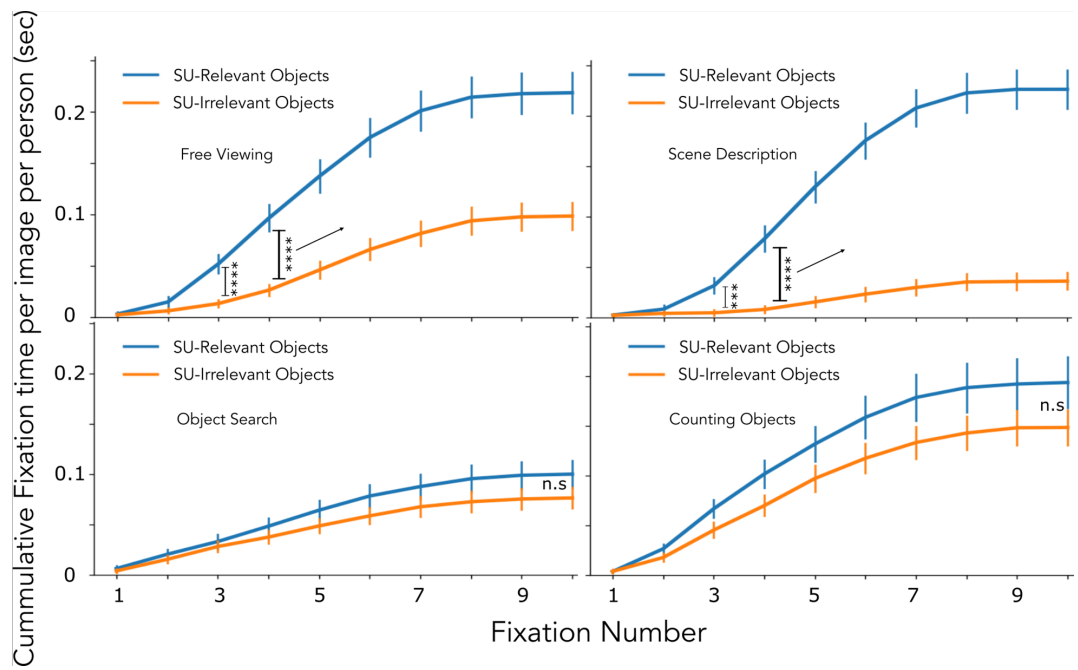


Figure 7: (a) Bounding boxes for objects in each Winograd pair were manually created, and the categories were assigned to them. The scene understanding relevant (SU-relevant) and the SU-irrelevant categories have the same set of objects across each Winograd pair. (b) Fixation frequencies based on the object categories in free viewing and scene description conditions show that the SU-relevant object was most fixated. The search target was most fixated in the object search condition. The SU-relevant object was fixated on the most for the counting objects, although the difference relative to the SU-irrelevant was smaller. (c) The difference in cumulative fixation frequency observed between SU-relevant and SU-irrelevant object categories was highest in the scene description condition, followed by free viewing. No significant difference was observed across any fixation for the other two conditions. (**** = $p < 0.0001$; *** = $p < 0.001$; ** = $p < 0.01$; * $p < 0.05$).

cal to scene understanding than when they were irrelevant for both the free viewing (SU-relevant vs. SU-irrelevant objects, bootstrap, $p < 0.0001$) and scene description conditions (bootstrap, $p < 0.0001$). Furthermore, in the free viewing and scene description conditions, the SU-relevant objects were fixated significantly more often than any of the top GBVS salient (bootstrap, $p < 0.0001$), DeepGaze (bootstrap, $p < 0.01$), or locally meaningful (bootstrap, $p < 0.0001$) objects. In the search condition, the searched target was fixated the most frequently (bootstrap, $p < 0.0001$). Still, a smaller but significant difference was observed in fixation frequency between SU object categories (higher frequency to SU-relevant vs. SU-irrelevant objects) for the object search (bootstrap, $p = 0.048$) and the counting conditions (bootstrap, $p = 0.029$).

To assess how observer fixation preferences develop through time, we analyzed the cumulative fixation frequencies to SU-relevant and SU-irrelevant objects for all conditions as a function of fixation number after image onset. For the free viewing and scene description conditions, the cumulative fixation frequency difference became significant after the 3rd fixation (bootstrap, $p < 0.0001$ for each condition, Figure 8a). No difference was observed for the other two conditions until the 10th fixation. Figure 7c compares the difference in fixation frequency across SU-relevant and SU-irrelevant objects as a function of fixation number. The difference in the scene description significantly deviated from that in free-viewing at the 6th fixation (bootstrap, $p = 0.043$), while the free-viewing deviated significantly from the object counting condition at the 5th fixation (bootstrap, $p = 0.049$). No significant difference was observed between object counting and object search conditions. An analysis that used time-weighted fixations rather than fixation frequency resulted in similar findings (Figure 8b & c).

a**b**

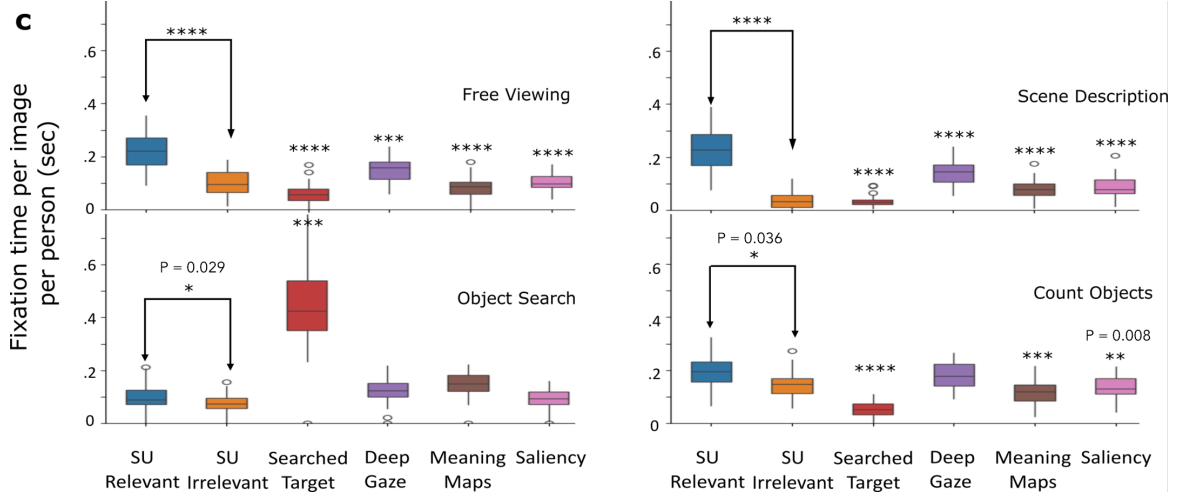


Figure 8: (a) Cumulative fixation frequency for SU-relevant and SU-irrelevant object categories shows a significant difference in free-viewing and scene description conditions from the 3rd fixation onwards. There is a trend of higher fixation frequency for SU-relevant vs. SU-irrelevant objects in search and counting-object conditions, but it did not reach statistical significance at any fixation measured (until the 10th). (b) Cumulative time-weighted fixations for SU-relevant and SU-irrelevant objects show similar result patterns for all four conditions. (c) Time-weighted fixation for object categories. This measure also shows the same trend as fixation frequency. Observers fixated longer on SU-relevant vs. SU-irrelevant objects. 25 participants per image, 36 images in all analysis. (**** = $p < 0.0001$; *** = $p < 0.001$; ** = $p < 0.01$; * $p < 0.05$). A one-tailed bootstrapped analysis was conducted to test the significance of all the results. Each analysis had its significance levels corrected for the False Discovery Rate ($\alpha=0.05$, 40 comparisons each in (a) and (b), 20 comparisons in (c)). In (a) and (b), the central measure is the mean, and the error bars show the 68% bootstrap confidence interval. In (c), the line within the box indicates the median. The box spans the interquartile range (IQR), and the whiskers extend to the most extreme values within $1.5 \times \text{IQR}$ across participants.

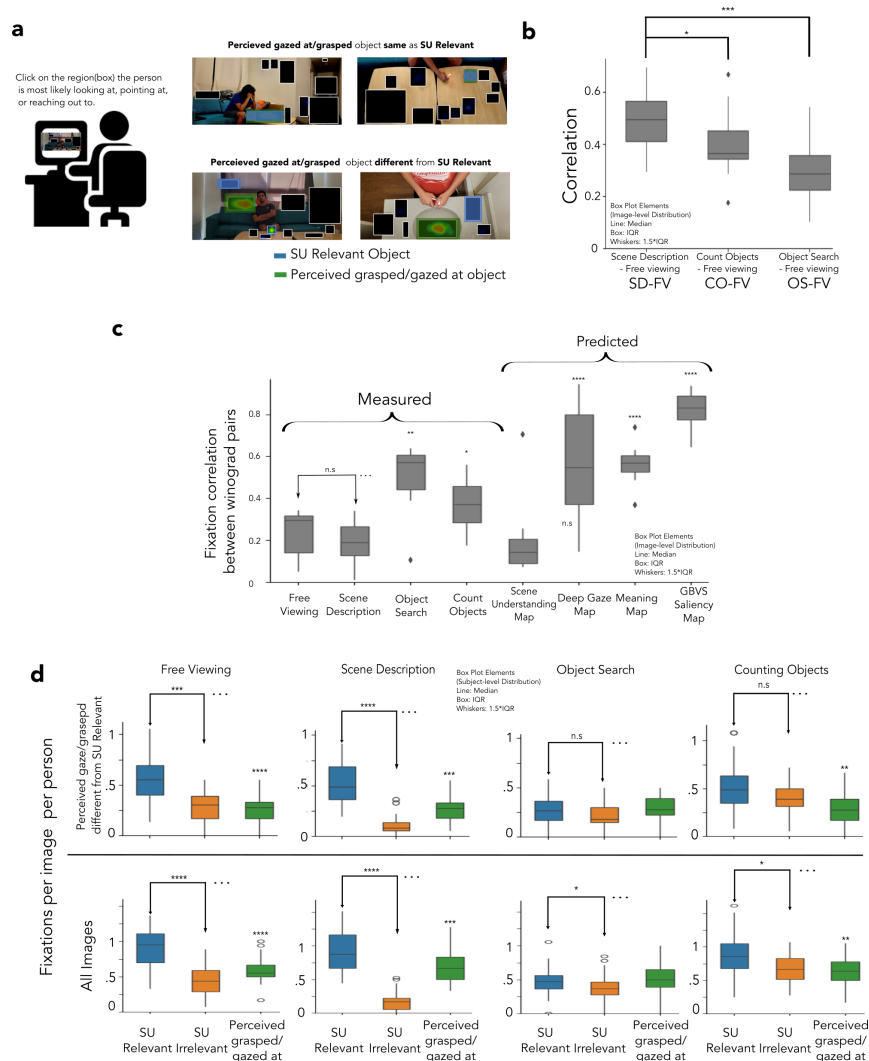


Figure 9: (a) Procedure to determine objects (left) that were perceived to be grasped/gazed at. Participants were asked to click on locations they perceived to be grasped/gazed at in each scene. Examples of images where the participants perceived the grasped/gazed-at object to be the same as the SU-relevant object (right, top 2) and images where it was different than the SU-relevant object (right, bottom 2, No gaze image subset). (b) Fixation heatmap correlations across conditions for the No gaze image subset were similar to the pattern of correlations for the entire image set (Figure 2c); (c) Fixation heatmap correlations between Winograd image pairs for the different conditions of the No gaze image set was similar to the entire image set (Figure 5b). (d) The distribution of fixation frequencies across object categories for the No Gaze image subset shows SU-relevant as the most fixated object category, significantly greater than objects perceived to be grasped/gazed at (top row) for scene description and free viewing conditions. This result holds even when all images were included (bottom row). However, the difference in fixation frequency between SU-relevant and SU-irrelevant objects was not significant for object search and counting object conditions in the No Gaze image subset.

The role of gaze, head and hand position in directing eye movements

Many of the images in our experiment present one or multiple individuals directing their gaze, head, or hands to objects that are typically relevant to the scene understanding. Thus, a possible explanation for our results is that observers' fixations on the SU-relevant locations and objects are a byproduct of the well-documented finding that observers often follow with their eyes the gaze [87, 88, 89, 63, 90] or anticipated action of others [91, 92, 93].

To show that the SU-relevant fixations are not solely a byproduct of following gaze, head, hand direction, or body posture but rather eye movements to objects critical to scene understanding, we conducted an experiment to assess which objects observers perceived to be grasped or gazed at for each scene. Figure 9a (left) details the experiment procedure. A separate group of participants ($n=25$) viewed the Winograd images with unlimited time and were asked to click the box they perceived to be grasped or gazed at by the person in the scene. Figure 9a(right) shows examples of a scene where a SU-relevant object was either perceived by the majority of participants to be grasped or gazed at (top two images), or it was not (bottom two images, No Gaze subset). We found 20 images out of the 36 images for which the SU-relevant objects were not considered the object to be grasped or gazed at by the human in the image.

Analysis of the No Gaze image subset showed a similar pattern of results to our main findings (Figure 2c, 5b & 7b). Observer fixation heat maps for free viewing were significantly more similar to the fixation heat maps from the scene description condition (FV-SD, $r = 0.5$) than to those from the object search (FV-OS, $r = 0.32$, bootstrap, $p < 0.001$ vs. FV-SD) or counting objects (FV-CO, $r = 0.41$, bootstrap, $p = 0.03$ vs. FV-SD) as shown in Figure 9b.

The observer fixation heat map correlation across Winograd images (Winograd

correlation) for the free viewing condition was no different than the scene description condition (bootstrap, $p = 0.22$) but significantly lower than that observed for the object search (bootstrap, $p = 0.006$) and counting objects (bootstrap, $p = 0.037$) conditions (Figure 9c). The Winograd correlation for the existing fixation prediction models (GBVS, meaning maps, and Deepgaze) was significantly higher than human correlations (bootstrap, $p < 0.0001$ vs. free-viewing), while the correlation for the scene understanding map (SUM) was the lowest and not different from that observed in human fixation heatmaps for the free viewing and scene description conditions (bootstrap, $p > 0.9$).

In addition, similar to what we observe in the entire data set, for the No Gaze subset, observers also execute more frequent fixations on SU-relevant objects in the free viewing and scene description conditions (bootstrap, $p < 0.0001$ for both conditions) compared to the SU-irrelevant objects. However, unlike the entire data set, the No Gaze subset does not show a significant difference between the fixation frequencies to the SU-relevant and SU-irrelevant object categories for the object search and counting objects conditions (bootstrap, $p > 0.17$ for both conditions), suggesting that for those two conditions, fixations on SU-relevant objects were driven mainly by gaze cues (where the gaze cues points to the SU-relevant objects in the images that were not part of the No Gaze Subset). Finally, fixation frequencies to other object categories and the cumulative fixation frequency analysis show similar results for the No Gaze subset as in the entire dataset (See Figure 10a & b).

To further assess whether observer eye movements are more guided by objects critical to scene understanding (SU-relevant objects) vs. objects perceived to be grasped/gazed at, we directly compared the frequency of fixations on the two categories. For the entire image dataset and the No Gaze image subset, SU-relevant objects were fixated more frequently than objects perceived to be grasped/gazed at (Figure 9d) in free-viewing (bootstrap, $p < 0.0001$), scene description (bootstrap, $p <$

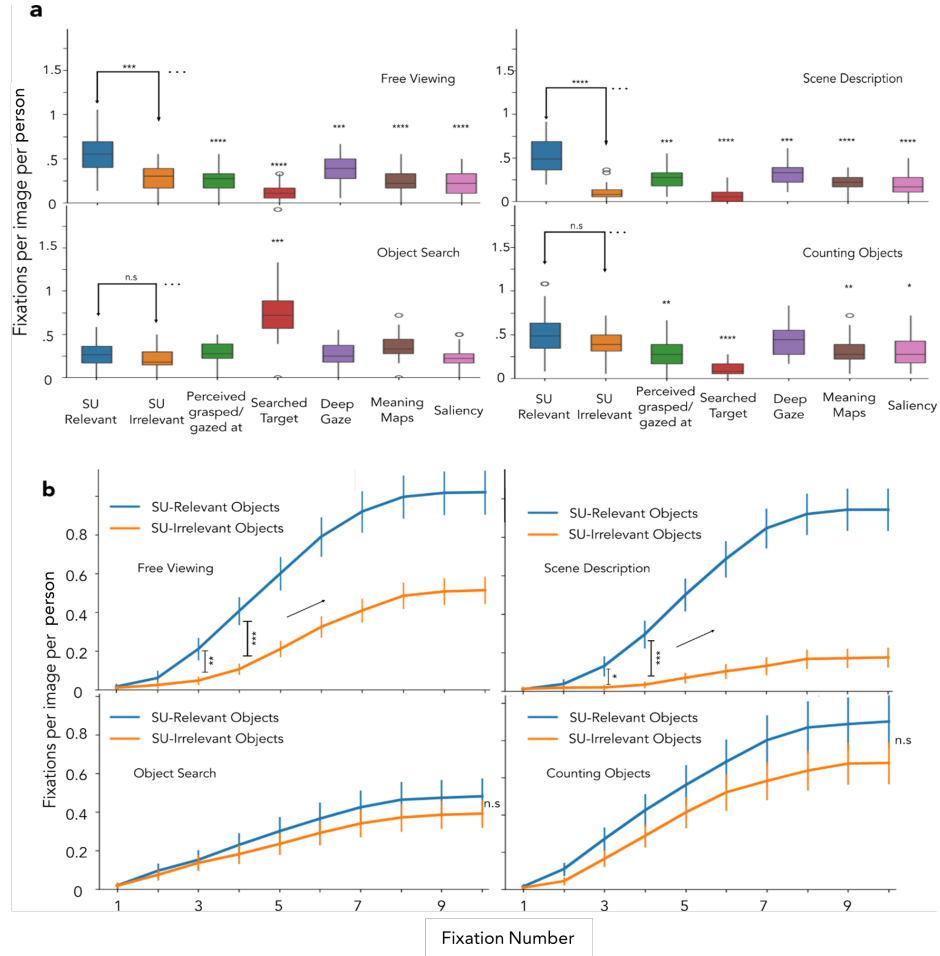


Figure 10: (a) Scenes with “No gaze” (scenes for which SU-relevant objects are different than objects judged to be grasped or gazed at) show a similar pattern of effects in free viewing and scene description conditions across all object categories. However, the difference in fixation frequency on SU-relevant and SU-irrelevant objects is no longer significant in the object search and counting objects conditions. (b) Cumulative fixation frequency shows that the difference between the fixation frequency on SU-relevant and SU-irrelevant categories is significant from the 3rd fixation onwards for scene description and free viewing conditions, and no significant difference was found for the other two conditions at least until the 10th fixation. 25 participants per image, 20 images in all analysis. (**** = $p < 0.0001$; *** = $p < 0.001$; ** = $p < 0.01$; * $p < 0.05$). A one-tailed bootstrapped analysis was conducted to test the significance of all the results. Each analysis had its significance levels corrected for the False Discovery Rate ($\alpha=0.05$, 24 comparisons in (a), 40 comparisons in (b)). In (b), the line within the box indicates the median. The box spans the interquartile range (IQR), and the whiskers extend to the most extreme values within $1.5 \times \text{IQR}$ across participants. In (a), the central measure is the mean, and the error bars show the 68% bootstrap confidence interval.

0.001) and counting objects conditions (bootstrap, $p < 0.01$), but not for the object search condition (bootstrap, $p > 0.7$).

Together, all the analyses suggest that during free viewing or while describing scenes, observers most frequently direct their eyes not just to objects inferred to be grasped or gazed at but to objects critical to understanding the scene.

Causal influence on scene understanding of fixating critical objects

More frequent fixation on SU-relevant objects during free viewing and scene description conditions might imply that processing those objects with the high-resolution fovea is functionally important to maximize the understanding of the scene. We separately analyzed the scene descriptions condition trials for which observers correctly described the scenes and those for which they incorrectly described them (using a threshold based on cosine similarity of LLM embeddings of descriptions, see the methods section for details). Across all images, the difference in fixation frequency to SU-relevant vs. SU-irrelevant objects was significantly greater for trials that resulted in correct descriptions than for trials with incorrect descriptions (bootstrap, $p = 0.0037$) (see Figure 12a for examples of descriptions near the threshold used for classifying a description to be correct/incorrect and see Figure 12b & c for the overall results and comparison of the threshold we used against human judgments and other possible thresholds). This association between eye movements and scene descriptions has been shown in previous studies. [65, 66].

However, the association does not necessarily imply a causal influence of fixation on SU-relevant objects on accurate scene understanding. An alternative explanation is that observers process the entire scene and extract its meaning before the eye movements. In this alternative explanation, observer fixations on SU-relevant objects are a

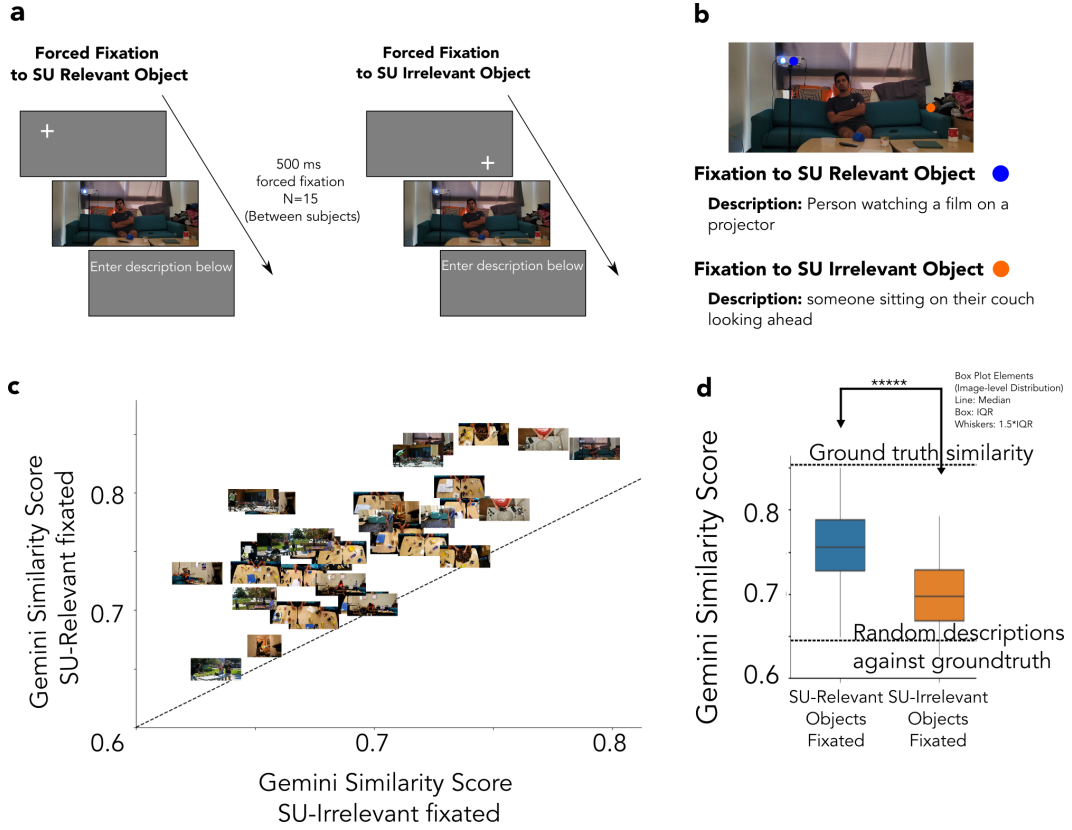


Figure 11: (a) A between-subjects (N=15) experimental design was used for the forced fixation condition. Subjects were instructed to maintain fixation at SU-relevant or SU-irrelevant objects for 500 ms. They were asked to provide a best-guess description of what was happening in the scene. (b) Examples show that observers' descriptions were more similar to the gold standard meaning of a scene when they fixated on the SU-relevant objects. (c) The scatter plot shows the similarity scores (Gemini, see Figure 3b & c for comparison of Gemini similarity scores and human ratings) of the scene descriptions relative to the gold standard for each image (averaged across observers) while fixating SU-relevant objects vs. while fixating SU-irrelevant objects. (d) Average similarity score relative to gold standard description (averaged across observers). *****($p < 0.00001$). The top dashed line indicates an upper bound established by the average similarity score across gold standards (inter-observer agreement of descriptions with unlimited time). The bottom dashed line indicates a lower bound calculated by permuting descriptions across images and comparing them to the gold standard from unmatched images.

subsequent process of looking at objects critical to the already extracted scene understanding. In this interpretation, fixations on the SU-relevant objects are not required to understand the scene accurately.

We conducted a separate experiment to evaluate the causal influence of fixations

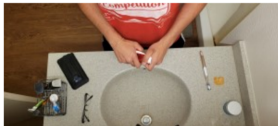
a



Description just above -1 SD from mean similarity
handling a wet napkin over a cluttered table, **Z-Score:** -1.46616034
Description just below -1 SD from mean similarity
man is cleaning up spilled wine, **Z-Score:** -0.93763191



Description just above -1 SD from mean similarity
a man is trying to put out something by using his hands., **Z-Score:** -1.10956765
Description just below -1 SD from mean similarity
someone is taping together ripped homework, **Z-Score:** -0.9953873



Description just above -1 SD from mean similarity
the person is capping their toothpaste, **Z-Score:** -1.33743852
Description just below -1 SD from mean similarity
brushing his teeth, **Z-Score:** -0.97218592



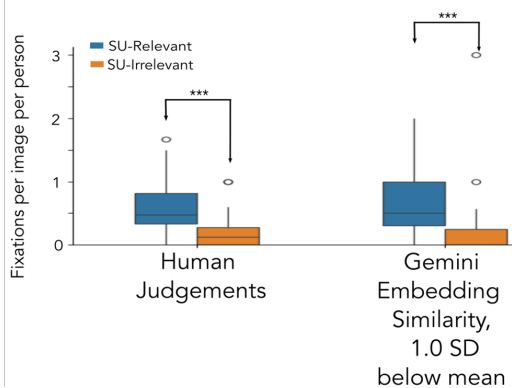
Description just above -1 SD from mean similarity
male college student folding shirt doing laundry, **Z-Score:** -1.34046465
Description just below -1 SD from mean similarity
man ironing a flannel shirt, **Z-Score:** -0.63857332

b

Scene Description - Fixations of participants with incorrect description



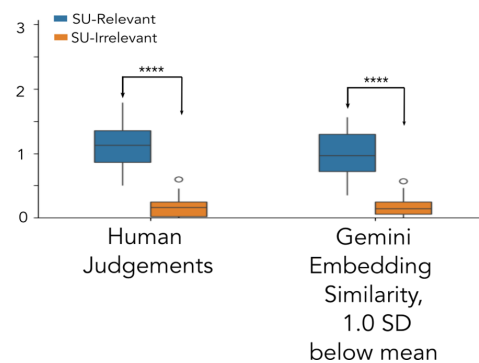
A guy seems sad looking at something



Scene Description - Fixations of participants with correct description



A man sitting on a couch watching something being shown by a projector



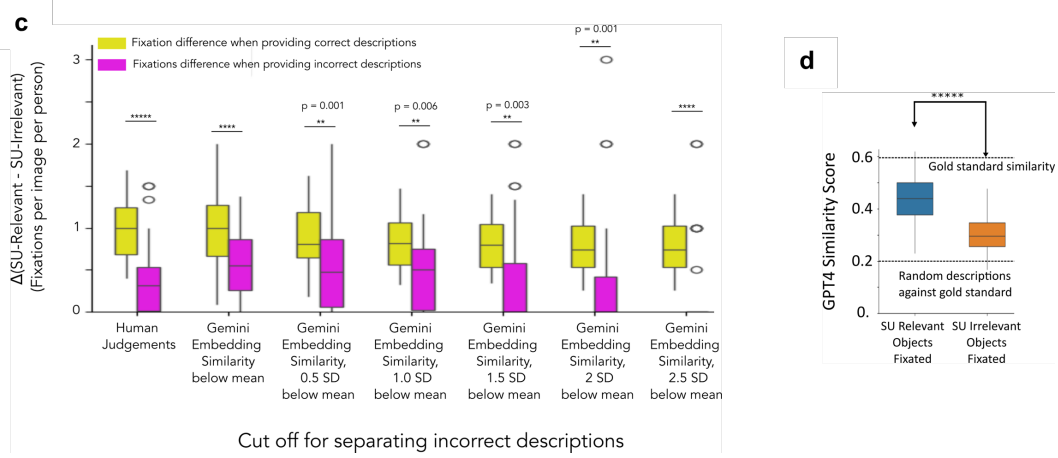


Figure 12: (a) Examples of descriptions with z-scores below and above -1 SD (threshold) from the mean description similarity rating provided by participants during the scene description condition for each image. Descriptions for all images near the decision threshold of -1 SD are available in Mendeley. (b) (Top) Example Winograd image pair with the fixations divided across groups of observers that correctly and incorrectly described images (see methods for the procedure to classify a description as correct or incorrect). (Bottom) Distribution of fixations on object categories in the scene description condition. Participants tend to fixate on the SU-relevant objects more frequently when accurately describing the scene. This result holds irrespective of whether one uses human judgments to determine whether the participants’ image descriptions were correct/incorrect (relative to the gold standard) or uses the LLM embedding similarity measure (1.0 standard deviation, SD from the mean similarity rating). (c) The difference in fixations on SU-relevant and SU-irrelevant object categories, split by participants who gave a correct versus an incorrect description, is shown as a function of the decision threshold for separating the incorrect and correct description sets. Images with correct descriptions, determined using various categorization criteria (varying SD from the mean cosine similarity of Gemini embeddings), always led to higher fixation frequency on SU-relevant vs. SU-irrelevant objects than images with incorrect descriptions. As SD increases, the threshold for what is considered an incorrect description increases, resulting in fewer data samples for the incorrect description category (pink bars). The fixation difference across SU-relevant and SU-irrelevant objects was significantly larger for correct vs. incorrect descriptions. This result generalized across all SD decision thresholds (for the LLM embedding similarity measure) and for the human correct/incorrect classification judgments. (d) The average similarity of descriptions from participants fixating at the SU-relevant and SU-irrelevant locations (15 participants per fixation location, 36 images) compared to the ground truth image descriptions, quantified using GPT4 embeddings to compute cosine similarity, shows similar results as Gemini embeddings, related to Figure 7. For all graphs: ***** = $p < 0.00001$; *** = $p < 0.001$; ** = $p < 0.01$. A one-tailed bootstrapped analysis was conducted to test the significance of all the results. In all figures, the line within the box indicates the median. The box spans the interquartile range (IQR), and the whiskers extend to the most extreme values within $1.5 \times \text{IQR}$ across participants for (b) and (c) and across images for (d).

on SU-relevant objects on scene understanding. Two new groups of 15 observers maintained fixation either at the SU-relevant or the SU-irrelevant objects during a 500 ms presentation of the same test images used in experiment 1. The observers then described what was happening in the scene. The collected descriptions were compared to the gold standard description (descriptions from observers with unlimited time and eye movements) using sentence similarity measures with the embeddings of Large Language Models (LLMs) (see methods for validation of Gemini similarity scores with human ratings). Figure 11a & b shows the procedural workflow of the experiment and an example description from observers fixating on the SU-relevant or SU-irrelevant object for a sample image.

Figure 11c & d shows that observers' descriptions were more similar to the gold standard description of the scene when they fixated on SU-relevant objects compared to when they fixated on SU-irrelevant objects (bootstrap, $p < 0.00001$). For reference, Figure 11d shows an upper bound of similarity of descriptions provided by the average score across different gold standard descriptions (different observers describing the scene while exploring the image with unlimited time). A lower bound on similarity scores was estimated by permuting scene descriptions across images and comparing their similarity to the gold standard from other unmatched images (see methods. Also, see Figure 12d for similar results obtained using embeddings from the GPT4 LLM model).

Together, the results suggest that fixating on SU-relevant objects during free viewing causally leads to more accurate descriptions of the scenes than when fixating on SU-irrelevant objects.

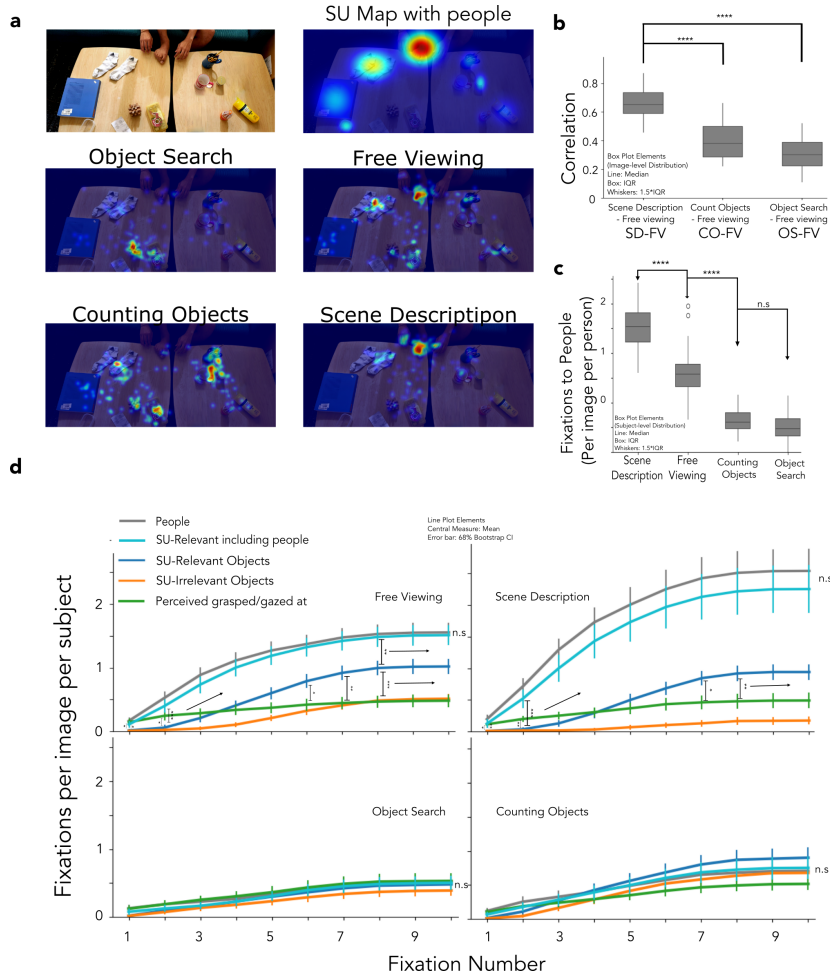


Figure 13: (a) Example scene understanding map (top right) when people in the scenes were included in the SUM generation procedure. The SUM predicts people as the most important (therefore most fixated) in this scene, followed by the object critical to understanding the scene (SU-Relevant). Similar fixation patterns were observed in both free-viewing and scene description conditions, which differed from those in object search or counting objects conditions. (b) After including people in the scenes, fixation analysis, observer fixation heat maps for the free viewing (FV) condition correlated the highest with the scene description (SD) condition. (c) The frequency of fixations on people in the scenes was highest in the scene description condition, followed by the free viewing condition. (d) The cumulative fixation distribution on various scene components for the No Gaze image subset shows that observers fixated on people in the scenes significantly more frequently than on other categories in free viewing and scene description conditions, starting with the first fixation. The SUM also predicts a higher frequency of fixations on people. Initially (in the first two fixations), observers fixated on the to-be-grasped/gazed at objects significantly more than the SU-relevant object; however, this trend eventually inverts, with cumulative fixations on SU-relevant objects significantly higher than those on the to-be-grasped or gazed at objects from the 6th (free-viewing) or 7th fixation (scene description).

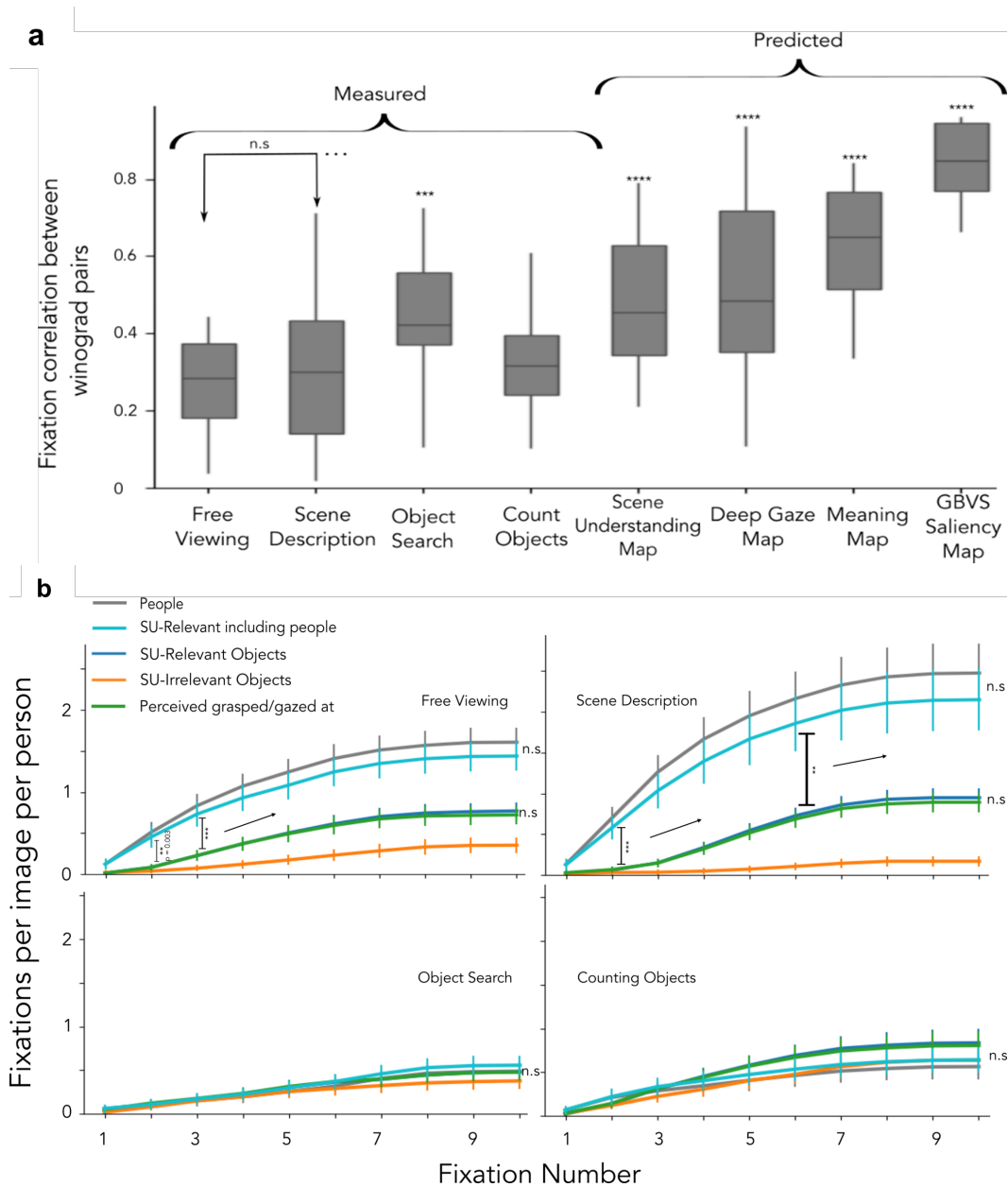
Fixations to people and their role in the understanding of scenes

All presented analyses involved fixations on objects in the scene and excluded the fixations on people (hands, faces, bodies) to isolate the influence on fixations of SU-relevant objects, which varied across Winograd images (while the people in the scene were present in both). However, the larger literature documenting the importance of fixations on people’s heads [37, 68, 94] and hands [95] motivated us to understand how such fixations fit within the theoretical framework of eye movements to maximize scene understanding.

We first re-estimated the SUMs by collecting descriptions of the scenes that included the deletion of people or people’s hands in the scenes (in addition to the deletion of objects). Figure 13a presents a sample SUM showing that the person’s hands in the scenes were the most critical (altered the scene description the most). We found that the people were the most critical in 77.78% (28 images) of the images in the dataset, while the SU-relevant object was in the remaining.

We then repeated all analyses from previous sections, but included the fixations on people in the scenes. As with the initial analyses excluding fixations on people, we found higher correlations of observer fixation heat maps of the free viewing with those of the scene description condition (FV-SD, $r = 0.66$) than with the fixation heat maps of counting objects and object search conditions (Figure 13b; bootstrap, $p < 0.0001$). The presence of fixations on people increased the correlations across fixation heat maps of the Winograd image pairs for all conditions and images. Although the results remain the same as the analysis excluding fixations on people (see Figure 14a), the fixation heat map correlation across the Winograd image pairs for the counting objects condition was no longer significantly different from that for the free viewing condition (bootstrap, $p = 0.26$).

We also re-analyzed the eye movement data for the four conditions to assess the



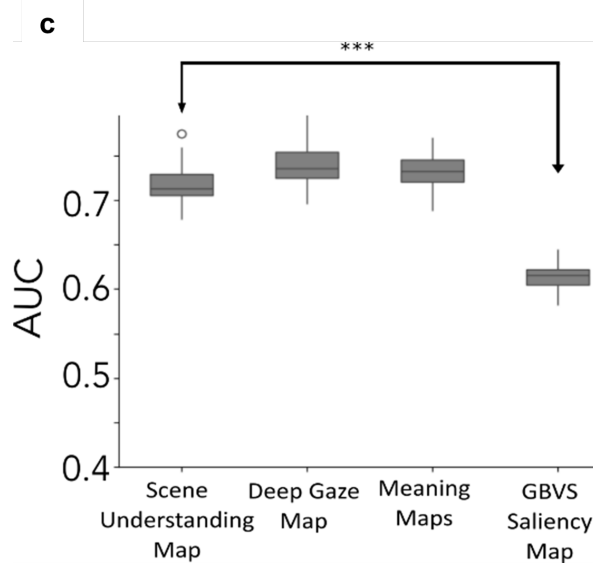


Figure 14: (a) Fixation heatmap correlation across Winograd pairs with human fixations/model fixation predictions to people included in the analysis (images=36, participants=25). The Winograd correlation for free viewing and scene description increases due to the many fixations on people in those two conditions (people are at similar locations in each Winograd pair). With the inclusion of fixations on people, the heatmap correlation across Winograd images for counting objects was not significantly different from that of free viewing. This also increases the SUM correlation across Winograd pairs, as the SUM predicts people to be the most important in 77% of the scenes. (b) Patterns similar to the main results (Figure 8d) are observed for the cumulative frequency of fixations on object categories for the image subset (images=20, participants=25), where the object perceived to be grasped/gazed at is the same as the SU-relevant object. The slight difference in SU-relevant and perceived grasped/gazed at is related to some images having more than one SU-relevant object. (c) A shuffled AUROC analysis for the entire image data set revealed that the SUM maps have similar performance in predicting human fixations during free viewing (AUC = 0.72) to meaning maps (AUC = 0.73, bootstrap, $p = 0.22$) or DeepGaze (AUC = 0.74, bootstrap, $p = 0.09$), and are significantly different from GBVS maps (AUC = 0.61, bootstrap, $p < 0.001$), (images=36, participants=25). (**** = $p < 0.0001$; *** = $p < 0.001$; ** = $p < 0.01$). A one-tailed bootstrapped analysis was conducted to test the significance of all the results. Analyses for comparisons in (b) had their significance levels corrected for the False Discovery Rate ($\alpha=0.05$, 30 comparisons in (a)). In figures (a) and (c), the line within the box indicates the median. The box spans the interquartile range (IQR), and the whiskers extend to the most extreme values within $1.5 \times \text{IQR}$ across images in (a) and across participants in (c). In (b), the central measure is the mean, and the error bars show the 68% bootstrap confidence interval.

fixation frequency on people in the scenes (their hands, bodies, or faces; see methods). Observers fixated significantly more frequently on people for the scene description condition (Figure 13c; bootstrap, $p < 0.0001$). Fixations on people in free viewing were significantly more frequent than those observed in the object counting and search conditions (bootstrap, $p < 0.0001$).

We calculated the cumulative fixation frequencies to understand the temporal dynamics of fixation selection on different scene components (people, SU-relevant objects, perceived to be grasped or gazed at objects, and SU-irrelevant objects). We used the No-Gaze image subset for which the perceived grasped/gazed-at objects were distinct from the SU-relevant objects (see Figure 14b for similar results for the set of images in our dataset where the grasped/gazed at objects were the same as the SU-relevant objects). Figure 13d shows that starting with the 1st fixation, observers most frequently fixated on people (or the highest value of the scene understanding map estimated with the inclusion of people, see SU-relevant including people in Figure 13d) for free viewing and scene description. For these conditions, early fixations (1st and 2nd) were also more frequently directed to objects perceived to be grasped or gazed at significantly above the SU-relevant object category (bootstrap, $p < 0.02$). Subsequently, the cumulative fixations on SU-relevant objects were significantly higher than the perceived grasped/gazed-at from the 6th fixation for free-viewing (bootstrap, $p < 0.03$) and the 7th fixation for scene description (bootstrap, $p < 0.04$) condition. We observed no statistically significant differences in cumulative fixations across scene components for the object counting and search conditions.

Similar to the AUROC analysis without people, AUROC analysis including fixations on people also resulted in a significant difference between SUM and GBVS saliency maps (bootstrap, $p < 0.001$) and no statistically significant difference for DeepGaze, meaning maps, and SUMs, though SUMs were lower than meaning maps and DeepGaze

(bootstrap, $p > 0.09$) (see Figure 14c).

1.5 Discussion

Eye movements during free viewing are directed to objects critical to understanding scenes

The purpose of our study was to gain an understanding of the image properties that guide eye movements during free viewing of scenes. Although there has been a long-held view that low-level saliency drives eye movements during free viewing [28, 29, 32, 33, 34], recent studies have supported the concept that eye movements are directed to regions/patches of a scene judged to be meaningful [64, 96, 43]. Here, we explored the hypothesis that humans aim to maximize scene understanding accuracy during free viewing and move their eyes to elements (people and objects) critical to scene understanding. The most locally meaningful region, object, or person (meaning map maximum) is not necessarily the most critical person/object for understanding a scene. The latter involves relationships between people, objects, and regions that contribute to the overall understanding of the scene, which is not present in meaning maps.

We first investigated human eye movements to objects for different instruction conditions with newly designed stimuli (Winograd images). The Winograd image pairs varied the objects critical to scene understanding while maintaining visual saliency and the meaningfulness of local regions. Three findings support our hypothesis. First, fixation distributions during free viewing were most similar to those of observers when describing the scenes. Their correlation was comparable to the inter-observer fixation distribution agreement for the free viewing condition and similar to previously reported correlations between the best models and human fixations (e.g., GBVS-human fixation correlation = 0.45, [97]). Second, small visual alterations to the images that changed

the meaning of the entire scene (Winograd images [77]) influenced human fixations during free viewing. Third, fixations were more frequently directed to the objects critical to understanding the scene rather than the most salient or locally meaningful region.

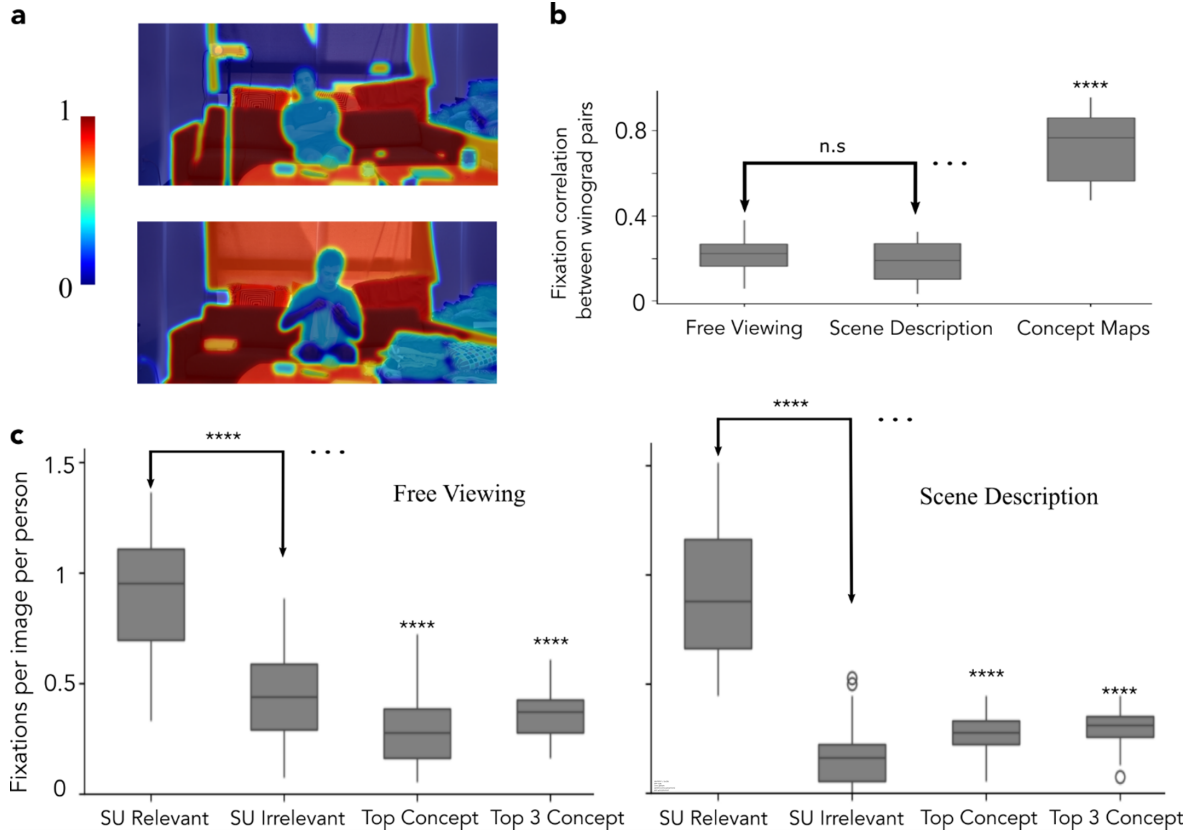


Figure 15: (a) Example concept maps[2] generated for a Winograd image pair. (b) The correlation across Winograd pairs for concept maps is significantly higher than that observed in the Winograd correlations of free viewing or scene description fixational heatmaps. This suggests that the concept maps do not vary greatly across Winograd image pairs. (c) The frequency of fixations on SU-relevant objects is significantly higher than the top or top 3 object predictions in the concept maps. (**** = $p < 0.0001$). A one-tailed bootstrapped analysis was conducted to test the significance of all the results. In figures (b) and (c), the line within the box indicates the median. The box spans the interquartile range (IQR), and the whiskers extend to the most extreme values within $1.5 \times \text{IQR}$ across images in (b) and across participants in (c).

Previous studies have shown how scene gist, object spatial relations, and object function relations guide eye movements when observers are searching for objects [98, 99, 100, 101, 102, 103, 104, 60, 105, 106, 107]. However, this had not been demonstrated for free viewing due to the difficulty of determining what components of the semantic

information present in the scene were relevant to free viewing eye movement guidance and how to quantify these semantic components. Here, we developed a method to quantify the semantic role of an object in free viewing by measuring how the object’s removal alters scene understanding (scene description). This is distinct from recent efforts to quantify the semantic relationship among objects in scenes (concept maps, [108]), which do not define their relationship to the scene’s understanding. Such concept maps do not predict the object observers fixate the most nor the influence of the Winograd images on human fixations in our study (see Figure 15a, b & c).

Eye movements to objects critical to understanding scenes vs. objects perceived to be grasped/gazed at

Some of our images contained a person directing their gaze or head toward the object critical to the scene understanding or the presence of hands, which could suggest an interpretation that observer eye movements are simply showing the well-documented behavior of following the gaze of other people in the scenes or fixating on objects perceived to-be-grasped/gazed-at [87, 88, 89, 91, 92, 93, 63, 90]

However, we analyzed images with no grasp/gaze cues and directly compared fixations on objects judged to be grasped/gazed vs. objects critical to scene understanding. Our findings show that objects critical to scene understanding are frequently looked at even when the gaze is judged to be directed elsewhere or the to-be-grasped/gazed-at object is other than SU-relevant objects. However, gaze, head, and body cues are important predictors of the spatial location of these critical objects, and observers initially fixate on perceived-to-be-grasped or gazed at objects, consistent with gaze cueing studies [109, 110]. However, even without the grasp/gaze cues, the observer’s eye movements eventually fixate on the SU-relevant objects. We argue that objects critical to scene understanding are a general concept guiding eye movement fixations

during free viewing and that the gaze of others, most often, but not always, serves as a predictive cue of the location of critical objects.

Functional role of fixations to maximize scene understanding

We found that trials in which a person did not fixate on the critical object were associated with less accurate scene descriptions. A follow-up experiment that manipulated the point of fixation in the scene evaluated the causal influence of fixation on extracting the information required to understand a scene. We showed that fixating on the object critical to scene understanding resulted in more accurate scene descriptions than fixating on a different object irrelevant to the scene understanding. This suggests that observers' fixations on objects during free viewing have a functional purpose to maximize scene understanding. This is consistent with many recent findings demonstrating the functional perceptual role of saccades [23, 19, 21, 111, 90], micro-saccades [112], and smooth pursuit eye movements [113, 16].

Our functional evaluation only evaluated a fixation on the object critical to the scene understanding, and another fixation on another object at a comparable distance from the center of the image. We did not evaluate an array of fixations that covered the entire image. It might not always be necessary to fixate on the center of an object. Fixating close to the critical object (within 2 degrees) may also result in maximal accuracy in describing the scenes. And for some scenes, it may be that center-of-mass fixations between the critical object and the hands might also provide accurate scene description [23, 114, 115].

The findings should not be interpreted to suggest that observers do not use peripheral information to guide their eye movements. For example, observers might identify a tool in the visual periphery but might not be able to specify the tool type. The peripheral information can guide the eye movement to the tool, and only after fixating

on the object can the observer identify it as a screwdriver and more accurately describe the scene.

The results do not suggest that a single fixation is sufficient to understand a scene. As with other tasks with scenes [116, 117, 118, 119, 120], many of our scenes might require multiple fixations to extract their meaning accurately. The planning of multiple fixations might also consider the motor cost of saccades and result in fixating irrelevant objects placed between two objects critical to scene understanding [121, 122].

The role of fixations on people and social cues in the theoretical framework of scene understanding

Humans looking at other people in scenes has been a universal finding across the eye movement literature [2, 37, 68, 95, 94]. Our results follow the classic finding. During free viewing and scene description, people in the scenes were fixated the most, well above any objects, including those critical to scene understanding. Such findings are consistent with the theoretical framework of eye movements to maximize scene understanding. Erasing people from the scenes had the largest impact on the scene description when we re-created the scene understanding maps (SUMs) to assess the influence of people.

Comparing scene understanding maps to saliency models, DeepGaze, and meaning maps

A saliency model based on low-level image features [6] failed to predict the high frequency of human fixations on objects critical to understanding the scene. This is because low-level saliency models do not incorporate any aspects of the object relationships or tasks. Meaning maps [64] also fall short for our images because they consider the local meaningfulness of objects and not their contributions to understand-

ing the entire scene. DeepGaze [49] is trained on human fixations and image features and thus, in principle, should be able to learn to predict fixations on the objects critical to scene understanding. DeepGaze’s lower prediction of frequent human fixations for our images may reflect the data used to train DeepGaze. The training data set might not include sufficient samples of images with complex behaviors as depicted in our Winograd images. Additionally, DeepGaze’s emphasis on fixation density maps may make it less capable of accurately predicting the most frequently fixated object. In addition, SUM’s advantage over the other models is not present when considering all human fixation density maps. SUMs were marginally better (though not significantly different) than DeepGaze and meaning maps when considering only fixations to objects, and were marginally lower (though not significantly different) than the two models when including fixations to people.

Importantly, even if DeepGaze predicted each scene’s most fixated object, it would not provide a unified theoretical understanding of what drives certain objects to be fixated the most for each image. Our experimental investigation provides such a theoretical account. It explains free viewing and frequent fixations on people and objects in terms of their contributions to understanding the scene.

Why do people seek to understand scenes during free viewing?

Our findings suggest that when observers view scenes without a specific task instructed by the investigator (free viewing), seeking visual information to understand the scene is one of the default tasks they engage in. Understanding a scene is essential for contextualizing the current visual input, making inferences about likely past events that have led to the current visual state, and, importantly, predicting future events [123, 124, 125, 126]. Making inferences requires an observer to consider all likely possibilities for that scene. In this framework, eye movements during free viewing of scenes are directed to locations that reduce the uncertainty of the possible states of the

visual world [127, 128], consistent with concepts of information sampling and curiosity [129].

Our findings also show a smaller but significant effect of fixating on objects critical to the scene’s understanding, even for the object-counting and search conditions. Our analysis reveals that these effects were primarily driven by grasp/gaze cues.

Consistent with our findings suggesting that understanding a scene is one of the default observer tasks during free viewing, a recent fMRI study showed that the semantic information of a scene (represented by the embeddings of a deep neural network visual-language model) predicted brain activity of people passively (no task) viewing scenes better than traditional object labels [130].

Generalization to other image types and influence of individual and cultural differences

All scenes in the Winograd pairs feature one or more individuals or a person’s hand, implying an action or social interaction. These types of images are fundamental to human daily lives. However, understanding these scenes can often change due to the presence or absence of a single object, thus requiring observers to explore the image with their eye movements. How would the analyses apply to images that contain objects spread across an image (e.g., a prototypical kitchen or forest image) with no implied future or past actions or social interactions? For such simpler images, there might not be a single object that is critical to their understanding. Furthermore, observers might be able to rapidly understand the scene with peripheral visual processing and without eye movement exploration [131, 132, 133]. Thus, we might expect fixations with those images to have a lower perceptual function and the most frequently fixated object to be less related to its relevance to scene understanding. Additionally, unlike

the Winograd image dataset, many real-world scenes feature multiple people, some of whom are critical to understanding the scene, while others are not. It is likely that the SUMs, but not the meaning maps, best predict which person or people in the scenes are fixated the most, but this will need to be tested.

In addition, a limitation of our approach is that it might not capture individual [134] and cross-cultural differences [135] in scene understanding that might influence the most fixated object in scenes. Individual and culture-specific SUMs may be a possible direction for predicting such inter-individual and inter-cultural differences in fixation behaviors.

To conclude, our findings suggest that during free viewing of scenes, eye movements are directed to people and objects that are critical to understanding the scene, rather than to regions that are visually salient or judged to be meaningful. The eye movement fixations serve a perceptual function, as they causally improve the accuracy of scene descriptions. The theory of eye movements to maximize scene understanding and its empirical implementation through scene understanding maps provide a unified account of free viewing of scenes. They predict which objects are fixated on frequently, explain the higher frequency of fixations on people over objects, and frequent fixations on objects that are perceived to be grasped or gazed at. Together, our findings suggest that an important default task for the human brain during free viewing is comprehending the visual world.

Chapter 2

Why We Look Where We Look: Emergent Human-like Fixations of a Foveated Visual Language Model Maximizing Scene Understanding

2.1 Abstract

When humans view scenes without a specific task (free viewing), they initially direct their eye movements toward the center of the scene and then fixate on people, text, objects being gazed at or grasped, and other semantically meaningful regions. What these signature fixation patterns reflect and whether they optimize an underlying perceptual task remain unknown. We show that a computational agent with simulated foveation, trained with reinforcement learning to optimize scene comprehension, exhibits all the emergent human fixation signature patterns. This alignment is observed both in our laboratory dataset of fifty participants and in a large-scale publicly available dataset comprising 6720 participants. In contrast, versions of the agent trained on other perceptual tasks, such as visual search or scene classification, or equipped with peripheral vision that was better or worse than human vision, predicted human fixation patterns less accurately. Another variation of the agent trained to minimize uncertainty across candidate scene descriptions, without ground-truth scene descriptions, also produces human-like fixations. These findings suggest that human free-viewing fixation patterns

may emerge as a functional byproduct of optimizing scene comprehension under the biological constraints of foveated vision.

2.2 Introduction

Humans explore the visual world through eye movements, directing central vision (the fovea) to regions of interest to process them with greater spatial detail. Decades of research show that these rapid eye movements are guided to locations and visual attributes that help us attain maximum accuracy when searching for targets [12, 23, 14, 25], identifying faces [21], or executing actions such as grasping and walking [119, 16, 18, 136]. But where humans look when free-viewing a scene without a particular task, such as when leisurely sitting at an outdoor cafe, glancing from a bus window, or entering a room, is not well understood. The notion that during free-viewing humans look at low-level salient regions defined by contrast, color, and orientation has been repeatedly challenged over the last 20 years [41, 72, 36, 27, 137]. There is converging evidence for a set of landmark signatures of human eye movement behavior during free-viewing. Humans look at other people [37, 38], text [39, 40], and objects [41, 42]. In particular, they look at objects judged meaningful, objects that are gazed at, and objects that are grasped [138, 89, 27]. And when they first look at an image, they tend to look at the center [44, 45], an effect known as the central bias. But how might the characteristic and pervasive human fixation behaviors arise? Could these emerge as byproducts of optimizing the visual system for a single task or a small number of specific tasks? We will call such a theory the task-optimization of free-viewing eye movements. It asserts that where humans look emerges from learning to efficiently explore a scene to attain the highest possible accuracy in performing some type of visuo-cognitive task. The frequent free-viewing fixation patterns to people, text, and objects are what we might expect from an organism whose visual system has varying spatial processing fidelity (foveation) and optimizes some underlying perceptual task.

If so, a computational agent equipped with human-like foveated vision and sequential eye movements optimized for a visual task should naturally produce fixation patterns that mirror those of humans.

There are various challenges in implementing such a computational agent. The first challenge is determining what task humans might be engaged in during free-viewing. One view is that different individuals are engaged in different subtasks based on their individual-specific interests or idiosyncratic tasks. A car lover might look at a '60s classic Citroen 12V, a bird-watcher on a purple martin, and a music lover on a vintage poster of Sonic Youth on the wall. Thus, we may be unable to define a common task that governs how most people view scenes naturally. Another view suggests that a default task most humans engage in involves comprehending scenes and events [123, 65, 66, 139]. Furthermore, a recent study showed that the locations humans look in scenes when free-viewing are similar to those where they look when explicitly instructed to describe the scene. [137].

Thus, a good starting point is to assess which fixation patterns are optimal for a foveated agent maximizing scene comprehension, and whether those patterns resemble human free-viewing eye movements. This leads to the second challenge in creating such a computational agent: a model that can accurately comprehend scenes and capture how different fixations (and foveations) translate into scene comprehension accuracy. Recent advances in compact visual language models (VLMs), which can describe scenes at human-level and run on local GPUs, provide a viable path to implement the model [140, 141, 142, 143, 144]. Implementing existing fixation-dependent visual loss techniques [145] on the image can degrade the model's scene comprehension accuracy with varying fixation points in the scene (Fig. 16a).

The third challenge is how to optimize the agent's eye movement exploration to maximize the understanding of scenes. This requires a metric to measure the accuracy



Model Description (No Foveation): The image captures a large stadium with a prominent sign announcing a performance by Bad Bunny tonight, set against a backdrop of a wet road and trees, with the sun shining through the clouds.'

Model Descriptions (after Foveation)



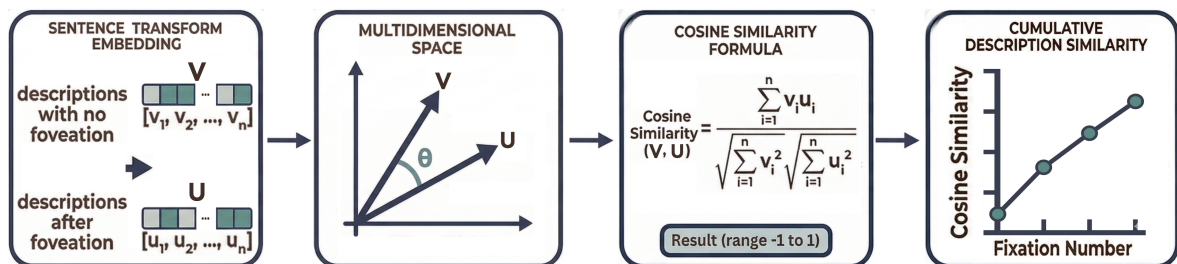
The scene captures a tranquil urban setting with a modern building featuring large windows and a tree providing shade on a sunny day.

A large, multi-story building with a modern architectural design stands prominently, surrounded by trees and a pathway, under a clear sky.'

A large, multi-story building with a modern architectural design stands prominently, surrounded by trees and a wet pavement, suggesting recent rain.

A concert venue is preparing for an upcoming performance by Bad Bunny, with the venue's exterior prominently displaying the announcement of the event.

Description Comparisons



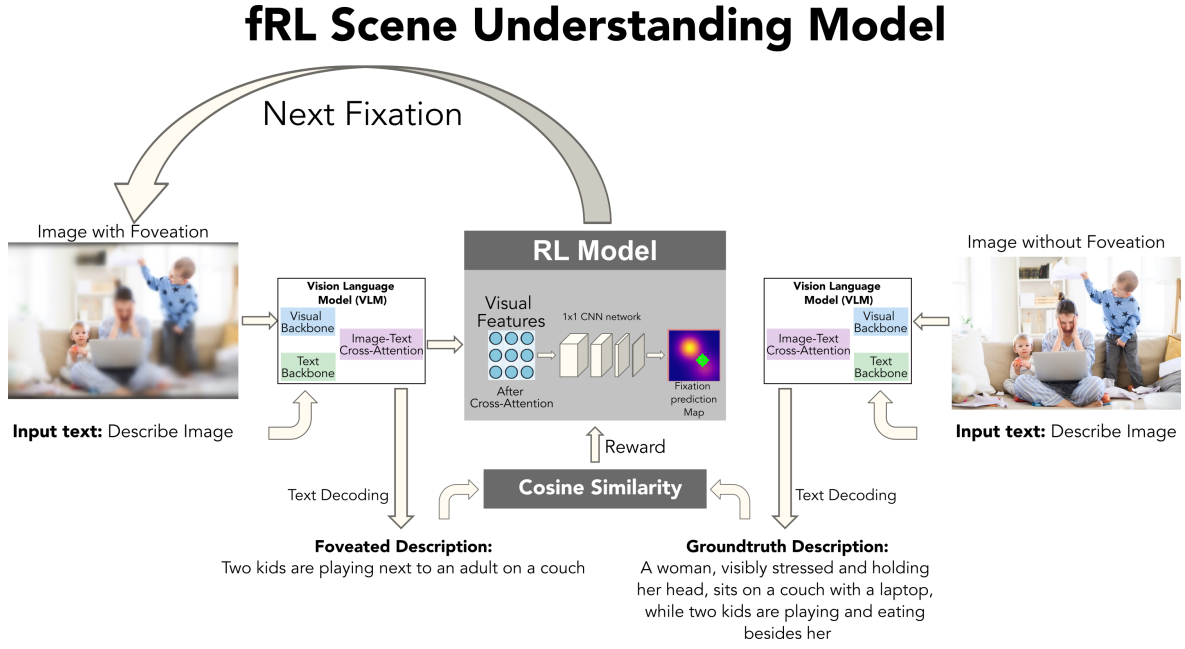


Figure 16: Semantic accuracy metric and foveated RL-scene understanding (fRL-SU) model training pipeline. (a) Schematic of a vision language model (VLM) description with and without foveation. As the number of fixations increases, the foveated description agrees more with the unfoveated description, matching it by the fourth fixation. Semantic accuracy is quantified as the cosine similarity between text-embedding vectors of the foveated and unfoveated descriptions. (b) Flowchart of the fRL-SU model’s training process. After each fixation, the semantic accuracy is computed and used as a reward for the RL model. The RL model takes hidden layer features corresponding to VLM visual tokens as input, processes them using a CNN, and outputs a probability map of possible fixation locations. The RL model then selects the next fixation, and the process repeats until the model has made 4 fixations following the initial fixation.

of scene descriptions. We measure semantic accuracy by representing both the agent’s scene description (after N fixations) and the ground-truth description (no foveation, full resolution) as text vectors (embeddings), then taking their cosine similarity (Fig. 16a). This metric correlates well with human judgments [137]. To maximize semantic accuracy, the model should use its current foveated view to select the next fixation that most improves scene description. Traditional approaches use a Bayesian ideal observer constrained by a foveated visual system and provide closed-form mathematical expressions to estimate accuracy-maximizing fixation locations or eye movement actions [23, 115, 146, 147]. These methods cannot be applied to real-world scenes without making a large number of assumptions because the probability density functions of the stimuli are unknown. Reinforcement learning techniques allow for optimization of actions by rewarding the association of actions and states [148, 149, 25]. Neural networks provide a method to learn a mapping between multi-dimensional states describing features of a scene and actions [150], and have been shown to approximate optimality of eye movements for simpler search tasks where the Bayesian ideal is computable [151].

Here, we combine a compact visual language model [140, 141], simulated foveation [145], and reinforcement learning (RL) to develop a model that learns eye movements to maximize scene understanding (the foveated RL for scene understanding, or fRL-SU, model). We hypothesize that the signature fixation patterns humans exhibit when viewing real-world scenes emerge from a near-optimal eye movement strategy for scene comprehension. We note that previous studies have developed deep neural network models that can be trained to successfully predict human fixation patterns on scenes (e.g., DeepGaze [46], ScanDiff [47], PathGan [48], etc.). An important distinction is that these models are trained on large datasets of human fixations. They do not execute any perceptual task, do not incorporate the foveated processing of the human visual system, and do not estimate task-optimal eye movements. They therefore do not address how these fixation patterns could arise in an organism whose foveated visual

system is optimized for ecologically relevant tasks. The fRL-SU model is not trained on human fixations. It is trained to generate fixations that optimize scene understanding under constraints of foveated vision.

We compare the fixations of the fRL-SU model to those of 50 human observers free-viewing 147 images. Our main metric is the frequency of fixations to different categories: people, text, objects relevant to scene understanding, and objects gazed at or grasped by people in the scene. We reason that if humans are optimizing scene comprehension during free-viewing, then training the fRL model to maximize accuracy on other tasks, such as identifying the smallest object in the scene or categorizing the scene [152], should predict fixations that do *not* match free-viewing human patterns as well. Furthermore, if the fRL-SU model fails to predict human fixations under either minimal or extreme foveation, this would indicate that human fixation patterns are a consequence of an interaction between the human foveated visual system and scene-comprehension optimization. As additional control models, we compare the fRL-SU model to saliency models (GBVS [6] and Itti [5]) and also DeepGaze, which serves as an upper-bound prediction model of human fixations, given that it is trained on human eye movement data. Finally, to assess the generalization of our model across observers and images, we evaluate the fRL-SU model on a recently published dataset that quantifies how fixations to text, people, and touched objects vary with age for 6,720 individuals [50].

2.3 Methods

Dataset

Training Set

To train a model to make eye movements while understanding scenes, we required a dataset of images depicting complex social activities, as such scenes would require the model to explore different regions of the image to understand what is happening. To this end, we combined images from three datasets: the MPII Human Pose dataset [153], developed by the Max Planck Institute for Informatics (MPII), which comprises 25,000 images featuring over 40,000 people performing 410 different activities; the FindingEmo dataset [154], consisting of 25,000 images from online sources that depict complex emotions among people in naturalistic settings; and the MSCOCO dataset [155], consisting of 330,000 images commonly used for image captioning. A subset of 8,000 images was selected from each dataset (see below for details on image selection), resulting in a final training set of 24,000 images.

Testing Set

To test our model, we curated a dataset comprising 147 images: 24 from the MS-COCO validation set, 47 screenshots of movie scenes depicting complex social interactions, and 76 images that were a mix of in-house photographs and AI-generated images created with tools such as Google Nano Banana Pro. The images were chosen so that the semantic accuracy of the VLM model’s descriptions relative to the human ground-truth description exceeded 0.5. We used embedding models and computed cosine similarity between description embeddings to determine the score. The images were also ensured to have at least 2 – *DVA* between the most salient and critical regions of the scene.

Eye Movement Study

Using the test dataset, we conducted a free-viewing study in which 50 participants viewed each scene for 2 seconds without any explicit task. We tracked eye movements using an EyeLink 1000+ desktop-mounted system (spatial resolution 0.01°) operating at 2,000 Hz. Participants were positioned 75 cm (D_{obs}) from a 19-inch display, which subtended a visual angle of $26.6^\circ \times 21.8^\circ$ at a resolution of $1,280 \times 1,024$ (pixel size (s) = 0.0293 cm). The stimulus images were scaled to fit the monitor width (without altering the aspect ratio) and were centrally positioned. Participants fixated a cross located 480 pixels (10.1°) below the center of the monitor and pressed the spacebar to begin each trial. To reduce head motion, participants used a chin rest. Saccades were identified using velocity ($22^\circ/s$) and acceleration ($4,000^\circ/s^2$) thresholds. Recordings were taken from the left eye, and the experiment was run using SR Research Experiment Builder. Participants provided written informed consent prior to data collection, including explicit consent for de-identified data to be published in the public domain. Personally identifiable information was removed from all eye-tracking datasets before archiving. No data that could identify individual participants is included in this study.

Modeling Foveation

Simulation:

To simulate foveation, we require a method that degrades the information available in an image as a function of eccentricity from the fixation location. To this end, we employed the method developed by Perry and Geisler [145] and later implemented by Jiang et al. [156, 157] for modeling foveation while exploring scenes through mouse movements. Their method applies Gaussian blurring to the image, with the blur bandwidth (half-max) increasing as a function of eccentricity from the fixation point. This function is referred to as the resolution map ($\mathbf{R}$). Jiang et al. [156] define the following

resolution map:

$$\begin{aligned} \mathbf{R}(\mathbf{x}, \mathbf{y}) &= \frac{\alpha}{\alpha + \theta(\mathbf{x}, \mathbf{y})}, \\ \theta(\mathbf{x}, \mathbf{y}) &= \frac{D(\mathbf{x}, \mathbf{y})}{p}, \end{aligned} \tag{2.1}$$

where,

$\mathbf{x}, \mathbf{y}$ are pixel coordinates in the image,

α is a hyper-parameter controlling how quickly resolution falls off with eccentricity (smaller α yields a faster drop),

$\theta(\mathbf{x}, \mathbf{y})$ is the eccentricity of pixel location $(\mathbf{x}, \mathbf{y})$ from the fixation point, expressed in degrees of visual angle,

$D(\mathbf{x}, \mathbf{y})$ is the Euclidean distance of location $(\mathbf{x}, \mathbf{y})$ from the fixation location,

p is the number of pixels per degree of visual angle, and

$\mathbf{R}(\mathbf{x}, \mathbf{y})$ is the resolution map (half-max bandwidth at $(\mathbf{x}, \mathbf{y})$).

In vision research, the number of pixels per degree (p) is a critical metric used to standardize stimulus size relative to the observer's perspective. It is calculated from the observer distance (D_{obs}) and the physical pixel size (s) using the following geometric relationship:

$$p = \left\lfloor \frac{D_{obs} \cdot \tan(1^\circ)}{s} \right\rfloor,$$

where,

p is the number of pixels per degree of visual angle, (2.2)

D_{obs} is the observer distance (distance from eye to screen),

s is the physical size of a single pixel, and

$\lfloor \dots \rfloor$ denotes the floor function (integer truncation).

Approximating Gaussian Blur Using Gaussian Pyramids:

Convolving an image with a Gaussian kernel to produce blur is computationally expensive, with even efficient implementations typically running in $O(N \log N)$ time, where N is the number of pixels in the image. To address this, Perry and Geisler [145] introduced the use of Gaussian pyramids, a multi-scale representation in which the image is iteratively downsampled by a factor of two and then upsampled back to the original resolution. Perry and Geisler empirically showed that this process approximates Gaussian blurring with specific half-max bandwidths. These functions are referred to as the transfer functions (T_i). The standard deviations of these transfer functions in spatial frequency are given by:

$$\sigma_{T_1} \approx 0.124 \text{ cycles/pixel},$$

$$\sigma_{T_2} \approx 0.056 \text{ cycles/pixel},$$

$$\sigma_{T_3} \approx 0.0267 \text{ cycles/pixel},$$

$$\sigma_{T_i} \approx \frac{\sigma}{2^i} \text{ cycles/pixel},$$

where,

the first three values are empirically derived from Perry & Geisler [145],

and the general formula approximates the pattern across pyramid levels.

σ_{T_i} denotes the standard deviation of transfer function T_i

in spatial frequency,

σ is the standard deviation of the base Gaussian blur, and

i indexes the Gaussian pyramid level.

(2.3)

T_0 corresponds to the original image and can be assumed to be part of the Gaussian pyramid, originating from an even larger pyramid. Therefore, the standard deviation of T_0 can be taken as 0.248. By interpolating between these predefined transfer functions, one can efficiently approximate a Gaussian blur of any desired half-max bandwidth, thereby achieving linear time complexity. Following Perry and Geisler, the transfer functions are expressed on a relative spatial frequency axis, where 1.0 corresponds to 0.25 cycles/pixel. The function used in our study is defined as:

$$T_i(\mathbf{f}) = \begin{cases} e^{-((\mathbf{f}/4)/(\sigma/2^i))^2}, & i = 1, \dots, 10, \\ 0, & i = 11, \end{cases}$$

where,

$T_i(\mathbf{f})$ is the transfer function at pyramid level i ,

$\mathbf{f}$ is the relative spatial frequency, normalized such that $f = 1.0$ (2.4)

corresponds to 0.25 cycles/pixel,

σ is the standard deviation of the base Gaussian blur in spatial

frequency ($\sigma = 0.248$ cycles/pixel for T_0), and

i is the index of the Gaussian pyramid level ($i = 1$ is the finest blur,

and larger i correspond to coarser scales.)

Interpolation Between Transfer Functions:

To achieve arbitrary levels of blurring, the following interpolation function is used:

$$B_i(\mathbf{x}, \mathbf{y}) = \begin{cases} 0, & R(\mathbf{x}, \mathbf{y}) \leq R_i, \\ \frac{0.5 - T_i(R(\mathbf{x}, \mathbf{y})/2)}{T_{i-1}(R(\mathbf{x}, \mathbf{y})/2) - T_i(R(\mathbf{x}, \mathbf{y})/2)}, & R_i < R(\mathbf{x}, \mathbf{y}) < R_{i-1}, \\ 1, & R(\mathbf{x}, \mathbf{y}) \geq R_{i-1}, \end{cases} \quad (2.5)$$

where,

$B_i(\mathbf{x}, \mathbf{y})$ is the interpolation weight at pixel (x, y) for pyramid level i ,

$R(\mathbf{x}, \mathbf{y})$ is the half-max bandwidth given by the resolution map,

$R(\mathbf{x}, \mathbf{y})/2$ converts this bandwidth to the normalized frequency axis used
by the transfer functions

R_i, R_{i-1} are the half-max bandwidths corresponding to transfer
functions T_i and T_{i-1} ,

$T_i\left(\frac{R(\mathbf{x}, \mathbf{y})}{2}\right)$ is the Gaussian transfer function at scale i evaluated at the
half-max bandwidth $R(x, y)$, and 0.5 is the half-maximum
sensitivity used to interpolate between adjacent scales.

Multiple Fixations:

To incorporate multiple fixations, the eccentricity maps $\theta_j(\mathbf{x}, \mathbf{y})$ generated by each fixation are combined using the following equation:

$$\theta_{\text{combined}}(\mathbf{x}, \mathbf{y}) = \min(\theta_1(\mathbf{x}, \mathbf{y}), \theta_2(\mathbf{x}, \mathbf{y}), \dots, \theta_N(\mathbf{x}, \mathbf{y})),$$

where, (2.6)

N is the number of fixations made, and

$\theta_j(\mathbf{x}, \mathbf{y})$ is the eccentricity map for the j^{th} fixation.

This combined eccentricity map effectively deblurs all previously fixated locations. This provides a convenient way of combining the information gained from multiple fixation locations into a single foveated representation.

Vision Language Model (VLM)

In this study, we use Ovis2.5 [140, 141] as our base vision-language model (VLM). Ovis2.5 is a decoder-only transformer that accepts multimodal inputs (text, images,

and videos) and produces multimodal outputs in the same domains. Here, we restrict the inputs to images (both foveated and unfoveated) and text, and generate two complementary outputs: natural-language descriptions of the scene (the standard capability of the model) and predictions of the next fixation location (our novel extension).

Internally, Ovis2.5 incorporates the Native Vision Transformer (NaViT) [158] for image encoding. NaViT partitions an input image into contiguous 16×16 pixel patches, transforms each patch into a feature vector, and then projects these vectors into a 4096-dimensional embedding space shared across modalities. For text, Ovis2.5 relies on the Qwen3 large language model [159], which likewise projects token embeddings into a 4096-dimensional space. The vision and language embeddings are concatenated and passed into a cross-attentional decoder. In Ovis2.5, the decoder produces a corresponding 4096-dimensional output vector for every input patch or text token embedding. The outputs aligned with image tokens are routed into our eye movement model to predict the next fixation, while the decoder outputs at text positions are passed through the language modeling head to generate the sentence.

Eye Movement Model

The eye movement model uses average spatial pooling with a stride length of two on the input image tokens from Ovis2.5, followed by a Convolutional Neural Network (CNN) with Pointwise Convolution, i.e., a convolution layer with 1×1 kernel, combined with layer normalization and ReLU activations, to sequentially project the 4096-dimensional decoder feature vector for each image token into a lower-dimensional space:

$$\begin{aligned}
\mathbf{Input} \in \mathbb{R}^{4096 \times 2H \times 2W} &\xrightarrow{\text{Spatial Pooling}} \mathbb{R}^{4096 \times H \times W} \\
&\mapsto \mathbb{R}^{1024 \times H \times W} \mapsto \mathbb{R}^{256 \times H \times W} \mapsto \mathbb{R}^{64 \times H \times W} \mapsto \mathbb{R}^{1 \times H \times W}
\end{aligned} \tag{2.7}$$

where,

H and W are the height and width of the action space.

Each decoder feature corresponding to an image patch encodes not only the visual content of that patch but also its semantic and spatial relationships with other patches. The initial average pooling halves the number of input image tokens, thereby reducing the total computational cost of the subsequent convolution layers. To predict eye movements, it is crucial to preserve the spatial alignment while extracting the relevant features. The Pointwise Convolutions ensure that information is processed independently for each token (i.e., without pooling across neighboring patches), while the dimensionality reduction progressively distills the features into a single scalar value per patch. These scalars form a map over the pooled image tokens ($\mathbf{M} \in \mathbb{R}^{H \times W}$), which is then passed through a softmax layer to produce a probability distribution over possible fixation locations $\mathbf{p}(\mathbf{x}, \mathbf{y})$. The model outputs the next fixation location $(\mathbf{X}, \mathbf{Y})$ by either sampling (during training) or greedy selection (during inference) (i.e., $(\mathbf{X}, \mathbf{Y}) = \arg \max_{\mathbf{x}, \mathbf{y}} \mathbf{p}(\mathbf{x}, \mathbf{y})$) from this probability distribution.

$$\mathbf{p}(\mathbf{x}, \mathbf{y}) = \frac{\exp(\mathbf{M}_{\mathbf{x}, \mathbf{y}} / \tau)}{\sum_{i=1}^H \sum_{j=1}^W \exp(\mathbf{M}_{i,j} / \tau)}$$

where, (2.8)

τ is the temperature parameter controlling
the sharpness of the softmax distribution.

To determine the size of the action space, we note that the VLM maps the input image from pixel space to token space with a reduction factor of 32, and the subsequent pooling introduces an additional reduction factor of 2. Thus, for an image of height H_{image} and width W_{image} , the action space is given by:

$$\text{action space} = \frac{H_{image} \times W_{image}}{64^2},$$

where, (2.9)

the factor 64 arises from a $32 \times$ tokenization reduction followed by
 $2 \times$ spatial pooling.

To standardize the action space across images of different sizes, we define a fixed field of view in degrees of visual angle (DVA) within which all images are embedded (see below).

Finally, to simplify training, we train a separate model for each fixation step, all sharing the same architecture described above. Thus, for N fixations, we train N independent models.

Reinforcement Policy

To train our model, we need to assign rewards to the actions taken by the model. The goal of our eye movement model is to make eye movements that maximize scene understanding. To facilitate this, we convert the VLM into a scene-understanding model by providing the following text prompts along with the image it has to interpret:

Text prompt 1: *"Provide your best guess of what is happening in the scene in a sentence. Do not mention the blur seen in the picture."*

Text prompt 2: *"In one sentence, describe what is happening in the scene and, if people are present, identify the person if possible, determine how they appear to feel, and their gender. Do not mention the blur seen in the picture. If you do not know who the person is, simply refer to them as man, woman, girl, boy, or person."*

The main analysis in this paper covers the results for prompt 1. The results for prompt 2 are provided in Fig. 31. The prompt includes an instruction not to mention blur because VLMs often explicitly describe visual artifacts such as blur when they are present in an image. Since our foveation method introduces blur in the peripheral regions of the image, we include this instruction to ensure that the model focuses on describing the scene content rather than commenting on the blur introduced by foveation.

Now suppose the model has made $\mathbf{j}$ fixations ($\mathbf{j} \leq \mathbf{N}$), where the initial fixation corresponds to $\mathbf{j} = \mathbf{0}$. Using the foveation technique discussed above, this corresponds to deblurring $\mathbf{j} + \mathbf{1}$ regions on the image (at $\mathbf{j}$ model fixations and one initial fixation). With this setup, an ideal reward function should satisfy the following properties:

- Fixating regions whose information content is important for understanding the scene should yield a high reward. Since we want the model to preferentially fixate such informative regions, the reward function should assign higher values to fixations that reveal useful scene information.
- The cumulative reward should be monotonically increasing as a function of the number of fixations. In other words, the reward after a single fixation should be the lowest, and subsequent fixations should increase or at least maintain the reward. This reflects the intuition that revealing additional regions of the image through deblurring should not decrease the information available about the scene.

Using the VLM’s descriptions after $\mathbf{j} + \mathbf{1}$ fixations, we explored the following options as possible reward functions that satisfy the above conditions.

Semantic Accuracy of Descriptions

We get semantic accuracy by computing similarity between the VLM’s descriptions when the image is foveated (an image with $\mathbf{j}+1$ deblurred regions), and the descriptions generated when the image is unfoveated (the intact original image), which represents the maximum level of understanding the model can achieve for that image.

For each image, we generate $\mathbf{M}$ descriptions for both the foveated and unfoveated conditions after each fixation. These descriptions are converted into embeddings using a Large Language Model called Jasper [160]. Jasper is an open-source model trained to produce sentence embeddings by distilling representations from larger closed-source embedding models. The embeddings are then used to compute the mean pairwise cosine similarity ($\mathbf{CS}$) between the foveated and unfoveated descriptions.

Let $\mathbf{E}_i$ and $\mathbf{E}_{ij}^{fov}$ denote the embeddings of the i^{th} unfoveated description and the i^{th} foveated description after $\mathbf{j}$ model fixations. Then the cosine similarity between the two sets of descriptions is given by:

$$CS(\mathbf{j}) = \frac{1}{M} \sum_{i=1}^M \frac{\mathbf{E}_{ij}^{fov} \cdot \mathbf{E}_i}{\|\mathbf{E}_{ij}^{fov}\| \|\mathbf{E}_i\|},$$

where,

$\mathbf{E}_i$ is the embedding of the i^{th} unfoveated description, (2.10)

$\mathbf{E}_{ij}^{fov}$ is the embedding of the i^{th} foveated description after

$\mathbf{j}$ model fixations, and

$\mathbf{M}$ is the number of generated descriptions.

The cosine similarity after $\mathbf{j}$ model fixations, defined above, acts as a cumulative metric

indicating how effective the first j model fixations are in revealing information needed to understand the scene. To compute the reward $R(j)$ for choosing model fixation j , we measure how much the current cosine similarity improves upon the best similarity obtained from previous fixations. In other words, a positive reward is obtained only when the current fixation reveals additional information that increases the similarity between the foveated descriptions and the unfoveated descriptions beyond what has already been achieved by earlier fixations.

$$R(j) = CS(j) - \max_{k < j} (CS(k)), \quad j \geq 1,$$

$$R(0) = CS(0), \quad \text{initial fixation}$$

where,

(2.11)

$CS(j)$ denotes the cosine similarity after j model fixations.

$R(j)$ denotes the semantic reward for choosing j^{th} model fixation.

Theoretically, as we increase the number of fixated locations, the model descriptions can ultimately approach the unfoveated descriptions (i.e., the case in which the entire image is deblurred), thereby representing the upper limit of scene understanding. Therefore, to standardize the computed rewards across images, we normalize the rewards using the mean pairwise cosine similarity among the unfoveated descriptions, denoted by $CS_{\text{upper limit}}$. To ensure boundedness, the normalized rewards are clipped to lie between 0 and 1:

$$\mathbf{R}_{\text{norm}}(\mathbf{j}) = \frac{\mathbf{R}(\mathbf{j})}{(\mathbf{CS}_{\text{upper limit}} - \mathbf{R}(0)) + \epsilon}, \quad \mathbf{j} \geq 1,$$

$$\mathbf{R}_{\text{norm}}(0) = 0, \quad \text{initial fixation}$$

$$\mathbf{R}_{\text{norm}}(\mathbf{j}) = 1, \quad \forall \mathbf{R}_{\text{norm}}(\mathbf{j}) > 1,$$

$$\mathbf{R}_{\text{norm}}(\mathbf{j}) = 0, \quad \forall \mathbf{R}_{\text{norm}}(\mathbf{j}) < 0,$$

$$\text{where,} \tag{2.12}$$

$\mathbf{CS}_{\text{upper limit}}$ is the mean pairwise cosine similarity computed among the unfoveated descriptions for a given image.

$\mathbf{R}_{\text{norm}}(\mathbf{j})$ is the normalized semantic reward for choosing the $\mathbf{j}^{th}$ model fixation.

ϵ is a small constant for numerical stability

Conditional Entropy

Another possible reward function stems from the idea that the uncertainty introduced by foveation in the image pixel space should translate into uncertainty in the model’s generated descriptions. Since VLMs such as Ovis2.5 are next-token predictors given an input, description generation involves predicting tokens that correspond to words. Each predicted token is associated with a probability distribution over the vocabulary, which can be used to compute the entropy rate.

Let $\mathbf{X}_0, \mathbf{X}_1, \dots, \mathbf{X}_T$ denote the sequence of tokens forming a generated sentence, where $\mathbf{X}_t$ is the t^{th} token predicted given input image $\mathbf{I}$ and the previously predicted tokens $\{\mathbf{X}_0, \mathbf{X}_1, \dots, \mathbf{X}_{t-1}\}$. The conditional entropy rate of the generated sequence can be

written as:

$$\begin{aligned}
H(\mathbf{X}_0, \mathbf{X}_1, \dots, \mathbf{X}_T \mid \mathbf{I}) &= \sum_{t=0}^T H(\mathbf{X}_t \mid \mathbf{X}_0, \dots, \mathbf{X}_{t-1}, \mathbf{I}), \\
H(\mathbf{X}_t \mid \mathbf{X}_0, \dots, \mathbf{X}_{t-1}, \mathbf{I}) &= - \sum_{\mathbf{k} \in \mathcal{V}} p_{\mathbf{k}} \log p_{\mathbf{k}},
\end{aligned}$$

where,

$$p_{\mathbf{k}} = p(x_t = k \mid X_0, \dots, X_{t-1}, I),$$

$\mathcal{V}$ is the vocabulary (i.e., the set of all possible tokens), and

$H(\cdot)$ denotes Shannon entropy.

(2.13)

For M generated sentences from the VLM after $\mathbf{j}$ model fixations, the average entropy rate $\bar{H}(\mathbf{j})$ is given by:

$$\bar{H}(\mathbf{j}) = \frac{1}{M} \sum_{m=1}^M \frac{1}{T_{mj}} H(\mathbf{X}_0^{(mj)}, \mathbf{X}_1^{(mj)}, \dots, \mathbf{X}_{T_{mj}}^{(mj)} \mid \mathbf{I}),$$

where,

$$\begin{aligned}
M &\text{ is the number of generated sentences,} \\
T_{mj} &\text{ is the length (number of tokens) of sentence } m \text{ obtained} \\
&\text{ from VLM after } \mathbf{j} \text{ fixations, and} \\
\mathbf{X}_t^{(mj)} &\text{ is the } t^{\text{th}} \text{ token in sentence } m \text{ obtained from VLM after} \\
&\mathbf{j} \text{ model fixations.}
\end{aligned}$$
(2.14)

Similar to the cosine similarity reward, we compute the reward at each fixation based on whether the current fixation reduces the model’s uncertainty beyond what has already been achieved by previous fixations. Specifically, we compare the average entropy at the current fixation to the minimum average entropy obtained across all earlier fixations. A positive reward is obtained only when the current fixation leads to a new minimum entropy, i.e., when it reduces uncertainty more than any previous fixation. If the current fixation does not improve upon the best (lowest) entropy observed so far, the reward becomes negative and is clipped to zero. Unlike the cosine similarity reward, this formulation does not rely on unfoveated VLM descriptions. Instead, it more directly captures the idea that scene exploration should be rewarded when it progressively reduces uncertainty about what is being seen. Therefore,

$$\begin{aligned} \mathbf{R}_e(j) &= \max \left(0, \min_{k < j} (\bar{H}(k)) - \bar{H}(j) \right), \quad j \geq 1, \\ \mathbf{R}_e(0) &= 0, \quad \text{initial fixation} \end{aligned} \tag{2.15}$$

where,

$\bar{H}(j)$ denotes the average entropy after j model fixations.

$\mathbf{R}_e(j)$ denotes the entropy reward for choosing j^{th} model fixation.

REINFORCE Algorithm

The fixation prediction problem can be formulated as a reinforcement learning problem in which the eye movement model represents a stochastic policy that selects fixation locations on the image. After each fixation, the model predicts a probability distribution over all possible image patches. Sampling from this distribution corresponds to selecting the next fixation location.

Given that we train separate models for each fixation step, let $\pi_{\phi_j}(\mathbf{a}_j|\mathbf{s}_{j-1})$ denote the probability of selecting the current fixation location $\mathbf{a}_j$ given the previous state $\mathbf{s}_{j-1}$, where ϕ_j represents the parameters of the eye movement model for predicting j^{th} model fixation. The state $\mathbf{s}_{j-1}$ corresponds to the foveated image after $j-1$ model fixations. The goal of training is to maximize the expected reward associated with the fixation actions.

Using the REINFORCE algorithm [161], the gradient of the expected reward with respect to the policy parameters can be written as:

$$\begin{aligned} \nabla_{\phi_j} J(\phi_j) &= \mathbb{E} \left[\mathbf{R}(j) \nabla_{\phi_j} \log \pi_{\phi_j}(\mathbf{a}_j|\mathbf{s}_{j-1}) \right], \\ \text{where,} & \\ J(\phi_j) &\text{ is the expected reward objective for the} \\ &\text{policy choosing } j^{\text{th}} \text{ model fixation.} \end{aligned} \tag{2.16}$$

In practice, this expectation is approximated using a batch of sampled fixation trajectories from $\mathbf{B}$ images. To reduce the variance of the gradient estimate, we subtract a baseline corresponding to the average reward across the batch:

$$\nabla_{\phi_j} \mathbf{J}(\phi_j) = \frac{1}{B} \sum_{i=1}^B (\mathbf{R}_i(\mathbf{j}) - \bar{\mathbf{R}}(\mathbf{j})) \nabla_{\phi_j} \log \pi_{\phi_j}(\mathbf{a}_{ij} | \mathbf{s}_{ij-1}),$$

$$\bar{\mathbf{R}}(\mathbf{j}) = \frac{1}{B} \sum_{i=1}^B \mathbf{R}_i(\mathbf{j}),$$

where,

$\mathbf{R}_i(\mathbf{j})$ is the reward for choosing image $\mathbf{j}^{th}$ model fixation for image $\mathbf{i}$, (2.17)

$\mathbf{a}_{ij}$ is the sampled fixation $\mathbf{j}$ for image $\mathbf{i}$,

$\mathbf{s}_{ij-1}$ is the image $\mathbf{i}$ foveated at the initial and $\mathbf{j} - 1$ previous
model fixation locations

$\bar{\mathbf{R}}(\mathbf{j})$ is the average reward across the batch for choosing fixation $\mathbf{j}$.

Since fixation locations correspond to spatial positions on the image grid, nearby patches often reveal similar visual information. To encourage spatially smooth learning, we apply Gaussian smoothing to the policy gradient. Instead of assigning the gradient update only to the sampled fixation location, the update is distributed to neighboring patches according to a spatial weighting function. Note that the weights are not normalized. The Gaussian kernel is used purely as a spatial decay function.

Let $\mathbf{k}$ denote a patch index and $\mathbf{x}_k$ the spatial coordinate of that patch. The Gaussian weighting function is defined as:

$$w(\mathbf{a}, \mathbf{k}) = \exp \left(-\frac{\|\mathbf{x}_a - \mathbf{x}_k\|^2}{2\sigma^2} \right),$$

where,

$\mathbf{x}_a$ is the spatial coordinate of the sampled fixation location a , (2.18)

$\mathbf{x}_k$ is the spatial coordinate of patch k , and

σ controls the spatial spread of the smoothing kernel.

Using this weighting, the smoothed policy gradient becomes:

$$\nabla_{\phi_j} \mathbf{J}_{\text{smooth}}(\phi_j) = \frac{1}{B} \sum_{i=1}^B \sum_{\mathbf{k}} w(\mathbf{a}_{ij}, \mathbf{k}) (\mathbf{R}_i(j) - \bar{\mathbf{R}}(j)) \nabla_{\phi_j} \log \pi_{\phi_j}(\mathbf{k} | \mathbf{s}_{ij-1}),$$

where, (2.19)

$w(\mathbf{a}_{ij}, \mathbf{k})$ distributes the update from sampled fixation a_{ij}
to neighboring patch k .

This smoothing distributes the gradient update to nearby spatial locations, reflecting the fact that neighboring patches often contain similar information relevant for scene understanding. In practice, this leads to more stable learning and encourages the model to learn coherent spatial fixation policies. Finally, the weights of the model corresponding to each fixation are updated to maximize reward by performing a gradient ascent

using the AdamW optimizer, as shown

$$\begin{aligned}
\mathbf{g}_t &= \nabla_{\phi_j} \mathbf{J}(\phi_j), \\
\mathbf{m}_t &= \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t, \\
\mathbf{v}_t &= \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t^2, \\
\hat{\mathbf{m}}_t &= \frac{\mathbf{m}_t}{1 - \beta_1^t}, \\
\hat{\mathbf{v}}_t &= \frac{\mathbf{v}_t}{1 - \beta_2^t}, \\
\phi_j^{(t+1)} &= \phi_j^{(t)} + \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon} - \eta \lambda \phi_j^{(t)},
\end{aligned} \tag{2.20}$$

where,

$\mathbf{g}_t$ is the gradient of the objective at iteration t ,

$\mathbf{m}_t$ and $\mathbf{v}_t$ are the first and second moment estimates,

β_1, β_2 are the exponential decay rates for these moments,

η is the learning rate,

λ is the weight decay coefficient, and

ϵ is a small constant for numerical stability.

Implementation details

Stimuli

To enable valid comparisons between human and model eye movements, it is necessary to ensure that the images used by the model correspond to approximately the same

visual angles as the images shown to participants during the eye-tracking study. To achieve this, the model’s field of view was set to a square region of size $26.6^\circ \times 26.6^\circ$ of visual angle, corresponding to a resolution of 1280×1280 pixels at an observer distance of 75 cms. All images were embedded within this space at the appropriate size and spatial location to match the presentation conditions used in the eye-tracking experiment. Unless stated otherwise, all analyses described in this Methods section are performed on images embedded within this standardized space.

The training dataset consists of 24,000 images compiled from three publicly available datasets (as described above). From each dataset, a subset of 8,000 images was selected to construct the final training set. Our goal was to ensure that the selected images required more than the initial fixation for the model to understand the scene. To approximate this criterion, we computed the cosine similarity between the descriptions generated from foveated and unfoveated versions of each image. For this comparison, the foveation was placed at the corner of the image. This ensured that the initially revealed region was unlikely to correspond to an informative part of the scene, thereby avoiding the possibility of accidentally selecting a location that might already contain critical information for that specific image. By using a corner location, we ensured that the revealed region was consistently uninformative across images.

Images were then ranked by cosine similarity, and we selected the 8,000 images with the lowest similarity values from each dataset. These images correspond to scenes in which the model’s understanding improves substantially when additional regions are revealed, thereby encouraging the model to perform exploratory fixations.

Hardware and Software

Model training was performed using seven NVIDIA RTX A6000 GPUs, each with 48 GB of memory. Since the VLM (OVIS2.5) is used only for inference during training, we

employed the vLLM library [162] to increase inference efficiency. We hosted the VLM on five GPUs and used the remaining two GPUs as client nodes to request outputs from the hosted VLM instances. These two GPUs were responsible for training the eye movement model. Training was performed synchronously across the two GPUs using PyTorch’s Distributed Data Parallel (DDP). We modified the vLLM implementation to additionally output the hidden states and token probabilities needed to train the eye movement model.

Description Sampling from VLM using vLLM library

Given that we generate M descriptions from the VLM, a common approach would be to use beam search to obtain the top candidate descriptions. However, beam search often produces sentences that follow very similar token trajectories, resulting in highly similar descriptions. This behavior is not ideal in our case because we wish to capture the uncertainty introduced by foveation in the image. To obtain more diverse yet deterministic descriptions, we modified the sampling procedure in vLLM. Specifically, we perform M separate greedy decoding runs, each seeded with one of the top-M first tokens. This produces M descriptions that differ in their initial token while remaining deterministic during generation, ensuring each description is obtained from a unique trajectory. This sampling method for descriptions was used for all training and analysis done as part of this study.

Foveation Parameters Calibration

To enable fair comparison with humans, we aimed to set the foveation parameters (discussed above) to approximate properties of human vision. An observer distance of 75 cm sets the pixels-per-degree value (P) to 44 pixels/deg, matching our human studies. A Gaussian pyramid of depth 10 was sufficient to cover the resolution drop at any arbitrary distance from the fixation point within the visual field ($26.6^\circ \times 26.6^\circ$).

To match the resolution map of the human eye, we adjusted the α value using a separately conducted limited-fixation (two or four) scene description study on humans ($N = 20$) with 277 images (86 overlapping with the free-viewing dataset). The participant’s eyes were tracked using the same setup as in the free-viewing study. A gaze-contingent display was implemented to ensure the image presentation stopped after four fixations for each participant. Since the VLM (OVIS2.5) may not describe all images as accurately as humans with unlimited viewing time, we first selected a subset of 124 images in which the average similarity of VLM descriptions (without foveation) matched the within-human description similarity (unlimited viewing time).

Using this subset, we chose α such that the cosine similarity between VLM descriptions of foveated images (based on human fixations from the scene description study) and human descriptions with unlimited viewing time matched the cosine similarity between human descriptions after two or four fixations and those with unlimited viewing time. The resulting value of α was 0.63.

Although participants in our study viewed stimuli from an observer distance of 75 cm, human visual expertise is honed through real-world exposure to objects and faces at specific visual angles. To ensure that our models were trained on ecologically valid scales, we referenced naturalistic observation data from Oruç et al. [163]. Their findings indicate that while the median face size in the adult "face-diet" is approximately $6.0 - DVA$, unfamiliar faces, which more closely resemble our test stimuli, subtend a median visual angle of $4.9 - DVA$.

In our training dataset, a sample of 1000 images containing people yielded an average face size of only $2.6 - DVA$ at the initial 75 cm observer distance. To align these stimuli with the median $4.9 - DVA$ value encountered in daily life, we recomputed the required

observer distance (D) to be 39.3 cm. Consequently, the models were trained using the human-calibrated foveation parameter ($\alpha = 0.63$), but at this adjusted observer distance, to better approximate the high-resolution features characteristic of natural human social interaction. Testing was still conducted at an observer distance of 75 cm to match human experiments.

Action space for the models

Given a fixed field of view of 1280×1280 pixels, the action space of the eye movement model consists of 400 possible locations ($1280^2/64^2$, see Methods), arranged as a 20×20 grid. Each grid cell corresponds to a square region of 64×64 pixels or $1.4 \times 1.4 - DVA$ overlaid on the image.

An action $(\mathbf{X}, \mathbf{Y})$ corresponds to selecting the center of a grid cell as the fixation location. Thus, $(\mathbf{X} = 0, \mathbf{Y} = 0)$ corresponds to the center of the top-left cell, while $(\mathbf{X} = 19, \mathbf{Y} = 19)$ corresponds to the center of the bottom-right cell.

Training procedure

The eye movement model is trained to generate four sequential fixations following an initial fixation, and each model is trained for a total of five epochs. The effective batch size across two GPUs is 60 images. For each batch, the initial fixation is randomly selected from five predefined locations: the four corners of the image and a location 300 pixels below the center of the field of view. With the current setup, it took 4 days to train a model. To ensure reproducibility, we trained five independent instances of the fRL-SU model and report fixation frequencies and metrics averaged across these instances. Control models (fRL-SC, fRL-Search, and fRL-Entropy) were each trained as a single instance.

Since each fixation is modeled by a separate but architecturally identical Convolutional

Neural Network (CNN), the models are trained sequentially. Specifically, each CNN model is trained on five batches (300 images). During training to predict the j^{th} fixation, the corresponding CNN model receives the intermediate visual output from the VLM, conditioned on the input image and the previous $j - 1$ model fixations. The CNN outputs are then transformed into a softmax distribution with temperature $\tau = 3$, from which the j^{th} fixation is sampled. The preceding $j - 1$ fixations are obtained via greedy selection from their corresponding frozen CNN models.

The learning rate (η) for the AdamW optimizer is set to 0.0002, with weight decay (λ) set to 0.0001, and all other hyperparameters use the default values provided by the PyTorch implementation of AdamW. The spatial smoothing Gaussian applied in the loss function has a standard deviation of $\sigma = 1.5$.

Controls

Task and foveation

To train the eye movement model, the input prompt to the VLM is fixed, instructing it to describe the given input image. As a control, we also train the eye movement model using alternative input prompts corresponding to different tasks and compare the resulting eye movements with human eye movements. Specifically, we trained two additional models: one performing a search task and another performing a scene categorization task based on categories from Anderson et al. [152]. We use the following prompts for the respective tasks.

Text prompt 1 (Search task):

“Provide your best description of the smallest man-made object present and clearly visible in the image in a sentence. Do not mention the blur seen in the picture.”

Text prompt 2 (Scene Classification task):

“Analyze this scene and classify it by selecting exactly one label from each of the three dimensions below. Do not use any outside vocabulary. Semantic Content: Nature, Road, Residence, Farm, Beach, Car Park 3D Spatial Structure: Cluttered, Closed Off, Flat, Tunnel 2D Image Appearance: Dark, Bright, Blue, Green, Brown Format your output strictly as: Semantic: [label] Spatial: [label] Appearance: [label]”

For simplicity, the models trained using the above prompts were optimized solely using the semantic accuracy reward. Additionally, we trained models with very high ($\alpha = 0.2$) and very low foveation ($\alpha = 20$). All models were trained, at an observer distance of 39.3 cm, to make four fixations after the initial fixation (see details above).

The Oracle model

To determine the theoretical upper bound of scene understanding for the foveated VLM, we developed an Oracle model that, unlike the fRL model, exhaustively executes and evaluates every possible fixation location within the visual field. The model constructs a fixation sequence greedily by selecting the location that maximizes semantic accuracy at each step, conditioned on prior fixations. Because the action space consists of 400 possible locations per fixation, this approach is computationally expensive, taking approximately three days to process the test dataset.

Existing fixation prediction models

DeepGaze

DeepGaze [49] is a neural network model that combines low-level image features from a ResNet-50 CNN and human fixation data to predict a fixation density map. This study uses an implementation provided by Kümmerer [81]. Since this results in a static map for each image, we adapt it to generate sequential fixations in our test dataset by

smoothing the DeepGaze map with a $2 - DVA$ circular average kernel and selecting four fixation locations sequentially using inhibition of return (IOR).

The IOR is modeled as a circular patch whose size is set to match the average saccade distance per fixation observed in human free-viewing data. For DeepGaze, the diameter of the IOR patch is $2.6 - DVA$.

Saliency Models

As part of this study, we compare two low-level saliency models: Itti Koch [5] and Graph-Based Visual Saliency (GBVS) [6]. The Itti-Koch model predicts fixation locations by combining multi-scale feature contrasts (color, intensity, orientation) into a bottom-up saliency map. GBVS extracts similar low-level features but constructs a graph over them and uses Markov chains to generate a saliency map. This study uses an implementation provided by Kümmerer [78].

Similar to DeepGaze, we generate sequential fixations by selecting four locations per image through iterative maximum selection on smoothed ($2 - DVA$ kernel) saliency maps with inhibition of return matched to human data. For GBVS, the IOR diameter is $2.8 - DVA$, while for the Itti-Koch model, it is $2 - DVA$. The differences in IOR diameters reflect the sparsity of the predicted saliency maps, with sparser maps requiring smaller IOR regions.

Random

For the random baseline, 1000 samples of four fixations were uniformly drawn from the pixels corresponding to the image region within the field of view. The resulting average inter-fixation distance was substantially larger ($11 - DVA$) than that observed in human data, even without applying inhibition of return. Nevertheless, we imposed an IOR constraint with a diameter of $2 - DVA$ during sampling.

Fixation prediction heatmaps for trained models

The eye movement model outputs a probability map of size 20×20 for each predicted fixation; thus, four fixations yield four such maps. To obtain a single static fixation prediction map, we compute the element-wise maximum across these maps. This operation aggregates the most probable regions from each fixation step, assuming that each map highlights distinct regions of interest. Taking the maximum, therefore, combines these high-probability regions into a single map, providing a consolidated estimate of likely fixation locations.

Measured fixations heatmaps

Fixations on an image were represented as pixel-wise white dots on a blank canvas and convolved with a Gaussian kernel with a standard deviation of 0.25° visual angle. OpenCV2's `GaussianBlur` function was used for this purpose. The convolved fixation points were superimposed onto a blank map at their corresponding (x, y) locations to generate the heatmap. To match the trained models, the heatmaps were generated for the first four measured fixations after the initial fixation.

Fixation prediction heatmap comparison metrics

In this study, we use the following standard metrics from the literature: Area Under ROC (AUC) and Correlation Coefficient.

- **AUC:** Measures how well a saliency map discriminates fixation locations from non-fixated locations.
- **CC:** Measures the linear similarity between the predicted saliency map and the human fixation density map.

This study uses an implementation of these metrics provided by Kümmerer [78].

Fixation frequency to categories comparisons

To compare fixation patterns, we classified the visual elements in each scene and quantified how frequently the first four subsequent fixations fell within these categories. We generated binary segments for each object or region, using a spatial tolerance of $0.7 - DVA$ from the segment edges to determine category membership (see Fig. 31 for generalization of results with other tolerances). This analysis was applied across humans, saliency models, and our trained and control models. The categories are as follows:

People:

This category corresponds to all people present in an image. The Segment Anything Model 3 (SAM3) [164] was used to detect and segment people in each scene, and the resulting masks were used. Of the 147 test images, 85 contain people.

Center bias:

This corresponds to the central $5 - DVA$ region in each image. A circular binary mask of size $5 - DVA$ centered in each image was used to count fixations falling into this category. All 147 images in the dataset include a center bias mask.

SU-R Object:

This corresponds to objects that are critical for understanding the given image. Based on previous work on scene understanding maps (SUMs) [137, 165], these critical objects were identified by quantifying the impact of removing objects/people from a scene on the scene descriptions generated by a Multi-Modal Large Language Model (MLLM). The object/people removal, scene description generation, and impact measurement using embedding-based description ratings were all automated (AUTOSUM), as detailed in [165]. To ensure consistency with the SU-R Object criteria, we excluded any peo-

ple present in the scene from the AUTOSUM predictions. The remaining scores were normalized relative to the most impactful object in each image. Segments achieving a normalized score of 0.95 or higher were classified as SU-Relevant (SU-R) Objects; these segments then served as masks for calculating fixation frequency.

The SU-R object category was further split into the following subcategories:

SU-R Object (Gaze/Grasp) These are images where the SU-R Object is either grasped or gazed at by people present in the scene. Naturally, this forms a subset of images that contain people. Of the 85 images with people, 53 include gaze or grasp cues directed toward the SU-R Object, and the SU-R masks for these 53 images are included in this category.

SU-R Object (No Gaze/Grasp) These are images in which the SU-R Object has no gaze or grasp cues directed toward it. All images without people naturally fall into this category. There are 45 such images that contain SU-R Objects with no gaze/grasp cues, and their masks are included here. We ignore images that contain people but lack gaze/grasp cues toward the SU-R Object due to the limited number of such cases (only 9 images).

SU-I Objects:

All objects with an AUTOSUM score below 0.95 were considered SU-Irrelevant (SU-I) Objects. Since there are typically many SU-I objects in each scene, their masks were normalized by the number of SU-I objects per image to ensure fair comparison with other categories.

Similar to SU-R Objects, SU-I Objects were split using the same set of images as the two SU-R categories (53 and 45 images, respectively), ensuring fair comparisons.

Text:

Images containing text were manually segmented, and 4-sided polygon masks were created for text regions. Of the 147 images, 51 contain text regions.

Saliency (GBVS):

This category corresponds to regions identified by the top prediction of the GBVS saliency map. For images containing an SU-R Object, a circular patch centered at the top prediction location, with an area matching the SU-R Object segment, was used as the mask. For images without an SU-R Object, a circular patch matching the average size of SU-R Objects across the test dataset was used. All 147 images include a saliency mask.

Comparison of model and human fixation frequency distributions

To assess how well the models predict the distribution of human fixation frequencies across these categories, we measure the divergence between the models’ predicted fixation frequencies and the empirical human distribution. We frame this as a Negative Log-Likelihood (NLL) problem. By modeling the human fixation frequencies across object categories as a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, we calculate the NLL of the model’s predictions under this human prior.

We compute two variants of NLL: one assuming the categories are independent (the main results), and the other assuming they are not (see Fig. 23). The independent variation is as follows:

$$\text{NLL}_{\text{indep}} = - \sum_{i=1}^k \left[\frac{(\mathbf{x}_i - \boldsymbol{\mu}_i)^2}{2\text{SE}_i^2} + \frac{1}{2} \ln(2\pi\text{SE}_i^2) \right]$$

where,

k is total number of object categories (2.21)

$\mathbf{x}_i$ is the model's predicted fixation frequency for category i

$\boldsymbol{\mu}_i$ is the mean empirical human fixation frequency for category i

SE_i is the standard error of the human frequencies for category i

In the dependent variation, we calculate the covariance between categories based on the fixation frequencies of 50 participants in the free-viewing study, averaged across all 147 images in the test dataset. The NLL is computed as follows:

$$\text{NLL}_{\text{mvn}} = -\left(\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) + \frac{1}{2} \ln |\boldsymbol{\Sigma}| + \frac{k}{2} \ln(2\pi)\right)$$

where,

$\mathbf{x}$ is the vector of the model's predicted fixation frequencies across all categories

$\boldsymbol{\mu}$ is the vector of the mean empirical human fixation frequencies

$\boldsymbol{\Sigma}$ is the covariance matrix of the empirical human data

$\boldsymbol{\Sigma}^{-1}$ is the inverse of the covariance matrix (precision matrix)

$|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix

T is the transpose operator

k is the total number of object categories (2.22)

To allow comparison across datasets, the NLL scores were normalized by the NLL score of the Itti-Koch saliency model, yielding the Normalized NLL (NNLL):

$$\text{NNLL} = \frac{\text{NLL}_{\text{model}}}{\text{NLL}_{\text{Itti-Koch}}} \quad (\text{S23})$$

A score closer to 0 indicates better agreement with human fixation frequencies, while a score of 1 indicates performance equivalent to the Itti-Koch baseline, the worst-performing model in this study.

Description similarity comparisons

Given that the VLM generates image descriptions in response to the text prompt (defined above) and that the eye movement model can be constrained to fixate at specified locations, we use fixation sequences from various models and participants in the test dataset and force the eye movement model to follow these sequences. We then record the model-generated descriptions for each fixation sequence.

These descriptions are compared to the unfoveated model descriptions of the same image using cosine similarity between sentence embeddings obtained from the Jasper embedding model [160]. To compute the description similarities for the random model, we used only 10 out of the 1000 generated samples (see Methods above) to keep the computation tractable.

Testing on publicly available dataset

To evaluate our models, we use an external dataset from a recent free-viewing study by Linka et al. [50]. This dataset comprises a subset of the OSIE dataset (40 images; OSIE 40, [166, 74]) and includes a substantial amount of participant data ($N = 6,720$) collected across different age groups. Linka et al. analyzed these data by calculating the proportion of fixations falling into four distinct categories: 1) Head, 2) Body, 3) Touched (objects actively touched by people present in the scene), and 4) Text.

Stimuli Dimensions

Because observer distance in the Linka et al. [50] study varied between 50 cm and 90 cm, the subtended visual angle of the images was not constant. Therefore, we maintained the original stimulus resolution of 1000×750 pixels and positioned the images centrally within the models' field of view. We fixed the observer distance to 75 cm to match our testing conditions. Given that the pixel pitch (0.294 mm) of the monitor used in their study (ProLite T1931SR-B5) is nearly identical to ours (0.293 mm), the centrally placed images subtend $21^\circ \times 16^\circ$ — *DVA* within the models' field of view.

Fixation Analysis

Participants in the Linka et al. [50] study initiated viewing from the screen center. To ensure a fair comparison, the initial fixation for all baseline models (DeepGaze, other saliency models, and the Random baseline), as well as the participants, is set to the screen center. In contrast, our trained models inherently predict the first fixation toward the center, so their starting point is 10.1° below the center of the field of view, consistent with our primary analyses. For all comparative analyses, these respective initial starting points are treated as the "first fixation" across all models and participants. Consistent with the methodology used for our primary dataset, we evaluate the first four fixations. Furthermore, we compute fixation frequencies—rather than proportions—for both the four original categories from Linka et al. For our categories, we use data from adults only (ages 18-72) and the SU-R Object and SU-I Object categories derived from AutoSUM. We also calculate fixation frequencies at the top-saliency location predicted by the GBVS model. Finally, we compute the cosine similarity across the first four fixations for all models and participants.

Bootstrap Analysis

Error bars for fixations to categories were obtained using 10000 bootstrap resampling of participants. The images were not bootstrapped as doing so would change the number of images in each category. For the AUCs and correlation coefficients, error bars were obtained using 1000 bootstrap resampling of participants and images.

2.4 Results

Foveated Reinforcement Learning Scene Understanding Model (fRL-SU)

We incorporated foveation using an image-based method that applies progressively stronger linear filtering with increasing retinal eccentricity (distance from fixation) and de-blurs the image with subsequent fixations [145]. This might be an incomplete model of foveation, but implementing other methods, such as spatial pooling of the model’s internal feature representations [167], required extensive retraining of the VLM. As fixations accumulate, progressively more of the image is represented at high resolution. Fig. 16a illustrates how the VLM’s description evolves as fixations accumulate. The model’s degradation with retinal eccentricity was calibrated to match that of humans. This was carried out using a separate psychophysical study in which participants (N=20) viewed 277 images on a gaze-contingent display and described each image after two or four fixations. The model’s foveation was adjusted so that, when it followed the measured fixations of individual humans, the accuracy of its generated scene descriptions matched that of humans (see Materials and Methods). Once the foveation parameters were set for the separate psychophysical dataset, we trained the fRL-SU model using a CNN (four 1x1 convolutional layers with ReLU activations and a softmax output). The CNN received VLM features extracted from an image foveated at

all previous fixation locations and output a probability distribution over the next fixation location that maximized scene-description accuracy. Fig. 16b depicts a detailed flow chart of the training process. Training used 24,000 images split evenly across MS-COCO [155], MPII Human Pose [153], and FindingEmo [154]. fRL-SU was then tested on the 147 test images (not seen during RL training), generating four fixations following an initial fixation at the bottom center, with ten descriptions generated after each fixation reflecting the cumulatively deblurred image up to that point (see Materials and Methods). People were present in 85 of those images, and the remaining 62 had objects and no people. The images depicted human actions, social situations, or implied actions based on the placement of objects within the scene. To dissociate low-level saliency from scene-relevant content, we selected images that contained salient regions distant from the image center and from objects or regions critical for scene understanding (see Materials and Methods). To ensure reproducibility, we trained five independent instances of the fRL-SU model.

Human and model comparison of fixation frequencies

We compared the fRL-SU fixations to those of 50 participants (different from those in the foveation calibration experiment) who viewed the 147 scenes for 2 s each, with no specific instructions other than to view them freely (see Fig. 17a). We also compared the fRL-SU model to saliency models (GBVS and Itti-Koch) and with fRL models trained to maximize accuracy on alternative tasks. The fRL-Scene Classification model was trained to execute eye movements that maximize accuracy in superordinate scene classification (e.g., semantic content, 3D spatial structure, and 2D image appearance [152]). The fRL-Search model was trained to execute eye movements to maximize accuracy in identifying the smallest object in the scene.

Regions within the images in our dataset were segmented as shown in Fig. 17b and categorized into scene elements: people, text, and objects. We quantified the frequency of fixations on these scene elements. Fixations were counted to a scene ele-

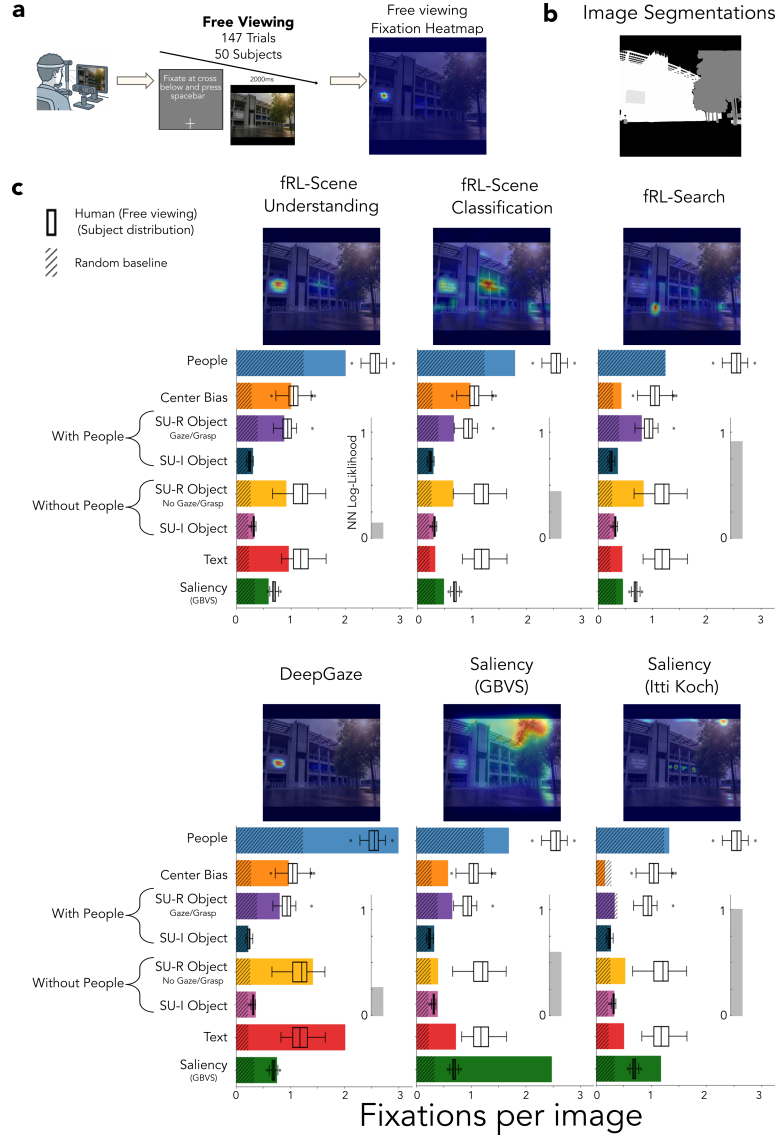


Figure 17: Fixation frequencies across scene-element categories. (a) Experimental flow for the free-viewing eye-tracking study. Fifty participants freely viewed 147 images for 2 s. The human icon on the left is AI-generated. (b) Example image segmentation (using SAM3) used to define scene-element categories for quantifying fixation frequencies. (c) Fixation frequencies across scene-element categories for participants (box plots) and models (bar plots). Hatched bars show a random-fixation baseline reflecting category size in the image. Heatmaps above each panel show example model predictions on the scene. The model predictions for these categories were compared with human free-viewing fixation distributions using the Negative Normalized Log Likelihood (NNLL), normalized to the distance observed between the Itti-Koch saliency model and human fixation distributions. The fRL-Scene Understanding (fRL-SU) model matches human data significantly better than fRL models trained on either search or scene classification. Furthermore, the fRL-SU model significantly outperforms DeepGaze and traditional saliency models.

ment if it fell within 0.7 deg of the segmentation edge to match the spatial sampling of the fRL model’s action space (see Fig. 18 for generalization of results for different spatial thresholds). Based on prior studies [168], center bias was defined as fixations falling within a 2.5 deg radius circle centered on the image (see Fig. 19 for generalization of results for different definitions of center bias). Objects were classified as scene understanding-relevant (SU-R object) or scene understanding-irrelevant (SU-I Object). SU-R Object was defined as an object that, when digitally removed from the image, maximally changed the VLM’s scene description. Scene-understanding-irrelevant objects (SU-I objects) were defined as objects that, when removed from the scene, did not impact the scene description. Classification of SU-R and SU-I objects was based on automatically segmenting objects, placing masks, deleting them, and assessing the impact of the object removal on scene description accuracy using the VLM (see Materials and Methods, and validation of method in [165]). Fixations for objects were quantified separately for scenes with and without people. In scenes containing people, the SU-R objects were typically gazed at or grasped by a person in the scene. SU-R objects in scenes without people did not include gaze or grasp cues.

Fig. 17c shows example fixation heatmaps of each model on the sample image from Fig. 16a. The heatmap shows how the fRL-SU and DeepGaze models fixate on the text sign on the left, matching human participants’ fixations. In contrast, the other models do so to a lesser extent (RL-SC and GBVS) or not at all (RL-Search and Itti-Koch saliency). Additional fixation heatmaps for each scene-element category across all models are shown in Figure 20. Our results showed that, averaged across images, the human fixations exhibited frequency patterns similar to those observed in prior studies. People were significantly fixated more often than any other category ($p < 0.0001$ for all comparisons). Objects critical to scene understanding were fixated more often than both unimportant objects and salient regions defined by GBVS

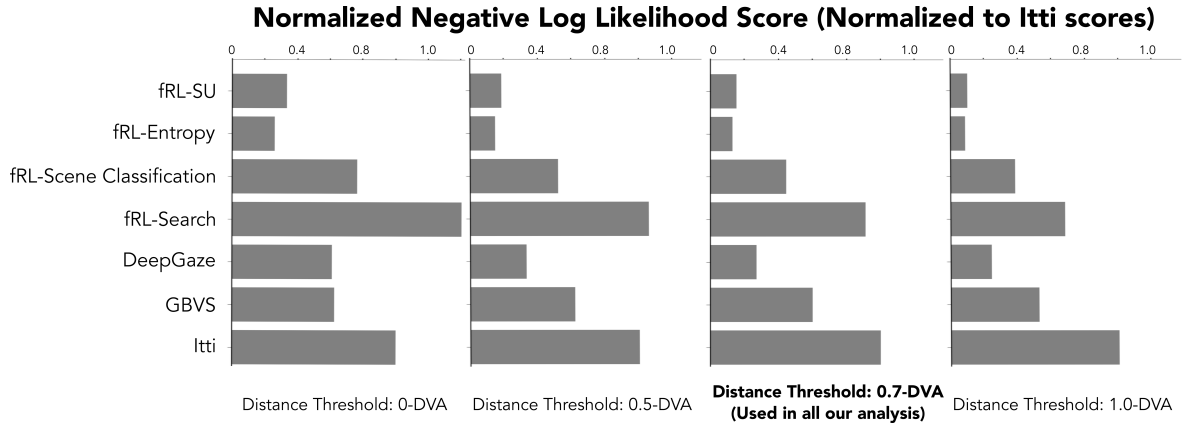


Figure 18: Model-human fixation frequency agreement across segmentation distance thresholds. Normalized Negative Log Likelihood (NNLL) scores, normalized to Itti-Koch Saliency scores, quantifying how closely each model’s fixation frequencies match those of human participants. Lower values indicate closer agreement. Each panel corresponds to a different distance threshold used to assign a fixation to a scene element. Across all thresholds, fRL-SU and fRL-Entropy show the closest agreement with human data, while fRL-Search and Itti-Koch show the largest deviations. The relative ordering of the models is the same regardless of the distance threshold, indicating that the main findings are not an artifact of the 0.7 DVA criterion.

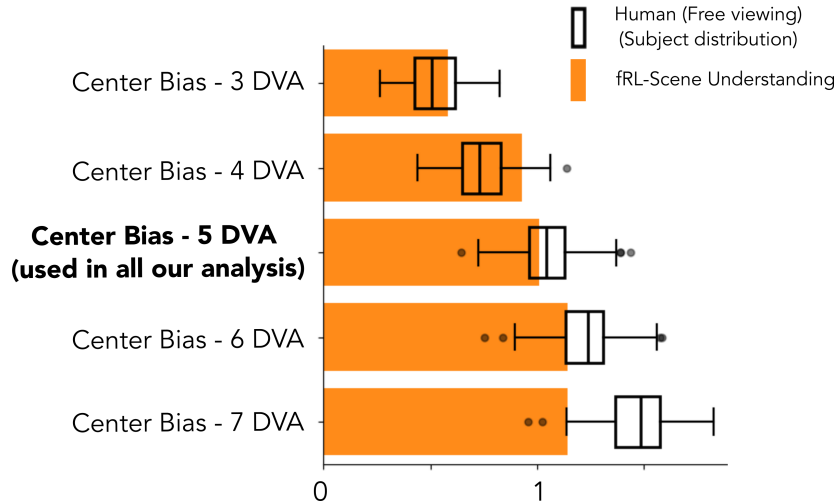


Figure 19: Center bias fixation frequency across various degrees of visual angle (DVA). Fixation frequency within the central region of the image for human free-viewing participants (box plots) and the fRL-SU model (orange bars), computed using circular central regions of 3,4,5,6 and 7 DVA. The DVA size of 5, used in all main analyses, was selected based on prior human fixation data. At this size, fRL-SU central fixation frequency falls within the human subject distribution.

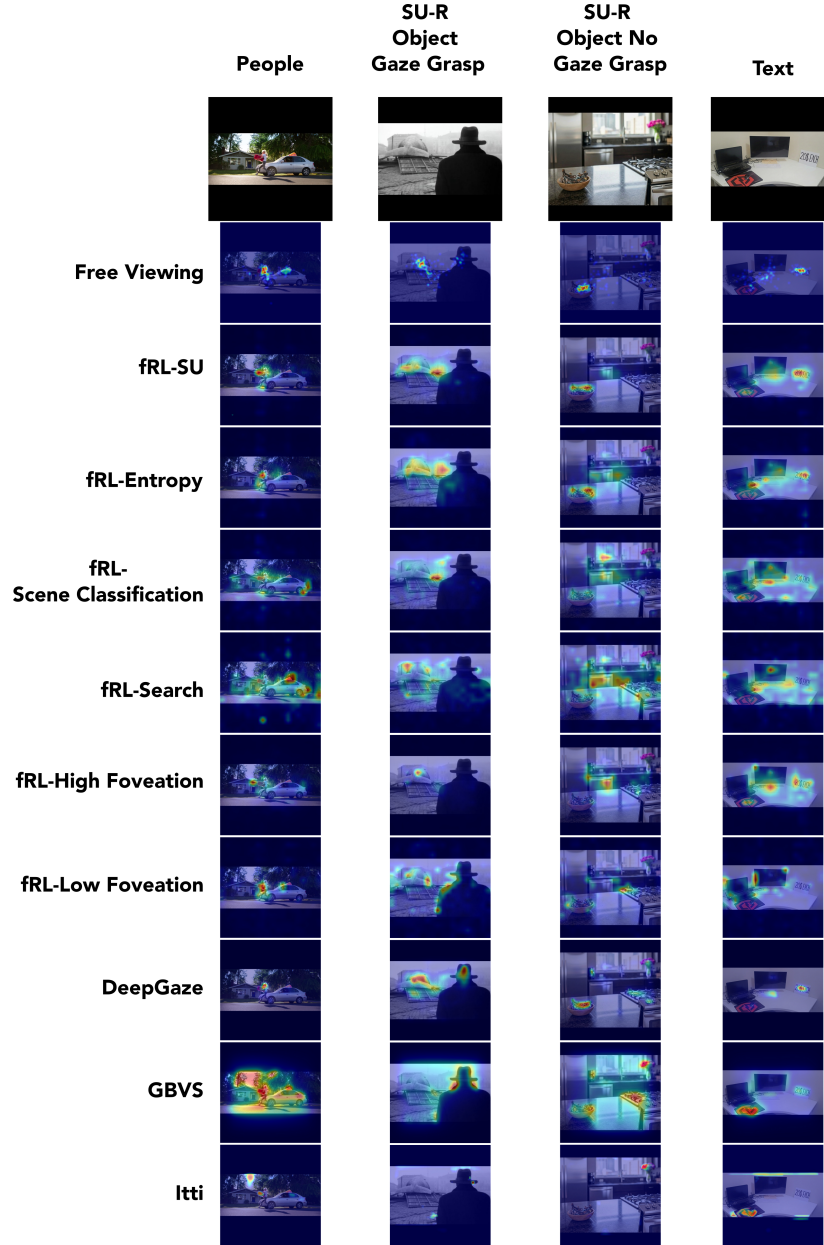


Figure 20: Fixation heatmap examples across models and scene element categories. Example images from four scene element categories are shown in columns: scenes with people, scenes with SU-relevant objects gazed or grasped, scenes with SU-relevant objects without gaze or grasp, and scenes containing text. Rows show fixation heatmaps generated by human free-viewing participants and each computational model.

($p < 0.0001$, for both comparisons). Regions with text were fixated more often than salient regions ($p < 0.0001$, all the comparisons above have FDR-adjusted p-values). The fRL-SU fixation frequencies closely matched those of human participants across all scene-element categories (Fig. 17c). The model primarily fixated on people (Supplementary Video S1), prioritizing objects they are looking at or holding over those they are not interacting with (Supplementary Video S2). In scenes without people, grasp or gaze cues, the fRL-SU model still fixated more on SU-R objects than SU-I objects (Supplementary Video S3). The model also captured the large proportion of text fixations (Supplementary Video S4). In addition, first fixations were biased towards the image center, exceeding the frequency expected under the random-fixation baseline. The largest human-fRL-SU discrepancy was the model’s lower frequency of fixations on people relative to participants (see Discussion). In contrast to the fRL-SU, the saliency models generated fixation frequencies that differed from the participants’ fixation frequencies. As expected, the saliency models primarily fixated on low-level salient regions and underpredicted fixation frequencies across most categories, except for SU-I objects. DeepGaze overpredicted human fixation frequencies on people, text, and SU-R objects. Fig. 21 shows the fixation distributions for a model with randomly initialized weights, demonstrating that the CNN does not possess any inherent prior knowledge regarding which locations to fixate for a given task. As a control, we also trained a model on visual features extracted directly from the VLM’s visual backbone, rather than on features obtained after cross-attention with the input text. This excludes any textual information from the RL models’ input during training, making the semantic accuracy reward the only context available for the model to learn optimal fixations for improved scene understanding. Finally, Fig. 22 illustrates that this control model exhibits similar fixation patterns and is not different from the fRL-SU model ($p = 0.43$)

Fig. 17c also shows that the fRL-Search and fRL-SC models (control models) generated patterns closer to those of humans than the saliency models, but to a lesser

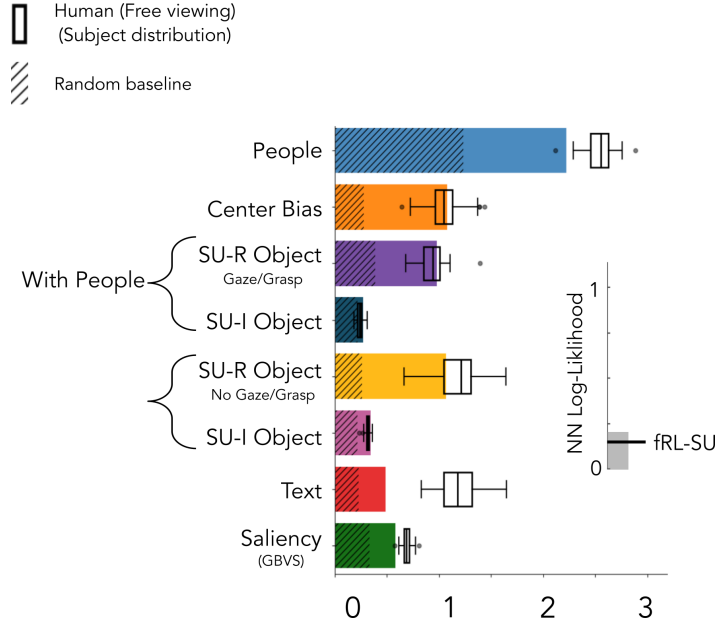


Figure 21: Fixation patterns observed for a model trained on visual features extracted from the VLM’s visual backbone. This configuration ensures that the RL model lacks access to the task’s context in its input, acquiring that information solely through reward signals from the semantic accuracy of its foveated descriptions. This model aligns well with human data and performs similarly to the fRL-SU model.

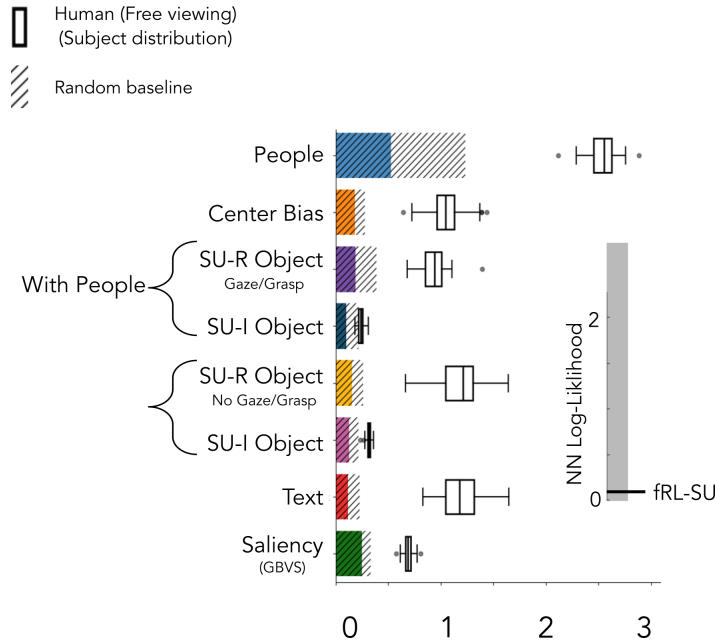


Figure 22: Fixation patterns observed for a model with randomly initialized weights. The model with random weights has no information about the task or the visual features to fixate on, and hence performs very poorly at matching human data.

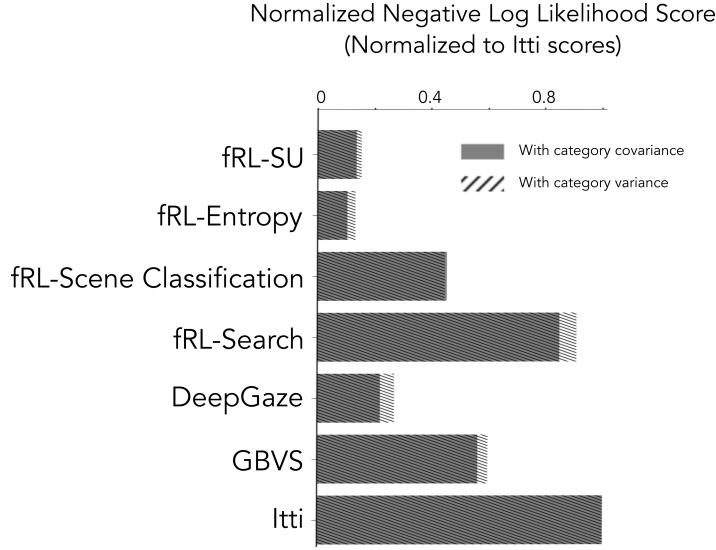


Figure 23: Category independent vs category dependent NNLL comparisons. NNLL scores, normalized to Itti-Koch, comparing each model’s fixation frequencies to human data. Solid bars show the NNLL computed with category covariance, treating categories as dependent. Hatched bars show the NNLL computed using only category-specific variances, treating categories as independent. The values are similar, indicating that dependencies within the categories minimally affect the NNLL scores.

degree than the fRL-SU model. The control RL models showed reduced fixations on text, SU-R objects (those not gazed at or grasped), and people. To quantify the distance between the models and human fixation frequencies (excluding those to the most salient objects to avoid over-penalizing GBVS), we calculated a negative normalized log-likelihood (NNLL). A lower number reflects better agreement between model and human data. The fRL-SU were closer to participant fixation frequencies than the fRL-SC, fRL-Search ($p < 0.0001$ vs. fRL-SC; $p < 0.0001$ vs. fRL-search), saliency models ($p < 0.0001$ vs. GBVS; $p < 0.0001$ vs. Itti-Koch), and DeepGaze ($p < 0.0001$) due to the higher fixation frequencies than participants to categories like people, SU-R objects, and text. The main analysis assumed that the categories were independent while computing the NNLL scores. The NNLL scores were similar even when dependencies between categories were taken into account (Fig. 23)

Model	AUC	CC	p-value
fRL-SU	0.9046 ± 0.0038	0.3715 ± 0.0121	-
fRL-Scene Classification	0.8707 ± 0.0049	0.2785 ± 0.0090	***
fRL-Search	0.8088 ± 0.0061	0.1832 ± 0.0097	***
DeepGaze	0.9312 ± 0.0036	0.6281 ± 0.0164	***
GBVS	0.8451 ± 0.0051	0.2382 ± 0.0086	***
Itti Koch	0.5329 ± 0.0048	0.0596 ± 0.0093	***

Table 1: AUROC and correlational coefficient (CC) metrics comparison of different models. fRL-SU is significantly better than other models in comparison, except DeepGaze. DeepGaze is the best-performing model across both metrics.

We complemented our fixation frequency analyses by using other metrics, such as AUROC and heatmap correlations. In both AUROC and heatmap correlations, DeepGaze performed best, followed by fRL-SU ($p < 0.001$). fRL-SU significantly outperformed fRL-SC, fRL-Search, and the saliency models on both metrics ($p < 0.001$ for all comparisons, Table 1).

Semantic Accuracy of VLM with Human and Model Fixation Selection

Theoretically, the fRL-SU should converge on the optimal eye-movement strategy that maximizes the VLM’s scene-description accuracy. If our implementation was successful, then exploring the scene with other fixation strategies for that level of foveation should result in lower accuracy of VLM scene description. In addition, if humans’ free-viewing eye movements are near-optimal, then using human fixations as input to the VLM should yield scene accuracy comparable to that of the fRL-SU (Fig. 24a). In contrast, saliency-based fixations should result in diminished accuracy in VLM scene description. Finally, Deep Gaze fixations should yield VLM accuracy comparable to humans’. To evaluate these predictions, we generated VLM scene descriptions for the 147 test images

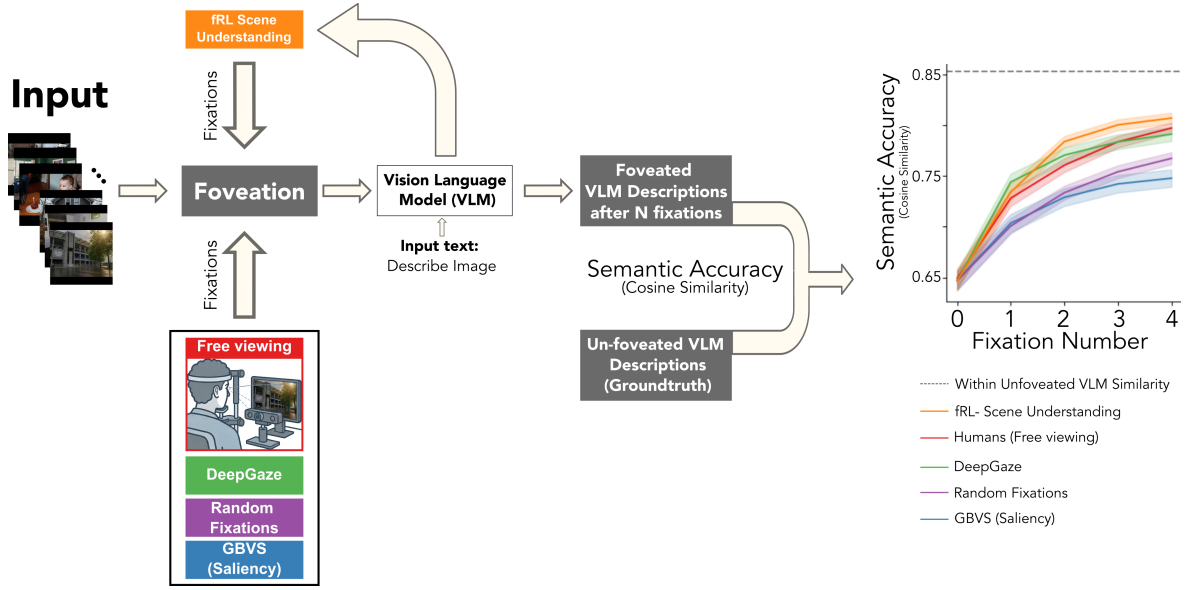


Figure 24: Semantic accuracy flowchart and results. Images are foveated using fixations from the fRL-SU model, DeepGaze, a saliency model (GBVS), a random baseline, and human free-viewing data. After each fixation, these foveated images are sequentially processed by a Vision-Language Model (VLM), and their semantic accuracy is computed against the unfoveated ground-truth description. Descriptions derived from fRL-SU, DeepGaze, and human fixations yield comparable semantic accuracies, consistently outperforming those based on traditional saliency or random fixations.

using the fixation predictions of the various models and those measured for our 50 participants. This entailed foveating the image using different model fixation sequences, inputting the foveated image after each fixation to the VLM, obtaining a description, and evaluating semantic accuracy as a function of fixation number. Fig. 24b shows that the semantic accuracies for the VLM with the various model/human fixation strategies are in line with our predictions.

The Influence of Foveation in fRL-SU Fixation Predictions

We hypothesized that the fRL-SU’s close correspondence to human fixation frequencies is an emergent property arising from an interaction between optimizing scene understanding, but also the foveated nature of the human visual system. This would predict that the agreement between the fRL-SU model and human fixation frequencies might be lower if the fRL-SU model had much lower foveation or more extreme foveation

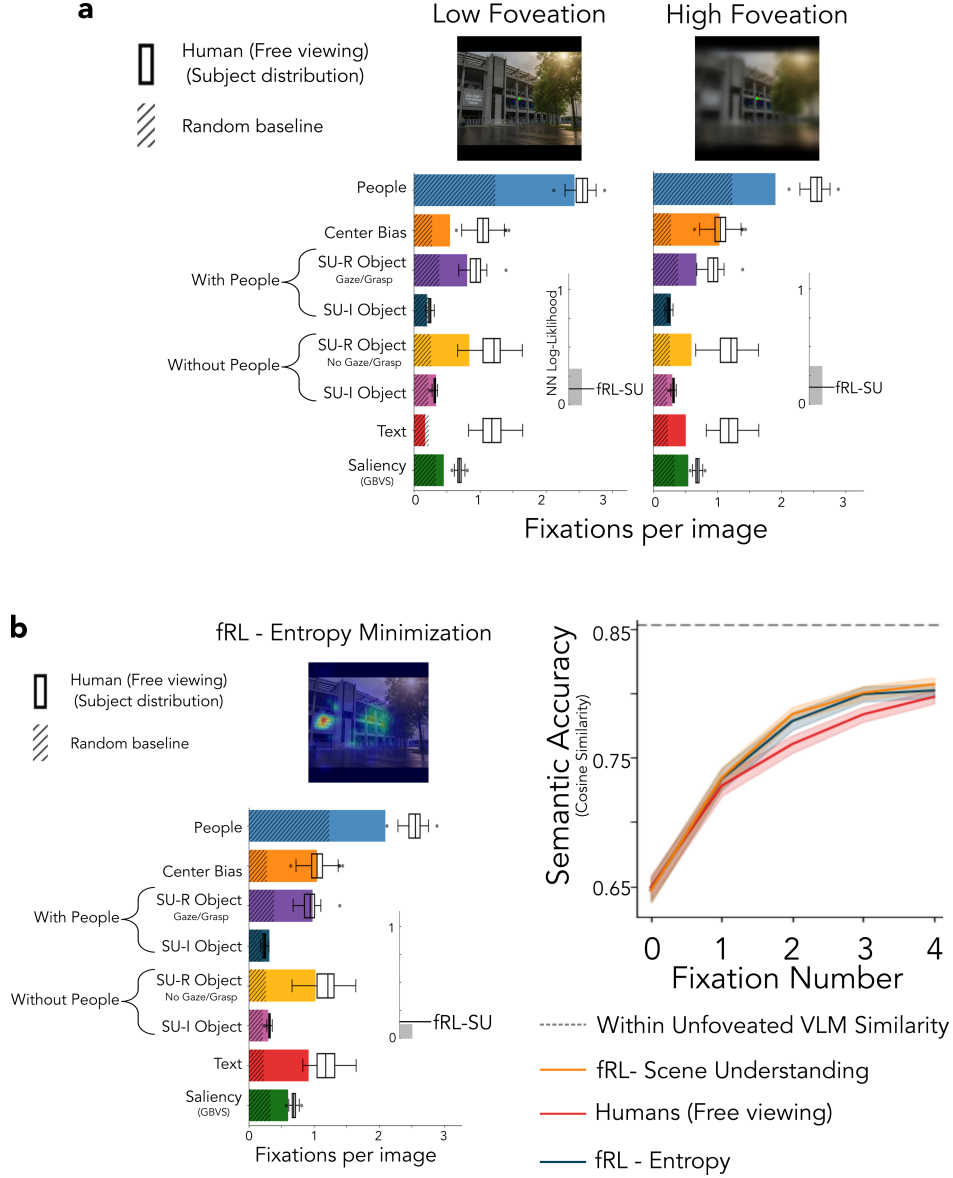


Figure 25: Effect of foveation and fRL-Entropy model results. (a) Models trained with either extremely low or high foveation (examples shown) perform significantly worse at matching human fixation data compared to the standard fRL-SU model. (b) Conversely, a model trained to minimize entropy (heatmap shown) matches human data significantly better than the fRL-SU model (left). Despite these differences in predicting human fixations, the semantic accuracy of descriptions generated using the fRL-Entropy model remains similar to that of the fRL-SU model (right).

than humans. To assess this prediction, we re-trained the model with extremely low ($\alpha = 20$) and high ($\alpha = 0.2$) foveation. Fig. 25a shows that an fRL-SU model with foveation much lower or higher than that of humans yields lower agreement with human fixation frequencies. In particular, fixations on text and objects have diminished,

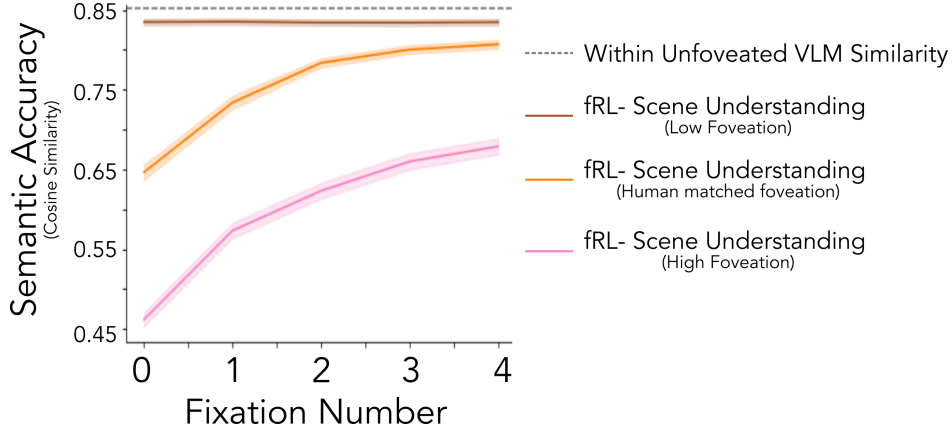


Figure 26: Semantic accuracy of fRL-SU models across varying foveation conditions. The low-foveation model exhibits a flat performance curve, indicating that semantic accuracy is insensitive to foveation at this level. Conversely, the high-foveation model yields lower accuracy scores compared to the baseline fRL-SU model trained with human-matched foveation.

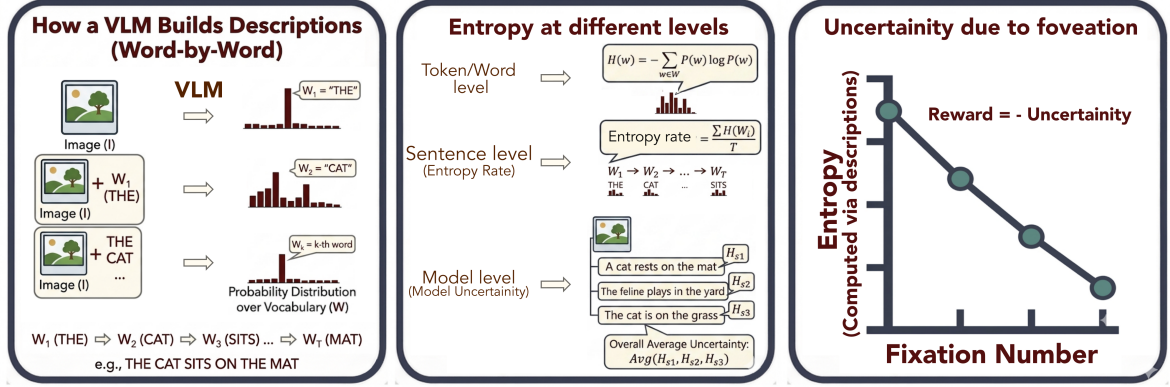
suggesting the importance of the constraints induced by foveation in realizing the patterns observed in human eye movements. Fig. 26 shows semantic accuracy graphs for the first 4 fixations made by the low foveation, high foveation, and fRL-SU model. The low foveation has a high flat line graph, implying that foveation has no effect, while the high foveation graph is much lower compared to the fRL-SU model.

No Ground Truth Training: Foveated Reinforcement Learning with Entropy Minimization

Our approach relied on ground-truth scene descriptions (no foveation) to train the fRL-SU model. Such a training scheme might parallel some real-world situations. A parent verbally foreshadowing what might be happening in a picture book or a real-world scene provides ground truth before children explore the scene through eye movements. However, access to ground-truth descriptions during training is uncommon and might be deemed to reflect the experience of young children unrealistically. Thus, can an agent with foveated vision learn to exhibit human-like eye movements without access to ground-truth descriptions? We evaluated this possibility by modifying the fRL-SU

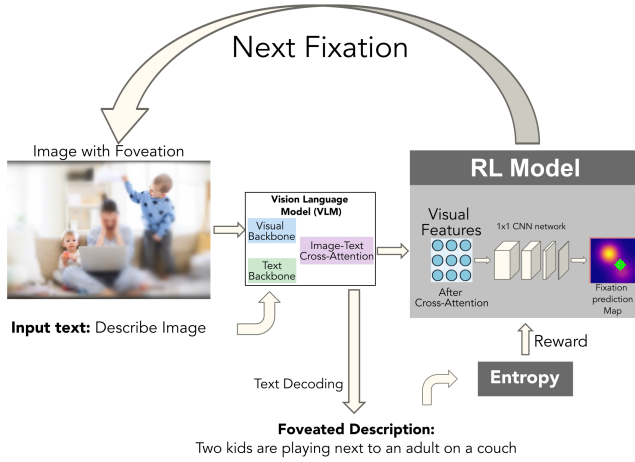
a

Entropy as a reward function



b

fRL Entropy Model



c

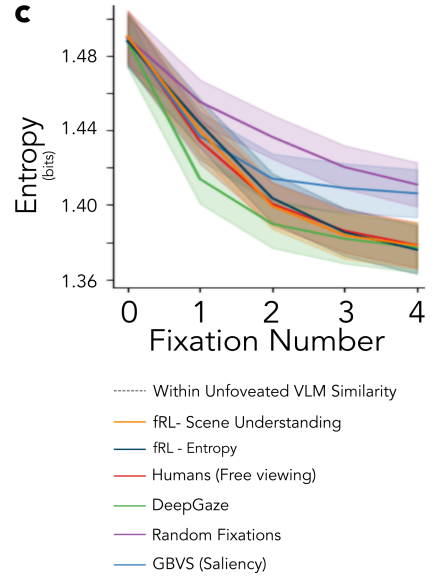


Figure 27: fRL-Entropy model architecture and training pipeline. (A) Entropy as a reward signal. A VLM generates descriptions word-by-word, producing a probability distribution at each word. Token-level entropy is averaged across tokens to obtain a sentence-level entropy score, then averaged across ten candidate descriptions to estimate overall uncertainty. Reward is defined as the negative of this uncertainty. (B) Schematic of the fRL-Entropy model, which shares the fRL-SU architecture but replaces the cosine-similarity reward with the entropy-based reward, removing the need for ground-truth descriptions during training. (C) Average description entropy as a function of fixation number across models.

loss function to minimize the entropy across ten high probability (but semantically unique) candidate VLM descriptions (see methods and Fig. 27). VLMs essentially output probability distributions over the next word (or word chunk) conditioned on the input and previously predicted tokens. We leverage this to compute the entropy rate of a description by averaging the conditional entropy of its individual tokens. After each fixation, we obtain multiple descriptions of the foveated image from the VLM and estimate the VLM model’s uncertainty by averaging the entropy rate across these descriptions (see methods). Thus, the model (fRL-Entropy or fRL-E) learned to make eye movements to minimize this uncertainty without access to the ground truth (unfoveated descriptions). Fig. 25b shows the fixation heatmap for our sample image fRL-Entropy. The model achieves slightly better agreement with human fixation frequencies than the fRL-SU model ($p = 0.001$).

Generalization to other images and eye movement datasets

To assess the generalization of our model and findings to other images and eye movement data sets, we assessed the fRL-SU model fixation frequencies for a recently published data set of free viewing fixations of 6,720 human observers with ages ranging from 5 to 72, to a set of 40 images) intended to study how age influences free-viewing fixation. The fRL-SU model (with no new training or fitting parameters) was evaluated by matching the upper and lower subtended angles of the images reported by the study. We first compared model and human fixation frequencies. The fRL-SU model captured the relative fixation frequency of the participants using the categories in the study: people, grasped (touched) objects, and text (Fig. 29a). We then analyzed fixation frequencies, including additional categories: SU-relevant objects, SU-irrelevant objects, and the most salient region (based on GBVS, Fig. 29b). The fRL-SU model’s agreement with human fixation was similar to that of our eye movement data set and resulted in better agreement than other control models (see Fig. 28 materials).

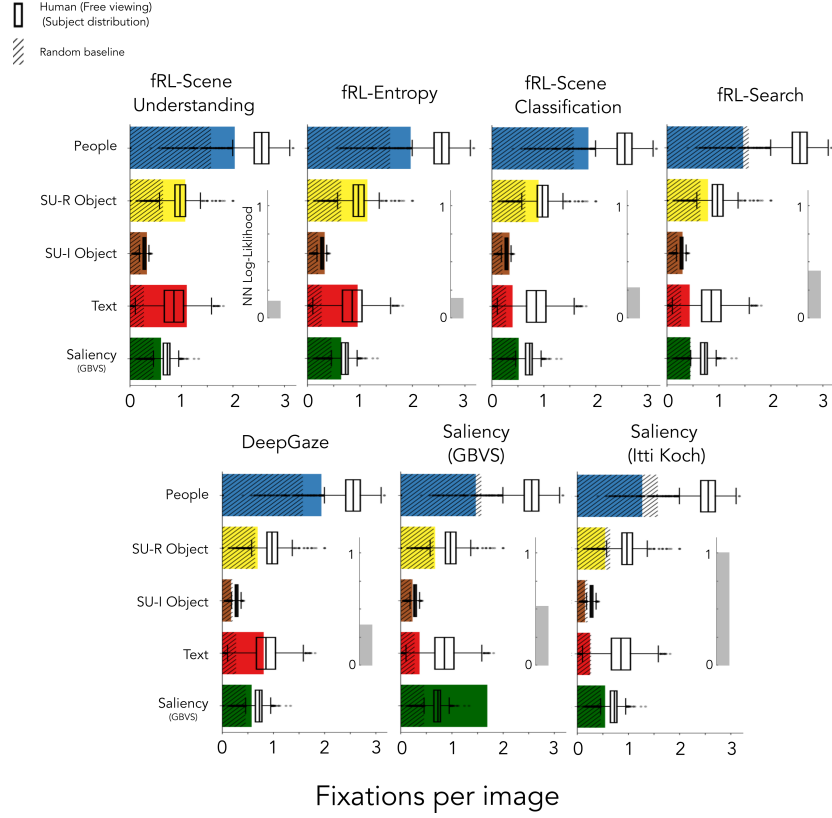


Figure 28: Generalization to the Linka et al. [50] dataset. (a) Fixation frequencies per image across scene element categories for each model compared to the De Haas et al. free-viewing dataset. Box plots show the human subject distribution; the bars show model predictions; hatched bars show the random baseline. (b) Semantic accuracy as a function of fixation number across models applied to the De Haas images.

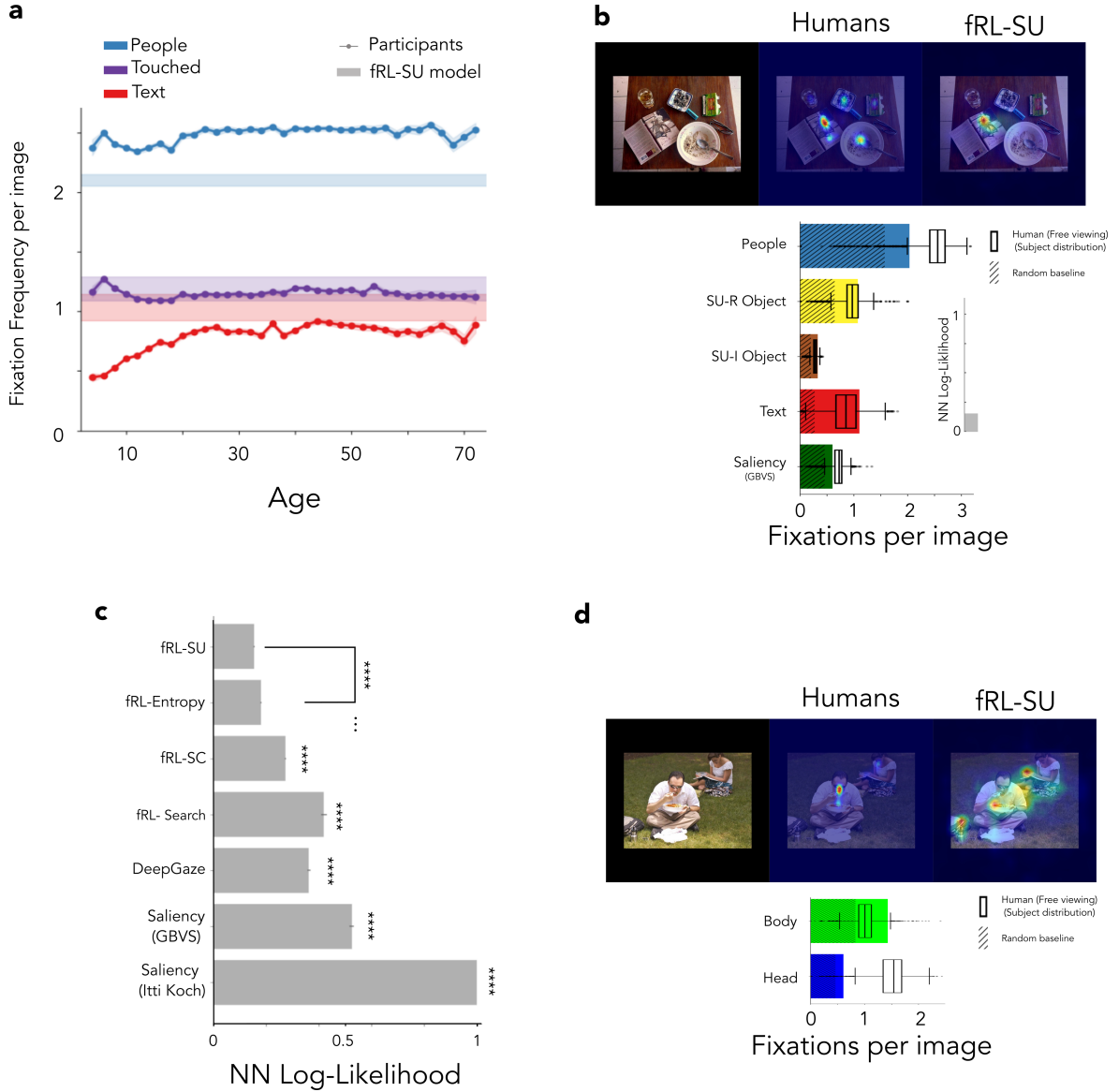


Figure 29: fRL-SU model results on a publicly available free-viewing dataset [50] (a) The dataset contains human free-viewing data across different ages, 5-72 years ($N=6,720$). Solid lines show human fixation frequencies across ages for three image categories: people, touched objects, and text. Dashed lines indicate fRL-SU model predictions. (b) Applying our category definitions to this dataset yields similar fixation patterns. The fRL-SU model successfully matches human data across these categories (measured and predicted heatmaps shown). (c) The NN-Loglikelihood score shows that the fRL-SU model best matches human data, consistent with the results observed in our dataset (d). However, the model exhibits mismatched predictions regarding fixations on specific body parts. While the model correctly prioritizes people present in the scene, it focuses disproportionately on the body, whereas humans typically fixate on the head (example heatmaps shown).

Comparison of fRL-SU to a know-it-all model, The Oracle

To establish an upper bound on scene understanding, we evaluated a model that fixates at every possible location within the visual field and selects the fixation sequence that maximizes semantic accuracy. We refer to this model as the Oracle model. For each subsequent fixation, we repeated this process greedily, always choosing the location that yielded the highest semantic accuracy given prior fixations. Unlike the fRL models we implemented, the Oracle is not limited by its peripheral processing when finding the semantic-accuracy-maximizing fixation points. It represents an upper-bound performance for an fRL-model in which fixation selection is limited by peripheral processing.

As shown in Fig. 30a, the Oracle model’s fixation patterns bear little resemblance to those of human participants, and the match is significantly worse than that of fRL-SU ($p < 0.0001$). Nevertheless, the semantic accuracy of descriptions derived from Oracle fixations exceeds that of both fRL-SU and human participants (see Fig. 30b).

This dissociation is informative: while the Oracle model achieves superior semantic accuracy, it does so with the benefit of full knowledge of each image. For some images, the optimal fixation location may lie between multiple critical objects, at a position that captures information from several regions simultaneously. The Oracle model can identify such locations by exhaustively evaluating all possible fixations. Humans and fRL-SU, having never seen the image before, lack this information and therefore cannot reliably land on these optimal intermediate locations, which explains their comparatively lower semantic accuracy.

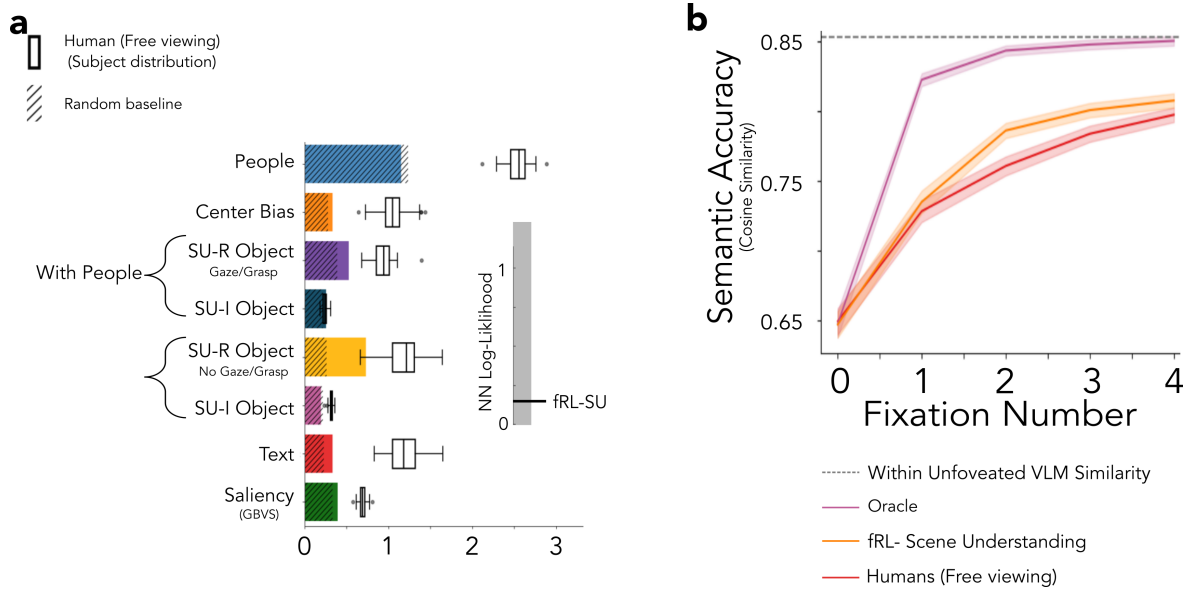


Figure 30: Performance of Oracle model on the test dataset (a) The fixations of the oracle model match poorly with the human fixation patterns, significantly worse than the fRL-SU model. (b) The semantic accuracy of the descriptions based on the Oracle model fixations is highest for images in the test dataset, exceeding that observed for fRL-SU or participants’ fixations.

2.5 Discussion

Why humans look at scenes the way they do during free-viewing and what they aim to accomplish remain poorly understood. Here, we addressed these questions by developing a computational agent with simulated human foveated vision that learns optimal (accuracy-maximizing) eye movements for different visual tasks on real-world scenes. The model optimized for scene understanding produced emergent fixation patterns that matched those of humans: increased fixation frequencies on people, text, scene-relevant objects (relative to scene-irrelevant objects), gazed or grasped objects, and even the initial center bias. This pattern held in both our own data (50 participants viewing 147 images) and a large, previously published dataset (6720 participants and 40 images). Training the model to optimize different tasks (e.g., searching for the smallest object or superordinate-level scene classification) resulted in worse predictions of human fixation frequencies. Training fRL-SU with very low or very high foveation also decreased

agreement with humans, specifically reducing the model’s fixations on text and scene-relevant objects. Notably, reducing the model’s foveation also weakened the initial center bias, suggesting that this prevalent human behavior is not solely a byproduct of photographer bias (placing content at image centers) [45, 169] but also reflects a near-optimal strategy given foveated vision: an initial central fixation minimizes the distance of all image regions to central vision. Together, these results suggest that human free-viewing fixations emerge from the interaction between optimization for scene understanding and the specific properties of the human foveated visual system.

The fRL-SU model outperformed DeepGaze in predicting human fixation frequencies. This finding might seem puzzling, since DeepGaze is trained on human fixations and would, by design, be expected to provide the best predictions. One explanation is that DeepGaze is trained on fixation maps accumulated over 3 to 5 seconds of viewing (about 10 to 12 fixations), whereas our analyses focus on only the first four fixations. Early human fixations are more strongly constrained by peripheral visibility and foveation than later fixations, which may explain why a model built on foveation principles (fRL-SU) better predicts early fixations than a model trained on later- or longer-time-course data (DeepGaze). This seems qualitatively consistent with the finding that DeepGaze fixates on text more frequently than humans and the fRL-SU in their first four fixations. Furthermore, when evaluating the precise location of the human fixations using AUROC and heat maps, DeepGaze outperformed the fRL-SU model, as expected.

An important finding is that an fRL model that learns eye movements with a reward function that minimizes sentence entropy also resulted in emergent human-like fixations. This might provide a way for humans to learn optimal fixations for scene comprehension without a priori access to the ground truth before eye-movement exploration.

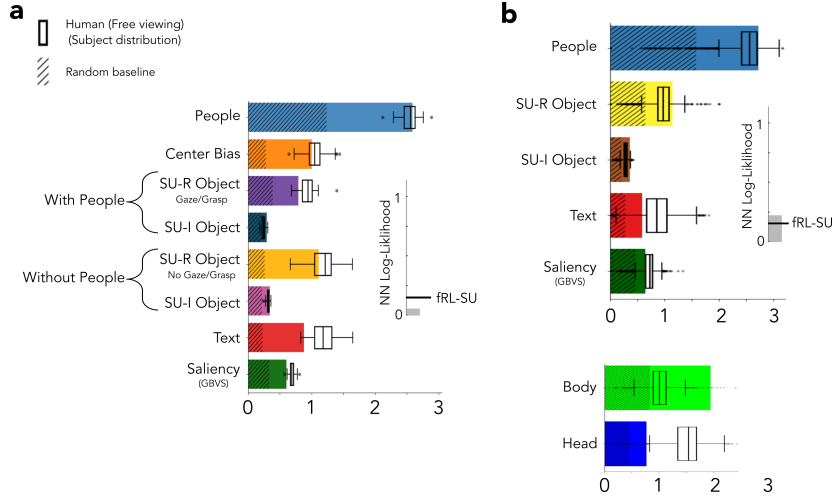


Figure 31: Model results for detailed prompt. The fRL-SU model trained on a prompt asking the model to provide the age, gender, and emotional expressions of people while describing the scene increased the fixations to people in both (a) our dataset and (b) the OSIE short dataset from [50]. However, the model still looks at bodies disproportionately more than faces.

The fRL-SU or fRL-Entropy results should not be interpreted as suggesting that language is needed to learn eye movements that optimize scene comprehension. That would imply that humans who never acquire language would be unable to learn to optimize eye movements for scene comprehension, which is unlikely. Language is a convenient representation of scene meaning for training, but visual representations aligned with language’s semantic space should yield similar emergent fixation patterns (except for text, which inherently requires a language representation).

One limitation of our modeling effort is that the fRL-SU model resulted in fewer fixations on people than human observers for both our dataset and that of Linka et al. [50]. Furthermore, a breakdown of fixations to bodies and heads shows a dissociation. The fRL-SU model fixated on bodies more often, while humans fixated on faces more frequently (see Fig. 29d). This suggests that humans might not be solely optimizing scene comprehension when free-viewing. There might be a complementary task to

identify people, their gender, age, and affective state. When the fRL model was implemented with two tasks: scene understanding and identifying a person’s age, emotion, and gender, it fixated on people more frequently, yielding results comparable to human gaze patterns (see Fig. 31). However, even that model did not attain the high fixation frequency on faces that humans showed. Two factors contribute to this limitation. One, the VLMs we used are compact, can be implemented in our local GPU cluster to speed up the RL training, but do not reach human understanding of the real world [170, 171], including not achieving human-level face-processing performance. Second, our model relies on a simplified linear filtering approach for foveation. As a result, it does not account for the non-linear effects of crowding, which selectively degrade the peripheral processing of text and faces [172] and might influence the learned eye movement strategy.

In conclusion, our results indicate that eye movements during free-viewing of scenes reflect a task-driven strategy shaped by the goal of comprehending the visual world. The characteristic pattern of fixations on people, text, scene-relevant objects, and gazed or grasped objects emerges from the interaction between scene-understanding optimization and the foveated architecture of the human visual system.

Conclusions and Future Work

The overall goal of this dissertation is to decode the free viewing paradigm and provide solid evidence on the factors that influence eye movements during free viewing. To this end, I have developed novel datasets, measured human eye movements, and built reinforcement learning agents to model free-viewing eye movements.

Chapter one compared human free-viewing eye movements with those from other tasks. Free-viewing eye-movement patterns closely resemble those of observers explicitly instructed to describe a scene. This implies that people default to understanding scenes when they free-view. To understand where people look during free viewing, we designed stimuli that dissociate semantic and saliency factors. We created Winograd images: pairs that are visually similar, maintaining nearly identical low-level saliency, but yield drastically different human descriptions, making them semantically distinct. During free viewing, humans look at objects/people critical to scene understanding, rather than solely salient or meaningful objects. I created scene-understanding maps (SUMs) to evaluate the criticality of objects/people in scene understanding by quantifying the impact of their removal on human scene descriptions. To determine whether eye movements during free viewing play an active role in scene understanding, we conducted a forced-fixation study, asking participants to fixate at locations that were either critical or noncritical to scene understanding. Participants provided descriptions closer to ground-truth human descriptions when fixating on a critical region than on non-critical regions. This implies that humans looking at critical regions of scenes is the result of active exploration to better understand the scene, rather than solely looking at salient or locally meaningful regions.

Some of the limitations of these studies are their inherent focus on Winograd pairs,

in which all scenes feature one or more people or a person’s hand, directly implying an action or social interaction. While such images are fundamental to daily life and often require active eye movement exploration to resolve critical object dependencies, they do not represent all visual environments. Chapter two expands its analysis using a much larger dataset, showing that participants look at important elements in a scene even in the absence of humans. But it is still possible to have scenes with diffusely distributed objects without implied actions or social interactions, and without a single critical object. In these simpler images, peripheral visual processing might enable rapid scene comprehension without extensive eye-movement exploration [131, 132, 133].

Chapter two showed that the signature human-like fixation patterns during free viewing are an emergent behavior of a near-optimal eye movement strategy aimed at comprehending scenes. I developed a foveated reinforcement learning model for scene understanding (fRL-SU) that combines a compact visual language model (VLM), simulated foveation, and reinforcement learning to execute eye movements that maximize the semantic accuracy of VLM-generated scene descriptions. The fRL-SU model produces fixation frequencies across various scene elements that closely mirror human free-viewing fixation patterns. In contrast, models trained to maximize accuracy on alternative tasks—such as classifying the scene or identifying the smallest object—failed to predict human fixations accurately, as did traditional low-level saliency models. The VLM descriptions generated by simulating foveation at human-fixated locations achieved semantic accuracy comparable to that obtained via fixations from the fRL-SU model. Furthermore, implementing foveation levels significantly lower or higher than human-level foveation resulted in diminished agreement with human fixation frequencies. These results suggest that human scene fixations are fundamentally a consequence of an interaction between the constraints of the human foveated visual system and the drive to optimize scene comprehension. Finally, I show that an agent can learn to exhibit these human-like eye movements even without access to ground-truth training

descriptions by minimizing the entropy (uncertainty) of its generated descriptions.

Together, this research demonstrates that eye movements during free viewing are not merely directed to salient or locally meaningful regions in a stimulus, but are an actively optimized, task-driven behavior fundamentally aimed at comprehending the visual world.

The specific goals and patterns of human eye fixations during free viewing are still not fully understood. However, our model offers a new perspective. Optimizing it to process scenes using simulated human foveated vision sheds light on the underlying cognitive functions involved in how people naturally observe real-world environments. Because of this, the model has the potential to serve as a biomarker, helping to establish a baseline for normal cognitive development.

Furthermore, the model can be manually guided to fixate on the exact locations where a human observer looked, making it a valuable tool for studying cognitive deficits. To do this, we can track the eye movements of individuals with cognitive deficits and force the model to follow those exact fixation paths. The model then generates a description of the scene based entirely on those specific locations. By comparing the semantic accuracy of these descriptions against a neurotypical baseline, we can quantify exactly how well the scene was comprehended.

Beyond free viewing, the model is highly adaptable. The ability to train it on different tasks using specific prompts vastly expands its potential applications. For instance, specialized versions could be developed to study human eye movement behaviors during active tasks, such as searching for an object, answering questions, or reading text. With recent technical advancements in virtual reality and the commercial availability of various VR devices, this model could potentially be integrated with VR

platforms to move beyond static stimuli. This integration would allow us to understand the influence of complex, dynamic factors such as retinal flow (motion) [173] on human eye-movement behavior during active tasks in immersive, 3D environments [174].

Furthermore, the model could even be adapted for clinical applications to study visual and cognitive impairments. For instance, by training the RL model under simulated low-vision foveation, such as artificially restricting the model’s simulated visual field to mimic specific ocular conditions, we can gain deeper, mechanistic insights into the compensatory eye-movement behaviors associated with low-vision individuals.

Similarly, the fRL-SU represents a normative theory of eye movements for a person with normally functioning cognition. The fixation heatmaps from the model can be used to those of patients to assess or identify cognitive impairments that influence eye movements, including conditions such as Alzheimer’s, Attention Deficit Hyperactivity Disorder, and Autism Spectrum Disorder. The model can be used not only to evaluate whether fixations depart from normative scene exploration eye movements for any scene but also to assess the functional impact of atypical fixation patterns on scene comprehension.

Bibliography

- [1] G. T. Buswell, *How people look at pictures: a study of the psychology and perception in art*, Univ. Chicago Press, Oxford, England (1935). Pages: 198.
- [2] A. L. Yarbus. *Eye Movements During Perception of Complex Objects*. In A. L. Yarbus, editor, *Eye Movements and Vision*, pages 171–211. Springer US, Boston, MA, (1967).
- [3] G. R. Loftus and N. H. Mackworth, *Cognitive determinants of fixation location during picture viewing.*, Journal of Experimental Psychology: Human Perception and Performance **4**, 565 (1980).
- [4] A. M. Treisman and G. Gelade, *A feature-integration theory of attention*, Cognitive Psychology **12**, 97–136 (1980).
- [5] L. Itti, C. Koch, and E. Niebur, *A model of saliency-based visual attention for rapid scene analysis*, IEEE Transactions on Pattern Analysis and Machine Intelligence **20**, 1254–1259 (1998).
- [6] J. Harel, C. Koch, and P. Perona, *Graph-based visual saliency*, Advances in Neural Information Processing Systems **19** (2006).
- [7] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, *SUN: A Bayesian framework for saliency using natural statistics*, Journal of Vision **8**, 32–32 (2008).
- [8] N. D. B. Bruce and J. K. Tsotsos, *Saliency, attention, and visual search: An information theoretic approach*, Journal of Vision **9**, 5–5 (2009).

- [9] J. Zhang and S. Sclaroff. *Saliency detection: A boolean map approach*. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 153–160, (2013).
- [10] E. Erdem and A. Erdem, *Visual saliency estimation by nonlinearly integrating features using region covariances*, *Journal of Vision* **13**, 11–11 (2013).
- [11] J. M. Findlay and I. D. Gilchrist. *Chapter 13 - Eye Guidance and Visual Search*. In G. Underwood, editor, *Eye Guidance in Reading and Scene Perception*, pages 295–312. Elsevier Science Ltd, Amsterdam, (1998).
- [12] M. P. Eckstein, B. R. Beutter, and L. S. Stone, *Quantifying the performance limits of human saccadic targeting during visual search*, *Perception* **30**, 1389–1401 (2001).
- [13] W. Einhäuser, U. Rutishauser, and C. Koch, *Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli*, *Journal of Vision* **8**, 2–2 (2008).
- [14] G. L. Malcolm and J. M. Henderson, *Combining top-down processes to guide eye movements during real-world scene search*, *Journal of Vision* **10**, 4–4 (2010).
- [15] M. Wischnewski and M. V. Peelen, *Causal neural mechanisms of context-based object recognition*, *Elife* **10**, e69736 (2021).
- [16] A. J. de Brouwer, J. R. Flanagan, and M. Spering, *Functional use of eye movements for an acting system*, *Trends in Cognitive Sciences* **25**, 252–263 (2021).
- [17] J. Fookien, P. Kreyenmeier, and M. Spering, *The role of eye movements in manual interception: A mini-review*, *Vision Research* **183**, 81–90 (2021).
- [18] M. M. Hayhoe and R. A. Lerch. *Visual Guidance of Natural Behavior*, (2022).

- [19] M. Hayhoe and D. Ballard, *Eye movements in natural behavior*, Trends in Cognitive Sciences **9**, 188–194 (2005).
- [20] M. M. Hayhoe, *Vision and action*, Annual Review of Vision Science **3**, 389–413 (2017).
- [21] M. F. Peterson and M. P. Eckstein, *Looking just below the eyes is optimal across face recognition tasks*, Proceedings of the National Academy of Sciences **109**, E3314–E3323 (2012).
- [22] L. W. Renninger, P. Verghese, and J. Coughlan, *Eye movements incorporate knowledge of part structure*, Journal of Vision **6**, 482–482 (2006).
- [23] J. Najemnik and W. S. Geisler, *Optimal eye movement strategies in visual search*, Nature **434**, 387–391 (2005).
- [24] E. Akbas and M. P. Eckstein, *Object detection through search with a foveated visual system*, PLOS Computational Biology **13**, e1005743 (2017).
- [25] D. Hoppe and C. A. Rothkopf, *Multi-step planning of eye movements in visual search*, Scientific reports **9**, 144 (2019).
- [26] D. Parkhurst, K. Law, and E. Niebur, *Modeling the role of salience in the allocation of overt visual attention*, Vision Research **42**, 107–123 (2002).
- [27] C. E. Peacock, T. R. Hayes, and J. M. Henderson, *The role of meaning in attentional guidance during free viewing of real-world scenes*, Acta Psychologica **198**, 102889 (2019).
- [28] N. Bruce and J. Tsotsos, *Saliency based on information maximization*, Advances in Neural Information Processing Systems **18** (2005).

- [29] L. Itti. *CHAPTER 94 - Models of Bottom-up Attention and Saliency*. In L. Itti, G. Rees, and J. K. Tsotsos, editors, *Neurobiology of Attention*, pages 576–582. Academic Press, Burlington, (2005).
- [30] O. Le Meur, D. Thoreau, P. Le Callet, and D. Barba. *A spatio-temporal model of the selective human visual attention*. In *IEEE International Conference on Image Processing 2005*, volume 3, pages III–1188. IEEE, (2005).
- [31] R. J. Peters, A. Iyer, L. Itti, and C. Koch, *Components of bottom-up gaze allocation in natural images*, *Vision Research* **45**, 2397–2416 (2005).
- [32] O. Le Meur, P. Le Callet, and D. Barba, *Predicting visual fixations on video based on low-level visual features*, *Vision Research* **47**, 2483–2498 (2007).
- [33] A. Borji and L. Itti, *State-of-the-art in visual attention modeling*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 185–207 (2012).
- [34] A. Borji, D. N. Sihite, and L. Itti, *What stands out in a scene? A study of human explicit saliency judgment*, *Vision Research* **91**, 62–77 (2013).
- [35] J. Stoll, M. Thrun, A. Nuthmann, and W. Einhäuser, *Overt attention in natural scenes: Objects dominate features*, *Vision Research* **107**, 36–48 (2015).
- [36] J. M. Henderson and T. R. Hayes, *Meaning-based guidance of attention in scenes as revealed by meaning maps*, *Nature human behaviour* **1**, 743–747 (2017).
- [37] M. Cerf, E. P. Frady, and C. Koch, *Faces and text attract gaze independent of the task: Experimental data and computer model*, *Journal of Vision* **9**, 10–10 (2009).
- [38] E. Birmingham, W. F. Bischof, and A. Kingstone, *Gaze selection in complex social scenes*, *Visual cognition* **16**, 341–355 (2008).
- [39] H.-C. Wang and M. Pomplun, *The attraction of visual attention to texts in real-world scenes*, *Journal of vision* **12**, 26–26 (2012).

- [40] B. De Haas, A. L. Iakovidis, D. S. Schwarzkopf, and K. R. Gegenfurtner, *Individual differences in visual salience vary along semantic dimensions*, Proceedings of the National Academy of Sciences **116**, 11687–11692 (2019).
- [41] W. Einhäuser, M. Spain, and P. Perona, *Objects predict fixations better than early saliency*, Journal of vision **8**, 18–18 (2008).
- [42] A. Nuthmann and J. M. Henderson, *Object-based attentional selection in scene viewing*, Journal of vision **10**, 20–20 (2010).
- [43] C. E. Peacock, T. R. Hayes, and J. M. Henderson, *The role of meaning in attentional guidance during free viewing of real-world scenes*, Acta Psychologica **198**, 102889 (2019).
- [44] K. H. Ruddock, D. S. Wooding, and S. K. Mannan, *The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images*, Spatial vision **10**, 165–188 (1996).
- [45] B. W. Tatler, *The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions*, Journal of vision **7**, 4–4 (2007).
- [46] M. Kümmerer, M. Bethge, and T. S. Wallis, *DeepGaze III: Modeling free-viewing human scanpaths with deep learning*, Journal of Vision **22**, 7–7 (2022).
- [47] G. Cartella, V. Cuculo, A. D’Amelio, M. Cornia, G. Boccignone, and R. Cucchiara. *Modeling Human Gaze Behavior with Diffusion Models for Unified Scanpath Prediction*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16206–16216, (2025).
- [48] M. Assens, X. Giro-i Nieto, K. McGuinness, and N. E. O’Connor. *PathGAN: Visual scanpath prediction with generative adversarial networks*. In *Proceedings*

- of the European Conference on Computer Vision (ECCV) Workshops, pages 0–0, (2018).
- [49] A. Linardos, M. Kümmerer, O. Press, and M. Bethge. *DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12919–12928, (2021).
 - [50] M. Linka, H. Karimpur, and B. de Haas, *Protracted development of gaze behaviour*, *Nature Human Behaviour* **9**, 1887–1897 (2025).
 - [51] M. A. Lago, C. K. Abbey, and M. P. Eckstein, *Foveated model observers for visual search in 3D medical images*, *IEEE transactions on medical imaging* **40**, 1021–1031 (2020).
 - [52] A. Oliva, T. Konkle, M. R. Greene, and A. Torralba, *Not all scene categories are created equal: The role of object and layout diagnosticity in scene gist understanding*, *Journal of Vision* **6**, 464–464 (2006).
 - [53] M. P. Eckstein, B. A. Drescher, and S. S. Shimozaki, *Attentional cues in real scenes, saccadic targeting, and Bayesian priors*, *Psychological Science* **17**, 973–980 (2006).
 - [54] M. B. Neider and G. J. Zelinsky, *Scene context guides eye movements during visual search*, *Vision Research* **46**, 614–621 (2006).
 - [55] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, *Eye guidance in natural vision: Reinterpreting salience*, *Journal of Vision* **11**, 5–5 (2011).
 - [56] A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner, *Eye movements and perception: A selective review*, *Journal of Vision* **11**, 9–9 (2011).

- [57] A. C. Schütz, J. Trommershäuser, and K. R. Gegenfurtner, *Dynamic integration of information about salience and value for saccadic eye movements*, Proceedings of the National Academy of Sciences **109**, 7547–7552 (2012).
- [58] M. Hayhoe and D. Ballard, *Modeling task control of eye movements*, Current Biology **24**, R622–R628 (2014).
- [59] T. Foulsham, *Eye movements and their functions in everyday tasks*, Eye **29**, 196–199 (2015).
- [60] K. Koehler and M. P. Eckstein, *Beyond scene gist: Objects guide search more than scene background.*, Journal of Experimental Psychology: Human Perception and Performance **43**, 1177 (2017).
- [61] M. P. Eckstein, K. Koehler, L. E. Welbourne, and E. Akbas, *Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes*, Current Biology **27**, 2827–2832.e3 (2017).
- [62] P. N. Chakravarthula, Y. Tsank, and M. P. Eckstein, *Eye movement strategies in face ethnicity categorization vs. face identification tasks*, Vision Research **186**, 59–70 (2021).
- [63] N. X. Han and M. P. Eckstein, *Head and body cues guide eye movements and facilitate target search in real-world videos*, Journal of Vision **23**, 5–5 (2023).
- [64] J. M. Henderson, T. R. Hayes, G. Rehrig, and F. Ferreira, *Meaning Guides Attention during Real-World Scene Description*, Scientific Reports **8**, 13504 (2018).
- [65] M. I. Coco and F. Keller, *Scan patterns predict sentence production in the cross-modal processing of visual scenes*, Cognitive science **36**, 1204–1223 (2012).
- [66] M. I. Coco and F. Keller, *Classification of visual and linguistic tasks using eye-movement features*, Journal of Vision **14**, 11–11 (2014).

- [67] Y. Esaulova, M. Penke, and S. Dolscheid, *Describing events: Changes in eye movements and language production due to visual and conceptual properties of scenes*, *Frontiers in Psychology* **10**, 835 (2019).
- [68] E. Birmingham, W. F. Bischof, and A. Kingstone, *Saliency does not account for fixations to eyes within social scenes*, *Vision Research* **49**, 2992–3000 (2009).
- [69] T. Judd, K. Ehinger, F. Durand, and A. Torralba. *Learning to predict where humans look*. In *2009 IEEE 12th International Conference on Computer Vision*, volume , pages 2106–2113. IEEE, (2009).
- [70] T. Judd, F. Durand, and A. Torralba. *A Benchmark of Computational Models of Saliency to Predict Human Fixations*. In *MIT Technical Report*, (2012).
- [71] A. Borji and L. Itti, *Cat2000: A large scale fixation dataset for boosting saliency research*, *arXiv preprint* (2015).
- [72] K. Koehler, F. Guo, S. Zhang, and M. P. Eckstein, *What do saliency models predict?*, *Journal of vision* **14**, 14–14 (2014).
- [73] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, *Intrinsic and extrinsic effects on image memorability*, *Vision Research* **116**, 165–178 (2015).
- [74] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, *Predicting human gaze beyond pixels*, *Journal of Vision* **14**, 28–28 (2014).
- [75] W. K. Hoh, F.-L. Zhang, and N. A. Dodgson, *Salient-centeredness and saliency size in computational aesthetics*, *ACM Transactions on Applied Perception* **20**, 1–23 (2023).
- [76] H. J. Levesque, E. Davis, and L. Morgenstern, *The Winograd schema challenge.*, *KR* **2012**, 3 (2012).

- [77] E. M. Murlidaran Shravan. *Winograd Images used in our study*, (2025). Available at <https://github.com/shravan1394/WinogradDataset>.
- [78] M. Kümmerer. *Saleincy Models Implementation*. <https://github.com/matthias-k/pysaliency>, (2015).
- [79] N. D. Bruce, C. Wloka, N. Frosst, S. Rahman, and J. K. Tsotsos, *On computational modeling of visual saliency: Examining what’s right, and what’s left*, Vision research **116**, 95–112 (2015).
- [80] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556 (2014).
- [81] M. Kümmerer. *Deep Gaze ii*. <https://github.com/matthias-k/DeepGaze>, (2016).
- [82] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti. *Analysis of scores, datasets, and models in visual saliency prediction*. In *Proceedings of the IEEE international conference on computer vision*, pages 921–928, (2013).
- [83] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, et al., *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*, arXiv preprint arXiv:2403.05530 (2024).
- [84] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., *Gpt-4 technical report*, arXiv preprint arXiv:2303.08774 (2023).
- [85] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society: series B (Methodological) **57**, 289–300 (1995).

- [86] Y. Benjamini and D. Yekutieli, *The control of the false discovery rate in multiple testing under dependency*, Annals of statistics , 1165–1188 (2001).
- [87] C. L. Kleinke, *Gaze and eye contact: a research review.*, Psychological Bulletin **100**, 78 (1986).
- [88] R. Brooks and A. N. Meltzoff, *The development of gaze following and its relation to language*, Developmental Science **8**, 535–543 (2005).
- [89] M. S. Castelhana, M. Wieth, and J. M. Henderson. *I see what you see: Eye movements in real-world scenes are affected by perceived direction of gaze*. In *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint: 4th International Workshop on Attention in Cognitive Systems, WAPCV 2007 Hyderabad, India, January 8, 2007 Revised Selected Papers 4*, pages 251–262. Springer, (2007).
- [90] N. X. Han and M. P. Eckstein, *Inferential eye movement control while following dynamic gaze*, Elife **12**, e83187 (2023).
- [91] R. Eshuis, K. R. Coventry, and M. Vulchanova, *Predictive eye movements are driven by goals, not by the mirror neuron system*, Psychological Science **20**, 438–440 (2009).
- [92] C. Fawcett and G. Gredebäck, *Infants use social context to bind actions into a collaborative sequence*, Developmental Science **16**, 841–849 (2013).
- [93] G. Gredebäck and T. Falck-Ytter, *Eye movements during action observation*, Perspectives on Psychological Science **10**, 591–598 (2015).
- [94] P. Borovska and B. de Haas, *Faces in scenes attract rapid saccades*, Journal of Vision **23**, 11–11 (2023).
- [95] C. M. Fausey, S. Jayaraman, and L. B. Smith, *From faces to hands: Changing visual input in the first two years*, Cognition **152**, 101–107 (2016).

- [96] C. E. Peacock, T. R. Hayes, and J. M. Henderson, *Meaning guides attention during scene viewing, even when it is irrelevant*, Attention, Perception, & Psychophysics **81**, 20–34 (2019).
- [97] A. Borji, D. N. Sihite, and L. Itti, *Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study*, IEEE Transactions on Image Processing **22**, 55–69 (2012).
- [98] A. Torralba, A. Oliva, M. S. Castelhana, and J. M. Henderson, *Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search.*, Psychological Review **113**, 766 (2006).
- [99] J. M. Henderson, G. L. Malcolm, and C. Schandl, *Searching in the dark: Cognitive relevance drives attention in real-world scenes*, Psychonomic Bulletin & Review **16**, 850–856 (2009).
- [100] M. L.-H. Võ and J. M. Henderson, *Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception*, Journal of Vision **9**, 24–24 (2009).
- [101] J. M. Wolfe, M. L.-H. Võ, K. K. Evans, and M. R. Greene, *Visual search in scenes involves selective and nonselective pathways*, Trends in Cognitive Sciences **15**, 77–84 (2011).
- [102] M. S. Castelhana and C. Heaven, *Scene context influences without scene gist: Eye movements guided by spatial associations in visual search*, Psychonomic Bulletin & Review **18**, 890–896 (2011).
- [103] S. C. Mack and M. P. Eckstein, *Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment*, Journal of Vision **11**, 9–9 (2011).

- [104] M. S. Castelhana and R. L. Witherspoon, *How you use it matters: Object function guides attention during visual search in scenes*, Psychological Science **27**, 606–621 (2016).
- [105] M. P. Eckstein, *Probabilistic computations for attention, eye movements, and search*, Annual Review of Vision Science **3**, 319–342 (2017).
- [106] M. L.-H. Võ, S. E. Boettcher, and D. Draschkow, *Reading scenes: How scene grammar guides attention and aids perception in real-world environments*, Current Opinion in Psychology **29**, 205–210 (2019).
- [107] A. Goettker, H. Pidaparthi, D. I. Braun, J. H. Elder, and K. R. Gegenfurtner, *Ice hockey spectators use contextual cues to guide predictive eye movements*, Current Biology **31**, R991–R992 (2021).
- [108] T. R. Hayes and J. M. Henderson, *Looking for semantic similarity: what a vector-space model of semantics can tell us about attention in real-world scenes*, Psychological Science **32**, 1262–1270 (2021).
- [109] C. K. Friesen and A. Kingstone, *The eyes have it! Reflexive orienting is triggered by nonpredictive gaze*, Psychonomic bulletin & review **5**, 490–495 (1998).
- [110] N. X. Han and M. P. Eckstein, *Gaze-cued shifts of attention and microsaccades are sustained for whole bodies but are transient for body parts*, Psychonomic Bulletin & Review **29**, 1854–1878 (2022).
- [111] M. Rucci and M. Poletti, *Control and functions of fixational eye movements*, Annual review of vision science **1**, 499–518 (2015).
- [112] H.-k. Ko, M. Poletti, and M. Rucci, *Microsaccades precisely relocate gaze in a high visual acuity task*, Nature neuroscience **13**, 1549–1553 (2010).

- [113] M. Spering, A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner, *Keep your eyes on the ball: smooth pursuit eye movements enhance prediction of visual motion*, Journal of Neurophysiology **105**, 1756–1767 (2011).
- [114] G. J. Zelinsky, *A theory of eye movements during target acquisition.*, Psychological Review **115**, 787 (2008).
- [115] M. P. Eckstein, W. Schoonveld, S. Zhang, S. C. Mack, and E. Akbas, *Optimal and human eye movements to clustered low value cues to increase decision rewards during search*, Vision Research **113**, 137–154 (2015).
- [116] A. Hollingworth and J. M. Henderson, *Accurate visual memory for previously attended objects in natural scenes.*, Journal of Experimental Psychology: Human Perception and Performance **28**, 113 (2002).
- [117] E. Brenner, J. J. Granzier, and J. B. Smeets, *Perceiving colour at a glimpse: The relevance of where one fixates*, Vision Research **47**, 2557–2568 (2007).
- [118] G. M. Huebner and K. R. Gegenfurtner, *Effects of viewing time, fixations, and viewing strategies on visual memory for briefly presented natural objects*, Quarterly Journal of Experimental Psychology **63**, 1398–1413 (2010).
- [119] K. R. Gegenfurtner, *The interaction between vision and eye movements*, Perception **45**, 1333–1357 (2016).
- [120] K. Koehler and M. P. Eckstein, *Temporal and peripheral extraction of contextual cues from scenes during visual search*, Journal of Vision **17**, 16–16 (2017).
- [121] C. Araujo, E. Kowler, and M. Pavel, *Eye movements during visual search: The costs of choosing the optimal path*, Vision Research **41**, 3613–3625 (2001).
- [122] E. Kowler, *Eye movements: The past 25 years*, Vision Research **51**, 1457–1483 (2011).

- [123] L. C. Loschky, A. M. Larson, T. J. Smith, and J. P. Magliano, *The scene perception & event comprehension theory (SPECT) applied to visual narratives*, Topics in Cognitive Science **12**, 311–351 (2020).
- [124] L. C. Loschky, M. E. Smith, P. Chandran, J. P. Hutson, T. J. Smith, and J. P. Magliano, *The Role of Event Understanding in Guiding Attentional Selection in Real-world Scenes: The Scene Perception & Event Comprehension Theory (SPECT)*, PsyArXiv Preprints (2024).
- [125] E. Berlot, L. Schmitt, C. Huber-Huber, M. Peelen, and F. de Lange, *I see! How narrative meaning influences gaze behaviour*, Conference on Cognitive Computational Neuroscience (2023).
- [126] N. Roth, J. McLaughlin, K. Obermayer, and M. Rolf, *Gaze Behavior Reveals Expectations of Potential Scene Changes*, Psychological science **35**, 1350–1363 (2024).
- [127] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes, *Information-seeking, curiosity, and attention: computational and neural mechanisms*, Trends in Cognitive Sciences **17**, 585–593 (2013).
- [128] J. Gottlieb and P.-Y. Oudeyer, *Towards a neuroscience of active sampling and curiosity*, Nature Reviews Neuroscience **19**, 758–770 (2018).
- [129] A. Baranes, P.-Y. Oudeyer, and J. Gottlieb, *Eye movements reveal epistemic curiosity in human observers*, Vision Research **117**, 81–90 (2015).
- [130] A. Doerig, T. C. Kietzmann, E. Allen, Y. Wu, T. Naselaris, K. Kay, and I. Charest, *Semantic scene descriptions as an objective of human vision*, arXiv preprint arXiv:2209.11737 (2022).

- [131] F. F. Li, R. VanRullen, C. Koch, and P. Perona, *Rapid natural scene categorization in the near absence of attention*, Proceedings of the National Academy of Sciences **99**, 9596–9601 (2002).
- [132] A. Oliva and A. Torralba, *Building the gist of a scene: The role of global image features in recognition*, Progress in Brain Research **155**, 23–36 (2006).
- [133] A. Jonnalagadda, W. Y. Wang, B. Manjunath, and M. P. Eckstein, *Foveater: Foveated transformer for image classification*, arXiv preprint arXiv:2105.14173 (2021).
- [134] D. Kollenda, A.-S. Reher, and B. de Haas, *Individual gaze predicts individual scene descriptions*, Scientific Reports **15**, 9443 (2025).
- [135] H. F. Chua, J. E. Boland, and R. E. Nisbett, *Cultural variation in eye movements during scene perception*, Proceedings of the national academy of sciences **102**, 12629–12633 (2005).
- [136] C. A. Rothkopf and M. M. Hayhoe, *Computational elements of natural vision*, Journal of Vision **25**, 4–4 (2025).
- [137] S. Murlidaran and M. P. Eckstein, *Eye movements during free viewing to maximize scene understanding*, Nat. Commun. **17**, 940 (2025).
- [138] M. Land, N. Mennie, and J. Rusted, *The roles of vision and eye movements in the control of activities of daily living*, Perception **28**, 1311–1328 (1999).
- [139] L. C. Loschky, M. E. Smith, P. Chandran, J. P. Hutson, T. J. Smith, and J. P. Magliano, *The role of event understanding in guiding attentional selection in real-world scenes: The scene perception & event comprehension theory (spect)*, Attention, Perception, & Psychophysics **88**, 92 (2026).

- [140] S. Lu, Y. Li, Q.-G. Chen, Z. Xu, W. Luo, K. Zhang, and H.-J. Ye, *Ovis: Structural Embedding Alignment for Multimodal Large Language Model*, arXiv:2405.20797 (2024).
- [141] S. Lu, Y. Li, Y. Xia, Y. Hu, S. Zhao, Y. Ma, Z. Wei, Y. Li, L. Duan, J. Zhao, Y. Han, H. Li, W. Chen, J. Tang, C. Hou, Z. Du, T. Zhou, W. Zhang, H. Ding, J. Li, W. Li, G. Hu, Y. Gu, S. Yang, J. Wang, H. Sun, Y. Wang, H. Sun, J. Huang, Y. He, S. Shi, W. Zhang, G. Zheng, J. Jiang, S. Gao, Y.-F. Wu, S. Chen, Y. Chen, Q.-G. Chen, Z. Xu, W. Luo, and K. Zhang, *Ovis2.5 Technical Report*, arXiv:2508.11737 (2025).
- [142] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, et al., *Minicpm-v: A gpt-4v level mllm on your phone*, arXiv preprint arXiv:2408.01800 (2024).
- [143] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, et al., *Qwen3-vl technical report*, arXiv preprint arXiv:2511.21631 (2025).
- [144] C. Clark, J. Zhang, Z. Ma, J. S. Park, M. Salehi, R. Tripathi, S. Lee, Z. Ren, C. D. Kim, Y. Yang, et al., *Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding*, arXiv preprint arXiv:2601.10611 (2026).
- [145] J. S. Perry and W. S. Geisler. *Gaze-contingent real-time simulation of arbitrary visual fields*. In *Human vision and electronic imaging VII*, volume 4662, pages 57–69. SPIE, (2002).
- [146] S. Zhang and M. P. Eckstein, *Evolution and optimality of similar neural mechanisms for perception and action during search*, PLoS Computational Biology **6**, e1000930 (2010).

- [147] J. F. Ackermann and M. S. Landy, *Choice of saccade endpoint under risk*, Journal of vision **13**, 27–27 (2013).
- [148] T. H. Weisswange, C. A. Rothkopf, T. Rodemann, and J. Triesch. *Can reinforcement learning explain the development of causal inference in multisensory integration?* In *2009 IEEE 8th International Conference on Development and Learning*, pages 1–7. IEEE, (2009).
- [149] B. Sullivan, L. Johnson, D. Ballard, and M. Hayhoe. *A modular reinforcement learning model for human visuomotor behaviour in a driving task*. In *Proc. AISB 2011 Symposium*, pages 33–40, (2011).
- [150] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., *Human-level control through deep reinforcement learning*, nature **518**, 529–533 (2015).
- [151] W. Zhou and M. P. Eckstein. *A deep Q-learning method for optimizing visual search strategies in backgrounds of dynamic noise*. In *Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment*, volume 12035, pages 60–67. SPIE, (2022).
- [152] M. D. Anderson, E. W. Graf, J. H. Elder, K. A. Ehinger, and W. J. Adams, *Category systems for real-world scenes*, Journal of vision **21**, 8–8 (2021).
- [153] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. *2D Human Pose Estimation: New Benchmark and State of the Art Analysis*. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2014).
- [154] L. Mertens, E. Yargholi, H. Op de Beeck, J. Van den Stock, and J. Vennekens, *Findingemo: An image dataset for emotion recognition in the wild*, Advances in Neural Information Processing Systems **37**, 4956–4996 (2024).

- [155] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, *Microsoft COCO: Common Objects in Context*, arXiv:1405.0312 [cs] (2015).
- [156] M. Jiang, S. Huang, J. Duan, and Q. Zhao. *SALICON: Saliency in Context*. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015).
- [157] Z. Ouyang. *Image_Foveation_Python: Python implementation of image foveation*. https://github.com/ouyangzhibo/Image_Foveation_Python, (2018). GitHub repository, accessed: 2018.
- [158] M. Dehghani, B. Mustafa, J. Djolonga, J. Heek, M. Minderer, M. Caron, A. Steiner, J. Puigcerver, R. Geirhos, I. M. Alabdulmohsin, et al., *Patch n’pack: Navit, a vision transformer for any aspect ratio and resolution*, *Advances in Neural Information Processing Systems* **36**, 2252–2274 (2023).
- [159] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., *Qwen3 technical report*, arXiv preprint arXiv:2505.09388 (2025).
- [160] D. Zhang, J. Li, Z. Zeng, and F. Wang, *Jasper and stella: distillation of sota embedding models*, arXiv preprint arXiv:2412.19048 (2024).
- [161] R. J. Williams, *Simple statistical gradient-following algorithms for connectionist reinforcement learning*, *Machine learning* **8**, 229–256 (1992).
- [162] W. Kwon. *vLLM: An Efficient Inference Engine for Large Language Models*. PhD thesis, UC Berkeley, (2025).
- [163] I. Oruc, F. Shafai, S. Murthy, P. Lages, and T. Ton, *The adult face-diet: A naturalistic observation study*, *Vision research* **157**, 222–229 (2019).

- [164] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, et al., *Sam 3: Segment anything with concepts*, arXiv preprint arXiv:2511.16719 (2025).
- [165] S. Murlidaran, Z. Wen, J. Skaza, W. Wang, and M. P. Eckstein, *Semantic Saliency from Multi-Modal Large Language Model Scene Understanding Maps*, arXiv preprint (2025).
- [166] M. Linka and B. de Haas, *OSIEshort: A small stimulus set can reliably estimate individual differences in semantic salience*, Journal of vision **20**, 13–13 (2020).
- [167] E. Akbas and M. P. Eckstein, *Object detection through search with a foveated visual system*, PLoS computational biology **13**, e1005743 (2017).
- [168] A. D. Clarke and B. W. Tatler, *Deriving an appropriate baseline for describing fixation behaviour*, Vision research **102**, 41–51 (2014).
- [169] P.-H. Tseng, R. Carmi, I. G. Cameron, D. P. Munoz, and L. Itti, *Quantifying center bias of observers in free viewing of dynamic natural scenes*, Journal of vision **9**, 4–4 (2009).
- [170] G. Rosenberg, S. Stadhard, B. C. Hansen, and M. R. Greene, *The Limits of Learning from Pictures and Text: Vision-Language Models and Embodied Scene Understanding*, arXiv preprint arXiv:2603.26589 (2026).
- [171] F. A. Wichmann and R. Geirhos, *Are deep neural networks adequate behavioral models of human visual perception?*, Annual review of vision science **9**, 501–524 (2023).
- [172] M. Martelli, N. J. Majaj, and D. G. Pelli, *Are faces processed like words? A diagnostic test for recognition by parts*, Journal of Vision **5**, 6–6 (2005).
- [173] N. V. Powell, D. Panfili, Y. Oh, J. S. Matthis, and M. Hayhoe, *The Role of Retinal Flow in Walking*, bioRxiv , 2026–02 (2026).

- [174] Y. Oh, N. Powell, D. Panfili, and M. Hayhoe, *Gaze in Dynamic Natural Environments*, Journal of Vision **25**, 2766–2766 (2025).