Motion-corrected eye tracking (MoCET) improves gaze accuracy during visual fMRI experiments

Jiwoong Park[1,2,3], Jae Young Jeon[1,3], Royoung Kim[1,3], Kendrick N. Kay[4], Won Mok Shim[1,2,3*]

*Corresponding author, Email: wonmokshim@skku.edu

[1] Center for Neuroscience Imaging Research, Institute for Basic Science (IBS), Republic of Korea

[2] Department of Biomedical Engineering, Sungkyunkwan University (SKKU), Republic of Korea

[3] Department of Intelligent Precision Healthcare Convergence, Sungkyunkwan University (SKKU), Republic of Korea

[4] Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota

Abstract

Human eye movements are essential for understanding cognition, yet achieving high-precision eye tracking in fMRI remains challenging. Even slight head shifts from the initial calibration position can introduce drift in eye tracking data, leading to substantial gaze inaccuracies. To address this, we introduce Motion-Corrected Eye Tracking (MoCET), a novel approach that corrects drift using head motion parameters derived from the preprocessing of fMRI data. MoCET requires no additional hardware and can be applied retrospectively to existing datasets. We show that it outperforms traditional detrending methods with respect to accuracy of gaze estimation and offers higher spatial and temporal precision compared to MR-based eye tracking approaches. By overcoming a key limitation in integrating eye tracking with fMRI, MoCET facilitates investigations of naturalistic vision and cognition in fMRI research.

Keywords: Eye tracking, Functional MRI, Motion correction, Computational modeling

Introduction

Human eye movements are closely linked to cognitive processes such as perception, attention, and memory[1–3]. Acquiring eye tracking data during functional magnetic resonance imaging (fMRI) experiments offers the possibility of better understanding these processes at both neural and behavioral levels. By combining eye tracking with fMRI data, we can link explicit behavioral signatures of mental states to underlying neural activity, thereby providing insights into how the brain responds to the visual world[4–8]. This integration becomes increasingly crucial as naturalistic free-viewing paradigms gain prominence in cognitive neuroscience: in naturalistic environments, precise eye tracking is necessary to understand the brain's dynamic processing of visual information[9–12].

Recent advances in eye tracking methods within fMRI research typically involve camera-based systems using endoscopic optic fiber[13,14]. These systems usually monitor one eye with a single camera, using computer vision techniques to detect pupil position. While this approach enables precise eye tracking data, it faces major technical challenges, such as hardware setup constraints in the MRI environment and limitations in calibration data quality.

Camera-based eye tracking accuracy relies critically on calibration that maps pupil positions to gaze positions. The MRI scanner's limited visual field makes precise calibration difficult, as participants fixate closely-spaced calibration points. More importantly, even slight head movements from the calibration position compromise accuracy. Unlike behavioral experiments that can use chin rests or bite bars, such head restraints are not feasible in MRI. Consequently, head shifts significantly reduce gaze accuracy, as the calibrated model is misaligned with eye tracking data collected during the experiment[15–17].

Behavioral experiments have recently introduced several methods for addressing head shifts, such as recording head movements through supplementary cameras, motion sensors, or a wide-field camera that captures both the participant's eye and head, allowing for separate tracking of eye and head movements[18–20]. However, these solutions are challenging to implement in fMRI experiments. Installing additional cameras or motion sensors in the head coil is not feasible in most cases, and the narrow field of view (FOV) of the camera, focused on a single eye, limits the ability to capture head movements. This challenge emphasizes the need for new

solutions that maintain accuracy despite the inevitable head movements that occur in fMRI experiments.

In this study, we investigate the impact of head motions on gaze accuracy during fMRI scans. We first develop computational simulations that confirm head movements significantly affect gaze estimation accuracy. Based on this finding, we then propose an eye tracking with head motion correction approach, which addresses head movement challenges without requiring additional hardware and can be retrospectively applied to existing data. Our approach, termed Motion-Corrected Eye tracking (MoCET), leverages head motion parameters derived from the preprocessing of fMRI data to correct errors in eye tracking data.

Using high-quality eye tracking data collected during free-viewing tasks, we show that MoCET consistently outperforms uncorrected data and traditional detrending methods. We also compare MoCET with magnetic resonance-based eye tracking[21], which estimates gaze position directly from eyeball region signals in fMRI data. We find that while MR-based methods successfully capture general gaze direction, they lack spatial precision. In contrast, we show that MoCET provides precise gaze locations, making it more suitable for detailed behavioral and neuroimaging analyses.

To facilitate the application of MoCET across diverse fMRI experiments, we provide a user-friendly Python package (https://github.com/jwparks/mocet). Additionally, we release the high-quality free-viewing eye tracking dataset used in this paper. This dataset, containing multiple calibration periods, enables both calibration and validation of eye tracking models, and may serve as a useful benchmark for advancing eye tracking methods, including MR-based eye tracking[21–24].

Results

Impact of head motion on gaze tracking accuracy

We simultaneously collected fMRI and eye tracking data while participants engaged in interactive Minecraft[25]-based video game tasks (Figure 1A). These tasks allowed free gaze movement (instead of central fixation), and were designed to provide a diverse range of visual stimuli and cognitive demands, enabling assessment of gaze tracking accuracy under different

conditions. To assist in eye tracking, two eye tracking calibration stages were included during the actual data collection. For each stage, participants fixated on sequentially appearing green dots. The first *calibration stage* involved 24 dots (two repetitions of a 12-dot sequence), while the second *validation stage* occurring at the end of the experiment required fixation on 12 dots. The data collected from the initial calibration stage were used to fit our eye tracking model (Figure 1B).

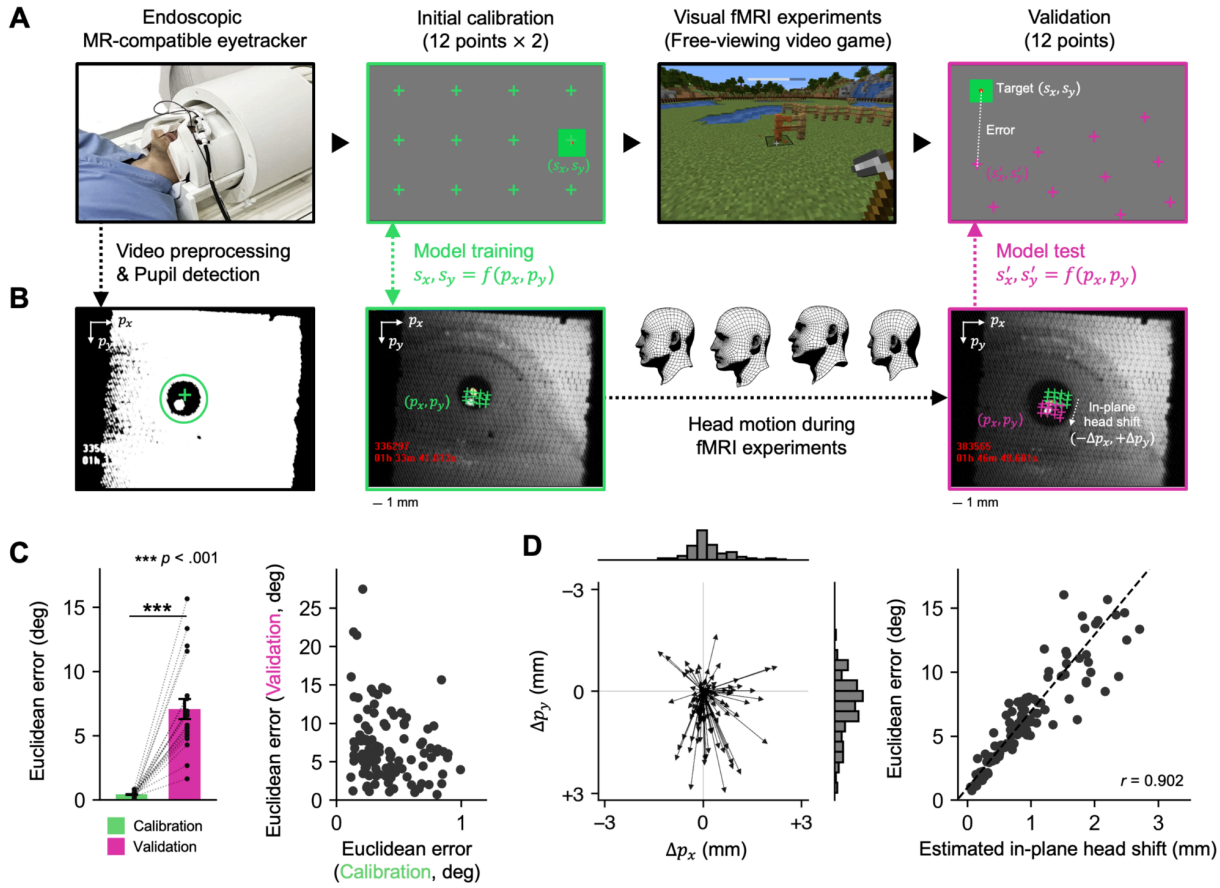

Figure 1. Eye tracking in fMRI experiments. (A) An endoscopic eye tracking camera captures participants' eye movements throughout the experiment. At an initial calibration stage at the beginning of each experiment, 12 calibration targets are presented sequentially on the screen, and participants are instructed to fixate on these targets. Gaze accuracy is evaluated at the end of the experiment in a validation stage by measuring the distance between the predicted gaze location and the actual target location. (B) Eye tracking video is preprocessed using low-pass filtering and binarization to isolate the pupil and remove spatial noise. Pupil coordinates recorded during the calibration stage are used to train the eye tracking model. In-plane head shifts, occurring within the p_x-p_y plane, are estimated by comparing pupil coordinates across the calibration and validation stages. (C) Gaze accuracy significantly decreases during the validation stage (Left: error bars indicate $\pm$ 1 s.e.m. across participants). No significant positive relationship is observed between calibration error and validation error (Right). (D) Downward shifts ($+\Delta p_y$) from the initial calibration are common, most likely caused by vertical head motions (e.g., pitching) that frequently occur during fMRI

experiments (Left). The extent of in-plane head shifts (Δp_x, Δp_y) shows a strong linear relationship with eye tracking error during the validation stage (Right). (C-D) Each arrow and dot represents eye tracking data from individual runs across all participants.

Our eye tracking approach used a radial basis interpolation model that is trained on the initial calibration data to map pupil positions to corresponding screen coordinates[26–28]. While model performance during initial calibration was accurate, performance significantly declined during the experimental sessions, showing degraded accuracy at the validation stage (Figure 1C, calibration error: 0.435 degree; validation error: 7.093 degree, $t(19) = -8.436$, $p < 0.001$). To determine whether the observed error during the validation stage stemmed from poor fixation during the calibration stage, we examined their correlation. The validation stage error did not show positive relationship with the calibration error (Figure 1C, $r = -0.221$), suggesting that gaze inaccuracy is not related to participants' inherent fixation instability but is likely driven by other sources.

We hypothesized that small head movements during fMRI sessions prevent the calibration model from accurately tracking pupil positions. Specifically, in-plane head movements—those occurring in the p_x–p_y plane perpendicular to the eye tracking camera—shift the recorded pupil coordinates, creating discrepancies between the calibration and validation stages. These shifts lead to inaccuracies in mapping pupil positions to their corresponding screen locations. Our data showed that pupil coordinates tend to shift downwards from their initial calibration positions. Importantly, we found that even minor head shifts typical in fMRI experiments can cause substantial inaccuracies in eye tracking results (Figure 1D, $r = 0.902$, $p < 0.001$).

Head motion is a recognized challenge in gaze tracking[15–17], often addressed by measuring head motion independently from eye movement, using motion sensors or video processing of visual features. However, implementing additional head motion sensors within an MRI scanner is technically challenging. Moreover, the limited FOV of endoscopic eye tracking cameras prevents capturing stable feature points, such as the nose, needed for head position estimation. Features visible within the camera's FOV, like the lacrimal caruncle or eyelid, are highly sensitive to blinks and eye movements, rendering them unreliable for head motion estimation. These limitations highlight a critical need for an effective method to capture and compensate for head motion, ensuring accurate gaze tracking throughout fMRI experiments.

To rigorously demonstrate that participant head motion is the primary source of eye tracking errors rather than errors in instrumental setup or suboptimal fixations during calibration, we developed a computational simulation using a 3D geometry-based eyeball model. This approach offers a controlled framework to isolate and quantify the impact of head motion on gaze accuracy. The model included head and eyeball components that could each move independently. The head-ball component allowed six degrees of freedom (DoF): translation along x, y, and z axes, and rotation including roll, pitch, and yaw. The eyeball was physically connected to and moved with the head, but was able to rotate independently to simulate gaze direction. Our simulation recorded pupil position on the eyeball using a virtual eye tracker and applied the same analysis procedures used for human participants (e.g., extract the pupil position, fit the eye tracking calibration model).

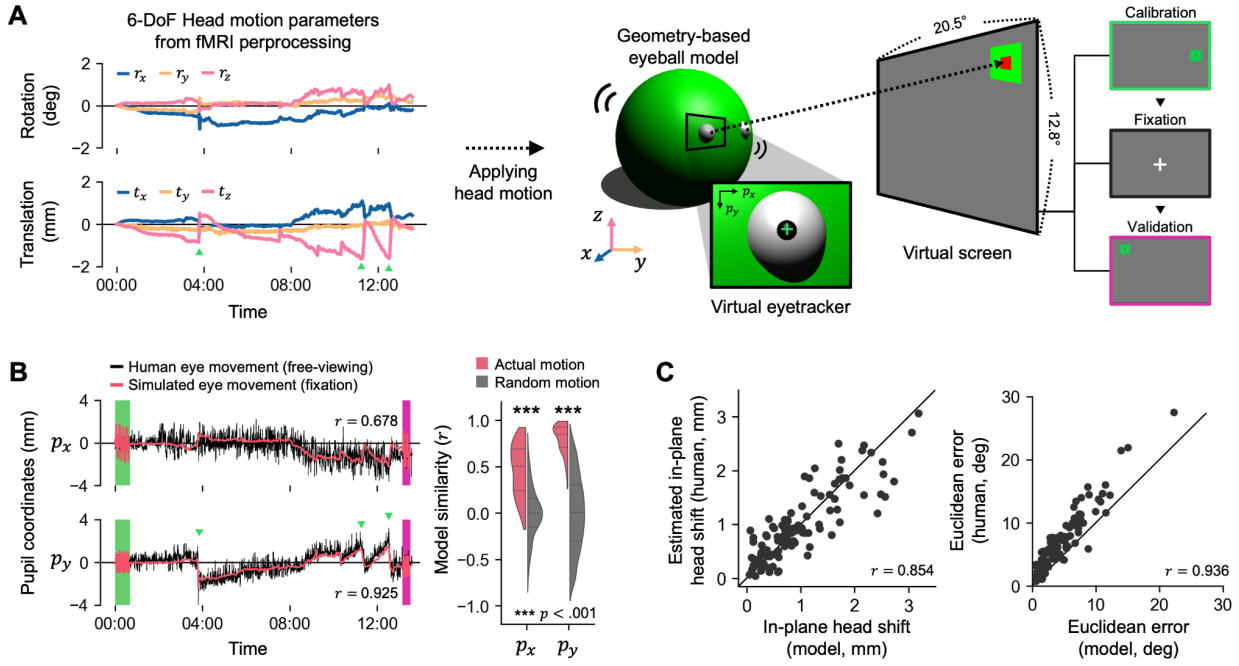

Figure 2. Computational simulation of eye tracking during the influence of head motion. (A) Six-degree-of-freedom (6 DoF) head motion parameters derived from the preprocessing of fMRI data were used to simulate eye movements with a geometry-based eyeball model. The model performed a calibration task requiring fixation on 12 calibration points on the screen (repeated twice), a central fixation task, and a validation task requiring fixation on the same 12 points. Simulated eye movements were rendered as images, and pupil coordinates were tracked using the same method applied to human participants. (B) Comparison of human and simulated eye movements. Despite the model performing a central fixation task, its global eye movement trends closely resembled those observed in human eye tracking data. Abrupt changes in the vertical pupil coordinate (p_y)—indicated by green triangles—occurred in opposite directions to vertical head motion along the z-axis as shown by translational motion in panel (A). The initial calibration and validation stages are highlighted by green and magenta rectangles, respectively (Left). Simulated horizontal and vertical pupil coordinates derived from participants' actual head motion demonstrated significantly higher similarity to human data compared to simulations using randomized head motion generated by

phase-shifting permutation (Right). Statistical significance was determined by comparing actual values to the null distribution of similarity, with horizontal lines indicating quantiles. (C) The model's estimated in-plane head shift (perpendicular to the eye tracking camera) exhibited a strong linear relationship with the estimated in-plane head shift observed in human data (Left). Simulated gaze inaccuracies were also positively correlated with those in human data, although human errors were generally larger due to the imprecision of actual human fixations (Right). Each dot represents eye tracking data from individual runs across all participants, and the diagonal line ($y = x$) represents perfect correspondence.

To simulate the impact of human head motion on gaze accuracy, we applied 6-DoF head motion parameters sampled from actual fMRI data to our model. The model performed an initial calibration task (same as human subjects) and then performed a central fixation task for about 13 minutes. (Note that during the central fixation task, pupil coordinate changes directly reflect the applied head motion effects.) Next, the model performed a validation task, which was used to evaluate gaze accuracy (Figure 2A). The model parameters (head/eyeball radius, eye location) matched average Asian head proportions[29–31], while the field of view of the camera, screen size, and eye-to-screen distance closely replicated our human fMRI experiment setup (See Supplementary figure 1, Supplementary Figure 2 and Methods).

Our eyeball simulation provides two key insights: First, comparing simulated and human gaze trajectories reveals the amount of variance in human eye tracking data that is attributable to head motion. Similar trajectories between the simulated model and human participants indicate that the observed drift in pupil coordinates is driven by head motion rather than actual eye movements. Second, although head motion affects pupil coordinates in a complex, non-linear manner, the simulation allows us to assess whether these effects can be effectively corrected using linear techniques. If effects are approximately linear, this would justify the use of fMRI-derived head motion parameters in a linear regression procedure to correct for head motion effects.

Our model simulation demonstrated that both horizontal (p_x, average model similarity: $r = 0.460$) and vertical (p_y, average model similarity: $r = 0.801$) eye movements can be replicated using motion parameters derived from participants' actual head motion, with vertical shifts (p_y) particularly well-aligned with model simulation due to characteristic pitching movements (Figure 2B). To assess statistical significance, we generated randomized head motion parameters by randomly shuffling the phase spectra, and simulated eye movements based on these randomized parameters, creating a null distribution of similarity values ($n = 100$ per simulation). Comparing actual simulation similarity to this distribution, with results aggregated

across datasets using Fisher's method[32], we find statistical significance for both horizontal and vertical pupil coordinates ($p < 0.001$, indicating that head motion drives global noise in human eye movements.

We found that the model simulation accurately replicated both the magnitude of head shift from initial calibration to final validation and the resulting gaze inaccuracies caused by head motion in individual human eye tracking data. Head shifts in the simulation showed a strong linear relationship with estimated human head shifts ($r = 0.854$). Similarly, the simulation reliably reproduced individual gaze inaccuracies, showing a strong positive correlation between simulated and actual gaze errors ($r = 0.936$). While human gaze errors were slightly larger than simulation gaze errors ($t_{paired}(110) = 14.47,\ p < 0.001$) this likely reflects the fact that actual human fixations have limited accuracy.

These findings highlight the effectiveness of the geometry-based model in isolating head motion effects on eye tracking data. By factoring out task-related variability in eye position and human-specific idiosyncrasies, the simulation confirms that even minor head movements can introduce substantial inaccuracies in eye tracking data.

Compensating for head movement with Motion-Corrected Eye Tracking

Our model simulation demonstrated that head motion alone can cause global drift in eye tracking data, even in the absence of actual eye movement. While many MR-compatible eye trackers offer drift correction, these methods assume that participants fixate on known locations during the experiment. This assumption fails in free-viewing experiments such as movie-viewing or playing video games, since the participant's gaze is not restricted to predefined targets. Without fixed fixation points, conventional drift correction cannot determine the offset between recorded and actual gaze positions.

Based on our computational simulation, which showed that head motion introduces systematic, approximately linear noise in pupil coordinates, we developed a correction approach. Our method, Motion-Corrected Eye Tracking (MoCET), uses head motion parameters as nuisance regressors to account for motion-induced drift, thereby helping to isolate actual eye movements (Figure 3A).

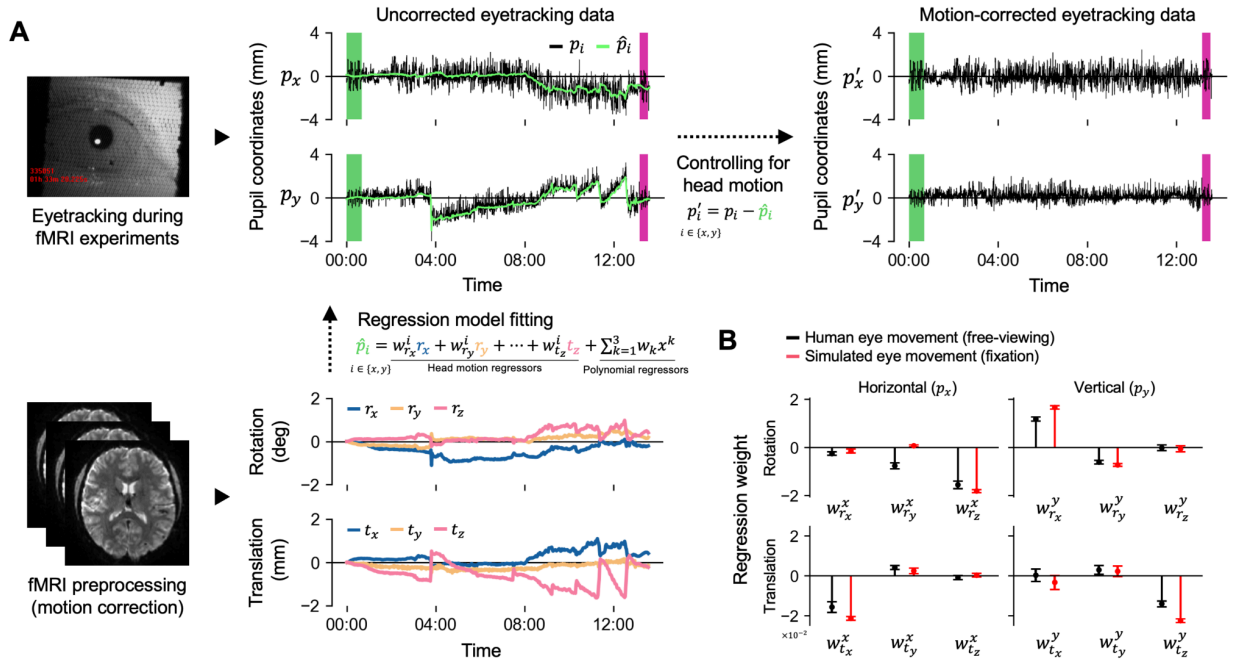

Figure 3. Motion-Corrected Eye Tracking (MoCET) pipeline. (A) MoCET uses head motion parameters to compensate for the effects of head motion on eye tracking data. A combination of head motion parameters and polynomial regressors is used to model the global trend of pupil coordinates ($\hat{p}_i$). After subtracting the fitted global trend, the motion-corrected eye tracking data exhibit a more stable structure while retaining variance (jitter) from actual eye movements. (B) Comparison of regression model weights between human and simulated eye movements. The relative magnitude of regression weights reflects the strength of each head motion parameter's influence on horizontal or vertical pupil coordinates. For instance, rotational motion along the x-axis (i.e., pitching motion) has a strong positive effect on vertical pupil coordinates ($w^y_{r_x}$) in both human and simulated eye movements. Error bars represent ± 1 s.e.m. across participants.

MoCET uses 6-DoF head motion parameters from fMRI data and polynomial regressors (up to the cubic order) to remove noise in eye tracking data. The polynomial regressors address very low-frequency drift not fully captured by head motion parameters, such as instrumental shifts and vibrations in the MRI environment affecting the stability of the eye tracking camera These drifts, unrelated to head motion, cannot be corrected by head motion parameters alone. By using both motion parameters and polynomial regressors, MoCET provides comprehensive correction. Head motion parameters effectively handle dynamic changes associated with head movements (high-frequency), while polynomial regressors address low-frequency drift from other sources unrelated to direct head motion, ensuring both dynamic and gradual noise are corrected.

To assess the validity of using linear regression to remove head motion-related noise, we compared the regression weights of six head motion parameters between human data and simulation containing only motion-related noise. The results demonstrate that MoCET effectively reduced gaze inaccuracy to a Euclidean error of 0.41 degrees (Figure 4A). In both human and simulated data, horizontal head motions, such as translational movement along the x-axis (left-to-right) and rotational movement around the z-axis (rolling), primarily affected global drift in the x-axis of pupil coordinates, while vertical head motions, including translational movement along the y-axis (up-and-down) and rotational movement around the x-axis (pitching), affected vertical drift in the y-axis (Figure 3B). These results indicate that the overall trends in regression weights across the six head motion parameters were consistent between human and simulation data. However, y-axis rotation (yawing) showed stronger effects in x-direction drift in human eye movement than in simulation, likely because the human brain sits above the eyes, unlike our model where the yawing center of the head was close to the eye position, allowing yawing motion to have a greater influence on horizontal eye position. Overall, the regression-based approach used in MoCET successfully removed effects of physical head motion in both human and simulated eye tracking data.

MoCET enables high-precision eye tracking during free-viewing experiments

Global drift in eye tracking data during fMRI experiments is a well-known issue and is commonly addressed using heuristic detrending methods such as linear[33,34] or quadratic polynomial detrending[21]. These methods assume that the average gaze position remains centered on the screen and drift occurs gradually rather than abruptly. A limitation of these methods is that they cannot correct transient effects such as those driven by abrupt head movements. Such head movements are especially pronounced in naturalistic tasks (e.g. free-viewing and video-game playing).

We compared the effectiveness of MoCET to conventional linear and polynomial detrending methods, as well as uncorrected data. Polynomial detrending up to the cubic order served as a baseline for comparison, while MoCET included additional 6-DoF head motion parameters. To evaluate drift correction performance, we employed three distinct metrics. First, we measured gaze accuracy during the validation stage, which evaluates eye tracking data by

requiring participants to fixate on specific points across the screen (Figure 4A). However, this metric cannot fully assess the quality of eye movements during the actual experiment and requires an additional validation stage.Thus, we introduced two additional performance metrics that do not require a validation stage and are also suitable for free-viewing experiments: the accuracy of eye tracking data in predicting participant behavior during the task (Figure 4B) and the accuracy of mapping visual stimuli to retinotopic representations in the visual cortex (Figure 4C and 4D). The latter constitutes a stringent test of eye tracking accuracy, since retinotopic representations are expressed relative to the exact gaze location.

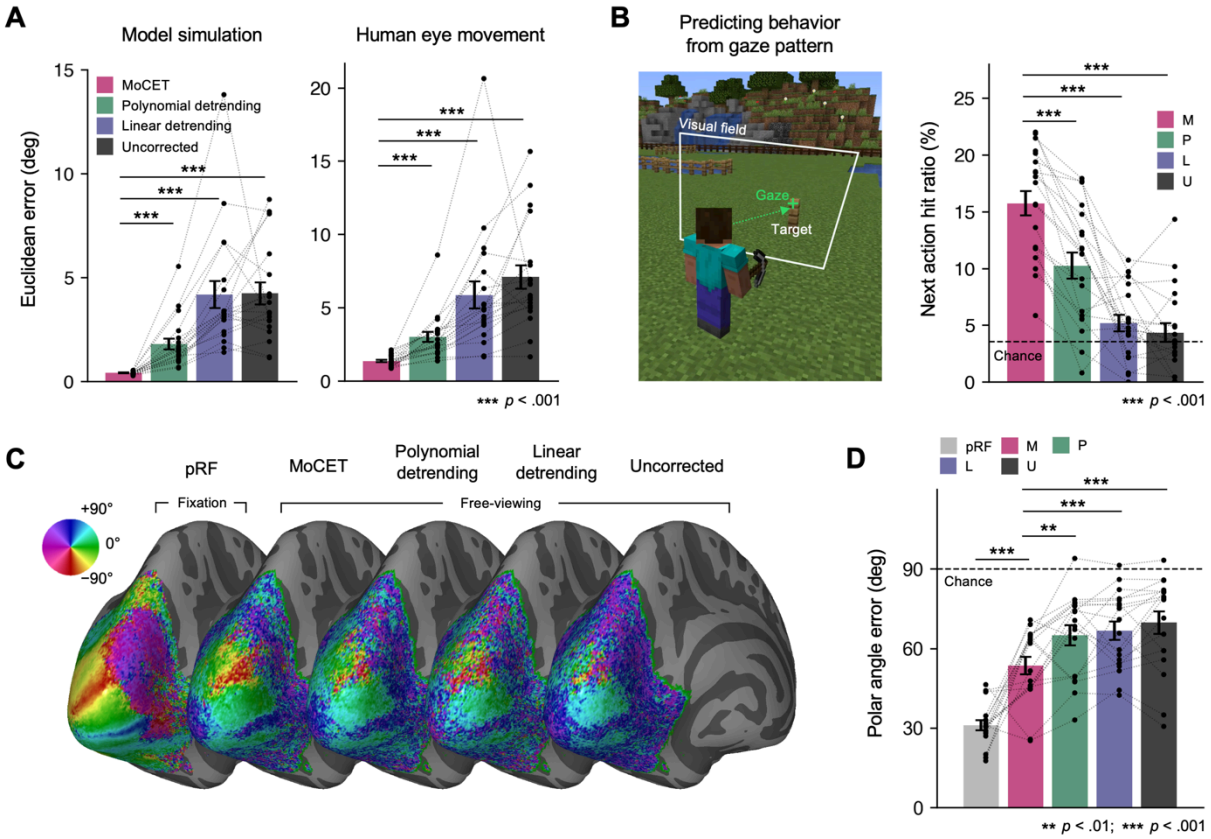

Figure 4. Comparison of drift correction performance between MoCET and traditional heuristic detrending methods. (A) MoCET significantly outperformed traditional detrending methods in reducing eye tracking error during the validation stage, as demonstrated in both model simulation data (Left) and human eye tracking data (Right). (B) MoCET showed superior performance in predicting subsequent participant behavior based on gaze patterns during the task. Specifically, while participants sequentially built or removed blocks in a Minecraft environment, the "next action hit ratio" was measured as the percentage of time participants gazed at the next target location before performing an action at the target. Chance level was calculated as the percentage of random gaze locations on the screen coinciding with the next target location. (C) Group-averaged polar angle representations in visual cortex ($N = 18$). Using population receptive field (pRF) modeling, polar angle representations were estimated based on data from the pRF experiment with retinotopic stimuli (pRF) and data from the free-viewing Minecraft experiment. For the latter, to compensate for eye movements, gaze-centered stimuli were used for pRF modeling. While the conventional pRF experiment, in which visual stimuli stimulate specific retinotopic locations during central fixation, provided the

clearest polar angle maps, MoCET exhibited the best performance in deriving retinotopic maps from free-viewing data. (D) Polar angle accuracy was assessed by calculating the error between predicted polar angles from pRF modeling and canonical polar angles. Ten cortical landmarks with representative polar angles—such as the calcarine sulcus and V1-V2 and V2-V3 boundaries (dorsal and ventral) in both hemispheres—were used for this analysis. MoCET applied to free-viewing experiments showed significantly better performance in mapping retinotopic visual fields compared to other methods. Chance level was determined as the average error from randomly selected angles (0–360°) compared to the target angle. Error bars indicate $\pm$ 1 s.e.m. across participants. Note that pRF analyses were conducted only in 18 subjects who participated in additional pRF experiments. In panels A, B, and D, each dot represents one subject.

For both model simulation and human eye movement data, MoCET outperformed other methods in reducing gaze inaccuracy during the validation stage (Figure 4A, All $ps < 0.001$). To ensure the observed improvements were not simply due to MoCET's use of a total of nine regressors compared to polynomial detrending's three, we included additional polynomial detrending regressors up to the 12th order. Results showed that, for both model simulation and human eye tracking data, the performance of polynomial detrending saturated after the 5th order. MoCET demonstrated superior effectiveness in drift correction, even when using fewer regressors compared to 12th-order polynomial detrending (Supplementary figure 3, Model simulation: All $ps < 0.05$, Human eye movement, All $ps < 0.05$ except for up to the 10th-order polynomial, $p = 0.053$).

To evaluate drift correction performance throughout the entire fMRI experiment, we assessed the ability of gaze data to predict participant behavior during the task. In the Minecraft-based video game task, participants were required to build or remove fence blocks inside a square arena. Previous research has shown that humans tend to fixate on the next target location before performing an action toward it[1,35–38]. Therefore, we hypothesized that motion-corrected gaze data should be able to predict participants' next building action or target-removal action before the action occurs. We calculated the average hit ratio over the 10-second period leading up to each action. A "hit" was defined as the participant's gaze falling within 2 degrees of visual angle from the next target location. While both linearly detrended and uncorrected gaze data performed at near chance levels, MoCET demonstrated significantly better prediction performance compared to all other detrending methods (Figure 2B; all $ps < 0.001$). These results suggest that the noise correction achieved by MoCET effectively removes non-gaze-related artifacts, such as head motion or instrumental instability, while preserving actual eye movements throughout the task.

As an additional performance metric, we evaluated the ability of gaze data to enable the derivation of retinotopy during free-viewing conditions. Retinotopic mapping in the visual cortex

refers to the spatial organization of neural representations of the visual field. Each region of the visual field is systematically represented in early visual areas, such as V1, V2, and V3, with maps that are defined relative to the gaze location[39,40]. This retinotopic organization provides a critical link between eye movements and neural activity, and deriving retinotopic organization during free-viewing requires high-accuracy gaze data. To transform free-viewing stimuli into gaze-centered stimuli, we corrected eye tracking data using various drift correction methods (MoCET, polynomial, linear detrending, and uncorrected data), and then calculated the spatiotemporal local contrast of the stimulus within small patches defined relative to gaze location[41,42]. Using this gaze-centered stimulus preparation, we fit a population receptive field (pRF) model[43,44] to recover spatial tuning (i.e., polar angle and eccentricity) in the early visual cortex (for details, see Supplementary Figure 4 and Methods). Results from free-viewing stimuli were compared to spatial tuning obtained from conventional pRF experiments using structured visual stimuli (wedges, rings, and bars) presented during central fixation.

While pRF experiments with structured stimuli provided the clearest polar angle maps, MoCET demonstrated the ability to accurately map free-viewing stimuli to brain activity and outperformed other detrending methods in recovering polar angles (all $ps < 0.01$). Although the video game stimuli showed limitations in mapping foveal representations, they more effectively mapped peripheral visual areas compared to the structured pRF experiment, even along the vertical meridian, where stimulus coverage was comparable to pRF stimuli. This advantage stems from free-viewing experiments naturally engaging the peripheral visual field, whereas structured pRF experiments conducted under central fixation are typically limited in their stimulation extent. Detailed retinotopic representations, including eccentricity maps, are presented in Supplementary Figure 5.

In summary, our results demonstrate that MoCET significantly improves eye tracking accuracy during free-viewing experiments compared to traditional detrending methods. By combining head motion parameters and polynomial regressors, MoCET not only reduces drift-related errors but also enhances the predictive power of gaze data for behavior and neural response characterization. MoCET consistently outperformed other methods in gaze accuracy during validation stages, behavioral prediction during dynamic tasks, and retinotopic mapping of free-viewing visual stimuli, particularly in peripheral visual areas. These findings establish

MoCET as a robust and versatile method for correcting eye tracking data, enabling more precise studies of behavioral and neural processes in naturalistic contexts.

Comparison of camera-based and MR-based eye tracking

Magnetic resonance (MR)-based eye tracking has been explored for decades as an alternative to traditional camera-based systems[17,22–24]. This approach eliminates the need for additional eye tracking hardware, as it relies solely on the fMRI data to infer gaze. Recent advancements, such as DeepMReye[21], have advanced MR-based eye tracking by leveraging deep neural networks to decode gaze coordinates directly from MR signals with promising levels of accuracy. MR-based eye tracking not only simplifies experimental setups but also enables retrospective analysis of preexisting datasets, broadening its application across studies[21,33].

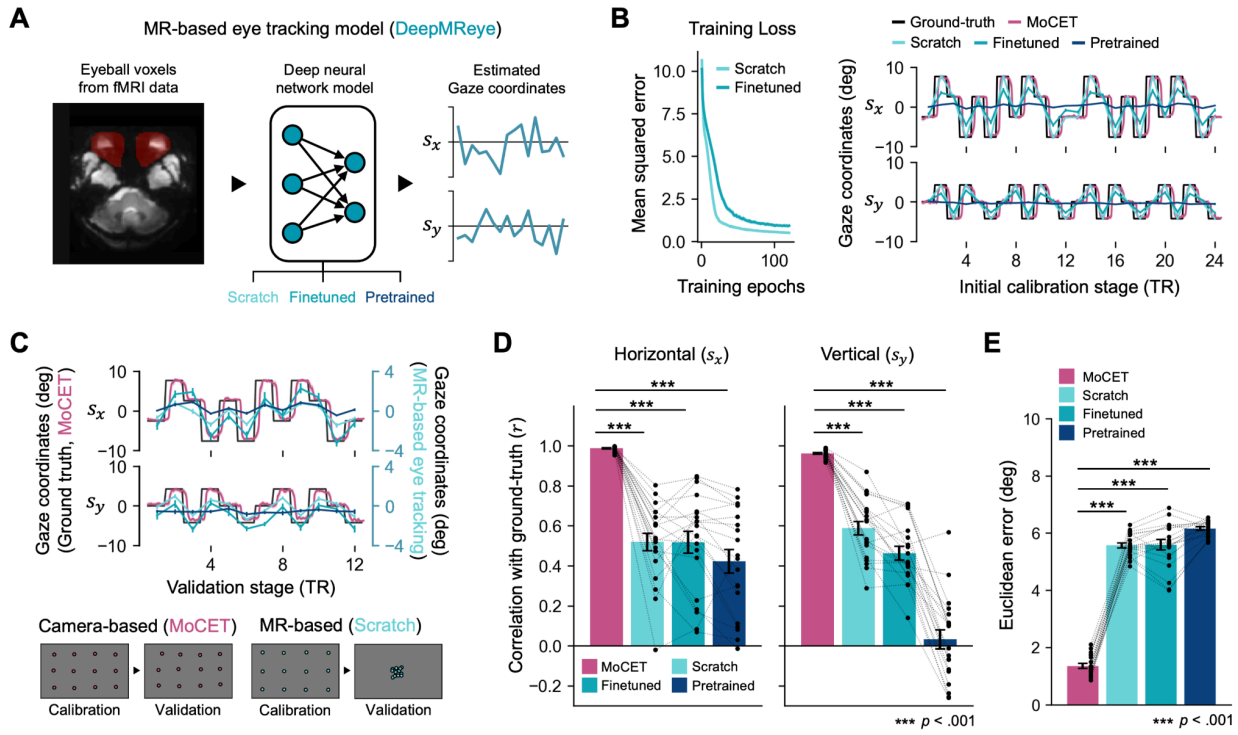

Figure 5. Comparison of camera-based and MR-based eye tracking methods. (A) DeepMReye[21] leverages deep neural networks to estimate gaze coordinates. Three variations of model weights were applied for subsequent analyses: 1) the original pre-trained DeepMReye model weights (*Pretrained*), 2) finetuned pre-trained weights using our dataset (*Finetuned*), and 3) model weights trained from scratch exclusively on our dataset (*Scratch*). (B) Both the Scratch and Finetuned models achieved low training losses (Left) and successfully recovered gaze coordinates during the calibration stage used for model training (Right). In contrast, the Pretrained model failed to recover accurate gaze coordinates. Notably, the camera-based eye tracking data enhanced with MoCET exhibited a clear delay relative to the onset of the calibration dots (ground truth) due to the reaction time of human

participants. (C) During the validation stage, all three MR-based eye tracking models followed the general trend of the ground truth gaze coordinates, but their magnitudes were substantially reduced. While camera-based eye tracking (MoCET) maintained spatially distributed gaze patterns close to the actual location of the dots, MR-based models produced gaze location estimates that clustered toward the center of the screen. Note that for visibility, different axis scales are used for the MR-based models. (D) For both horizontal and vertical gaze coordinates, all three MR-based models showed trends in recovering gaze direction during the validation stage. However, due to the limited vertical field of view of the screen, the *Pretrained* model showed markedly reduced performance in the vertical direction. (E) Despite significant trends in gaze direction recovery, all three MR-based models showed limitations in mapping precise gaze locations, resulting in higher Euclidean errors compared to camera-based eye tracking. (C-E) Error bars represent $\pm$ 1 s.e.m. across participants.

We compared the performance of camera-based and MR-based eye tracking methods. For the MR-based approach, we examined three DeepMReye models using (1) the original pretrained model weights (*Pretrained*), (2) pretrained weights finetuned with our dataset (*Finetuned*), and (3) weights trained from scratch exclusively on our dataset (*Scratch*). Both the Finetuned and Scratch models were trained on 24 calibration points during the calibration stage (Figure 5A). While the *Scratch* and *Finetuned* models demonstrated low training losses and effectively captured gaze coordinates during calibration, the Pretrained model struggled to fit the training data, showing near-zero responses (Figure 5B, Euclidean error: *Scratch*: 0.48 deg; *Finetuned*: 3.87 deg; *Pretrained*: 6.22 deg).

Although the re-trained MR-based models performed well on the calibration data, they exhibited limitations when applied to the rest of the experiments. Specifically, while the DeepMReye models successfully recovered general gaze directions (e.g., looking up, down, left, or right), demonstrating their ability to capture trends in gaze movement derived from fMRI data (Figure 5C, 5D; Pearson's correlation of horizontal direction: *Scratch*: $r = 0.519$, *Finetuned*: $r = 0.518$, *Pretrained*: $r = 0.422$; vertical direction: *Scratch*: $r = 0.588$, *Finetuned*: $r = 0.463$, *Pretrained*: $r = 0.031$), they showed reduced precision in mapping gaze locations, often predicting positions that deviated significantly from the ground truth (Figure 5C, 5E; Euclidean error: *Scratch*: 5.57 deg, *Finetuned*: 5.59 deg, *Pretrained*: 6.16 deg).

In contrast, the camera-based eye tracking method, enhanced with MoCET, consistently outperformed MR-based methods, maintaining both precise gaze direction and spatial accuracy of gaze coordinates across the entire experiment ($ps < 0.001$ for all comparisons). We observed that MR-based models tended to predict gaze positions near the screen's center, suggesting limited performance when generalizing to novel data. This pattern indicates that the deep neural network models had difficulty accurately mapping gaze locations beyond the

distribution of their training data, leading to small-magnitude or near-zero responses when encountering unfamiliar fMRI data. The combination of high gaze direction accuracy but low spatial precision observed in our study is consistent with findings from previous studies utilizing MR-based eye tracking models[21,33]. These results suggest that while MR-based models hold promise for investigating gaze patterns related to cognitive processes in fMRI studies, they are currently less suitable for applications requiring eye tracking with high spatial and temporal precision.

Discussion

This study highlights the critical impact of head motion on gaze accuracy in fMRI experiments and introduces Motion-Corrected Eye Tracking (MoCET) as a practical and effective solution. Our computational simulation confirmed that head motion systematically generates substantial drift in eye tracking data even without actual eye movements. By using six degrees of freedom (6 DoF) head motion parameters and polynomial regressors, MoCET demonstrated its ability to robustly correct motion-induced errors in eye tracking data, significantly outperforming traditional detrending methods.

MoCET proved highly effective in fMRI experiments requiring high-precision eye tracking data. By enabling reliable prediction of participants' future actions based on eye tracking data, MoCET demonstrates its potential for studying complex, visually guided behaviors in fMRI experiments. Furthermore, MoCET enabled accurate retinotopic mapping of free-viewing stimuli, excelling in peripheral visual areas where conventional pRF experiments often face limitations. These findings establish MoCET's broad applicability in visual neuroscience research, particularly for naturalistic, free-viewing paradigms.

We compared MoCET with the recently introduced MR-based eye tracking method DeepMReye[21], which leverages deep neural networks to estimate gaze coordinates. While DeepMREye demonstrated the ability to capture gaze direction, it exhibited significant limitations in spatial precision and often exhibited a center bias. In contrast, motion-corrected camera-based eye tracking (MoCET) consistently maintained precise spatial accuracy, making it

particularly well-suited for free-viewing fMRI experiments and tasks demanding high-precision gaze data.

Although MR-based eye tracking offers unique advantages, such as not requiring additional hardware, camera-based systems retain critical benefits in terms of precision and resolution. One interesting question is what specific factors might have contributed to the poor performance of MR-based eye tracking in our data. A key difference lies in the timing of fixation presentation, which was used to train and evaluate model performance. In the original DeepMReye study, each fixation point remained on the screen for 4 seconds, allowing five fMRI volumes per fixation (TR: 800 ms). This prolonged duration helped stabilize eyeball images, facilitating more accurate gaze predictions. In contrast, our calibration presented each fixation point for only one TR, meaning that by the time a volume was acquired, the participant's eyes may have already shifted toward the next target. This likely contributed to inaccurate gaze estimates, with a bias toward the center of the screen.

Additionally, the DeepMReye authors note potential issues with Posterior-Anterior (PA) phase encoding (see https://github.com/DeepMReye/DeepMReye/wiki), where eyeball images can be compressed, making them difficult to distinguish. While ultra-high magnetic field (7T) typically increases spatial distortion in fMRI images, visual inspection confirmed the robust quality of eyeball shapes in our preprocessed functional images. Future studies should further explore specific conditions that impact the performance of MR-based eye tracking.

MoCET's head motion correction without additional hardware particularly benefits fMRI research, where space and setup constraints often prevent installing motion sensors. The ability to be applied retrospectively to existing datasets further extends MoCET's utility, providing researchers with a versatile tool to enhance visual neuroscience research. By integrating MoCET into naturalistic experimental paradigms, it becomes possible to align neural and behavioral data more effectively, advancing our understanding of ecologically valid human cognition.

One potential concern in applying MoCET is the discrepancy between the sampling rates of eye tracking data and fMRI-derived head motion estimates. While eye tracking data have high temporal resolution, head motion parameters from fMRI preprocessing are limited by the lower sampling rate of fMRI (typically 1–2 seconds per volume). The slower sampling of head motion ensures that rapid eye movements, such as saccades, are not mistakenly regressed out as head

motion artifacts. At the same time, lower temporal resolution may raise concerns about whether head motion effects are fully captured and corrected. We note that head motion in fMRI is primarily characterized by low-frequency components, with most power below 0.1 Hz due to small voluntary and involuntary non-periodic, transient movements. Respiration-related motion introduces additional fluctuations around 0.2–0.3 Hz[45]. Thus, apparent pupil position shifts from both transient shifts in head position and respiration-related periodic head motion can be effectively corrected using fMRI-derived head motion estimates.

Despite its strengths, MoCET has some limitations that should be acknowledged. MoCET assumes linear relationships between head motion and gaze noise, which may become inadequate in cases of extreme head movement or when the eye tracking camera is installed at a steep tilt relative to the gaze direction. In such conditions, the nonlinear distortion of pupil coordinates can introduce errors that are not fully accounted for by linear regression. However, in typical fMRI setups, including 7T MRI scanners where the camera is slightly tilted due to spatial constraints, the linear assumption remains effective for correcting the effect of head motion on eye tracking data. This is partly because large head movements are physically restricted by stabilization devices such as foam padding, minimizing extreme deviations that could introduce nonlinear distortions.

Furthermore, while its retrospective nature is beneficial for existing data, it limits its use in real-time closed-loop experiments requiring gaze information[46–49]. In future work, MoCET could be adapted for real-time fMRI experiments that require precise gaze information. By directly integrating head motion parameters extracted in real time from MR image reconstruction computers, MoCET could dynamically correct gaze coordinates during ongoing experiments. This would extend its application to real-time paradigms, such as neurofeedback or task-adaptive experimental designs, where immediate gaze accuracy is critical.

Finally, a notable contribution of this work is that we release a high-quality eye tracking dataset with multiple calibration periods contained within the dataset itself. Thus, we believe this dataset could serve as a useful benchmark for developing and evaluating new eye tracking methods for fMRI experiments. The high-precision camera-based eye tracking data can also provide a reliable reference for improving MR-based eye tracking approaches. Together, we hope that MoCET and the dataset will enhance the integration of eye tracking with neuroimaging,

providing crucial support for cognitive neuroscience research that requires precise gaze measurements.

Methods

Participants

Twenty-one participants (7 females, mean age 24.18 ± 3.15 years) were initially recruited for the study. All participants provided written informed consent, with the study protocol approved by the Institutional Review Board of Sungkyunkwan University, and were monetarily compensated. Thirteen participants completed two fMRI sessions (12 runs), while the remaining eight completed a single session (6 runs). Eye tracking data from 6 to 12 runs per participant were evaluated based on specific validation criteria (see Methods - Eye tracking accuracy). One participant was excluded from the analysis as none of their runs met these criteria. As a result, data from twenty participants (average number of valid runs: 5.55 ± 2.87) were included in the final analysis. Among these twenty participants, eighteen participants also underwent an additional fMRI session involving pRF experiments, which served as a reference for validating the retinotopic mapping results.

Camera-based eye tracking system

An endoscopic eye tracking camera and illuminator from Avotec Incorporated were installed in the 7T MRI scanner, positioned in front of the participant's eye, recording gaze position at a frequency of 60 Hz. Post hoc image processing was applied to reduce high-frequency spatial noise in the eye video and to binarize the image for improved pupil detection. For accurate pupil detection and tracking, we implemented the Pupil Reconstructor with Subsequent Tracking (*PuReST*) algorithm[50]. The resulting pupil coordinates were then preprocessed to exclude low-confidence data caused by blinks, defined as frames with a pupil confidence score below 0.75. Additionally, we removed abnormal spikes in the pupil coordinates that exceeded three standard deviations from the local mean, which typically occurred when the

algorithm mistakenly detected non-pupil dark, roundish areas during eye closure. The identical pupil detection and tracking procedure was applied to simulated eye tracking data.

Experimental procedure

For the main analysis, we used eye tracking data collected from participants as they played a 13-minute 3D Minecraft[25]-based video game task in a 7T MRI scanner. Each participant completed one or two sessions, with each session consisting of six runs in which participants played the video game on different maps. At the beginning and end of each run, participants underwent eye tracking calibration, in which they fixated on green dots sequentially displayed on the screen. The initial calibration involved two repetitions of a 12-dot sequence, resulting in participants fixating on a total of 24 dots. At the end of the video game, a validation stage required participants to fixate again on the same 12 dots. During the game, participant gaze was unrestricted, allowing for natural, free-viewing behavior. The video game task included a three-minute period during which participants built new blocks and removed existing blocks in a 3D environment. Eye tracking data from this period was used to predict subsequent building or breaking actions (See Methods – Predicting behaviors from gaze patterns for more details).

For the retinotopic mapping analysis, eighteen participants completed an additional fMRI session of pRF experiments. During the experiment, participants maintained fixation on a central dot and pressed a button whenever they detected a change in the dot's color. The visual stimuli consisted of colorful cartoon images (toonotopy) that were viewed through moving wedge, ring, and bar apertures, designed to stimulate localized regions of the visual field[51,52]. The wedge runs used rotating wedges (counterclockwise or clockwise) to stimulate specific angular regions of the visual field, while the ring runs involved expanding or contracting concentric rings to map eccentricity. During the bar runs, bars with varying orientations (e.g., horizontal, vertical, or oblique) swept across the visual field in multiple directions. The pRF session comprised six runs, with a total duration of approximately 38 minutes.

fMRI data acquisition and preprocessing

Functional neuroimaging data were collected using a 7T Siemens MAGNETOM Terra MRI scanner equipped with a 32-channel Nova head coil, located at Sungkyunkwan University and the Institute for Basic Science, Center for Neuroscience Imaging Research. Blood oxygenation level-dependent (BOLD) contrast was measured using T2*-weighted functional images obtained with a dual-polarity GRAPPA (DPG) sequence (voxel size: 1.5 mm isotropic; TR: 1600 ms; TE: 21 ms; FOV: 210 × 210 mm; 96 slices covering the whole brain; flip angle: 50°). High-resolution anatomical images were acquired using an MP2RAGE sequence (voxel size: 0.7 mm isotropic; TR: 5888 ms; TE: 2.44 ms; FOV: 224 × 224 mm; 320 slices; flip angle 1: 4°; flip angle 2: 5°).

Preprocessing of both functional and anatomical data was conducted using fMRIprep[53]. Since the default skull-stripping for MP2RAGE data in fMRIprep was suboptimal, the skull was manually removed prior to preprocessing. Anatomical data were then processed for intensity non-uniformity correction, brain segmentation, and surface reconstruction. Functional data underwent motion correction and were aligned to the MNI152 standard space for analysis with DeepMReye. For voxel-wise pRF modeling, functional data were maintained in their native space to preserve precise spatial resolution. No additional preprocessing steps were applied.

Eye tracking analysis

Eye tracking model calibration

During the calibration period, we collected pupil coordinates at 24 points (4×3 grid, repeating each point twice) presented on the screen. Each calibration point appeared for 1.6 seconds as a green square with a central red dot. The square expanded to its largest size at 0.8 seconds and then began to shrink, creating a clear visual anchor for participants. To consider the saccadic delay and maximize calibration precision, we focused on a 0.5-second period during the middle of each presentation (when the square was largest) and averaged the x and y coordinates of the pupil within this time window. This resulted in a set of 24 averaged pupil coordinates paired with known screen coordinates for training the eye tracking model.

To predict gaze location from pupil coordinates, we employed a radial basis function (RBF) interpolation method[26–28], which is widely used for mapping spatial coordinates. RBF interpolation is particularly effective for non-linear mappings when the relationship between input and output coordinates may not follow a simple linear path, such as mapping recorded pupil coordinates from an oblique field-of-view eye tracking camera to flattened screen coordinates. RBF interpolation calculates interpolated values as a weighted average of distances between known calibration points and input pupil coordinate, providing a smooth and continuous mapping from pupil coordinates to gaze location on the screen. The calibrated RBF model was applied to the entire set of pupil coordinates collected during the experiment, converting each pupil position to a predicted gaze location on the screen.

Eye tracking accuracy

To assess the accuracy of the eye tracking data, we calculated the Euclidean distance between the predicted gaze location (averaged within each 0.5-second window) and the ground truth, defined as the actual location of each calibration point on the screen. Since gaze location was averaged after applying the interpolation model, calibration error could exceed zero due to the inherent variability in gaze position within each 0.5-second calibration window.

We verified that participants' eye movements were sufficiently accurate during both calibration and validation stages, using separate models trained on the data for each stage. Only eye tracking data achieving a gaze error of less than 1.0 visual degrees from calibration points in both calibration and validation phases were included in subsequent analyses. This criterion allowed us to exclude errors stemming from fixation instability or individual gaze variability, thereby ensuring that any remaining inaccuracies were not simply due to inconsistent eye movements.

To examine the relationship between head motion and eye tracking error, we estimated the magnitude of head shift indirectly, as variations in the camera-to-eye distance across participants prevented direct measurement of actual distances from the camera. To approximate head shift, we assumed an average pupil diameter of 5 mm, based on the typical pupil size under the dim lighting conditions of the MRI scanner[54]. The pupil size measured in pixels using the pupil detection algorithm allowed us to calculate a scaling factor in millimeters per pixel. The estimated head shift (in mm) during the experiment was then derived by multiplying this scaling

factor by the average change in pupil coordinates (in pixels) from the initial calibration to the validation stage. This method provided a relative measure of head shift, which was used to analyze its relationship with eye tracking error.

Motion-Corrected Eye tracking (MoCET)

To compensate for head motion effects in eye tracking data, we used participants' head motion parameters obtained from fMRI preprocessing. These parameters consist of six degrees of freedom (6 DoF) motions, capturing translations along the x, y, and z axes as well as rotations around these axes. These parameters, used during the motion correction step, reflect the relative displacement of the head from a reference template image. To ensure that the head motion parameters calculated from the BOLD reference image estimated during fMRIprep preprocessing, aligned with the initial eye tracking calibration at the beginning of the experiment, we re-centered them by subtracting the head motion values of the first volume from all subsequent volumes. This adjustment yielded head motion displacements relative to the initial calibration position, which were then used to correct for head motion-induced noise in the eye tracking data and to generate simulated eye movement under head motion in our computational model simulation.

To mitigate head motion-related noise in the eye tracking data, we applied a signal denoising method based on linear regression. For the main analysis, our regressors included the six head motion parameters as well as polynomial terms up to the third degree (cubic). The head motion parameters were upsampled using linear interpolation to match the temporal resolution of the eye tracking data at 60 Hz. Using these combined regressors, we constructed a linear regression model to predict the x and y pupil coordinates. The model predictions were then subtracted from the original pupil coordinates, resulting in motion-corrected pupil coordinates. We applied polynomial detrending using a similar approach to compare eye tracking accuracy across different detrending methods.

Computational simulation of eye movement

Anthropometric parameters of model simulation

Anthropometric measurements of adult human heads adopted from previous studies[29–31] were slightly adjusted to create an isotropic, spherical head model for our simulation. For example, three head dimensions—head breadth (ear-to-ear distance), horizontal depth (back of head to nose), and vertical length (chin to top of head)—were averaged to define a single parameter representing the diameter of the model's head sphere. The model's eyeballs are physically attached to this head sphere, such that head movement affects the 3D geometric position of the eyeballs, while each eyeball can rotate independently to simulate gaze toward target locations. As the pupil is attached to the eyeball, its 3D coordinates change in accordance with the eyeball's rotation, maintaining alignment along the direct line from the eyeball center to the target gaze location. Supplementary Figure 1 provides detailed anthropometric parameters of the head and eyeball used in the computational model.

Optimization of eye tracking camera parameters

A virtual eye tracking camera was positioned in front of the model's left eye to capture simulated eye movements. Due to variations in participant head size, eye position, and setup alignment, the direction of the actual camera installed within an MRI scanner may not be completely perpendicular to the eyeball, often resulting in slight tilts. To replicate realistic eye tracking as closely as possible, we customized the virtual camera's direction, roll angle, and distance from the model's eyeball, calibrating these parameters to match participant-specific eye tracking data. This was achieved by simulating a range of configurations for four camera parameters. For the camera's orientation, the baseline direction was set to face the model's eyeball directly; yaw and pitch angles were then varied by up to ± 5 degrees along both horizontal and vertical axes. The camera's roll angle was parameterized within a range of -90 to $+90$ degrees. Finally, the camera-to-eyeball distance was adjusted from 4.5 cm to 15 cm to account for natural variations in participant head and eye positioning relative to the fixed camera position in the head coil. This setup created approximately 36,000 unique combinations of four camera parameters (yaw, pitch, roll, and distance).

To determine the optimal camera parameters for each participant, we simulated the model's gaze directed toward 12 calibration points on a screen, producing twelve 2D pupil locations as captured by the virtual eye tracking camera. These simulated pupil locations were then compared to the participant's actual pupil coordinates recorded for the same 12 calibration points. For each of the 36,000 parameter combinations, we calculated the discrepancy between the model's and participant's pupil locations, and selected the optimal parameters for each participant through a grid search that minimized this difference. The optimized camera parameters averaged across participants and examples of parameters for individual participants are shown in Supplementary Figures 1 and 2.

Incorporating head motion parameters

At each step of the simulation, we constructed a 3D rotational transformation matrix and a 3D translational transformation matrix from head motion parameters. These matrices were then applied to the coordinates of the model's head and attached eyeball, enabling our simulation to precisely reproduce participants' head motions in the simulation environment.

Analysis of simulated gaze

The model's eye movement was recorded as a sequence of images using a virtual eye tracking camera. Ray casting was employed to generate 2D projections of 3D objects (head, eyeball, and pupil) by tracing rays from the virtual camera until they intersected with object surfaces, determining their visible positions[55]. For visualization (e.g., Figure 2 and Supplementary Figure 1), a light source and the Blinn-Phong reflection model[56,57] were used to render realistic colors and shadows. However, for efficient simulation and accurate pupil tracking, rendering was simplified by assigning fixed colors to objects without light reflection, ensuring only the closest intersecting object was represented. This optimized both simulation speed and accuracy while preserving essential pupil location details.

Simulated eye tracking data were preprocessed and analyzed using the same procedure as the human eye tracking data. To examine the relationship between the magnitude of head motion and eye tracking error, we calculated the extent of head shift occurring within the p_x–p_y plane, from initial calibration to final validation. While the magnitude of head shift in human eye

tracking data was estimated relative to the assumed pupil size, the model simulation allowed for the precise measurement of the actual physical distance of head shift in 3D space.

Permuted random head motion simulation

To assess the statistical significance of the relationship between head motion and pupil coordinates, we generated permuted random head motion parameters. The randomization was performed using a phase-shuffling approach, which preserves the power spectrum and temporal autocorrelation of the original head motion time series. Specifically, the Fourier transform of each head motion parameter was computed, and the phase components were randomly shuffled while keeping the amplitude spectrum intact. The inverse Fourier transform was then applied to reconstruct the randomized head motion time series.

This procedure ensures that the randomized head motion retains the same temporal structure as the original data but removes any relationships with the actual experimental conditions. We applied this method independently to each of the six head motion parameters. These randomized head motion parameters were then used to simulate pupil movements in the geometry-based eyeball model, producing a null distribution of gaze similarity metrics. Statistical significance was determined by comparing the actual simulation results with the null distribution obtained from 100 permutations for each dataset. To evaluate overall significance across the dataset, we aggregated individual p-values using Fisher's method[32], providing a robust summary statistic for group-level analyses.

Predicting behaviors from gaze patterns

To evaluate the predictive power of drift-corrected gaze data for participant behavior, we analyzed data from the Minecraft-based video game task. In this task, participants sequentially built new fences or removed existing fences within a square arena to create a shortcut from their starting locations to a designated destination. Each participant's decisions and actions were logged in-game, allowing us to identify their next target block (fence) based on their strategy. We used this behavioral dataset to compare the predictive performance of eye tracking data corrected using MoCET, linear detrending, polynomial detrending, and uncorrected data.

The in-game log file, recorded at 20 frames per second, included the 3D physical coordinates of each target block (fence) in the Minecraft world. To determine the target's corresponding location on the screen, we used the participant's in-game camera position, including their locations, yaw (horizontal rotation), and pitch (vertical rotation). These parameters were used to project the 3D world coordinates of the next target block onto 2D screen coordinates relative to the participant's field of view (FOV). This process was repeated for each frame to account for participant movement or camera adjustments during the task.

For each moment, the next target block was identified based on the in-game log, and the distance between the participant's gaze and the 2D screen coordinates of the target was calculated for a 10-second period preceding the action (building or removing a block). A "hit" was defined as a gaze position falling within a 2-degree visual angle from the target's location on the screen. The hit ratio of each action was then calculated as the percentage of frames within the 10-second window (i.e., 200 frames) in which a hit occurred. Chance-level performance was estimated based on the probability of a random gaze location falling within a 2-degree visual angle around the target. Specifically, the circular area corresponding to a 2-degree visual angle (πr^2) was divided by the total screen area, assuming a uniform distribution of gaze locations across the screen. This provided a baseline probability of a random gaze intersecting the target. The overall hit ratio for each participant was averaged across all actions. Statistical comparisons of hit ratios between the different eye tracking data correction methods were performed using paired t-tests.

Visual field mapping from free-viewing visual stimuli

To evaluate the accuracy of retinotopic mapping using free-viewing visual stimuli, we used drift-corrected gaze data for population receptive field (pRF) modeling[43,44]. The free-viewing stimuli consisted of the Minecraft video game display as participants played the task. Gaze data corrected using MoCET, polynomial detrending, linear detrending, and uncorrected data were used to generate gaze-centered stimulus representations and assess their impact on retinotopic mapping in the visual cortex.

Stimulus representation and spatiotemporal local contrast

For each frame of the video game display (original resolution: 1600×1000 pixels), we generated gaze-centered spatiotemporal local contrast by extracting stimulus patches relative to the participant's gaze. To account for potential gaze coordinates outside the visible screen, the original frame was padded with gray borders (3200 pixels on each side), creating an expanded frame of 4800×4200 pixels. Using the gaze coordinates, a 3200×3200 pixel patch was cropped from the padded image, ensuring the gaze remained at the center of the extracted stimulus.

The cropped patch was then upsampled to a resolution of 3360×3360 pixels to align with the dimensions required for subsequent analyses (a multiple of 224). To compute spatiotemporal local contrast, each upsampled frame was divided into small 3D patches of 15×15 pixels spatially and 1.6 seconds temporally (48 frames). Standard deviation values were calculated within each 3D patch, resulting in a spatiotemporal local contrast representation with a resolution of 224×224×510. This stimulus energy signal served as the input to the pRF model[42].

pRF model fitting and analysis

The pRF model estimated polar angle representations in the visual cortex, by fitting a Gaussian receptive field centered on the predicted gaze location to voxelwise BOLD responses. The visual cortex mask was extracted from a resting-state functional atlas[58]. Model fitting utilized the spatiotemporal local contrast signal derived from gaze-corrected free-viewing stimuli as the input and voxelwise time-series data as the output. The Gaussian receptive field model predicted neural responses to the spatial location of stimuli, enabling the reconstruction of polar angle maps. Model performance was quantified by calculating angular errors between the estimated polar angles and canonical polar angles for ten cortical landmarks, including the calcarine sulcus and V1-V2 and V2-V3 boundaries (dorsal and ventral) in both hemispheres. Cortical landmarks were identified using retinotopic ROIs from the HCP dataset[52], where key reference vertices (5–10 per landmark) were manually selected along the cortical boundaries. A geodesic pathfinding algorithm was then used to interpolate and connect these reference vertices, forming a continuous representation of each landmark. The extracted geodesic lines were used to derive ground-truth polar angles for evaluating model accuracy.

Training and validation of MR-based eye tracking models

To train and evaluate the MR-based DeepMReye models[21], we first extracted eyeball voxel masks from the fMRI data. The masks were generated using the protocol described in the original DeepMReye reference, and their quality was manually inspected by researchers to ensure proper delineation of the eyeball regions. Three MR-based eye tracking models were compared: the *Pretrained* model, which utilized the original pre-trained DeepMReye weights without modification; the *Finetuned* model, where the pretrained weights were further trained using our dataset to adapt to the current experimental settings; and the *Scratch* model, which was trained from scratch using only the calibration data from this study.

For the initial calibration stage, 24 volumes corresponding to the presentation of sequential calibration points were used to train the *Finetuned* and *Scratch* models. The calibration points appeared one per volume, providing labeled data for training. Most of the model training parameters, including the neural network architecture, loss function weights, batch size, and augmentation settings, were retained from the original configuration. However, specific adjustments were made to the number of training epochs and the learning rate decay schedule to ensure optimal model performance on the calibration data. These adjustments allowed the model to achieve sufficient convergence as assessed by the loss history during training.

After training, the models were applied to predict gaze coordinates across the entire fMRI experiment. The MR-based models produced one predicted gaze coordinate per fMRI volume. Performance validation was conducted for both the calibration and validation stages. During the validation stage, 12 gaze coordinates were predicted for the 12 validation points presented sequentially. Performance metrics, including gaze inaccuracy (Euclidean error) and directional accuracy (correlation with the ground truth), were calculated in the same manner as for camera-based eye tracking.

Data availability

The eye tracking data and head motion parameters are available at
https://doi.org/10.5281/zenodo.14892081

Code availability

The analysis scripts used in this study, along with the MoCET Python package, are publicly available at https://github.com/jwparks/mocet. The package can be installed via PyPI using the command: 'pip install mocet'.

Acknowledgements

We would like to thank Tianjiao Zhang and Jack L. Gallant for their initial development of eye tracking software. PECON lab members for their invaluable feedback on the MoCET python package. Boohee Choi, Kyubeen Ahn, Seoyu Kim, Sunhyun Min for technical support in fMRI and eye tracking data collection. This work was supported by NIH grant (R01EY034118), the Institute for Basic Science Grant (IBS-R015-D2), the National Research Foundation of Korea (RS-2024-00348130), and the Fourth Stage of Brain Korea 21 Project (S-2023-0794-000).

Author contributions

J.P., K.N.K., and W.M.S. conceptualized and designed the research. J.Y.J. developed the eye tracking hardware setup tailored for the 7T MRI scanner. J.P., J.Y.J., R.K., K.N.K., and W.M.S. conducted the research. J.P. developed the simulation model. J.P. analyzed the eye tracking data. J.P. and R.K. analyzed the neuroimaging data. J.P. drafted the initial version of the manuscript, and J.P., J.Y.J., R.K., K.N.K., and W.M.S. collaboratively revised and edited the manuscript.

Declaration of interests

The authors declare no competing interests.

References

1. Hayhoe, M. & Ballard, D. Eye movements in natural behavior. *Trends Cogn. Sci.* **9**, 188–194 (2005).
2. Henderson, J. M. Gaze Control as Prediction. Trends Cogn. Sci. **21**, 15–23 (2017).
3. Henderson, J. M. Human gaze control during real-world scene perception. Trends Cogn. Sci. 7, 498–504 (2003).
4. Tangtartharakul, G., Morgan, C. A., Rushton, S. K. & Schwarzkopf, D. S. Retinotopic connectivity maps of human visual cortex with unconstrained eye movements. *Hum. Brain Mapp.* **44**, 5221–5237 (2023).
5. Bonhage, C. E., Mueller, J. L., Friederici, A. D. & Fiebach, C. J. Combined eye tracking and fMRI reveals neural basis of linguistic predictions during sentence comprehension. *Cortex* **68**, 33–47 (2015).
6. Wagner, I. C. *et al.* Entorhinal grid-like codes and time-locked network dynamics track others navigating through space. *Nat. Commun.* **14**, 231 (2023).
7. Hayhoe, M. M. & Matthis, J. S. Control of gaze in natural environments: effects of rewards and costs, uncertainty and memory in target selection. *Interface Focus* **8**, 20180009 (2018).
8. Lakshminarasimhan, K. J. et al. Tracking the Mind's Eye: Primate Gaze Behavior during Virtual Visuomotor Navigation Reflects Belief Dynamics. *Neuron* **106**, 662-674.e5 (2020).
9. Hanke, M. et al. A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. *Sci. Data* **3**, 160092 (2016).
10. Telesford, Q. K. et al. An open-access dataset of naturalistic viewing using simultaneous EEG-fMRI. Sci. Data 10, 554 (2023).
11. Schoffelen, J.-M. et al. A 204-subject multimodal neuroimaging dataset to study language processing. Sci. Data 6, **17** (2019).
12. Zhang, B., Wang, F., Zhang, Q. & Naya, Y. Distinct networks coupled with parietal cortex for spatial representations inside and outside the visual field. *NeuroImage* **252**, 119041 (2022).
13. Kanowski, M., Rieger, J. W., Noesselt, T., Tempelmann, C. & Hinrichs, H. Endoscopic eye tracking system for fMRI. *J Neurosci Meth* **160**, 10–15 (2007).
14. Gitelman, D. R., Parrish, T. B., LaBar, K. S. & Mesulam, M.-M. Real-Time Monitoring of Eye Movements Using Infrared Video-oculography during Functional Magnetic Resonance Imaging of the Frontal Eye Fields. NeuroImage 11, 58–65 (2000).
15. Morimoto, C. H. & Mimica, M. R. M. Eye gaze tracking techniques for interactive applications. *Comput. Vis. Image Underst.* **98**, 4–24 (2005).
16. Banks, A., Abdelaal, A. E. & Salcudean, S. Head motion-corrected eye gaze tracking with the da Vinci surgical system. *Int. J. Comput. Assist. Radiol. Surg.* 1–9 (2024) doi:10.1007/s11548-024-03173-4.
17. Son, J. et al. Evaluating fMRI-Based Estimation of Eye Gaze During Naturalistic Viewing. *Cereb Cortex* **30**, 1171–1184 (2019).
18. Ohno, T. & Mukawa, N. A free-head, simple calibration, gaze tracking system that enables

gaze-based interaction. *Proc Eye Track Res Appl Symposium Eye Track Res Appl - Etra'2004* 115–122 (2004) doi:10.1145/968363.968387.
19. Huang, C.-W., Tseng, S.-C., Jiang, Z.-S. & Hu, C.-W. Projective Mapping Compensation for the Head Movement during Eye Tracking. *2014 Ieee Int Conf Consumer Electron - Taiwan* 131–132 (2014) doi:10.1109/icce-tw.2014.6904021.
20. Zhu, Z. & Ji, Q. Novel Eye Gaze Tracking Techniques Under Natural Head Movement. *IEEE Trans. Biomed. Eng.* **54**, 2246–2260 (2007).
21. Frey, M., Nau, M. & Doeller, C. F. Magnetic resonance-based eye tracking using deep neural networks. *Nat. Neurosci.* **24**, 1772–1779 (2021).
22. Kimmig, H., Greenlee, M. W., Huethe, F. & Mergner, T. MR-Eyetracker: a new method for eye movement recording in functional magnetic resonance imaging. *Exp. Brain Res.* **126**, 443–449 (1999).
23. Tregellas, J. R., Tanabe, J. L., Miller, D. E. & Freedman, R. Monitoring eye movements during fMRI tasks with echo planar images. Hum. Brain Mapp. 17, 237–243 (2002).
24. Beauchamp, M. S. Detection of eye movements from fMRI data. *Magn. Reson. Med.* **49**, 376–380 (2003).
25. Studios, M. *Minecraft*. (Mojang Studios, Xbox Game Studios, 2011).
26. Hardy, R. L. Multiquadric equations of topography and other irregular surfaces. *J. Geophys. Res.* **76**, 1905–1915 (1971).
27. Kiat, L. C. & Ranganath, S. One-Time Calibration Eye Gaze Detection System. *2004 Int. Conf. Image Process., 2004 ICIP '04* **2**, 873–876 (2004).
28. Sheela, S. V. & Vijaya, P. A. Mapping Functions in Gaze Tracking. *Int. J. Comput. Appl.* **26**, 36–42 (2011).
29. Rashid, A. B. & Showva, N.-N. Design and fabrication of a biodegradable face shield by using cleaner technologies for the protection of direct splash and airborne pathogens during the COVID-19 pandemic. *Clean. Eng. Technol.* **13**, 100615 (2023).
30. Villoing, D. et al. KOREAN PEDIATRIC AND ADULT HEAD COMPUTATIONAL PHANTOMS AND APPLICATION TO PHOTON SPECIFIC ABSORBED FRACTIONS CALCULATIONS. *Radiat. Prot. Dosim.* **176**, 294–301 (2017).
31. Lee, W. et al. A 3D anthropometric sizing analysis system based on North American CAESAR 3D scan data for design of head wearable products. *Comput. Ind. Eng.* **117**, 121–130 (2018).
32. Fisher & R., A. Statistical Methods for Research Workers. 66–70 (1992) doi:10.1007/978-1-4612-4380-9_6.
33. Nau, M. et al. Neural and behavioral reinstatement jointly reflect retrieval of narrative events. *bioRxiv* 2024.10.19.619187 (2024) doi:10.1101/2024.10.19.619187.
34. Kay, K. N., Weiner, K. S. & Grill-Spector, K. Attention Reduces Spatial Uncertainty in Human Ventral Temporal Cortex. *Curr. Biol.* **25**, 595–600 (2015).
35. Hayhoe, M. M. Vision and Action. *Annual Review of Vision Science* (2017) doi:10.1146/annurev-.

36. Triesch, J., Ballard, D. H., Hayhoe, M. M. & Sullivan, B. T. What you see is what you need. *J. Vis.* **3**, 9–9 (2003).
37. Land, M. F. & Lee, D. N. Where we look when we steer. *Nature* **369**, 742–744 (1994).
38. Johansson, R. S., Westling, G., Bäckström, A. & Flanagan, J. R. Eye–Hand Coordination in Object Manipulation. *J. Neurosci.* **21**, 6917–6932 (2001).
39. Engel, S. A., Glover, G. H. & Wandell, B. A. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cereb. cortex (N. York, NY : 1991)* **7**, **18**1–192 (1997).
40. Wandell, B. A., Dumoulin, S. O. & Brewer, A. A. Visual Field Maps in Human Cortex. Neuron **56**, 366–383 (2007).
41. Albrecht, D. G. & Hamilton, D. B. Striate cortex of monkey and cat: contrast response function. *J. Neurophysiol.* **48**, 217–237 (1982).
42. Allen, E. J. et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat Neurosci* **25**, 116–126 (2022).
43. Dumoulin, S. O. & Wandell, B. A. Population receptive field estimates in human visual cortex. NeuroImage 39, 647–660 (2008).
44. Kay, K. N., Winawer, J., Mezer, A. & Wandell, B. A. Compressive spatial summation in human visual cortex. J. Neurophysiol. **110**, 481–494 (2013).
45. Fair, D. A. et al. Correction of respiratory artifacts in MRI head motion estimates. NeuroImage **208**, 116400 (2020).
46. Glimcher, P. W. THE NEUROBIOLOGY OF VISUAL-SACCADIC DECISION MAKING. *Neuroscience* **26**, 133–179 (2003).
47. Stewart, N., Hermens, F. & Matthews, W. J. Eye Movements in Risky Choice. *J. Behav. Decis. Mak.* **29**, 116–136 (2016).
48. Krajbich, I., Armel, C. & Rangel, A. Visual fixations and the computation and comparison of value in simple choice. Nat. Neurosci. 13, 1292–1298 (2010).
49. Hanke, M. et al. Spatial Learning and Attention Guidance. *Neuromethods* 291–305 (2019) doi:10.1007/7657_2019_31.
50. Santini, T., Fuhl, W. & Kasneci, E. PuReST: robust pupil tracking for real-time pervasive eye tracking. *Proc. 2018 ACM Symp. Eye Track. Res. Appl.* 1–5 (2018) doi:10.1145/3204493.3204578.
51. Finzi, D. et al. Differential spatial computations in ventral and lateral face-selective regions are scaffolded by structural connections. Nat. Commun. 12, 2278 (2021).
52. Benson, N. C. et al. The Human Connectome Project 7 Tesla retinotopy dataset: Description and population receptive field analysis. J. Vis. 18, 23 (2018).
53. Esteban, O. et al. FMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat Methods* **16**, 111–116 (2019).
54. Spector, R. H. The pupils. (1990).
55. Roth, S. D. Ray casting for modeling solids. *Comput. Graph. Image Process.* **18**, 109–144 (1982).

56. Phong, B. T. Illumination for computer generated pictures. *Commun. ACM* **18**, 311–317 (1975).
57. Blinn, J. F. Models of light reflection for computer synthesized pictures. *ACM SIGGRAPH Comput. Graph.* **11**, 192–198 (1977).
58. Yeo, B. T. T. et al. Functional Specialization and Flexibility in Human Association Cortex. Cereb Cortex 25, 3654–3672 (2015).