EEG-based Decoding of Selective Visual Attention in Superimposed Videos

Yuanyuan Yao, Wout De Swaef, Simon Geirnaert, and Alexander Bertrand, *Senior Member, IEEE*

***Abstract*— Selective attention enables humans to efficiently process visual stimuli by enhancing important locations or objects and filtering out irrelevant information. Locating visual attention is a fundamental problem in neuroscience with potential applications in brain-computer interfaces. Conventional paradigms often use synthetic stimuli or static images, but visual stimuli in real life contain smooth and highly irregular dynamics. In this study, we show that these irregular dynamics in natural videos can be decoded from electroencephalography (EEG) signals to perform selective visual attention decoding. To this end, we propose an experimental paradigm in which participants attend to one of two superimposed videos, each showing a center-aligned person performing a stage act. We then train a stimulus-informed decoder to extract EEG signal components that are correlated with the motion patterns of the attended object, and show that this decoder can be used on unseen data to detect which of both objects is attended. Eye movements are also found to be correlated to the motion patterns in the attended video, despite the spatial overlap between the target and the distractor. We further show that these eye movements do not dominantly drive the EEG-based decoding and that complementary information exists in EEG and gaze data. Moreover, our results indicate that EEG also captures information about unattended objects. To our knowledge, this study is the first to explore EEG-based selective visual attention decoding on natural videos, opening new possibilities for experiment design in related fields.**

***Index Terms*— brain-computer interface, EEG, selective visual attention decoding, video stimuli**

I. Introduction

IN everyday life, humans are constantly exposed to a vast amount of visual information. To process this with limited resources, the brain has developed a mechanism known as selective visual attention, which enables individuals to prioritize stimuli of interest in the visual field while suppressing others [1]. Decoding selective visual attention has been a popular research topic in neuroscience and brain-computer interface communities, providing insights into the neural basis of attention and offering potential applications in various fields. For example, it can aid communication and control for individuals with severe paralysis [2], [3], diagnosis of attention and consciousness-related disorders [4], [5], [6], rehabilitation of cognitive deficits [7], [8], and optimization of streaming processes in virtual reality [9].

Extensive research has been conducted on the mechanisms underlying selective visual attention. These studies have shown that, although the sensory representations of both attended and unattended stimuli are present in the visual field, the attended stimulus elicits stronger cortical responses [10], [11]. This modulation effect of attention on neural responses enables the neural-based decoding of selective visual attention. For example, Kelly et al. [2] successfully decoded covert left/right spatial attention from steady-state visual evoked potentials elicited by flickering stimuli. The classification was based on the amplitude of the evoked potentials at the flicker frequency, which was approximately doubled when the flickering stimulus was attended.

Apart from spatial locations, selective attention also frequently targets specific objects. Attention enhances the features of the attended object, such as its motion, color, or shape, even when attended and unattended objects are superimposed [11], [12]. Early studies using functional magnetic resonance imaging (fMRI) have shown that it is possible to decode the category of the attended object in an image with superimposed objects from different categories [8], [13]. In more recent studies such as [14], the categories of unattended objects were also decoded and compared with the decoding accuracy of attended objects, showing that the latter were more accurately decodable. These studies trained classifiers only on brain signals to decode the object category, whereas Horikawa et al. [15] incorporated image features, decoding these features from the fMRI voxels using linear regression. Additionally, a deep generator network was appended after the regression model to generate images from the output features. The overall model predicted the features of the superimposed images and generated corresponding images that, as shown in the study, were similar to the attended object. Efforts have also been made to decode selective attention on superimposed images using electroencephalography (EEG) [5], [16], as EEG has much more potential for real-world applications due to its

This research is funded by the Research Foundation - Flanders (FWO) project No G081722N, junior postdoctoral fellowship fundamental research of the FWO (for S. Geirnaert, No. 1242524N), the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 101138304), Internal Funds KU Leuven (project IDN/23/006), and the Flemish Government (AI Research Program). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the granting authorities. Neither the European Union nor the granting authorities can be held responsible for them.

All authors are with KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics and with Leuven.AI - KU Leuven institute for AI, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium (e-mail: yuanyuan.yao@kuleuven.be, wout.deswaef@student.kuleuven.be, simon.geirnaert@kuleuven.be, alexander.bertrand@kuleuven.be).

S. Geirnaert is with KU Leuven, Department of Neurosciences, Research Group ExpORL, Herestraat 49 box 721, B-3000 Leuven, Belgium.

affordability and portability. Additionally, the high temporal resolution of EEG enables the capture and analysis of neural responses not only to static images but also to images presented in rapid succession [16].

Previous studies provide valuable insights into the neural basis of selective visual attention and demonstrate the feasibility of decoding selective attention based on brain signals. However, these studies primarily focus on synthetic stimuli and static or rapid serial images of various objects or scenes, which do not reflect the dynamic and continuous visual stimuli encountered in real life. This motivates us to explore natural and more dynamic visual stimuli: videos. EEG has a high temporal resolution such that it can capture the fast dynamics of neural responses elicited by the time-varying features within the video [17]. Using EEG signals, we aim to decode the attended object in videos with two moving objects (persons) that spatially overlap. To the best of our knowledge, no study has yet attempted to decode selective visual attention based on EEG signals when viewing natural videos. This work may lay the foundation for more naturalistic experiment design in cognitive neuroscience and brain-computer interfaces, more patient-friendly assessments of attention-related disorders, and attention tracking systems, e.g., for education and neuromarketing [18], [19]. It could also find applications in augmented reality, such as identifying whether the user's attention is focused on the (superimposed) virtual or real world [20].

In this work, the selective attention decoding is based on stimulus reconstruction, identifying the attended object by comparing the temporal correlations between the EEG responses and the features of the object(s) in the video. We choose this correlation-based paradigm since direct classification based on instantaneous features could overfit to confounds such as trial-dependent EEG feature shifts [21], [22]. To achieve our goal, it is essential to find a video feature that is correlated with EEG signals and modulated by attention. The motion-encoding object-based optical flow proposed in [17] extracts a time series that contains the average optical flow within the object at each time point, and was found to have significant correlations with EEG signals. In this study, we show that this feature is modulated by attention, allowing us to decode selective visual attention. Since participants are allowed to freely watch the videos without fixating on a specific point (though tracking only one of two center-aligned superimposed video objects), we also investigate the possibility of decoding selective attention by correlating the per-object optical flow time series with the eye movements, comparing their performance with EEG-based decoding.

The rest of this paper is organized as follows: Section II details the experimental setup, data preprocessing and feature extraction, introduces the analysis tools, and describes the evaluation tasks and the practicalities of our implementation. Section III presents the results and their implications, with a more in-depth discussion in Section IV. Section V concludes the paper.

II. Materials and Methods

A. Stimuli

This study primarily focuses on the neural decoding of selective visual attention when viewing natural videos with two moving objects. There is an important restriction on the experiment videos: they must be single-shot, i.e., with static camera angles and no scene changes. It is motivated by previous studies that have shown that the discontinuities due to shot cuts in videos can elicit strong neural responses [17], [23], [24], which are absent in natural visual stimuli, and which may lead to over-optimistic decoding performance [17].

We create video stimuli by merging pairs of single-shot videos, each containing a single moving person. We do not start from two-object videos to avoid interactions between the objects causing, e.g., correlated motion patterns and possible occlusions. The objects in both videos are first centered based on the center of mass of their bounding box and then superimposed with 50% transparency (Fig. 1). This transparent overlay avoids that the two objects are at different locations in the scene. In a pilot experiment we have observed that such a spatial separation places the interfering object in the peripheral visual field when focusing on the other object, which significantly reduces the neural responses to the former, making an EEG-based selective attention decoding rather trivial. The superimposed setting targeted in this study can therefore be viewed as a worst-case scenario in which the two objects always overlap in space.

14 single-shot, single-object videos with an average length of $305\,\mathrm{s}$ are selected, most of which are inherited from a previous study [17]. The frame rate is $30\,\mathrm{Hz}$, and the resolution is 1920×1080. The content of these videos includes a person performing a specific stage act, such as dancing, acrobatics, magic shows, and mime shows. The 14 videos are paired into 7 pairs: $(\mathrm{video}_{1,i}, \mathrm{video}_{2,i}), i \in \{1, 2, ..., 7\}$. The videos in each pair are not necessarily from different content categories but are distinct enough both visually and in terms of motion patterns, such that it is relatively easy for a participant to focus on one object while ignoring the other.

Fig. 2 illustrates the procedure of creating the experiment stimuli from these 7 pairs. In each pair, we truncate the videos to the minimum length of the two, and superimpose them with 50% transparency, except for the first two minutes. In these first two minutes, only the video of the attended object is visible (i.e., with a 100% transparency for the unattended video). The transition from single video to 50%-50% superimposed videos is made smooth by linearly changing the transparency over two seconds. Each video pair is presented twice, where the attended object switches in both presentations. This means that the first two minutes (showing only the attended object) is different in both presentations, yet the remaining part (showing 50%-50% superimposed videos) is exactly the same stimulus in both presentations.

Instruction frames are prepended in each video. These instruction frames contain a QR code for synchronization and an instruction text asking participants to always focus on the first object presented during the first two minutes of the video. The number of instruction frames is set to ensure they last longer than 30 seconds and make the total video length a

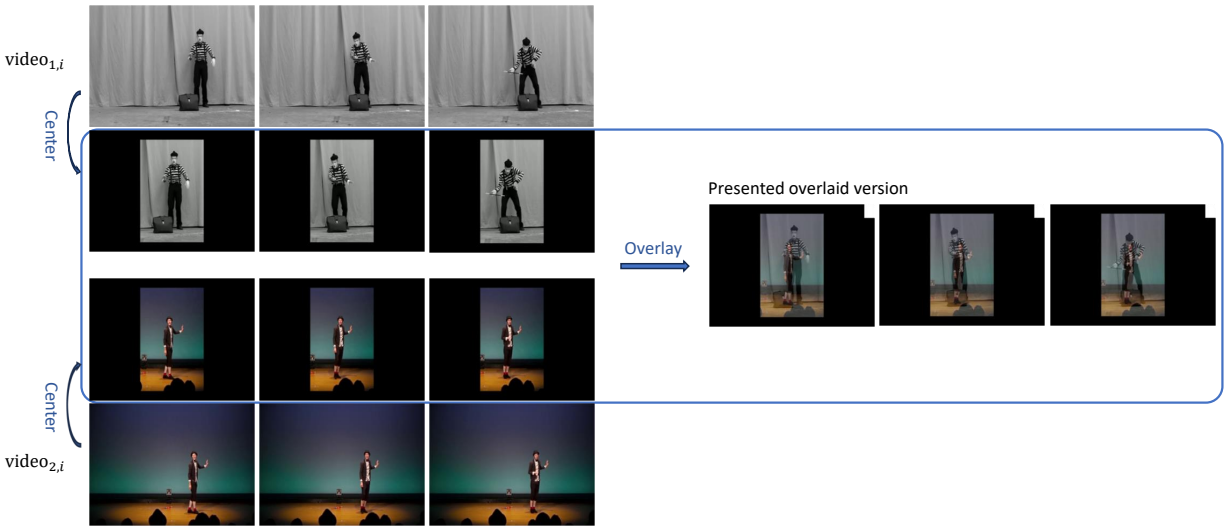


Fig. 1: An illustration of creating superimposed videos. The objects in two single-shot videos are centered and overlaid with 50% transparency. A white box is inserted in the top right corner to indicate the content-playing stage.

multiple of one minute. A progress bar is embedded to indicate the start of the video playback. The experiment consists of two trials of 42 minutes during which each video pair is presented once (randomized across participants). In the second trial, the 7 pairs are presented again, but with attention to the other object.

B. Participants and data acquisition

19 young, healthy adults participated in the experiment. All have normal or corrected-to-normal vision and no history of neurological disorders. The study is approved by the KU Leuven Social and Societal Ethics Committee, and all participants provided written informed consent before the experiment.

During the experiment, participants are seated comfortably in a quiet, separate room, approximately 90 cm from the screen. The stimuli (videos) are presented on a 22-inch monitor with a resolution of 1920×1080 pixels and a refresh rate of 60 Hz. Participants are instructed to watch the videos attentively and naturally, and to minimize unnecessary movements for better data quality. While participants watch the videos, their EEG data are recorded using a BioSemi ActiveTwo system (BioSemi B.V., Amsterdam) with 64 channels and a sampling rate of 2048 Hz. Eye movements are coregistered with EEG using four electrooculogram (EOG) electrodes placed above and below the right eye and on the outer canthi of both eyes. Participants also wear a NEON eye tracker (Pupil Labs GmbH, Berlin) to acquire gaze data at 200 Hz. Four markers are placed around the screen for defining the surface that the gaze data are mapped to.

The experiment videos consist of interleaving instruction-showing and content-playing stages, as detailed in Section II-A. A small box is embedded in the top right corner of the videos without occluding any content, serving as an indicator of the two different stages. The box is black during the instruction stage and turns white when the content starts to play, which is captured by a photodiode fixed at the corresponding region and connected to the EEG recorder. Synchronization between the EEG data and the video stimuli is achieved by detecting the upper edges of the photodiode signal. The embedded box is covered with black tape to prevent distraction.

To synchronize the eye tracker data with the video stimuli, a QR code is encoded during the instruction stage and is detected post hoc from the videos recorded by the world camera of the eye tracker. The time points at which the QR code appears are identified and aligned with the corresponding time points during the instruction stage. The first time point when the QR code disappears from the video is the synchronization point between the eye tracker data and the video stimuli.

The three data sources (EEG recorder, eye tracker, and video stimuli) are thus synchronized. This synchronization is performed per short video clip ($4 \sim 8$ minutes) in each trial, rather than only once at the beginning, to prevent non-negligible time lags caused by differences in the time clocks of each device.

C. Data preprocessing

The EEG and EOG data are first segmented based on the photodiode signals, which indicate the start of each content-playing stage. Basic preprocessing is applied to each segment, including interpolation of bad channels, average re-referencing, high-pass filtering with a cutoff frequency of 0.5 Hz to remove drifts, notch filtering at 50 Hz to remove powerline noise, and downsampling to 30 Hz (including anti-aliasing) to match the video frame rate. The filters are zero-phase and thus no delays are introduced.

From the eye tracker, we export the gaze coordinates mapped to the screen surface, the start and end points of fixations, and the time points of eye blinks. Saccades are identified as the end of fixations and are represented as a binary time series, indicating whether a saccade occurs at each frame. Eye blinks in the gaze data are linearly interpolated, and the gaze data are downsampled to 30 Hz.

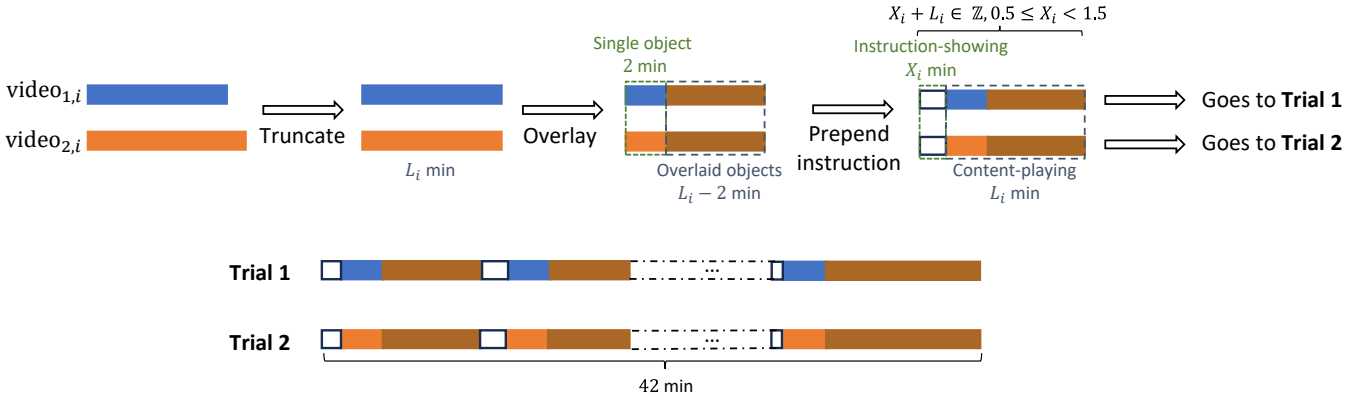


Fig. 2: An illustration of creating experimental videos: The original video pairs are truncated, overlaid from the second minute, and prepended with instruction frames. The videos in each pair are assigned to different trials.

Eye movements can also be informative, as participants' gaze may track the movement pattern of the attended object. We therefore extract a feature from both the gaze and the EOG data that is representative for the velocity of the eye movements:

$$\text{velocity} = \sqrt{(a(t) - a(t-1))^2 + (b(t) - b(t-1))^2}, \quad (1)$$

where $a(t)$ and $b(t)$ represent the horizontal and vertical gaze coordinates or the horizontal and vertical EOG channels at time t, respectively.

Overall, six data modalities are extracted for further analysis, as summarized in Table I. The data are further divided into two sets based on whether a single-object video or an superimposed-object video is playing. The fade-in periods are excluded from the analysis. Additionally, the first and last second of each video segment are also excluded to avoid potential effects caused by video onset and offset. In the end, the single-object dataset contains 19 subjects $\times$ 14 videos $\times$ 2 minutes of data. The superimposed-object dataset contains 19 subjects, 14 videos with an average video length of 166 s and a standard deviation (STD) of 76 s, totaling approximately 19 subjects $\times$ 38 minutes of data. The instruction-showing stage at the beginning of each video also serves as a short break for the participants and is a total of 18 minutes long.

TABLE I: Data modalities extracted from the recorded data.

Modality	Abbreviation
64-channel EEG signal	EEG
4-channel EOG signal	EOG
2D gaze coordinate	$GAZE$
Saccade (binary time series)	$SACC$
Velocity (1) calculated from gaze	$GAZE_V$
Velocity (1) calculated from EOG	EOG_V

D. Video feature extraction

Our approach for decoding selective attention is by identifying the temporal correlation between the dynamics in the video and the stimulus-following neural responses that follow these time-varying features. However, video data are high-dimensional, leading to an explosion of model parameters if it would be used in its raw format. Therefore, it is crucial to first extract relevant features that elicit strong neural responses in order to reduce data dimensionality. In [17], object-based optical flow (*ObjFlow*) and object-based temporal contrast (*ObjTempCtr*) were found to be correlated with EEG signals. We select *ObjFlow* for this study, as the performance of both features is comparable.

ObjFlow is defined as the average optical flow magnitude within the object of interest:

$$\text{ObjFlow} = \frac{1}{|\mathcal{O}|} \sum_{\mathbf{z} \in \mathcal{O}} |\mathbf{v}(\mathbf{z}, t)|, \quad (2)$$

where $|\mathbf{v}(\mathbf{z}, t)|$ denotes the magnitude of pixel velocity at position $\mathbf{z}$, time t, and $|\mathcal{O}|$ represents the number of pixels in the object mask. In practice, the object mask is obtained by applying a pre-trained object segmentation model, Mask R-CNN [25]. The optical flow is calculated using the Gunnar-Farneback method [26]. All videos are downsampled to 640×360 pixels to reduce computational cost before feature extraction.

E. Correlation analysis

Correlation analysis can be conducted on two or more views to measure their temporal coupling, optionally controlling for the effects of certain variables. In this section, we briefly review canonical correlation analysis and its two extensions, partial canonical correlation analysis and generalized canonical correlation analysis, and explain their application to our data. All data modalities and extracted features are centered before correlation analysis.

1) Canonical Correlation Analysis (CCA): CCA is a method for finding correlations between two sets of variables [27]. In this study, it is used to quantify the correlations between the video stimuli and the various data modalities introduced in Section II-C and Table I. When correlating a data modality (e.g., EEG signals) $\mathbf{x}(t) \in \mathbb{R}^{D_x}$ with video features $\mathbf{y}(t) \in \mathbb{R}^{D_y}$, CCA finds linear maps $\mathbf{w}_x \in \mathbb{R}^{D_x}$ and $\mathbf{w}_y \in \mathbb{R}^{D_y}$ that maximize the correlation between the transformed signals

$\mathbf{w}_x^{\mathrm{T}}\mathbf{x}(t)$ and $\mathbf{w}_y^{\mathrm{T}}\mathbf{y}(t)$. Mathematically, this can be expressed as the following optimization problem [27]:

$$\begin{aligned} \underset{\mathbf{w}_x,\mathbf{w}_y}{\text{maximize}} \quad & \mathbb{E}\{[\mathbf{w}_x^{\mathrm{T}}\mathbf{x}(t)][\mathbf{w}_y^{\mathrm{T}}\mathbf{y}(t)]\} \\ \text{subject to} \quad & \mathbb{E}\{[\mathbf{w}_x^{\mathrm{T}}\mathbf{x}(t)]^2\} = 1, \\ & \mathbb{E}\{[\mathbf{w}_y^{\mathrm{T}}\mathbf{y}(t)]^2\} = 1, \end{aligned} \tag{3}$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation operator. Correlations between neighboring samples can also be incorporated by extending $\mathbf{x}(t)$ and $\mathbf{y}(t)$ with $L_x - 1$ and $L_y - 1$ time-lagged copies, respectively, which also allows to automatically correct for relative time delays between $\mathbf{x}(t)$ and $\mathbf{y}(t)$. $\mathbf{w}_x$ and $\mathbf{w}_y$ then become spatial-temporal, with dimensions D_xL_x and D_yL_y. Solving problem (3) requires estimating the covariance matrices $\mathbf{R}_{xy} = \mathbb{E}\{\mathbf{x}(t)\mathbf{y}(t)^{\mathrm{T}}\} \in \mathbb{R}^{D_xL_x \times D_yL_y}$, $\mathbf{R}_{xx} = \mathbb{E}\{\mathbf{x}(t)\mathbf{x}(t)^{\mathrm{T}}\} \in \mathbb{R}^{D_xL_x \times D_xL_x}$, and $\mathbf{R}_{yy} = \mathbb{E}\{\mathbf{y}(t)\mathbf{y}(t)^{\mathrm{T}}\} \in \mathbb{R}^{D_yL_y \times D_yL_y}$, which can be approximated by the sample covariance matrices.

While (3) only aims to find a single canonical component $\mathbf{w}_x$ and $\mathbf{w}_y$, often higher-order canonical components are jointly estimated. Let $\mathbf{w}_x^k$ and $\mathbf{w}_y^k$ denote the k-th order canonical components, and $\mathbf{x}(t)^{\mathrm{T}}\mathbf{w}_x^k$ and $\mathbf{y}(t)^{\mathrm{T}}\mathbf{w}_y^k$ represent the k-th canonical directions. These components, $\mathbf{w}_x^k$ and $\mathbf{w}_y^k$, are linear maps applied to the data such that the transformed signals are orthogonal to all preceding canonical directions and are maximally correlated. In compact form, the multi-component version of (3) can be formulated as:

$$\begin{aligned} \underset{\mathbf{W}_x,\mathbf{W}_y}{\text{maximize}} \quad & \mathrm{Tr}\left(\mathbf{W}_x^{\mathrm{T}}\mathbf{R}_{xy}\mathbf{W}_y\right) \\ \text{subject to} \quad & \mathbf{W}_x^{\mathrm{T}}\mathbf{R}_{xx}\mathbf{W}_x = \mathbf{I}_K, \\ & \mathbf{W}_y^{\mathrm{T}}\mathbf{R}_{yy}\mathbf{W}_y = \mathbf{I}_K, \end{aligned} \tag{4}$$

where K is the number of components, $\mathbf{I}_K$ is the identity matrix of size K, $\mathbf{W}_x \in \mathbb{R}^{D_xL_x \times K}$ and $\mathbf{W}_y \in \mathbb{R}^{D_yL_y \times K}$ store the canonical components as columns, and $\mathrm{Tr}\,(\cdot)$ denotes the trace of a matrix. It can be shown that the solution to (4) can be obtained by solving the following generalized eigenvalue decomposition (GEVD) problem [28]:

$$\begin{bmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xy} \\ \mathbf{R}_{yx} & \mathbf{R}_{yy} \end{bmatrix} \begin{bmatrix} \mathbf{W}_x \\ \mathbf{W}_y \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{yy} \end{bmatrix} \begin{bmatrix} \mathbf{W}_x \\ \mathbf{W}_y \end{bmatrix} \mathbf{\Lambda}, \tag{5}$$

where $\mathbf{\Lambda} \in \mathbb{R}^{K \times K}$ is a diagonal matrix containing the generalized eigenvalues (GEVLs). The first K canonical components are the generalized eigenvectors (GEVCs) corresponding to the K largest GEVLs. The components (columns of $\mathbf{W}_x$ and $\mathbf{W}_y$) are rescaled to satisfy the constraints in (4).

2) Partial Canonical Correlation Analysis (PCCA): PCCA was proposed by Rao in [29] as an extension of CCA to account for the effects of confounding variables $\mathbf{c} \in \mathbb{R}^{D_cL_c}$, where D_c represents the dimension of the confounds and L_c denotes the number of time-lagged copies. Consider the problem discussed in Section II-E.1, where we aim to quantify the correlation between EEG signals and video features. Eye movements, often considered artefacts in EEG signals, may also correlate with video features, as specific patterns in the video may provoke particular eye movements. Therefore, it may be necessary to control for the effects of eye movements, which are captured by EOG signals.

PCCA involves one additional step compared to CCA: removing the effects of the confounds $\mathbf{c}$ from $\mathbf{x}$ and $\mathbf{y}$ by linear regression. Aggregating the samples of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{c}$ into matrices $\mathbf{X} \in \mathbb{R}^{T \times D_xL_x}$, $\mathbf{Y} \in \mathbb{R}^{T \times D_yL_y}$, and $\mathbf{C} \in \mathbb{R}^{T \times D_cL_c}$, respectively, the residuals can be written as:

$$\mathbf{X}_r = \mathbf{X} - \mathbf{P}_c\mathbf{X}, \tag{6a}$$

$$\mathbf{Y}_r = \mathbf{Y} - \mathbf{P}_c\mathbf{Y}, \tag{6b}$$

where $\mathbf{P}_c = \mathbf{C}(\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}$ is the projection matrix onto the column space of $\mathbf{C}$. The residuals $\mathbf{X}_r$ and $\mathbf{Y}_r$ are then fed to the input of the CCA method.

3) Generalized Canonical Correlation Analysis (GCCA): GCCA is a generalization of CCA that can handle more than two views, making it useful for applications such as finding coherent EEG signals across multiple subjects. Two well-known formulations of GCCA are SUMCORR and MAXVAR [30]. In this study, we select MAXVAR because the SUMCORR formulation does not have a closed-form solution [31].

MAXVAR-GCCA optimizes the decoders $\{\mathbf{W}_n\}_{n=1}^N$ applied to different views $\{\mathbf{X}_n\}_{n=1}^N$ to minimize the pairwise distances between the transformed views. An auxiliary variable $\mathbf{S}$ is introduced to represent the shared subspace among the views, and the optimization problem is formulated as:

$$\begin{aligned} \underset{\mathbf{W}_1,\ldots,\mathbf{W}_N,\mathbf{S}}{\text{minimize}} \quad & \sum_{n=1}^{N} \|\mathbf{S} - \mathbf{X}_n\mathbf{W}_n\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & \mathbf{S}^{\mathrm{T}}\mathbf{S} = \mathbf{I}_K. \end{aligned} \tag{7}$$

For the joint analysis of EEG signals from all subjects, the views $\mathbf{X}_n \in \mathbb{R}^{T \times D_xL_x}$ represent the EEG signals from different subjects, the matrices $\mathbf{W}_n \in \mathbb{R}^{D_xL_x \times K}$ are the per-subject decoders applied to these signals, and the shared subspace $\mathbf{S} \in \mathbb{R}^{T \times K}$ can be interpreted as the coherent EEG components across subjects. Other modalities in Table I can be analyzed in a similar way.

Denote the covariance matrix between two views $\mathbf{X}_i, \mathbf{X}_j$ as $\mathbf{R}_{ij}$. It can be shown that the solution to (7) can again be written as a GEVD problem similar to (5) [32]:

$$\mathbf{R}\mathbf{W} = \mathbf{D}\mathbf{W}\mathbf{\Lambda}, \tag{8}$$

where

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_N \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \cdots & \mathbf{R}_{1N} \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{N1} & \cdots & \mathbf{R}_{NN} \end{bmatrix},$$
$$\mathbf{D} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{R}_{NN} \end{bmatrix}. \tag{9}$$

The columns of $\mathbf{W}$ are the GEVCs corresponding to the K largest GEVLs. The shared subspace $\mathbf{S}$ can be obtained as the sum of the transformed views:

$$\mathbf{S} = \sum_{n=1}^{N} \mathbf{X}_n\mathbf{W}_n, \tag{10}$$

with scalings applied to each column of $\mathbf{S}$ to ensure that $\mathbf{S}^{\top}\mathbf{S} = \mathbf{I}_K$.

Analogous to PCCA, the effects of confounds such as eye movements can be removed from each view by regressing out the confounds and then applying GCCA to the residuals. When performing group-level analysis using GCCA, inter-subject correlation (ISC) can be used to assess the overall similarity of the extracted components across different views [33]. ISC is defined as the average pairwise correlation between all component pairs:

$$\mathrm{ISC}_k = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathrm{corr}(\mathbf{X}_i \mathbf{w}_i^k, \mathbf{X}_j \mathbf{w}_j^k), \quad (11)$$

where $\mathrm{corr}(\cdot,\cdot)$ denotes the Pearson correlation, and $\mathbf{w}_i^k$ and $\mathbf{w}_j^k$ are the k-th columns of $\mathbf{W}_i$ and $\mathbf{W}_j$, respectively.

F. Evaluation

The correlations obtained from (P)CCA provide insight into how well the data modalities and video features are temporally coupled. On the other hand, the ISCs calculated from GCCA indicate how well neural responses or eye movements are synchronized across subjects. In addition to these measures, we also perform two tasks to evaluate the decodability of attended objects, which is the primary focus of this study: a selective visual attention decoding (SVAD) task and a match-mismatch (MM) task.

The objective of both decoding tasks is to distinguish the attended video segment from an imposter using various data modalities. For example, with EEG data, the EEG decoders $\mathbf{W}_x$ and stimulus encoders $\mathbf{W}_y$ are trained on the EEG data and attended video features using (P)CCA. During the testing phase, the previously obtained encoders and decoders are applied to the held-out test data, and the target video segment is identified by selecting the video that shows the strongest correlation with the EEG across the CCA components. In both tasks, the video segment that should be identified is the attended one. The only difference lies in whether the imposter is the observed but unattended competing video segment (SVAD) or an unobserved non-competing segment from a different time point in the same test set (MM). The chance level for both tasks is 50%.

Intuitively, SVAD is more challenging because information from the unattended stimuli might also be decodable using the trained filters, making discrimination between the attended and unattended stimuli more difficult. In contrast, MM is easier since the imposter is not observed by the participant and can therefore not generate any correlated signal components in any of the recorded modalities. Therefore, evaluating both tasks together not only indicates how well the attended stimuli can be decoded but also provides insights into whether the analyzed data modality captures information from the unattended stimuli.

G. Practicalities

a) Cross-validation: The accuracies and correlations reported in the following sections are cross-validated using a leave-one-pair-out scheme. Specifically, data from one video pair are left out for testing (this includes both presentations of the same pair), while data from the remaining pairs are used for training. This process is repeated 7 times, corresponding to the 7 video pairs, and the results are averaged across all pairs. In the single-object dataset, the training set and test set have a fixed length of $24\,\mathrm{min}$ and $4\,\mathrm{min}$, respectively (2×2 min per pair). In the superimposed-object dataset, the training set has an average length of $33.2\,\mathrm{min}$ (STD = $2.5\,\mathrm{min}$), and the test set has an average length of $5.5\,\mathrm{min}$ (STD = $2.5\,\mathrm{min}$).

b) Time lags: The number of time lags in the CCA procedures are set in accordance with [17]. Specifically, the video feature encoders have $L_y = 15$ lags, capturing video features from approximately the past $0.5\,\mathrm{s}$ to the current time point. For EEG decoders, the (P)CCA model uses $L_x = 3$ lags centered around the current time point, spanning approximately from $-33\,\mathrm{ms}$ to $33\,\mathrm{ms}$, while the GCCA model employs $L_x = 5$ lags, covering approximately $-67\,\mathrm{ms}$ to $67\,\mathrm{ms}$. These numbers are also applied to other data modalities in Table I, as a grid search indicates that the results are not highly sensitive to the choice of the number of time lags.

c) Classifier: Multiple components can be extracted from (P)CCA, with each canonical component pair having a corresponding correlation value. To measure the "closeness" of the data to the video candidates, we sum the correlations of the first two component pairs, as these often show statistically significant correlation values [17]. The video segment with the higher score is then identified as the attended video. Although simple classifiers such as support vector machines or random forests can be applied to the complete set of obtained correlations, they do not yield significant improvements, and therefore the added complexity and increased risk for overfitting is not justified.

d) Statistical tests: We assess the significance of the correlations using a permutation test with phase scrambling. In phase scrambling, the phase components in the frequency domain are randomized to disrupt the temporal structure of the data while preserving the power spectrum [34]. A null distribution of correlations is generated by repeating the correlation analysis on the phase-scrambled data 500 times per fold. P-values are calculated as the proportion of correlations in the null distribution that are more extreme than the observed correlation (two-tailed). Note that this provides a relatively rigorous bound since the null distribution is computed from data with the same power spectrum. For comparing performance between different tasks or data modalities, we employ the Wilcoxon signed-rank test. When multiple comparisons are involved, p-values are adjusted using the Benjamini-Hochberg (BH) method [35]. Results are considered significant if the (adjusted) p-value is less than 0.05.

III. Results

A. Correlations are modulated by attention

A core assumption in our method is that the correlations between the video features and the collected data modalities are modulated by attention. If this assumption holds, the attended object can be identified by comparing the relative strengths of these correlations between the attended and unattended objects. In this experiment (using the superimposed-object

dataset), the encoders and decoders are trained on the *ObjFlow* feature of the attended object and each data modality using CCA. The correlations are then computed on the test set for the *ObjFlow* features of both the attended and unattended objects. The obtained correlation coefficients for the first two canonical components are shown in Fig. 3.

The first observation is that the correlations with attended features (in blue) are generally higher than those with unattended features (in orange), especially for the first canonical component. Therefore, we can conclude that correlations are modulated by attention, justifying the design of our classifier (Section II-G). Moreover, the correlations with unattended features are non-significant for all modalities, whereas for the attended case, the significance level is around or below the median for the modalities EEG, $GAZE_V$, and $SACC$. It is also worth noting that the correlations between different modalities are not directly comparable, as the significance levels differ due to the different spectral characteristics of each modality. Their performance can be better compared in specific tasks, as specified below.

B. Selective visual attention is decodable from EEG and eye movements

Since the assumption that correlations are modulated by attention holds, identifying the attended video segment based on these correlations appears to be feasible. To evaluate how well the attended video segment can be decoded from the data, we perform the SVAD task on the superimposed-object dataset as described in Section II-F and estimate the decoding accuracy using bootstrapping. Specifically, in each cross-validation fold, 30-second test segments are randomly sampled $2V_l$ times, where V_l is the length of the test set in seconds. The number of test segments is approximately 660 per fold on average. Over each $30\,\mathrm{s}$ test segment, we compute the correlation between the tested modality and the *ObjFlow* feature of both the attended and unattended object, and select the one exhibiting the highest correlation. The decoding accuracy is calculated as the proportion of times the attended video segment is correctly identified.

The results are shown in Fig. 4. Among the tested modalities, EEG, $GAZE_V$, and $SACC$ stand out with higher decoding accuracies, with approximately 75% of subjects reaching 60% accuracy or higher, and medians around 63.5%, 68.0%, and 67.0%, respectively. There is no significant difference in performance when comparing these three modalities.

Although the main focus of this study is EEG-based SVAD, it is noteworthy that this decoding can be achieved at least equally well using the gaze velocity or saccade information obtained from an eye tracker. The good performance of gaze velocity and saccade indicates that specific eye movement patterns elicited by the video stimuli can be informative for SVAD, even when the objects are spatially overlapping. The superior performance of gaze velocity amplitude over original gaze coordinates may be attributed to the fact that the *ObjFlow* feature is also based on (pixel) velocity magnitude. Additionally, we notice that EEG does not outperform gaze velocity and saccades, which raises an important question: does EEG-based decoding primarily rely on eye movement artefacts in the EEG recordings?

C. EEG-based decoding is not dominantly driven by eye movement artefacts

For some use cases, it may not be necessary to disentangle the effects of eye movements, as the primary goal is high decoding accuracy. However, since we also aim to enable novel experimental paradigms in neuroscience, where the focus is on neural responses, it is crucial to understand the role of eye movements in EEG-based decoding under free-viewing conditions. In this section, we suppress the effects of eye movement artefacts in the EEG (including saccades) in three ways: by regressing out eye movements (Section III-C.1), by using only EEG channels in the visual cortex (Section III-C.2), and by analyzing data free from saccades (Section III-C.3). The decoding accuracies are recomputed on the superimposed-object dataset under these three cases, collectively suggesting that EEG-based decoding is largely independent of eye movements.

1) Visual attention is decodable after regressing out eye movements: To control for the effects of eye movements, we apply PCCA (Section II-E.2) to correlate EEG signals with video features, setting the EOG and gaze velocity as confounds. More specifically, EOG and gaze velocity are concatenated along the channel axis and regressed out from both the EEG signals and the attended/unattended video features as in (6). CCA is then applied to the residuals to find the canonical components. Note that saccade information, which is encoded in gaze velocity as sudden changes in coordinates leading to peaks in velocity (Section II-C), is also implicitly suppressed after regression. The accuracies before and after controlling for eye movements are shown in Fig. 5. After regression, the attended video segment can still be decoded with a median accuracy of around 61.6%, which is only 1.9% lower than before. This result suggests that while eye movement information in EEG can assist SVAD, it does not primarily drive the decoding performance.

However, a limitation of the above analysis is that the "eye movement information" considered here includes only EOG and gaze velocity, and not all information related to eye movements. For instance, there could be other nonlinear transformations of the gaze coordinates that correlate with visual stimuli and EEG signals, thereby boosting the decoding performance. Another way to disentangle or reduce the influence of eye movements is to use EEG channels that are less affected by eye movements, which is discussed in the next section.

2) Visual attention is decodable using channels in visual cortex: It is known that eye movements primarily affect EEG channels in the frontal region, with the strength of these artefacts decreasing as they propagate towards the back of the head [36]. Therefore, if EEG-based decoding is mainly driven by eye movements, the decoding performance when using channels in the frontal region should be better compared to other regions, especially the parietal-occipital region, where the visual cortex is located, being the furthest away from

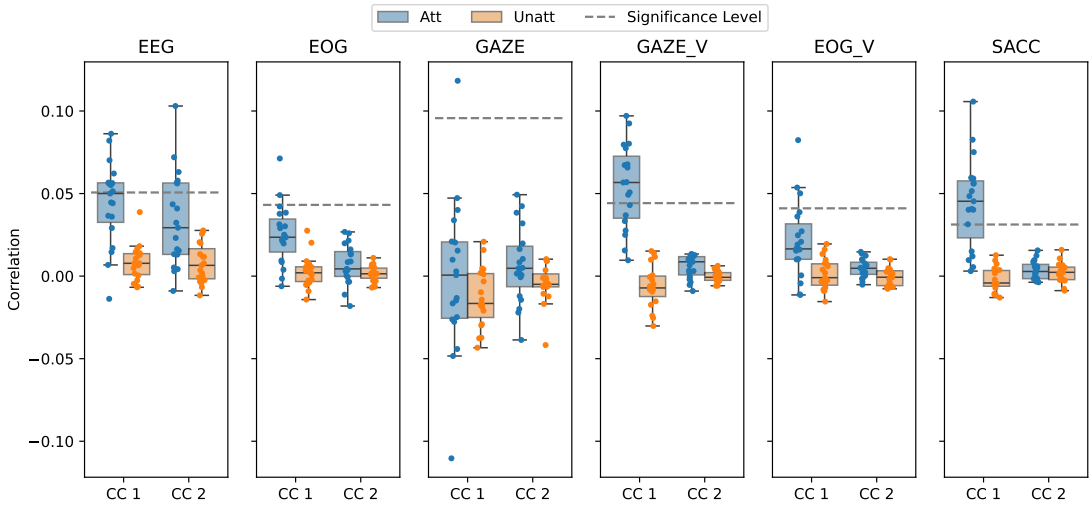


Fig. 3: The correlations of each data modality with the *ObjFlow* features of the attended and unattended object in the superimposed-object dataset are shown in blue and orange, respectively. "CC 1" and "CC 2" denote the first and second canonical components. The dots represent the mean correlations of individual subjects across folds. The boxes indicate the median and the interquartile range, while the whiskers extend to the most extreme data points not considered outliers. The dotted lines represent the significance level pooled across subjects and components.

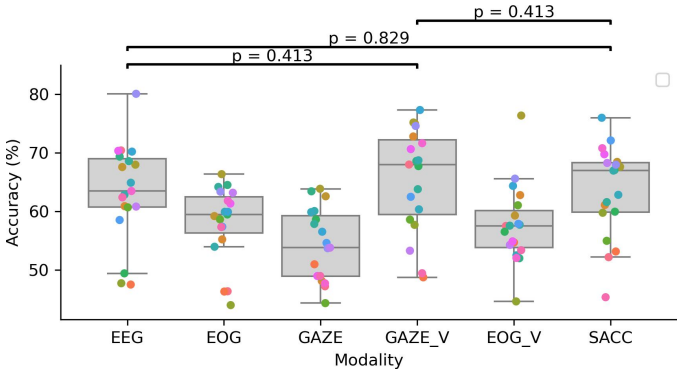


Fig. 4: Accuracies of the SVAD task, i.e., identifying the attended video segment from the unattended video segment using data segments from different modalities. The models are trained and tested on the superimposed-object dataset and the test segments are 30 s long. The dots in different colors represent the accuracies of individual subjects. The boxes show the median and the interquartile range, with the whiskers extending to the most extreme data points. Wilcoxon signed-rank tests are performed to determine if the distributions are significantly different, and the p-values are indicated on the figure (BH-adjusted).

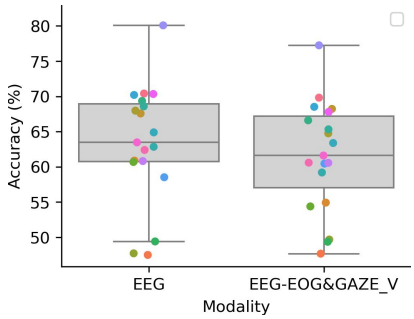
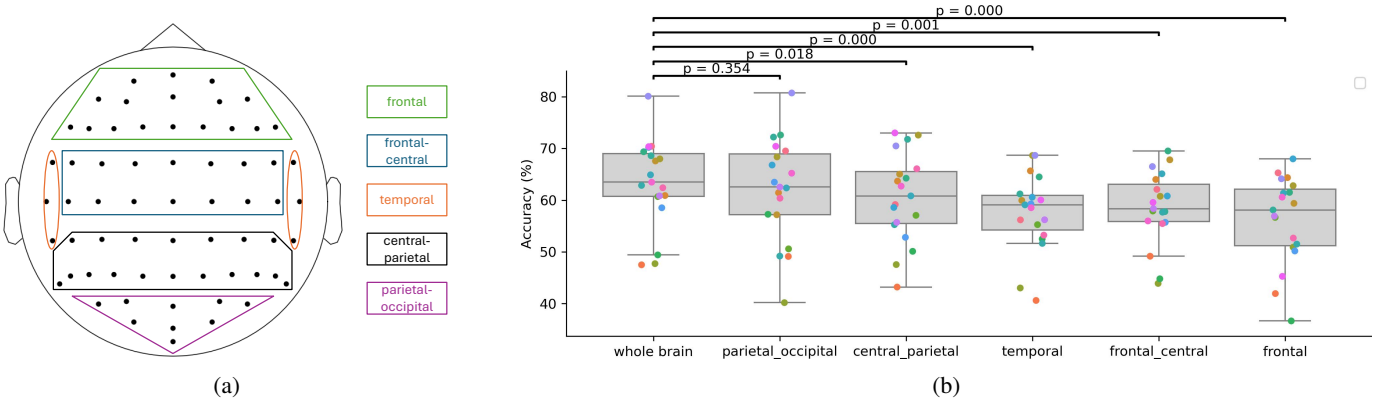


Fig. 5: Accuracies of the SVAD task before and after regressing out EOG and gaze velocity from EEG. The latter is denoted by "$EEG - EOG\&GAZE_V$". The models are trained and tested on the superimposed-object dataset and the test segments are 30 s long. The dots denote the individual (per-subject) accuracies, and the boxes show the median and the interquartile range.

the eyes. To test this hypothesis, we divide EEG channels into different groups based on their locations (Fig. 6a), and perform the SVAD task using each group. The results are presented in Fig. 6b. Contrary to the hypothesis, the decoding accuracy using channels in the parietal-occipital region is comparable to the performance using whole-brain signals, while the accuracy for all other regions drops significantly. This result suggests that EEG-based decoding is primarily driven by neural responses in the visual cortex rather than eye movements.

We also conduct an exploratory analysis to reduce the number of channels in the visual cortex to 4 without significant performance loss, whereas randomly selecting 4 channels from other regions results in significantly worse performance (Supplementary Material, Section I). This result not only supports the previous conclusion but also demonstrates that effective EEG-based decoding can be achieved with much fewer channels without compromising performance, which is advantageous for practical applications.

3) Visual attention is decodable after removing saccades: Saccades also elicit neural responses, which have been found to be dominant across the entire brain under a free-viewing setup similar to ours [24]. As mentioned in Section III-C.1, the effect of saccades is suppressed after regressing out EOG and gaze velocity from EEG and video features. However, a safer option is to remove the data segments around saccades



(a) (b)

Fig. 6: (a) Topographic maps of the EEG channels in different brain regions. (b) Accuracies of the SVAD task using EEG channels in different brain regions. The models are trained and tested on the superimposed-object dataset and the test segments are 30 s long. The dots denote the individual (per-subject) accuracies, and the boxes show the median and the interquartile range. Wilcoxon signed-rank tests are performed to determine if using whole-brain signals is significantly better than using signals from a specific region. The p-values are BH-adjusted.

and analyze the remaining data, as regression might not fully eliminate the event-related potentials elicited by saccades.

In our experiment, we remove data points from 0.33 s before to 1 s after the saccade onset, resulting in a data loss ranging from 23% to 82%, depending on the subject. To ensure sufficient data for training and testing, we train the CCA decoders on the single-object dataset in a subject-independent manner (i.e. concatenating the data from all subjects) and test on the superimposed-object dataset for individual subjects. For a fair comparison, we create control groups by randomly removing the same amount of data not necessarily around saccades. Wilcoxon signed-rank tests are performed to determine if the performance after removing data around saccades is significantly worse than the control groups. The BH-adjusted p-values are all above 0.05 (0.276, 0.410, 0.252, 0.225, 0.225, 0.252, 0.414, 0.225, 0.225, 0.225), indicating no strong evidence that EEG-based decoding is primarily driven by saccades.

D. EEG also captures information of the unattended object

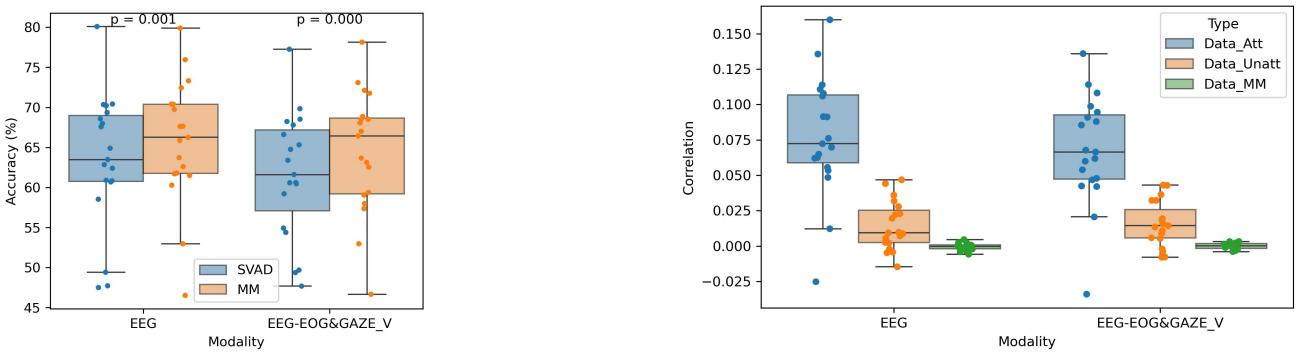
In Section III-A, we observe that the correlations with the unattended object are not only lower but also non-significant for all data modalities. This is remarkable, especially for EEG, since the unattended object is present in the same location in the visual field as the attended object. This suggests that the brain is able to separate the visual streams of both objects and suppress one of them in favor of the other. Since the correlation with the unattended object is not significant, the question remains whether the EEG actually contains signal components that encode the unattended object and whether these can be captured by the CCA model. To address this question, we conduct the match-mismatch (MM) task (Section II-F), where the attended video segment remains the same as in SVAD, and the unattended video segment is an unobserved segment at a random time point in the same test set. We apply the same bootstrapping and cross-validation procedure, with video segment lengths still set to 30 s, and compare the decoding accuracies of the SVAD and MM tasks on the superimposed-object dataset.

The results are shown in Fig. 7a. Both before and after regressing out eye movements, the EEG-based decoding accuracies of the MM task are significantly higher than those of the SVAD task, with medians around 67.0% and 65.3%, respectively. This indicates that EEG captures information about the unattended video, which confuses the SVAD decision. Further evidence is provided by the results in Fig. 7b, which show the sum of the first two canonical correlations between EEG (with or without eye movement regressed out) and the *ObjFlow* feature of the attended, unattended, and mismatch object. The significant difference between the correlations with the unattended object versus the mismatch object implies that the model also extracts responses correlated with the unattended object. Note that for eye-related data modalities, the performance of the MM task is comparable to or even worse than that of the SVAD task, suggesting that the unattended object is hardly captured by these modalities (Supplementary Material, Section II).

E. Complementary information exists in EEG and gaze features

In Section III-C.1, we have demonstrated that regressing out gaze velocity (and EOG) from EEG signals does not drastically affect decoding performance. Therefore, it is reasonable to assume that these two data modalities capture complementary information. This raises a natural question: can combining them improve decoding performance? The combination can be achieved by simply concatenating gaze velocity to EEG signals as an extra channel. We then apply the decoding pipeline to the combined data and compare the performance with using EEG and gaze velocity separately. The accuracies of the two tasks on the superimposed-object dataset using EEG, gaze velocity, and their combination are shown in Fig. 8.

In the SVAD task, using combined modalities significantly outperforms using EEG alone but not gaze velocity alone,



(a) (b)

Fig. 7: (a) Accuracies of the SVAD and MM tasks. Wilcoxon signed-rank tests are performed to determine if the accuracies of SVAD tasks are significantly lower than that of MM tasks using EEG. The p-values are BH-adjusted. (b) The sum of first two canonical correlations between each modality and the *ObjFlow* feature of the attended object (in blue), the unattended object (in orange) and the mismatch object (in green). The models are trained and tested on the superimposed-object dataset and the test segments are 30 s long. The dots denote the per-subject results averaged across all trials, and the boxes show the median and the interquartile range.

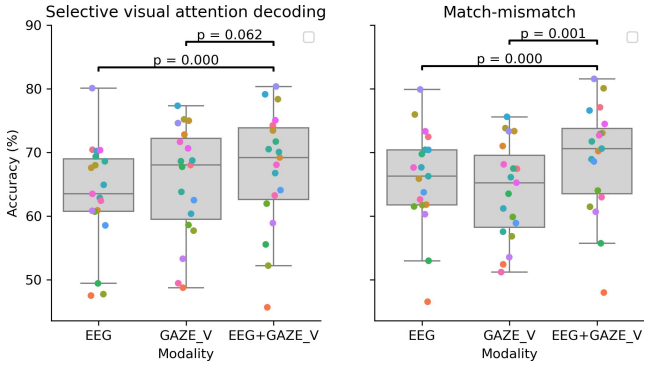


Fig. 8: Accuracies of the SVAD and MM tasks using EEG, gaze velocity, and their combination. The models are trained and tested on the superimposed-object dataset and the test segments are 30 s long. The dots denote the individual (per-subject) accuracies, and the boxes show the median and the interquartile range. Wilcoxon signed-rank tests are performed to determine if the accuracies of using EEG and gaze velocity separately are significantly lower than that of using them together. The p-values are BH-adjusted.

with a median accuracy around 68.9%. In the MM task, the performance of using combined modalities is significantly higher than using each of them separately, with a median accuracy around 70.5%. These results suggest that the information captured by EEG and gaze velocity is complementary and can lead to better decoding performance, especially in the MM task. However, for the SVAD task, the additional information from EEG may not be as discriminative as the information in gaze velocity (as already explained in Section III-D), leading to limited improvement.

F. Synchronization among subjects decreases in the presence of a distractor

In the previous sections, we have focused on stimulus-aware individual-level analysis, correlating video features with data modalities and identifying the attended video segment. Now, we shift our focus to group-level analysis, which bypasses video feature extraction, quantifies the synchronization level among participants, and provides insights into group attention or engagement [33], [37], [32]. Specifically, we apply GCCA (Section II-E.3) to each data modality and compute the inter-subject correlation (ISC) for the first canonical component for both datasets: single-object and superimposed-object. The ISCs are cross-validated using a leave-one-pair-out scheme.

The results for each fold are presented in Fig. 9 [1], from which we can observe that the ISCs of EEG (with eye movements regressed out) are significantly lower in the superimposed-object dataset. A decreasing trend in ISCs is also evident for the other modalities, although this decrease is not statistically significant, potentially due to the limited sample size. This indicates that synchronization among subjects decreases when a distractor is present. A possible explanation is that subjects may be distracted by the unattended object and this distraction can happen at different points in time for different subjects, leading to a reduced synchronicity. This poses a challenge for stimulus-unaware ISC-based measurements of attention when viewing naturalistic videos: the attention of the participants is more scattered in the presence of multiple objects, and lower ISCs do not necessarily imply lower absolute attention levels to the overall stimuli. For example, in a tennis match recording, participants might focus on different players at different times, resulting in lower ISCs even if their attention levels are high. Another interesting

[1]Note that ISCs should not be compared across modalities as they heavily depend on the spectral characteristics and signal-to-noise ratio of the underlying signals.

observation is that despite the high synchronization of EOG and gaze coordinates across subjects, they do not perform well in the SVAD and MM tasks (Fig. 7a), whose performance depends more on the correlation between the data and the extracted video features.

IV. Discussion

A. Is the ground truth reliable?

In this study, we have assumed that participants always follow the instruction, and we use the object they are asked to attend to as the ground truth in the SVAD task. Although this may not always be the case, we expect the ground truth to remain reliable assuming participants only occasionally attend to the distractor.

Additional evidence for this assumption can be found in Fig. 10. Here, we repeat the analysis described in Section III-B for the best three modalities where the CCA decoders are this time trained on the single-object dataset, for which the ground truth is certain due to the absence of a distracting object. From Fig. 10, we conclude that the impact is relatively mild; for the $GAZE_V$ and $SACC$ modalities the difference is not significant, and in the case of EEG, training with the single-object data actually leads to significantly worse results, despite the availability of an exact ground truth. In this case, training with superimposed objects results in higher accuracies, which would be unlikely if the ground truth in this data set would be unreliable.

The decrease in performance when training the decoders on the single-object data in the case of EEG might be explained by the fact that the decoder can not learn to suppress neural responses to the unattended object, as these responses are not present in the training set. Another possible explanation is the fact that the task of attending to a target object in the superimposed videos is more challenging, which could result in stronger neural responses, which are more easily decodable. A similar effect has been described in the context of selective attention decoding with speech stimuli, where more difficult tasks, i.e., in more challenging acoustic conditions, can result in better decoding accuracies [38], [39].

B. Eye tracker or EEG

In selective visual attention decoding, eye trackers are perhaps a more straightforward and popular choice. Their advantage in overt attention decoding is evident as they directly provide gaze information with high spatial and temporal resolution. Therefore, gaze maps exported from eye trackers are usually considered the ground truth for attention in many fields such as video saliency prediction [40], neuromarketing [41], and cognitive workload measurement [42]. Additionally, eye trackers have been found useful in quantifying the absolute attention level. For example, in [37], ISCs of gaze and pupil size were used as markers of attention and were predictive of students' test scores on a group-level.

A limitation of eye trackers is that they cannot measure covert attention, which can be problematic when participants attend to objects in their peripheral vision without moving their eyes. However, this is less of a concern in free-viewing scenarios where overt attention is more prevalent. Eye trackers may also struggle when objects are close to each other, as the gaze coordinates might be ambiguous. Nevertheless, in this study we demonstrated that gaze velocity and saccades can still be informative even when objects are spatially overlapped, as long as the objects have distinct motion patterns.

Another limitation of eye trackers in the context of (selective) attention decoding is that eye movements are an indirect measure of attention since there is a gap between merely "looking at" and "paying attention", whereas EEG directly captures neural responses to stimuli, allowing for arguably more reliable inference of attention, independent of gaze patterns. This could be particularly useful when one aims to decode not only the attended object but also how attentive the participant is to the object. Take the results of Subject 11 and 17 (Table II) as an example: the gaze velocity-based decoding accuracy is comparable for both subjects, but the EEG-based decoding accuracy (with eye movements regressed out) is 9% lower for Subject 11. Additionally, the EEG signals of Subject 11 exhibit higher correlations with the unattended object. This suggests that Subject 11 may track the attended object similarly to Subject 17 but is less successful in suppressing the distractor that spatially overlaps with the target object. Together with the fact that the correlation between EEG and the attended object is also lower for Subject 11, this indicates that Subject 11 might be less attentive overall than Subject 17.

TABLE II: Comparison of accuracies between gaze velocity-based and EEG-based decoding (with eye movements regressed out) in the SVAD task on the superimposed-object dataset for two selected subjects. The sum of the first two canonical correlations between EEG and attended video features, and EEG and unattended video features are also shown.

Subject ID	Accuracy (SVAD, %)		Correlation	
	Gaze Velocity-Based	EEG-Based	EEG-Att	EEG-Unatt
11	70.3	59.3	0.058	0.041
17	70.7	68.3	0.097	0.011

Neural-based decoding can also be advantageous for understanding the underlying neural mechanisms of selective attention, such as the timing of attentional effects and sources of attentional control signals [14], [16], [43], although EEG-based paradigms are not yet prevalent. Therefore, although eye trackers seem to dominate visual attention decoding in human-computer interaction, studying EEG-based paradigms still has unique assets that offer additional possibilities.

C. Gaze-informed attention decoding

In Section III-E, we have combined EEG and gaze velocity by concatenating them along the channel axis. Another way to incorporate gaze information is during video feature extraction, by weighting or selecting video features based on gaze coordinates. The goal is then not only to identify the attended object but also to measure the absolute attention level to the object in the gaze direction. This approach is motivated by the selective attention mechanism: we assume

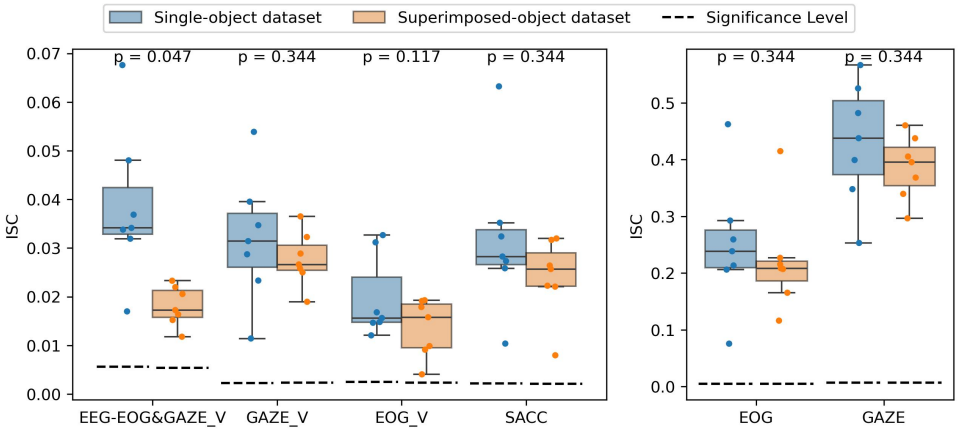


Fig. 9: Inter-subject correlations (ISCs) of the first canonical component across different data modalities for the single-object dataset (blue) and the superimposed-object dataset (orange). The ISCs for EOG and $GAZE$ are plotted separately because their correlation values are much higher due to their particular signal characteristics, i.e., they are much lower in frequency and are approximately piecewise constant. Each dot represents the ISC for an individual fold, and the boxes display the median and interquartile range. Wilcoxon signed-rank tests are performed to assess whether ISCs in the single-object dataset are significantly higher than those in the superimposed-object dataset, with p-values adjusted using the BH procedure.

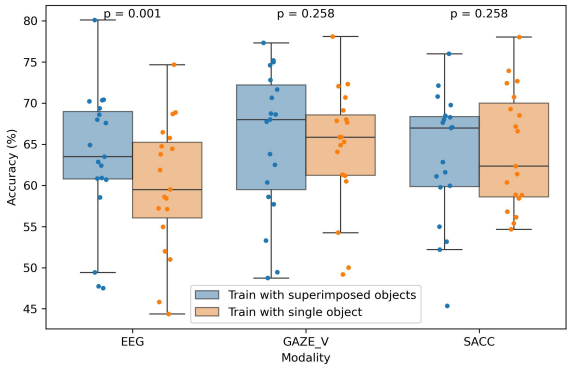


Fig. 10: Accuracies of the SVAD task using decoders trained in the superimposed-object dataset (in blue) and the single-object dataset (in orange). The test segments are 30 s long. The dots denote the individual (per-subject) accuracies, and the boxes show the median and the interquartile range. Two-sided Wilcoxon signed-rank tests are performed and the p-values are BH-adjusted.

participants direct their gaze to the object of interest in free-viewing scenarios, and features around the gaze coordinates should be emphasized to find strong correlations with EEG signals, as attended features are enhanced in the brain. This is particularly useful when multiple objects are interacting in the scene and each participant's attention is unknown beforehand. A region defined based on the gaze map could potentially replace the bounding boxes of objects in the object-based features proposed in [17], circumventing the need for feature fusion in multi-object scenarios.

A similar idea has been applied in a neural encoding model, where the goal was to predict the shared fMRI responses across subjects elicited by natural videos [44]. The authors demonstrated that using gaze fixation maps as attention maps to weigh the video features significantly improved the prediction. They further proposed a trainable attention module that could learn the attention maps end-to-end, ultimately generating attention maps similar to the gaze fixation maps. For video-EEG encoding and decoding, the neural signals are noisier and the availability of effective video features is limited, making it more attractive to incorporate gaze information.

D. Gaze-driven EEG components versus stimulus-driven neural responses

Returning to the proposed EEG-based decoder, although the analysis in Section III-C shows that the decoding performance is unlikely to be driven by eye movements, we cannot make definitive arguments as not all confounds are completely removed. Methods of eye movement artifact removal, such as the linear regression in our PCCA procedure, only suppress the artefacts without necessarily fully eliminating them, and the residuals might still affect decoding performance. Analyzing only EEG channels in the occipital-parietal region also mitigates ocular contamination, but neural responses elicited by saccades are still present in that region [24]. Cutting out segments around saccades appears to be a reliable way of obtaining "clean" data, but the resulting discontinuities in the signals might introduce new confounds. Additionally, it is difficult to disentangle the EEG signals related to the motor control of eye movements.

In essence, eye movement is a complex process that involves multiple brain regions and is closely linked to attention. Consequently, it is not feasible to fully disentangle all related confounds. While the results of this study suggest that EEG-based decoding is not dominantly driven by eye movements, we acknowledge that, depending on the research question and application area, gaze fixation may be necessary to better disentangle the effects of eye movements from EEG signals,

as opposed to the free-viewing paradigm that was used in this study.

V. Conclusion

In this paper, we have taken a step towards more natural experimental paradigms for selective visual attention decoding by transitioning from static images to dynamic videos, where participants attend to one of the objects while freely viewing. We demonstrate that it is possible to decode the object of interest from EEG signals, even when using only visual cortex channels and when the two objects are co-located. Apart from EEG, eye gaze data have also been used to decode attention. Although the attended and unattended objects are superimposed, the target is still decodable from gaze data since the gaze velocity and saccades are related to the movement pattern of the attended object.

To better understand the role of eye movements in EEG-based decoding, we have conducted a series of experiments to disentangle possible eye gaze confounds in the EEG signals. The results indicate that EEG-based decoding is not dominantly driven by eye movements. We have also demonstrated that EEG captures information about both the attended and unattended objects, which makes the EEG-based decoder less discriminative. This finding may explain why adding EEG to gaze data does not significantly improve SVAD performance, despite the existence of complementary information. Furthermore, group-level analysis reveals that the participants' attention is more scattered when a distractor is present, making stimulus-unaware group-level attention metrics such as ISC less reliable with increased stimulus complexity.

References

[1] M. Carrasco, "Visual attention: The past 25 years," *Vision research*, vol. 51, no. 13, pp. 1484–1525, 2011.

[2] S. Kelly, E. Lalor, C. Finucane, G. McDarby, and R. Reilly, "Visual spatial attention control in an independent brain-computer interface," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 9, pp. 1588–1596, Sep. 2005.

[3] C. Reichert, I. F. Tellez Ceja, C. M. Sweeney-Reed, H.-J. Heinze, H. Hinrichs, and S. Dürschmid, "Impact of stimulus features on the performance of a gaze-independent brain-computer interface based on covert spatial attention shifts," *Frontiers in Neuroscience*, vol. 14, p. 591777, 2020.

[4] D. Li, X. Luo, J. Guo, Y. Kong, Y. Hu, Y. Chen, Y. Zhu, Y. Wang, L. Sun, and Y. Song, "Information-based multivariate decoding reveals imprecise neural encoding in children with attention deficit hyperactivity disorder during visual selective attention," *Human Brain Mapping*, vol. 44, no. 3, pp. 937–947, 2023.

[5] R. Abiri, S. Borhani, Y. Jiang, and X. Zhao, "Decoding attentional state to faces and scenes using EEG brainwaves," *Complexity*, vol. 2019, no. 1, p. 6862031, 2019.

[6] M. M. Monti, J. D. Pickard, and A. M. Owen, "Visual cognition in disorders of consciousness: From V1 to top-down attention," *Human Brain Mapping*, vol. 34, no. 6, pp. 1245–1253, 2013.

[7] E. Astrand, C. Wardak, and S. Ben Hamed, "Selective visual attention to drive cognitive brain–machine interfaces: from concepts to neurofeedback and rehabilitation applications," *Frontiers in Systems Neuroscience*, vol. 8, Aug. 2014.

[8] M. T. Debettencourt, J. D. Cohen, R. F. Lee, K. A. Norman, and N. B. Turk-Browne, "Closed-loop training of attention with real-time brain imaging," *Nature neuroscience*, vol. 18, no. 3, pp. 470–475, 2015.

[9] C. Ozcinar, J. Cabrera, and A. Smolic, "Visual Attention-Aware Omnidirectional Video Streaming Using Optimal Tiles for Virtual Reality," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 217–230, Mar. 2019.

[10] J. Moran and R. Desimone, "Selective attention gates visual processing in the extrastriate cortex," *Science*, vol. 229, no. 4715, pp. 782–784, 1985.

[11] S. Yantis and J. T. Serences, "Cortical mechanisms of space-based and object-based attentional control," *Current opinion in neurobiology*, vol. 13, no. 2, pp. 187–193, 2003.

[12] F. Tong and M. S. Pratte, "Decoding patterns of human brain activity," *Annual review of psychology*, vol. 63, pp. 483–509, 2012.

[13] A. M. Niazi, P. L. C. van den Broek, S. Klanke, M. Barth, M. Poel, P. Desain, and M. A. J. van Gerven, "Online decoding of object-based attention using real-time fMRI," *European Journal of Neuroscience*, vol. 39, no. 2, pp. 319–329, 2014.

[14] A. S. Keller, A. V. Jagadeesh, L. Bugatus, L. M. Williams, and K. Grill-Spector, "Attention enhances category representations across the brain with strengthened residual correlations to ventral temporal cortex," *NeuroImage*, vol. 249, p. 118900, Apr. 2022.

[15] T. Horikawa and Y. Kamitani, "Attention modulates neural representation to render reconstructions according to subjective appearance," *Communications Biology*, vol. 5, no. 1, pp. 1–12, Jan. 2022.

[16] T. Grootswagers, A. K. Robinson, S. M. Shatek, and T. A. Carlson, "The neural dynamics underlying prioritisation of task-relevant information," *Neurons, Behavior, Data analysis, and Theory*, vol. 5, no. 1, Feb. 2021.

[17] Y. Yao, A. Stebner, T. Tuytelaars, S. Geirnaert, and A. Bertrand, "Identifying temporal correlations between natural single-shot videos and EEG signals," *Journal of Neural Engineering*, vol. 21, no. 1, p. 016018, 2024.

[18] M. Gavaret, A. Iftimovici, and E. Pruvost-Robieux, "EEG: Current relevance and promising quantitative analyses," *Revue Neurologique*, vol. 179, no. 4, pp. 352–360, 2023.

[19] N. Jamil, A. N. Belkacem, S. Ouhbi, and C. Guger, "Cognitive and affective brain–computer interfaces for improving learning strategies and enhancing student capabilities: A systematic literature review," *Ieee Access*, vol. 9, pp. 134 122–134 147, 2021.

[20] H. Si-Mohammed, J. Petit, C. Jeunet, F. Argelaguet, F. Spindler, A. Evain, N. Roussel, G. Casiez, and A. Lécuyer, "Towards BCI-based interfaces for augmented reality: feasibility, design and evaluation," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 3, pp. 1608–1621, 2018.

[21] R. Li, J. S. Johansen, H. Ahmed, T. V. Ilyevsky, R. B. Wilbur, H. M. Bharadwaj, and J. M. Siskind, "The perils and pitfalls of block design for eeg classification experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 316–333, 2020.

[22] I. Rotaru, S. Geirnaert, N. Heintz, I. Van de Ryck, A. Bertrand, and T. Francart, "What are we really decoding? Unveiling biases in EEG-based decoding of the spatial focus of auditory attention," *Journal of Neural Engineering*, vol. 21, no. 1, p. 016017, 2024.

[23] A. Herbec, J.-P. Kauppi, C. Jola, J. Tohka, and F. E. Pollick, "Differences in fMRI intersubject correlation while viewing unedited and edited videos of dance performance," *Cortex*, vol. 71, pp. 341–348, Oct. 2015.

[24] M. Nentwich, M. Leszczynski, B. E. Russ, L. Hirsch, N. Markowitz, K. Sapru, C. E. Schroeder, A. D. Mehta, S. Bickel, and L. C. Parra, "Semantic novelty modulates neural responses to visual change across the human brain," *Nature Communications*, vol. 14, no. 1, p. 2910, 2023.

[25] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, Feb. 2020.

[26] G. Farnebäck, "Two-Frame Motion Estimation Based on Polynomial Expansion," in *Image Analysis*, G. Goos, J. Hartmanis, J. Van Leeuwen, J. Bigun, and T. Gustavsson, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, vol. 2749, pp. 363–370.

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

[28] E. B. Corrochano, T. De Bie, N. Cristianini, and R. Rosipal, "Eigenproblems in pattern recognition," *Handbook of Geometric Computing: Applications in Pattern Recognition, Computer Vision, Neuralcomputing, and Robotics*, pp. 129–167, 2005.

[29] B. R. Rao, "Partial canonical correlations," *Trabajos de estadistica y de investigación operativa*, vol. 20, pp. 211–219, 1969.

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

[31] X. Fu, K. Huang, E. E. Papalexakis, H.-A. Song, P. P. Talukdar, N. D. Sidiropoulos, C. Faloutsos, and T. Mitchell, "Efficient and distributed algorithms for large-scale generalized canonical correlations analysis,"

in *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 2016, pp. 871–876.

[32] S. Geirnaert, T. Francart, and A. Bertrand, "Stimulus-Informed Generalized Canonical Correlation Analysis of Stimulus-Following Brain Responses," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023.

[33] J. P. Dmochowski, P. Sajda, J. Dias, and L. C. Parra, "Correlated components of ongoing EEG point to emotionally laden attention–a possible marker of engagement?" *Frontiers in human neuroscience*, vol. 6, p. 112, 2012.

[34] D. Prichard and J. Theiler, "Generating surrogate data for time series with several simultaneously measured variables," *Physical review letters*, vol. 73, no. 7, p. 951, 1994.

[35] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, pp. 1165–1188, 2001.

[36] S. Romero, M. A. Mañanas, and M. J. Barbanoj, "A comparative study of automatic techniques for ocular artifact reduction in spontaneous EEG signals based on clinical target variables: a simulation case," *Computers in biology and medicine*, vol. 38, no. 3, pp. 348–360, 2008.

[37] J. Madsen, S. U. Júlio, P. J. Gucik, R. Steinberg, and L. C. Parra, "Synchronized eye movements predict test scores in online video education," *Proceedings of the National Academy of Sciences*, vol. 118, no. 5, p. e2016980118, 2021.

[38] N. Das, W. Biesmans, A. Bertrand, and T. Francart, "The effect of head-related filtering and ear-specific decoding bias on auditory attention detection," *Journal of neural engineering*, vol. 13, no. 5, p. 056014, 2016.

[39] N. Das, A. Bertrand, and T. Francart, "Eeg-based auditory attention detection: boundary conditions for background noise and speaker positions," *Journal of neural engineering*, vol. 15, no. 6, p. 066017, 2018.

[40] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "Deepvs: A deep learning based video saliency prediction approach," in *Proceedings of the european conference on computer vision (eccv)*, 2018, pp. 602–617.

[41] R. d. O. J. dos Santos, J. H. C. de Oliveira, J. B. Rocha, and J. d. M. E. Giraldi, "Eye tracking in neuromarketing: a research agenda for marketing studies," *International journal of psychological studies*, vol. 7, no. 1, p. 32, 2015.

[42] J. Zagermann, U. Pfeil, and H. Reiterer, "Measuring cognitive load using eye tracking technology in visual computing," in *Proceedings of the sixth workshop on beyond time and errors on novel evaluation methods for visualization*, 2016, pp. 78–85.

[43] E. Goddard, T. A. Carlson, and A. Woolgar, "Spatial and feature-selective attention have distinct, interacting effects on population-level tuning," *Journal of cognitive neuroscience*, vol. 34, no. 2, pp. 290–312, 2022.

[44] M. Khosla, G. Ngo, K. Jamison, A. Kuceyeski, and M. Sabuncu, "Neural encoding with visual attention," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 942–15 953, 2020.