ECCENTRICITY CONFOUND IN EEG-BASED VISUAL ATTENTION DECODING FROM GAZE-FIXATED NEURAL TRACKING OF MOTION IN NATURAL VIDEOS

Yuanyuan Yao[1], *Celina Salamanca González*[1], *Simon Geirnaert*[1,2], *Céline R. Gillebert*[3,5], *Tinne Tuytelaars*[4], *Alexander Bertrand*[1]

[1]KU Leuven, Department of Electrical Engineering,
STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Belgium
[2]KU Leuven, Department of Neurosciences, Research Group ExpORL, Belgium
[3]KU Leuven, Brain and Cognition, Belgium
[4]KU Leuven, Department of Electrical Engineering, Processing Speech and Images (PSI), Belgium
[5]Leuven Brain Institute (LBI), Belgium

ABSTRACT

Objective. Decoding visual attention from brain signals during naturalistic video viewing has emerged as a new direction in brain-computer interface research. Current methods assume that stronger coupling between object motion and neural activity indicates higher attention, but this can be confounded by eye movement artifacts and stimulus properties. This study investigates how visual eccentricity—the distance between a visual object and the fixation point—affects neural responses when eye movement artifacts are controlled. *Approach.* EEG signals were recorded across three tasks that manipulated object eccentricity and attention conditions while participants maintained gaze fixation. Correlation analysis and match-mismatch decoding were performed to quantify the neural tracking of object motion. *Main results.* The analysis supports three conclusions: (1) neural tracking of object motion in natural videos works under gaze fixation; (2) the strength of neural tracking under gaze fixation is predictive of attention; and (3) there exists a significant eccentricity confound in the EEG responses, with poorer neural tracking of motion at larger eccentricities. *Significance.* These results provide critical evidence that findings from previous free-viewing studies reflect genuine neural processing rather than mere oculomotor artifacts. However, the identified eccentricity effect highlights a major limitation for current decoding approaches that assume coupling strength reflects attention levels alone.

***Index Terms*—** visual attention decoding, EEG, neural tracking, gaze fixation, eccentricity confound

1. INTRODUCTION

Neural activity is modulated by attention: neurons processing task-relevant information exhibit increased firing rates and reduced variability, while those processing irrelevant information are suppressed [1]. Therefore, stronger neural responses are expected with attended stimuli, leading to stronger temporal coupling between stimulus and neural activity. This modulation has been observed not only with simple synthetic stimuli but also with naturalistic speech, images, and videos [2, 3, 4, 5]. These findings form the foundation for numerous studies aiming to decode visual or auditory attention from brain signals [6, 7, 8], where stronger stimulus-response coupling is typically interpreted as evidence of higher attention levels towards the stimulus.

However, attention is not the only factor that influences stimulus-brain coupling strength. When using naturalistic stimuli, various confounding factors can modulate this coupling. For example, Li et al. [9] reconstructed speech envelopes from electroencephalography (EEG) signals and used the correlation between reconstructed and actual envelopes as a metric for auditory attention. While this score was higher when participants were attentive and lower during distraction, it also correlated with the time-varying standard deviation of the speech envelope, suggesting that the score was also sensitive to stimulus dynamics rather than attention alone. When this confound was not addressed, decoding performance deteriorated. This exemplifies that stimulus-brain coupling is not a pure measure of top-down goal-directed attention, but is also dependent on bottom-up sensory stimulus properties [10, 11]. Moreover, such bottom-up influences may not cancel out in complex naturalistic conditions and could systematically bias decoding results in real-world applications.

These concerns extend naturally to newly emerging EEG-based visual attention decoding paradigms with potential ap-

This research is funded by the Research Foundation - Flanders (FWO) project No G081722N and G026026N, junior postdoctoral fellowship fundamental research of the FWO (for S. Geirnaert, No. 1242524N), Internal Funds KU Leuven (projects IDN/23/006, C14/25/108, C3/25/107, and METH/24/003), and the Flemish Government (AI Research Program).
Y. Yao, S. Geirnaert, T. Tuytelaars and A. Bertrand are also affiliated with Leuven.AI - KU Leuven institute for AI, Belgium.

plications in education, neuromarketing, and clinical diagnostics [4, 5, 6]. We hypothesize that the spatial location of visual objects relative to gaze position (i.e., eccentricity, calculated from the center of the object's bounding box) could be a potential confound: a higher stimulus-brain coupling may not necessarily indicate higher attention, but could also result from the object being located at a more central location in the visual field. This is motivated by the well-established nonuniformity of visual processing, with maximal visual acuity and pattern recognition capabilities at the fovea and a rapid decline toward the periphery [12, 13]. A direct link between eccentricity and neural responses has been demonstrated previously. For instance, in a functional magnetic resonance imaging study by Wang et al. [14], participants were presented with images of different categories (faces, houses, etc.) at different eccentricity positions. They reported that the neural activation was most intense when stimuli were in the central visual field, and the magnitude of responses declined as eccentricity increased.

Beyond these stimulus-dependent effects, eye movements present another persistent confound in free-viewing paradigms. Under natural conditions, attention is typically accompanied by overt behavior such as directing gaze toward the attended object, which is known as overt attention. This is particularly relevant for paradigms relying on the motion information in naturalistic videos, as eye movements (saccades and smooth pursuit) are naturally evoked by the movements in the videos, and thus may strongly correlate with the video motion features. These eye movements can influence stimulus-EEG correlations through several mechanisms: (1) (residual) electrooculogram (EOG) artifacts in the EEG signals, (2) neural motor activity associated with the planning and execution of eye movements, and (3) neural responses to gaze-induced visual motion, akin to a moving-camera effect where the eyes serve as the camera. This ambiguity undermines findings from prior studies that reported significant neural tracking of video motion under free-viewing conditions [4, 5], since it remains unclear to what extent these effects reflect genuine neural processing of motion information rather than oculomotor contributions. In [6], eye gaze data were explicitly shown to be correlated with video motion features, and using gaze data alone could achieve comparable attention decoding performance to EEG-based decoders, which again raises concerns about the neural origin of the reported effects. Similarly, Ki et al. [15] showed that the neural responses are spatially selective and are enhanced for task-related locations, but since their paradigm relied on overt attention, doubts remain as to whether the reported enhancement was simply an artifact of eye gaze bias toward task-relevant locations, rather than a genuine neural correlate of covert attentional selection.

Since disentangling genuine neural correlates of attention from oculomotor confounds is particularly challenging under free-viewing conditions, we follow a common strategy in cognitive neuroscience: employing a covert attention paradigm. Participants maintain fixation while attending to stimuli in the visual periphery, thereby eliminating eye-movement confounds at the cost of reduced ecological validity. This controlled approach allows us to rigorously test two key questions: (1) whether neural tracking of video motion persists in the absence of eye movement contributions, and (2) how eccentricity of visual stimulation systematically influences neural tracking of naturalistic videos.

The remainder of this paper is organized as follows. Section 2 describes the experimental protocol, including visual stimuli, tasks, and data acquisition. Section 3 details preprocessing and correlation analysis methods. Section 4 presents results on the effects of attention and eccentricity on single-subject decoding performance and inter-subject correlation (ISC). Section 5 discusses the implications of our findings. Finally, Section 6 concludes the paper.

2. EXPERIMENTAL PROTOCOL

14 healthy young adults with normal or corrected-to-normal vision participated in the study, each providing written informed consent. Three visual attention tasks were designed to investigate the effects of attention and eccentricity confounds. This protocol was approved by the KU Leuven Social and Societal Ethics Committee (Application No. G-2022-4765-R3).

2.1. Visual content

The stimuli consisted of seven single-shot natural videos taken from the video set used in [5, 6]. Each video featured a person performing a stage act and was truncated to $3\,\mathrm{min}$ duration with a frame rate of $30\,\mathrm{Hz}$. To better control the performer's eccentricity relative to the gaze fixation point, we centered the performer in each video by (1) detecting the performer's bounding box in each frame using a pre-trained Mask R-CNN model [16], (2) defining the bounding box size as the maximum width and height across all frames, (3) smoothing the bounding box center positions to reduce abrupt movements, and (4) cropping the video frames based on the smoothed centers and fixed bounding box size. The resulting cropped videos varied in size and were resized, with aspect ratios preserved, to a height of 540 pixels for display on a 1920×1080 pixel canvas with a black background. Three tasks were designed (see Section 2.2). In Tasks 1 and 2, each cropped video gradually shifted from the screen center toward either the left or right side (randomly assigned) at a constant speed of 0.1 pixels per frame, while in Task 3, the videos remained centered (see Figure 2).

Throughout all tasks and videos, a fixation cross (60×60 pixels) was presented at the screen center. Its luminance decreased and returned to normal at irregular intervals. Additionally, a red circle (radius: 50 pixels) appeared and faded out at random times and at random locations within each

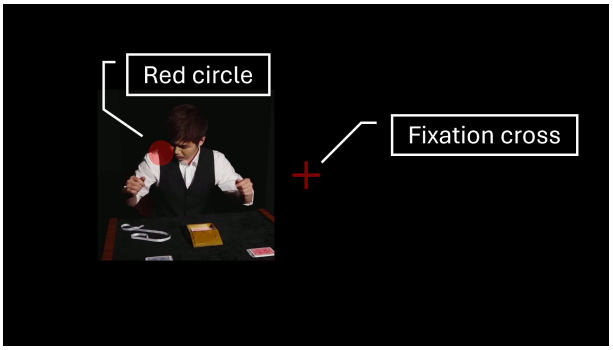

Fig. 1: Illustration of a frame showing the central fixation cross and a red circle appearing at a random location in the video.

cropped video. An example frame with a circle appearing is shown in Figure 1. These two visual events never occurred simultaneously, and each of these fade-out events lasted 2 seconds to avoid inducing strong time-locked event-related potentials. To keep participants engaged, they were instructed to respond to these two different types of events in different tasks (see Section 2.2). Each video was prefaced with instruction frames. The seven edited videos were then concatenated into a single 21-minute (excluding instructions and questions) video per task, resulting in a total of 63 minutes of experimental data per participant.

2.2. Tasks

Participants were instructed to maintain fixation on the central cross throughout all tasks. In Task 1 (ignore condition), they were instructed to ignore the video content. To discourage attending to the video, they were also instructed to press the "+" key with their preferred hand on a keyboard whenever the fixation cross dimmed. In Tasks 2 and 3 (attend conditions), participants attended to the videos while continuing to fixate. Because object eccentricity varies in Task 2 but remains constant in Task 3, the latter serves as a control to isolate the effects of eccentricity. To maintain consistent engagement with the video content, participants performed a secondary vigilance task requiring them to press the "o" key with their preferred hand whenever a subtle red circle appeared within the video. Both the task order and the video order within each task were randomized across participants. An illustration of the attention and eccentricity conditions across the three tasks is shown in Figure 2, and a summary of the tasks is provided in Table 1.

2.3. Data acquisition

Stimuli were presented on a 22-inch monitor (1920×1080, $60\,\mathrm{Hz}$ refresh rate). Participants sat approximately $130\,\mathrm{cm}$ from the screen and were instructed to maintain a stable posture and minimize head and body movements during video playback.

EEG data were recorded using a 64-channel BioSemi ActiveTwo system at $2048\,\mathrm{Hz}$. Four electrooculogram (EOG) electrodes were placed above/below the right eye and at the outer canthi of both eyes. Video onset was synchronized with EEG using a photodiode placed at the top-right corner of the screen, which detected a black-to-white box transition at the start of each video. This area was covered with black tape to avoid distracting the participants.

Gaze data were recorded using the Pupil Labs NEON eye tracker ($200\,\mathrm{Hz}$). A world-facing camera recorded the screen content from the participant's perspective. Gaze-video synchronization was achieved using a QR code embedded in the instruction frames. The time point when the QR code disappeared in the world-facing camera recording was marked as video onset, and gaze timestamps were aligned to this reference point.

3. METHODS

3.1. Preprocessing

The EEG and EOG data were preprocessed using standard techniques as in Yao et al. [5]: segmentation based on photodiode synchronization signals, bad channel interpolation, average referencing, high-pass filtering ($0.5\,\mathrm{Hz}$ cutoff), and notch filtering ($50\,\mathrm{Hz}$). The EOG data were then linearly regressed out from EEG to suppress ocular artifacts. All data were downsampled to $30\,\mathrm{Hz}$ with anti-aliasing to match the video frame rate. Rather than using raw video data, we extracted the object-based optical flow, which is a feature that quantifies the overall motion of a single object[1] in the video at each moment in time, resulting in a 1D time series with the same length as the original video [5]. The EEG signals and video features were centered to have zero mean per video.

Gaze coordinates, along with blink and saccade annotations, were exported from the eye tracker to identify non-fixation periods. First, gaze coordinates were clustered with the density-based spatial clustering algorithm DBSCAN [17], which is well suited for identifying dense fixation regions without requiring a predefined number of clusters. The largest cluster, corresponding to the most frequently viewed region, was interpreted as the fixation cluster. Samples falling outside this cluster were classified as non-fixations, except when blink annotations indicated that the deviation was caused by an eye blink. Saccade annotations served as a secondary check for non-fixation samples identified by DBSCAN. Any sample annotated as a saccade was categorized as a non-fixation, overriding its cluster assignment. An example of the gaze samples and the resulting fixation cluster for a representative participant in Task 1 is shown in Figure 3. The mean and standard deviation of the proportion of detected

[1] The videos in our experiment only contain one object, namely the human performer.

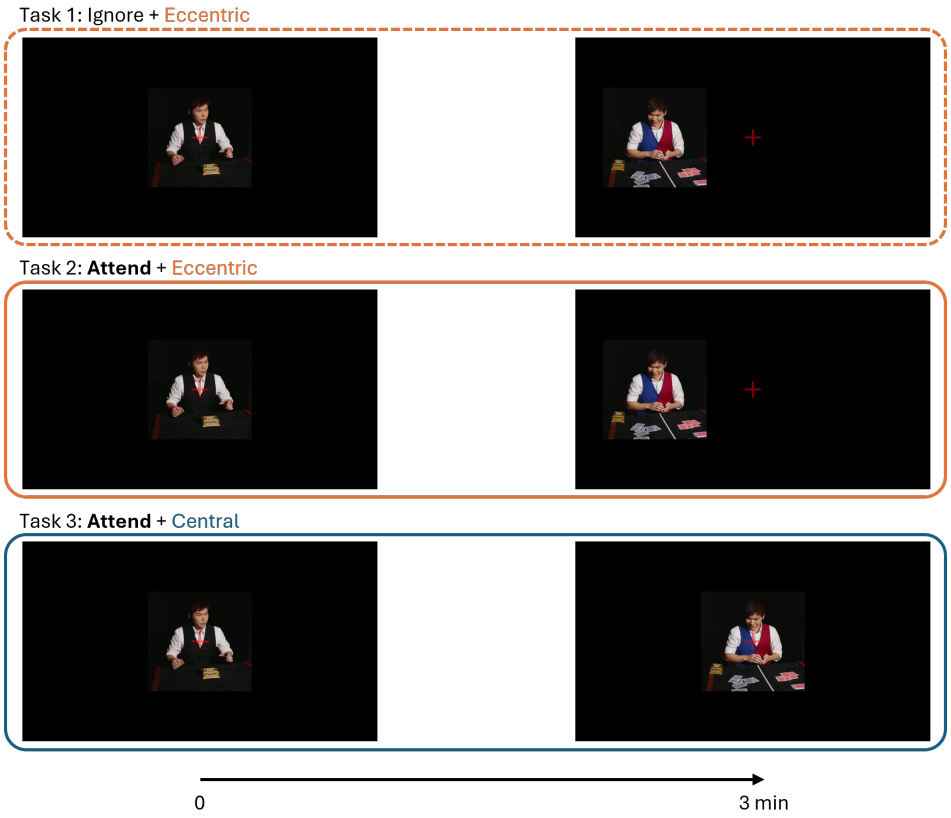

Fig. 2: Illustration of the three experimental tasks. In Task 1, participants ignored the videos while fixating centrally as the videos moved to the periphery. In Task 2, they attended to the videos while fixating centrally as the videos moved to the periphery. In Task 3, they attended to the videos while fixating centrally with the videos remaining in the center. Note that the foreground video does not always have a black or dark background as shown in this example.

gaze shifts across participants and videos for each task are summarized in Table 2.

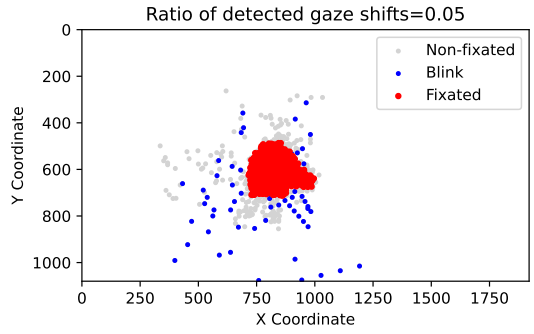

Fig. 3: Illustration of non-fixation detection for a representative participant (subject 4) during a video in Task 1. Dots indicate gaze positions with colors representing fixations (red), eye blinks (blue), and non-fixations (gray).

From Table 2, we conclude that the overall percentage of non-fixation samples is low, with Task 2 exhibiting a slightly higher proportion of gaze shifts. This increase may be attributable to the greater difficulty of maintaining central fixation when the attended object moves toward the periphery.

3.2. Correlation analysis

Two types of correlation analyses were performed: (1) stimulus-response correlation between EEG and video stimuli within each individual participant using canonical correlation analysis (CCA) [5], and (2) inter-subject correlation of EEG responses across participants using correlated component analysis (CorrCA) [18]. For completeness, we briefly summarize both methods here.

Denote the EEG signals as $\mathbf{X} \in \mathbb{R}^{T \times D_x}$ and video stimuli as $\mathbf{Y} \in \mathbb{R}^{T \times D_y}$, where T is the number of samples, D_x is the number of EEG channels, and D_y is the dimension of the video features. In our case $D_x = 64$ as we collected 64-channel EEG, and $D_y = 1$ as we used the 1-dimensional object-based optical flow feature. However, these dimensions were expanded by adding time-lagged copies of the signals, enabling a data-driven temporal alignment between the two modalities and realizing a data-driven finite impulse response filtering that suppresses irrelevant frequency bands. Specifi-

Table 1: Summary of experimental tasks

Task	Attention	Eccentricity	Gaze	Response trigger	Response
Task 1	Ignore	Central to peripheral	Fixed	Fixation cross dimming	Press "+" key
Task 2	Attend	Central to peripheral	Fixed	Red circle appearance	Press "o" key
Task 3	Attend	Central	Fixed	Red circle appearance	Press "o" key

Table 2: Gaze shift statistics across participants and videos per task.

	Percentage of detected gaze shift samples	
Task	Mean (%)	Std (%)
Task 1 (ignore, eccentric)	3.7	3.3
Task 2 (attend, eccentric)	5.9	5.6
Task 3 (attend, central)	3.6	3.5

cally, including $L_x - 1$ time lags for an EEG sample $\mathbf{x}(t) = [x_1(t), ..., x_{D_x}(t)] \in \mathbb{R}^{1\times D_x}$ yields the extended vector:

$$\tilde{\mathbf{x}}(t) = [x_1(t), x_1(t-1), \ldots, x_1(t-L_x+1), \ldots, x_{D_x}(t), \\ x_{D_x}(t-1), \ldots, x_{D_x}(t-L_x+1)] \in \mathbb{R}^{1\times D_x L_x}. \tag{1}$$

Similarly, the video features were extended with $L_y - 1$ time lags. We also included offsets to account for latencies between the EEG and the stimulus signal:

$$\tilde{\mathbf{x}}(t) = [x_1(t+\Delta_x), x_1(t-1+\Delta_x), \ldots, \\ x_1(t-L_x+1+\Delta_x), \ldots, x_{D_x}(t+\Delta_x), \\ x_{D_x}(t-1+\Delta_x), \ldots, x_{D_x}(t-L_x+1+\Delta_x)], \tag{2}$$

where Δ_x is the offset for EEG that can take positive or negative integer values. The inclusion of these fixed offsets is particularly useful to compensate for a delay in the neural response to the video while avoiding the introduction of a large number of time lags, which would unnecessarily expand the dimension of $\mathbf{X}$ and $\mathbf{Y}$. A similar operation was applied to video features with offset Δ_y. After this extension, the dimensions of $\mathbf{X}$ and $\mathbf{Y}$ became $T \times D_x L_x$ and $T \times D_y L_y$, respectively.

Recall that the signals were centered to have zero mean. The CCA model finds linear maps $\mathbf{w}_x$ and $\mathbf{w}_y$ such that the correlation between the transformed signals is maximized [19]:

$$\begin{aligned} \underset{\mathbf{w}_x, \mathbf{w}_y}{\text{maximize}} \quad & \mathbf{w}_x^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{Y} \mathbf{w}_y \\ \text{subject to} \quad & \mathbf{w}_x^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{w}_x = 1, \\ & \mathbf{w}_y^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{Y} \mathbf{w}_y = 1. \end{aligned} \tag{3}$$

The optimized $\mathbf{w}_x$ and $\mathbf{w}_y$ are the first canonical components, yielding the first canonical directions $\mathbf{X}\mathbf{w}_x$ and $\mathbf{Y}\mathbf{w}_y$. A higher correlation between these canonical directions indicates stronger neural tracking of the video features.

Higher-order components can be obtained by solving the same optimization problem with an additional constraint that these higher-order canonical directions are orthogonal to all previous canonical directions. Aggregating canonical components from first to K-th order as $\mathbf{W}_x \in \mathbb{R}^{D_x L_x \times K}$, $\mathbf{W}_y \in \mathbb{R}^{D_y L_y \times K}$, the solution can be obtained by solving a generalized eigenvalue decomposition (GEVD) problem:

$$\begin{aligned} \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{W}_x &= \mathbf{R}_{xx}\mathbf{W}_x\mathbf{\Lambda}, \\ \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{W}_y &= \mathbf{R}_{yy}\mathbf{W}_y\mathbf{\Lambda}, \end{aligned} \tag{4}$$

where $\mathbf{R}_{xy} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$, $\mathbf{R}_{yx} = \mathbf{Y}^{\mathrm{T}}\mathbf{X}$, $\mathbf{R}_{xx} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$, $\mathbf{R}_{yy} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y}$, and $\mathbf{\Lambda}$ is a diagonal matrix with eigenvalues in descending order.

CorrCA extends CCA to a group-level analysis, enabling the extraction of common neural components from the EEG data of multiple (more than two) participants. Since the EEG signals of all participants are time-locked on the same videos, these common components are expected to be neural responses to these videos [18]. Using subscript n to denote the subject index, EEG signals from different subjects are denoted as $\mathbf{X}_n$, where $n = 1, ..., N$. CorrCA finds a shared[2] spatial filter $\mathbf{W}_s$ that maximizes the approximate average pairwise correlations between all subject pairs:

$$\begin{aligned} \underset{\mathbf{W}_s}{\text{maximize}} \quad & \sum_{i=1, i\neq j}^{N} \sum_{j=1}^{N} \mathrm{Tr}\left(\mathbf{W}_s^{\mathrm{T}} \mathbf{R}_{ij} \mathbf{W}_s\right) \\ \text{subject to} \quad & \sum_{i=1}^{N} \mathbf{W}_s^{\mathrm{T}} \mathbf{R}_{ii} \mathbf{W}_s = \mathbf{I}_K, \end{aligned} \tag{5}$$

where $\mathrm{Tr}\,(\cdot)$ denotes the trace, $\mathbf{W}_s$ contains the first K canonical components, and $\mathbf{R}_{ij} = \mathbf{X}_i^{\mathrm{T}}\mathbf{X}_j$ is the cross-covariance matrix between participants i and j. This optimization problem can also be solved via a GEVD:

$$\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{R}_{ij}\right) \mathbf{W}_s = \left(\sum_{i=1}^{N} \mathbf{R}_{ii}\right) \mathbf{W}_s \mathbf{\Lambda}. \tag{6}$$

The eigenvalues in $\mathbf{\Lambda}$ are in descending order. Inter-subject correlation (ISC) for the k-th component is defined as the av-

[2] Other generalizations of CCA exist, which allow using a different spatial filter for each participant [20, 21]. We opted for CorrCA here because of the limited dataset size, since the reduction in degrees of freedom (i.e., using the same $\mathbf{W}_s$ for all participants) has a regularizing effect on the transformation [21].

erage correlation between all subject pairs:

$$\mathrm{ISC}_k = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\mathbf{w}_{s,k}^{\mathrm{T}} \mathbf{R}_{ij} \mathbf{w}_{s,k}}{\sqrt{\mathbf{w}_{s,k}^{\mathrm{T}} \mathbf{R}_{ii} \mathbf{w}_{s,k}} \sqrt{\mathbf{w}_{s,k}^{\mathrm{T}} \mathbf{R}_{jj} \mathbf{w}_{s,k}}}, \tag{7}$$

where $\mathbf{w}_{s,k}$ is the k-th column of $\mathbf{W}_s$. Higher ISC indicates stronger neural synchronization across participants, which has been linked to higher attention levels [18, 22].

3.3. Parameters and evaluation

Temporal lag parameters were set according to Yao et al. [5]. For CCA, we used $L_x = 3$ with $\Delta_x = 1$, capturing EEG information from $-33\,\mathrm{ms}$ to $33\,\mathrm{ms}$ around the current time point, and $L_y = 15$ with $\Delta_y = 0$, capturing video features from approximately 500 ms of history to the current time. For CorrCA, $L_x = 5$ and $\Delta_x = 2$ were used, covering $-67\,\mathrm{ms}$ to $67\,\mathrm{ms}$.

We adopted leave-one-video-out cross-validation in all experiments. Unless stated otherwise, training was performed on the combined data from all three experimental tasks (54 minutes per participant: 3 tasks $\times$ 6 videos $\times$ 3 minutes). This was an empirical decision, as aggregating data across all tasks yielded better decoding performance than training on individual tasks alone. A possible explanation is that even in Task 1, where participants were instructed to ignore the video, passive visual processing still engages the same neural pathways, providing useful information for estimating the spatial filters. In the stimulus-response correlation analysis, non-fixation periods identified in Section 3.1 were excluded (along with a 0.5-second margin on either side) from testing to strictly control for eye-movement confounds. Therefore, the size of the test set varied across tasks and participants. For CorrCA, all data were retained because the proportion of non-fixation samples was small and removing non-fixation periods would disrupt temporal alignment across participants. Test segments were defined as 45-second windows sampled from the test set with overlap, randomly drawn with an average interval of 2 seconds.

For CCA, instead of using raw correlation coefficients, which are rather noisy, we performed match-mismatch (MM) decoding to quantify stimulus-response coupling strength [23]. Given an EEG segment, the task is to identify the corresponding video segment from two candidates: the correct match and a mismatch (a segment from the same video but different time point). The procedure applies the trained EEG decoder $\mathbf{w}_x$ and video encoder $\mathbf{w}_y$ to compute canonical correlations between the test EEG segment and both candidate video segments. The candidate with the higher sum of the first two canonical correlation components was selected as the predicted match. The first two components were used because they were generally found to be significant across most participants, and aggregating them yielded better decoding than using only the first component. Decoding accuracy was then calculated across all test trials. To assess whether decoding performance was significantly above chance, we constructed a null distribution by circularly permuting the sequence of video segments such that the temporal alignment between EEG and video segments was disrupted and both MM candidates became mismatches (100 permutations per participant). The significance bound was set at the 97.5th percentile of this null distribution.

For CorrCA, ISC values were computed and averaged across all test segments. Significance was assessed using phase scrambling [24]. For each participant, the phase of every channel's Fourier coefficients was randomized while preserving the amplitude spectrum, thereby disrupting temporal correlations while maintaining the original power spectrum. ISC was then recomputed on these phase-scrambled data (500 times per fold), and the significance threshold was set at the 97.5th percentile of the resulting null distribution, corresponding to a two-sided test with $\alpha = 0.05$.

4. RESULTS

4.1. Effect of eccentricity on decoding performance

Tasks 1 and 2 both had videos gradually moving from center to periphery, differing only in attention condition: participants ignored video content in Task 1 but attended covertly in Task 2. Comparing these two tasks reveals whether attentional modulation persists when the visual objects are not centrally located. Task 3 maintained videos at the center throughout while the participants attended to the content. Comparing Task 3 with Task 2 isolates the effect of eccentricity on decoding performance. MM accuracies for all three tasks are shown in Figure 4.

In Task 1, more than half of the participants (8 out of 14) failed to achieve above-chance decoding accuracy, with a median accuracy of 0.55. Tasks 2 and 3 yielded median accuracies of 0.60 and 0.65, respectively, with most participants achieving significantly above-chance performance. To evaluate the distinct effects of attention and eccentricity, we conducted two planned comparisons using Wilcoxon signed-rank tests[3]. The Task 1 accuracy was significantly lower than that of Task 2 ($p = 0.034$), confirming that attentional modulation persists even when objects move away from the visual field center. Additionally, Task 2 accuracy was significantly lower than that of Task 3 ($p = 0.034$), demonstrating that shifting the visual stimulus to peripheral locations significantly impairs decoding performance even under identical attentional conditions. This provides direct evidence that eccentricity acts as a confounding factor in visual attention decoding.

[3]Because these comparisons were planned a priori to test different hypotheses (the effect of attentional modulation and the effect of visual eccentricity), no correction for multiple comparisons was applied.

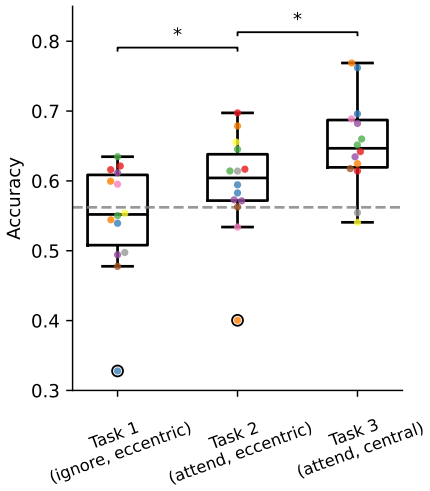

Fig. 4: MM-decoding accuracy across three experimental tasks on 45-second test segments. Individual participant accuracies are shown as colored dots, with statistical outliers circled. Box plots display median and interquartile range, with whiskers extending to extreme values. Dashed line indicates the significance bound for above-chance performance. Asterisks denote significant between-task differences (Wilcoxon signed-rank test, $p < 0.05$). Task 1 < Task 2 < Task 3, demonstrating both attentional effects and eccentricity confounds.

4.2. Gaze fixation versus free viewing

In this section, we compared the decoding performance under our gaze-fixation protocol with previous free-viewing studies. Using the free-viewing dataset from Yao et al. [5][4], we selected only the data corresponding to the same video stimuli used in our current study, yielding 21 minutes per participant (for both datasets). EOG data were linearly regressed out from EEG during preprocessing, as in the present dataset. For fair comparison, only the data from Task 3 (attention condition with object in the center) were used from the current study in both training and testing to have an equal amount of test and train data, and to remove effects of eccentricity. The same leave-one-video-out cross-validation was performed. Results are presented in Figure 5.

The gaze-fixation protocol yielded a mean accuracy of 0.62, compared to 0.68 for free viewing. To test whether this reduction was significant, we used a Mann-Whitney U test since the participant groups differed between the two protocols. The p-value was 0.024, suggesting significant performance loss under gaze fixation. Additionally, more participants failed to exceed the significance level under gaze fixation (4 out of 14) versus free viewing (1 out of 19). We note that the results in Figure 5 differ from the Task 3 results shown in Figure 4, which is due to the difference in the amount of training data in both analyses (the results in Figure 4 use 3 times more training data, as training data is aggregated across the 3 tasks).

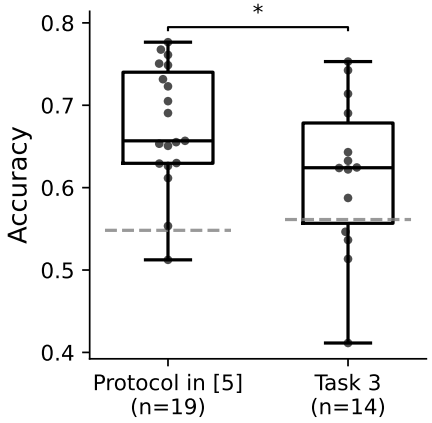

Fig. 5: Comparison of MM decoding accuracy on 45-second test segments under free-viewing (19 participants, [5, 25]) and gaze-fixation (14 participants) protocols. Dots represent the accuracies of individual participants. Box plots display median and interquartile range, with whiskers extending to extreme values. Dashed lines indicate significance bounds for above-chance performance. Gaze fixation significantly reduced decoding accuracy compared to free viewing (Mann-Whitney U test, $p = 0.024$).

4.3. Effect on group analysis

The previous sections investigated the stimulus-response coupling on a per-subject basis. In this section, we performed a group analysis based on the ISC to investigate the coherence of the neural responses across participants, without explicitly linking these to the stimulus signal. We computed ISCs for all three tasks in the current protocol, excluding time points surrounding visual events of the artificially added markers (fading of the fixation cross and red circle) that require key presses to avoid potential contamination from synchronized motor artifacts. The results are shown in Figure 6, alongside the free-viewing protocol results from [5] for comparison. The free-viewing results were computed by randomly selecting 14 participants (repeated 5 times) from the original dataset [5, 25] to match the current study's participant count, as the number of participants influences the ISC values. As in Section 4.2, only data corresponding to the same video stimuli as in the current gaze-fixation protocol were used, yielding 18×14 minutes of training data per fold. Note that this was only one-third of the training data available in the gaze-fixation protocol, where data from all three tasks were aggregated for training.

The ISC pattern mirrored the decoding accuracy trends from Section 4.1: Task 1 showed the lowest ISC (below significance threshold), followed by Task 2, with Task 3 exhibiting the highest ISC. This confirmed both attention- and

[4]The original study included 20 subjects, with 19 consenting to make their data publicly available in [25].

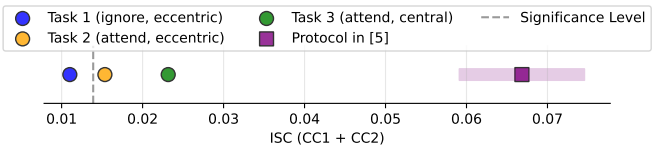

Fig. 6: Sum of the ISCs of the first two canonical components across experimental conditions. ISC values are computed separately for each task under the gaze-fixation protocol and compared with the free-viewing protocol. To match the participant count of the current study, the free-viewing reference is generated by repeatedly drawing 14 participants at random from the original dataset [5, 25] (5 independent draws) and computing the ISC for each draw. The shaded region represents the standard deviation across these draws and the square represents the mean. The vertical dashed line marks the significance threshold.

eccentricity-related attenuation of the stimulus-evoked neural responses, reducing the ISC across participants. However, Task 2 ISC was only marginally above significance, showing a larger gap with Task 3 than with Task 1. This suggests that eccentricity may have an even stronger impact on ISC than attention effects. Additionally, free-viewing protocol results were substantially higher despite a smaller training data size, which aligns with the different decoding performance observed in Section 4.2.

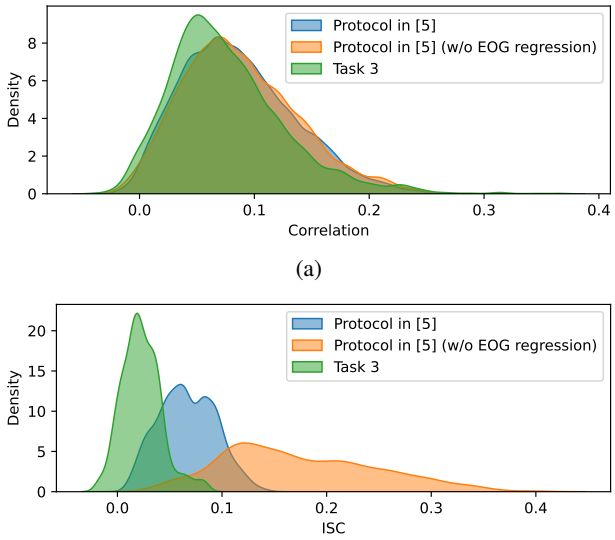

Fig. 7: Distributions of (a) stimulus-response correlations and (b) inter-subject correlations across participants for 45-second segments under the free-viewing and gaze-fixation protocols. The values represent the sum of the first two components. For comparison, results obtained from EEG without EOG regression are also included. For the gaze-fixation condition, only Task 3 (attend, central) is shown.

5. DISCUSSION

5.1. Group analysis is more susceptible to eye movement artifacts than stimulus-response analysis

From Figure 6, we observed that the decrease in ISC from free-viewing to gaze-fixation protocols appeared more pronounced than the corresponding decrease in decoding accuracy shown in Figure 5. However, note that Figure 5 displays MM accuracy, whereas ISC is a correlation metric. To more directly compare the effects of gaze fixation on group-level and individual-level analyses, we plotted the distributions of stimulus-response correlations (aggregated across participants) and ISCs in Figure 7 for both the free-viewing protocol and the gaze-fixation protocol (Task 3: attend, central). For reference, we also included results obtained without EOG regression under the free-viewing protocol.

For the individual stimulus-response analysis, the distributions with and without EOG regression under the free-viewing protocol overlapped substantially. Under the gaze-fixation protocol (with EOG regression), the distribution shifted slightly leftward with a thinner right tail, indicating that gaze fixation reduced stimulus-response correlations rather mildly. In contrast, the group analysis showed a much stronger effect. After regressing out EOG, the free-viewing ISC distribution already exhibited a pronounced leftward shift, indicating that EOG-capturable ocular artifacts contributed substantially to ISC. Under the gaze-fixation protocol, ISC values decreased further and the distribution became narrower. This additional reduction beyond EOG regression may be attributed to neural activities such as motor signals from saccade planning and execution that are not captured by EOG channels but are nonetheless synchronized across participants viewing the same dynamic content, thereby inflating ISC under free-viewing conditions. Together, these results suggest that group analysis is more susceptible to eye-movement confounds than individual stimulus-response analysis.

One likely reason is that eye movements tend to be synchronized across participants when viewing the same dynamic visual stimuli, artificially inflating ISC. Stimulus-response analysis, on the other hand, focuses on the relationship between the stimulus and each participant's neural response. Although eye-movement-related activity in the EEG can also contribute to stimulus-response correlations when using motion-based video features, its impact is less pronounced, likely because the relationship between the stimulus and eye movements is indirect and more difficult to capture.

5.2. Free-viewing studies likely rely on genuine neural responses to object motion

In previous free-viewing studies that explicitly addressed eye-movement confounds [5, 6], several analyses were conducted to demonstrate that the observed stimulus-response correlations and above-chance decoding accuracies were not solely driven by eye movement artifacts. These included regressing out EOG signals, restricting analyses to EEG channels less affected by eye movements, and removing data segments containing saccades. Together, these approaches provide strong evidence that EEG-based decoding of visual attention captures genuine neural responses to object motion. However, since it was never possible to completely rule out all eye-movement confounds in free-viewing paradigms, some uncertainty remained.

In the present study, strict gaze fixation was imposed to directly eliminate eye-movement confounds. The fact that significant decoding accuracy and ISC persisted under these conditions provides additional support that previous free-viewing results largely reflect genuine neural responses to object motion.

5.3. Why does gaze fixation reduce decoding accuracy and ISC?

While gaze fixation clearly reduces decoding accuracy and ISC, as shown in Sections 4.2 and 4.3, the underlying mechanisms remain unclear. The reduction could arise from the loss of synchronized eye-movement information that aids decoding and boosts ISC in free-viewing paradigms, and/or from the increased difficulty of maintaining sustained attention when gaze is constrained. The covert attention paradigm used here, while necessary to control for eye movement confounds, is less naturalistic than overt attention, where eye movements naturally accompany and facilitate engagement. This reduced ecological validity, combined with increased cognitive demand, may also contribute to weaker neural responses. Indeed, most participants reported that maintaining fixation while attending to moving objects was challenging. However, since the ground truth of a participant's actual attentional state is unavailable, disentangling these two factors is impossible here. Future studies could incorporate performance-tracked secondary tasks to objectively measure attentional consistency.

6. CONCLUSION

This study revealed a spatial confound in EEG-based visual attention decoding by manipulating stimulus eccentricity under gaze fixation. When objects moved toward peripheral locations, lower decoding accuracies and lower ISCs were observed, demonstrating that peripheral objects are more difficult to decode than central objects. This spatial confound represents an important limitation for current attention decoding approaches that assume coupling strength reflects attention levels alone. Comparison with free-viewing conditions showed that gaze fixation reduced both decoding accuracy and ISC, yet both measures remained statistically significant. The latter suggests that neural tracking of video motion still exists and is predictive of attention under strict gaze fixation, and that previous free-viewing studies also reflect genuine neural responses to object motion and not only gaze-induced components such as residual EOG artifacts, gaze-related motor neural activity, or gaze shift-evoked neural responses. While our sample size ($N = 14$) is relatively small, the observed attention and eccentricity effects were consistent across both individual-level stimulus-response analysis and group-level inter-subject correlation analysis, lending confidence to the robustness of our findings.

7. REFERENCES

[1] Alexander Thiele and Mark A Bellgrove, "Neuromodulation of attention," *Neuron*, vol. 97, no. 4, pp. 769–785, 2018.

[2] Nima Mesgarani and Edward F Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[3] Daniel Baldauf and Robert Desimone, "Neural mechanisms of object-based attention," *Science*, vol. 344, no. 6182, pp. 424–427, 2014.

[4] Jacek P Dmochowski, Jason J Ki, Paul DeGuzman, Paul Sajda, and Lucas C Parra, "Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity," *NeuroImage*, vol. 180, pp. 134–146, 2018.

[5] Yuanyuan Yao, Axel Stebner, Tinne Tuytelaars, Simon Geirnaert, and Alexander Bertrand, "Identifying temporal correlations between natural single-shot videos and EEG signals," *Journal of Neural Engineering*, vol. 21, no. 1, pp. 016018, 2024.

[6] Yuanyuan Yao, Wout De Swaef, Simon Geirnaert, and Alexander Bertrand, "EEG-based decoding of selective visual attention in superimposed videos," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[7] Wouter Biesmans, Neetha Das, Tom Francart, and Alexander Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 25, no. 5, pp. 402–412, 2016.

[8] Arnout Roebben, Nicolas Heintz, Simon Geirnaert, Tom Francart, and Alexander Bertrand, "'Are you even listening?' - EEG-based decoding of absolute auditory attention to natural speech," *Journal of Neural Engineering*, vol. 21, no. 3, pp. 036046, jun 2024.

[9] Tianyi Li, Simon Geirnaert, Davina Van de Broek, Elien Bellon, Bert De Smedt, and Alexander Bertrand, "Temporal variation in the acoustic dynamic range is a confounding factor in EEG-based tracking of absolute auditory attention to speech," in *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025, pp. 1–4.

[10] Maurizio Corbetta and Gordon L Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature reviews neuroscience*, vol. 3, no. 3, pp. 201–215, 2002.

[11] Laurent Itti and Christof Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[12] Christine A Curcio, Kenneth R Sloan, Robert E Kalina, and Anita E Hendrickson, "Human photoreceptor topography," *Journal of comparative neurology*, vol. 292, no. 4, pp. 497–523, 1990.

[13] Hans Strasburger, Ingo Rentschler, and Martin Jüttner, "Peripheral vision and pattern recognition: A review," *Journal of vision*, vol. 11, no. 5, pp. 13–13, 2011.

[14] Bin Wang, Jiayue Guo, Tianyi Yan, Seiichiro Ohno, Susumu Kanazawa, Qiang Huang, and Jinglong Wu, "Neural responses to central and peripheral objects in the lateral occipital cortex," *Frontiers in human neuroscience*, vol. 10, pp. 54, 2016.

[15] Jason J Ki, Jacek P Dmochowski, Jonathan Touryan, and Lucas C Parra, "Neural responses to natural visual motion are spatially selective across the visual field, with selectivity differing across brain areas and task," *European Journal of Neuroscience*, vol. 54, no. 10, pp. 7609–7625, 2021.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[17] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu, "Dbscan revisited, revisited: why and how you should (still) use dbscan," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[18] Jacek P Dmochowski, Paul Sajda, Joao Dias, and Lucas C Parra, "Correlated components of ongoing EEG point to emotionally laden attention–a possible marker of engagement?," *Frontiers in human neuroscience*, vol. 6, pp. 112, 2012.

[19] Harold Hotelling, "Relations between two sets of variates," *Breakthroughs in statistics: methodology and distribution*, pp. 162–190, 1992.

[20] Jon R Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[21] Simon Geirnaert, Yuanyuan Yao, Tom Francart, and Alexander Bertrand, "Stimulus-informed generalized canonical correlation analysis for group analysis of neural responses to natural stimuli," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 2, pp. 970–983, 2025.

[22] Jason J Ki, Simon P Kelly, and Lucas C Parra, "Attention strongly modulates reliability of neural responses to naturalistic narrative stimuli," *Journal of Neuroscience*, vol. 36, no. 10, pp. 3092–3101, 2016.

[23] Alain De Cheveigné, Malcolm Slaney, Søren A Fuglsang, and Jens Hjortkjaer, "Auditory stimulus-response modeling with a match-mismatch task," *Journal of Neural Engineering*, vol. 18, no. 4, pp. 046040, 2021.

[24] Dean Prichard and James Theiler, "Generating surrogate data for time series with several simultaneously measured variables," *Physical review letters*, vol. 73, no. 7, pp. 951, 1994.

[25] Yuanyuan Yao, Axel Stebner, Tinne Tuytelaars, Simon Geirnaert, and Alexander Bertrand, "Video-EEG encoding-decoding dataset KU Leuven," `https://doi.org/10.5281/zenodo.10512414`, Jan. 2024.