A Framework for Leveraging LLMs for Scene Analysis and Cognitive Processing

CATARINA MOREIRA*, University of Technology Sydney, Australia
JEFFREY COCKBURN, University of Iowa, USA
MONICA S. CASTELHANO, Queen's University, Canada

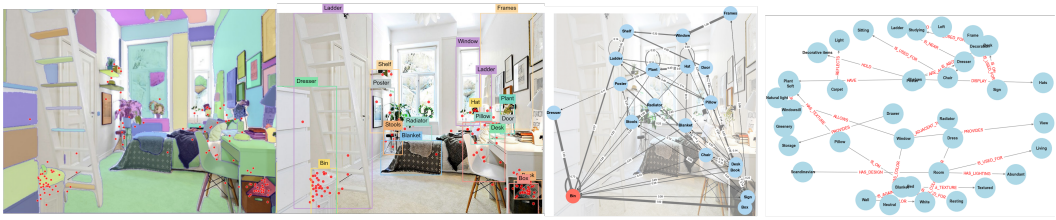

Fig. 1. A gaze-driven segmentation of a complex scene labeled using large language models. Each segmented region reflects human-like attention patterns to identify likely object locations, illustrating the framework's ability to generate a semantically rich and human-centered scene representations

In everyday visual search tasks, humans rely on prior knowledge of object placements in scenes to efficiently locate target objects. This ability is evidenced by eye movement patterns, where individuals focus on areas that are more likely to contain the target, such as searching for a cup on a table or shoes on the floor. Building on this, we propose a new annotation pipeline that leverages these priors by extracting a knowledge graph from images based on automatically annotated objects. This knowledge graph is then used with large language models (LLMs) to predict the most likely locations of a specific target object in an image. Our approach is the first instance of using LLMs to identify relevant prior knowledge in images and to bridge the gap between human scene understanding and computational models.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Knowledge representation and reasoning**; **Image segmentation**.

Additional Key Words and Phrases: Visual search, Eye movement analysis, Knowledge graph reasoning, Large language models

ACM Reference Format:
Catarina Moreira, Jeffrey Cockburn, and Monica S. Castelhano. 2025. A Framework for Leveraging LLMs for Scene Analysis and Cognitive Processing. *Proc. ACM Comput. Graph. Interact. Tech.* 8, 2, Article 27 (June 2025), 18 pages. https://doi.org/10.1145/3729414

*Also with INESC-ID Lisboa.

Authors' Contact Information: Catarina Moreira, University of Technology Sydney, Data Science Institute, Sydney, Australia, catarina.pintomoreira@uts.edu.au; Jeffrey Cockburn, University of Iowa, Iowa City, Iowa, USA, jeffrey-cockburn@uiowa.edu; Monica S. Castelhano, Queen's University, Psychology, Kingston, Ontario, Canada, monica.castelhano@queensu.ca.

ACM 2577-6193/2025/6-ART27
https://doi.org/10.1145/3729414

1 Introduction

Human ability to locate objects rapidly and efficiently within their environment is rooted in extensive prior knowledge of typical scene categories and spatial associations. This understanding enables individuals to use familiar spatial patterns to guide attention to likely object locations, such as searching for a stove in a kitchen or a monitor on a desk [Castelhano and Heaven 2011; Oliva and Torralba 2007]. Eye movements, which direct attention, are thus driven not only by visual saliency but also by these learned spatial relationships, which reflect the complex interplay between scene context and object function [Castelhano and Krzyś 2020; Pereira and Castelhano 2014; Torralba et al. 2006; Williams and Castelhano 2019; Wolfe et al. 2011]. Research has shown that when objects are located in their expected contexts, search performance improves significantly, whereas misplaced objects—such as a pot found in a bathroom—impair search efficiency as attention must adjust to unexpected arrangements [Castelhano and Heaven 2011; Castelhano and Witherspoon 2016; Neider and Zelinsky 2006; Võ and Wolfe 2013].

This reliance on spatial context extends beyond general scene understanding to include highly specific expectations about object placements [Biederman 1972; Henderson 2003]. Studies on visual search indicate that these spatial associations are so ingrained that they guide attention even without the full scene gist, underscoring the importance of a priori knowledge [Castelhano and Heaven 2011; Võ and Wolfe 2013]. More recent research further reveals that eye movements during visual search are proactive, informed by both scene structure and accumulated experiences, which direct gaze toward functionally relevant regions like countertops in kitchens or desks in offices [Krzyś et al. 2024, 2023; Pereira and Castelhano 2019; Zelinsky et al. 2019]. This proactive behavior highlights the predictive power of scene context, which optimizes search strategies by focusing attention where target objects are most likely to be.

Despite substantial progress in understanding human visual search, existing computational models are often limited by their reliance on saliency or scene-context models that do not fully integrate the rich spatial and semantic associations that underpin human search behaviors [Cornia et al. 2018; Harel et al. 2006; Itti and Koch 2000, 2001; Judd et al. 2009; Kümmerer et al. 2016; Torralba et al. 2006]. These models, although increasingly capable of capturing low-level features and saliency patterns, still fall short of representing scene semantics and spatial organization in a way that mirrors the human approach to scene understanding. The scientific community has expressed the need for models that more accurately reflect the depth of human cognitive strategies by leveraging scene semantics, prior knowledge, and spatial associations to improve computational efficiency and semantic relevance in visual search tasks.

To address this gap, our work proposes a novel framework that incorporates human prior knowledge into gaze-driven scene analysis. Unlike traditional predictive models of visual search, our framework does not aim to simulate search behaviors directly from eye-tracking data. Instead, we leverage human gaze patterns to segment scenes and construct structured knowledge representations that integrate human prior knowledge, as encoded in LLMs, to analyze object placements in a computationally formalized way, leading to a formal scene representation to study how spatial and semantic relationships influence human visual search. We achieve this by employing Meta's Segment Anything Model (SAM) for scene segmentation and using fixation data from eye-tracking to identify regions of interest, our model generates segmentation masks aligned with participants' attention patterns. Each segmented region is then labeled by a large language model (LLM), producing a knowledge graph that captures the semantic associations between objects within the scene. This approach enables our framework to prioritize probable object locations based on prior knowledge and to mimic human-like search strategies through context-aware labeling and knowledge graph generation. This knowledge graph serves as a structured representation of prior knowledge,

allowing us to analyze how well LLM-derived priors align with human expectations of object placements. To the best of our knowledge, no prior work has formalized this type of knowledge representation within a scene graph framework.

Our main contributions to the field are as follows:

(1) **Integration of Gaze-Driven Segmentation with LLM-Based Labeling:** Our framework is the first to combine gaze-driven segmentation with LLM-based contextual labeling, producing segmentation masks that align with human attention patterns, thereby enhancing contextual relevance;
(2) **Generation of Semantically Rich Knowledge Graphs:** The framework constructs knowledge graphs representing spatial, functional, and semantic associations among objects, closely aligning with human cognitive heuristics and scene understanding;
(3) **Prioritization of (Un)Likely Object Locations:** Leveraging the LLM's prior knowledge, our model identifies *likely* and *unlikely* object locations, enhancing scene querying to reflect human perceptual biases;
(4) **Structured Analysis of Spatial and Semantic Relationships in Visual Search**: The framework provides a basis for analyzing how spatial and semantic relationships guide human search patterns, offering a powerful tool for advancing human-centered scene analysis in applications such as augmented reality, robotics, and human-computer interaction.

2 Related Work

Numerous computational models have been developed to predict eye movement patterns in scene viewing. These models range from those based on low-level image properties, such as saliency, to models that incorporate high-level scene context and semantic information.

The saliency-based model by Itti and Koch [2000, 2001] uses low-level visual features, like color and orientation, to create a saliency map that highlights visually prominent regions likely to attract eye movements. This foundational model demonstrated that attention is drawn to areas with high visual contrast. Building on this, the graph-based visual saliency (GBVS) model by Harel et al. [2006] represents each pixel as a node in a graph, connecting similar pixels to enhance salient regions and suppress noise. GBVS provides refined predictions of eye movements across various scenes. Judd et al. [2009] advanced saliency modeling by combining low-, mid-, and high-level features, using machine learning to weigh each feature's contribution. This multi-level integration enhances the model's ability to predict eye movements more accurately in natural scenes. Other works in the literature also attempt to use saliency-maps for more targeted object detection [Hsieh et al. 2023; Neves et al. 2024].

Saliency models fall short of capturing high-level context and semantic factors that also guide attention in scenes. Addressing this, the contextual guidance model by Torralba et al. [2006] combines global scene statistics, such as spatial layout, with saliency to prioritize task-relevant areas. This model was the first to bridge low-level saliency with learned scene knowledge. More recent is the DeepGaze II model by Kümmerer et al. [2016], which uses a convolutional neural network to capture high-level features, producing saliency maps that align with human gaze patterns in natural scenes. This model combines deep learning with traditional saliency, adding semantic relevance. Lastly, the SAM-RES model by Cornia et al. [2018] applies an encoder-decoder architecture to generate saliency maps that reflect scene structure and object relationships. Through semantic segmentation, SAM-RES recognizes categories and spatial arrangements, producing nuanced gaze predictions based on object relevance and expected locations.

Despite these advancements, there remains a need for models that fully capture scene semantics and the spatial organization of objects in a way that reflects human scene understanding. Our

proposed model, which uses segmentation to define distinct regions within a scene and integrates eye movement data to identify relevant areas, offers a novel approach to addressing this gap. By labeling these segmented regions using a language model and constructing a knowledge graph, we can generate a rich representation of the scene's semantic structure. This knowledge graph captures both the likely and unlikely locations of objects based on typical spatial associations. This approach represents a step forward by combining the strengths of segmentation, scene semantics, and machine learning to yield a model that reflects the a priori knowledge humans use to understand scenes.

3 Our Framework

Our framework is designed to analyze gaze-driven attention patterns in complex scenes and generate knowledge graphs to capture the semantic relationships between objects fixated by participants. The approach uses deep learning models and LLMs to integrate image processing, fixation-driven segmentation, scene label generation, and knowledge graph construction. Below is a detailed description of the steps in the proposed framework presented in Figure 2.

3.1 Image Processing and Prompt Generation

The Image Processing and Prompt Generation phase is the critical first step in the proposed framework, designed to bridge human eye-tracking data and deep learning-driven image segmentation. This phase processes raw gaze data, identifies regions of interest (ROIs) in the scene based on participants' fixations, and generates structured prompts for segmentation using Meta's Segment Anything Model (SAM). The goal is to accurately represent and the regions participants focus on in order to provide input for downstream semantic analysis.

Eye Fixations and Uncertainty Representation. Participants in the study were asked to search for a pre-specified target object within a scene image while their eye movements were recorded. Fixations f, which represent points of focused visual attention, were used as the foundation for generating prompts. However, due to the inherent imprecision in eye-tracking data and in the human visual system such that fixation coordinates may not perfectly align with the object perceived, a method for representing uncertainty around each fixation was required.

This approach ensures that visual information surrounding the fixation point is included, capturing a broader region of interest and mitigating fixation imprecision.

To account for this uncertainty, we extend the fixations into *focus areas* of *22 pixels*, which corresponded to 1 degree of visual angle, based on the characteristics of the eye-tracking system and the resolution of the image. This choice is supported by studies demonstrating that visual attention extends beyond the fixation point into the parafoveal and peripheral regions, influencing perception and search behavior [Rayner 2009]. This approach ensured that the visual information surrounding the fixation point were included. This added uncertainty allowed us to capture a more comprehensive region of interest.

Fixation Prompts: Four Geometric Representations. To facilitate the interaction between the gaze data and the SAM foundation model, we represented each fixation using one of four geometric shapes—referred to as *fixation prompts*. These were designed to reflect different levels of precision and spatial extent around the fixation coordinates. The four possible prompts are:

- **Pointwise:** This is the most precise representation and corresponds directly to a single fixation point f_i. It serves as an exact match of where the participant's gaze is recorded without any surrounding area being considered.

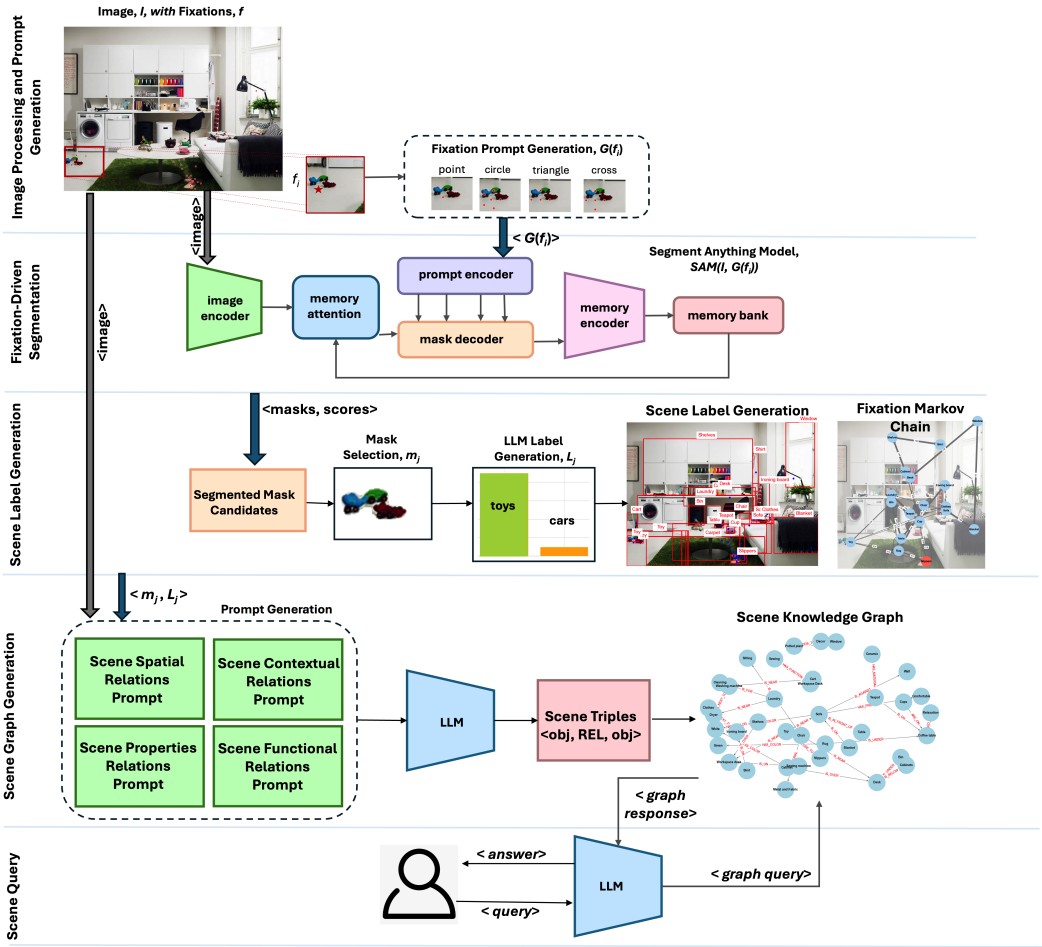

Fig. 2. Framework for Gaze-Driven Scene Segmentation and LLM-Augmented Knowledge Graph Generation

- **Circle:** The circle has a *22-pixel radius* centered at the fixation point. This representation captures a broader region around the fixation, accommodating slight inaccuracies in eye-tracking data. It aligns with known parafoveal processing effects, where objects are perceives beyond their immediate fixation [Rayner 2009].
- **Triangle:** This prompt extends *22 pixels* from the center fixation point in three directions—upward, rightward, and leftward—forming a triangular region. This shape models the possibility that the participant may focus on multiple adjacent regions within a scene, reflecting a more distributed fixation pattern [Rayner 2009].
- **Cross:** It extends *22 pixels* in four directions (up, down, left, and right) from the fixation centre, forming a ”+” shape. This prompt is used when it is assumed that the participant’s attention is spread in multiple directions simultaneously, representing a larger area of interest without fully encompassing the circular region. This shape is particularly useful when dealing with cluttered or complex scenes, as it aligns with findings that fixation transitions are influenced by both spatial and semantic scene structures [Rayner 2009].

Each fixation prompt $G(f_i)$ is described by the fixation point (x_i, y_i) and the geometric shape that was chosen to represent the spatial extent of attention. These prompts encode possible areas of focus and are used in the next phase to guide the segmentation process. Figure 3 illustrates an example of a participant fixating on the floor near the toys, highlighting how different fixation prompts capture varying levels of attention spread.

The choice of multiple geometric representations for fixation prompts was motivated by the need to model different types of attention. While the *pointwise* prompt captures highly focused attention, the *circle*, *triangle*, and *cross* prompts provide flexibility in representing broader or directional attention patterns. This approach is well-aligned with empirical studies in scene perception, which indicate that attention extends dynamically beyond fixation points to aid object recognition and localization [Rayner 2009]. This is important when there is uncertainty about the object the participant was fixating on or when they were scanning a region with multiple objects. The different prompt representations help the segmentation model generate more contextually appropriate masks for each ROI.

3.2 Fixation-Driven Segmentation

The Fixation-Driven Segmentation phase leverages the fixation prompts generated in the previous step to direct the segmentation process. This phase uses Meta's Segment Anything Model (SAM) [Kirillov et al. 2023] to create segmentation masks that are aligned with participants' visual focus, ensuring that the segmentation is human-centered and contextually relevant. Using fixation data as input, SAM's segmentation capabilities are tuned to focus on regions of interest, leading to more efficient and interpretable segmentations.

SAM is a powerful, general-purpose segmentation model that produces accurate segmentation masks based on different prompts, such as points, bounding boxes, or textual descriptions. In this framework, we take advantage of SAM's flexibility by using the fixation prompts $G(f_i)$ as the input that directs SAM to the specific regions of the image where participants focused their gaze. Each fixation prompt encodes the participant's point of attention and is passed to SAM, which processes the input through its *image encoder* and *prompt encoder* modules. The image encoder extracts features from the image. In contrast, the prompt encoder uses the fixation data to guide the segmentation process, ensuring that SAM focuses on the regions most relevant to the participant's visual search.

The core of this interaction is the *attention mechanism* used by SAM, which enables the model to allocate computational resources to specific image regions, which we based on the fixation prompts. This attention mechanism ensures that only the areas around the fixations are processed in depth, effectively reducing the *search space* for segmentation. By narrowing down the regions of interest based on fixation data, we significantly enhance the efficiency of the segmentation process.

Generation of Segmentation Masks. For each fixation prompt, SAM produces a set of *segmentation mask candidates*, each representing a potential object or region at the focus of the participant's attention. These masks are evaluated based on the SAM's internal scoring mechanism, which assesses how well the mask corresponds to the visual features surrounding the fixation point. This measure was used as an indicator of quality, and the best candidate m_j is selected for further processing.

This segmentation process yields masks highly relevant to the visual task, reflecting the objects or areas that participants were likely paying attention to during the search. The masks are directly tied to human visual behavior, ensuring that segmentations are accurate and meaningful within the scene context. Figure 3 presents an example of segmentation masks generated by SAM for different fixation prompts.

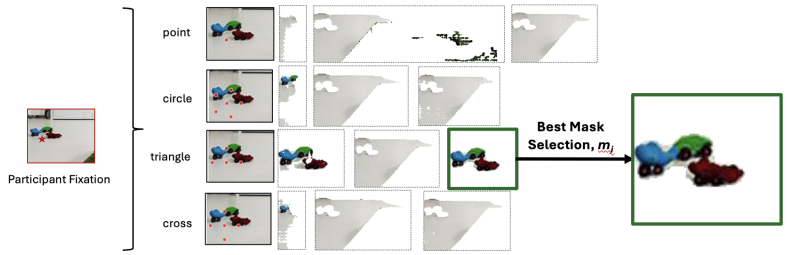

Fig. 3. Segmentations generated by SAM for different fixation prompts. Each segmentation mask is associated with a score. The segmentation with the highest score is selected as the *best mask candidate* for the participant's fixation.

Grounding Segmentation in Human Behavior. This phase ensures the generated segmentation masks are accurate and contextually relevant to the scene by grounding SAM's segmentation process in human gaze behavior. The resulting masks provide valuable insight into what participants perceived during the visual search task, enabling downstream analysis of attention patterns, object interactions, and task performance. The Fixation-Driven Segmentation phase ensures the segmentation process is tightly coupled with human visual attention, producing efficient, interpretable, and context-aware segmentation masks. By focusing on fixation-based prompts, the framework reduces computational overhead and enhances the quality of the segmentations and lays the groundwork for the subsequent scene understanding and knowledge graph generation.

3.3 Scene Label Generation

The Scene Label Generation phase of the framework assigns semantic labels to the segmented regions identified by the fixation-driven Segment Anything Model (SAM). After the best candidate segmentation masks are selected based on their quality and relevance to the participants' fixations, they are passed to an LLM for labeling. This phase leverages the LLM's semantic knowledge structure, which provides contextually accurate labels for each mask by analyzing the overall scene and the segmented regions. The process uses a chain of thought approach that retains the LLM's understanding of the scene while labeling individual objects. Each mask is associated with a bounding box that localizes the object or region of interest within the broader scene. We used *Open AI's GPT 4o as our LLM*. No fine-tuning was performed on the LLM or SAM; instead, we leveraged its pre-trained knowledge to generate scene descriptions, annotate segmentation masks, and construct the knowledge graph. The LLM was prompted with structured queries to ensure consistency in labeling and relationship extraction.

To ensure that the LLM assigns appropriate labels, the process is divided into two main steps: **(1) Scene Context Generation.** The LLM is first provided with the full image of the scene. Using a system prompt, the LLM is instructed to describe the scene in as much detail as possible. This scene description serves as the context for subsequent steps. The LLM's understanding of the scene image is critical to generating relevant and accurate labels for individual objects; and **(2) Mask Labeling Based on Context.** Once the LLM has processed and described the entire scene, the next step involves providing it with each segmented mask in isolation. For each mask, the LLM is prompted to label the object within the mask. Crucially, the LLM is asked to use the description generated for the full scene image to inform its labeling decision, ensuring that the label for each mask is contextually grounded.

Initial experiments comparing two approaches—mask-only labeling versus mask labeling with a textual scene description—revealed that without a textual summary of the scene, the LLM struggled to generate accurate and contextually relevant labels. This suggests that scene context is critical for correct object identification, as isolated segmentation masks often lack sufficient visual information to be reliably classified. By incorporating a full-scene textual description, the LLM retains an understanding of the broader scene, leading to more precise and meaningful annotations.

The output of this phase is a set of labeled masks, each accompanied by a corresponding bounding box. As the LLM generates these labels based on the contextual understanding of the full scene image, it is able to provide context-aware classifications for each segmented mask. The resulting output can be represented as $\langle m_j, L_j \rangle$, where m_j is the segmentation mask, and L_j is the label provided by the LLM. This labeling process generates high-quality, semantically relevant labels that reflect both the local visual content and the global context of the scene.

3.4 Scene Graph Generation

The Scene Graph Generation phase formalizes the transformation of segmented and labeled scene elements into a structured knowledge graph. This knowledge graph represents the relationships among objects in the scene, providing a semantic structure that facilitates computational understanding of the scene's spatial and functional organization.

The process begins by supplying an LLM with two sources of contextual information: *a detailed description of the scene* and *bounding box coordinates for each segmented object*. The scene description, generated in previous phases, offers a high-level overview of the image, detailing the types and arrangements of objects within it. The bounding box information, on the other hand, specifies the spatial positions and relative locations of each object. Together, these inputs allow the LLM to incorporate both the semantic content and the spatial structure of the scene into its analysis.

Upon receiving these contextual inputs, the LLM is prompted to generate a set of relational triples that capture the various interactions and associations among objects within a specific scene context. Specifically, the LLM is instructed to consider a range of relationship types that contribute to this comprehensive scene representation. These include spatial relationships, which describe the positioning and arrangement of objects (e.g., "the chair is next to the table" or the "cup is on the counter"), as well as functional relationships that capture the purpose or use of objects (e.g., "the cup is used for drinking"). The model also identifies semantic relationships, which link objects commonly associated or used together, and property relationships, which describe inherent attributes of objects such as color and shape.

Further, the LLM generates contextual relationships that associate objects with the broader scene context, as well as distance and size relationships that specify proximity and relative dimensions. For example, it might determine that "the rug is larger than the slippers" or "the plant is near the window." By generating these triples in the format (subject, RELATION, object), the LLM outputs a systematic coding of the structural and functional layout of the scene.

The resulting triples are organized into a scene knowledge graph, where each object is represented as a node and each relationship as an edge connecting nodes. This graph structure provides a formalized and machine-readable representation of the scene, in which objects are connected by specific relational types that reflect both local visual attributes and broader contextual associations. This representation supports querying and reasoning about the scene context and facilitates further analysis and downstream applications that require a structured understanding of object interactions and spatial configurations. For example, our framework can model Markov Chains to explore human cognitive patterns during visualization tasks [Moreira et al. 2023]. Figure 4 shows an example of a gaze-driven Markov Chain directly extracted from the scene labels and the LLM-enriched scene knowledge graph.

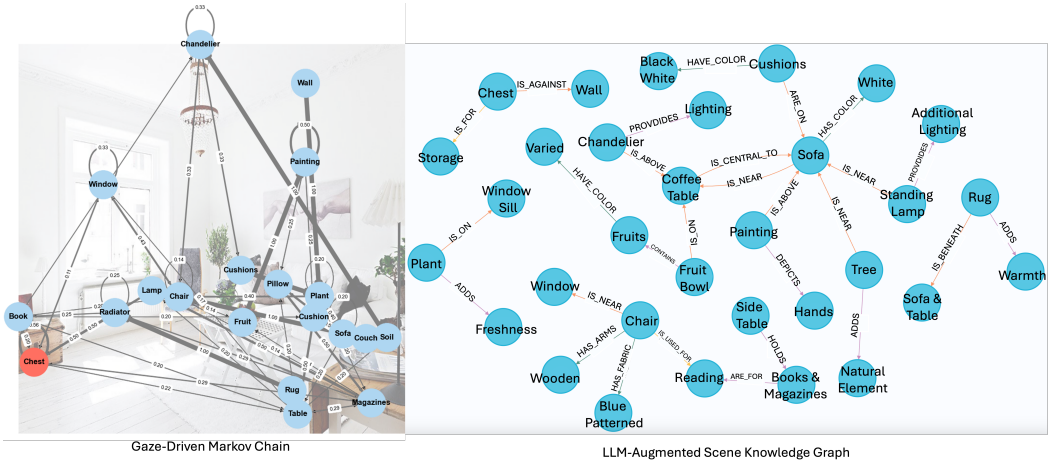

Fig. 4. Example of Scene Analysis with Gaze-Driven Markov Chain and the corresponding LLM-Augmented Scence Knowledge Graph.

3.5 Scene Querying and Response

The knowledge graph generated by the framework serves as a structured basis for investigating human visual search behaviors, explicitly examining whether individuals rely on spatial relationships, semantic relationships, or a combination of both when exploring a scene. Users can pose queries about the scene, such as ”Where are the most likely places to find ’shoes’?”. The LLM answers this question by drawing on the structured information within the knowledge graph and its extensive prior knowledge about typical object arrangements. This combination enables the LLM to produce contextually relevant responses that reflect common patterns of object placement, which are informed by human expectations and scene semantics.

A key advantage of using the LLM’s prior knowledge is its ability to provide insights into semantic associations in human eye movement guidance during visual search. The LLM understands generalized concepts about how people expect objects to be organized in familiar spaces (e.g., shoes are often found near the door or on the floor, cups on tables). This background knowledge allows the system to interpret the spatial structure within the knowledge graph in ways that align with how humans might conceptually approach the scene.

The main advantage of our framework is that it supports a deeper analysis of human visual search strategies by incorporating these semantic associations alongside the structured data from the knowledge graph. It enables an LLM to assess how participants would engage with different scene areas based on physical proximity and meaningful associations between objects. This approach can offer responses to specific spatial queries, even if those associations are not part of the original set of masks. The framework is flexible and allows for inquires about common semantic connections, such as searching near the doorway for shoes or on the table for a cup based on a specific images context, thus incorporating high-level information known to influence viewing strategies in human participants.

4 Experimental Setup

This study aims to determine whether participants exhibit distinct perceptual patterns when searching for target objects in scenes, specifically focusing on ”likely” and ”unlikely” locations. We

used a systematic approach to calculate and evaluate the semantic similarity between different rankings of these locations.

Data Collection. Participants were instructed to search for a pre-specified target object in a scene and to press a response button once located. Target objects included items such as mugs, lamps, and boots, were embedded within photographs of complex real-world scenes depicting indoor environments, which included spaces such as kitchens, living rooms, offices, and bedrooms. The target objects appeared in semantically expected locations (e.g., a mug on a kitchen counter, a shoe on the floor). Targets were fully visible, and varied in size, orientation, and relative positioning across trials. Across all objects and scenes, spatial regions for target placement were defined in accordance with the Surface Guidance Framework first proposed by Pereira and Castelhano [2019] as upper (e.g., top of cabinets, wall shelves and ceiling), middle (e.g., countertops, desktops), and lower regions (e.g., floor, lower portion of the walls). The categories of target objects were selected such that across trials, targets appeared equally across all spatial regions. The definition of spatial placement categories also allowed us to roughly categorize the a priori knowledge of each target location. To record eye movements, participants were first calibrated on the eye tracker using a nine-point calibration screen (average spatial error < .5°). Calibration was checked prior to every trial using a five-point calibration screen. Participants were seated approximately 60 cm in front of the monitor with their heads stabilized by a head and chin rest. The stimuli were displayed on a 21' CRT monitor with a refresh rate of 100 Hz. Scene images subtended a visual angle of 38.1° x 28.6°. Viewing was binocular, however eye movements were sampled only from the right eye at 1000 Hz using a tower mounted EyeLink1000 eye-tracker (S.R. Research Ltd, Canada). For each trial, a target word was presented in the center of the screen for 2s, followed by a fixation cross presented for 500ms. The search scene was then displayed until a button press or 15s had elapsed. Targets were presented either in likely or unlikely locations across two groups of participants, each of who saw the scenes in one of the conditions. This was done to maximize or dampen reliance on a priori knowledge. The experiment took ≈40 min to complete.

Metrics. To quantify the differences between fixation patterns across ranks, we employed Jensen-Shannon Divergence (JSD) as a measure of semantic similarity. JSD was calculated for all combinations of ranks within likely and unlikely groups, as well as across groups. Each JSD calculation aimed to capture the degree of semantic overlap in participants' fixation distributions between ranks. JSD is given by

$$\mathrm{JSD}(P||Q) = \sqrt{\frac{1}{2}\left(D_{\mathrm{KL}}(P||M) + D_{\mathrm{KL}}(Q||M)\right)},$$

where P and Q are the probability distributions of fixation patterns over the objects for each rank, $M = \frac{1}{2}(P + Q)$ is the average distribution, and $D_{\mathrm{KL}}(P||M)$ denotes the Kullback-Leibler divergence between P and M.

Statistical Testing. To evaluate the significance of the observed JSD values, we conducted permutation tests. For each pair of distributions, labels were shuffled between likely and unlikely categories, and a null distribution of JSD values was created from 5,000 permutations. By comparing the observed JSD against this null distribution, we derived p-values that indicate the likelihood of obtaining the observed JSD by chance. This statistical approach allowed us to identify significant differences in fixation patterns across ranks and between likely and unlikely groups, highlighting whether certain locations were consistently perceived as more likely or unlikely for a specific target object.

5 Results

The following section presents the results of our experiments, which evaluate the effectiveness of our gaze-driven, LLM-enhanced framework across several dimensions of scene analysis.

5.1 Experiment 1: Object Detection and Semantic Labels

This experiment evaluated the effectiveness of our gaze-driven, LLM-enhanced approach for object segmentation and labeling, with comparisons to fully-automated baseline models Detectron2 [Wu et al. 2019] and SAM [Kirillov et al. 2023]. Our goal was to assess whether our framework could deliver more humam-centric, contextually relevant and targeted segmentations, reflecting participant fixation patterns and utilizing LLM contextualization to improve labeling consistency. Figure 5 illustrates qualitative segmentation results for three example images across the different approaches.

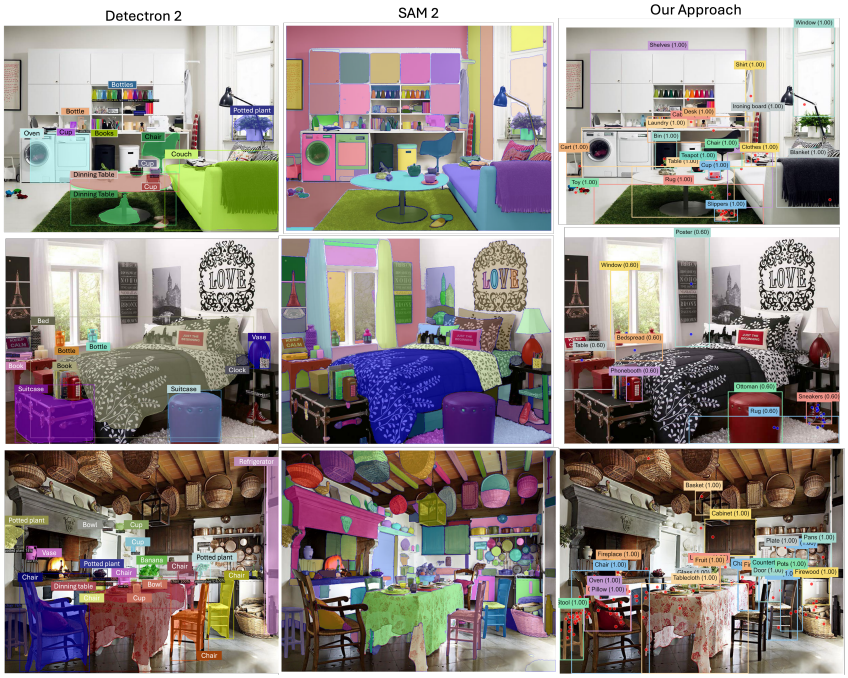

Fig. 5. Qualitative comparison of object detection and labeling using Detectron2, SAM, and our gaze-driven approach.

Detectron2, constrained by its COCO dataset training (80 object classes), struggled to detect many of the objects participants fixated on due to its limited label set. This limitation often led to incomplete segmentations and frequent mismatched masks, underscoring the model's dependency on the limited class diversity in its training set. Thus, Detectron2 is not sufficiently adaptable for comprehensive scene understanding in real-world environments. SAM, while more flexible in its mask generation, produced a high volume of masks per region without accompanying semantic labels, making it challenging to interpret and semantically align each mask with human-relevant objects.

Our framework, by contrast, applies gaze-driven prompts with different geometric representations to capture varying degrees of visual focus, generating segmentations that prioritize areas that

were directly fixated. This approach reduced extraneous masks, enhancing computational efficiency while targeting meaningful regions for segmentation. Additionally, by providing the full scene as context before task-specific mask labeling, the LLM was able to produce labels with improved consistency and semantic relevance. This scene-aware labeling mimics human cognitive processing: much like humans require global scene context to identify smaller objects accurately, our method benefits from this contextual understanding to reflect participant intent more effectively.

5.2 Experiment 2: Quality of the Generated Knowledge Graph

In this experiment, we assessed the quality of knowledge graphs generated by our framework to determine how accurately they captured objects, relationships, and attributes within scenes. To conduct this evaluation, we used 30 images, each evaluated by three independent experts who scored the quality of each triple (object-relationship-object) produced by our model on six criteria. Each criterion was rated on a Likert scale from 1 (Strongly Disagree) to 6 (Strongly Agree), with the option of N/A (Not Applicable) for any criteria deemed irrelevant to specific triples.

The six criteria used for evaluation were as follows: **Object Detection Accuracy** is the accuracy in identifying objects within the scene; **Relationship Accuracy** is the correctness of relational assignments between objects; **Attribute Accuracy** measures the precision in attributing features such as color and shape to objects; **Spatial Accuracy** is the degree to which spatial and positional relationships were accurately captured; **Functional Accuracy** is the accuracy describing functional relationships or purposes of objects within the scene; **Overall Triple Plausibility** corresponds to the perceived plausibility of each triple in reflecting the scene structure.

Table 1. Evaluation of Knowledge Graph Quality Based on Human Ratings Across Multiple Criteria.

	Object Detection	Relationship Accuracy	Attribute Accuracy	Spatial Accuracy	Functional Accuracy	Overall Plausibility
Mean	5.7119	5.6092	5.6667	5.3396	5.7592	5.5319
Std	1.1398	1.2905	1.1547	1.6366	1.0134	1.3615

The results of this evaluation are summarized in Table 1, which shows high scores across most criteria, indicating that our framework successfully generated plausible and accurate triples. Object Detection, Relationship Accuracy, Attribute Accuracy, and Functional Accuracy scored above 5.5 on average, suggesting strong overall performance. While Spatial and Positional Accuracy received a slightly lower mean score of 5.29, reflecting minor spatial misalignments, the framework maintained a robust performance. The high mean scores for Overall Triple Plausibility affirm that the triples were largely perceived as believable and reflective of real-world relationships.

These findings suggest that our framework effectively integrates the LLM's prior knowledge to enhance object segments with spatial and attribute information, yielding a semantically rich knowledge graph. Moreover, the model exhibited minimal hallucinations, underscoring the accuracy and reliability of our generated knowledge graphs.

5.3 Experiment 3: Alignment of Visual Fixations with Semantic Associations

This experiment investigates whether participants' gaze patterns are influenced by semantic relationships in the scene, suggesting that prior knowledge of object associations may guide search strategies. If such an alignment exists, it would imply that participants rely on both spatial and semantic associations between objects during search.

To evaluate this, we mapped each fixation transition in the Markov network to its corresponding semantic relationship in the knowledge graph. This approach allowed us to determine the extent to which participants' visual transitions align with predefined semantic relationships. Figure 6 presents the results of this analysis.

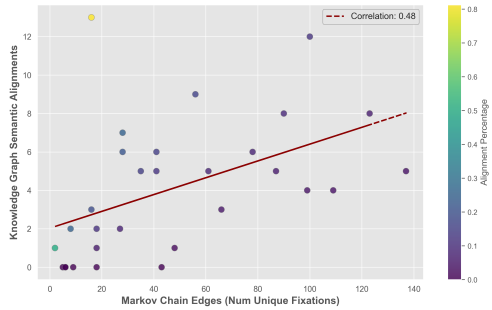

Fig. 6. Relationship between Markov Chain Edges (representing the total number of unique fixation transitions) and Knowledge Graph Semantic Alignments (representing the count of fixations aligning with semantic associations in the knowledge graph).

The results indicate that in more complex scenes—those with a greater number of unique fixation transitions (MC Edges)—participants demonstrated more extensive gaze movements. This behavior likely reflects the need to search through a larger number of elements to identify the target. Conversely, in simpler scenes with fewer visual elements, participants exhibited fewer fixations, indicating they could more quickly locate the target. While a moderate correlation ($r = 0.48$) exists between the number of fixations and semantic alignment, this tends to decrease as scene complexity increases. This suggests that in complex scenes, participants rely less on predefined semantic relationships between objects when searching for the target, possibly due to a broader search strategy that is less semantically-driven. In these cases, participants might prioritize other information such as visual features over functional or relational cues. For simpler scenes, however, there is a higher alignment, meaning participants' gaze patterns more closely follow known spatial or functional associations between objects. This implies that in less complex environments, participants are more likely to rely on contextual or semantic associations in their visual search strategy, allowing for an efficient and targeted approach to locating the target.

It is important to note that the relationships generated by the LLM do not serve as an absolute ground truth but rather as a structured representation of prior knowledge about object relationships. This structured format allows us to analyze whether gaze transitions follow a pattern that aligns with these prior expectations. We acknowledge that semantic relationships can be fluid and context-dependent; however, our aim is not to impose strict classifications but to explore whether general trends in gaze behavior reflect structured relational knowledge.

5.4 Experiment 4: Semantic Differentiation in Visual Search Patterns

This experiment explored the ability of the LLM to semantically distinguish between likely and unlikely locations of target objects within scenes, leveraging its prior knowledge to assign rankings and assess their significance. This analysis provides insights into whether the LLM's rankings align with human search patterns and scene understanding, particularly in differentiating objects and regions based on their of serving as potential target locations. This experiment does not aim to predict object positions but rather investigates how prior knowledge can be computationally structured in a way that reflects human cognitive biases in visual search. Figure 7 presents an example of the distributions generated for the *likely* and *unliekly* locations in the 1*st* rank for the target 'Teddy Bear'.

We measured JSD *within-groups* where we compared the rankings of each *likely* and *unlikely* object group, and we also compared JSD across groups *likely vs. unlikely*. Our experimental results

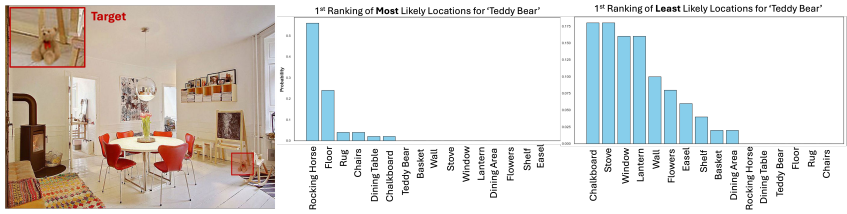

Fig. 7. Distribution of Likely and Unlikely Locations in a Scene Based on Semantic Context. The left plot shows likely target locations based on LLM's prior knowledge, while the right plot highlights less relevant areas for finding the target.

are detailed in Table 2, which provides evidence of the LLM's capacity to separate likely and unlikely locations based on semantic relevance. In the *within-group comparisons*, permutation tests yielded non-significant p-values for comparisons among ranks within the "likely" category, suggesting that the LLM perceived ranks 1, 2, and 3 as having similar semantic relevance to the target object. This uniformity implies that the LLM treated all ranks within the likely category as equivalently plausible. A similar pattern was observed within the "unlikely" category, where the LLM exhibited minimal differentiation between ranks, indicating a uniform perception of low relevance for these locations.

Table 2. Jensen-Shannon Divergence results for within-group and between-group comparisons of *likely* and *unlikely* rankings with statistically significant p-values in green.

	Likely 1st	**Likely 2nd**	**Likely 3rd**	**Unlikely 1st**	**Unlikely 2nd**	**Unlikely 3rd**
Likely 1st	0.000 ± 0.000	0.469 ± 0.128	0.617 ± 0.075	0.806 ± 0.035	0.801 ± 0.039	0.786 ± 0.039
Likely 2nd	0.469 ± 0.128	0.000 ± 0.000	0.386 ± 0.086	0.767 ± 0.070	0.744 ± 0.070	0.717 ± 0.065
Likely 3rd	0.617 ± 0.075	0.386 ± 0.086	0.000 ± 0.000	0.719 ± 0.090	0.689 ± 0.086	0.641 ± 0.078
Unlikely 1st	0.806 ± 0.035	0.767 ± 0.070	0.719 ± 0.090	0.000 ± 0.000	0.382 ± 0.119	0.468 ± 0.133
Unlikely 2nd	0.801 ± 0.039	0.744 ± 0.070	0.689 ± 0.086	0.382 ± 0.119	0.000 ± 0.000	0.365 ± 0.094
Unlikely 3rd	0.786 ± 0.039	0.717 ± 0.065	0.641 ± 0.078	0.468 ± 0.133	0.365 ± 0.094	0.000 ± 0.000

On the other hand, *between-group comparisons* yielded significant p-values for most comparisons between likely and unlikely ranks. This finding highlights a perceptual boundary maintained by the LLM, which consistently perceived objects in the "likely" group as more probable locations for target objects than those in the "unlikely" group. This clear distinction suggests that the LLM categorizes objects with sensitivity to likely semantic contexts. The significant differences observed between *likely* and *unlikely* groups suggest a systematic bias in the LLM's rankings, informed by prior knowledge that reflects an intuitive understanding of scene semantics. Certain objects were consistently regarded as probable locations, while others were dismissed as unlikely, indicating that the LLM effectively mimics human-like cognitive heuristics in assigning semantic relevance. Non-significant results within each group reinforce the LLM's perception of likely and unlikely objects as distinct clusters within the scene, further supporting the framework's alignment with human visual and semantic expectations.

6 Discussion

The results of our experiments support the strengths of combining gaze-driven prompts with LLM-based contextual labeling to produce semantically rich and human-centered representations of scenes. Our findings reveal that leveraging an LLM's prior knowledge, informed by human gaze

data, can achieve nuanced scene understanding that aligns closely with human visual perception and cognitive strategies. Our main findings are the following.

Gaze-Driven LLM Labeling Enhances Contextual Relevance and Targeted Segmentation. Our framework's use of gaze-driven prompts combined with LLM-based contextual labeling produced more accurate and human-centered segmentations than traditional methods. By focusing on regions actively sought out by participants and incorporating the full scene context for labeling, our approach achieved semantically consistent object identification. This process mirrors human cognitive strategies, offering a targeted segmentation solution that reduces computational overhead. A key design decision in our framework was the use of textual scene descriptions to supplement mask labeling. Our initial experiments revealed that when the LLM was given only a segmented mask, its labels were often ambiguous or incorrect. By incorporating a full-scene textual summary, the LLM produced significantly more accurate and meaningful annotations, aligning better with human interpretations.

High Plausibility of Generated Knowledge Graphs. The LLM's ability to distinguish between likely and unlikely object locations demonstrated a strong alignment with human visual search patterns. The significant separation between likely and unlikely groups shows the LLM's ability to assign relevance to scene areas based on prior semantic knowledge reinforcing the model's categorization of likely and unlikely regions as cohesive semantic clusters.

Semantic alignment is stronger in simpler scenes, while complex scenes prompt broader, less context-driven search strategies. In simpler scenes, participants' gaze patterns align more closely with semantic relationships in the knowledge graph, suggesting that participants rely on known object associations to guide their search effectively. This reliance on context and functionality allows for more targeted visual search, making it easier to locate targets in straightforward environments. In contrast, in complex scenes, alignment with semantic associations decreases as participants adopt a broader search strategy, covering more of the scene rather than focusing on specific object relationships. This shift likely results from the higher cognitive load of processing more visual elements, reducing the influence of semantic cues. The moderate correlation ($r = 0.48$) between fixation transitions and semantic alignment indicates that while semantic associations play a role, their influence diminishes as scene complexity increases.

Expected Alignment Between Fixation-Based Transitions and KG Structure. Our framework models scene understanding from a human perspective, so the alignment between fixation-based transitions and the knowledge graph's semantic structure is an expected outcome, not a bias. The KG reflects participants' search patterns, validating our approach. However, this alignment varies: it is stronger in simpler scenes but weaker in complex ones, where participants rely more on exploratory search rather than predefined semantic relationships. This suggests that the KG adapts to different search behaviors rather than being artificially constrained by fixation-based segmentation. While the LLM provides a consistent framework for scene relationships, it does not establish definitive semantic boundaries. Instead, it enables a structured comparison between gaze behavior and expected object associations.

Scalability considerations for real-time and large-scale applications. While our framework effectively integrates gaze-driven segmentation with LLM-based reasoning, its scalability for real-time or large-scale applications presents challenges. The sequential nature of processing fixation-driven segmentation, knowledge graph construction, and LLM inference introduces computational overhead.

Optimizing these processes through parallelization, lightweight LLMs, and efficient segmentation techniques could enhance performance. Additionally, integrating real-time eye-tracking with edge computing may further improve feasibility for applications in augmented reality, robotics, and assistive technologies. Future work will explore these optimizations to balance computational efficiency with interpretability in large-scale deployments.

7 Conclusions

This study introduced a novel framework that combines gaze-driven segmentation with LLM-enhanced scene representation to create a semantically enriched, human-centered knowledge graph. To our knowledge, this is the first framework to integrate these methods for nuanced scene understanding aligned with human visual perception and cognitive strategies. Experimental results demonstrate that the gaze-driven, LLM-labeled segmentation outperforms traditional models like Detectron2 and SAM, producing contextually relevant segmentations and knowledge graphs that capture detailed spatial, functional, and semantic relationships. Our findings also indicate that participants' gaze patterns often align with known object associations, particularly in less complex scenes, where contextual and semantic cues more effectively guide visual search. In more complex scenes, participants may adopt a broader, feature-driven search strategy, relying less on predefined associations. This framework offers a robust tool for advancing human-centered scene analysis, with potential applications in fields such as augmented reality, robotics, human-computer interaction, and cognitive psychology. By aligning computational models with human perceptual patterns and leveraging the semantic capabilities of LLMs, this approach contributes to the development of more contextually aware systems that incorporate both spatial and semantic cues in visual search behaviors.

8 Ethics Statement

This study uses anonymized, securely stored eye-tracking data for gaze-driven scene analysis and knowledge graph generation, with no collection of personally identifiable information. The research aligns with institutional review board guidelines and presents minimal risk of misuse, focusing solely on computational advancements. We acknowledge, however, that future applications of gaze-driven technology may raise privacy concerns and emphasize the importance of ongoing ethical considerations.

9 Code Availability

The prompts utilized for extracting mask labels from LLMs, along with the code underlying our framework and the procedures to replicate our findings, are available in https://github.com/catarina-moreira/human_patterns_exploration.

Acknowledgments

This work was supported by the UNESCO Chair on AI&XR; and the Portuguese *Fundação para a Ciência e a Tecnologia (FCT)* with references DOI:10.54499/UIDB/50021/2020, DOI:10.54499/DL57/2016/CP1368/CT0002 and 2022.09212.PTDC (XAVIER project) to CM and the Natural Sciences and Engineering Research Council of Canada,RGPAS-2018-522460,RGPIN-2018-05166 to MSC.

References

Irving Biederman. 1972. Perceiving Real-World Scenes. *Science* 177, 4043 (July 1972), 77–80. doi:10.1126/science.177.4043.77

Monica S. Castelhano and Chelsea Heaven. 2011. Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. *Psychonomic Bulletin & Review* 18, 5 (Oct. 2011), 890–896. doi:10.3758/s13423-011-0107-8

Monica S. Castelhano and Karolina Krzyś. 2020. Rethinking Space: A Review of Perception, Attention, and Memory in Scene Processing. *Annual Review of Vision Science* 6, 1 (Sept. 2020), 563–586. doi:10.1146/annurev-vision-121219-081745

Monica S Castelhano and Richelle L Witherspoon. 2016. How you use it matters: Object function guides attention during visual search in scenes. *Psychological Science* 27, 5 (2016), 606–621.

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing* 27, 10 (2018), 5142–5154.

Jonathan Harel, Christof Koch, and Pietro Perona. 2006. Graph-Based Visual Saliency. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, Bernhard Schoelkopf, John C. Platt, and Thomas Hofmann (Eds.). Vancouver, British Columbia, Canada, 545–552.

J. M. Henderson. 2003. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences* 7, 11 (2003), 498–504. doi:10.1016/j.tics.2003.09.006

Chihcheng Hsieh, Isabel Blanco Nobre, Sandra Costa Sousa, Chun Ouyang, Margot Brereton, Jacinto C Nascimento, Joaquim Jorge, and Catarina Moreira. 2023. MDF-Net for abnormality detection by fusing X-rays with clinical data. *Scientific Reports* 13, 1 (2023), 15873.

L. Itti and C. Koch. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research* 40, 10–12 (2000), 1489–1506.

L. Itti and C. Koch. 2001. Computational modelling of visual attention. *Nature Reviews Neuroscience* 2, 3 (March 2001), 194–203. doi:10.1038/35058500

Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*. IEEE, 2106–2113.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

Karolina J. Krzyś, Louisa L. Y. Man, Jeffrey D. Wammes, and Monica S. Castelhano. 2024. Foreground bias: Semantic consistency effects modulated when searching across depth. *Psychonomic Bulletin & Review* (May 2024). doi:10.3758/s13423-024-02515-2

Karolina J. Krzyś, Mubeena Mistry, Tyler Q. Yan, and Monica S. Castelhano. 2023. Predicting the Allocation of Attention: Using contextual guidance of eye movements to examine the distribution of attention. In *2023 Symposium on Eye Tracking Research and Applications*. ACM, New York, NY, USA, 1–10. doi:10.1145/3588015.3588405

Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. 2016. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563* (2016).

Catarina Moreira, Diogo Miguel Alvito, Sandra Costa Sousa, Isabel Maria Gomes Blanco Nobre, Chun Ouyang, Regis Kopper, Andrew Duchowski, and Joaquim Jorge. 2023. Comparing visual search patterns in chest x-ray diagnostics. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*. 1–6.

Mark B. Neider and Gregory J. Zelinsky. 2006. Searching for camouflaged targets: effects of target-background similarity on visual search. *Vision Research* 46, 14 (July 2006), 2217–2235. doi:10.1016/j.visres.2006.01.006

José Neves, Chihcheng Hsieh, Isabel Blanco Nobre, Sandra Costa Sousa, Chun Ouyang, Anderson Maciel, Andrew Duchowski, Joaquim Jorge, and Catarina Moreira. 2024. Shedding light on ai in radiology: A systematic review and taxonomy of eye gaze-driven interpretability in deep learning. *European Journal of Radiology* (2024), 111341.

Aude Oliva and Antonio Torralba. 2007. The role of context in object recognition. *Trends in Cognitive Sciences* 11, 12 (Dec. 2007), 520–527. doi:10.1016/j.tics.2007.09.009

Effie J. Pereira and Monica S. Castelhano. 2014. Peripheral guidance in scenes: The interaction of scene context and object content. *Journal of Experimental Psychology: Human Perception and Performance* 40, 5 (Oct. 2014), 2056–2072. doi:10.1037/a0037524

Effie J. Pereira and Monica S. Castelhano. 2019. Attentional capture is contingent on scene region: Using surface guidance framework to explore attentional mechanisms during search. *Psychonomic Bulletin & Review* 26, 4 (Aug. 2019), 1273–1281. doi:10.3758/s13423-019-01610-z

Keith Rayner. 2009. The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology* 62, 8 (2009), 1457–1506.

Antonio Torralba, Aude Oliva, Monica S. Castelhano, and J. M. Henderson. 2006. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review* 113, 4 (2006), 766–786. doi:10.1037/0033-295X.113.4.766

Melissa L.-H. Võ and Jeremy M. Wolfe. 2013. The interplay of episodic and semantic memory in guiding repeated search in scenes. *Cognition* 126, 2 (Feb. 2013), 198–212. doi:10.1016/j.cognition.2012.09.017

Carrick C Williams and Monica S Castelhano. 2019. The changing landscape: High-level influences on eye movement guidance in scenes. *vision* 3, 3 (2019), 33.

J. M. Wolfe, Melissa L.-H. Võ, Karla K. Evans, and Michelle R. Greene. 2011. Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences* 15, 2 (Feb. 2011), 77–84. doi:10.1016/j.tics.2010.12.001

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Gregory Zelinsky, Zhibo Yang, Lihan Huang, Yupei Chen, Seoyoung Ahn, Zijun Wei, Hossein Adeli, Dimitris Samaras, and Minh Hoai. 2019. Benchmarking Gaze Prediction for Categorical Visual Search. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Vol. 2019-June. IEEE, 828–836. doi:10.1109/CVPRW.2019.00111