Distinct roles of central and peripheral vision in rapid scene understanding

Byron A. Johnson Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA, USA

Ansh K. Soni Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA, USA

Shravan Murlidaran Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA, USA

Michael Beyeler Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA, USA
Department of Computer Science, University of California, Santa Barbara, CA, USA

Miguel P. Eckstein Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA, USA
Department of Computer Science, University of California, Santa Barbara, CA, USA

Central and peripheral vision loss, caused by conditions such as age-related macular degeneration and retinitis pigmentosa, disrupt visual processing in distinct ways, yet their impact on real-world scene perception remains poorly understood. Here, we used a real-time, gaze-contingent simulation to examine how central vision loss and peripheral vision loss alter eye movements and scene understanding. Sighted participants ($n = 32$, five males) viewed 120 real-world scenes (50% social interaction, 50% neutral) under one- or three-saccade constraints and described each scene; description quality was quantified via semantic similarity to ground-truth responses. Peripheral vision loss observers produced significantly less informative descriptions than both central vision loss and control participants, particularly for social interaction scenes, suggesting that peripheral vision is critical for rapid extraction of scene semantics. In contrast, central vision loss primarily disrupted oculomotor behavior, including increased saccade amplitudes, delayed saccade initiation, and reduced intersubject fixation consistency. Description quality was not predicted by fixation similarity to controls but by fixations to labeled humans and critical objects, underscoring the role of semantically informative sampling for real-world scenes that include people. These results reveal a dissociation between perceptual and oculomotor consequences of vision loss and highlight the importance of peripheral input for real-world scene understanding.

Introduction

Chronic vision loss alters both perceptual processing and eye movement behavior, yet its impact on real-world scene understanding remains poorly understood. Three leading causes of progressive blindness (age-related macular degeneration, retinitis pigmentosa, and glaucoma) produce bilateral scotomas that impair vision either centrally (central vision loss [CVL]) or peripherally (peripheral vision loss [PVL]) (Shintani, Shechtman, & Gurwood, 2009; Vandersnickt et al., 2024; Verghese, Vullings, & Shanidze, 2021). These conditions disrupt the typical mapping between visual input and retinal location, often leading to altered gaze strategies (Seiple, Rosen, & Garcia, 2013; Janssen & Verghese, 2016; Legge & Chung, 2016; McDonald, Stevenson, Kersten, & Danesh-Meyer, 2022; Vullings, Lively, & Verghese, 2022; Guadron et al., 2023).

Citation: Johnson, B. A., Soni, A. K., Murlidaran, S., Beyeler, M., & Eckstein, M. P. (2026). Distinct roles of central and peripheral vision in rapid scene understanding. *Journal of Vision*, *26*(6):6, 1–35, https://doi.org/10.1167/jov.26.6.6.

https://doi.org/10.1167/jov.26.6.6 Received July 23, 2025; published June 12, 2026 ISSN 1534-7362

Previous studies have assessed the perceptual consequences of low vision using psychophysical tasks such as reading, face identification, and visual search (McIlreavy, Fiser, & Bex, 2012; Geringswald & Pollmann, 2015; Vice, Biles, Maniglia, & Visscher, 2022). Others have examined scene categorization (or "gist") by presenting real-world images to patients with CVL or PVL (Tran, Rambaud, Despretz, & Boucart, 2010; Thibaut, Tran, Szaffarczyk, & Boucart, 2014; Peyrin, Ramanoël, Roux-Sibilon, Chokron, & Hera, 2017). Yet these tasks capture only coarse semantic judgments and fail to reveal how vision loss affects higher-level understanding of real-world content. Early examples of scene perception work involved measuring how scene presentation time interacted with comprehension of global scene context. Biederman and colleagues (1974) presented scenes under two conditions (intact and scrambled) for brief presentations (20–300 ms) and found that participants could select the accurate descriptor among pairs of descriptors for the scene above chance performance with as little as 20 ms. While accuracy was highest for intact scenes, longer presentation times showed a benefit for both scene types, suggesting that rapid scene comprehension was possible even when global scene context was manipulated (Biederman, Rabinowitz, Glass, & Stacy, 1974). Performance for the scrambled scenes was worse due to altered context, which impacted participants' ability to form a *schema* (the visual input, semantic meaning, and reasoning needed to generate a representation) for each scene (Biederman, 1977).

Other experiments have explored how semantic meaning interacts with visual perception. A defining characteristic of real-world scenes is that they can contain meaningful information even when highly irregular. While low-level visual features (i.e., "saliency") can be used to predict where observers will look, meaningful regions of scenes were found to be better predictors for scene-viewing behavior (Henderson & Hayes, 2017; Hayes & Henderson, 2025). Recent experiments suggest that participants were more likely to fixate people and objects that are critical to scene understanding (Murlidaran & Eckstein, 2025). Similarly, Fei-Fei and colleagues (2007) used the "full report" method to measure the reasoning (or inference) component of schema formation, in which participants provided typed descriptions of real-world scenes after brief (27–500 ms) presentations. In a follow-up study using normal and visually similar but highly improbable scenes, they found that descriptions with rich scene interpretation required extracting object relationships, context, and meaning beyond simple identification (Greene, Botros, Beck, & Fei-Fei, 2015). In a different experiment that refined the "full report" method, participant descriptions for a single real-world photo varied based on priming but ultimately focused on relational concepts between the people and objects in the scene (Sanocki, Nguyen, Shultz, & Defant, 2023).

Only a handful of studies have approached the question of how vision loss impacts scene understanding directly. Costela and colleagues found that patients with CVL gave less consistent verbal descriptions of dynamic scenes compared to normally sighted controls, suggesting impaired scene understanding (Costela et al. 2019). However, their study lacked eye-tracking, limiting insight into how visual sampling may have contributed. In a different study recording eye movements, Titchener and colleagues reported that participants with severe PVL were unable to accurately describe a naturalistic video and made fewer shifts in gaze compared to controls and participants with early to moderate PVL, but the study did not measure scene understanding directly (Titchener et al., 2019). More generally, the link between gaze behavior and semantic scene interpretation in vision loss remains unclear. This is relevant for people with low vision, as studies have found that vision loss leads to worse detection of people while walking and can limit participation in social activities (Houston, Bowers, Peli, & Woods, 2018; Klauke, Sondocie, & Fine, 2023). Agent–patient scenes, in which a person or character is interacting with another person or object, have been used to measure rapid extraction of meaning and reasoning (Dobel, Gumnior, Bölte, & Zwitserlood, 2007; Hafri, Papafragou, & Trueswell, 2013; Glanemann, Zwitserlood, Bölte, & Dobel, 2016). The study by Hafri et al. (2013) specifically involved collecting typed descriptions of real-world agent–patient scenes and used participant reports of action phrases as a measure of agreement between descriptions (referred to as a norming study) (Hafri et al., 2013). The use of agent–patient scenes and typed descriptions have allowed for studying the formation of a scene schema in relation to the scene *script*: the natural language representation, or description, of an agent and other entities in the scene (Abelson, 1981).

The Scene Perception and Event Comprehension Theory (SPECT) suggests a cognitive model describing how eye movements and memory interact with time to influence narrative comprehension across various contexts with agent–patient scenes (Loschky, Larson, Smith, & Magliano, 2020). The theory emphasizes how the context of what an observer immediately sees upon scene presentation onset ("front-end" stimulus features) influences how they extract information with subsequent fixations ("back-end" processes), which impacts understanding of the entire scene. The presence of a person in a scene has predictive effects on the observer, known as the "person bias." People tend to look at other people in scenes, have better memory for scenes

that contain people, and follow other people's gaze (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Humphrey & Underwood, 2010; Zwickel & Vo, 2010; Murlidaran & Eckstein, 2025; Karmakar & Eckstein, 2025).

Gaze-contingent displays, a psychophysical method involving dynamic changes in stimulus appearance based on when and where an observer is looking, offer a powerful tool to simulate central or peripheral vision loss in a rapid and controlled manner (McConkie & Rayner, 1975; Reingold, Loschky, McConkie, & Stampe, 2003; Duchowski & Çöltekin, 2007; Kasowski, Johnson, Neydavood, Akkaraju, & Beyeler, 2023). Based on the earlier work by McConkie and Rayner (1975), these paradigms preserve naturalistic viewing while selectively removing information from specific retinal regions. The effects of gaze-contingent displays tested on sighted participants with simulated scotomas located over the fovea or periphery have been studied in different tasks ("gaze-" or "eye-contingent" [Rayner, 1998; Cajar, Schneeweiss, Engbert, & Laubrock, 2016; Foulsham, Teszka, & Kingstone, 2011] or "blindspots" and "windows" [Shioiri & Ikeda, 1989; Henderson & Hollingworth, 1998; Loschky & McConkie, 2002; Nuthmann, 2014; Nuthmann & Malcolm, 2016; Nuthmann & Canas-Bajo, 2022]).

Gaze-contingent simulation experiments can be categorized as measuring sighted participant performance with CVL (McIlreavy et al., 2012; Kwon, Nandy, & Tjan, 2013; Tsank & Eckstein, 2017; Vice et al., 2022; Agaoglu, Fung, & Chung, 2022), PVL (Shioiri & Ikeda, 1989; Loschky & McConkie, 2002; Loschky, McConkie, Yang, & Miller, 2005), or both (Larson & Loschky, 2009; Nuthmann, 2014; Geringswald & Pollmann, 2015; Cajar et al., 2016; Loschky, Szaffarczyk, Beugnet, Young, & Boucart, 2019; Pollmann, Geringswald, Wei, & Porracin, 2020; Nuthmann & Canas-Bajo, 2022; Yu & Kwon, 2023). All of these studies reported variations in saccade amplitudes (distance between the beginning and end of a saccade) and latencies or fixation durations (amount of time needed before making a saccade) with different effects based on scotoma or clear window sizes tested for CVL and PVL, respectively (Nuthmann, 2014; Cajar et al., 2016; Yu & Kwon, 2023). The central scotoma for CVL conditions increased saccade amplitudes, while the clear window and peripheral scotoma for PVL decreased amplitudes, showing how simulated vision loss covering different parts of the retina can have inverse effects on saccade behavior (Shioiri & Ikeda, 1989; Loschky & McConkie, 2002; Larson & Loschky, 2009; Nuthmann, 2014; Cajar et al., 2016). Both CVL and PVL conditions increased latencies and fixation durations, suggesting that more time is needed to plan the next saccade (Loschky et al., 2005; Nuthmann, 2014; Cajar et al., 2016). The critical radii of the scotoma/clear window conditions were found to be 7.4°–8° for scene gist categorization and 1°–3° for reading, suggesting that the level of impairment needed to change task performance could be task dependent (Loschky et al., 2019; Yu & Kwon, 2023). This is also related to changes in resolution across the retina, in which awareness of PVL is more likely to be perceived and lead to longer fixation durations with low spatial frequency filters (Loschky et al., 2005). These effects have been captured across a range of tasks and stimulus types for real-world scenes, including static frames and dynamic videos (Nuthmann & Canas-Bajo, 2022). However, prior studies have not tested simulation methods for the "full-report" scene-understanding task or directly contrasted the effects of central and peripheral loss on semantic interpretation.

To fill this gap, we combined biologically plausible gaze-contingent scotoma simulations with a rapid scene description task involving real-world images. Participants viewed scenes for one or three saccades under simulated CVL, PVL, or control conditions and were asked to describe the scene content. Eye movements were recorded throughout, allowing us to test how visual sampling, fixation patterns, and semantic understanding are affected by different types of visual field loss. Importantly, we wanted to test how scene understanding would be impacted by the presence of people interacting in an agent–patient context compared to scenes used for object recognition. Based on the previous literature, we predict that scene-understanding ability and eye movements will be impaired with CVL and PVL compared to controls. We hypothesize that CVL participants will produce worse descriptions than PVL participants due to removal of detailed information. Alternatively, PVL participants will produce worse descriptions due to the removal of global context, reducing overall comprehension. For eye movements, we predict that CVL participants will have larger saccade amplitudes while PVL participants will have reduced amplitudes. Previous studies suggest that saccade latencies will increase equally for both CVL and PVL participants because of limited information and reduced area for planning their next saccade. In an effort to link scene understanding to eye movement behavior, we expect that scene-understanding ability will be predicted by similar patterns in visual sampling but could be influenced by making eye movements to specific annotated areas of interest (AOIs). We hypothesize that impacts on scene understanding and eye movements will be strongest for social interaction scenes compared to "neutral" object recognition scenes. Our approach bridges the gap between low-level gaze metrics and high-level scene comprehension, offering new insights into how central and peripheral vision support everyday perception.

Methods

Participants

We recruited 32 undergraduate students (5 male, 27 female, mean age of 19.88 years). Participants were recruited from the Psychological and Brain Sciences Department at the University of California, Santa Barbara and completed the study for course credit. All participants provided informed consent. Experimental protocols were approved in accordance with the Institutional Review Board (IRB) of the University of California, Santa Barbara. All participants had normal or corrected-to-normal vision. None reported having any known visual impairments. Any potential subject who could not pass the initial calibration test did not advance to the first trial and was excluded from the experiment.

Apparatus

Python was used to program and run the experiment with scripts developed in-house. We specifically used PsychoPy (Peirce et al., 2019) to present visual stimuli and control the eye tracker. The monitor used was an Alienware AW2524H monitor with 1,280 × 1,024-pixel resolution (31.3° × 24.8°) and a refresh speed of 480 Hz. The experiment was run on a Windows PC with an AMD Ryzen 7 3700X Processor (3.60 GHz). Participants were seated 68.6 cm away from the monitor with a headrest and chair to minimize head movements.

Participants' eyes were tracked monocularly using an SR Research Eyelink 1000 plus Tower Mount (SR Research ltd., Ontario, Canada) eye tracker. If calibration was unsuccessful with the right eye for whatever reason, we recorded from the left eye for that participant for the entirety of the experiment. A velocity threshold of 30° s^{-1} and an acceleration threshold of 9,500° s^{-2} were used to detect saccade events. The sampling rate for the eye tracker was 2000 Hz. Before each block, all participants performed a 9-point calibration test with their normal or corrected-to-normal vision. If two or more broken fixations occurred during the forced fixation portion of each trial, participants were recalibrated.

Stimuli

Participants completed a rapid scene-understanding task. Each image was a real-world scene scaled to 881.2 ± 61.7 × 600 pixels (21.8° × 14.7°). A total of 120 real-world scenes were used:

- "Social interaction" scenes: Half of the scenes were from a novel dataset of photos collected at the University of California, Santa Barbara. In this dataset, a premise for each scene was generated based on one or more actors interacting with one or more objects or people, motivated by agent–patient action scenes (Dobel et al., 2007; Hafriet et al., 2013; Glanemann et al., 2016) and studies that evaluated the effects of people in real-world scenes (Fletcher-Watson et al., 2008; Humphrey & Underwood, 2010). Photographs of people acting as characters for these roles were taken to generate a dataset of social interaction photos. These scenes could also contain actors in the background. There were one to nine people included in each scene, with most (95%) scenes having at least two people. Example "social interaction" scenes are shown in the top row of Figure 1. A list of action phrases for each scene can be found in Appendix Table A1.1.
- "Neutral" scenes: The other half of the scenes was pseudo-randomly selected from the publicly available Microsoft Common Objects in Context (MSCOCO) image dataset (Lin et al., 2014) used for machine learning and object recognition tasks. While half of the MSCOCO images contained at least one person ($n = 30$), others only contained empty rooms or animals. Importantly, the MSCOCO dataset was specifically designed for benchmarking computer vision model performance for object recognition. The bottom row of Figure 1 shows example "neutral" scenes.

To explore the effects of low-vision viewing conditions on scene understanding, participants were assigned to one of three conditions: none (control, $n = 10$), PVL ($n = 11$), and CVL ($n = 11$). One participant from the control group was unable to finish the study, so they were excluded from the dataset. Participants viewed every scene in their assigned condition (e.g., between-subjects design). Participants were not aware of the experimental hypotheses and were not explicitly told the details of their viewing condition.

To simulate PVL, concentric blur was applied to the image, leaving a clear window with a radius of 5°. While previous simulations have used a gray mask (Yu & Kwon, 2023; Vice et al., 2022), we used Gaussian blur from OpenCV (Bradski, 2000) for the simulated scotomas. This is because previous work with patients has shown that they are not always aware of the presence or location of their scotoma (Hartong, Berson, & Dryja, 2006; Fletcher, Schuchard, & Renninger, 2012; Peli, Goldstein, & Jung, 2023). A solid filled mask has sharper edges that are more salient against a blank background, while Gaussian blur can make it more difficult for participants to detect the edges of the scotoma layered on a real-world scene. The blur used in this study had a kernel shape of approximately 2.45°

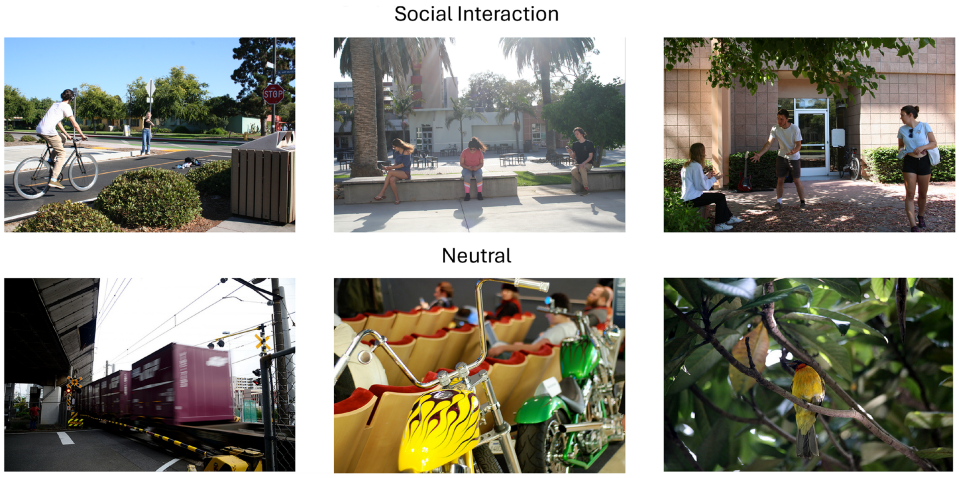

Figure 1. Representative images from the two scene categories used in the study. The top row shows social interaction scenes characterized by human presence and observable social behavior. The bottom row includes neutral scenes from the MSCOCO dataset, containing inanimate objects or animals without overt human interaction.

by 2.45° (99 by 99 pixels) with a standard deviation of 2.47° (100 pixels). For CVL, an inner circular patch was blurred with the same Gaussian blur used for PVL. The size of the central scotoma matched the size of the clear window for the PVL condition (5°). This size was based on previous studies that simulated CVL for a visual search task (Kwon et al., 2013; Geringswald & Pollmann, 2015; Cajar et al., 2016). Figure 2 shows an example scene of all three viewing conditions.

The display updated based on real-time recordings from the eye tracker to make the PVL and CVL simulations gaze-contingent. Before each trial, the stimulus was preprocessed with the scotoma area filtered at different locations of the image by subsampling every 40 pixels (1°) of the scene in order to present the simulation with minimal lag. This was done repeatedly on a participant-by-participant basis with no more than 20 s for loading each scene. For methods that precomputed scene data offline, see Loschky & McConkie (2000), Loschky & McConkie (2002), and Loschky et al. (2005). For all participants, the eye tracker indicated where the fovea (central scotoma in CVL, peripheral scotoma in PVL) is relative to the image. Once the scene appeared, the center of the clear window or central scotoma appeared on the screen at the participants' current fixation location and moved as participants made eye movements. The average delay between the eye tracker and the experiment monitor was 6.89 ± 1.21 ms (see Figure A1.1).

Procedure

Scene presentation

An overview of the experiment can be seen in Figure 3. The experiment was broken up into 12 blocks with 10 trials each, resulting in 120 trials. Before the experiment, scenes were randomly assigned to each block. While every participant had the same block sequence, the order of scenes within the block was randomized for each participant. The experiment was programmed to close at the end of each block.

Figure 2. Simulated viewing conditions for a representative social interaction scene. The left panel shows the original image as seen under unimpaired vision. The center panel simulates central vision loss (CVL), with a central scotoma spanning 5° of visual angle, obscuring the person and backpack on the right. The right panel simulates peripheral vision loss (PVL), with a 5° clear central window, obscuring the approaching biker on the left. The full scene subtends 21.8° × 14.7° of visual angle.

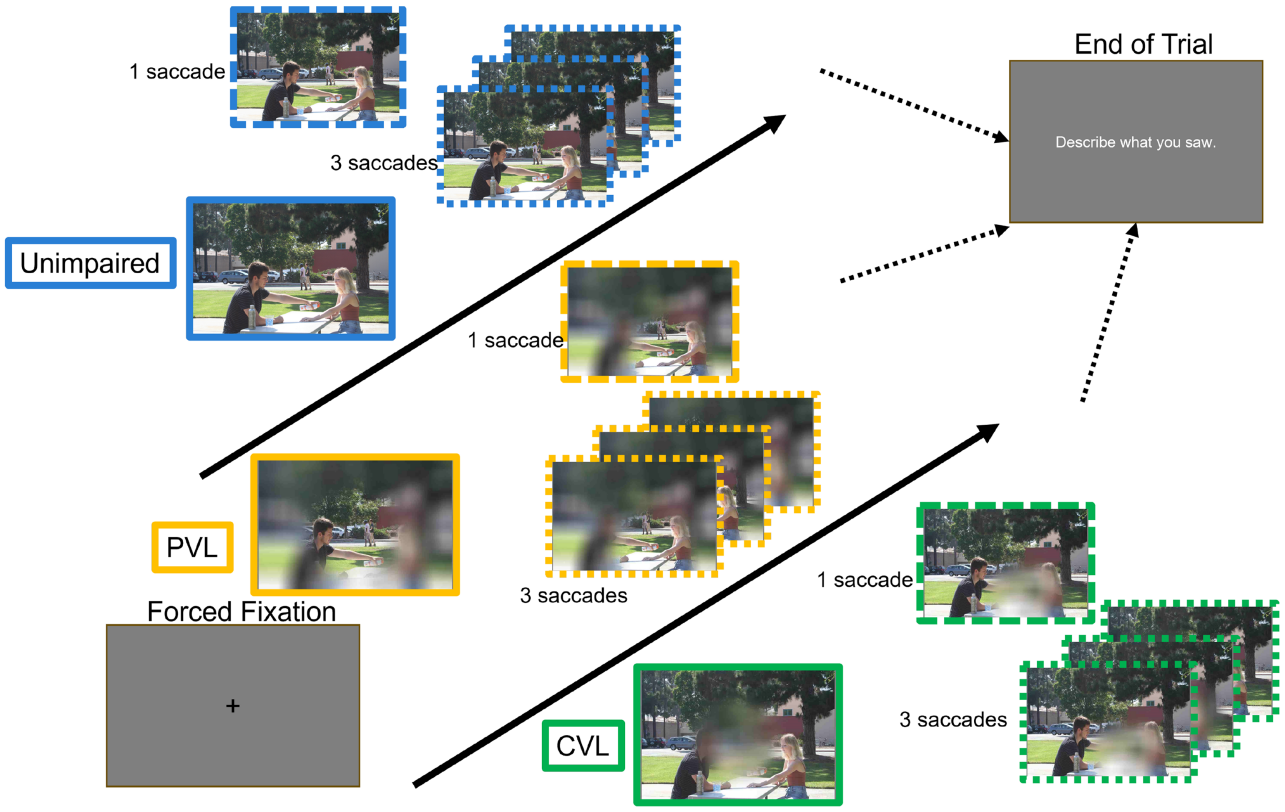

Figure 3. Overview of the experimental procedure. Each trial began with a central fixation cross. Participants ($n = 32$) viewed a natural scene under one of three simulated viewing conditions: unimpaired, peripheral vision loss (PVL), or central vision loss (CVL). Depending on the trial, participants were allowed either one or three saccades before the scene disappeared (randomized). They were then prompted to describe what they saw. No feedback was given. The same set of scenes was used across conditions and saccade limits, with presentation order counterbalanced across participants.

Participants were allowed to take as many breaks as they requested throughout the experiment. Experimental sessions were 1 to 2 hours in length and took place over 2 or more days.

Participants completed a rapid scene-understanding task viewing all stimuli in their assigned simulation condition. After successful calibration, participants were shown a fixation cross in the middle of the screen. To start the trial, participants had to press the "space" key while looking at the fixation cross for a random amount of time selected from a uniform distribution between 250 and 750 ms. The purpose of altering the fixation cross duration was to prevent participants from knowing when exactly the scene would appear and to prevent participants from making eye movements during the central fixation. If the participant moved their eyes away from the fixation cross for any reason, a "broken fixation" message would appear. Participants were recalibrated after frequent broken fixations. Participant data were automatically saved after each trial. If the participant had to take a break or recalibrate between trials, they would exit the program. Upon restarting, they would recalibrate and continue with the next trial in sequence.

If the participant fixated the central cross successfully (250 to 750 ms), the scene would appear. The amount of time for the stimulus presentation was dependent on the number of saccades made by the participant while viewing the scene. Before the experiment, half of the scenes were randomly assigned to allow for one saccade or three saccades. If a scene was assigned to allow for one saccade, upon stimulus onset, participants could make one saccade, starting from the initial central fixation upon scene onset to their second fixation. If a scene was assigned to allow for three saccades, participants could make four fixations (initial central fixation at scene onset until completion of the fourth fixation). The simulated scotoma would update based on the participant's current fixation location. Once the saccade limit was reached, upon the start of the next saccade, the scene disappeared and a gray screen was shown. If the saccade limit was not reached, the total maximum scene presentation time was 10 s. This fast presentation for scene perception is referred to as "rapid scene understanding."

Response

After the scene disappeared, participants were instructed to type a description of the scene in English. There was no strict length limit to their response, but participants were encouraged to keep their descriptions shorter than four sentences. Text would appear on the screen as participants typed. No other feedback was provided about the quality of their descriptions.

Verbal instructions for the scene description were given at the beginning of the experiment. The following instructions were presented on screen at the beginning of the experiment and every block:

1. Double check descriptions for errors before submitting.
2. Use complete sentences only. No one-word sentences.
3. Do not use proper nouns or names.
4. Gender and pronouns can be used.
5. Start your description with "There was a...."
6. Only describe what you saw. Do not put "unclear" or "could not see it."

Metrics

Scene descriptions

The semantic similarity between a ground-truth description of the scene and each participant response was computed as follows:

- Generating ground-truth descriptions of each scene: To generate ground-truth descriptions of each scene, a different group of five participants (five female, mean age of 20.2 years) was given the same scene task with unlimited viewing time and no simulated impairment. This provided five sets of ground-truth descriptions to be used to evaluate the semantic similarity of responses from the rapid scene-understanding task. To evaluate the upper limits of the ground-truth descriptions, a separate group of five participants (five female, mean age of 19.8 years) compared four sets of ground-truth descriptions for all scenes to one set of ground-truth descriptions. One ground-truth description set was duplicated, randomly assigned to scenes, and also included to test the lower limit of the ground-truth descriptions. All of the descriptions used to calculate the lower bound were randomly assigned to a different scene (none matched the original scene). Additionally, three descriptions from the main experiment (one control, one CVL, and one PVL description) were included to provide a range of responses when comparing the ground-truth descriptions (totaling eight descriptions to compare with one ground-truth). Participants rated the semantic similarity between each description and the ground-truth description on a scale of 1 to 10, with 1 meaning no similarity and 10 meaning high similarity. Raters did not see the descriptions with the scene image.
- Rating participant descriptions: Two measures were used to evaluate the quality of participant descriptions. The following measures are specific to the saccade-limited descriptions in the main experiment and were done separately from the upper- and lower-bound analyses.
  - The first measure is based on human ratings of descriptions and was inspired by work that examined how humans judge the similarity of pairs of words and text (Dobel et al., 2007; Hafri et al., 2013; Nguyen, Trieschnigg, & Theune, 2014; Wang et al., 2023). A third group of four participants (one male, mean age of 20 years) compared every response to the ground-truth description for each scene. Raters saw all saccade-limited participant descriptions for one scene at the same time, presented in a randomized order to prevent expectation of description quality. Every response was rated on a scale of 1 to 10, where 1 meant that the response did not semantically match the ground-truth description and 10 meant that the response semantically matched the ground-truth description. Each rater ($n = 4$) provided a score for every participant response set ($n = 32$) for all scenes ($n = 120$), totaling 15,360 ratings. The average rating between the four raters provided 3,840 measures of semantic similarity to the ground-truth descriptions. Human ratings were generated from the descriptions only (the scene image was not provided to the raters). Importantly, raters did not see the scene image, and only one ground-truth description set was used for the human ratings. For example, each rater would view one among all saccade-limited descriptions, such as "There is a boy and a girl. They are both sitting on a long chair with a gray colored table surface. The boy is pouring an orange drink in a cup." with the ground-truth description "A man and a woman are sitting on a picnic table. The man is spraying sunscreen on her arm. There are two other people walking together in the background." If each of the four raters scored the saccade-limited description as 8, 4, 6, and 7, respectively, then the average rating would be 6. To determine variability in ratings, interrater reliability and Pearson's correlations were calculated between raters. Interrater reliability calculates Cohen's kappa between each rater to determine the level of agreement on categorical data (scores 1–10) as:

$$\kappa \equiv \frac{p_o - p_e}{1 - p_o} = 1 - \frac{1 - p_o}{1 - p_e}, \qquad (1)$$

where p_o is the agreement between two raters, and p_e is hypothetical chance agreement between the 10 scores, resulting in $\kappa \in [0, 1]$, with 0 meaning no agreement as predicted by chance and 1 meaning complete agreement.

- The second measure for analyzing participant responses generates a numerical representation of each typed description (referred to as embeddings) based on Chat Generative Pre-Trained Transformer (GPT)–4 (Open AI, Inc., San Francisco, California, United States). GPT-4 can be used to compare text in an embedding space. Text embeddings are numerical representations of text transformed into vectors (long sequences of numbers) within a high-dimensional space. They are designed to capture the semantic meaning and context of the text, rather than just its literal, word-for-word structure. These embeddings can be used to compute cosine similarities between pairs of ground-truth descriptions and participant responses for one scene. GPT-4 is popular for its impressive language reasoning abilities based on large language models (LLMs) and can be used to quantify potential differences in participant responses. We used GPT-4 to generate scores of semantic similarity between ground-truth descriptions as candidate text and participant responses as reference sentences:

$$\text{Cosine similarity} = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}||||\mathbf{B}||} = \frac{\sum_{i=1}^{n} A_i B_{i,j}}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_{i,j}^2}} \quad (2)$$

where, for each scene i and each participant response j, the candidate ground-truth description A_i is paired with the participant response for that scene $B_{i,j}$. The dot product of the components of each sentence is normalized by the magnitude of both descriptions combined in Euclidean space, providing the cosine similarity score. This value can range from −1 to 1, where −1 represents high dissimilarity, and 1 represents high similarity.

Eye movement metrics

Saccade amplitude is the third measure and was calculated as the distance between the start of a saccade (end of a fixation) and the end of a saccade (start of a fixation). The fourth measure, saccade latency, was defined as the total amount of time between the scene appearing and the participant making a saccade. This was calculated for every trial and every recorded eye movement. If more than one saccade was allowed to view the scene, the latency was measured as the time between the end of the previous saccade and the beginning of the next saccade (i.e., fixation dwell time).

Fixation heat maps

Analysis of participant eye movement behavior was based on five metrics to capture the spatial and temporal features of saccades. Distributions of participant fixation locations for a scene were measured by transforming fixation locations for every scene into separate heat maps for each viewing condition. To assess how fixations varied across viewing conditions (controls, PVL, and CVL). We compared fixation heat map correlations across viewing conditions to groups of different subjects in the control condition. Each heat map was generated by randomly sampling subjects (n = 4) into four groups for testing: one group of controls ("control comparability group"), a different group of controls to compare to the control comparability group, a group of PVL, and a group of CVL. All fixations recorded during the stimulus presentation were convolved with a Gaussian filter (39 × 39 pixels, or approximately 1° × 1°; SD = 1; 60 bins) using the Astropy package (The Astropy Collaboration et al., 2013), providing a spatial representation of where participants fixated during each trial. Each heat map has a color map representing fixation density, where red indicates a high number of fixations relative to the rest of the map. Correlations between each control comparability group heat map and each viewing condition fixation heat map were computed for every scene.

Scene segmentation and frequency of fixations to annotated areas of interest

Since we used real-world scenes containing people and objects, we wanted to understand if fixations to specific objects were affected by the viewing condition. The second measure, based on eye movement behavior, used AOIs to determine how often participants fixated objects. One annotator (one male, 19 years old) had unlimited time to outline faces and bodies of humans and animals, as well as critical objects for all scenes. The critical object was defined as the most important object (or group of objects) needed to accurately describe the scene. Each critical object was strongly inferred from comparing the five ground-truth description sentences to each scene photograph and selecting the object most relevant for an accurate description. The AOIs were created using Make Sense (Skalski, 2019) and saved into a .JSON file. Fixation locations separated by viewing condition and AOIs were layered onto each scene, then the number of fixations located within a region was counted. Examples of labeled social interaction and neutral scenes are in Figure A1.10.

Given that 61 of the 120 scenes presented (45 social interaction and 16 neutral) contained people located in the periphery (greater than 5 degrees visual angle) with respect to the fovea at the start of the trial, we classified each scene (ad hoc) with a binary label to determine

if the effects of viewing condition on descriptions were affected by peripheral arrangements of people. Figure A1.6 shows an example scene classified as having a person located in the periphery upon stimulus onset. We repeated this labeling process for critical objects in the scenes. Thirty-three of 120 scenes (21 social interaction, 12 neutral) were labeled as including a critical object located in the periphery of the scene; the remaining scenes had critical objects located in the center.

Results

Effects of simulated vision loss on semantic scene understanding

To measure how ratings of descriptions of real-world scenes would differ between viewing conditions, a linear mixed-model lme4 (Bates, Mächler, Bolker, & Walker, 2015) was used to assess the variance in participant descriptions. Separate tests were done for human description ratings and GPT-4 cosine similarity values as the outcome variable. Viewing condition (none, PVL, CVL), scene type (social interaction or neutral), and number of saccades allowed (one or three) were included as fixed effects. Participants' unique identifiers and unique scene identities were included as random factors to account for variability in semantic similarity across subjects. Post hoc comparisons were performed using Šidák correction to test for pairwise differences between conditions.

To test for potential learning effects, we compared human ratings and GPT-based cosine similarities across the first and last experimental blocks (Figure A1.3). A two-way analysis of variance (ANOVA) found no effect of block number for human ratings ($F(1, 6) = 1.27$, $p = 0.303$) but did show an effect for GPT-4 cosine similarity values ($F(1, 6) = 25.11$, $p = 0.002$). Post hoc analysis for GPT-4 cosine similarity values revealed that control participant descriptions for social interactions scenes in the first block were significantly lower than control participant descriptions for neutral scenes in the last block ($M = -0.25$, $p = 0.03$, $d = 0.26$), but no other comparisons were significant, suggesting that there was no learning between conditions.

Human ratings of semantic similarity

Figures 4A, 4B shows how participant descriptions varied by viewing condition (control, PVL, CVL), scene type (social or neutral), and number of saccades allowed (one or three). For human ratings of ground-truth social interaction scene descriptions, the upper bound was $M = 7.21 \pm 0.12$, and the lower bound was $M = 1.74 \pm 0.12$. For neutral scene ground-truth descriptions, the upper bound was $M = 6.92 \pm 0.13$, and the lower bound was $M = 1.17 \pm 0.09$ (upper and lower dashed blue lines and shaded areas in each figure). A linear mixed-effects model was used to predict average human similarity ratings aggregated across four raters for each trial (see Appendix for rating distribution and interrater reliability characteristics). Fixed effects included viewing condition, scene type, and saccade number, and random intercepts were included for both participant and scene. The model using unique subject and scene identifiers as random intercepts significantly outperformed a subject intercept-only baseline ($\chi^2(1) = 1{,}465.95$, $p < 0.001$; conditional $R^2 = 0.37$, marginal $R^2 = 0.07$).

A three-way ANOVA on the fitted model revealed a significant interaction between viewing condition, scene type, and number of saccades ($F(2, 3{,}712) = 3.99$, $p = 0.019$), indicating that the quality of scene descriptions depended jointly on visual loss type, content type, and allowed viewing time. There was a significant interaction between viewing condition and scene type for the human ratings ($F(2, 3{,}712) = 6.97$, $p = 0.001$), but no other interactions with saccades allowed were significant. Analyses revealed significant main effects for viewing condition ($F(2, 3{,}712) = 43.23$, $p < 0.001$), scene type ($F(1, 116) = 16.12$, $p < 0.001$), and number of saccades allowed ($F(1, 116) = 7.89$, $p = 0.006$) on human ratings for descriptions.

To validate assumptions for the model, we first tested for homogeneity of variance across groups using Levene's test, which indicated unequal variances ($F(2, 3{,}837) = 8.83$, $p < 0.001$). A Welch's ANOVA confirmed a significant difference between at least one pair of group means ($F(2{,}2553.84) = 35.48$, $p < 0.001$). Inspection of residuals revealed no substantial relationship between fitted values and residual variance, supporting the use of mixed-effects modeling for subsequent analyses.

To interpret the significant three-way interaction, we conducted Šidák-corrected pairwise comparisons. For social scenes with only one saccade allowed, both PVL and CVL groups produced significantly lower similarity ratings than controls (PVL: $M = -0.52$, $p < 0.001$, $d = 0.52$; CVL: $M = -0.21$, $p = 0.02$, $d = 0.22$), and PVL ratings were also significantly lower than CVL ($M = -0.31$, $p < 0.001$, $d = 0.30$). With three saccades, both impaired groups remained significantly lower than controls (PVL: $M = -0.39$, $p < 0.001$, $d = 0.39$; CVL: $M = -0.37$, $p < 0.001$, $d = 0.37$), with no difference between PVL and CVL ($M = 0.03$, $p = 1$). For neutral scenes, the only significant difference emerged when three saccades were allowed: the PVL group rated significantly lower than controls ($M = -0.28$, $p < 0.001$, $d = 0.28$). No other pairwise differences were significant for neutral scenes (p values = 0.28–1). These results are consistent with the hypothesis predicting

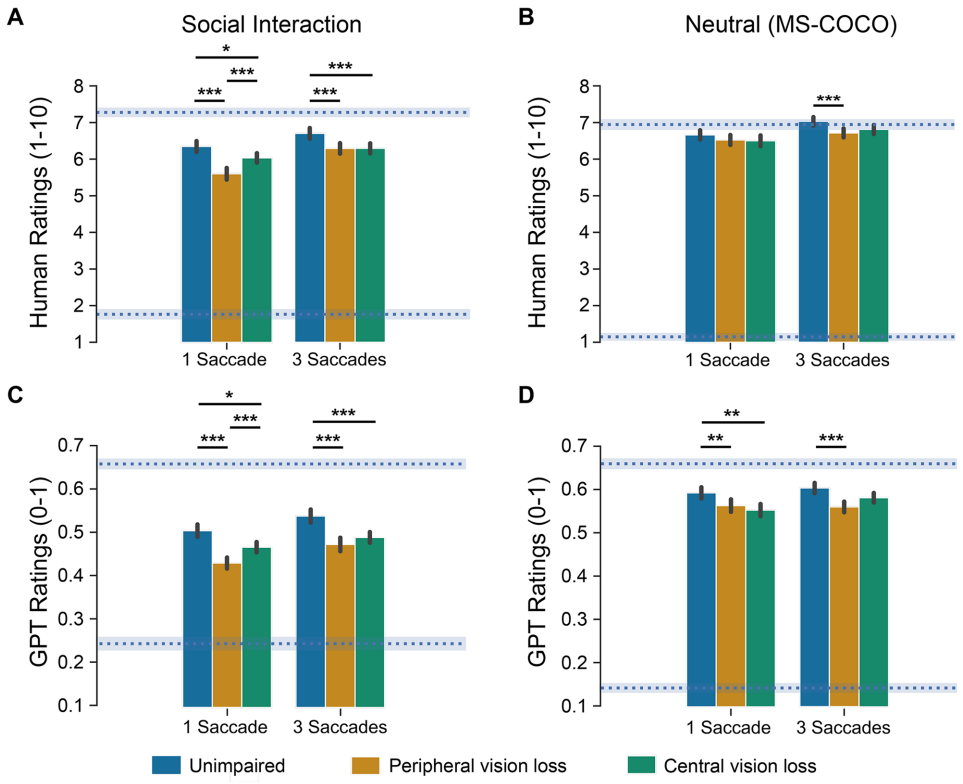

Figure 4. Semantic similarity scores for participant descriptions across viewing conditions, scene types, and number of saccades allowed. (A, B) Mean human ratings of description quality. (C, D) GPT-4 cosine similarity scores comparing participant descriptions to ground-truth. For social interaction scenes with one saccade, PVL descriptions were rated significantly lower than both controls and CVL by human raters (A) and received lower GPT-4 scores (C). With three saccades, both PVL and CVL descriptions were rated lower than controls for social scenes. For neutral scenes, PVL descriptions were rated lower than controls by human raters with one saccade (B) and by GPT-4 with either one or three saccades (D). Horizontal dashed blue lines (and shaded areas) denote the range of ground-truth scores for each scene type. Human rating ground-truths: social = 7.21 ± 0.12 (upper), 1.74 ± 0.12 (lower); neutral = 6.92 ± 0.13/1.17 ± 0.09. GPT-4 scores: social = 0.65 ± 0.01/0.25 ± 0.01; neutral = 0.65 ± 0.01/0.17 ± 0.01.

the negative impact of CVL and PVL on scene understanding, with PVL performing worse than CVL and controls for social interaction scenes with one saccade allowed and neutral scenes with three saccades allowed.

LLM-based ratings using GPT-4 embeddings

To complement the human ratings, we computed semantic similarity scores using GPT-4 sentence embeddings (Figures 4C, 4D). Cosine similarity was calculated to compare the ground-truth descriptions. For social interaction scenes, the upper bound was $M = 0.65 \pm 0.01$, and the lower bound was $M = 0.25 \pm 0.01$. For neutral scene, the upper bound was $M = 0.65 \pm 0.01$, and the lower bound was $M = 0.17 \pm 0.01$. For each participant description, we computed the cosine similarity to the corresponding ground-truth description. These model-based scores were analyzed using the same linear mixed-effects model structure as the human ratings, with fixed effects for viewing condition, scene type, and number of saccades and random intercepts for participant and scene.

The model revealed a significant three-way interaction between viewing condition, scene type, and number of saccades ($F(2, 3{,}712) = 4.77$, $p = 0.009$), consistent with the human rating results. There was a significant interaction between viewing condition and scene type for the GPT-4 cosine similarities ($F(2, 3{,}712) = 14.26$, $p < 0.001$), but no other interactions with saccades allowed were significant. Analyses revealed significant main effects for viewing condition ($F(2, 3{,}712) = 86.07$, $p < 0.001$) and scene type ($F(1, 116) = 53.02$, $p < 0.001$) for cosine similarity values, but number of saccades allowed was not significant ($p > 0.05$). GPT-based similarity scores were significantly correlated with average human ratings across trials ($r = 0.73$, $p < 0.001$), indicating convergent validity between the two approaches.

Šidák-corrected pairwise comparisons revealed that, for social scenes with one saccade, both PVL and CVL groups produced significantly lower cosine similarity

scores than controls (PVL: $M = -0.06$, $p < 0.001$, $d = 0.83$; CVL: $M = -0.02$, $p < 0.001$, $d = 0.29$), with PVL lower than CVL ($M = -0.04$, $p < 0.001$, $d = 0.54$). When three saccades were allowed, both vision loss groups again differed from controls (PVL: $M = -0.07$, $p < 0.001$, $d = 0.92$; CVL: $M = -0.05$, $p < 0.001$, $d = 0.66$), with no difference between PVL and CVL ($M = 0$, $p = 1$).

For neutral scenes with one saccade, both PVL and CVL groups scored lower than controls (PVL: $M = -0.03$, $p = 0.001$, $d = 0.38$; CVL: $M = -0.03$, $p = 0.001$, $d = 0.37$) but did not differ from each other ($M = 0$, $p = 1$). With three saccades, only PVL differed significantly from controls ($M = -0.03$, $p < 0.001$, $d = 0.39$); the difference between PVL and CVL was not significant ($M = -0.01$, $p = 0.30$). The GPT-4 cosine similarity results are consistent with the hypothesis that scene-understanding ability will be negatively impacted due to the CVL and PVL conditions, and post hoc analysis shows a strong alignment with human ratings.

Effects on eye movements

Two participants were excluded from the eye movement analyses due to missing data, leaving 30 participants (10 control, 11 PVL, 9 CVL) included in all analyses reported below.

Saccade amplitude and latency

To determine how eye movements would differ between viewing conditions, we analyzed the effect of viewing condition on saccade amplitude and latency using a linear mixed-effects model. Prior work suggests that individuals with CVL tend to make varied saccades to compensate for their scotoma (Seiple et al., 2013; Vullings et al., 2022), while those with PVL or CVL may exhibit delayed saccades due to the scotoma (Van Der Stigchel et al., 2013; Guadron et al., 2023). Participants' first saccade amplitudes and first saccade latencies were included as outcome measures for two different *F*-tests to assess whether these basic eye movement metrics varied across viewing conditions. Separate ANOVAs were done to test viewing condition as the only main effect and then test interactions between viewing condition, scene type, and saccades allowed as fixed effects. Participant and scene identities were included as random effects for each test. Šidák correction for multiple comparisons was used to measure differences between viewing conditions.

Viewing condition was found to have a significant interaction with scene type for saccade amplitude ($F(2, 3{,}472) = 3.20$, $p < 0.04$). Control amplitudes for social interaction scenes were significantly larger than neutral scenes ($M = 0.37$ dva, $p = 0.03$, $d = 0.29$), but there were no significant differences in amplitudes between scene type for CVL ($p > 0.05$) and PVL ($p > 0.05$).

As shown in Figure 5A, a separate ANOVA showed that viewing condition was a significant main effect for saccade amplitude ($F(2, 3{,}478) = 160.71$, $p < 0.001$). Participants with CVL made larger first saccades than both controls ($M = -1.06$ dva, $p < 0.001$, $d = 0.65$) and PVL ($M = -1.20$ dva, $p < 0.001$, $d = 0.73$), with no difference between controls and PVL ($M = 0.13$ dva, $p = 0.56$). Histograms of distributions of saccade amplitudes by viewing condition and the differences between distributions based on PVL and CVL are shown in Figure A1.4. These results are consistent with the prediction that amplitudes would increase for CVL. PVL amplitudes were not significantly smaller than controls, which is inconsistent with our initial hypothesis. The lack of interaction between CVL and PVL amplitudes and scene type suggests that the magnitude of where participants move their eyes upon scene appearance is independent of scene context.

Figure 5B shows that saccade latency also varied by condition ($F(2, 3{,}478) = 125.28$, $p < 0.001$): Both PVL and CVL participants were slower than controls (PVL: $M = +0.06$ s, $p < 0.001$, $d = 0.71$; CVL: $M = +0.05$ s, $p < 0.001$, $d = 0.63$), with no difference between the vision loss groups ($M = 0.01$ s, $p = 0.34$). A separate ANOVA revealed that scene type was found to have a significant main effect on saccade latency ($F(1, 116) = 7.76$, $p = 0.006$). Post hoc analyses showed that PVL latencies were be shorter for social interaction scenes compared to neutral scenes ($M = -0.03$ s, $p = 0.003$, $d = 0.34$) but not for control or CVL participants ($p > 0.05$. These results are consistent with the hypothesis predicting longer saccade latencies for both CVL and PVL, with PVL participants having faster latencies for social interaction scenes compared to neutral scenes.

Finally, we assessed whether amplitude and latency were correlated (Figures 5C, 5D). Across all scenes and participants, these measures were modestly correlated ($r = 0.17$, $p = 0.001$), driven primarily by the CVL group ($r = 0.26$, $p = 0.003$). The relationship held for social scenes ($r = 0.30$, $p < 0.001$) but not neutral scenes ($r = 0.09$, $p = 0.23$) and was strongest for controls ($r = 0.41$, $p = 0.001$) and CVL participants ($r = 0.39$, $p = 0.001$) viewing social scenes.

Fixation heat maps

To assess how visual sampling varied across viewing conditions, we compared heat map similarities using a linear mixed-effects model. Viewing condition, scene type, and number of saccades allowed were included as fixed effects with unique scene identifiers as a random effect. Heat maps were generated by averaging data from multiple participants in each assigned viewing condition, so participant identifiers were not included as a random effect for heat maps analyses. We

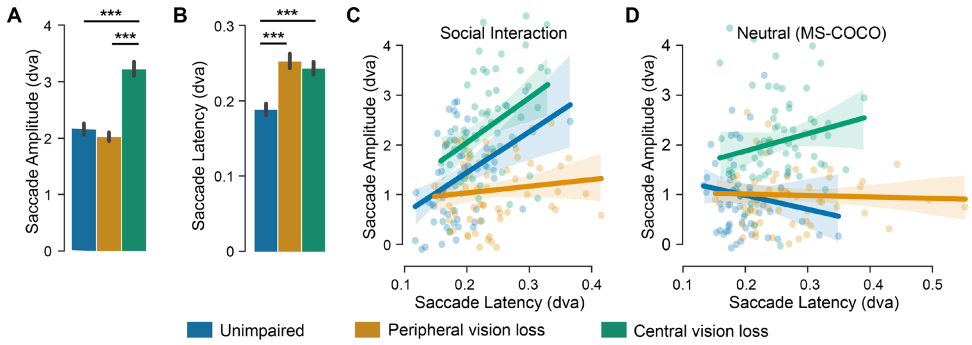

Figure 5. First saccade amplitude and latency. (A) Participants in the CVL condition made significantly larger first saccades than both unimpaired and PVL participants. No significant difference was observed between unimpaired and PVL participants. (B) First saccade latency was significantly longer for both PVL and CVL participants compared to unimpaired controls. There was no significant difference in latency between PVL and CVL conditions. (C, D) Relationship between first saccade amplitude and latency, separated by scene type. A significant correlation was observed for social interaction scenes (C) but not for neutral scenes (D).

calculated a large number of heat map correlations between the three experimental viewing conditions and a control comparability group for comparison. We used a resampling (aka bootstrap) approach in which we randomly resampled the heat maps from groups of four participants from each viewing condition for a given image. For each iteration, we computed the correlation between the heat maps for each viewing condition and the control comparability group. This procedure was repeated 100 times for a given scene to generate a large distribution of correlations and calculate a stable average correlation for each viewing condition and scene. Fixation locations were convolved with a Gaussian filter to produce a heat map for each scene and group, allowing for easier computation in the two-dimensional (2D) image space. Heat maps for scenes in which only one saccade was allowed were more sparse than heat maps with three saccades allowed, but the number of participants selected to compute the correlation was always $n = 4$. Multiple comparisons were corrected using the Šidák method. A secondary analysis tested whether heat map similarity differed based on whether people or critical objects were located in the periphery. Given that the CVL viewing condition led to larger saccade amplitudes (Figures 5 and A1.4), this variation in eye movements should lead to differences in spatial correlations of heat maps.

Figures 6B, 6C shows that CVL participants consistently produced less control-like heat maps than either controls or PVL across both social (Panel B) and neutral scenes (Panel C).

A linear mixed-effect model was fit to determine if heat map correlations differed based on experimental conditions. Viewing condition, saccades allowed, scene type, and their interactions were used as fixed effects. Unique scene identifiers were included as random effects. Maximum likelihood estimation was used and p values were calculated using Satterthwaite's method of approximation. Scaled residuals for the model were centered on 0 (minimum $= -2.42$, first quantile $= -0.6$, median $= -0.06$, third quantile $= 0.54$, maximum $= 2.92$) and conditional $R^2 = 0.58$ and marginal $R^2 = 0.38$. Results of the linear mixed model show a significant difference in heat map correlation between controls and CVL participants ($\beta = -0.12$, $t(240) = -6.17$, $SE = 0.02$, $p < 0.001$). The negative β coefficient suggests that CVL fixation heat maps were less correlated than controls by -0.12 on average. The difference in heat map correlations was not significant between controls and PVL participants ($\beta = -0.02$, $t(240) = -1.03$, $SE = 0.02$, $p = 0.3$).

The linear mixed model can only provide results for simple effects differences between controls and PVL or controls and CVL. A second linear mixed model was done to compare differences between PVL with controls and PVL with CVL. This second model revealed a significant difference in heat map correlation between PVL and CVL participants ($\beta = -0.1$, $t(240) = -5.146$, $SE = 0.02$, $p < 0.001$), but not between PVL and controls ($\beta = 0.02$, $t(240) = 1.02$, $SE = 0.02$, $p = 0.31$), suggesting that CVL fixation heat maps were less correlated than PVL, but PVL correlations were not significantly different from controls.

The difference in heat map correlations between one and three saccades allowed was significant ($\beta = 0.14$, $t(296.5) = 5.57$, $SE = 0.02$, $p < 0.001$), suggesting that heat maps were 0.14 more correlated when three saccades were allowed instead of one. The difference in heat map correlation based on saccades allowed is likely due to the sparseness of maps, in which scenes when one saccade was allowed would have half the number of coordinates compared to scenes when three saccades were allowed (even though subsampled groups were always based on $n = 4$). There was no significant difference in heat map correlations between social interaction and neutral scenes ($\beta = -0.01$, $t(296.5)$

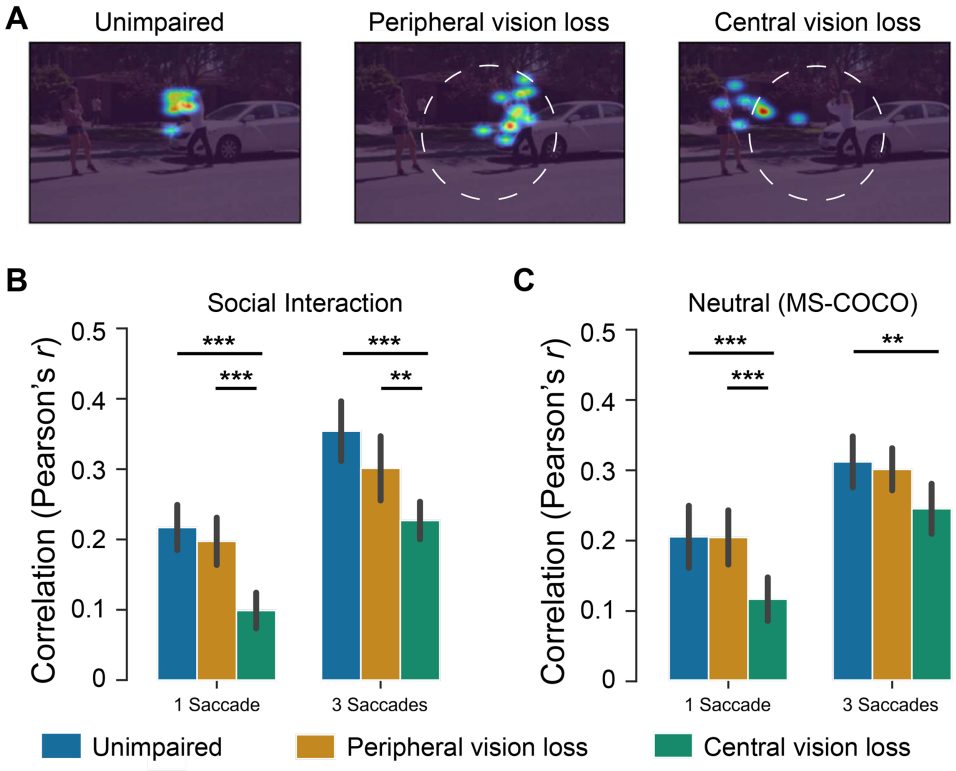

Figure 6. Fixation heat map similarity across viewing conditions. (A) Example fixation heat maps for a social interaction scene, shown for unimpaired (left), PVL (middle), and CVL (right) participants. Heat maps reflect where participants looked during the scene; dashed circles denote the approximate region obscured (CVL) or visible (PVL). Additional heat map examples are shown in Figures A1.6–A1.8. (B, C) Average Pearson correlation between individual participant heat maps and a control group reference heat map (subsampled, $n = 4$ participants per condition, repeated 100 times). For both social interaction scenes (B) and neutral MSCOCO scenes (C), heat map similarity was significantly reduced in the CVL group compared to both controls and PVL, across both fixation limits.

$= -0.54$, $SE = 0.02$, $p = 0.59$). The fixed effects for the intercept were significant ($\beta = 0.22$, $t(296.5) = 13.16$, $SE = 0.02$, $p < 0.001$), suggesting that heat maps were correlated when fixed effects were ignored. No interactions in the linear mixed model were significant for fixation heat map correlations.

An ANOVA was conducted on the linear mixed model to test all possible comparisons between groups. The ANOVA revealed a significant main effect of viewing condition on heat map correlations ($F(2, 232) = 13.91$, $p < 0.001$). There was also a significant main effect of saccades allowed ($F(1, 120) = 81.85$, $p < 0.001$), but not scene type ($F(1, 120) = 0.07$, $p = 0.80$). The ANOVA did not reveal any significant interactions for the model.

Post hoc comparisons (Šidák-corrected) revealed that CVL heat maps were significantly less correlated with controls than both control and PVL groups across nearly all conditions. For social scenes, heat map similarity was lower with one saccade (vs. control: $M = 0.12$, $p < 0.001$, $d = 1.52$; vs. PVL: $M = 0.10$, $p < 0.001$, $d = 1.27$) and three saccades (vs. control: $M = 0.13$, $p < 0.001$, $d = 1.64$; vs. PVL: $M = 0.07$, $p < 0.001$, $d = 1.42$). For neutral scenes, CVL groups also differed from both control and PVL with one saccade (control: $M = 0.09$, $p < 0.001$, $d = 1.14$; PVL: $M = 0.09$, $p < 0.001$, $d = 1.13$), and from controls with three saccades ($M = -0.07$, $p = 0.009$, $d = 0.85$), but not significantly from PVL ($M = 0.06$, $p = 0.05$). No differences were found between control and PVL participants ($p = 0.16$–1).

ANOVA results for the fixation heat map analysis should be interpreted with caution due to the chance of increased false positives (Type I error) that can occur from the subsampling process. The linear mixed models and ANOVA post hoc results show that CVL participant heat maps were significantly less correlated than controls and PVL participants.

More example heat maps can be found in Figures A1.6–A1.8.

Fixations to annotated areas of interest

To assess how the average number of fixations to AOIs would differ between viewing conditions, we counted the number of fixations to labeled humans and critical objects for each scene and viewing condition. A linear mixed model was used to test whether the

number of object-directed fixations to AOIs varied. Viewing condition and number of saccades allowed were included as fixed effects, and scene ID was included as a random factor. The number of fixations to AOIs was done by averaging data from all participants in each assigned viewing condition, so participant identities were not included as random effects. Post hoc Šidák-corrected comparisons were performed across conditions, and object arrangement (e.g., peripheral layout) was included as an additional binary predictor. As scene type did not significantly interact with any other factors in our model, results are collapsed across social and neutral scenes for clarity (Figure 7B).

A linear mixed-effects model revealed a significant interaction between viewing condition and number of saccades allowed ($F(2, 236) = 7.01$, $p = 0.001$). Both viewing condition ($F(2, 236) = 24.19$, $p < 0.001$) and number of saccades allowed ($F(1, 118) = 50.18$, $p < 0.001$) were significant main effects on fixations to AOIs. When participants were allowed three saccades (Panel C), those with CVL made significantly fewer fixations to AOIs than both controls ($M = 0.44$, $p < 0.001$, $d = 1.31$) and PVL participants ($M = 0.35$, $p < 0.001$, $d = 1.06$). There was no difference between control and PVL groups ($M = 0.08$, $p = 0.75$) and no significant group differences when only one saccade was allowed (Panel B; $p = 0.17$–0.88).

Linking gaze behavior to scene understanding

To examine the relationship between scene understanding and eye movement behavior, we used Pearson's correlation to measure the relationship between average human description ratings with across-condition fixation heat map correlations on a per-scene basis. This allowed us to assess whether higher description ratings were associated with more control-like fixation patterns. Additional correlations were computed between description ratings and other eye movement metrics to determine how viewing condition modulated behavior across modalities.

As shown in Figure 8A, there was no significant correlation between average human ratings and global fixation heat map similarity ($r = 0.09$, $p = 0.06$), suggesting that the overall spatial distribution of fixations was not predictive of scene comprehension. In contrast, higher description ratings were significantly associated with increased fixations to AOIs ($r = 0.21$, $p < 0.001$; Figure 8B). This relationship held across all viewing conditions (controls: $r = 0.26$, $p = 0.004$; PVL: $r = 0.20$, $p = 0.03$; CVL: $r = 0.18$, $p = 0.04$), suggesting that the informativeness of participant descriptions was better predicted by socially or semantically relevant content than by the overall shape of their gaze distribution. The data in Figure 8B are skewed to fewer average fixations (between 0 and 2, same data as Figure 7B), with critical fixation values larger than 4 likely representing scenes in which areas were fixated by multiple participants or multiple AOIs overlapped. Taken together, these results are surprisingly inconsistent with our hypothesis that visual sampling patterns (heat maps) can be used to predict scene-understanding ability but consistent with the hypothesis that the number of fixations to AOIs is a better predictor of real-world scene comprehension.

Summary of behavioral measure relationships

To explore how different behavioral and eye movement variables related to each other, we computed a correlation matrix using scene-level averages across all viewing conditions. The relationships between measures are shown in Figure 9. Description ratings were most strongly associated with the number of fixations to AOIs. In contrast, there was no reliable relationship between description ratings and global fixation similarity (i.e., heat map correlation). Interestingly, heat map correlation was negatively associated with

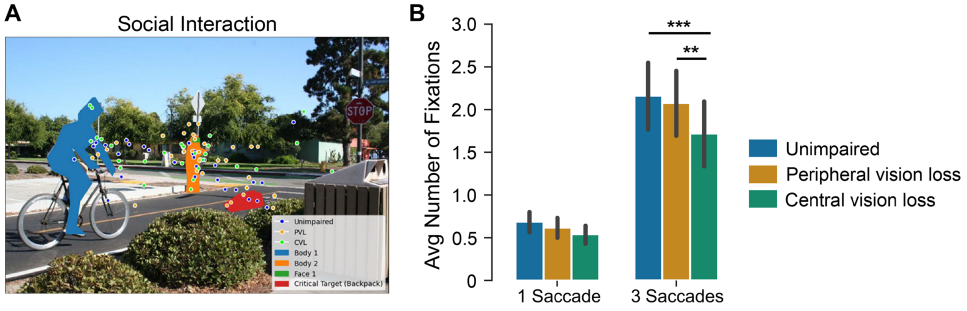

Figure 7. (A) Example scene with labeled humans (orange and blue) and critical task-relevant objects (*red*), used to quantify visual access to meaningful scene content (see Methods). (B) Average number of fixations to AOIs by viewing condition and number of saccades allowed. With three saccades, participants in the CVL condition fixated labeled content significantly less often than both controls and PVL participants, regardless of scene type. No significant differences were observed when only one saccade was allowed.

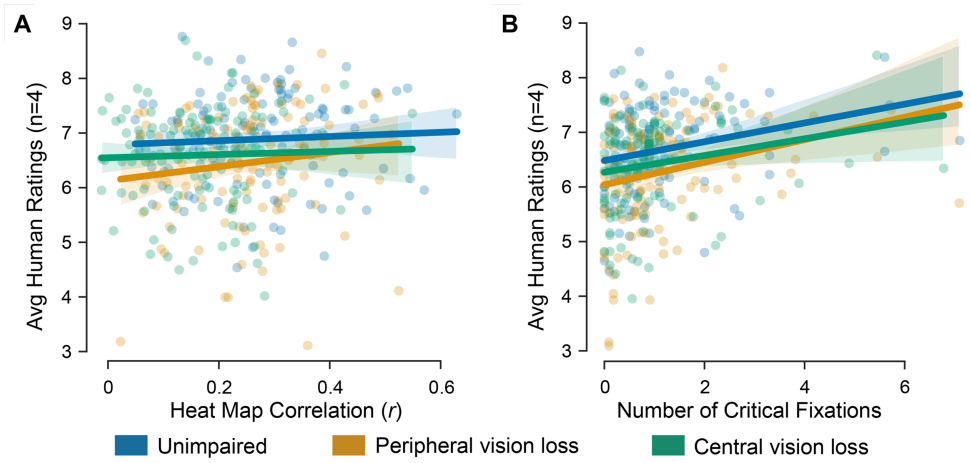

Figure 8. (A) Relationship between average human ratings of scene descriptions and fixation heat map correlations across all scenes. No significant association was observed, suggesting that global gaze distribution alone did not predict scene understanding. (B) Description ratings were significantly correlated with the number of critical fixations (i.e., fixations to labeled humans and critical objects), indicating that access to semantically meaningful regions of the scene contributed to higher-quality descriptions.

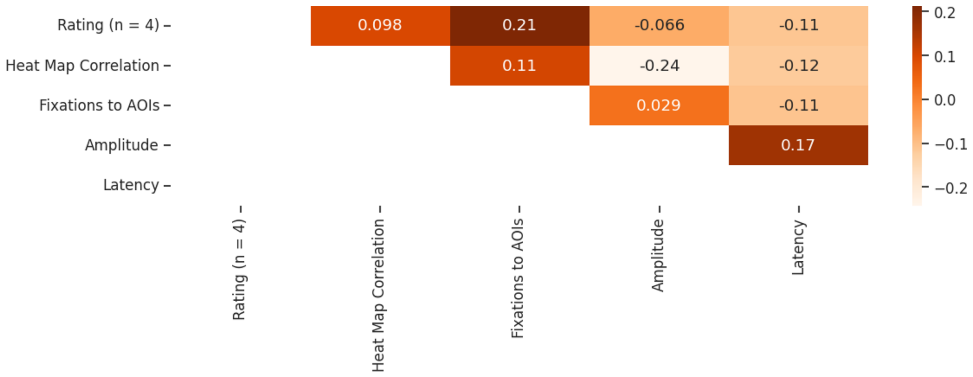

Figure 9. Correlation matrix of key behavioral and oculomotor measures.

first saccade amplitude, while amplitude was positively associated with first saccade latency.

Discussion

This study demonstrates that simulated vision loss (both central and peripheral) impairs rapid scene understanding and alters early eye movement behavior for real-world scenes. By combining gaze-contingent visual impairment with human-rated scene descriptions, we show that the type of vision loss produces distinct effects: CVL primarily disrupted where participants looked, while PVL strongly impacted their understanding of the scene. Importantly, comprehension suffered most in social scenes, where describing meaningful interactions requires integration of multiple scene elements. These findings suggest that rapid scene understanding relies on both foveal detail and peripheral context and that different forms of visual impairment disrupt this process in complementary ways.

Our results extend prior work on scene categorization and simulated low vision (Loschky et al., 2019; Trouilloud et al., 2020) by using more ecologically valid stimuli and tasks. Descriptions from PVL and CVL participants were rated worse than controls, particularly for social interaction scenes (Figure 4). Viewing condition, scene type, and number of saccades allowed were significant main predictors for ratings and cosine similarity values of participant descriptions. This suggests that understanding social interaction scenes required accurate and detailed script representations compared to neutral scenes. Descriptions with three saccades were rated as better quality than descriptions with one saccade, highlighting the role of eye movements for information acquisition. This replicates and expands

on prior evidence that peripheral vision supports rapid gist extraction (Larson & Loschky, 2009; Thibaut et al., 2014), and suggests that central vision is critical for resolving the social and semantic detail required to describe interactions. Notably, the neutral scenes (which often contained isolated objects or passive humans) produced smaller differences, likely because description accuracy for the neutral scenes required recognition instead of high-level interpretation of multiple elements.

Scene type (social interaction vs. neutral) impacted scene-understanding ability independent of viewing condition and number of saccades allowed. The effects of scene context are shown in Figures 4A, 4B, in which descriptions for social interaction scenes were rated lower than neutral scenes. The social interaction scenes contained multiple meaningful relationships between AOIs by design, which would require a level of information extraction not used for the neutral scenes. Previous studies testing comprehension of agent–patient scenes, in which a person is acting on another person or object, required participants to describe relationships between different characters in the same scene (similar to different human AOIs in the current study). Understanding social interaction and agent–patient scenes may involve forming detailed scene scripts that would not be required for neutral scenes (Abelson, 1981; Dobel et al., 2007; Hafri et al., 2013). In the current study, social interaction scenes always included one or multiple characters completing a specific action or communicating with another person. The relationships between AOIs and the presence of people in scenes suggest differences in how the person bias effect interacts with scene context: The relationships between people and AOIs in social interaction scenes required more detail than identifying people in neutral scenes (Fletcher-Watson et al., 2008; Humphrey & Underwood, 2010). The SPECT model, for example, suggests that the medium- to high-level front-end stimulus features relevant for social interaction scenes could be reflected in back-end semantic processes related to information extraction (Loschky et al., 2020). Interestingly, results of the current study revealed stronger differences in effects for scene understanding compared to eye movement metrics: Post hoc analyses for the main effect of scene type on amplitude, latency, heat maps, and fixations to AOIs showed minimal differences between viewing conditions.

Simulated vision loss also disrupted fixation behavior for real-world scenes. CVL observers made larger first saccades (Figure 5A) and had lower fixation heat map similarity to controls across all conditions (Figures 6B, 6C). Previous studies have shown that scotoma/clear window sizes for CVL and PVL simulations have a direct impact on saccade amplitudes, with larger central scotomas leading to larger amplitudes and smaller clear windows reducing amplitudes (Loschky & McConkie, 2002; Nuthmann, 2014; Cajar et al., 2016; Nuthmann & Canas-Bajo, 2022; Yu & Kwon, 2023). Using the "full-report" method, in which participants had minimal expectations about where to look for each scene, we found a significant increase in amplitudes for CVL participants but no change in amplitudes for PVL participants. Impacts of viewing condition on saccade amplitude may be task and experience dependent. Participants viewing scenes with simulations, especially without training, will rely on the remaining intact areas, causing participants with CVL to look outside the scotoma, resulting in larger amplitudes. The effects of CVL on first saccade amplitudes in the current study should be carefully interpreted in comparison with prior studies reporting smaller saccades in patients with binocular central scotomas performing visual search tasks (Vullings et al., 2022) or adaptive reductions in saccade range with training involving simulations (Kwon et al., 2013; Vice et al., 2022).

Our results suggest that without training, participants with central scotomas adopt a compensatory strategy of quickly fixating peripheral regions to avoid the blind spot. As shown in Figures 5A and A1.4, PVL participant amplitudes were slightly smaller but not significantly different from controls. Other experiments have found that PVL leads to smaller saccade amplitudes, suggesting that a clear window with a radius smaller than 5° would cause PVL participants to make smaller saccades for scene understanding tasks compared to control participants (Loschky & McConkie, 2002; Loschky et al., 2005; Nuthmann, 2014; Cajar et al., 2016). While only one scotoma/clear window size and level of blur were tested in the current study, the visual search and peripheral blur detection experiments by Loschky and McConkie (2002), Loschky et al. (2005), and Cajar et al. (2016) found different effects based on the size and severity of blur for the scotoma/clear window conditions (Loschky & McConkie, 2002; Loschky et al., 2005; Cajar et al., 2016). Participants in this study were not given information about what to look for in each scene, so first saccade amplitudes may vary for different task instructions ("Report the number of people in the scene." or "Was the scene located indoors or outdoors?"). The within-scene locations and distances between AOIs in the social interaction scenes suggest different impacts on first saccade amplitude and latency (see Figure A1.5), but central and peripheral location of AOIs were not controlled for in the social interaction dataset used.

Both PVL and CVL participants showed increased saccade latencies (Figure 5B), consistent with prior studies of visual search and reading in patient populations and simulation paradigms (Nuthmann, 2014; Cajar et al., 2016; Janssen & Verghese, 2016; Agaoglu et al., 2022; Vice et al., 2022; Yu & Kwon, 2023). The increase in saccade latencies due to reduced

information was likely due to limited planning and extraction of scene context (Loschky & McConkie, 2002; Cajar et al., 2016). Scene type had a significant main effect on the relationship between amplitude and latency, as shown in Figures 5C, 5D. For social interaction scenes, there was a strong correlation between amplitude and latency for control and CVL participants, but not PVL participants. This association was less prominent for neutral scenes, including a weaker relationship between amplitude and latency for the unimpaired viewing conditions. Interestingly, CVL participants made slightly faster and significantly larger saccades than PVL participants, yet still produced higher-rated descriptions. One interpretation is that the central scotoma disrupted foveal detail but preserved access to scene context, whereas PVL observers struggled to localize targets and integrate spatial relationships due to a lack of peripheral preview. Supporting this view, saccade amplitude and latency were positively correlated for scenes with humans in the periphery (Figure A1.5), particularly for CVL participants, suggesting that increased uncertainty delayed planning and triggered larger compensatory eye movements.

Heat map similarity to controls did not predict description quality, even among control participants (Figure 8A), suggesting that meaningful scene understanding depends not on looking in the same place but on extracting the right *information*. CVL participants were more likely to miss AOIs when three saccades were allowed, similar to results in which participants with simulated CVL missed targets (showed increased errors) during visual search tasks (Nuthmann, 2014; Nuthmann & Malcolm, 2016; Nuthmann & Canas-Bajo, 2022). While heat maps were not predictive of scene-understanding ability compared to fixations to AOIs, distributions of saccade amplitudes for CVL matched differences in heat map correlations, showing the relationship between viewing condition, amplitude, and visual sampling behavior (Loschky & McConkie, 2002; Cajar et al., 2016). Participants in the PVL condition produced fixation heat maps that were more similar to controls, but their descriptions were rated worse, showing the importance of contextual integration for a scene across varying locations in the current field of view. Extraction of the right information may involve fixating and accurately describing the relationship between two or more AOIs instead, as shown in Figures 7B and 8B. The larger the distance between two AOIs, particularly between a critical object and another labeled region, the more PVL was likely impacted. Similarly, smaller distances between two AOIs could make it more difficult for participants in the CVL conditions to discern scene content. Supporting this, fixations to AOIs were consistently linked to higher-rated descriptions across all groups (Figure 8B), underscoring the importance of semantically informative fixations, particularly in social scenes.

Limitations

The goal of this study was to measure scene-understanding ability by simulating different forms of vision loss. Collecting human ratings of descriptions posed multiple challenges. Given that collecting ratings of typed descriptions as a measure of scene understanding is new, we did not conduct a power analysis. Our sample size of $n = 32$ is relatively small. Having multiple raters for each participant description provided multiple observations ($n = 4$) for each trial, which helped show a better representation of scene understanding for each description but would have also benefited from more human raters and more ground-truth descriptions.

Our scene-understanding between-subjects design is inspired by the within-subjects designs used in the previous literature. Studies that used a within-subjects design could be categorized as either having normally sighted participants tested with stimuli for both CVL and PVL simulation conditions (Nuthmann, 2014; Nuthmann & Canas-Bajo, 2022; Loschky & McConkie, 2002; Yu & Kwon, 2023) or measuring changes in oculomotor behavior with CVL (Kwon et al., 2013; Tsank & Eckstein, 2017; Agaoglu et al., 2022; Vice et al., 2022). Two experiments from Geringswald and Pollman (2015) and Pollman et al. (2020) used a between-subjects design with larger sample sizes of sighted participants ($n = 75$ and $n = 60$, respectively) completing visual search tasks for letters and objects (Geringswald & Pollmann, 2015; Pollmann et al., 2020). A larger number of subjects could have increased statistical power for the current study but would have also increased the differences in effects already demonstrated by the viewing conditions.

We had five sets of ground-truth descriptions rated by five people. While the goal was to capture the highest and lowest scores for ideal and incorrect descriptions relative to the saccade-limited descriptions in the main experiment, there were logistical challenges to get ratings for both ground-truth and saccade-limited descriptions simultaneously. Ratings for ground-truth descriptions were done after data from the main saccade-limited experiment were collected. A larger number of ground-truth descriptions would have provided a more normative representation of ideal description quality. Only one ground-truth was randomly assigned to a different scene for the lower-bound calculation. More ground-truths would likely lead to a lower ceiling for the upper bound due to variations in descriptions, specifically for social interaction scenes, given the level of detail needed for a higher-rated description. Using two or more randomly assigned sets to calculate the lower bound would increase the floor, due to the increased likelihood of descriptions being assigned to different scenes that share similar critical targets or details. The same could be predicted for the

descriptions in the main experiment as a larger number of saccade-limited descriptions would pull the average rating for each viewing condition down, likely increasing the differences in ratings by viewing condition.

The central starting fixation position at the beginning of each trial determined the amount of information available for both CVL and PVL viewing conditions. While this helps to explain the results for the current study, changing the starting location to a peripheral location could lead to different results. The heat map correlation measure used can be sensitive to small spatial shifts, kernel size, and grid resolution. When testing different kernel sizes, standard deviations, and number of bins, the correlation values were different, but the results were not changed (CVL heat map correlations were lower than PVL and control heat maps; see Figure A1.9). The subsampling method used in our analysis reduced bias, but including heat maps from the same participant could have introduced statistical dependence for the average correlation through randomization.

The number of saccades allowed was matched between participants. While it is expected that more saccades allowed would improve scene-understanding ability, counterbalancing one and three saccades for the same image could help control for sparsity in heat maps between the different amounts of saccades allowed. While Pearson's correlation between description ratings and the average number of fixations to AOIs was significant, the distribution of the number of fixations to AOIs was not normal. Spearman's rank correlation and a nonparametric test revealed similar effects (Spearman's: $r = 0.25$, $p < 0.001$; Kendall's tau: $\tau = 0.17$, $p < 0.001$), but results should be interpreted with caution.

Future work

Taken together, our findings suggest that peripheral vision supports rapid orienting and context estimation, while central vision provides access to high-acuity details needed for interpretation. Both are necessary for functional scene understanding in the real world. Simulations provide an informative approximation of patient behavior, but future studies may benefit from measuring scene understanding with clinical populations. The central scotoma and clear window used in this study were inspired by Kwon et al. (2013) and other studies that used a radii of 5° (Kwon et al., 2013; Geringswald & Pollmann, 2015; Cajar et al., 2016). The impact of CVL and PVL simulation size has been studied for scene gist (Larson & Loschky, 2009; Loschky et al., 2019), visual search (Nuthmann, 2014; Nuthmann & Canas-Bajo, 2022), and reading (Yu & Kwon, 2023). The effects of different scotoma and clear window sizes on "full-report" scene understanding have not been explored.

Previous studies have attempted to quantify how much adaptation or training can mitigate the observed deficits with simulations (McIlreavy et al., 2012; Agaoglu et al., 2022; Nuthmann & Canas-Bajo, 2022; Vice et al., 2022) or how training gaze strategies can influence experience-dependent plasticity in patients with chronic vision loss (Janssen & Verghese, 2016; Maniglia, Soler, & Trotter, 2020). Participants using CVL simulations might have tried to avoid the central scotoma area entirely instead of fixate on one of the relevant AOIs. Training participants with simulated CVL before testing could improve stability of fixations and measure description accuracy based on peripheral viewing strategies (Kwon et al., 2013; Agaoglu et al., 2022). Likewise, training for PVL could lead to more voluntary saccades into the peripheral scotoma area. Previous work using simulations of CVL reported changes in eccentric viewing behaviors after 12 to 30 hours (Kwon et al., 2013; Vice et al., 2022) or thousands of training trials (Mazyar & Tjan, 2016). Longitudinal studies and eye movement–based interventions could reveal whether patients can learn to optimize gaze strategies under naturalistic conditions.

This work also highlights the value of naturalistic tasks, such as scene description, in assessing functional vision. Traditional clinical metrics like acuity and contrast sensitivity may not account for impairments in real-world understanding. Integrating eye tracking and semantically rich tasks could yield more sensitive and ecologically valid tools for evaluating patient outcomes and assistive device efficacy. Furthermore, future work should explore how dynamic stimuli and active tasks, such as activities of daily living, interact with vision loss. Our findings also emphasize the unique role of social information (i.e., faces, interactions, and people) in driving both visual behavior and semantic understanding. Functional assessments that ignore these dimensions may underestimate the real-world impact of vision loss.

Finally, while simulations are not a substitute for patient studies, they offer a safe and scalable way to prototype interventions and evaluate scene comprehension under controlled conditions. By anchoring behavioral performance in both gaze behavior and semantic interpretation, our findings help bridge the gap between sensory impairment and functional vision and point toward more ecologically valid tools for vision assessment and rehabilitation.

Conclusions

Understanding how vision loss affects real-world perception requires disentangling the distinct contributions of central and peripheral vision. Using gaze-contingent simulations of biologically plausible

scotomas, we show that peripheral vision loss degrades scene understanding more severely than central loss, particularly for socially meaningful content. In contrast, central vision loss led to more pronounced changes in eye movement planning, including larger and delayed saccades. These findings reveal a dissociation between perceptual and oculomotor consequences of visual field loss, suggesting that peripheral input plays a critical role in rapid semantic processing of natural scenes. Our results underscore the importance of evaluating low vision conditions with semantically rich information. Future strategies in low vision rehabilitation can benefit by showing the connections between different types of vision loss, eye movements, and their effects on perception.

Keywords: low vision, scotoma, real-world scenes, saccades

Acknowledgments

The authors thank Isabella Sarullo, Madeline Kaplan, Lucia Alem, Jin Hee Yang, Emma Bowen, and Jianna Wong for their assistance with data collection.

Supported by a project from the Noyce Foundation.

Commercial relationships: none.
Corresponding author: Byron Johnson.
Email: byron.johnson@psych.ucsb.edu.
Address: Department of Psychological & Brain Sciences, University of California, Santa Barbara, Santa Barbara, CA 93106-9660, USA.

References

Agaoglu, M. N., Fung, W., & Chung, S. T. L. (2022). Oculomotor responses of the visual system to an artificial central scotoma may not represent genuine visuomotor adaptation. *Journal of Vision, 22*(10), 17, https://doi.org/10.1167/jov.22.10.17.

Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist, 36*(7), 715–729.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Biederman, I. (1977). On processing information from a glance at a scene: Some implications for a syntax and semantics of visual processing. In *Proceedings of the ACM/SIGGRAPH workshop on user-oriented design of interactive graphics systems - UODICS '76, 75*. Pittsburgh, PA: ACM Press.

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology, 103*(3), 597–600.

Bradski, G. (2000). The OpenCV library. *Dr Dobb's Journal of Software Tools, 25*(11), 120–125.

Cajar, A., Schneeweiss, P., Engbert, R., & Laubrock, J. (2016). Coupling of attention and saccades when viewing scenes with central and peripheral degradation. *Journal of Vision, 16*(2), 8, https://doi.org/10.1167/16.2.8.

Costela, F. M., Saunders, D. R., Rose, D. J., Katjezovic, S., Reeves, S. M., & Woods, R. L. (2019). People with central vision loss have difficulty watching videos. *Investigative Ophthalmology & Visual Science, 60*(1), 358.

Dobel, C., Gumnior, H., Bölte, J., & Zwitserlood, P. (2007). Describing scenes hardly seen. *Acta Psychologica, 125*(2), 129–143.

Duchowski, A. T., & Çöltekin, A. (2007). Foveated gaze-contingent displays for peripheral LOD management, 3D visualization, and stereo imaging. *ACM Transactions on Multimedia Computing, Communications, and Applications, 3*(4), 1–18.

Fletcher, D. C., Schuchard, R. A., & Renninger, L. W. (2012). Patient awareness of binocular central scotoma in age-related macular degeneration. *Optometry and Vision Science, 89*(9), 1395–1398.

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception, 37*(4), 571–583.

Foulsham, T., Teszka, R., & Kingstone, A. (2011). Saccade control in natural images is shaped by the information visible at fixation: Evidence from asymmetric gaze-contingent windows. *Attention, Perception, & Psychophysics, 73*(1), 266–283.

Geringswald, F., & Pollmann, S. (2015). Central and peripheral vision loss differentially affects contextual cueing in visual search. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(5), 1485–1496.

Glanemann, R., Zwitserlood, P., Bölte, J., & Dobel, C. (2016). Rapid apprehension of the coherence of action scenes. *Psychonomic Bulletin & Review, 23*(5), 1566–1575.

Greene, M. R., Botros, A. P., Beck, D. M., & Fei-Fei, L. (2015). What you see is what you expect: Rapid scene understanding benefits from prior experience. *Attention, Perception, & Psychophysics, 77*(4), 1239–1251.

Guadron, L., Titchener, S. A., Abbott, C. J., Ayton, L. N., Van Opstal, J., Petoe, M. A., . . . Goossens, J. (2023). The saccade main sequence in patients

with retinitis pigmentosa and advanced age-related macular degeneration. *Investigative Opthalmology & Visual Science, 64*(3), 1.

Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. *Journal of Experimental Psychology: General, 142*(3), 880–905.

Hartong, D. T., Berson, E. L., & Dryja, T. P. (2006). Retinitis pigmentosa. *The Lancet, 368*(9549), 1795–1809.

Hayes, T. R., & Henderson, J. M. (2025). DeepMeaning: Estimating and interpreting scene meaning for attention using a vision-language transformer. *Open Mind, 9*, 1020–1036.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour, 1*(10), 743–747.

Henderson, J. M., & Hollingworth, A. (1998). Eye movements during scene viewing. In G Underwood. *Eye Guidance in Reading and Scene Perception* (pp. 269–293). Oxford, England: Elsevier.

Houston, K. E., Bowers, A. R., Peli, E., & Woods, R. L. (2018). Peripheral prisms improve obstacle detection during simulated walking for patients with left hemispatial neglect and hemianopia. *Optometry and Vision Science, 95*(9), 795–804.

Humphrey, K., & Underwood, G. (2010). The potency of people in pictures: Evidence from sequences of eye fixations. *Journal of Vision, 10*(10), 19, https://doi.org/10.1167/10.10.19.

Janssen, C. P., & Verghese, P. (2016). Training eye movements for visual search in individuals with macular degeneration. *Journal of Vision, 16*(15), 29, https://doi.org/10.1167/16.15.29.

Karmakar, S., & Eckstein, M. P. (2025). The psychophysics of dynamic gaze following saccades during search. *Journal of Vision, 25*(14), 14, https://doi.org/10.1167/jov.25.14.14.

Kasowski, J., Johnson, B. A., Neydavood, R., Akkaraju, A., & Beyeler, M. (2023). A systematic review of extended reality (XR) for understanding and augmenting vision loss. *Journal of Vision, 23*(5), 5, https://doi.org/10.1167/jov.23.5.5.

Klauke, S., Sondocie, C., & Fine, I. (2023). The impact of low vision on social function: The potential importance of lost visual social cues. *Journal of Optometry, 16*(1), 3–11.

Kwon, M., Nandy, A., & Tjan, B. (2013). Rapid and persistent adaptability of human oculomotor control in response to simulated central vision loss. *Current Biology, 23*(17), 1663–1669.

Larson, A. M., & Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision, 9*(10), 6, https://doi.org/10.1167/9.10.6.

Legge, G. E., & Chung, S. T. (2016). Low vision and plasticity: Implications for rehabilitation. *Annual Review of Vision Science, 2*(1), 321–343.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., . . . Dollár, P. (2014). Microsoft COCO: Common objects in context. arXiv preprint, https://doi.org/10.48550/ARXIV.1405.0312.

Loschky, L., McConkie, G., Yang, J., & Miller, M. (2005). The limits of visual resolution in natural scene viewing. *Visual Cognition, 12*(6), 1057–1092.

Loschky, L. C., Larson, A. M., Smith, T. J., & Magliano, J. P. (2020). The Scene Perception & Event Comprehension Theory (SPECT) applied to visual narratives. *Topics in Cognitive Science, 12*(1), 311–351.

Loschky, L. C., & McConkie, G. W. (2000). User performance with gaze contingent multiresolutional displays. In A. T. Duchowski *Proceedings of the Symposium on Eye Tracking Research & Applications - ETRA '00* (pp. 97–103). Palm Beach Gardens, FL: ACM Press.

Loschky, L. C., & McConkie, G. W. (2002). Investigating spatial vision and dynamic attentional selection using a gaze-contingent multiresolutional display. *Journal of Experimental Psychology: Applied, 8*(2), 99–117.

Loschky, L. C., Szaffarczyk, S., Beugnet, C., Young, M. E., & Boucart, M. (2019). The contributions of central and peripheral vision to scene-gist recognition with a 180° visual field. *Journal of Vision, 19*(5), 15, https://doi.org/10.1167/19.5.15.

Maniglia, M., Soler, V., & Trotter, Y. (2020). Combining fixation and lateral masking training enhances perceptual learning effects in patients with macular degeneration. *Journal of Vision, 20*(10), 19, https://doi.org/10.1167/jov.20.10.19.

Mazyar, H., & Tjan, B. (2016). In search of the visual and oculomotor factors that determine the location of a preferred retinal locus. *Journal of Vision, 16*(12), 1338, https://doi.org/10.1167/16.12.1338.

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics, 17*(6), 578–586.

McDonald, M. A., Stevenson, C. H., Kersten, H. M., & Danesh-Meyer, H. V. (2022). Eye movement abnormalities in glaucoma patients: A review. *Eye and Brain, 14*, 83–114.

McIlreavy, L., Fiser, J., & Bex, P. J. (2012). Impact of simulated central scotomas on visual search in natural scenes. *Optometry and Vision Science, 89*(9), 1385–1394.

Murlidaran, S., & Eckstein, M. P. (2025). Eye movements during free viewing to maximize scene understanding. *Nature Communications, 17*(1), 940.

Nguyen, D., Trieschnigg, D., & Theune, M. (2014). Using crowdsourcing to investigate perception of narrative similarity. In J. Li, X. S. Wang, M. Garofalakis, I. Soboroff, T. Suel & M. Wang *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 321–330). Shanghai, China: ACM.

Nuthmann, A. (2014). How do the regions of the visual field contribute to object search in real-world scenes? Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance, 40*(1), 342–360.

Nuthmann, A., & Canas-Bajo, T. (2022). Visual search in naturalistic scenes from foveal to peripheral vision: A comparison between dynamic and static displays. *Journal of Vision, 22*(1), 10, https://doi.org/10.1167/jov.22.1.10.

Nuthmann, A., & Malcolm, G. L. (2016). Eye guidance during real-world scene search: The role color plays in central and peripheral vision. *Journal of Vision, 16*(2), 3, https://doi.org/10.1167/16.2.3.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., . . . Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203.

Peli, E., Goldstein, R., & Jung, J.-H. (2023). The invisibility of scotomas I: The carving hypothesis. *Optometry and Vision Science, 100*(8), 515–529.

Peyrin, C., Ramanoël, S., Roux-Sibilon, A., Chokron, S., & Hera, R. (2017). Scene perception in age-related macular degeneration: Effect of spatial frequencies and contrast in residual vision. *Vision Research, 130*, 36–47.

Pollmann, S., Geringswald, F., Wei, P., & Porracin, E. (2020). Intact contextual cueing for search in realistic scenes with simulated central or peripheral vision loss. *Translational Vision Science & Technology, 9*(8), 15.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.

Reingold, E. M., Loschky, L. C., McConkie, G. W., & Stampe, D. M. (2003). Gaze-contingent multi-resolutional displays: An integrative review. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 45*(2), 307–328.

Sanocki, T., Nguyen, T., Shultz, S., & Defant, J. (2023). Novel scene understanding, from gist to elaboration. *Visual Cognition, 31*(3), 188–215.

Seiple, W., Rosen, R. B., & Garcia, P. M. (2013). Abnormal fixation in individuals with age-related macular degeneration when viewing an image of a face. *Optometry and Vision Science, 90*(1), 45–56.

Shintani, K., Shechtman, D. L., & Gurwood, A. S. (2009). Review and update: Current treatment trends for patients with retinitis pigmentosa. *Optometry, 80*(7), 384–401.

Shioiri, S., & Ikeda, M. (1989). Useful resolution for picture perception as a function of eccentricity. *Perception, 18*(3), 347–361.

Skalski, P. (2019). Make Sense (Version 1.11.0-alpha) [Computer software]. GitHub. https://github.com/SkalskiP/make-sense/.

The Astropy Collaboration,Robitaille, T. P., Tollerud, E. J., Greenfield, P., Droettboom, M., Bray, E., Aldcroft, T., . . . Streicher, O. (2013). Astropy: A community Python package for astronomy. *Astronomy & Astrophysics, 558*, A33.

Thibaut, M., Tran, T. H. C., Szaffarczyk, S., & Boucart, M. (2014). The contribution of central and peripheral vision in scene categorization: A study on people with central vision loss. *Vision Research, 98*, 46–53.

Titchener, S. A., Ayton, L. N., Abbott, C. J., Fallon, J. B., Shivdasani, M. N., Caruso, E., . . . Petoe, M. A. (2019). Head and gaze behavior in retinitis pigmentosa. *Investigative Opthalmology & Visual Science, 60*(6), 2263.

Tran, T. H. C., Rambaud, C., Despretz, P., & Boucart, M. (2010). Scene perception in age-related macular degeneration. *Investigative Opthalmology & Visual Science, 51*(12), 6868.

Trouilloud, A., Kauffmann, L., Roux-Sibilon, A., Rossel, P., Boucart, M., Mermillod, M., . . . Peyrin, C. (2020). Rapid scene categorization: From coarse peripheral vision to fine central vision. *Vision Research, 170*, 60–72.

Tsank, Y., & Eckstein, M. P. (2017). Domain specificity of oculomotor learning after changes in sensory processing. *The Journal of Neuroscience, 37*(47), 11469–11484.

Van Der Stigchel, S., Bethlehem, R. A. I., Klein, B. P., Berendschot, T. T. J. M., Nijboer, T. C. W., & Dumoulin, S. O. (2013). Macular degeneration affects eye movement behavior during visual search. *Frontiers in Psychology, 4*, 1–9.

Vandersnickt, M. F., Van Eijgen, J., Lemmens, S., Stalmans, I., Pinto, L. A., & Vandewalle, E. M. (2024). Visualfield patterns in glaucoma: A systematic review. *Saudi Journal of Ophthalmology, 38*(4), 306–315.

Verghese, P., Vullings, C., & Shanidze, N. (2021). Eye movements in macular degeneration. *Annual Review of Vision Science, 7*(1), 773–791.

Vice, J. E., Biles, M. K., Maniglia, M., & Visscher, K. M. (2022). Oculomotor changes following learned use of an eccentric retinal locus. *Vision Research, 201*, 108126.

Vullings, C., Lively, Z., & Verghese, P. (2022). Saccades during visual search in macular degeneration. *Vision Research, 201*, 108113.

Wang, Y., Tao, S., Xie, N., Yang, H., Baldwin, T., & Verspoor, K. (2023). Collective human opinions in semantic textual similarity. *Transactions of the Association for Computational Linguistics, 11*, 997–1013.

Yu, H., & Kwon, M. (2023). Altered eye movements during reading with simulated central and peripheral visual field defects. *Investigative Opthalmology & Visual Science, 64*(13), 21.

Zwickel, J., & Vo, M. L. H. (2010). How the presence of persons biases eye movements. *Psychonomic Bulletin & Review, 17*(2), 257–262.

Appendix

Action phrases and number of people in social interaction scenes

The action phrase(s) (verbs) and number of people present in each scene from the social interaction scene dataset.

Timing validation for gaze-contingent display

To estimate the delay between eye position updates and stimulus rendering, we inserted timestamped logging commands into the experiment's Python code. For 10 randomly selected scenes, we recorded the time elapsed between receiving gaze coordinates from the eye tracker and rendering the updated frame with the appropriate scotoma filter. The average delay was 6.89 $\pm$ 1.21 ms across all tested trials (Figure A1.1).

Human rater reliability and rating distributions

To assess variability and agreement in human semantic similarity judgments, we analyzed the distribution of ratings, pairwise interrater reliability, and rating correlations between all four raters.

Figure A1.2A shows the distribution of ratings assigned by each rater across all scene descriptions. A Shapiro–Wilk test indicated that the average ratings were not normally distributed ($W = 0.98$, $p < 0.001$). The mean rating across all descriptions was 6.44 (SEM = 0.02), and both the mode and median were 6.5.

Figure A1.2B shows pairwise interrater reliability (IRR) scores between raters, calculated as the proportion of agreement in ordinal ranking across all rated items. IRR values ranged from 0.042 (Rater 1 and Rater 4) to 0.23 (Rater 2 and Rater 4), indicating modest agreement in relative similarity judgments.

Figure A1.2C presents a Pearson correlation matrix across all rater pairs. All correlations were statistically significant ($p < 0.001$), with coefficients ranging from $r = 0.32$ (Rater 1 and Rater 4) to $r = 0.74$ (Rater 2 and Rater 3). These analyses demonstrate moderate

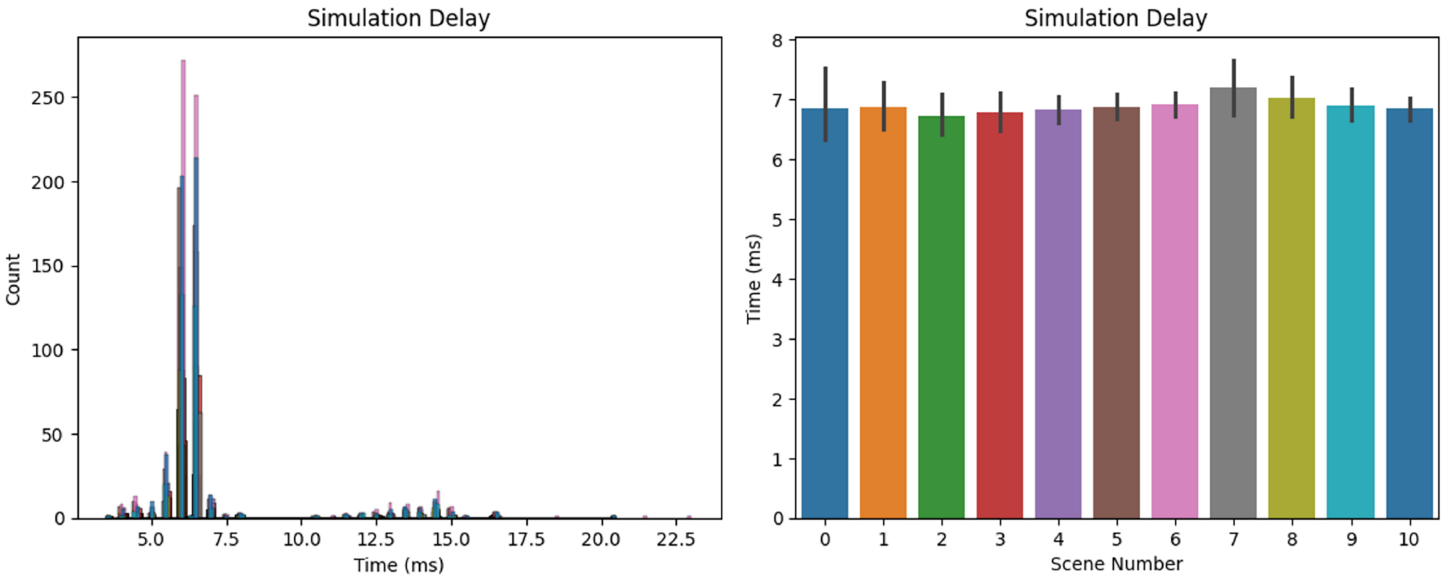

Figure A1.1. Timing validation of gaze-contingent display updates. The histogram (*left*) shows the distribution of delays across trials and scenes. The right panel shows the mean and standard deviation per trial for each scene.

Action (prominent)	Action (background)	Number of people
Grabbing, standing		1
Holding, kneeling		1
Holding, sitting		1
Stealing	Standing	2
Putting on makeup	Opening door	2
Jogging, listening	Walking	2
Repairing, kneeling	Standing	2
Playing basketball		2
Hiding	Looking at phone, walking	2
Fishing, standing		2
Cycling, listening	Pointing	2
Reading, pointing		2
Surprising, sitting		2
Relaxing	Playing soccer	3
Tying shoe, listening	Sitting, talking	3
Relaxing	Looking at phone, sitting	3
Running	Cheering	3
Repairing, kneeling	Looking at phone, sitting	3
Crying	Sitting, talking	3
Waving	Sitting	3
Waving	Walking	3
Looking in backpack, standing	Sitting, looking at phone	3
Running, walking, listening		3
Standing, pointing	Walking	3
Screaming, swinging	Walking	3
Cleaning	Walking	3
Standing, talking		3
Brushing hair	Walking	3
Grabbing, standing	Walking	3
Grabbing, cutting, eating		3
Playing harmonica, sitting		3
Throwing, catching		3
Throwing, defending	Yelling	3
Sitting	Pointing, holding	3
Taking a photo, standing	Walking	3
Standing, pointing		3
Arm wrestling, clapping	Walking	3
Playing card game, grabbing		3
Holding, grabbing, sitting		3
Talking, sitting		3
Sitting, studying	Looking at phone	3
Looking through door	Standing	3
Sitting, looking at phone		3
Holding, clapping	Running	3
Falling	Standing	3
Eating, sitting		3
Grabbing	Eating, talking	3
Recording, talking	Sitting, talking	4
Yoga, studying	Sitting, talking	4
Thinking	Walking	4
Spraying, sitting	Walking	4
Playing table tennis, clapping	Walking	4
Grabbing, jumping	Sitting	4

Table A1.1. List of action phrases and number of people in each social interaction scene.

Action (prominent)	Action (background)	Number of people
Waving, walking		4
Exercising	Talking, sitting, walking	4
Relaxing, studying	Walking	5
Studying, yoga	Walking	5
Thinking	Talking, sitting, studying	5
Tying shoe, pointing	Studying	5
Recording, yelling, studying	Sitting, walking	9

Table A1.1. Continued

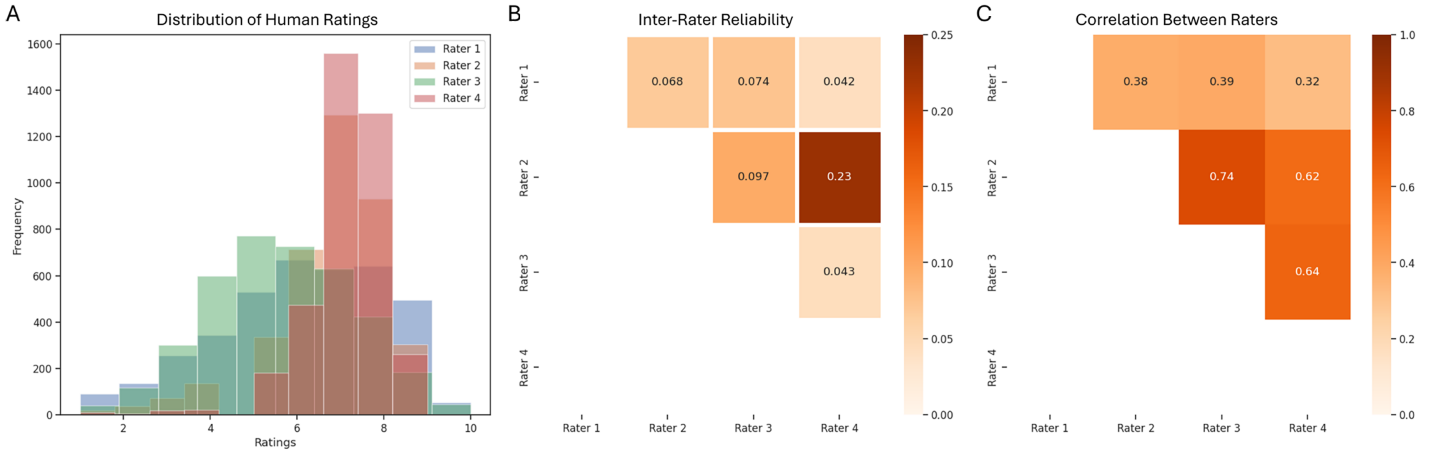

Figure A1.2. (A) Distribution of semantic similarity ratings by each of the four human raters across all 3,840 scene descriptions. (B) Pairwise interrater reliability (IRR) values based on ordinal agreement across all descriptions. Values indicate modest agreement, especially between Rater 2 and the other raters. (C) Pearson correlation matrix between all human raters. All pairwise correlations were significant at $p < 0.001$, with strongest agreement between Rater 2 and Rater 3.

consistency across raters and support the use of average ratings in subsequent analyses.

Stability of scene description performance over time

To assess whether participants improved their descriptions over time, we compared semantic similarity scores from the first and last experimental blocks, separately for each viewing condition. Figure A1.3 shows the average human ratings (top) and GPT-4 cosine similarity scores (bottom) across blocks. A two-way ANOVA found no effect of block number for human ratings ($F(1, 6) = 1.27$, $p = 0.303$) but did show an effect for GPT-4 cosine similarity values ($F(1, 6) = 25.11$, $p = 0.002$) . Post hoc analysis for GPT-4 cosine similarity values revealed that control participant descriptions for social interactions scenes in the first block were significantly lower than control participant descriptions for neutral scenes in the last block ($M = -0.25$, $p = 0.03$, $d = 0.26$), but no other comparisons were significant, suggesting that there was no learning between conditions.

Distribution of saccade amplitudes

Figure A1.4 shows visualization of saccade amplitude distributions for all three viewing conditions and differences in distributions for PVL and CVL. PVL amplitudes were similar to control participants, but CVL amplitudes shifted away from the uninformative locations under the scotoma (more saccades with amplitudes of 3 or larger). Figure A1.4B shows the redistribution of saccades based on each low vision viewing condition compared to control participants.

Saccade metrics for scenes with and without peripheral humans

To further examine the role of peripheral social cues, we analyzed the relationship between first saccade amplitude and latency, stratified by scene type (social interactions vs. neutral), presence of humans in the periphery, and viewing condition (Appendix Figure A1.5). This analysis supports the idea that peripheral social content increases oculomotor uncertainty and planning time, particularly under central vision loss.

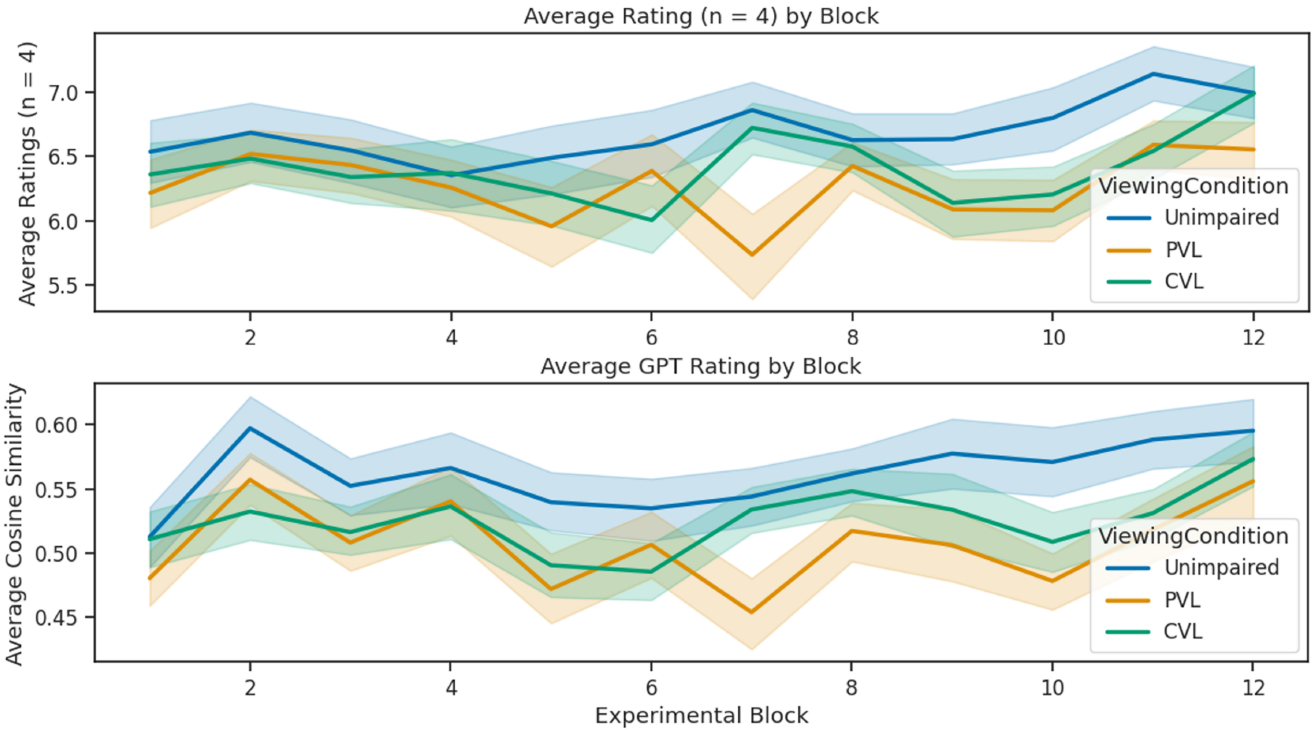

Figure A1.3. Average semantic similarity ratings by block number and viewing condition. Top: human ratings; bottom: GPT-4 cosine similarity. No meaningful learning effects were observed between viewing conditions.

Example heat maps and descriptions

To illustrate the structure and variability of our stimulus set, this section provides representative examples of group-level fixation heat maps and descriptions.

To visualize group-level gaze behavior, fixation locations were transformed into fixation heat maps for each scene. Heat maps were generated by randomly sampling groups of four participants from each viewing condition (control, PVL, CVL), as well as from a separate "control comparability group." Fixation locations were convolved with a Gaussian kernel (39 × 39 pixels, approximately 1° × 1°) to estimate spatial fixation density. Red areas indicate regions of high fixation density. Figures A1.6, A1.7, and A1.8 show example heat maps for one scene across the four groups. Participant descriptions and the average rating and cosine similarity values for each description are listed in a table after each heat map scene example.

These examples illustrate how viewing conditions influenced fixation behavior relative to the spatial distribution of social and task-relevant content.

Heat map correlations with different sizes and grid densities

The heat map method used for Figures 6 and A1.6–A1.8 was repeated to check for differences in effects based on varying parameters for Gaussian Blur (see Figure A1.9). Larger Gaussian kernels (79 × 79; 159 × 159), larger standard deviation (2 and 4), and smaller and larger amounts of bins (30 and 120) were tested. Results were similar to values reported in the main text: The average CVL heat map correlation was always lower than controls and PVL participants. Increasing the standard deviation of the kernel from 1 to 4 led to higher correlation values. Increasing the number of bins resulted in lower correlation values.

Example annotated areas of interest

Each of the 120 video scenes was manually labeled to identify the faces and bodies of humans or animals, as well as critical task-relevant objects. AOIs were created using Make Sense (Skalski, 2019) and saved in JSON format. A critical object was defined as the object (or group of objects) most relevant for accurately describing the scene. AOIs were also categorized as either central or peripheral, based on whether they appeared beyond 5 degrees of visual angle from the initial fixation point. Of the 120 scenes, 61 included people in the periphery (45 social, 16 neutral), and 33 were labeled as containing peripheral critical objects (21 social, 12 neutral). Examples of labeled frames from both social interaction and neutral scenes are shown in Figure A1.10.

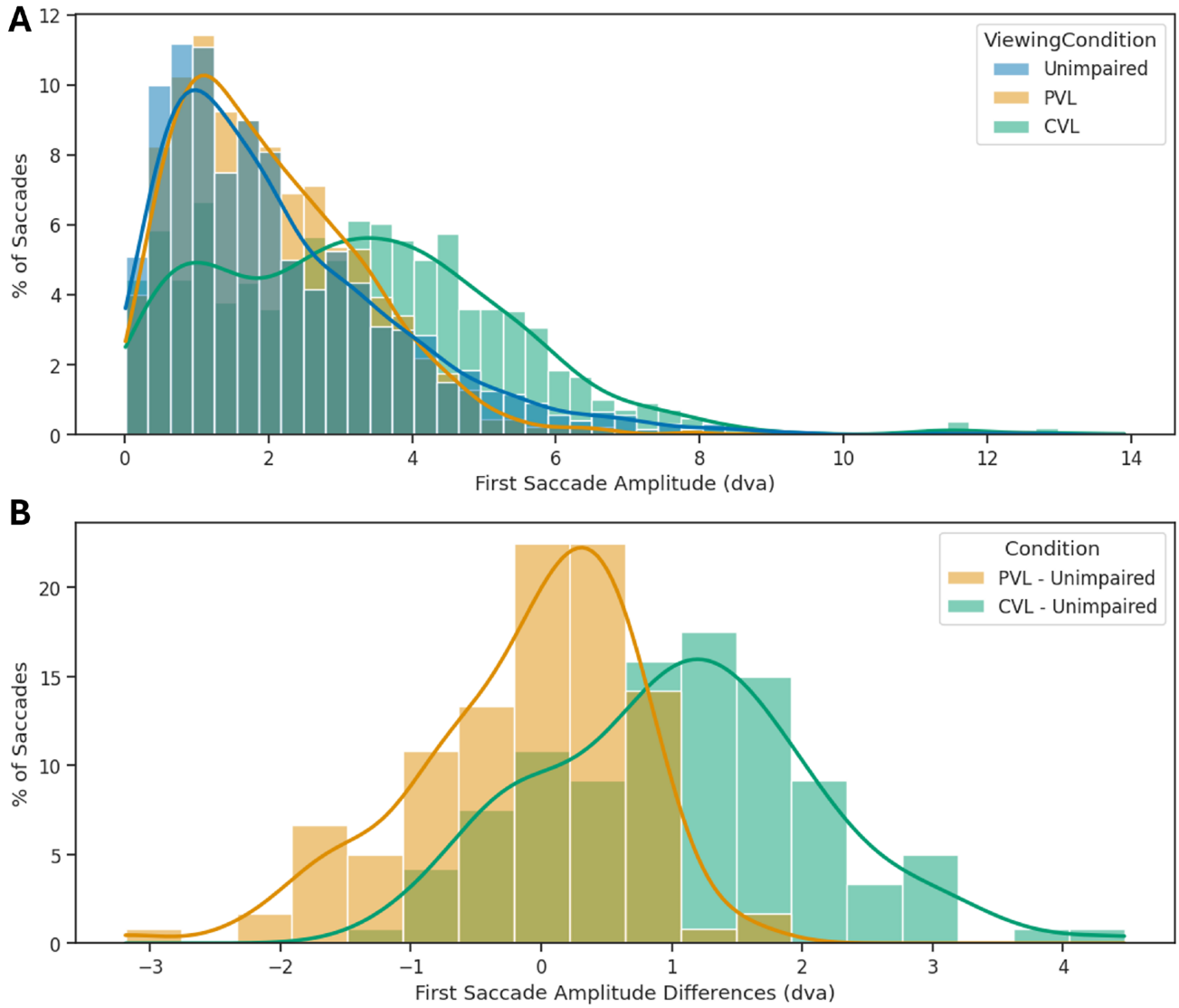

Figure A1.4. (A) Distribution of saccade amplitudes by viewing condition. (B) Difference in distribution of saccade amplitudes between PVL and control participants (orange) and CVL and control participants (green). Positive values indicate larger differences in amplitude between viewing condition and control participants; negative values indicate when participants in the assigned viewing condition made saccades smaller than control participants.

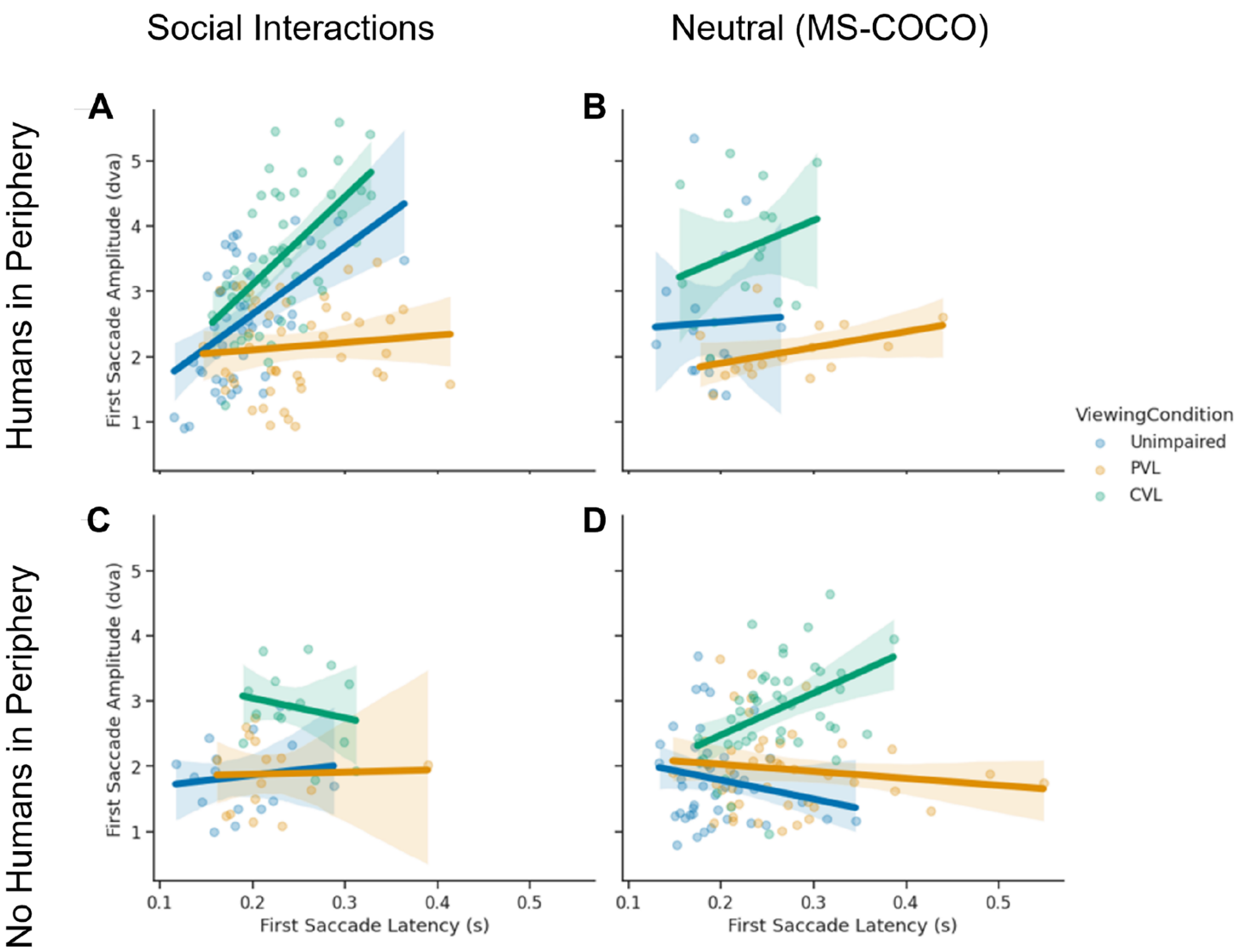

Figure A1.5. Relationship between first saccade amplitude and latency as a function of scene type, peripheral human presence, and viewing condition. In scenes with humans in the periphery (A, B), saccade amplitude and latency were positively correlated, especially for CVL participants, suggesting delayed saccade planning and compensatory eye movements. This relationship was weaker or absent in scenes without peripheral humans (C, D).

Figure A1.6. Example scene with raw fixation data and heat maps by viewing condition: (top left) original scene; (top right) scene with fixation data based on viewing condition; (second row) fixation data for controls, PVL, and CVL; (third row) convolving the fixation data with a Gaussian filter; (fourth row) heat map data with representation of experimental viewing conditions at stimulus onset. For an example trial of the actual experiment, refer to Figure 2.

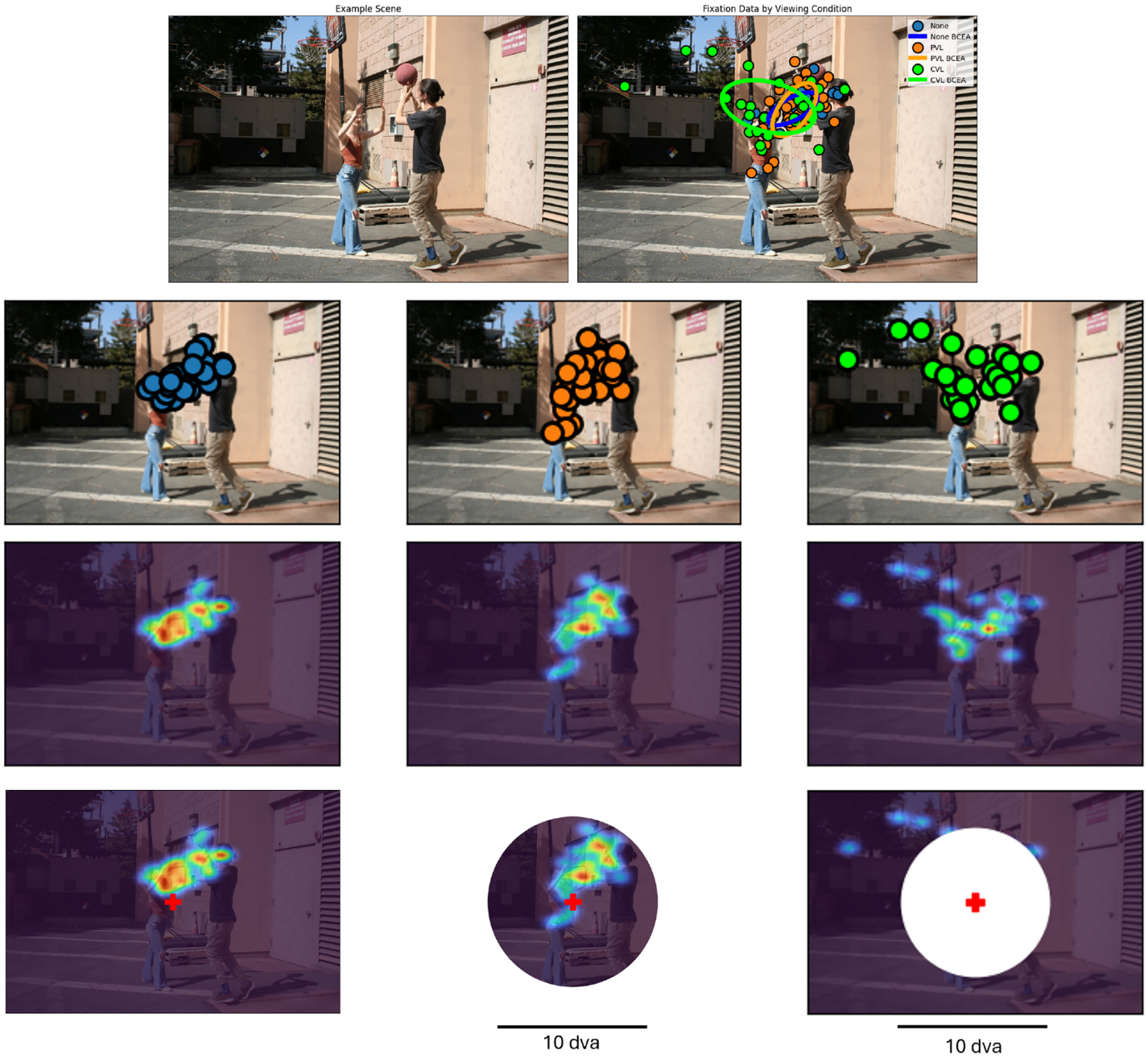

Figure A1.7. Example social interaction scene of people arranged centrally with raw fixation data and heat maps by viewing condition: (top left) original scene; (top right) scene with fixation data based on viewing condition; (second row) fixation data for controls, PVL, and CVL; (third row) convolving the fixation data with a Gaussian filter; (fourth row) heat map data with representation of experimental viewing conditions at stimulus onset. For an example trial of the actual experiment, refer to Figure 2.

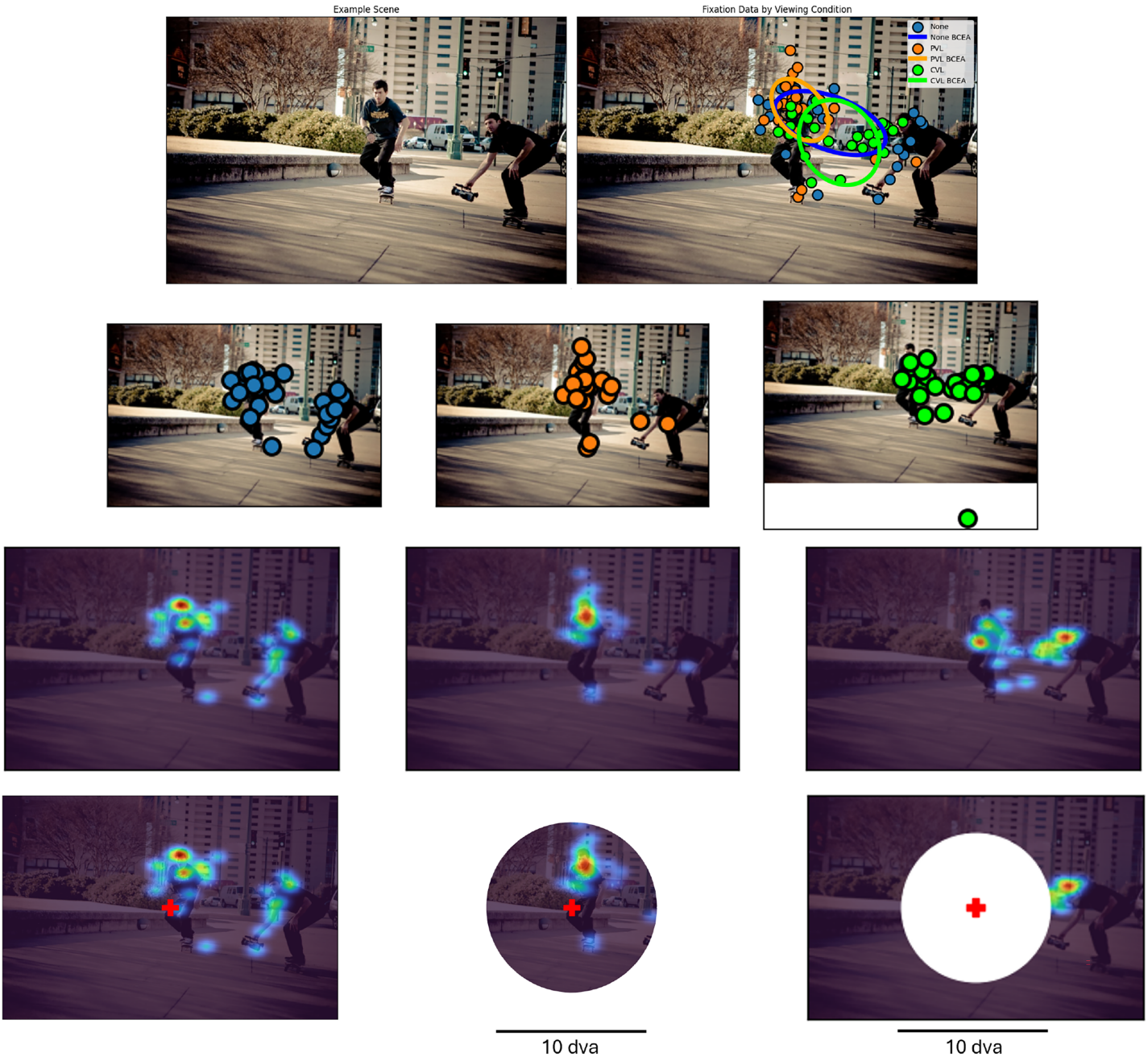

Figure A1.8. Example neutral scene with a person arranged peripherally with raw fixation data and heat maps by viewing condition: (top left) original scene; (top right) scene with fixation data based on viewing condition; (second row) fixation data for controls, PVL, and CVL; (third row) convolving the fixation data with a Gaussian filter; (fourth row) heat map data with representation of experimental viewing conditions at stimulus onset. For an example trial of the actual experiment, refer to Figure 2.

Condition	Response	Rating	GPT
Unimpaired	A person is attempting to hit another with a blue bat	7.25	0.573
Unimpaired	Guy and girl standing on the road next to a car while the girl is holding a blue baseball bat towards the direction of the guy	6.75	0.477
Unimpaired	There are two people outside on the road and one of them is a girl holding a baseball bat	6.25	0.551
Unimpaired	A girl is swinging a blue bat and another girl is watching her. There is a white car in the background	6.25	0.443
Unimpaired	There is a woman with a baseball bat standing on the street outside looking at another person	6	0.559
Unimpaired	A girl in the middle of swinging a blue bat	5.25	0.419
Unimpaired	A girl was holding a blue baseball bat, playing with a boy in front of a white car	5	0.404
Unimpaired	A woman is holding a blue baseball bat in front of a car	5	0.438
Unimpaired	A girl with a blue baseball bat standing near a white car that is parked on a curb, other person standing across from her is ready to throw her a pitch	4.25	0.458
Unimpaired	People playing baseball on the road	3.5	0.300
PVL	There is a girl in a white crewneck standing in front of a white car. She is holding a blue baseball bat and appears to be swinging it. Behind it there is a brown or tan building with grass in front	5.75	0.440
PVL	There is a girl holding a blue baseball bat in the middle of the road next to a white car	5.5	0.432
PVL	A young girl stands outdoors with a bat and prepares to swing, with buildings in the back	5.5	0.394
PVL	A woman holding a bat in preperation to swing. Behind her is a white car	5.5	0.480
PVL	There is a woman holding up a bat infront of a white car	5.25	0.465
PVL	There is a woman holding a baseball bat in front of a car	5.25	0.496
PVL	There is a girl in front of a white car swinging a blue bat in the street	5	0.417
PVL	There is a woman playing baseball in front of a white car	4.5	0.321
PVL	Women holding something up in the air and she is standing in front of a house	3.75	0.222
PVL	There is a woman who wears white shirts standing in front of a white car and trying to catch a ball	3.25	0.233
PVL	There is a building with a lawn in front of it. A white car is parked in front of the building. People are walking by the building, and there is a girl in a white sweatshirt and black pants walking next to the car	3	0.204
CVL	Two women standing in the center of the street. One woman was holding a blue baseball bat and appeared to be doing a striking down motion	6.75	0.435
CVL	A woman holding a baseball bat next to another woman in the street	6	0.504
CVL	I see two people on the street near a white car arguing facing each other, one was a woman and the other a man	5	0.334
CVL	Two people are standing in the street. One is holding something in her hand near her head. Next to them is a car	4.75	0.274
CVL	Two girls standing in front of a car outside. oNe is throwing something blue at the other	4.5	0.298
CVL	This scene is in a neighboorhood. The girl on the right is blonde is is throwing something at the girl on the left with brown hair	4.25	0.358
CVL	Two people standing on a street with a white car in the background	4.25	0.258
CVL	There is a white car by the sidewalk. Two people are standing on the sides of the street	4	0.205
CVL	There is aperson talking to someone in front of a building near cars	3.75	0.290
CVL	A white car was parked on the street. Two people were standing near by	3.5	0.216
CVL	There was a white car on the right side of the photo. There were two people walking across the photo. One was a boy in the red shirt. he was on the left side. The other person was a blonde girl. She was walking toward the left. This was on the road. The car was parked	3.5	0.230

Condition	Response	Rating	GPT
Unimpaired	A guy and girl playing basketball. The guy is about to shoot the ball and the girl has her hands up trying to block him	9	0.677
Unimpaired	There is a man and a woman playing basketball outside. The man has the ball and is facing the basketball hoop and the woman has her arms up in front of him	8.25	0.756
Unimpaired	A girl wearing an orange tank top and jeans is standing near a guy and she has her hands up in front of her face as though she is about to grab the ball from him. He is holding a basketball and there is a basketball hoop in the background	8.25	0.534
Unimpaired	A girl and boy are playing basketball. He is holding the ball above his head	7.25	0.482
Unimpaired	A man and a women are playing basketball on the outside of a building	7	0.676
Unimpaired	Two people dressed in casual clothing are playing with a basketball and a basketball hoop outside of a building	7	0.639
Unimpaired	Two people, a girl and a boy were playing a red ball. They were on pavement, in a parking lot next to a building	6.75	0.441
Unimpaired	Two people playing a game with one of them kinda throwing a basketball	6.75	0.558
Unimpaired	There is a man and woman playing with a basketball	6.5	0.611
Unimpaired	Two people playing basketball	6.25	0.623
PVL	A boy in a black shirt and a girl in a red tank top play basketball. The boy is about to shoot the basketball at the hoop on the left. The girl is standing between him and the hoop, trying to block him from shooting	9	0.636
PVL	Man shooting a basketball with a women in front of him trying to block	8.5	0.686
PVL	A boy and girl are playing basketball. He is about to shoot while the girl is attempting to block him	8.5	0.699
PVL	A man is holding a basketball ready to shoot it. A woman is holding her hands up in attempt to block him. Behind her is the hoop	8.25	0.744
PVL	There is a boy and a girl passing a basketball to each other. There is a basketball hoop in the background and it seems like they are outside	7.75	0.581
PVL	A young man and woman toss eachother a red ball outdoors near a building	6.75	0.455
PVL	There is a man and a woman standing in front of a black table. They are passing a basketball between them. There is an orange building behind them. The girl is wearing an orange shirt and blue jeans and the boy is wearing black pants and a black shirt	6.5	0.587
PVL	There is a man and a woman playing basketball	6.25	0.637
PVL	There was a man bouncing a ball upwards towards a woman	6	0.457
PVL	There is a guy and a girl and there is a basketball in the air between them	5.75	0.541
PVL	There are two women playing basketball in the playground	5.25	0.537
CVL	A man and a woman playing basketball the man is shooting while the woman is blocking him	9	0.765
CVL	There are two people playing basketball behind a building. One person is going up for a shot while the other person is trying to defend them	8.5	0.667
CVL	There were two people playing basketball behind a building. The man was much taller than the woman. The man had the basketball in his hand. It was daytime	7.75	0.650
CVL	There are two people playing basketball in a makeshift court behind a building. One person is significantly taller than the other who is a girl. She is wearing a salmon colored top. She is reaching upwards towards the ball which is in the hand of the taller person	7.75	0.599
CVL	This image is of two people possibly playing basketball. The guy is holding something in the air and the girl is trying to reach for it	7.25	0.561
CVL	There are two people playing basketball behind a building	7.25	0.603
CVL	A man and a woman playing basketball	7	0.667

Condition	Response	Rating	GPT
CVL	I see two people, a guy and a girl playing with a ball that is orange and has black stripes with a hoop behind them on the street	6.75	0.563
CVL	Two people were playing basketball. The boy in the gray shirt was shooting the ball	6.75	0.567
CVL	There are two people playing basketball behind a building. One of them is holding the ball in the air and the other one is trying to catch the ball	6.25	0.589
Unimpaired	A stakeboarder was riding a skateboard while somebody else filmed him with a camera	8.5	0.653
Unimpaired	A person is riding a skateboard while the other person is recording his ride	8.5	0.714
Unimpaired	A man is videotaping the other man skateboarding	8.25	0.656
Unimpaired	There is a man riding his skateboard outside and there is another man following and filming his skills	8	0.667
Unimpaired	A skateboarder about to do a trick and another one filming it	7.75	0.656
Unimpaired	Two men riding skateboards on a sidewalk through the city, one of them is performing a jump	7.5	0.607
Unimpaired	Person skateboarding outside	7	0.547
Unimpaired	Two guys skateboarding next to each other	6.75	0.541
Unimpaired	Two men dressed in dark colors are skateboarding in the city	6.25	0.627
Unimpaired	There are two men rollerskating outside	5.75	0.503
PVL	A man in the middle is skating. A man on the right is also skating while filming the man in the middle	8.25	0.568
PVL	There is a men doing skateboarding tricks	6.5	0.440
PVL	There are two guys riding skateboards	6.5	0.555
PVL	There is a man wearing a black shirt and black pants on a skateboard on the sidewalk. There are buildings behind him and a road with a streetlight	6.25	0.408
PVL	Man riding a skateboard on a sidewalk	6.25	0.485
PVL	There is a man on a skateboard with a stoplight and some tall buildings behind him	6	0.465
PVL	There is a boy riding a skateboard. He has dark hair and a mustache	6	0.421
PVL	There is a man in dark colored clothing skateboarding down a sidewalk. There are tall tan buildings behind him that have a parking lot in front	5.5	0.460
PVL	There is a man in a blue jacket riding a skateboard	5.25	0.438
PVL	A young man in black skates in an oudoor city setting	5	0.521
PVL	There is man who wears blak shirt running on the road	3.25	0.150
CVL	One man is riding a skateboard appearing to be doing a trick. Another man is standing next to him holding a camera, recording the man doing his skateboard tricks	8.25	0.687
CVL	Two men were skate boarding. The one on the right was videotaping the one on the left. They were located in a city setting	8	0.737
CVL	There are two guys. One of them is skateboarding on a wooden surface and the other person is holding a camera below him, filming him	7.75	0.734
CVL	I see two guys riding on skateboards beside each other in the city, both are wearing dark colored clothing	7.5	0.619
CVL	Two men skateboarding on the sidewalk	7	0.533
CVL	This scene is of 2 people. There is a guy skating and the other guy is recording the guy skating	6.5	0.678
CVL	Two men are skateboarding on a wooden surface	6.5	0.505
CVL	Two guys riding on skateboards, one has something in his hand	6.25	0.582
CVL	Two men skateboarding	6.25	0.563
CVL	Two people skateboarding in the city. They are on the sidewalk	6.25	0.688
CVL	There were two skateboarders. One was on his skateboard. The other was looking at the ground, but not on his skateboard. They were near a curb but still on the sidewalk	5.5	0.494

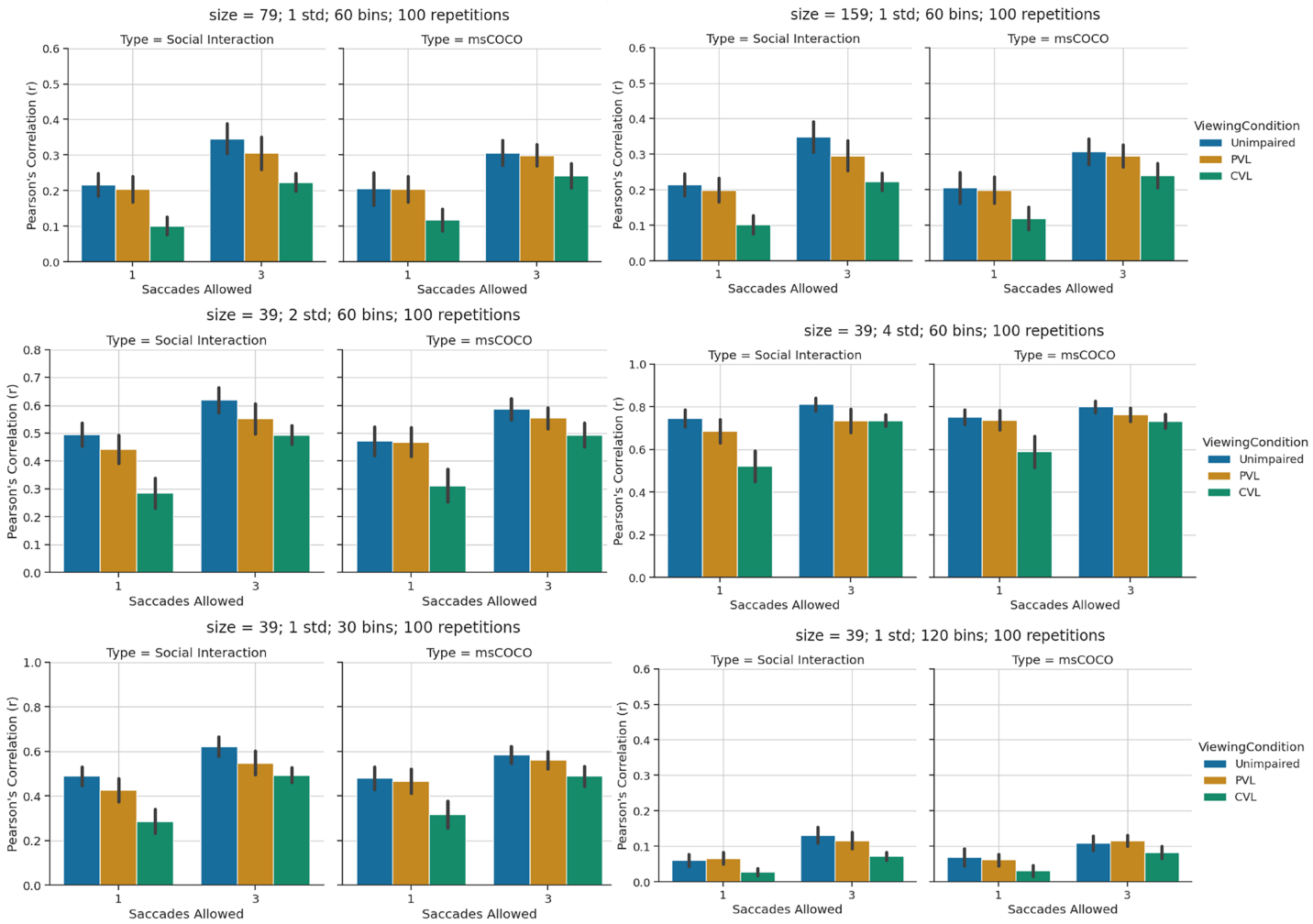

Figure A1.9. Heat map correlation results for different kernel sizes, standard deviations, and number of bins.

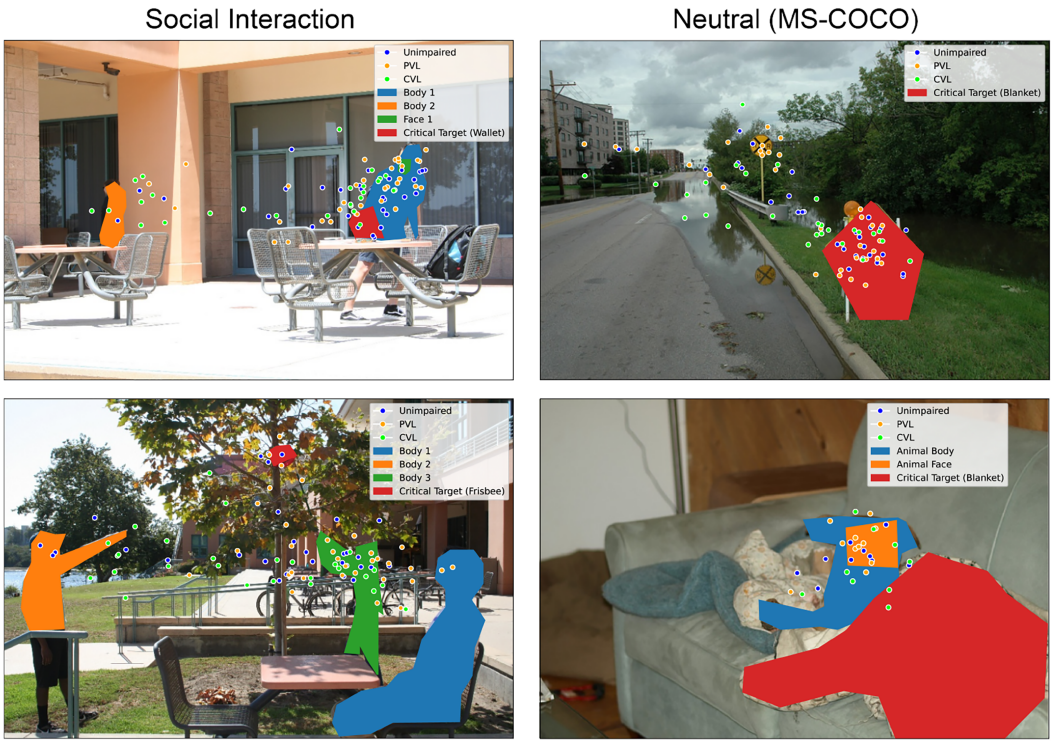

Figure A1.10. Example scenes with fixation locations and labels for humans and critical objects, for social scenes (*left*) and neutral scenes (*right*).