Modeling human social vision with cinematic stimuli: An integrative approach

Severi Santavirta Turku PET Centre, University of Turku, Turku, Finland
Turku University Hospital, Turku, Finland

Birgitta Paranko Turku PET Centre, University of Turku, Turku, Finland
Turku University Hospital, Turku, Finland

Kerttu Seppälä Turku PET Centre, University of Turku, Turku, Finland
Turku University Hospital, Turku, Finland
Department of Medical Physics, Turku University Hospital, Turku, Finland

Jukka Hyönä Department of Psychology and Speech-Language Pathology, University of Turku, Turku, Finland

Lauri Nummenmaa Turku PET Centre, University of Turku, Turku, Finland
Turku University Hospital, Turku, Finland
Department of Psychology and Speech-Language Pathology, University of Turku, Turku, Finland

Sociability is central for humans. Visual information ranging from low-level physical features (e.g., luminance) to mid-level semantic information (e.g., face recognition) and high-level social inference (e.g., emotional valence of social interactions) is constantly sampled for navigating the social world. In this study, we utilized large-scale eye tracking during natural vision for mapping how different levels of visual information guide the perception of socially relevant features (social vision) simultaneously. In three experiments, participants (N = 166) watched full-length films and short movie clips with varying social content (total duration: 193 minutes) during eye tracking. To model the association between perceptual features and spatiotemporal gaze parameters (gaze position, gaze synchronization, pupil size and blinking), we extracted 39 stimulus features from the movies, including low-level audiovisual features (e.g., luminance, motion), presence and location of mid-level semantic categories (e.g., faces, objects), and high-level social information (e.g., body movements, pleasantness). Integrative analysis techniques with cross-validation were developed to simultaneously associate the perceptual features with the gaze behavior. Pupil size was modulated by luminance, scene cuts, and emotional arousal while gaze position was most accurately predicted by a combination of the presence of human faces, local motion, and entropy. Faces and eyes were prioritized over other semantic categories, and blinking rate decreased during periods of attentional engagement. Altogether, the results show that human social vision is primarily guided by low-level physical features and mid-level semantic categories, while high-level social features such as emotional arousal primarily modulate pupillary responses.

Introduction

Visual information about other people is central to human social interaction. Visual information is constantly sampled from the environment by controlling gaze position, modulating visual sampling by adjusting fixation frequency, segmenting visual input into chunks through blinking, and controlling the number of photons landing on the retina through pupillary control. For example, we quickly recognize affective content from scenes (Nummenmaa, Hyönä, & Calvo, 2010) and facial expressions (Calvo & Nummenmaa, 2008) and preferentially allocate attention to social signals ranging from faces to reproductive cues (Morrisey, Hofrichter, & Rutherford, 2019). This indicates that human vision is particularly tuned to filter and process social information from the environment.

Citation: Santavirta, S., Paranko, B., Seppälä, K., Hyönä, J., & Nummenmaa, L. (2026). Modeling human social vision with cinematic stimuli: An integrative approach. *Journal of Vision*, *26*(5):9, 1–26, https://doi.org/10.1167/jov.26.5.9.

https://doi.org/10.1167/jov.26.5.9 Received March 21, 2025; published May 27, 2026 ISSN 1534-7362

The abundance of information in the real-life visual world is a major challenge for controlled eye-tracking studies that investigate how external stimulus features modulate the gaze behavior. Traditionally, controlled studies limit the available visual information by using static images and isolate the investigated phenomena (e.g., emotion, face perception) from overlapping confounding features. However, these simplifications may overlook important aspects of real-life visual sampling where multiple external features occur at different yet overlapping time scales, such as parallel processing of sensory features and monitoring the ever-changing locations of objects in dynamic scenes. It has thus been debated whether the results from such simplified conditions transfer to real-life vision operating in a dynamic social environment (Williams & Castelhano, 2019), particularly because gaze is differently influenced by static versus dynamic stimuli (Dorr, Martinetz, Gegenfurtner, & Barth, 2010). Furthermore, different external stimulus features may have unique, additive, or interactive associations with spatial and temporal gaze parameters (e.g., pupillary response, gaze position, blinking), yet these features are rarely analyzed in unitary models.

Thus, our understanding of how multiple external features modulate the human gaze simultaneously in dynamic environments is currently limited. By using uncontrolled movie watching during eye tracking, the present study aims to expand the understanding of this topic from the social vision perspective. Due to serious concerns about the replicability of results in psychology (Open Science Collaboration, 2015), it is critical to conduct large-scale studies using cross-validated models and report replicable findings across independent samples. Cross-validation also enables estimating the predictive power of the trained models, which is especially important for estimating the effect sizes within psychological science, where the prediction power of models is usually moderate at best. Hence, this study incorporates novel analytical and data-driven approaches by extracting a multitude of stimulus features and then using them for predicting different eye movement parameters, ranging from gaze location to fixation and saccade metrics and intersubject synchronization of eye movements.

Stimulus features that may affect how we attend dynamic social scenes require different levels of sensory and cognitive processing. Simple physical properties, such as luminance or audio intensity, are quickly processed in the primary visual and auditory cortices of the human brain. Recognizing a human face already requires pattern recognition, and faces are represented in associative brain cortices such as the lateral occipitotemporal cortex and the fusiform gyrus, which are closely connected to the primary visual cortex (Kessler et al., 2011; Pitcher, Dilks, Saxe, Triantafyllou, & Kanwisher, 2011). Interpreting social and emotional contents requires more detailed associative processing, which is undertaken in the brain's social and emotional networks, primarily in polysensory temporal and frontal cortices (Saarimäki et al., 2025; Santavirta et al., 2023). Studies have revealed hierarchical temporal receptive windows in the brain, ranging from occipitotemporal sensory processes to frontoparietal cortices and temporal associative areas integrating information over longer time scales (Hasson, Yang, Vallines, Heeger, & Rubin, 2008). For example, object categorization precedes its affective evaluation (Nummenmaa et al., 2010). EEG measurements indicate that brain responses are first associated with visual features of a stimulus, followed by observed actions and later by socio-affective features, which support hierarchical processing (Dima, Tomita, Honey, & Isik, 2022). In the present study, we conceptualized the stimulus features that may guide visual sampling in three levels based on the cognitive processing hierarchy, from (1) low-level feature processing (e.g., luminance) to (2) mid-level semantic categorization warranting associative processing (e.g., faces) and to (3) complex social inference (e.g., scene pleasantness) requiring contextual processing and accessing existing social and emotional information stored into long-term memory. This categorization allows us to investigate whether human gaze behavior is merely associated with changes in low- and mid-level stimulus features or whether also high-level social information has unique associations with gaze orienting, pupil size, and blinking in dynamic social scenes.

While low-level physical properties and semantic object categories can be extracted using mathematical algorithms, higher-level social information requires human evaluation. There have been multiple attempts at characterizing the main dimensions or categories in human social perception (Abele & Wojciszke, 2014; Fiske, 2018; Oosterhof & Todorov, 2008; Osgood & Suci, 1955; Parrigon, Woo, Tay, & Wang, 2017; Rauthmann et al., 2014; Sutherland & Young, 2022). Recently, behavioral work using naturalistic audiovisual stimuli has concluded that social perception is aligned along a limited set of main axes or dimensions, such as the perception of pleasant versus unpleasant situations, social hierarchy, or communication (Santavirta, Malén, Erdemli, & Nummenmaa, 2024), and functional neuroimaging has established that large-scale occipitotemporal brain networks are involved in processing of these features (Lahnakoski et al., 2012; Santavirta et al., 2023). It, however, remains unresolved whether these perceptually relevant social features associate with human gaze behavior.

This study aims to simultaneously model human gaze direction, fixation and saccade properties, pupil size changes, and blinking behavior in dynamic social scenes. During dynamic vision, the gaze positions synchronize between individuals (Dorr et al., 2010;

Franchak, Heeger, Hasson, & Adolph, 2016; Smith & Mital, 2013). The synchronization of gaze indicates shared attention, which is useful for a common understanding of events and for cooperation. Synchronization can be measured using intersubject correlation (ISC), which can be as high as 0.4 to 0.6 during movie watching (Hasson et al., 2008; Wang, Freeman, Merriam, Hasson, & Heeger, 2012), indicating high attentional coherence and a major influence of the stimulus properties in directing the gaze. A large body of studies has addressed how gaze is directed by external or intrinsic features during scene perception, and both bottom-up and top-down models for gaze control have been proposed (Henderson, 2003; Schütt, Rothkegel, Trukenbrod, Engbert, & Wichmann, 2019). Bottom-up models predict gaze direction by computing weighted saliency maps from color, intensity, and orientation information (Itti & Koch, 2000). However, such bottom-up models cannot explain the strong human attentional preferences for human faces and eyes that are not physically salient based on pure bottom-up saliency maps or task-dependent top-down guidance of visual attention (Birmingham, Bischof, & Kingstone, 2008; Birmingham, Bischof, & Kingstone, 2009). Additionally, oculomotor and centrality biases decrease the likelihood of orienting to peripheral locations (Dorr et al., 2010; Tatler & Vincent, 2009). The debate whether gaze directions are best modeled with low-level saliency or object-based models is ongoing (Stoll, Thrun, Nuthmann, & Einhäuser, 2015), and the latest evidence suggests that saliency and object information together yield better predictive models than either of them alone (Nuthmann, Schütz, & Einhäuser, 2020; Roth, Rolfs, Hellwich, & Obermayer, 2023). Recent deep learning models for gaze patterns can yield high predictive performance without predefined features for images (Cornia, Baraldi, Serra, & Cucchiara, 2018; Lou, Lin, Marshall, Saupe, & Liu, 2022) and videos (Bellitto et al., 2021; Jain et al., 2021). However, as the models do not rely on predefined stimulus features, it is difficult to know exactly what information is used to predict scene saliency. Indirect methods can be utilized to interpret the outputs of these complex models (Hayes & Henderson, 2021), but currently they offer little theoretical insight about gaze behavior.

In addition to gaze orienting, pupil size changes are dynamically aligned with external and internal state changes. The pupil automatically regulates the amount of light reaching the retina by contracting in luminous conditions, but pupil size also fluctuates as a function of internal states, such as emotional arousal and cognitive effort. Seminal results indicated that the pupil dilates when watching pleasant images (Hess & Polt, 1960), while recent studies have shown that pleasant and unpleasant emotions trigger pupil dilation when viewing images (Bradley, Miccoli, Escrig, & Lang, 2008), listening to sounds (Oliva & Anikin, 2018; Partala & Surakka, 2003), and imagining emotional episodes (R. R. Henderson, Bradley, & Lang, 2018). This arousal-driven pupillary response is usually stronger for unpleasant than pleasant stimuli (Babiker, Faye, & Malik, 2013; Kawai, Takano, & Nakamura, 2013). Pupil dilation is also associated with cognitive effort in various cognitive control tasks (Ayres, Lee, Paas, & van Merriënboer, 2021; Hyönä, Tommola, & Alaja, 1995; Kahneman & Beatty, 1966; van der Wel & van Steenbergen, 2018). In turn, pupils have been found to constrict as a function of the attractiveness of faces and natural scenes (Liao, Kashino, & Shimojo, 2021). Pupil constriction predicts how well people memorize images and reveals the novelty of a scene (Naber, Frässle, Rutishauser, & Einhäuser, 2013). Pupils seem to constrict even during sudden transitions of simple stimuli in isoluminous conditions (Kimura, Abe, & Goryo, 2014). These luminance-independent pupillary responses are likely mediated by the adrenergic and cholinergic neurotransmitter systems that are engaged during both emotional arousal and cognitive effort to maximize attentional prioritization and arousal levels in response to environmental demands (Joshi, Li, Kalwani, & Gold, 2016; Reimer et al., 2016). Real-life pupillary response is a complex combination of pupillary light reflex and higher-order effects such as those induced by emotions (Cherng, Baird, Chen, & Wang, 2020; Steinhauer, Siegle, Condray, & Pless, 2004), but it is not yet established how multiple factors simultaneously modulate the pupil size during natural vision.

Eyeblinks keep the eyes lubricated and remove irritants from the eye surface. The automatic eyeblink reflex is also a part of the startle response, where the orbicularis oculi muscle contracts rapidly after brief and intense auditory, visual, or tactile stimuli (Grillon & Baas, 2003). Blinking is also influenced by cognitive and affective factors. Blinking and pupil dilation are suggested to be the most sensitive physiological measures of cognitive load (Ayres et al., 2021). In a naturalistic study of *Mastermind* TV quiz contestants, blinking was strongly modulated by the task so that more blinks occurred at attentional or cognitive breakpoints (Wyly et al., 2024). In addition, blink rate varies as a function of the emotional content, indicating motivational relevance and biological significance of the stimuli (Maffei & Angrilli, 2019). Blinking is found to synchronize between participants during shared dynamic stimuli, indicating shared attention (Nakano, Yamamoto, Kitajo, Takahashi, & Kitazawa, 2009); moreover, blink synchronization is higher among participants who are more interested in the stimuli (Nakano & Miyazaki, 2019). Thus, blinking reflects cognitive load, attention, and interest levels of the participants and possibly the emotional context. Functional neuroimaging results have also established a link between blinking and attentional engagement, as decreased brain activity in the dorsal attention network

after blink onset has been identified simultaneously with increased activity in the default-mode network (Nakano, Kato, Morito, Itoi, & Kitazawa, 2013).

The current study

The objective of this study was to establish an integrated framework for linking spatial and temporal aspects of gaze behavior to low-, mid-, and high-level scene features when watching dynamic social scenes. We studied moment-to-moment gaze positions, pupillary response, and blinking behavior while participants freely viewed social movies. We aimed to simultaneously investigate the contributions of low-level physical features, semantic perceptual features, and high-level perceived social information on the spatial and temporal dynamics of human social vision. Movie scenes were utilized as dynamic stimuli since they depict highly engaging social and emotional contents and effectively synchronize human gaze positions (Hasson et al., 2008), which makes them ideal naturalistic stimuli used in the laboratory (Adolphs, Nummenmaa, Todorov, & Haxby, 2016; Saarimäki, 2021).

We conducted three experiments with a total of 166 participants and 193 minutes of movie stimuli to allow replicability testing for results across the datasets. A total of 39 different stimulus features from low-level audiovisual properties (e.g., luminance, motion) to mid-level semantic objects (e.g., eyes, faces, and objects) and high-level perceived socioemotional features (e.g., pleasantness, talking) were dynamically extracted from the movies. A stimulus model including 16 predictors was generated from the extracted stimulus features after studying the covariance structure of the original 39 features. Four main analysis techniques (total gaze time analysis, multistep regression analysis, gaze prediction analysis, and scene cut effect analysis) were developed to link the extracted stimulus features with the eye-tracking parameters. Replicability of the results was ensured with cross-validation across datasets.

The results show that human social vision is primarily driven by low-level physical features and mid-level semantic categories, while high-level social features, such as emotional arousal, primarily predict pupillary responses. Importantly, gaze direction, pupillary responses, and blinking all indexed unique response patterns in relation to external stimulus features during dynamic social scenes.

Materials and methods

Experimental design

To investigate how the visual perceptual system is modulated by dynamic social contexts, we set up three independent eye-tracking experiments, in which participants watched different movie stimuli rich in social interaction. Three experiments were conducted to allow cross-validation of the results across the independent dataset with semantically different contents and measuring out-of-sample prediction performance. Different participants were recruited for each experiment. In Experiment 1, we used our previously validated socioemotional "localizer" paradigm that allows us to present variable social content (Karjalainen et al., 2017; Karjalainen et al., 2018; Lahnakoski et al., 2012; Nummenmaa et al., 2021; Santavirta et al., 2023). The experimental design and stimulus selection were described in detail in the original study with this setup (Lahnakoski et al., 2012). Briefly, the participants viewed a set of 68 movie clips (median duration: 11.2 s, range: 5.3–27.8 s, total duration: 14 min 26 s) that have been curated to contain large variability of social and emotional content. The videos were extracted from mainstream Hollywood movies with audio tracks in English. See Supplementary Table SI-1 for short descriptions of the movie clips. In Experiment 2, the participants watched a Finnish historical drama film *Käsky* (Louhimies, 2008; 70 min 14 s), and in Experiment 3, the participants watched a full-length horror movie *The Conjuring 2* (Wan, 2016; 109 min 3 s). The total stimulus duration of the stimuli was 193 minutes 43 seconds in the three independent experiments. The variability of the movie stimuli ensures that the findings would generalize across different social perceptual contexts, at least within cinematic stimuli.

Participants, eye-tracking data, and quality control

A total of 166 participants were recruited (Experiment 1: 110 participants, Experiment 2: 28 participants, Experiment 3: 28 participants), and the data were collected in Turku, Finland, between 2019 and 2021. Normal or corrected-to-normal vision was required. Participants gave informed consent prior to

Experiment	Subjects	M/F	Mean age (range)
Experiment 1, movie clips	106	40/66	27.1 (19–74)
Experiment 2, feature film	21	2/19	23.6 (19–38)
Experiment 3, horror movie	24	5/19	27.0 (19–57)

Table 1. Demographic information of the participant sample.

taking part in the experiment. Seven participants were excluded due to incomplete data (three for incomplete data and four for corrupted data files resulting in partial data loss). We also excluded participants from the analyses based on data-driven quality control of the eye-tracking data. We first calculated the total time of fixations and blinks as well as the total time of gaze within the presentation monitor (time in presentation area/total experiment duration) for each participant. Then we fitted beta distributions to these data and identified the 2% probability cutoff points from the long tail of the fitted distributions (Supplementary Figure SI-6). The participants whose total fixation time, total blink time, or total gaze time fell into the 2% probability in the long tail of the distribution were considered outliers, and their data were excluded. Four participants were excluded because they had both atypically low total fixation times and high blink durations (fixations less than 76% and blinks over 12% of the total time), two participants based on short fixation time alone (fixations less than 76% of the total time), and two participants based on short total time within video area (gaze position within video area less than 93% of the time). The final sample thus included 151 out of 166 participants (Table 1).

Eye movement recordings and data preprocessing

In Experiment 1, short movie clips whose contents were unrelated to each other were presented to the participants in fixed, initially randomized order without breaks to enable synchronization analyses. The eye tracker was calibrated and validated using a 5-point calibration, and validation was repeated with 1 point before the experiment. Validation was successful if gaze position error was below 1°. Validation was repeated three times during the experiment (after every 17th movie clip). The stimuli were presented with a 27-in. Retina 5K monitor at a 90-cm distance from the eyes. The eye-tracking data were collected with the SR EyeLink 1000 Plus (SR Research, Ontario, Canada) eye tracker with the following setup: v5.15 January 24, 2018; eyes: right; file filtering level: extra; and pupil tracking algorithm: Centroid. In Experiments 2 and 3, the movie was presented in short segments in a fixed chronological order (Experiment 2: 26 trials, median trial duration: 159 s, range: 68–256 s; Experiment 3: 30 trials, median trial duration: 215 s, range: 146–300 s) to allow repeated drift correction and camera calibration throughout the experiment and to maximize participant comfort. Before each trial (i.e., movie segment), a 5-point calibration and validation and a subsequent 1-point validation (error <1°) were conducted. The stimuli were presented with a 24-in. BenQ XL2411Z monitor at a 70-cm distance from the eyes. The eye-tracking data were collected with the EyeLink 1000 (SR Research) eye tracker with the following setup: v4.594 July 6, 2012; eyes: right; file filtering level: extra; and pupil tracking algorithm: Ellipse. Fixation and saccade reports were generated with EyeLink DataViewer 4.1.1 software (https://www.sr-research.com/data-viewer/). Fixations shorter than 80 ms were considered unreliable, and previous reliable fixation was extrapolated to continue until the next reliable fixation to create a continuous time series of fixation information.

Eye-tracking parameters

We extracted pupil size, gaze position, saccades, and blink timings from the EyeLink DataViewer reports. The outcome variables in the regression analysis (see section "Regression analysis") were pupil size, intersubject correlation of gaze (ISC), fixation rate, and blink rate. ISC was calculated dynamically to index the moment-to-moment gaze position synchronization across subjects using eISC-toolbox for MATLAB (Nummenmaa et al., 2014). ISC was based on first computing participant-specific fixation heatmaps in predefined time windows. Then, momentary ISC was calculated as a spatial correlation of the fixation heatmaps for all possible pairs of subjects. Finally, for each subject and time window, a single ISC value was obtained by averaging the ISC values of all pairs involving that subject. Hence, ISC is a subject-specific measure estimating how strongly the gaze of a subject is synchronized with others, and it changes dynamically in time. Fixation and blink rates were calculated separately for each participant in specific time windows. For the gaze prediction analysis (see section "Gaze prediction analysis"), population average gaze heatmaps were also generated using the eISC-toolbox. Blink synchronization was calculated for estimating the effect of scene cuts for blinking (see section "Scene cut influence analysis"), and it was estimated by calculating how many participants blinked within a given time window.

A single analysis time window such as 500 ms or 1,000 ms is not justifiable in uncontrolled movie perception because different eye movement parameters (e.g., pupillary responses or gaze position changes) have different intrinsic time scales and because the predictor variables also had different time scales (from a 40-ms resolution for physical and semantic features to a 4,000-ms resolution for the perceived social features; see "Perceptual models derived from movie stimuli"). Therefore, different analysis time scales may capture associations between the gaze metrics and the stimulus features. To estimate the effect of the chosen temporal scale, the eye-tracking parameters were sampled and analyzed across different temporal scales (200 ms, 500 ms, 1,000 ms, 2,000 ms, 4,000 ms). As the median fixation duration in our data was around 300 ms

(Figure 1) and the pupillary response lag was around 1,000 ms (Supplementary Figure SI-7), we decided to report the results in a 500-ms analysis scale in the main text.

Perceptual models derived from the movie stimuli

Stimulus features varying from low-level audiovisual properties (e.g., luminance and audio intensity) to semantic information (e.g., faces and objects) and to the perceived socioemotional information (e.g., (un)pleasant interaction and arousal) were estimated from the video stimuli to model the simultaneous effects of perceived features across distinct processing levels in human cognition. We extracted multiple low-level features and their time derivatives from the audiovisual domain (24 features), identified semantic features from the video frames using computer vision (7 features), and collected dynamic perceptual ratings of high-level socioemotional features from human annotators (8 features). The predictors were then sampled to different temporal scales (200 ms, 500 ms, 1,000 ms, 2,000 ms, 4,000 ms) matching with the eye-tracking data (see above).

Low-level audiovisual features

Low-level scene features such as luminance, contrast, and motion have been previously associated with pupil size and visual attention (Abrams & Christ, 2003; Itti & Koch, 2000; Kimura et al., 2014; Pratt, Radulescu, Guo, & Abrams, 2010; Tatler, Baddeley, & Gilchrist, 2005), but also auditory stimuli can have modulatory effects (Partala & Surakka, 2003). For a detailed description of the low-level information space, several low-level visual and auditory features were dynamically extracted from the movie stimuli. Visual features were extracted for each frame (static features) or consecutive frame pair (features that measure change) from the video stimulus. Visual features included luminance, visual entropy, optic flow, spatial energy with two distinct Fourier filters for edge detection, and differential energy for measuring total change between consecutive frames. Auditory features included audio intensity (RMS), properties of the frequency spectrum (geometric mean, standard deviation, entropy, and high-frequency energy), waveform sign change rate or "noisiness," and sensory dissonance or "roughness." The auditory features were extracted using MIRToolbox (Lartillot & Toiviainen, 2007). Since the video frame rate was 25/s, the auditory features were extracted in matching

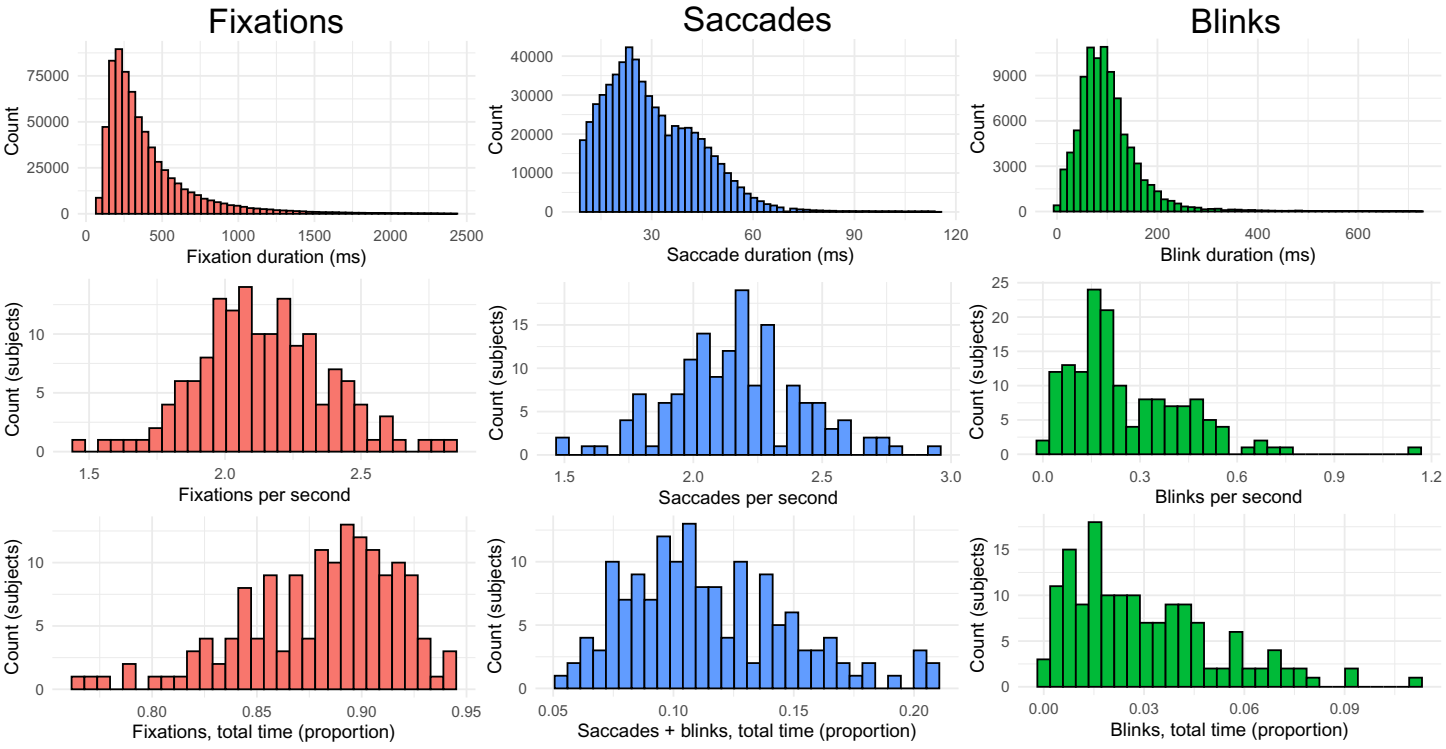

Figure 1. Distributions of fixations, saccades, and blinks during movie viewing. The top row shows distributions for fixation, saccade, and blink durations over all participants and datasets. The middle row shows the distributions of participant-wise mean fixation/saccade/blink rates, and the bottom row shows the distributions of participant-wise proportions of time allocated for fixations, saccades, or blinks.

40-ms time windows. For each audiovisual feature (except optical flow and differential energy, which are already measured between frames), the difference between consecutive frames was also calculated. Visual information was extracted locally within each frame and then averaged across pixels. Detailed properties of the feature extraction are provided in the Supplementary materials (see section "Low-level feature extraction").

Semantic features

Based on the known importance of semantic information, such as human faces and especially eyes, for visual attention (Birmingham et al., 2009; Morrisey et al., 2019; Varela, Towler, Kemp, & White, 2023; Võ, Smith, Mital, & Henderson, 2012), we segmented the stimulus films frame-by-frame to extract semantic information from the videos using pretrained open-source computer vision models. Each pixel was assigned to a class (eye area, mouth area, other face area, body parts, animals, objects, or background) or left unknown if the model's prediction confidence was under 50%. To first segment the whole image, we used a panoptic feature pyramid network (FPN) segmentation model (https://github.com/facebookresearch/detectron2/blob/main/configs/COCO-PanopticSegmentation/panoptic_fpn_R_101_3x.yaml) from the Deceptron2 Python library (Wu, Kirillov, Massa, Lo, & Girshick, 2019). The FPN included a Mask R-CNN architecture with an ImageNet-pretrained ResNet101 backbone, and it was trained on the COCO dataset to segment the images into 134 initial categories (Kirillov, Girshick, He, & Dollar, 2019; Lin et al., 2014). After segmentation, the initial categories were assigned to broad semantic classes (bodies, animals, objects, background, and unknown) for the analyses. Next, we used the RetinaFace face detection model (Deng, Guo, Ververas, Kotsia, & Zafeiriou, 2020) following the implementation of a previous eye-tracking study of autism to segment rectangular face, eye, and mouth areas from the videos (Keles et al., 2022). Eye and mouth areas were excluded from the whole-face segmentations to divide faces into three segments (eyes, mouth, and face excluding eyes and mouth).

The accuracy of the automatic segmentations was validated using human reference in the stimulus for Experiment 1. Judging the accuracy of all segmentations from each stimulus frame would have been overly laborious. Hence, the human annotator evaluated the accuracy of the segmentations in the most interesting areas (i.e., areas receiving most fixations from observers) of a subset of the video frames as follows. One frame per second was extracted from the stimulus, and the segmentations for this frame were overlaid onto the frame. A 1-second interval was considered sufficiently short for accurate estimations of the segmentation data quality, and the work was not overly laborious for manual checking at that interval. Next, the thresholded (95th percentile) population average gaze position heatmap was overlaid onto the frame to identify the areas that were the most interesting to the participants around the extracted frame (100-ms time window around the frame). A human annotator checked these images one by one and judged whether the segmented classes within the interest areas (under the gaze position heatmap) were correct or not and then marked the right class based on her opinion. The model's average positive predictive values were as follows: eye area 99%, mouth area 87%, other face area 99%, bodies 69%, animals 65%, objects 79%, and background 83%. The sensitivities were as follows: eye area 92%, mouth area 95%, other face area 77%, body parts 89%, animals 59%, objects 44%, and background 82%. The confusion matrix of the quality control results and the sample frames of the segmentations are shown in Figure 2. Since animals were infrequently present and the positive predictive value of this category was the lowest, we decided not to include this category in the further analyses.

Scene cuts

Unlike real-life situations, movies contain scene cuts that result in major semantic and perceptual changes. The cuts also impact eye movements considerably (Bruckert et al., 2023). We used a scene filter from ffmpeg to detect scene changes in the stimulus (https://ffmpeg.org/ffmpeg-filters.html#select_002c-aselect, filter command: "select='gt(scene,threshold)'"). The filter first assigns a score to each frame, where higher scores indicate a higher probability of the frame starting a new scene. Cuts are then identified using a detection threshold for the score. To optimize the detection threshold and validate the performance of the automatic scene cut detection method, three human annotators identified the scene cuts in the Experiment 1 stimulus. Using the manual annotations as the ground truth, the detection threshold of the automatic scene filter was optimized to balance the positive predictive value and sensitivity of the scene cut detection. A temporal distance of 200 ms between the algorithm and human annotators was chosen as the maximum distance for confirming a match. Optimized scene cut detection achieved 96% positive predictive value (PPV) and 95% sensitivity in the Experiment 1 stimulus. The optimized scene filter was used to detect scene cuts in all three experiment stimuli.

Perceived social features

We selected a set of perceived socioemotional features identified as important perceptual features in

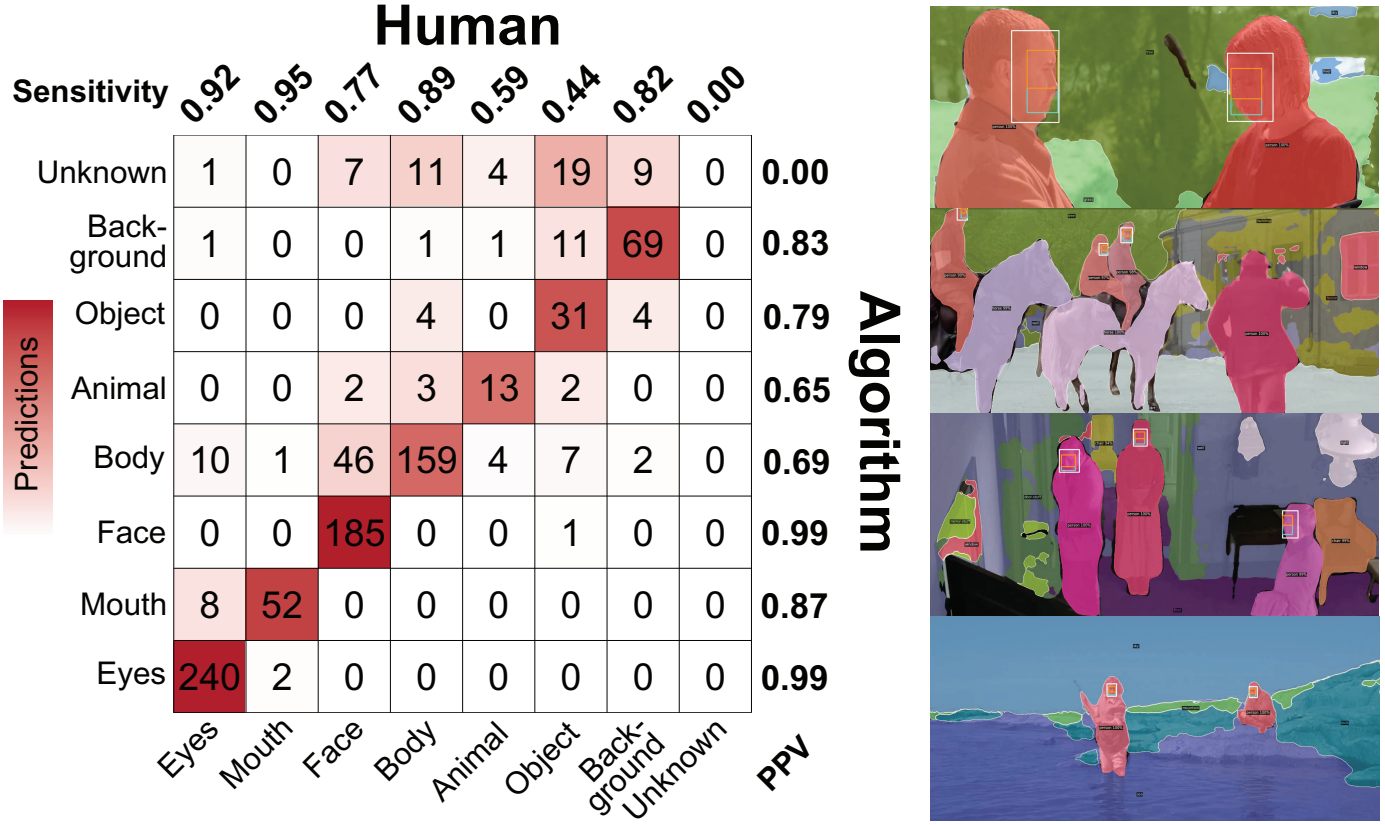

Figure 2. Semantic segmentation quality control. The confusion matrix shows the agreement of a human annotator and the segmentation algorithm in the Experiment 1 (movie clips) stimulus. Examples of the segmentation results are visualized alongside the confusion matrix. The sample images are from the Experiment 2 movie stimuli (Louhimies, 2008).

our previous studies of social perception (Santavirta et al., 2023; Santavirta et al., 2024). These features are also annotated with high between-participant consistency (Santavirta et al., 2024), which is crucial since different participants annotated the stimuli and participated in the eye-tracking experiments. These features were perceived pleasant feelings, unpleasant feelings, arousal, pain, talking, body movement, feeding, and playfulness. The ratings were collected in a 4-second temporal resolution with an abstract slider scale from "absent" to "very much." Five annotators were recruited to the laboratory for annotating these features for Experiments 1 and 2 using the Onni platform (https://onni.utu.fi/) (Heikkilä, Laine, Savela, & Nummenmaa, 2021). The annotations for Experiment 3 were collected online using the Gorilla platform (https://gorilla.sc/), and the participants were recruited through Prolific (https://www.prolific.com/). To decrease the workload of online participants, one participant annotated only a subset of the video clips. The online experiment began with a bot check (in addition to Prolific's internal bot checks for their participant pool) to increase the data quality. Quality control of the online participant data was conducted as described previously (Santavirta et al., 2024), and data were collected until five reliable annotations were retrieved for each clip and feature. Finally, we took the average over the individual annotators to obtain the high-level predictors for the eye-tracking analyses.

Final predictor set

As described in the above sections, we extracted a total of 39 perceptual features from the video stimulus to get a detailed description of the stimulus contents in multiple processing levels, from pure audiovisual processing to semantic categorization and perceived social information. A correlation analysis between the extracted features indicated that some of the features were strongly associated with each other. Reliable and interpretable mapping of the eye-tracking data with the available features requires that the predictors provide relatively independent information. We thus decreased the correlations between predictors by hierarchically clustering the correlation matrix of the extracted features using the default UPGMA clustering algorithm

(https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/hclust). The clustering identified 16 clusters (7 clusters for low-level features, 5 clusters for perceptual semantic categories, and 4 clusters for perceived social information) from the 39 extracted features, which were then used as the predictors in the analyses instead of the original features. The cluster predictors were formed by calculating the average over the standardized values of the features within each cluster. See section "Dimension reduction" and Supplementary Figure SI-5 in the supplementary materials for a detailed description of the dimension reduction.

Statistical analyses

The aim of this project was to establish how external low-, mid-, and high-level social perceptual features together guide the human vision. The following research questions were studied with unique analytical techniques developed for each of them:

1. Do people prioritize certain semantic categories (e.g., faces) over other semantic categories in social scenes?
2. How are the eye-tracking parameters dynamically modulated by low-level audiovisual information, semantic category information, and perceived social information in social scenes?
3. How accurately can the population-level gaze directions be dynamically predicted in social scenes with an interpretable model that incorporates low-level information, semantic category information, and perceived social information?
4. How do scene cuts influence the eye-tracking parameter dynamics immediately after a scene cut?

To address each of these questions, we developed a set of analytical methods and tested whether the results generalize across the three independent datasets. First, we conducted a gaze time analysis to assess if people have attentional priorities for some semantic categories (Question 1). Second, we used a cross-validated multiple regression analysis to assess which perceptual predictors have an independent association with pupil size, ISC, fixation rates, and blink rates (Question 2). Third, we built cross-validated random forest models for moment-to-moment pixelwise predictions of the population-level gaze heatmaps to understand how well the actual gaze positions can be modeled with the available predictors and to deepen the understanding of the attentional priorities when observing dynamic social scenes (Question 3). Finally, we investigated the average dynamics of the pupil size, ISC, and blink rates after each scene cut (Question 4).

Gaze time analysis

To investigate whether people have an attentional preference toward certain semantic classes, we conducted a gaze time analysis. The pixel-to-pixel semantic class information from the computer vision was used to identify which class (eyes, mouth, face excluding eyes and mouth, body parts, object, or background) a participant was watching at each time point. From these data, we calculated the total gaze time for each class and averaged the gaze times over the participants within each experiment to get population-level gaze times for each class. The analysis was conducted separately for each experiment.

High gaze time for a class does not itself reveal an attentional preference for the class but could merely indicate that the class is often present occupying large parts of the screen. Additionally, centrally presented contents are likely to receive more fixations as observers prefer looking at the center of the stimulus (Dorr et al., 2010; Tatler & Vincent, 2009). To control for the location, frequency, and total presented area for the semantic classes in our stimulus, we used permutation testing to differentiate whether the true gaze times are significantly different from those expected by chance. The permutation testing was conducted by bootstrapping the true fixation gaze data circularly to break the temporal synchrony between the stimulus and the gaze coordinates. Total gaze times for semantic classes were then calculated for the bootstrapped random data. The bootstrapping was repeated 500 times to get the estimation of the gaze time null distributions. Finally, a p value for the null hypothesis that the true gaze time does not differ from chance was calculated by ranking the real gaze times to the corresponding null distributions. Since true fixation coordinates were used in the permutation testing instead of simulated random coordinates, the only difference between random data and true data is the temporal asynchrony between the gaze coordinates and the stimuli (distributions of gaze coordinates and fixation durations do not differ).

Regression analysis

To investigate the associations between the extracted predictors and eye-tracking parameters while controlling for all other predictors, we conducted a multistep regression analysis. Even after dimension reduction, the 16 features in the predictor space were not fully independent. Repeated simple regressions or one multiple regression with all the predictors in the same model could result in mixed findings due to possible collinearity between predictors. Instead, we chose a multistep regression approach to achieve more support that the identified associations are truly

independent of the effects of other included predictors. For pupil size modeling, predictors were shifted 1,000 ms forward to correct for the lag in pupillary response (Supplementary Figure SI-7) while ISC, fixation rate, and blink rate were modeled without temporal shift.

First, we ran simple regression for all predictors in a leave-one-experiment-out cross-validation process where models were fit to the data of two experiments and one experiment's data were left as the testing set. The predictors where the sign of the association (regression coefficient) was consistent across all cross-validations indicating a replicable sign of association between experiments were then selected for the next analysis step. If the simple regression associations were inconsistent between the cross-validation rounds, we concluded that the results do not support an association between the eye-tracking data and the feature. Second, we ranked the consistent predictors by their out-of-sample prediction accuracy (correlation of the model's predictions with the leave-out experiment data) to prioritize the features with higher predictive power in the simple regression setting over the features with lower predictive power. Finally, we built an additive regression model where we added candidate predictors one by one to a multiple regression design, starting with the features whose predictive power ranked the highest in the previous regression model. After every addition of a new predictor to the model, we tested whether the out-of-sample prediction accuracy (correlation with the leave-out experiment data) improved more than would be expected by chance. If the prediction accuracy after adding a new predictor was lower than expected by chance, we concluded that the predictor does not have an independent association, and the predictor was dropped from the design. Hence, the final model tests all consistent predictors and returns the final model with only predictors that improve the prediction accuracy significantly.

The order in which the predictors are added to the multiple regression influences the results when predictors are not fully independent, but testing all possible combinations and orders is computationally prohibitive with up to 16 predictors. Hence, the initial prediction accuracy in the simple regression was chosen to determine the order of addition. This approach penalizes the predictors that initially showed a low prediction accuracy but does not rule out the possibility of an independent effect if the predictor's effect is not better modeled by the previously added predictors. This approach assumes that if multiple correlated predictors had an association with the variable of interest in the simple regression setup, the one with the highest prediction accuracy would be the one most likely associated with the variable of interest. To ensure that the simple regression and the stepwise multiple regression results did not yield mixed findings, we confirmed that the sign of association of the final set of significant predictors stayed the same in the initial simple regression and in the multiple regressions.

To test whether adding a predictor to the multiple regression design yields better predictions than expected by chance in the left-out-experiment data, we conducted a permutation test after each predictor addition. Only the newly added column of the design matrix was bootstrapped circularly to preserve the covariance structure of the previously validated model. Bootstrapping was repeated 500 times to generate the null models. The null distribution of the out-of-sample prediction accuracy was generated from the prediction accuracies of these null models. The p value for the hypothesis that the newly added feature did not improve the model's prediction accuracy more than would be expected by chance was then calculated by ranking the true prediction accuracy against the null distribution. The predictor was considered to improve the prediction accuracy significantly when $p < 0.05$.

Gaze prediction analysis

We built a data-driven model to predict the population average gaze position heatmaps to assess how well the current predictors can explain the momentary gaze directions of the subjects. Pixelwise predictor values were used to predict the gaze probability of each pixel. To lower computational demands, the gaze heatmaps and predictors were downsampled to 64-pixel × 64-pixel resolution prior to training the gaze prediction models. We did not model the gaze heatmaps for each frame since a 40-ms time interval is unnecessarily short. Instead, we calculated the gaze heatmaps in 200-ms time windows for this analysis to preserve high temporal resolution while allowing the temporal accumulation of information between a few adjacent frames. This temporal resolution was close to the rate of gaze change as the typical fixation duration was approximately 300 ms in our data (Figure 1).

We used a random forest regression model that preserves some interpretability, allows for a more complex association to be considered than ordinary linear regression, and is computationally efficient, unlike the deep convolutional neural network, which would be computationally prohibitive and difficult to interpret in this context. The random forest models were trained using the fitrensemble function in MATLAB (https://se.mathworks.com/help/stats/fitrensemble.html). The fitrensemble function randomly selects N out of N observations of data with replacement and one third of the predictors for training each individual decision tree. The number of individual decision trees and the number of branches in each tree were optimized (see Supplementary materials, section "Random forest regression optimization"). Based on the optimization results, 50 decision trees and 63 branches (6 branches

from the tree trunk to the leaf) were selected. Separate models were trained for each of the three datasets to allow out-of-sample performance testing and comparison of the trained models between datasets.

The performance of the gaze prediction models was evaluated by testing the trained models on the two datasets that were not included in the model training. Two performance metrics were calculated. First, we calculated the correlation between the true heatmap values and the predicted heatmap values over all time windows in the tested dataset. The correlation measures the overall similarity of the heatmap distributions but is not sensitive to measuring how well the model predicts the most salient area (the area with the peak heatmap value) on the screen. To assess how well the model can predict the most salient areas, we also calculated the distance between the true peak value and the predicted peak value for each 200-ms time window. This peak prediction distance was then averaged over all time windows and is reported as percentage of the image width. If the peak prediction distance is low, then the model is able to predict well the most salient part of the screen, regardless of the overall similarity of the true and predicted heatmap distributions.

To interpret the trained random forest models, we extracted the relative feature importance (https://se.mathworks.com/help/stats/classreg.learning.classif.compactclassificationensemble.predictorimportance.html) for each predictor, which states how influential the predictor was for the model's prediction with a range between 0 and 1. The relative importance does not reveal whether an increase in the predictor value would increase the predicted gaze probability or vice versa. Hence, we simulated how changing the value of one predictor influences the model's predictions when all other predictor values are held constant. First, we randomly selected a pixel from the training dataset and extracted the real predictor values for that pixel. For categorical predictors, we simulated how changing the value from 0 to 1 would influence the gaze probability prediction in that pixel. For standardized continuous predictors, we simulated new predictions with 20 uniformly distributed random predictor values (z scores) between –3 and 3. This simulation was repeated for 200,000 randomly selected pixels, and the simulation was done separately for each trained model (one for each dataset) before pooling the simulation results together.

Scene cut influence analysis

We extracted the eye-tracking data for the 3,600-ms time window around each identified scene cut (from cut –600 ms to cut +3,000 ms) for each participant. Pupil size, ISC, and blink rate were then extracted for these time windows. Pupil size was extracted continuously (1 kHz), and ISC and blink rates were calculated in 200-ms time windows (5 Hz). Since pupil size is an arbitrary participant-specific measure, pupil size time series were normalized with the mean pupil size in a 600-ms interval before scene cut. To estimate the replicability of the scene cut effect, the data were averaged over the participants separately for each experiment.

To estimate whether the population-level change in eye-tracking measures after the scene cut was significantly higher than would be expected by chance, a permutation test was applied. We sampled 100 time series (3,600 ms) at random time points from the eye-tracking data for each participant and averaged these over all participants to get a random sample of the population-level average time series for the 3,600-ms time period. This sampling was repeated 500 times to get a null distribution of the population-level random time series, and thus the null distribution was generated from 50,000 random samples (100 × 500). The true population-level time series calculated from participant-level time series around scene cuts were ranked within this null distribution to get the exact p value for each time point around the scene cut. A significant cut effect is reported for time points with $p < 0.05$ based on this permutation test.

Results

Eye movements during movie viewing

Figure 1 shows the histograms for fixations, saccades, and blinks during movie viewing. The data are combined over all three experiments. The fixation, saccade, and blink durations were similar across the experiments (Supplementary Fig. SI-1). The median fixation duration was 309 ms (q5%–q95%: 135–1,047 ms), the median saccade duration was 27 ms (q5%–q95%: 11–55 ms), and the median blink duration was 91 ms (q5%–q95%: 23–217 ms). The fixation, saccade, and blink rates varied between participants. The median fixation rate was 2.12/s (q5%–q95%: 1.77–2.52/s), and the median blink rate was 0.20/s (q5%–q95%: 0.05–0.55/s). The median total fixation time (total duration of fixations/total stimulus duration) was 0.89 (q5%–q95%: 0.82–0.93), and the median total blink time (total duration of blinks/total stimulus duration) was 0.02 (q5%–q95%: 0.00–0.07). The gaze patterns were moderately consistent across subjects (mean ISC: 0.37, SD: 0.20). The eye-tracking measures (pupil size, ISC, fixation rate, and blink synchronization) were mostly uncorrelated with each other (Figure 3 & Supplementary Figure SI-2). In the 200 and 500 ms temporal scale, a negative correlation was observed between fixation rate and ISC ($r = –0.14$) and between blinking and ISC ($r = –0.09$) consistently

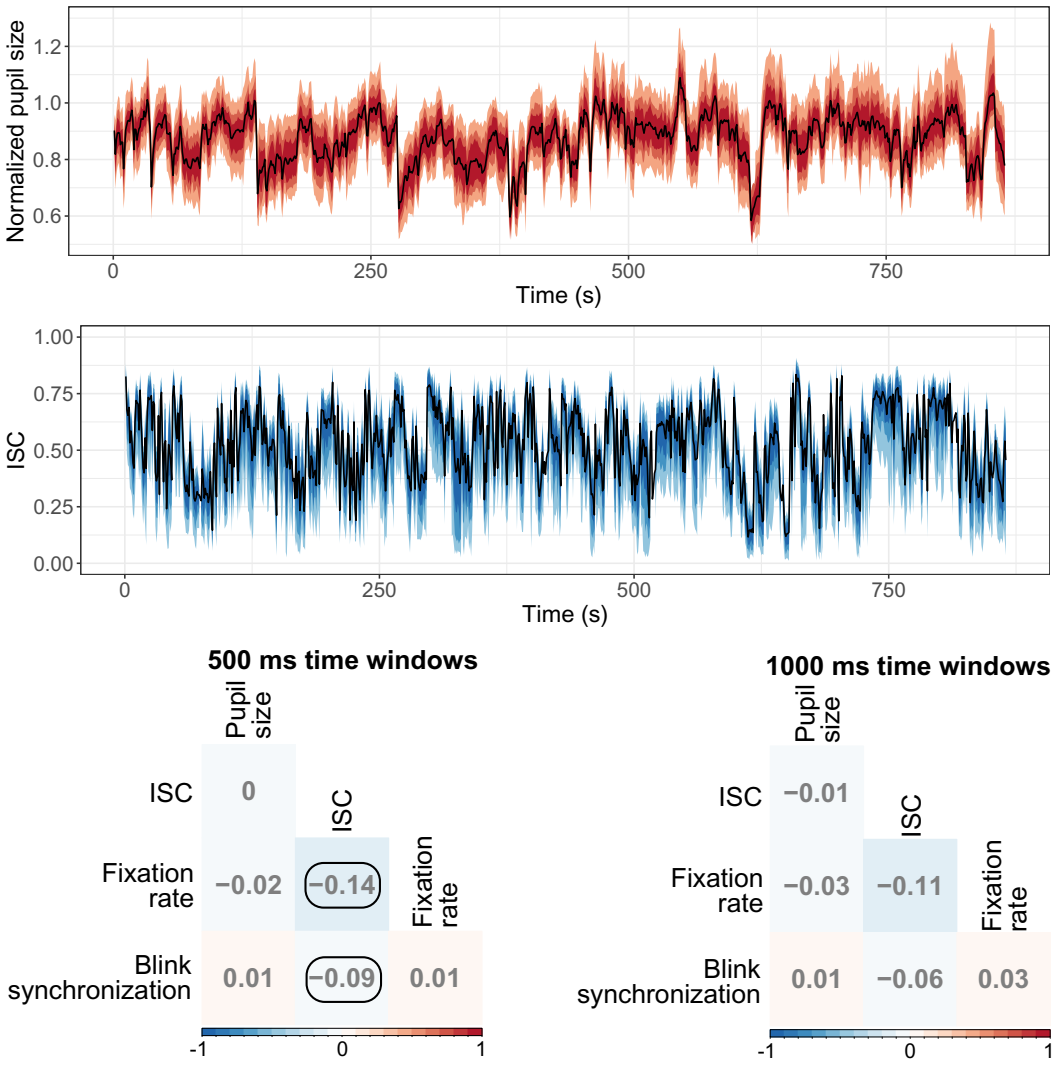

Figure 3. Top: Time series of median pupil size and ISC across subjects in the Experiment 1 data (colored areas mark the 40%, 60%, and 80% quantile intervals; see Supplementary Figure SI-3 for time series of Experiment 2 & 3 data). Bottom: Correlations between the eye-tracking parameters. The correlations were calculated for each experiment separately and then averaged over datasets. Correlations that had consistent signs over all experiments are circled. The correlations were calculated in multiple different time scales (200–4,000 ms; see Supplementary Figure SI-2). Blink synchronization indicates how many participants blinked in a given time window.

across the three experiments, but this weak correlation was not consistent across datasets when analyzed in longer temporal scales (1,000–4,000 ms).

Gaze time analysis

The participants showed an attentional preference in the dynamic stimuli to the eye and mouth area in all three datasets (Figure 4). The eyes were viewed for 21% to 33% of the total stimulus time, which was 16 to 24 percentage points more than expected by chance ($p < 0.005$). The mouth area was viewed 10% to 11% of the time, which was 7 to 8 percentage points more than expected by chance ($p < 0.005$). The bodies were looked at for 18% to 28% of the total time, which was 4 to 7 percentage points less than expected by chance ($p < 0.005$). Similarly, background was viewed for 15% to 23% of the total time, which was 17 to 23 percentage points less than expected by chance ($p < 0.005$). The difference in gaze time for objects and faces excluding the eye and mouth area was small, and the direction of the difference was inconsistent between the datasets. People also gazed at unclassified areas, for which the computer vision algorithm failed to assign a reliable category for 10% to 14% of the time, which was 1% to 2% less than expected by chance ($p < 0.005$). This indicates that the computer vision algorithm did not

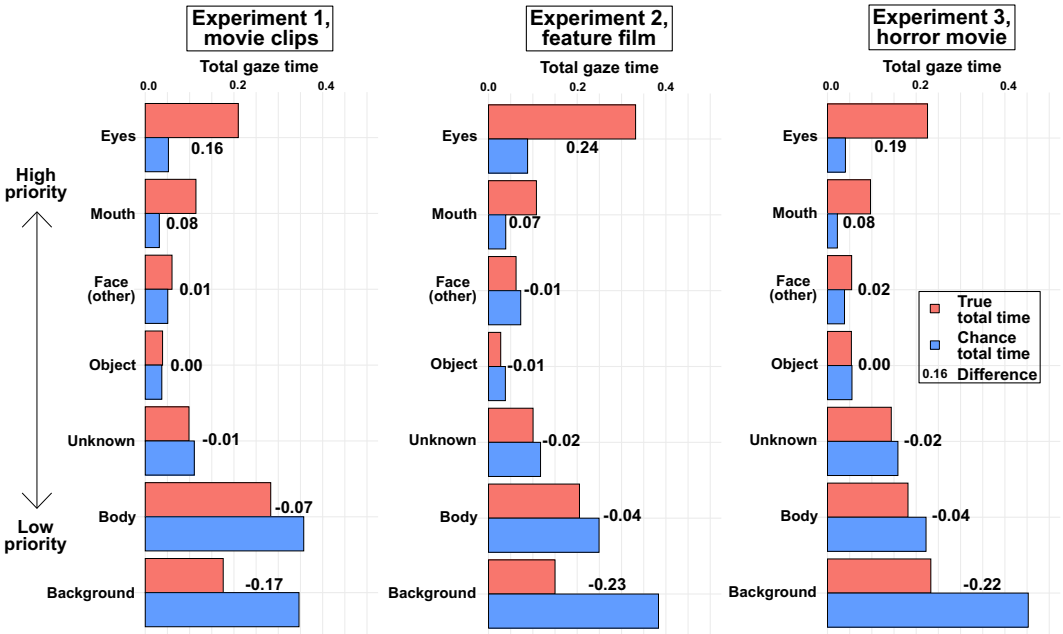

Figure 4. Results of the gaze time analysis separately for each experiment. The red bars show the true total gaze times (proportional to the experiment duration), while the blue bars show the total times that would be expected by chance for different semantic classes. The numbers show the difference between true gaze time and estimated chance gaze time. All differences between true and chance gaze times were statistically significant based on a permutation test ($p < 0.005$).

miss highly prioritized information from the stimuli. People gazed at areas outside the video less than 1% of the time. Since the real gaze coordinates were randomly sampled for the permuted chance time estimation, the gaze time outside the video area cannot be differentiated from chance with the implemented test.

Multistep regression analysis

We next established links between spatiotemporal gaze parameters and low-, mid-, and high-level stimulus features. To that end, pupil size, ISC, fixation rate, and blink rate were modeled with a stimulus model of 16 low-level, semantic, and social perceptual features using a multistep regression approach. The stimulus model was generated from the extracted 39 features based on the covariance structure between the original features. Figure 5 summarizes the regression results for the representative 500-ms temporal analysis scale. The results were consistent when the analyses were repeated across different time scales (Supplementary Figure SI-4). While most of the predictors had a consistent sign of association between the eye-tracking variables in simple regression over the cross-validation rounds (leave-one-experiment-out cross-validation, not tested for statistical significance), only a limited number of predictors increased the subsequent multiple regression models' out-of-sample prediction accuracies significantly.

Pupil size was most consistently associated with low-level features and perceived unpleasantness (Figure 5 & Supplementary Figure SI-4). In the 500-ms time scale, adding consistent predictors one by one into a multiple regression model starting from the predictor with the best out-of-sample prediction accuracy in simple regression, the predictors "Luminance/entropy" (negative association), "Luminance/entropy, time derivative" (negative association), "Audio intensity/roughness" (positive association), "Scene cut" (negative association), and "Unpleasant situation" (positive association) significantly increased the model's out-of-sample prediction accuracy compared to the permuted chance level ($p < 0.05$). The final models including these five features produced predictions in the test set (leave-one-experiment-out) that correlated highly with the true pupil size (Experiment 1 as the test set: $r = 0.36$, Experiment 2 as the test set: $r = 0.49$, Experiment 3 as the test set: $r = 0.50$).

ISC of gaze (i.e., synchononization of gaze with other subjects) was most consistently associated with mid-level features and motion (Figure 5 & Supplementary Figure SI-4). In the 500-ms time scale, the predictors "Gaze on face" (positive association), "Gaze on body" (negative association), "Gaze on background" (negative association), and "Visual movement" (negative association) significantly increased the multiple regression model's out-of-sample prediction accuracy ($p < 0.05$). The final models including these four features produced predictions in

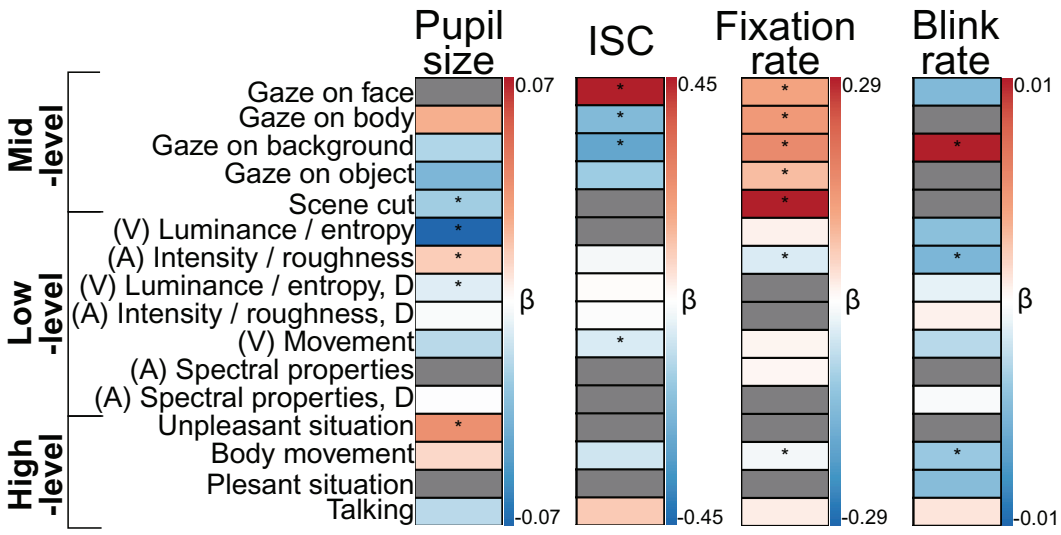

Figure 5. The results of multistep regression analyses for pupil size, ISC, fixation rate, and blink rate in the 500-ms time scale. Coefficients from the simple regressions are visualized with a blue (negative)–red (positive) color gradient. Gray color indicates that the direction of the association was inconsistent between the simple regression cross-validation rounds (leave-one-experiment-out cross-validation). The final multiple regression model included the features marked with asterisks. These features significantly increased the model's out-of-sample prediction accuracy compared to chance ($p < 0.05$) when added to a multiple regression model. (V) denotes low-level visual predictors, (A) denotes auditory low-level predictors, and D denotes the time derivative of the feature.

the test set (leave-one-experiment-out) that correlated highly with the true ISC (Experiment 1 as the test set: $r = 0.43$, Experiment 2 as the test set: $r = 0.33$, Experiment 3 as the test set: $r = 0.40$).

Fixation rate was most consistently associated with mid-level features and audio intensity/roughness (Figure 5 & Supplementary Figure SI-4). In the 500-ms time scale, the predictors "Scene cut" (positive association), "Gaze on face" (positive association), "Gaze on body" (positive association), "Gaze on background" (positive association), "Gaze on object" (positive association), "Audio intensity/roughness" (negative association), and "Body movement" (negative association) significantly increased the multiple regression model's out-of-sample prediction accuracy ($p < 0.05$). The final models including these seven features produced predictions in the test set (leave-one-experiment-out) that correlated moderately with the true fixation rate (Experiment 1 as the test set: $r = 0.26$, Experiment 2 as the test set: $r = 0.21$, Experiment 3 as the test set: $r = 0.24$).

Blink rate was mainly associated with "Gaze on background" (positive association), "Audio intensity/roughness" (negative association), and "Body movement" (negative association). These variables significantly increased the model's out-of-sample prediction accuracy ($p < 0.05$) in the 500-ms time scale. The final models including these three features were not able to predict much variation in the leave-one-experiment-out test set (Experiment 1 as the test set: $r = 0.03$, Experiment 2 as the test set: $r = 0.02$, Experiment 3 as the test set: $r = 0.03$).

All the above-reported associations between perceptual features and eye-tracking measures were consistent between the initial simple regressions and the multistep regression rounds, indicating that the addition of other predictors in the model did not yield interpretational difficulties.

Performance of the gaze prediction model

Random forest models using all available low-level, semantic, and social perceptual information were trained separately for each of the three datasets to predict the moment-to-moment population-level gaze heatmaps. The performance of the trained models was tested with the two datasets that were left out of the model training. The performance of the models was evaluated by calculating the correlation between the true and predicted gaze heatmaps and by measuring how far, on average, the predicted peak value was from the true heatmap peak value (Figure 6). The evaluation metrics were consistent across the models trained with different datasets. Correlations between predicted heatmaps and true heatmaps in the testing dataset ranged between 0.41 and 0.47, while the peak value distances ranged between 10% and 16% of the image width, indicating robust out-of-sample prediction performance.

Relative importance measures how influential (between 0 and 1) a given predictor is for the model's predictions. The relative importances were consistent across the models trained on different datasets (Figure 7, bar plot). The presence of eyes in a given voxel showed by far the highest importance for predictions (Experiment 1: 0.48, Experiment 2: 0.67, Experiment 3: 0.46). After eyes, the presence of mouth (Experiment 1: 0.20, Experiment 2: 0.08, Experiment 3: 0.18), visual movement (Experiment 1: 0.09, Experiment 2: 0.10, Experiment 3: 0.06), and luminance/entropy (Experiment 1: 0.03, Experiment 2: 0.06, Experiment 3: 0.09) showed higher relative importances than the rest of the predictors.

To establish whether the important predictors have a positive or negative effect on the models' predictions, we simulated how the models' predictions would change when all other predictors are held constant, but the values of only one predictor are changed (Figure 7, violin and density plots). The random forest models predicted higher gaze probability when eyes or mouth were present than when they were absent. Additionally, higher-than-average local movement and visual luminance/entropy resulted in increased gaze probability prediction compared to lower-than-average values. The relative importance and change in simulated predictions with different predictor values were negligible for the rest of the predictors. The high-level social features are constant within each time window and thus cannot have an independent main effect on the gaze location. However, high-level social information can influence how the model predicts gaze distributions

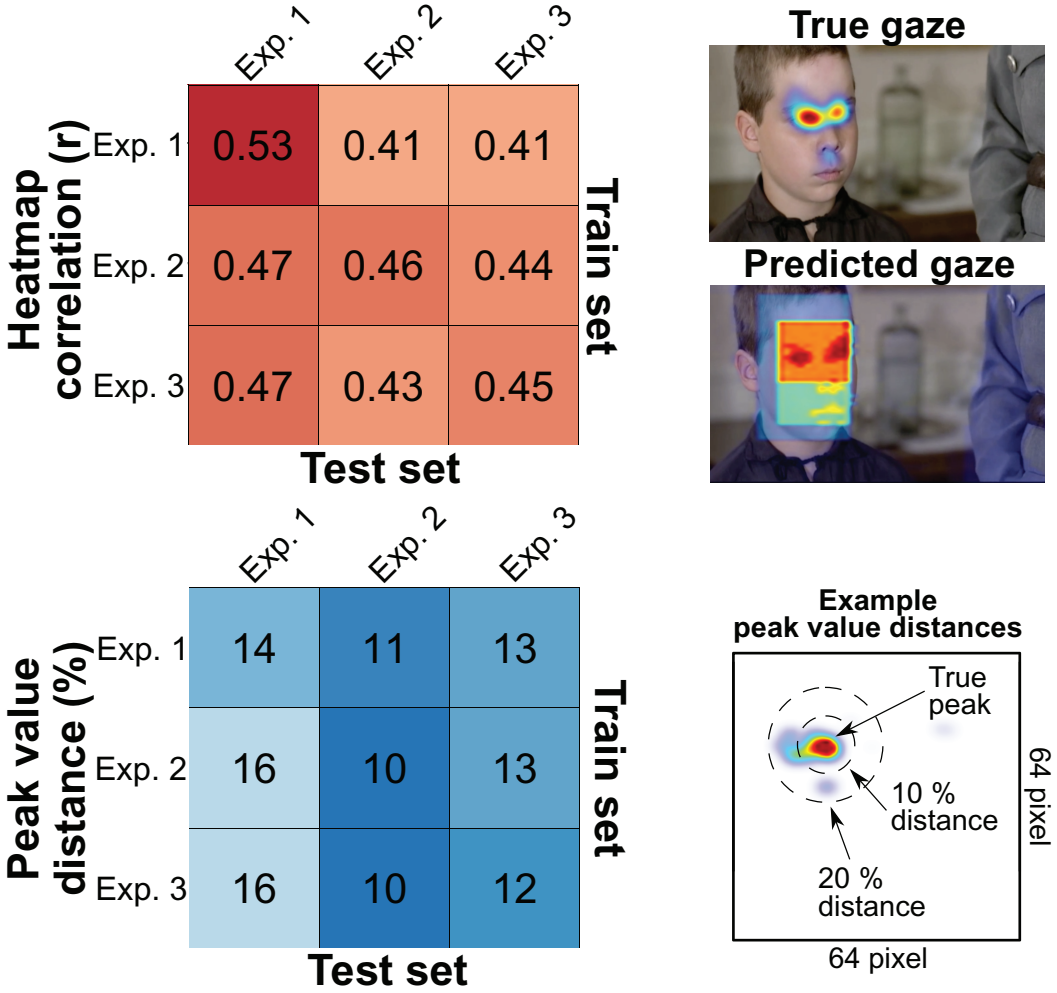

Figure 6. Performance of the gaze prediction model. The random forest model was trained with each dataset separately, and the model's performance was tested with the two other datasets. Top left: A confusion matrix showing the correlation of the true versus predicted gaze heatmaps for the trained models. Top right: True and predicted gaze heatmaps for one representative 200-ms time window. Bottom left: A confusion matrix showing how far, on average, the predicted peak heatmap value was from the true heatmap peak. Bottom right: Example peak value distances for 10% and 20% levels are visually provided for guiding the evaluation of the reported peak value distances. The sample images are from the Experiment 2 movie stimuli (Louhimies, 2008).

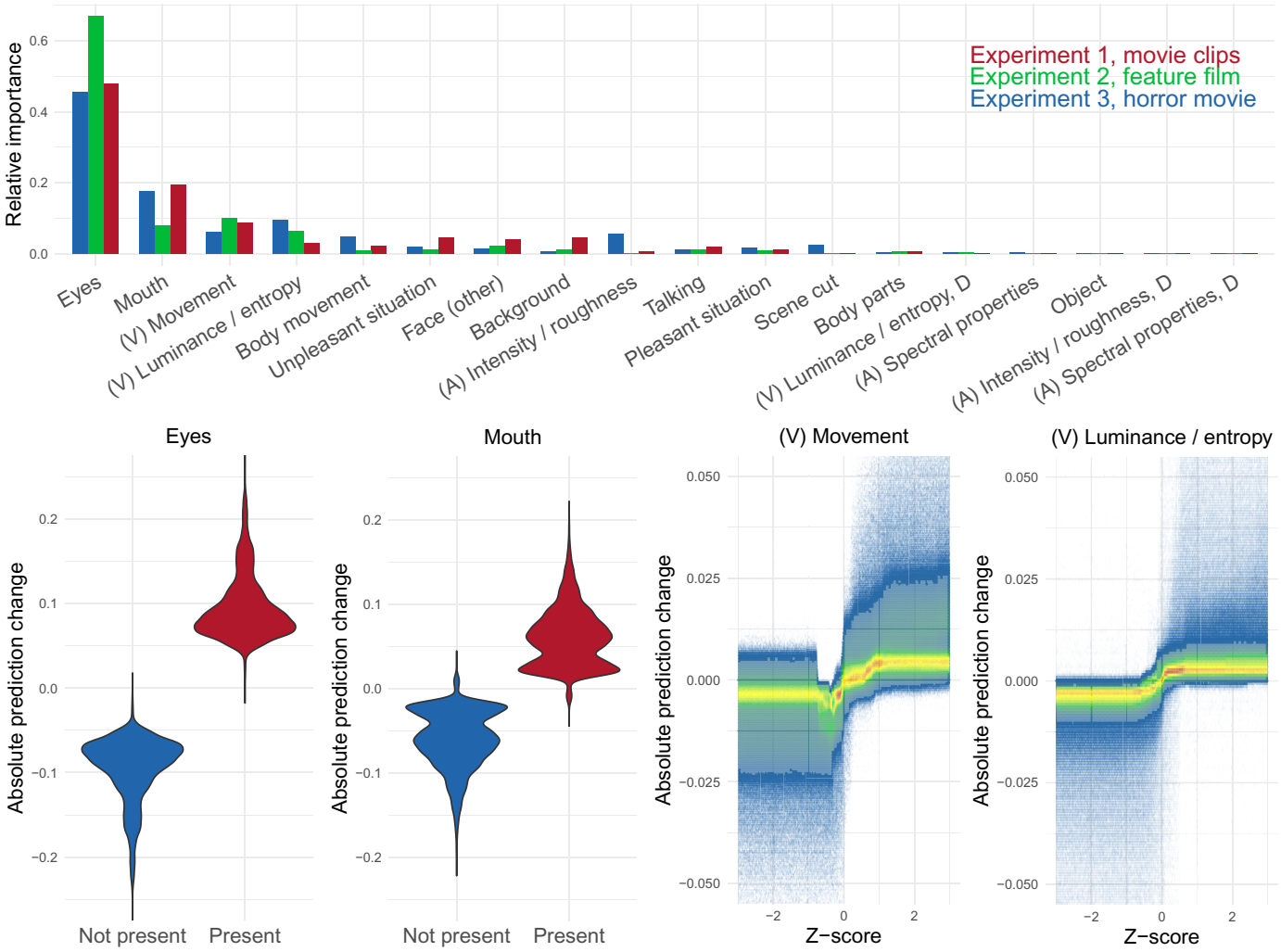

Figure 7. Interpretation of the gaze prediction models. Bar plots visualize the relative importance of each predictor in the trained models. Violin plots for categorical and density plots for continuous predictors visualize the simulated influence of each predictor on the gaze probability prediction when other predictors' values are held constant. (V) denotes low-level visual predictors, (A) denotes auditory low-level predictors, and D denotes the time derivative of the feature.

across time windows, and high-level information can also have an interaction with some pixelwise predictor (e.g., eyes are watched more closely during unpleasant situations). However, exploratory simulations of interactions between the perceived social predictors and the four most important predictors did not indicate any clear interaction effects.

Scene change effect

Figure 8 shows the scene change effect on pupil size, ISC, and blink synchronization. The findings were consistent across the datasets. Pupil size started to decrease rapidly after the scene changed, and the pupil size was significantly decreased compared to the permuted baseline between 350 and 1,150 ms after the scene cut in all datasets ($p < 0.05$, based on a permutation test). The minimum pupil size (97–98% of the baseline on average) was reached between 500 and 800 ms after the scene cut. ISC increased briefly after the scene changed. For full movies (Experiments 2 and 3), ISC increased significantly up to 800 ms after the scene transition (Experiment 2: between 400 and 800 ms after cut, Experiment 3: between 200 and 800 ms after cut, $p < 0.05$), while for short movie clips (Experiment 1), ISC remained elevated for a longer period of time (between 400 and 1,400 ms after the scene cut, $p < 0.05$).

The percentage of participants who blinked in the 200-ms time windows decreased shortly following the scene change in all datasets. For full movies, the percentage of participants blinking was briefly decreased at single time windows (Experiment 2:

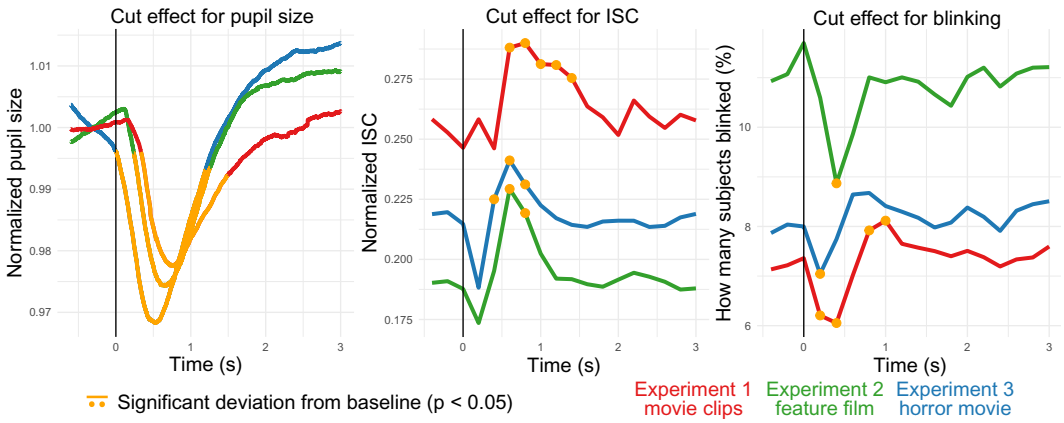

Figure 8. Temporal dynamics of pupil size, ISC, and blinking after a scene cut separately for each experiment. The scene cut time is marked as a vertical line, and the permuted statistically significant deviation period (pupil) or time points (ISC & blinking) from baseline are marked in orange ($p < 0.05$). Normalized ISC (ISC * stimulus area/total monitor area) is plotted since it accounts for the stimulus size differences between the experiments (stimulus area/monitor area, Experiment 1: 0.53, Experiment 2: 0.94, Experiment 3: 0.95).

200–400 ms after the scene cut, Experiment 3: 0–200 ms after the scene cut, $p < 0.05$). For the movie clip dataset (Experiment 1), blinking first decreased between 0 and 400 ms after the scene cut and then increased between 600 and 1,000 ms ($p < 0.05$). The decrease, followed by an increase in blinking, was not reliably observed in the movie datasets, although a similar increasing pattern after an initial decrease in blinking was also suggested by the Experiment 3 data (Figure 8).

Discussion

Our results established consistent relationships between low-level and semantic scene features with gaze, pupil diameter, and blinking. During natural vision, the gaze positions were relatively consistent across observers (mean ISC = 0.37) and most consistently predicted by mid-level semantic social information and low-level audiovisual features, whereas high-level social context contributed only minimally. In turn, pupil size was modulated by low-level sensory features as well as emotional content and scene transitions, while blinking rate was indicative of attentional engagement on the scenes. These results generalized across three independent datasets with different dynamic stimuli and participants. Our results thus suggest that gaze position, pupillary response, and blinking are at least partly determined by independent, yet partially overlapping sets of external features, and that human social vision is primarily guided by low-level physical features and mid-level socially relevant semantic categories (most notably faces), while high-level socioemotional features, such as emotional arousal, primarily modulate pupillary responses (Figure 9).

Gaze direction depends on the presence of faces, motion, and visual information

The regression analysis for ISC and gaze prediction analysis both achieved moderate to high out-of-sample prediction accuracies (up to $r = 0.43$ for ISC and up to $r = 0.47$ for gaze heatmaps), indicating that external features consistently predict what we attend to in dynamic social scenes. Across the analyses, we found that the eyes and mouths of the characters were the best predictors of dynamic gaze positions and the synchronization of gaze between participants. Across all datasets, people fixated on human eyes and mouth areas much more often than would be expected by chance (Figure 4). The developed permutation analysis ensured that the attentional priority for faces was not due to a centrality bias (Dorr et al., 2010; Tatler & Vincent, 2009) or just a result of movies presenting faces disproportionately more often than other contents. High priority for faces resulted in fewer fixations on the bodies of the characters and on the background/surroundings, which were gazed at much less than expected by chance. These results were confirmed in regression analyses that also accounted for low-level audiovisual features and perceived social information of the stimuli.

The subject's gaze synchronization with others (ISC) increased on average when the subject watched faces and decreased when the subject gazed at bodies or background (Figure 5). This increased synchrony indicates that the faces, as a biologically salient signal, could serve as a cue that captures viewers' attention (Morrisey et al., 2019), thus resulting in increased gaze synchronization. While the ISC and gaze targets were both calculated from the gaze data, they were not necessarily dependent on each other since any

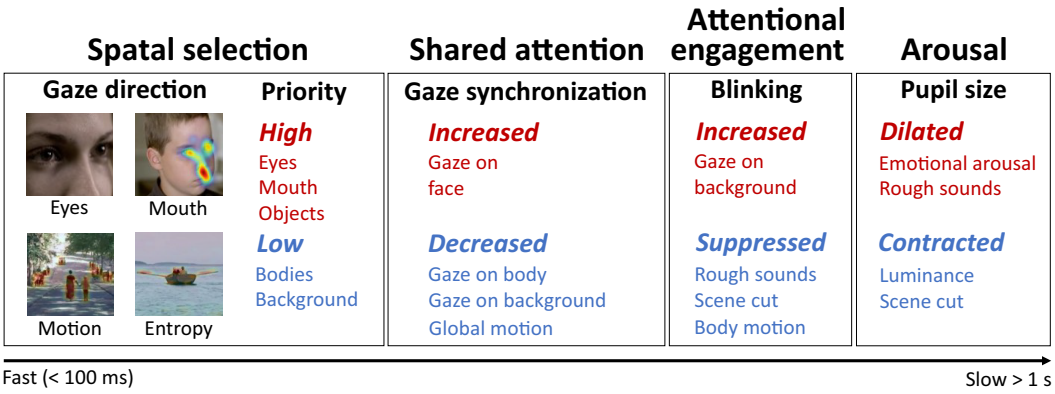

Figure 9. Summary of the findings. Investigation of the gaze direction, semantic category priority, and gaze synchronization reveals that the presence of human faces, especially eyes, drives the fast visual sampling simultaneously with low-level visual features. Blinking is suppressed in intense scenes and during scene transitions, indicating attentional engagement to the scenes, while pupil size is modulated by emotional arousal, scene transitions, and luminance. The findings are based on analyses with 16 features derived from the movie stimuli (16 feature clusters were identified from the extracted 39 features). The sample images are from the Experiment 2 movie stimuli (Louhimies, 2008).

subject's gaze target is not dependent on what the others are watching. As additional evidence, the population gaze prediction analysis showed that eyes and mouths were the most important predictors of population-level gaze positions in cinematic social scenes, surpassing low-level features (Figure 7). This is in line with previous findings suggesting that social information is prioritized over low-level features in social scenes (End & Gamer, 2017). Selective attention to human faces has been well established in controlled study designs (Bindemann, Burton, Hooge, Jenkins, & de Haan, 2005; Morrisey et al., 2019; Theeuwes & Van der Stigchel, 2006). Other people are noticed quickly and independently of the task, indicating that the identification of our conspecifics happens automatically (Rösler, End, & Gamer, 2017; Smith & Mital, 2013). Talking faces are looked at more often than listeners during conversation, which indicates that social information conveyed by faces is prioritized when multiple faces are present (Coutrot & Guyader, 2014). Our results provide novel evidence on how faces are prioritized over other available information in dynamic and uncontrolled movie watching, and the current results provide an estimated "effect size" of the face priority when watching dynamic social scenes. The participants watched the faces 38% (Experiment 1), 50% (Experiment 2), and 38% (Experiment 3) of the total stimulus presentation time when only 14%, 20%, and 10% gaze times would have been expected if participants randomly inspected the scenes.

In addition to social semantic features, local visual luminance/entropy predicted the moment-to-moment gaze positions (Figure 7). This visual feature cluster combined the information of "luminance" (a measure of local lighting), "visual entropy" (a measure of local randomness of pixel intensities), and "spatial energy" (a measure that detects local edges). Therefore, it indexes the amount of local visual information that the area conveys, as visual information requires light (luminance), local variation (entropy), and detectable forms (visual energy) to stand out from the background. High local contrast and low correlation between local pixel intensities (as measures of local information) have been previously shown to draw attention in natural still images (Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; Reinagel & Zador, 1999), supporting the interpretation that visually information-rich areas stand out from the rest and capture attention effectively.

While human faces were the strongest predictors of ISC and moment-to-moment gaze positions, *global* scene motion (Figure 5 & Supplementary Figure SI-4) was associated with lower ISC, indicating that the overall motion desynchronized the gaze positions between participants. The decrease in synchronization was most likely due to the need for increased visual sampling, which resulted in diverging temporal gaze patterns between participants in these highly dynamic scenes. Nevertheless, the gaze prediction analysis indicated that areas with *local* motion draw attention (Figure 7). Increased attention to motion has been previously reported in humans (Abrams & Christ, 2003; Bruckert, Christie, & Le Meur, 2023) and also in macaques (Mahapatra, Winkler, & Yen, 2008), and motion has been shown to capture attention task-independently, suggesting automated attention toward motion (Smith & Mital, 2013). In sum, scenes with plenty of movement (real movement or camera movement) result in generally more scattered gaze positions between participants, but local areas with

high movement draw attention irrespective of the global amount of movement.

The analyses related to scene cuts indicated that ISC increased briefly, starting 200 ms after the scene cut, and, depending on the stimulus, it returned to baseline between 800 ms (full movie stimulus) and 1,400 ms (movie clips) after the scene cut (Figure 8). Higher ISC after scene transition suggests that introduction of a new scene leads to an orientation response that is tightly time-locked across subjects, while after this initial response, the interindividual variation in visual sampling increases. This accords with prior studies investigating the effect of scene cuts on gaze positioning and synchronization (Coutrot, Guyader, Ionescu, & Caplier, 2012; Mital, Smith, Hill, & Henderson, 2011). However, the regression analysis did not indicate a significant association between ISC and scene transition, although a consistent positive association was apparent in multiple different time windows (Supplementary Figure SI-4). Since the regression results were carefully controlled for other predictors, unlike the scene cut effect analysis, it is most likely that the observed increase in ISC after a scene change was more accurately modeled with other features, such as the presence of human faces.

Finally, fixation rate was higher when the participants gazed at semantic categories compared to unknown areas or areas outside the presentation window (Figure 5). Scene transitions were also associated with an increased number of fixations, but during intense and rough sounds, the number of fixations decreased. Fewer saccades (decreased fixation rate) have been found to occur during high visual attention and with stationary stimuli (Mahanama et al., 2022). Hence, our results indicate that recognizing and tracking semantic objects and orienting to new scenes require more fixations (and saccades) compared to stationary moments with more static information. Intense and possibly emotionally arousing scenes are most likely visually engaging, and saccades are inhibited to closely monitor the most important source of information.

Pupil size depends on low-level information, emotional arousal, and scene changes

The multiple regression analysis for pupil size achieved high out-of-sample prediction performance (up to $r = 0.5$), indicating a major influence of external stimulus features on pupil size. In the regression models, the low-level feature luminance/entropy (including "luminance," "visual entropy," and "spatial energy") was negatively associated with pupil size, as the main function of the pupil is to regulate how much light enters the eye (Figure 5 & Supplementary Figure SI-4). Perceived unpleasantness of the scene was positively associated with pupil size, and this effect was independent of the overall scene luminosity. Based on the clustering analysis, the unpleasantness predictor was combined from the evaluated features of "unpleasantness," "aroused," "aggressive," and "pain." This result indicates that unpleasant scenes were highly arousing in the movie stimuli. Additionally, we found that audio intensity/roughness was positively associated with pupil size. Intense sounds are arousing and alerting (Dean, Bailes, & Schubert, 2011; Di Stefano & Spence, 2022; Ilie & Thompson, 2006; Trevor, Arnal, & Frühholz, 2020), which was also supported by the moderately positive association between perceived arousal and measured audio intensity ($r = 0.39$, Supplementary Figure SI-5). The observed effect of emotional arousal on pupil size is in line with the previous controlled studies showing that both highly unpleasant and pleasant conditions dilate the pupil compared to neutral conditions (Bradley et al., 2008; R. R. Henderson et al., 2018; Partala & Surakka, 2003). Some studies also indicate that negative emotions may show a more robust effect on pupil size than positive emotions (Babiker et al., 2013; Kawai et al., 2013), a result consistent with the present study.

The analysis of scene transition dynamics (Figure 8) showed that the pupil begins to contract rapidly after a scene transition, with the peak contraction reached between 500 and 800 ms after the cut. The pupil dilates back to the baseline between 1,150 and 1,500 ms after the scene transition. This effect occurred on the same temporal scale as the ISC response to scene transition. Importantly, the regression analysis also found a negative association between pupillary response and scene transition after controlling for the effects of other stimulus features. Pupil constriction at scene transitions may eventually reflect a more general response for abrupt changes in the visual input, since the same phenomenon is found with controlled and simple stimulus changes under isoluminous conditions (Kimura et al., 2014). Nevertheless, the scene transitions should be taken into consideration in future studies using cinematic stimuli.

Blinking as an indicator of attentional engagement

Participants lost only a few percent (median 2%) of the total stimulus viewing time due to blinking. Fewer participants blinked at the very beginning of a new scene (<400 ms from the scene cut) compared to the later time points. The increase in ISC (and pupillary response) after scene transition began simultaneously with the inhibition of blinking, providing further evidence that when a scene changes, people really attend to the most important/interesting stimuli first, and only afterward can they blink without the

risk of losing visual information (Figure 8). In line with this, we also observed a weak but consistent negative correlation (500-ms scale: $r = -0.14$) between blinking and ISC when analyzing the full time series of the experiments (Figure 3). In the regression analysis, we did not observe a significant association between participant-specific blink rates and scene cuts (Figure 5). People blink relatively infrequently when watching movies, around once per 5 seconds in our data, and this sparsity of blinks over the entire time course most likely explains why significant associations between participant-specific blink rate and scene change were not observed in the regression analysis. However, the blink rates were negatively associated with audio intensity/rough sounds and perceived body movement, which could be indicative of behaviorally intense moments in the scenes. This is also suggested by the relatively high correlations between arousal, aggression, body movement, auditory roughness, and audio intensity in the movie stimuli (Supplementary Figure SI-5). People also blinked more often when they were gazing at the background, which could indicate that blinks occur during scenes with fewer important features, such as faces. In contrast with other modeled eye parameters, the regression models were poor at predicting the out-of-sample blink rates, suggesting that changes in blink rates may be more intrinsic than stimulus driven.

All in all, the results suggest that attentional engagement modulates the blinking behavior. Previous studies focusing on blinks have reported that blinking is inhibited during attentional engagement (Ranti, Jones, Klin, & Shultz, 2020; Shin et al., 2015), blinking decreases as a function of attentional demand (Oksama & Hyönä, 2016), and blinks tend to occur at attentional breakpoints (Nakano et al., 2009; Wyly et al., 2024), supporting the current results derived from dynamic modeling of complex social scenes. Video stimuli with a narrative storyline synchronize blinking and yield lower blink rates compared to documentaries (Nakano et al., 2009; Shin et al., 2015), and blinking becomes increasingly synchronous across participants who are highly interested in the topic compared to those that are not (Nakano & Miyazaki, 2019). Functional brain imaging has demonstrated that the brain activity shifts from the dorsal attention network to the default-mode network after blink onset, which suggests that blinking is associated with attentional disengagement (Nakano et al., 2013). Our results thus add evidence for the view that blinking dynamically reveals the participant's attention or interest when they watch complex social scenes, but high-level social information has only negligible effects on blinking. Based on these results, synchronized blinks could be used as indicators of attentional breakpoints in the stimulus, and periods without (synchronized) blinks would indicate attentional engagement.

Intense and rough sounds as an indirect measure of emotional arousal

The combined audio intensity and roughness consistently modulated pupil size, fixation rate, and blink rate in movie scenes. The most likely interpretation is that rough sounds indirectly capture intense periods in the movie stimuli, leading to people paying close attention to these events (suppressed blinking and lower fixation rate) and becoming emotionally aroused (pupil dilation). The audio intensity and roughness are physical properties that are easily extracted from the audio stream and can thus be incorporated in any temporal analysis scale. Perceptual measures, such as evaluated emotional arousal, are subjective, and the temporal resolution depends on how the perceptual ratings are collected. Hence, audio intensity and roughness could serve as an indirect and objective measure of the behavioral intensity of movie scenes and emotion arousal of perceivers. Similar findings demonstrating that the auditory roughness relates to emotionally arousing (negative) events have been reported (Dean et al., 2011; Di Stefano & Spence, 2022; Ilie & Thompson, 2006; Trevor et al., 2020).

Limitations and future directions

The stimulus contained movies—more specifically, socially rich Hollywood movie clips and two full-movie stimuli (an American horror movie and a Finnish feature film)—which do not entirely reflect real-world social situations. Movies have been intentionally directed and edited to capture attention, and with editing and camera work, the viewers' visual attention is externally modulated in ways that diverge from how visual attention operates in real life. However, movies have the advantage of being controllable dynamic stimuli in the laboratory, since traditional eye tracking still requires a stationary measurement protocol. Viewing movies is a form of passive observation, while vision is also used for guiding actions as well as for both gathering information from others and signaling back to them (Risko, Richardson, & Kingstone, 2016). This dual function of gaze differentiates gaze patterns in interactive situations from those during passive observation (Gobel, Kim, & Richardson, 2015; Risko & Kingstone, 2011). Future studies could collect first-person footage of everyday social interaction with wearable cameras, while measuring the participants' eye movements with wearable eye trackers.

The initial feature space for the stimulus models was designed by researchers. Hence, the results do not rule out the possibility that some other perceptual information can have a significant influence on visual attention. Additionally, the time window approach adopted in the current analyses treated consecutive time

windows as independent samples. Since the consecutive samples are never truly independent, even more detailed information could be extracted from the data if the temporal patterns over multiple time windows were also analyzed. Future studies could investigate whether the social context could yield specific temporal gaze patterns with dynamic stimuli. Finally, the results delineate the shared external influences of the gaze behavior, especially in social contexts. In doing so, they describe the average associations between external influences and gaze metrics over individuals. Hence, they do not reveal any participant-specific factors or between-participant differences in how people pay attention to social contexts. Future studies on attentional differences among individuals are needed to understand how well current models predict gaze behavior in certain individuals and which individual-level factors would improve those predictions.

Conclusions

Our results yield a comprehensive model on how gaze orienting, pupillary response, and blinking behavior are uniquely influenced by low-, mid-, and high-level stimulus features in dynamic social scenes in cinema. Human faces, especially eyes, and low-level information (motion and visual information density) are associated with gaze orientation, while high-level socioemotional context has a negligible association with gaze. Pupillary response, in turn, was modulated by low-level information (mostly luminance) and emotional arousal, while fewer blinks occurred after scene transitions and during behaviorally intense scenes (loud noises, body movement). This work advances our understanding of visual processing in social contexts by modeling different gaze parameters dynamically and simultaneously. Future research should explore how gaze behavior is modulated outside the laboratory to bridge the gap between controlled studies and vision modulation in everyday life.

Keywords: social vision, visual attention, visual cognition, social perception, naturalistic stimuli

Acknowledgments

The authors thank Elina Lewandowski for her help in the eye-tracking data collection and Maya Rassouli for the image segmentation quality control.

Supported by Turku University Foundation and Alfred Kordelin Foundation grants to SS and by Finnish Governmental Research Funding for Turku University Hospital and for the Western Finland collaborative area to SS and LN and by the European Research Council (ERC-ADV #101141656), Jane and Aatos Erkko Foundation to LN.

Data and code availability: The analysis scripts are freely available in the project's GitHub repository (https://github.com/santavis/social-vision-in-cinema) (Santavirta, 2024). The subjects' permissions for public distribution of the subject-level eye-tracking data were not collected and hence the eye-tracking data cannot be distributed. The stimulus movies can be made available for researchers upon request, but copyrights preclude public redistribution of the stimulus set. Supplementary Table SI-1 provides short descriptions of the movie clips in Experiment 1. This study was not preregistered.

CrediT authorship contribution statements: S.S.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing–original draft, Writing–review & editing, Visualization, Project administration. **B.P.**: Conceptualization, Investigation, Resources, Data curation, Writing–review & editing. **K.S.**: Conceptualization, Investigation, Resources, Data curation, Writing–review & editing. **J.H.**: Conceptualization, Methodology, Resources, Writing–review & editing. **L.N.**: Conceptualization, Methodology, Resources, Writing–original draft, Writing–review & editing, Visualization, Supervision, Project administration.

Commercial relationships: none.
Corresponding author: Severi Santavirta.
Email: svtsan@utu.fi.
Address: Turku PET Centre, Kiinamyllynkatu 4–6, 20520 University of Turku, Finland.

References

Abele, A. E., & Wojciszke, B. (2014). Communal and agentic content in social cognition: A dual perspective model. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. *50*, pp. 195–255). New York: Academic Press, https://doi.org/10.1016/B978-0-12-800284-1.00004-7.

Abrams, R. A., & Christ, S. E. (2003). Motion onset captures attention. *Psychological Science, 14*(5), 427–432, https://doi.org/10.1111/1467-9280.01458.

Adolphs, R., Nummenmaa, L., Todorov, A., & Haxby, J. V. (2016). Data-driven approaches in the investigation of social perception. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 371*(1693), https://doi.org/10.1098/rstb.2015.0367.

Ayres, P., Lee, J. Y., Paas, F., & van Merriënboer, J. J. G. (2021). The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in Psychology, 12*, 702538, https://doi.org/10.3389/fpsyg.2021.702538.

Babiker, A., Faye, I., & Malik, A. (2013). Pupillary behavior in positive and negative emotions. In *2013 IEEE International Conference on Signal and Image Processing Applications* (pp. 379–383). New York: IEEE, https://doi.org/10.1109/ICSIPA.2013.6708037.

Bellitto, G., Proietto Salanitri, F., Palazzo, S., Rundo, F., Giordano, D., & Spampinato, C. (2021). Hierarchical domain-adapted feature learning for video saliency prediction. *International Journal of Computer Vision, 129*(12), 3216–3232, https://doi.org/10.1007/s11263-021-01519-y.

Bindemann, M., Burton, A. M., Hooge, I. T. C., Jenkins, R., & de Haan, E. H. F. (2005). Faces retain attention. *Psychonomic Bulletin & Review, 12*(6), 1048–1053, https://doi.org/10.3758/BF03206442.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008). Gaze selection in complex social scenes. *Visual Cognition, 16*(2–3), 341–355, https://doi.org/10.1080/13506280701434532.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision Research, 49*(24), 2992–3000, https://doi.org/10.1016/j.visres.2009.09.014.

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45*(4), 602–607, https://doi.org/10.1111/j.1469-8986.2008.00654.x.

Bruckert, A., Christie, M., & Le Meur, O. (2023). Where to look at the movies: Analyzing visual attention to understand movie editing. *Behavior Research Methods, 55*(6), 2940–2959, https://doi.org/10.3758/s13428-022-01949-7.

Calvo, M. G., & Nummenmaa, L. (2008). Detection of emotional faces: Salient physical features guide effective visual search. *Journal of Experimental Psychology. General, 137*(3), 471–494, https://doi.org/10.1037/a0012771.

Cherng, Y.-G., Baird, T., Chen, J.-T., & Wang, C.-A. (2020). Background luminance effects on pupil size associated with emotion and saccade preparation. *Scientific Reports, 10*(1), 15718, https://doi.org/10.1038/s41598-020-72954-z.

Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2018). Predicting human eye fixations via an LSTM-based saliency attentive model. *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 27*(10), 5142–5154, https://doi.org/10.1109/TIP.2018.2851672.

Coutrot, A., & Guyader, N. (2014). How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of Vision, 14*(8), 5, https://doi.org/10.1167/14.8.5.

Coutrot, A., Guyader, N., Ionescu, G., & Caplier, A. (2012). Influence of soundtrack on eye movements during video exploration. *Journal of Eye Movement Research, 5*(4), 1–10, https://doi.org/10.16910/jemr.5.4.2.

Dean, R. T., Bailes, F., & Schubert, E. (2011). Acoustic intensity causes perceived changes in arousal levels in music: An experimental investigation. *PLoS One, 6*(4), e18591, https://doi.org/10.1371/journal.pone.0018591.

Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). RetinaFace: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5203–5212). New York: IEEE, https://doi.org/10.1109/cvpr42600.2020.00525.

Di Stefano, N., & Spence, C. (2022). Roughness perception: A multisensory/crossmodal perspective. *Attention, Perception & Psychophysics, 84*(7), 2087–2114, https://doi.org/10.3758/s13414-022-02550-y.

Dima, D. C., Tomita, T. M., Honey, C. J., & Isik, L. (2022). Social-affective features drive human representations of observed actions. *ELife, 11*, https://doi.org/10.7554/eLife.75027.

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision, 10*(10), 28, https://doi.org/10.1167/10.10.28.

End, A., & Gamer, M. (2017). Preferential processing of social features and their interplay with physical saliency in complex naturalistic scenes. *Frontiers in Psychology, 8*, 418, https://doi.org/10.3389/fpsyg.2017.00418.

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science, 27*(2), 67–73, https://doi.org/10.1177/0963721417738825.

Franchak, J. M., Heeger, D. J., Hasson, U., & Adolph, K. E. (2016). Free viewing gaze behavior in infants and adults. *Infancy, 21*(3), 262–287, https://doi.org/10.1111/infa.12119.

Gobel, M. S., Kim, H. S., & Richardson, D. C. (2015). The dual function of social gaze. *Cognition, 136*, 359–364, https://doi.org/10.1016/j.cognition.2014.11.040.

Grillon, C., & Baas, J. (2003). A review of the modulation of the startle reflex by affective states and its application in psychiatry. *Clinical Neurophysiology, 114*(9), 1557–1579, https://doi.org/10.1016/S1388-2457(03)00202-5.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *The Journal of Neuroscience, 28*(10), 2539–2550, https://doi.org/10.1523/JNEUROSCI.5487-07.2008.

Hayes, T. R., & Henderson, J. M. (2021). Deep saliency models learn low-, mid-, and high-level features to predict scene attention. *Scientific Reports, 11*(1), 18434, https://doi.org/10.1038/s41598-021-97879-z.

Heikkilä, T., Laine, O., Savela, J., & Nummenmaa, L. (2021). *Onni: An online experiment platform for research*. Zenodo. https://doi.org/10.5281/zenodo.4719220.

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

Henderson, R. R., Bradley, M. M., & Lang, P. J. (2018). Emotional imagery and pupil diameter. *Psychophysiology, 55*(6), e13050, https://doi.org/10.1111/psyp.13050.

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science, 132*(3423), 349–350, https://doi.org/10.1126/science.132.3423.349.

Hyönä, J., Tommola, J., & Alaja, A. M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology A, Human Experimental Psychology, 48*(3), 598–612, https://doi.org/10.1080/14640749508401407.

Ilie, G., & Thompson, W. F. (2006). A Comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception, 23*(4), 319–329, https://doi.org/10.1525/mp.2006.23.4.319.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10–12), 1489–1506, https://doi.org/10.1016/s0042-6989(99)00163-7.

Jain, S., Yarlagadda, P., Jyoti, S., Karthik, S., Subramanian, R., & Gandhi, V. (2021). ViNet: Pushing the limits of visual modality for audio-visual saliency prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 3520–3527). New York: IEEE, https://doi.org/10.1109/IROS51168.2021.9635989.

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron, 89*(1), 221–234, https://doi.org/10.1016/j.neuron.2015.11.028.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154*(3756), 1583–1585, https://doi.org/10.1126/science.154.3756.1583.

Karjalainen, T., Karlsson, H. K., Lahnakoski, J. M., Glerean, E., Nuutila, P., Jaaskelainen, I. P., … Nummenmaa, L. (2017). Dissociable roles of cerebral mu-opioid and type 2 dopamine receptors in vicarious pain: A combined PET-fMRI study. *Cerebral Cortex, 27*(8), 4257–4266, https://doi.org/10.1093/cercor/bhx129.

Karjalainen, T., Seppala, K., Glerean, E., Karlsson, H. K., Lahnakoski, J. M., Nuutila, P., … Nummenmaa, L. (2018). Opioidergic regulation of emotional arousal: A combined PET-fMRI study. *Cerebral Cortex, 29*(9), 4006–4016, https://doi.org/10.1093/cercor/bhy281.

Kawai, S., Takano, H., & Nakamura, K. (2013). Pupil diameter variation in positive and negative emotions with visual stimulus. In *2013 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 4179–4183). New York: IEEE, https://doi.org/10.1109/SMC.2013.712.

Keles, U., Kliemann, D., Byrge, L., Saarimäki, H., Paul, L. K., Kennedy, D. P., … Adolphs, R. (2022). Atypical gaze patterns in autistic adults are heterogeneous across but reliable within individuals. *Molecular Autism, 13*(1), 39, https://doi.org/10.1186/s13229-022-00517-2.

Kessler, H., Doyen-Waldecker, C., Hofer, C., Hoffmann, H., Traue, H. C., & Abler, B. (2011). Neural correlates of the perception of dynamic versus static facial expressions of emotion. *Psycho-Social Medicine, 8*, Doc03, https://doi.org/10.3205/psm000072.

Kimura, E., Abe, S., & Goryo, K. (2014). Attenuation of the pupillary response to luminance and color changes during interocular suppression. *Journal of Vision, 14*(5), 14, https://doi.org/10.1167/14.5.14.

Kirillov, A., Girshick, R., He, K., & Dollar, P. (2019). Panoptic feature pyramid networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6399–6408). New York: IEEE, https://doi.org/10.1109/cvpr.2019.00656.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision, 13*(2–3), 201–214, https://doi.org/10.1163/156856800741216.

Lahnakoski, J. M., Glerean, E., Salmi, J., Jaaskelainen, I. P., Sams, M., Hari, R., … Nummenmaa, L. (2012). Naturalistic FMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception.

Frontiers in Human Neuroscience, 6, 233, https://doi.org/10.3389/fnhum.2012.00233.

Lartillot, O., & Toiviainen, P. (2007). A Matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects*. https://www.dafx.de/paper-archive/2007/Papers/p237.pdf.

Liao, H.-I., Kashino, M., & Shimojo, S. (2021). Attractiveness in the eyes: A possibility of positive loop between transient pupil constriction and facial attraction. *Journal of Cognitive Neuroscience, 33*(2), 315–340, https://doi.org/10.1162/jocn_a_01649.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Computer Vision—ECCV 2014,* 740–755. New York: Springer, https://doi.org/10.1007/978-3-319-10602-1_48.

Lou, J., Lin, H., Marshall, D., Saupe, D., & Liu, H. (2022). TranSalNet: Towards perceptually relevant visual saliency prediction. *Neurocomputing, 494*, 455–467, https://doi.org/10.1016/j.neucom.2022.04.080.

Louhimies, A., (Director). (2008). *Käsky [Film]*. Helsinki-Filmi, Two Thirty Five, Mogador Film.

Maffei, A., & Angrilli, A. (2019). Spontaneous blink rate as an index of attention and emotion during film clips viewing. *Physiology & Behavior, 204*, 256–263, https://doi.org/10.1016/j.physbeh.2019.02.037.

Mahanama, B., Jayawardana, Y., Rengarajan, S., Jayawardena, G., Chukoskie, L., Snider, J., . . . Jayarathna, S. (2022). Eye movement and pupil measures: A review. *Frontiers in Computer Science, 3*, https://doi.org/10.3389/fcomp.2021.733531.

Mahapatra, D., Winkler, S., & Yen, S.-C. (2008). Motion saliency outweighs other low-level features while watching videos. *Human Vision and Electronic Imaging XIII, 6806*, 246–255, https://doi.org/10.1117/12.766243.

Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation, 3*(1), 5–24, https://doi.org/10.1007/s12559-010-9074-z.

Morrisey, M. N., Hofrichter, R., & Rutherford, M. D. (2019). Human faces capture attention and attract first saccades without longer fixation. *Visual Cognition, 27*(2), 158–170, https://doi.org/10.1080/13506285.2019.1631925.

Naber, M., Frässle, S., Rutishauser, U., & Einhäuser, W. (2013). Pupil size signals novelty and predicts later retrieval success for declarative memories of natural scenes. *Journal of Vision, 13*(2), 11, https://doi.org/10.1167/13.2.11.

Nakano, T., Kato, M., Morito, Y., Itoi, S., & Kitazawa, S. (2013). Blink-related momentary activation of the default mode network while viewing videos. *Proceedings of the National Academy of Sciences of the United States of America, 110*(2), 702–706, https://doi.org/10.1073/pnas.1214804110.

Nakano, T., & Miyazaki, Y. (2019). Blink synchronization is an indicator of interest while viewing videos. *International Journal of Psychophysiology, 135*, 1–11, https://doi.org/10.1016/j.ijpsycho.2018.10.012.

Nakano, T., Yamamoto, Y., Kitajo, K., Takahashi, T., & Kitazawa, S. (2009). Synchronization of spontaneous eyeblinks while viewing video stories. *Proceedings. Biological Sciences /The Royal Society, 276*(1673), 3635–3644, https://doi.org/10.1098/rspb.2009.0828.

Nummenmaa, L., Hyönä, J., & Calvo, M. G. (2010). Semantic categorization precedes affective evaluation of visual scenes. *Journal of Experimental Psychology. General, 139*(2), 222–246, https://doi.org/10.1037/a0018858.

Nummenmaa, L., Lukkarinen, L., Sun, L., Putkinen, V., Seppala, K., Karjalainen, T., . . . Tiihonen, J. (2021). Brain basis of psychopathy in criminal offenders and general population. *Cerebral Cortex, 31*(9), 4104–4114, https://doi.org/10.1093/cercor/bhab072.

Nummenmaa, L., Smirnov, D., Lahnakoski, J. M., Glerean, E., Jaaskelainen, I. P., Sams, M., . . . Hari, R. (2014). Mental action simulation synchronizes action-observation circuits across individuals. *The Journal of Neuroscience, 34*(3), 748–757, https://doi.org/10.1523/JNEUROSCI.0352-13.2014.

Nuthmann, A., Schütz, I., & Einhäuser, W. (2020). Salience-based object prioritization during active viewing of naturalistic scenes in young and older adults. *Scientific Reports, 10*(1), 22057, https://doi.org/10.1038/s41598-020-78203-7.

Oksama, L., & Hyönä, J. (2016). Position tracking and identity tracking are separate systems: Evidence from eye movements. *Cognition, 146*, 393–409, https://doi.org/10.1016/j.cognition.2015.10.016.

Oliva, M., & Anikin, A. (2018). Pupil dilation reflects the time course of emotion recognition in human vocalizations. *Scientific Reports, 8*(1), 4871, https://doi.org/10.1038/s41598-018-23265-x.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America, 105*(32), 11087–11092, https://doi.org/10.1073/pnas.0805664105.

Open Science Collaboration. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science (New York, N.Y.), 349*(6251),

aac4716, https://doi.org/10.1126/science.aac4716.

Osgood, C. E., & Suci, G. J. (1955). Factor analysis of meaning. *Journal of Experimental Psychology, 50*(5), 325–338, https://doi.org/10.1037/h0043965.

Parrigon, S., Woo, S. E., Tay, L., & Wang, T. (2017). CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *Journal of Personality and Social Psychology, 112*(4), 642–681, https://doi.org/10.1037/pspp0000111.

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies, 59*(1–2), 185–198, https://doi.org/10.1016/s1071-5819(03)00017-x.

Pitcher, D., Dilks, D. D., Saxe, R. R., Triantafyllou, C., & Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage, 56*(4), 2356–2363, https://doi.org/10.1016/j.neuroimage.2011.03.067.

Pratt, J., Radulescu, P. V., Guo, R. M., & Abrams, R. A. (2010). It's alive! animate motion captures visual attention. *Psychological Science, 21*(11), 1724–1730, https://doi.org/10.1177/0956797610387440.

Ranti, C., Jones, W., Klin, A., & Shultz, S. (2020). Blink rate patterns provide a reliable measure of individual engagement with scene content. *Scientific Reports, 10*(1), 8267, https://doi.org/10.1038/s41598-020-64999-x.

Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., … Funder, D. C. (2014). The Situational Eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology, 107*(4), 677–718, https://doi.org/10.1037/a0037250.

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., … Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications, 7*, 13289, https://doi.org/10.1038/ncomms13289.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network, 10*(4), 341–350, https://www.ncbi.nlm.nih.gov/pubmed/10695763.

Risko, E. F., & Kingstone, A. (2011). Eyes wide shut: Implied social presence, eye tracking and attention. *Attention, Perception & Psychophysics, 73*(2), 291–296, https://doi.org/10.3758/s13414-010-0042-1.

Risko, Evan F., Richardson, D. C., & Kingstone, A. (2016). Breaking the fourth wall of cognitive science. *Current Directions in Psychological Science, 25*(1), 70–74, https://doi.org/10.1177/0963721415617806.

Rösler, L., End, A., & Gamer, M. (2017). Orienting towards social features in naturalistic scenes is reflexive. *PLoS One, 12*(7), e0182037, https://doi.org/10.1371/journal.pone.0182037.

Roth, N., Rolfs, M., Hellwich, O., & Obermayer, K. (2023). Objects guide human gaze behavior in dynamic real-world scenes. *PLoS Computational Biology, 19*(10), e1011512, https://doi.org/10.1371/journal.pcbi.1011512.

Saarimäki, H. (2021). Naturalistic stimuli in affective neuroimaging: A review. *Frontiers in Human Neuroscience, 15*, 675068, https://doi.org/10.3389/fnhum.2021.675068.

Saarimäki, H., Nummenmaa, L., Volynets, S., Santavirta, S., Aksiuto, A., Sams, M., … Lahnakoski, J. M. (2025). Cerebral topographies of perceived and felt emotions. *Imaging Neuroscience, 3*, https://doi.org/10.1162/imag_a_00517.

Santavirta, S. (2024). *Modelling human social vision with cinematic stimuli*. GitHub. https://github.com/santavis/social-vision-in-cinema.

Santavirta, S., Karjalainen, T., Nazari-Farsani, S., Hudson, M., Putkinen, V., Seppälä, K., … Nummenmaa, L. (2023). Functional organization of social perception in the human brain. *NeuroImage, 272*, 120025, https://doi.org/10.1016/j.neuroimage.2023.120025.

Santavirta, S., Malén, T., Erdemli, A., & Nummenmaa, L. (2024). A taxonomy for human social perception: Data-driven modeling with cinematic stimuli. *Journal of Personality and Social Psychology, 127*(6), 1146–1171, https://doi.org/10.1037/pspa0000415.

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Engbert, R., & Wichmann, F. A. (2019). Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. *Journal of Vision, 19*(3), 1, https://doi.org/10.1167/19.3.1.

Shin, Y. S., Chang, W.-D., Park, J., Im, C.-H., Lee, S. I., Kim, I. Y., … Jang, D. P. (2015). Correlation between inter-blink interval and episodic encoding during movie watching. *PLoS One, 10*(11), e0141242, https://doi.org/10.1371/journal.pone.0141242.

Smith, T. J., & Mital, P. K. (2013). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision, 13*(8), 16, https://doi.org/10.1167/13.8.16.

Steinhauer, S. R., Siegle, G. J., Condray, R., & Pless, M. (2004). Sympathetic and parasympathetic innervation of pupillary dilation during

sustained processing. *International Journal of Psychophysiology, 52*(1), 77–86, https://doi.org/10.1016/j.ijpsycho.2003.12.005.

Stoll, J., Thrun, M., Nuthmann, A., & Einhäuser, W. (2015). Overt attention in natural scenes: Objects dominate features. *Vision Research, 107*, 36–48, https://doi.org/10.1016/j.visres.2014.11.006.

Sutherland, C. A. M., & Young, A. W. (2022). Understanding trait impressions from faces. *British Journal of Psychology, 113*(4), 1056–1078, https://doi.org/10.1111/bjop.12583.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*(5), 643–659, https://doi.org/10.1016/j.visres.2004.09.017.

Tatler, B. W., & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition, 17*(6–7), 1029–1054, https://doi.org/10.1080/13506280902764539.

Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition, 13*(6), 657–665, https://doi.org/10.1080/13506280500410949.

Trevor, C., Arnal, L. H., & Frühholz, S. (2020). Terrifying film music mimics alarming acoustic feature of human screams. *The Journal of the Acoustical Society of America, 147*(6), EL540, https://doi.org/10.1121/10.0001459.

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review, 25*(6), 2005–2015, https://doi.org/10.3758/s13423-018-1432-y.

Varela, V. P. L., Towler, A., Kemp, R. I., & White, D. (2023). Looking at faces in the wild. *Scientific Reports, 13*(1), 783. https://doi.org/10.1038/s41598-022-25268-1.

Võ, M. L.-H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision, 12*(13), 1–14, https://doi.org/10.1167/12.13.3.

Wan, J., (Director). (2016). *The Conjuring 2 [Film]. New Line Cinema, RatPac-Dune Entertainment*, The Safran Company, Atomic Monster.

Wang, H. X., Freeman, J., Merriam, E. P., Hasson, U., & Heeger, D. J. (2012). Temporal eye movement strategies during naturalistic viewing. *Journal of Vision, 12*(1), 16, https://doi.org/10.1167/12.1.16.

Williams, C. C., & Castelhano, M. S. (2019). The changing landscape: High-level influences on eye movement guidance in scenes. *Vision (Basel, Switzerland), 3*(3), https://doi.org/10.3390/vision3030033.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). *Detectron2*. GitHub. https://github.com/facebookresearch/detectron2.

Wyly, S., Jinon, N., Francis, T., Evans, H., Kao, T. L., Lambert, S., … Wilson, R. C. (2024). The psychophysiology of Mastermind: Characterizing response times and blinking in a high-stakes television game show. *Psychophysiology, 61*(3), e14485, https://doi.org/10.1111/psyp.14485.