Multimodal Detection of External and Internal Attention in Virtual Reality using EEG and Eye Tracking Features

Xingyu Long
xingyu.long@univie.ac.at
LMU Munich
Munich, Germany
University of Vienna
Vienna, Austria

Sven Mayer
info@sven-mayer.com
LMU Munich
Munich, Germany

Francesco Chiossi
francesco.chiossi@um.ifi.lmu.de
LMU Munich
Munich, Germany

ABSTRACT

Future VR environments will sense users' context, enabling a wide range of intelligent interactions, thus enabling diverse applications and improving usability through attention-aware VR systems. However, attention-aware VR systems based on EEG data suffer from long training periods, hindering generalizability and widespread adoption. At the same time, there remains a gap in research regarding which physiological features (EEG and eye tracking) are most effective for decoding attention direction in the VR paradigm. We addressed this issue by evaluating several classification models using EEG and eye tracking data. We recorded that training data simultaneously during tasks that required internal attention in an N-Back task or external attention allocation in Visual Monitoring. We used linear and deep learning models to compare classification performance under several uni- and multimodal feature sets alongside different window sizes. Our results indicate that multimodal features improve prediction for classical and modern classification models. We discuss approaches to assess the importance of physiological features and achieve automatic, robust, and individualized feature selection.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Virtual Reality, Attention, EEG, Eye Tracking, Physiological Computing, Machine Learning

ACM Reference Format:
Xingyu Long, Sven Mayer, and Francesco Chiossi. 2024. Multimodal Detection of External and Internal Attention in Virtual Reality using EEG and Eye Tracking Features. In *Proceedings of Mensch und Computer 2024 (MuC '24), September 1–4, 2024, Karlsruhe, Germany*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3670653.3670657

MuC '24, September 1–4, 2024, Karlsruhe, Germany

ACM ISBN 979-8-4007-0998-2/24/09
https://doi.org/10.1145/3670653.3670657

1 INTRODUCTION

Recent advancements in Virtual Reality (VR) technology have expanded its applications across numerous fields, such as gaming [27], healthcare [56], and training [28]. With the evolution of VR, there's a growing emphasis on creating adaptive systems capable of intelligently responding to user states in real-time [8]. This shift towards personalized and dynamic experiences aims to enrich VR interactions significantly [21]. The importance of adaptive interaction in delivering compelling VR experiences has become increasingly recognized. Unlike traditional VR, which often relied on fixed interactions and scenarios, adaptive systems promise a higher degree of personalization and applicability by adjusting to users' affective [12], attentional [16], and engagement states [19]. Physiological computing emerges as a key approach for enabling these adaptive VR experiences [31]. It utilizes human physiological signals as interactive inputs, providing insights into users' cognitive and affective states and allowing for the customization of the VR experience to meet their immediate needs and objectives. In VR environments, users encounter diverse stimuli requiring both external attention, where attentional resources are allocated to the outer environment (e.g., processing visual information), and internal attention, where resources are directed inward to internal representations of information, e.g. for tasks such as memory recall [67] and mental arithmetic [4]. The distinction between internal and external attention is crucial for various VR tasks [51, 60, 61], as attention levels may vary due to internal and external factors, impacting the quality of the interaction. Attentional mechanisms are essential for prioritizing relevant information and filtering out distractions. VR's controlled settings allow for the adaptation of content to manage and mitigate potentially distracting elements, aligning virtual content with the user's attentional state and the application's goals. In cases where users might be internally focused, adaptations can enhance the saliency of external stimuli to help maintain task focus.

While there is previous work in Augmented Reality (AR) [59, 71, 73] showing that attention decoding is possible and that adaptive systems can be designed, there is a lack of evaluating detection of attentional fluctuations in VR. A recent review from Nwagu et al. [53] highlights the advancements in EEG-based brain–computer interfaces (BCI) in VR are lacking hybrid interaction based on multimodal data and that most applications for BCIs are mostly limited to biofeedback [62] and in-game difficulty adjustments rather than user state detection [79].

To address this gap, our study explores machine learning techniques to differentiate between internal and external attention

states implicitly, leveraging electroencephalographic (EEG) and eye tracking data. Directly asking participants to identify their focus of attention can prove difficult, given these states' often subconscious or fluctuating nature. Recognizing the limitations of relying solely on participant recall, which is subject to human error and memory lapses, we investigate the potential for automatic detection of attention states to overcome these obstacles. We chose EEG and eye tracking based on their proven capabilities in prior research. EEG is particularly valued for its ability to detect alpha and theta activity changes, which reflect shifts between internal and external attention by showing increases during internal focus and decreases during engagement with external tasks [13, 23]. Eye tracking complements this by providing real-time, behavioral indicators of attentional direction, such as changes in gaze patterns and eye closures associated with internal attention [5]. Together, these modalities offer a multi-faceted view of attention, combining the depth of EEG with the behavioral insights of eye tracking to provide a robust framework for attention decoding in dynamic settings like VR [59, 69, 73]. Our investigation involved a dual strategy, incorporating feature engineering based on established protocols and automatic feature extraction with deep neural networks. Our study went beyond merely presenting the model prediction results by attempting to interpret the learned parameters of our machine-learning models.

Our contribution is fivefold: First, we introduce benchmark tasks for investigating external and internal attention in VR (I). Second, we demonstrate that combining EEG and eye tracking features enables reliable detection and prediction of external attention levels on a subject-dependent basis, achieving an accuracy exceeding 85% (II). Third, we show that multimodal fusion of EEG signals and eye tracking data elevates the accuracy of attention classification models by 5%–15%, as compared to their unimodal counterparts, in both subject-dependent and subject-independent settings (III). Fourth, we identify frontal theta power as the most significant predictor for internal attention, drafting potential applications for adaptive systems (IV). Lastly, we make our analysis approach and preprocessed datasets openly available enabling other researchers to replicate, extend, and innovate based on our work (V).

2 RELATED WORK

In this section, we review relevant existing literature and highlight the relevance of investigating internal and external attentional states for VR. Then we discuss their EEG correlates in terms of alpha and theta frequency bands. Finally, we summarize previous work that employed EEG as input for adaptation in VR.

2.1 Attention Classification using EEG

In VR environments, the immersive experience continuously stimulates our senses, primarily through visual and auditory channels, significantly influencing our attentional state. This constant stimulation necessitates distinguishing between external and internal attention mechanisms to tailor adaptive VR experiences effectively [22]. External attention is oriented towards stimuli in the environment, either voluntarily by focusing on task-relevant aspects or involuntarily by the saliency of external events [23]. Conversely, internal attention involves processing information within our mental framework, such as memory retrieval or mental calculations, and is often guided by our goals or knowledge [16]. The delineation between external and internal attention is crucial in VR, where the visually dominant environment can either facilitate or hinder the direction of attention, impacting user engagement and task performance.

The system sometimes negatively impacts the user's attentional state. For example, it can be frustrating when a user attempts to focus on a task (internal attention) but is interrupted by visually stimulating content (external attention), or conversely, when peacefully engaged in the virtual environment, they are suddenly tasked with mentally demanding activities. Implementing attention-aware adaptive systems that rely on passive physiological measures, such as EEG, could prevent such interference or leverage it to support user experience more effectively [15, 17].

Empirical evidence spanning multiple studies consistently suggests that alpha and theta activities in the EEG can predict attentional direction, with variations in these signals indicating shifts between internal and external attention. Specifically, Chiossi et al. [16] implemented an EEG-based adaptive VR system by assessing the relative change in parietal alpha and frontal theta power within a fixed-length time window, successfully demonstrating that this approach could effectively support attention allocation by distinguishing between internal and external attention states. This distinction is vital for adapting VR environments in real-time to enhance user performance, engagement, and reduce perceived workload.

Further studies, such as those by Aliakbaryhosseinabadi et al. [2], Alirezaei & Sardouie [3], and Sharma et al. [66], have expanded the classification of attention levels and types through EEG in various contexts, reinforcing the potential of EEG in understanding and enhancing user interaction within VR. Vortmann et al.'s work [71] in AR, and subsequent studies, highlight the feasibility of classifying internal versus external attention through EEG, though research in VR settings remains less explored.

This subsection lays the groundwork for further investigation into EEG-based attention classification within VR, aiming to bridge the gap identified by prior studies and exploring the application of this technology to enhance the VR user experience across a wider array of visual and task load settings.

2.2 Attention Classification using Eye Tracking

Recognizing the importance of eye gaze behavior in gaining insights into the human mental state, researchers have long been exploring the potential use of eye tracking data to classify attentional directions. Zarour et al. [77] utilized eye tracking in VR to monitor visual distraction levels of learners during cognitive tasks and achieved high performance. Benedek et al. [11] reported that the direction of cognition correlates significantly with eye features such as pupil diameter. Supporting these findings, Annerer-Walcher et al. [6] confirmed the indicative role of pupil diameter in distinguishing between external and internal attention. Nonetheless, the researchers acknowledged that, while attentional states can be effectively classified using eye features, it is challenging to generalize this capability across different tasks, as the concrete task type usually moderates these features.

In accordance with these observations, Vortmann & Putze [73] demonstrated that, within a subject-dependent framework, the incorporation of eye tracking features marginally improved the accuracy of predicting an individual's attentional state when combined with EEG features (cf. Section 2.3). Nonetheless, this improvement did not extend to a subject-independent setting. This outcome suggests the presence of significant variability in eye features across individuals. It has also proven less effective to reliably detect attentional states based solely on these eye features, whether in subject-dependent or subject-independent settings.

Vortmann et al. [70] then presented a novel approach for eye-based attention detection. The authors transformed eye tracking time series into images and trained deep models to classify those images instead of the original time series or any explicitly extracted feature set thereof. Intuitively, this imaging can be viewed as an initial decoding step that "disperses" some latent information embedded in the time series, and it has a key advantage that it facilitates the utilization of modern deep learning models specialized in image processing, which obviates the need for explicit feature engineering. The researchers achieved a high accuracy using this approach. Subsequently, Vortmann & Putze [72] found that, in a subject-independent setting, this approach is also more robust compared to the explicit feature engineering.

2.3 Multimodal Attention Classification

While either of EEG and eye tracking independently captures specific facets of human attention, the true predictive power may lie in the synergy achieved when a model coherently integrates the information from both modalities. For example, Sharma et al. [65] combined EEG with eye tracking for classifying navigational and informational search intents, and achieved high accuracy in the subject-independent setting.

In the two methodologically related studies [73, 74], authors extracted 12–14 eye tracking features (cf. Section 2.2) and 160–192 EEG features based on power spectral densities from various channels and frequency bands. For each data instance, its associated eye tracking features and EEG features were combined into a comprehensive feature vector. Then, machine learning models were trained to classify these combined feature vectors, and the classification outcomes were compared with results obtained using unimodal feature vectors, consisting exclusively of either EEG features or eye tracking features. It was observed that EEG features and eye tracking features exhibited a relatively weak correlation, indicating that they do encode different aspects of the same cognitive process. Classifiers trained with the multimodal feature set also performed better than those trained with unimodal features.

The image representation techniques from [70] and [72], and the multimodal methodologies outlined in [74] and [73] were then integrated by Vortmann et al. in [69]. The authors evaluated two data representation formats without explicit feature extraction, alongside four feature fusion strategies. The study revealed that multimodality by image channel concatenation demonstrated inferior performance compared to simple unimodal approaches.

3 USER STUDY

We aim to classify externally and internally directed attentional states using automatic, individual feature selection. Thus, we compared three Visual Complexity levels (No, Low, High) while performing two Tasks, either allocating external attention, i.e., Visual Monitoring, or internal attention, i.e., N-Back task. Based on previous work by Chiossi et al. [14, 16, 20], we implemented a within-participants study where the level of Visual Complexity can be manipulated by adjusting the number of Non-Player Characters (NPCs) within the virtual environment. Therefore, independent variables were manipulated using a 3×2 experimental design, see Figure 1.

3.1 Procedure

Upon arrival, participants were briefed on the study protocol and addressed any questions before signing informed consent. An explanation of tasks followed the EEG cap setup. Post EEG and VR headset preparation, a five-point eye calibration was performed (cf. Section 3.4.2), and detailed instructions were provided before each block. A ca. one-minute preliminary phase in the default, neutral VR environment allowed participants to familiarize themselves with the visual feeling (e.g., distance) within the VR headset as well as the correct operation of the controller. Following this, the experimental procedure started. The procedure commenced with the Individual Alpha Frequency (IAF) block, entailing a 2-minute eyes-closed session, detailed in Section 3.4.1. Subsequently, a 6-minute resting-state block started, where participants sat motionless in the VR setting, devoid of NPCs or tasks. The experiment progressed then through six randomized blocks (Visual Monitoring - No/Low/High Visual Complexity, N-Back - No/Low/High Visual Complexity), each lasting 6 minutes. Between blocks, participants evaluated their workload using the NASA-TLX questionnaire [36] and engagement via the Game Experience Questionnaire (GEQ) Core Module [38]. We collected Competence, Immersion, and Positive Affection subscales, as those subscales showed the highest content validity, following the recommendations of Law et al. [43]. We do not report results on subjective scores in this work. The total experiment duration was one hour and thirty minutes.

3.2 Tasks

For each participant, the study began with an *IAF Block*, where we asked the participant to keep their eyes closed for two minutes, during which their EEG signals are recorded for later computation of individual alpha frequency (IAF) [24]. Then, the experiment moved to the *Resting Block*, where we asked participants to seat comfortably, relax, and stare in a neutral VR environment. In this block, we acquired EEG data from participants in a resting position for later EEG normalization. Thereafter, the participants underwent the same six "experiment blocks," with the order of these blocks individually randomized. These six blocks were categorized into two groups: *Visual Monitoring* and *N-Back*, as summarized in Figure 1. Visual capture of the experimental conditions is presented in Figure 2.

3.2.1 Visual Monitoring. In each of the three *Visual Monitoring* blocks, participants were engaged in a VR task that is assumed to

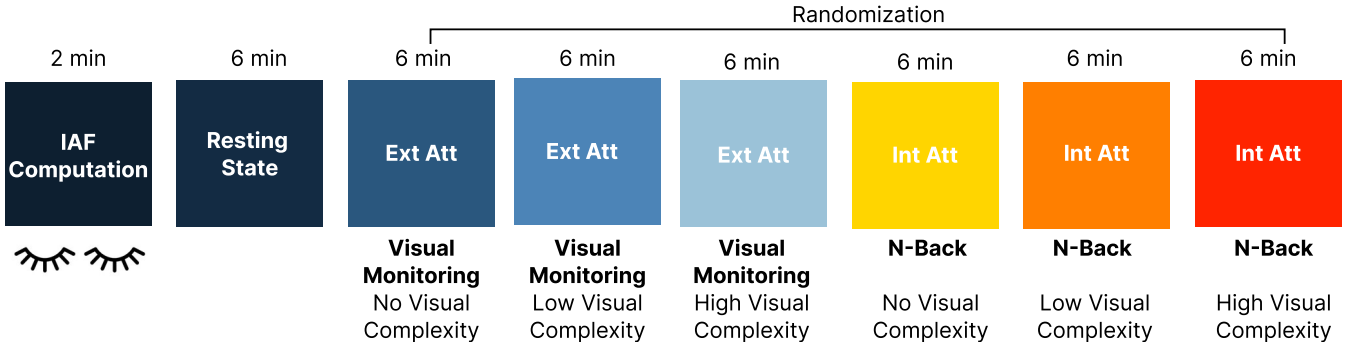
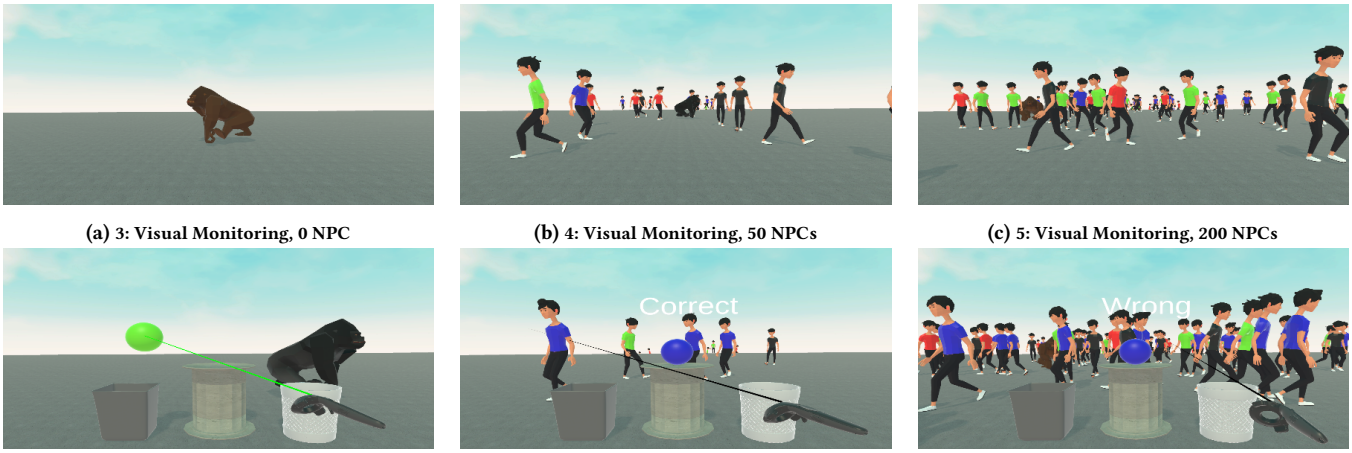


Figure 1: *Experiment Procedure.* The experiment encompassed eight different blocks. Participants closed their eyes for 2 minutes in the first block, allowing us to compute the individual alpha frequency (IAF). Then, they seated comfortably and stared in a neutral VR environment for 6 minutes, as we acquired EEG data from participants in a resting position for later EEG normalization. Finally, the experimental blocks started, manipulating Task and Visual Complexity. Refer to Section 3.1 and Section 3.2 for a complete description of the experimental conditions.



(a) 3: Visual Monitoring, 0 NPC **(b) 4: Visual Monitoring, 50 NPCs** **(c) 5: Visual Monitoring, 200 NPCs**

(d) 6: N-Back, 0 NPC **(e) 7: N-Back, 50 NPCs** **(f) 8: N-Back, 200 NPCs**

Figure 2: Screenshots of the experiment blocks: Two independent variables, Task (Visual Monitoring, N-Back) and Visual Complexity (No, Low, High), were manipulated using this 2×3 experimental design.

evoke external attention processes with different levels of visual complexity. Participants were presented with a stream of NPCs and tasked with vigilantly monitoring the NPC stream. The NPCs were spawned from the horizon and then walked roughly towards the viewer along certain random waypoints, and passing to the left and right of the viewer. The participants aimed to identify a deviant NPC, notably a gorilla, and track its trajectory until it disappeared from the VR scene. Participants remained seated throughout the entire experiment. We implemented three different levels of visual complexity across these Visual Monitoring blocks to systematically manipulate the demands on external attention resources. This task was designed based on previous cognitive tasks to recruit external attention resources [16, 23]. In the *No Visual Complexity* block, there was no NPC, and the participant only needed to identify and visually follow the gorilla. Another block maintained a *Low Visual Complexity*, limiting the NPC count to a maximum of 50 at any given time. In the other block, the visual environment featured a *High Visual Complexity*, with a maximum of 200 NPCs concurrently present, necessitating increased external attention from participants. We expected this intentional variation to allow us to explore the differential demands on external attention resources. We used NPC count to manipulate visual complexity, instead of a neutral, non-social setting, as their distracting effect is based on previous work [18, 19]. Furthermore, our NPCs had a skeleton and were realistically animated. so they are more realistic than abstact shapes (e.g., cubes, spheres, pyramids), thus increasing the immersive properties of VR simulations.

3.2.2 N-Back. In each of the three *N-Back* blocks, participants were asked to perform a visual working memory task known as the N-Back task [52] with $N = 2$. In this task, participants were required to decide whether the color of the sphere presented two steps earlier matched the color of the currently presented sphere. The execution involved picking up the sphere and placing it into one of

two buckets that were positioned on the left and right sides of the user. In cases where the colors matched, participants were directed to deposit the sphere into the right bucket; conversely, in case of a mismatch, participants were instructed to place the sphere in the left bucket. Sphere colors included green, red, blue, or black. To promote task engagement, immediate visual feedback was provided after each decision, indicating correctness or incorrectness. Additionally, after every 20 trials, the total accuracy of these trials was briefly displayed, allowing participants to gauge their performance. Participants were encouraged to try to maintain an accuracy level over 90%. This N-Back task effectively recruited internal attention through maintaining and updating relevant information (sphere color) in working memory for decision-making. By manipulating the level of visual complexity and incorporating task-irrelevant visual information (NPCs) mirroring the Visual Monitoring blocks, we expected to realize a competition between the external and internal attention states.

3.3 Participants

We recruited 24 participants (M = 23.54, SD = 3.55; 15 female, 9 male, none diverse) through convenience sampling and social media. Participants provided written informed consent before their participation. None of the participants reported a history of neurological, psychological, or psychiatric symptoms.

3.4 Apparatus and Data Recording

The N-Back and Visual Monitoring tasks were implemented in Unity (Version 2020.3.8 LTS). We presented the VR environment using an HTC VIVE Eye Pro headset with a display resolution of 2880×1600 pixels combined (Field of View: 110°). For environment tracking, we used two HTC Vive lighthouses 2.0. We acquired two physiological measurements: EEG signal using two LiveAmp amplifiers connected via Bluetooth (BrainProducts GmbH, Germany, 500 Hz) and eye tracking data via the VR headset (120 Hz). Physiological data were streamed within the Unity VR environment within the Lab Streaming Layer (LSL) framework[1] to the acquisition PC (Windows 10, Intel Core i7-11700K, 3.60 GHz, 16GB RAM). Figure 3 shows a participant wearing the experiment apparatus.

3.4.1 EEG Recording & Preprocessing. EEG data were recorded from 64 Ag/AgCl pin-type passive electrodes mounted over a water-based EEG cap ("64Ch Wet-Sponge R-Net for LiveAmp", Brain Products GmbH, Germany) at the following electrode locations: Fp1, Fz, F3, F7, F9, FC5, FC1, C3, T7, CP5, CP1, Pz, P3, P7, P9, O1, Oz, O2, P10, P8, P4, CP2, CP6, T8, C4, Cz, FC2, FC6, F10, F8, F4, Fp2, AF7, AF3, AFz, F1, F5, FT7, FC3, C1, C5, TP7, CP3, P1, P5, PO7, PO3, Iz, POz, PO4, PO8, P6, P2, CPz, CP4, TP8, C6, C2, FC4, FT8, F6, F2, AF4, AF8 according to the 10–20 system. Two wireless, Bluetooth-based LiveAmp amplifiers acquired EEG signals with a sampling rate of 500 Hz. All electrode impedances were kept below 20 kΩ according to the manufacturer's prescription. During the recording, we referenced channels to FCz and chose AFz as the ground. As the first preprocessing step, we removed the first 6 seconds and the last 2 seconds from each block. This was motivated by the need to mitigate the influence of transitional artifacts that typically occur

[1] https://github.com/labstreaminglayer/

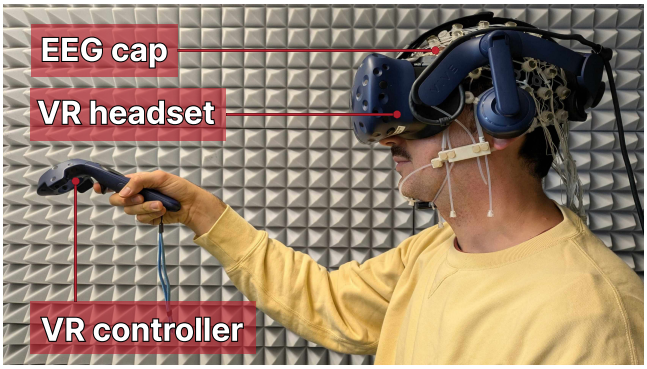


Figure 3: The experimental setup encompassed a desktop computer while the VR scene was deployed using an HTC Vive Pro Eye to collect eye tracking data (90 Hz). EEG data were collected using a 64-channel R-Net with two wireless amplifiers (500 Hz). The experiment took place in a distraction-free and soundproof laboratory.

when a participant is either adjusting to the start of a new task block or transitioning away from it at the end [46]. We then implemented our further preprocessing pipeline in MNE Python [33]. A band-pass FIR filtering was performed from 1 Hz to 70 Hz to remove low- and high-frequency noise. Then, the signal was notch-filtered at 50 Hz to remove power-line interference and finally re-referenced to the common average reference. We then applied Independent Component Analysis (ICA) with extended `InfoMax` [9, 44]. To facilitate automatic classification and correction of ICA components, we employed the MNE plugin `MNE-ICALabel` [45].

Regions of Interest (ROI). We selected our ROIs guided by prior research and literature based on alpha and theta Power Spectral Density (PSD) computation [10, 16, 51]. Their studies employed parieto-occipital channels for computing alpha oscillations, while fronto-central channels were chosen for theta analysis. To streamline our ROIs further, we have excluded four channels susceptible to artifacts from VR headsets (Fp1, Fp2, AF3, AF4) while incorporating four additional channels (P1, P2, P4, Oz) for compactness and symmetry. Following Putman et al. [58], we chose our ROI for beta to be the same as that for theta. Finally, for delta and gamma, we oriented ourselves based on further related work on attention [26, 35] and its detection [71, 73]. We used all five ROIs for our first classification approach (see Section 4.1.3), while for another classification approach we only considered the ROIs for alpha and theta (see Section 4.2.4). This resulted in following ROIs: Fz, F3, F4, F7, F8, Cz, C3, C4, Pz, P3, P4, Oz, O1, O2 for delta; Fz, F1, F2, F3, F4, FC1, FC2 for theta and beta; Pz, P1, P2, P3, P4, POz, PO3, PO4, Oz, O1, O2 for alpha; Fz, F1, F2, F3, F4, FC1, FC2 for beta; Fz, F3, F4, FT7, FT8, Cz, C3, C4, Pz, P3, P4, PO7, PO8, Oz for gamma.

IAF computation. We adopted the methodology from Corcoran et al. [24] and Klimesch [42] for calculating IAF, allowing for alpha band determination on a personal level by accounting for inter-individual differences. IAF was computed using electrodes of the alpha ROI, resulting in an average lower alpha range of 7.28 Hz (SD = 0.98) and an upper range of 11.89 Hz (SD = 0.71). Utilizing

the individual alpha lower bound as a point of reference, we delineated the individual theta frequency range. The alpha lower bound functions as the theta upper bound, while the theta lower bound is established by subtracting 4 Hz from the alpha lower bound.

3.4.2 Eye Tracking Recording & Preprocessing. For eye tracking data, we utilized the integrated eye tracker of the VR headset, coupled with the SRanipal eye tracking SDK[2]. For this study, we focused on the pupil diameter. To calculate features related to pupil diameter, as outlined in prior studies [40, 54], we first interpolated the recorded time series of pupil diameter linearly to maintain a uniform sampling rate of 120 Hz, and then determined the average of the normalized sizes of both the left and right pupils, termed combined pupil diameter. These preprocessed pupillary time series serve as the foundation for subsequent analysis.

4 CLASSIFICATION APPROACHES

In this section, we delineate our dual classification strategy, which juxtaposes the precision of classical linear models with the adaptability of contemporary deep learning techniques. The choice of Linear Discriminant Analysis (LDA) is grounded in its established efficacy for datasets of limited size, as noted by Lotte et al. [50] offering a robust baseline for feature-based classification including PSDs. On the other hand, the deployment of neural networks, particularly convolutional neural networks, is inspired by their superior performance in complex pattern recognition tasks within EEG data, as supported by findings from Appriou et al. [7]. This approach enables the direct processing of raw time-series data, capturing intricate temporal and spatial dependencies without the prerequisite of manual feature extraction. The integration of these methodologies allows us to harness both the interpretability of traditional statistical models and the comprehensive feature extraction capabilities of deep learning, aiming for the most effective detection of attention states in VR.

4.1 LDA Classification

Our goal is to explore the subject-independent case, focusing on models that can reliably detect attentional states across different individuals. This approach seeks to obviate the need for training a separate model for each person, thereby providing a scalable solution. While the overarching objective is a subject-independent analysis, we also conducted subject-dependent studies to facilitate comparative feature selection and model performance analyses.

We considered three different *window sizes*: 4 seconds, 8 seconds, and 11 seconds. In each setting, the signals were segmented based on the window size, yielding three distinct datasets with 88, 44, and 32 non-overlapping data samples per block per participant, respectively. In all three settings, we conducted the same feature extraction steps and classification experiments, as described below.

4.1.1 EEG Feature Engineering. For each EEG data segment, we calculated the PSDs using Welch's periodogram method [76] for delta (1 Hz to $IAF.lower - 4$), theta ($IAF.lower - 4$ to $IAF.lower$), alpha ($IAF.lower$ to $IAF.upper$), beta ($IAF.upper$ to 30 Hz), and gamma (30 to 45 Hz) frequencies based on our predefined ROIs and computed IAF ranges (see Section 3.4.1). In addition, we incorporated normalized versions of these values by subtracting the segment's PSD from the corresponding PSD value computed based on the entire Resting Block.

4.1.2 Eye Tracking Feature Engineering. For each eye tracking data segment, we computed the average pupil diameter, standard deviation of pupil diameter, and the index of pupillary activity (IPA) [30] using the combined time series of left and right pupil diameters.

4.1.3 LDA Feature Sets. This led to the identification of five distinct feature groups in the subject-independent case:

(1) *EEG only*: PSDs of Alpha, Theta, Delta, Beta, and Gamma
(2) *Normalized EEG only*: Normalized versions of the five PSDs
(3) *Pupil only*: Pupil Diameter Average, Pupil Diameter Standard Deviation, IPA
(4) *Pupil + EEG*: 3 *Pupil only* features plus 5 *EEG only* features
(5) *Pupil + normalized EEG*: 3 *Pupil only* features plus 5 *Normalized EEG only* features

In the subject-dependent case, the individual normalization of EEG based on the Resting Block does not yield significant differences in principle. This is attributed to the fact that each person is analyzed individually in this scenario, where `StandardScaler` from the Python data science library `scikit-learn` [57] was consistently applied to standardize both the training and validation data, and then transforming the test data accordingly. As a result, in the subject-dependent case, there are only three distinct feature groups instead of five.

4.1.4 LDA Classification Setup. The classification analysis aimed to differentiate physiological data recorded in the Visual Monitoring blocks from those recorded in the N-Back blocks. In this context, LDA has demonstrated good performance in classifying EEG and/or eye tracking features in various previous studies [16, 71, 73]. We used the `scikit-learn` implementation of LDA [57]. For subject-dependent and subject-independent analyses, we applied two different procedures as follows.

Subject-Dependent Classification. In this case, each participant's data was analyzed separately. For each person, the data was initially segmented into windows, with 1/6 of these segments randomly selected and set aside as a test set. Then, a 5-fold stratified random permutation cross-validator `StratifiedShuffleSplit` with a training size of 80% was employed to further partition the remaining 5/6 of the data into training and validation sets. This resulted in approximately 2/3 of the total data for training, 1/6 for validation, and a fixed 1/6 for testing. Combining the 5-fold shuffle split and `GridSearchCV` for hyperparameter optimization, the model achieving the best cross-validation performance was then scored on the test set. This entire process was repeated 20 times for each person, involving 20 random splits into 1/6 data for testing and 5/6 for training and validation. Consequently, each individual yielded 20 cross-validation scores and 20 test scores. This procedure was replicated for three different window sizes (4 seconds, 8 seconds, and 11 seconds) and three feature groups (Pupil only, EEG only, Pupil + EEG), i.e., 9 configurations in total. Altogether, each configuration of feature group and window size produced $24 \times 20 = 480$ cross-validation scores and test scores, respectively. For scoring, the

[2] https://developer.vive.com/resources/vive-sense/

balanced accuracy metric was employed. We did not use F_1 score, because the two classes, internal and external attention, carry equal weight in our case. Alongside the scores, the optimal hyperparameter combination found, and the refitted feature coefficients of the LDA model were recorded. The overall number of evaluable score points is given as:

$$3 \text{ window sizes} \times 3 \text{ feature groups} \times 24 \text{ subjects} \times 20 \text{ iterations}$$

Subject-Independent Classification. In this case, we aggregated participants' data and analyzed them collectively. In each iteration, 4 participants were randomly selected as the test set, while the data from the remaining 20 participants were standard-scaled, and the LDA model was then trained and cross-validated using `GridSearchCV` and 5-fold `StratifiedGroupKFold` (16 participants for training and 4 participants for cross-validation) on the standardized data. Next, the optimized model was scored on the test set, which consisted of the 4 persons randomly chosen beforehand. The data of these four individuals were also transformed using the standard scaler fitted on the other 20 persons. This iteration was repeated 480 times, yielding the same number of cross-validation scores, test scores, hyperparameters, and vectors of feature coefficients as in the subject-dependent experiment. Note that in the subject-independent case, there are five feature groups instead of three, since in a cross-individual analysis, the normalization of individual EEG based on the Resting Block becomes relevant. Thus, the total number of score points can be calculated using the formula:

$$3 \text{ window sizes} \times 5 \text{ feature groups} \times 480 \text{ iterations}$$

4.2 Deep Classification

We then addressed the attention detection problem using deep learning. In this study, we worked with EEG signals and eye tracking data, both of which can be essentially regarded as multivariate non-stationary time series. Despite these time series fully encoding all information about the respective signals, many interesting characteristics remain hidden inside their temporal dynamics. To exploit the benefits of various well-established deep learning techniques prevalent in the rapidly advancing fields of computer vision and image recognition, we first transformed these one-dimensional time series into two-dimensional images. In this context, we utilized an approach to transform physiological time series into images [69, 70] based on two algorithms from Wang and Oates [75].

4.2.1 Markov Transition Field. The first technique for imaging time series is *Markov Transition Field* (MTF), which transforms a time series into a matrix using transition probabilities, as illustrated in Figure 4. The main diagonal represents the self-transition probability at each timestamp. The MTF representation manifests larger squares in regions where the time series exhibit minimal magnitude variations over time. Conversely, thin lines are indicative of segments in the time series that share similar temporal dynamics.

4.2.2 Gramian Angular Field. The second technique for imaging time series is *Gramian Angular Field* (GAF), including two variants: *Gramian Angular Difference Field* (GADF) and *Gramian Angular Summation Field* (GASF), as shown in Figure 5. Each cell $[i, j]$ represents the trigonometric difference or trigonometric sum of the points x_i and x_j with respect to the time interval. On the main diagonal, each cell contains the original angular information and could be used to reconstruct the original time series. In this work, we only worked with the GASF images.

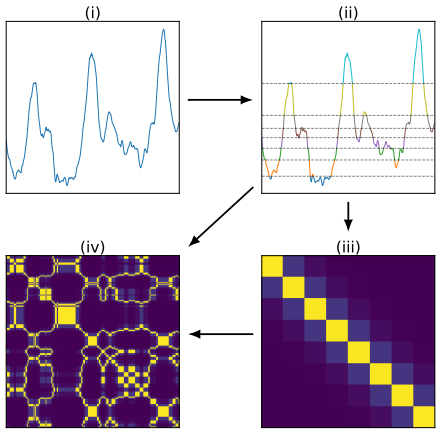


Figure 4: Markov Transition Fields: The values of time series X with T time points (i) are partitioned into Q quantiles $q_1, q_2, \ldots$, and each data point $x_t \in X$ is assigned to a quantile (ii). Then, we count the transitions from x_t to x_{t+1} between quantiles along the time axis and construct a normalized $(Q \times Q)$ weighed adjacency matrix W, termed Markov transition matrix (iii). Finally, W is "broadcasted" among the magnitude axis considering the temporal positions, producing the $(T \times T)$-image M, using the formula $M_{[i,j]} := W_{[a,b]}$ where $q_a \ni x_i$ and $q_b \ni x_j$ (iv).

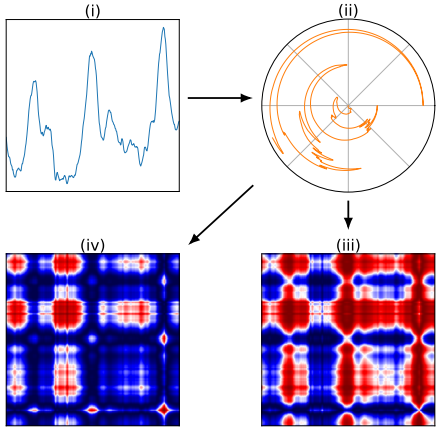


Figure 5: Gramian Angular Fields: The values of the time series X with T time points (i) are first rescaled to $[-1, 1]$ and then represented in polar coordinates by encoding the magnitudes x_t as the angular cosine $\phi_t = \arccos(x_t) \in [0, \pi]$ and with radius $r_t = (t-1)/(T-1) \in [0, 1]$ (ii). Then, we identify the temporal correlation within time intervals by $\sin(\phi_i - \phi_j)$ pairwise between the points, resulting in a $(T \times T)$-image GADF (iii), or calculating $\cos(\phi_i + \phi_j)$ pairwise between the points, resulting in a $(T \times T)$-image GASF (iv).

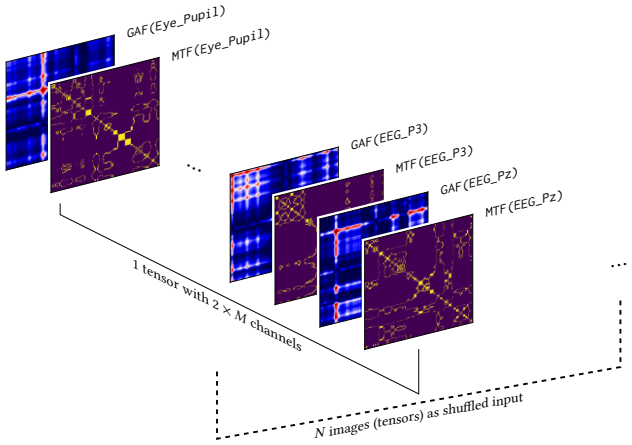


Figure 6: Each 4-second segment consists of M signal sources (e.g., "left pupil diameter", "Pz", etc.). Each is transformed in two (224×224)-images, MTF and GASF. Then, these $2M$ images are concatenated along the channel dimension. Therefore, each data instance is a tensor of shape $2M \times 224 \times 224$.

4.2.3 Channel Concatenation. Assume we have M time series from each subject, either from one or different modalities, e.g., they can be a pupillary time series, a Pz time series, a P3 time series, etc. and we have established temporal synchronization between them, and segmented each of these time-synchronized time series into N short intervals based on timestamps. For each segment and each source of time series, we used MTF and GAF to transform it into 2 images, resulting in $2 \times M$ images per segment. Across all segments, this yields a total of $2 \times M \times N$ images per subject. The images are arranged as illustrated in Figure 6.

4.2.4 Deep Classification Datasets. We examined three modalities:

(1) *Pupil only* feature set consists of 3 time series: Left Pupil Diameter, Right Pupil Diameter, Average Pupil Diameter.
(2) *EEG only* feature set consists of a variable set of EEG time series (see below).
(3) *Pupil + EEG*: Multimodal feature set consists of 3 *Pupil only* time series plus *EEG only* time series.

For EEG, we additionally differentiated between three options:

- **FC**: 7 frontal and central channels (ROI for theta and beta).
- **PO**: 11 parietal and occipital channels (ROI for alpha).
- **FCPO**: 18 channels from **FC** and **PO**.

For each configuration, relevant features' time series were extracted from all participants across all blocks. The signals were then segmented into 4-second non-overlapping intervals. This process resulted in 352/4 = 88 signal segments per signal source per block. In total, 24 subjects × 6 blocks × 88 = 12672 tensors were generated for each configuration. The sizes of datasets and numbers of channels per tensor are detailed in Table 1.

4.2.5 Deep Classification Model. For image classification, we used ResNet-18 [37] whose residual connections enable robust feature extraction, potentially facilitating superior classification performance even in complex non-natural image datasets. PyTorch (`torch`) [55]

Table 1: Overview of deep classification experiments.

Feature Group	# Instances	# Tensor Channels
Pupil only	12672 (= 24 · 6 · 88)	6
EEG only (FC)		14
EEG only (PO)		22
EEG only (FCPO)		36
Pupil + EEG (FC)		20
Pupil + EEG (PO)		28
Pupil + EEG (FCPO)		42

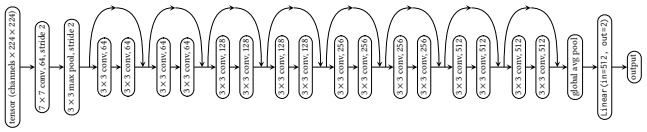


Figure 7: ResNet-18 architecture used for deep classification.

offers an implementation of ResNet-18. We then modified its first convolutional layer to adjust the number of channels from 3 to our specified number of channels (see Table 1). We also modified the last fully-connected layer to have 2 output features, standing for internal and external attention. The overall architecture of our ResNet-18 model is depicted in Figure 7. We did not use pre-trained weights because MTFs and GAFs are not natural images.

PyTorch automatically resizes input tensors for ResNet-18 to a fixed shape of 224 × 224 × #Channels. To prevent information loss caused by an additional scaling process, we generated images directly in this resolution. Each 4-second signal segment was first linearly interpolated to a length of 224 and then transformed using the `pyts` toolbox [32] into two images: one MTF (with 16 bins) and one GASF, both sized at 224 × 224.

4.2.6 Deep Classification Setup. For each of the 7 feature groups, the following procedure was repeated 20 times independently: A ResNet-18 model was initialized and trained on a random subset totaling 70% of the images, to classify image tensors from blocks 3-5 against image tensors from blocks 6-8. Validation was performed on the remaining 30% of images. The train-validation split was consistently stratified based on the block number, resulting in 6 strata. Throughout all experiments, we used the Cross-Entropy Loss function, the Adam optimizer [41] with a constant learning rate of `1e-4`, and maintained a batch size of 32. Early stopping was triggered if the validation loss continuously increased over the last three epochs. In total, we trained 7×20 = 140 models for evaluation.

5 RESULTS

In this section, we present the results produced from our extensive classification experiments.

5.1 Subject-Dependent LDA Classification

5.1.1 Accuracy. Table 2 shows the LDA classification accuracy within a subject-dependent setting. When combining pupil and EEG features and using 11-second windows, the median accuracy on the test set reached 86.7% . In particular, there are two interesting

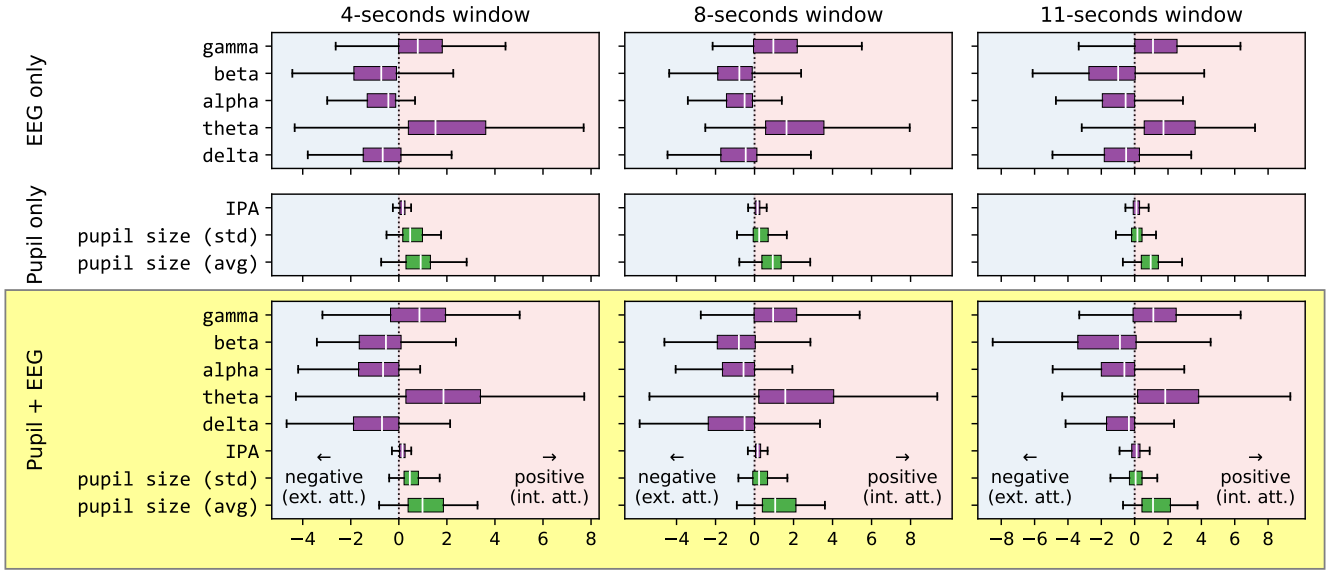


Figure 8: Overview of subject-dependent LDA classifier weights: Each row stands for one modality, and each column for one window size. Each subplot shows the feature weights (`.coef_`) of fitted LDA models from the respective configuration. In each subplot, the dotted vertical line indicates 0, the left area with light blue background indicates < 0, and the light pink area indicates > 0. In our particular setting, features that have predominantly negative weights are supposed to correlate more with external attention (e.g., `beta`, `alpha`, `delta`), and predominantly positive weights are indicative of correlation with internal attention (e.g., `gamma`, `theta`, `pupil_avg`). Furthermore, the larger the magnitude, the stronger the correlation. The modality with best overall classification accuracy [Pupil + EEG] is additionally highlighted with a yellow background.

Table 2: Overview of subject-dependent LDA classification accuracy: Regardless of window size, the multimodal feature sets [Pupil + EEG] generally resulted in higher accuracy scores than the unimodal counterparts [EEG only] & [Pupil only]. The pupil-only modality achieved the weakest performance. Underlined numbers denote the three top median accuracy scores. The best result is additionally highlighted in bold. The 95% confidence intervals were computed using the bootstrap method with 10,000 resamples.

		EEG		Pupil		Pupil + EEG	
		Med	**CI_{95}**	**Med**	**CI_{95}**	**Med**	**CI_{95}**
CV	**4s**	76.3	[75.7–77.0]	70.8	[69.6–71.9]	83.5	[82.3–84.8]
	8s	78.6	[77.3–80.5]	70.2	[69.1–71.2]	85.3	[83.2–87.0]
	11s	80.3	[78.9–81.2]	70.0	[68.8–70.6]	**86.2**	[83.8–87.7]
Test	**4s**	77.1	[75.7–78.1]	69.8	[68.9–71.2]	84.1	[82.8–85.1]
	8s	79.5	[77.8–81.2]	70.4	[68.9–71.3]	84.6	[83.1–86.3]
	11s	81.0	[79.0–81.6]	68.8	[67.1–69.8]	**86.7**	[85.3–87.5]

patterns: First, regardless of the window size employed, multimodality consistently outperforms the EEG-only approach, which again consistently outperforms the pupil-only approach. Second, with EEG features, performance improves with a longer window size or fewer data points. However, when only pupil features are utilized, the performance remains relatively stable or even deteriorates with this trend.

5.1.2 Weights. The training data was consistently standard-scaled, allowing for meaningful comparisons of weight coefficients. Larger magnitudes indicate a higher influence of the respective feature on the overall outcome of the LDA model. In our classification setup, we labeled data from Blocks 3, 4, and 5 as **0**, while Blocks 6, 7, and 8 were labeled as **1**. Consequently, any higher weight from the classifier is indicative of internal, while any lower weight points toward external attention. In each of the 9 configurations, we collected 24 subjects $\times$ 20 iterations $=$ 480 weight vectors (cf. Section 4.1.4), which are visualized in Figure 8.

In the most performative case of **Pupil + EEG** (see Section 5.1.1), i.e., the last row in Figure 8, the most informative feature is **theta**, which is positively associated with internal attention. This aligns with our assumption that theta power increases during internal tasks (see Section 2.1). Other features are also significant to a certain extent, including **average pupil diameter** and **gamma**, whose increase indicates internal attention, and **delta**, **alpha**, as well as **beta**, whose increase directs more towards external attention.

5.2 Subject-Independent LDA Classification

5.2.1 Accuracy. We now transition to the subject-independent analyses. Table 3 presents the classification outcomes in a subject-independent setting.

Among all feature combinations, the multimodal approaches that combine pupil and EEG features almost always outperform their unimodal counterparts. The only outlier is using pupil-only features with a 4-second window length. Irrespective, all multimodal

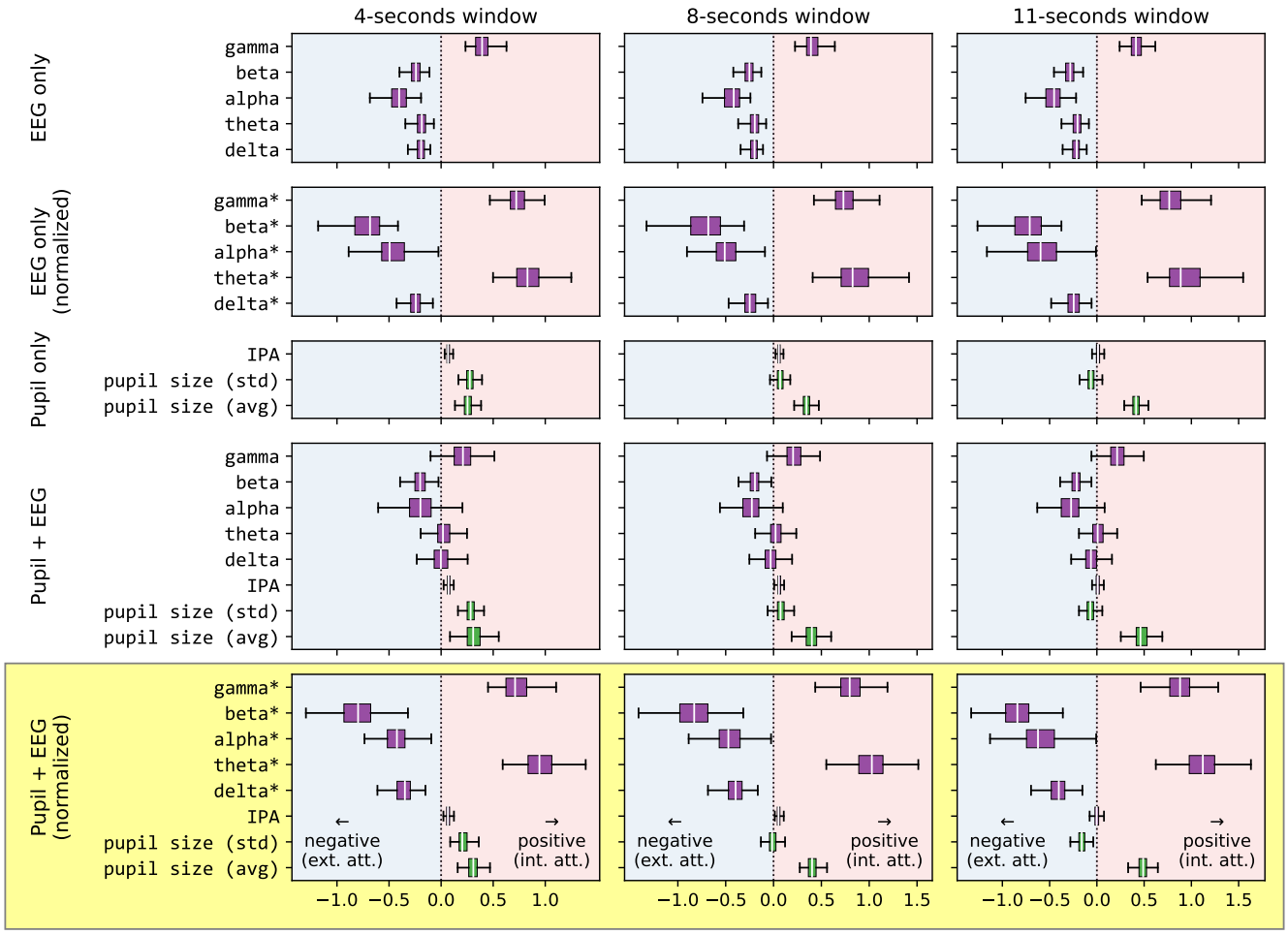


Figure 9: Overview of subject-independent LDA classifier weights: For explanation, please refer to the caption of Figure 8. The modality with the best overall classification accuracy is [Pupil + normalized EEG].

Table 3: Overview of subject-independent LDA classification accuracy: Normalized EEG is indicated by an asterisk. On the test sets, the multimodal feature sets with normalized EEG generally outperformed the other variants, with (4s, Pupil-only) being the only exception. Irrespective of that, the general trend of [Pupil + normalized EEG] > [Pupil + EEG] > [Pupil only] > [normalized EEG only] > [EEG only] can be observed on the test sets. Underlined numbers denote the five top median accuracy scores. The best result is additionally highlighted in bold. The 95% confidence intervals were computed using the bootstrap method with 10,000 resamples.

		EEG		EEG*		Pupil		Pupil + EEG		Pupil + EEG*	
		Med	CI_{95}	Med	CI_{95}	Med	CI_{95}	Med	CI_{95}	Med	CI_{95}
CV	**4s**	56.8	[56.7–57.0]	58.7	[58.6–58.8]	60.7	[60.6–60.8]	61.7	[61.5–61.9]	**62.9**	[62.7–63.0]
	8s	56.8	[56.7–57.0]	59.1	[59.0–59.2]	58.8	[58.7–59.0]	60.7	[60.6–60.9]	62.5	[62.4–62.7]
	11s	56.4	[56.2–56.6]	59.1	[59.0–59.2]	58.3	[58.1–58.4]	60.8	[60.5–60.9]	62.8	[62.7–63.1]
Test	**4s**	55.6	[55.0–55.9]	55.8	[55.3–56.2]	60.4	[59.8–61.0]	59.4	[58.8–59.9]	60.1	[59.5–60.6]
	8s	55.2	[54.6–55.7]	55.8	[55.3–56.3]	58.1	[57.5–58.5]	58.6	[58.0–59.4]	60.2	[59.6–60.7]
	11s	55.1	[54.6–55.7]	55.9	[55.3–56.2]	57.4	[56.8–57.9]	59.0	[58.2–59.9]	**61.2**	[60.2–61.6]

approaches incorporating normalized EEG achieved a median accuracy of greater than 60% on the test set, surpassing the performance of all other modalities. We can draw several interesting observations from the results. First, the combination of EEG and pupil features

consistently outperformed approaches using only pupil features, which, in turn, is better than relying solely on EEG features, that is, **Multimodal > Pupil > EEG**. However, in the subject-dependent (see Section 5.1.1) classification, the ranking is **Multimodal > EEG > Pupil**. Second, normalized EEG consistently outperformed their unnormalized counterparts, demonstrating the individual variability inherent in EEG characteristics and the necessity of normalization prior to further analysis.

5.2.2 Weights. The weights of the optimized subject-independent LDA models are displayed in Figure 9. As before, we consider only the most performative case of **Pupil + normalized EEG**, i.e., the last row in Figure 9: In general, EEG features play a more significant role than pupil features. Among the EEG features, **theta** stands out as the most prominent, with its increase strongly indicating internally directed attentional state. Also **gamma** plays a similar role. Conversely, increased **alpha**, **delta**, and **beta** power suggests tendencies in the externally directed attentional state. Regarding pupil features, the **average pupil diameter** appears to be a positive indicator for internal attention.

5.3 Subject-Independent Deep Classification

Building upon the insights gained from the feature engineering approach, we proceeded to explore deep learning methodologies. Here, we focused on subject-independent classification and a window size of 4 seconds. The results are presented in Figure 10. Overall, combining EEG features with pupil features consistently leads to higher median classification accuracy, surpassing the results obtained from using EEG or pupil features alone. In particular, the best unimodal result is still inferior to the worst multimodal one. Furthermore, utilizing a larger number of EEG channels or regions (combining FC and PO channels) also leads to improved outcomes compared to using either FC or PO channels alone, both in uni- and multimodal cases.

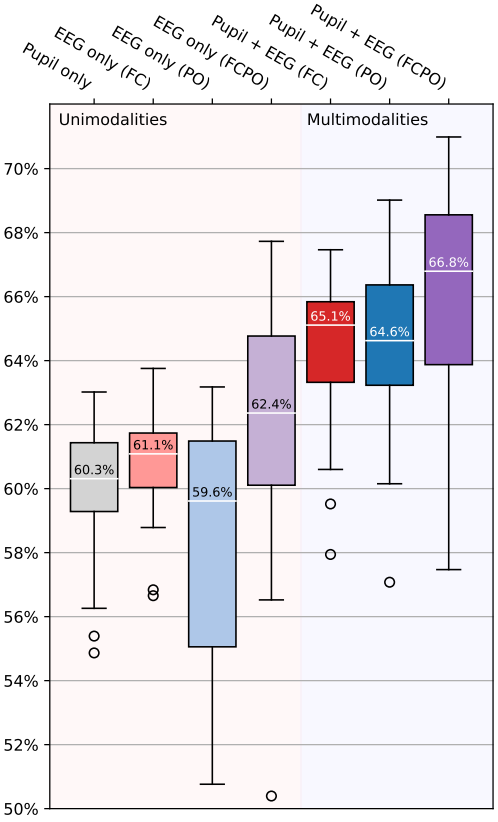


Figure 10: ResNet-18 validation accuracy. For each of the 7 feature groups (cf. Table 1), we trained 20 models independently (cf. Section 4.2.6). For each iteration, we selected the epoch with the lowest observed validation loss, and extracted validation accuracy from that epoch. The corresponding 20 validation accuracies from these selected epochs for every feature group are depicted here. For example, the median accuracy of [Pupil + EEG (FCPO)] is 66.8%, considerably higher than that of [EEG only (PO)], which is 59.6%. In general, we observe elevated accuracy in all multimodal cases as compared to their unimodal "constituents".

6 DISCUSSION

We have examined different configurations and methods for detecting internal and external attention within VR environments. We aimed to advance the development of a real-time, user-agnostic brain-computer interface that integrates seamlessly with VR technology. We evaluated the effectiveness of EEG and eye tracking as input methods, analyzed the performance of two distinct classification algorithms, and tested the classification performance of various data collection windows, modalities and feature sets.

6.1 Summary of Results

In our study, we investigated attention states decoding in VR using both subject-dependent and -independent LDA classifications, complemented by subject-independent deep learning analysis. The integration of EEG and pupil data consistently emerged as the most accurate method for classifying attention states, surpassing singular modality approaches across all models. This multimodal strategy achieved a peak accuracy of 86.7% in subject-dependent scenarios and demonstrated significant generalizability in subject-independent contexts, with median accuracies exceeding 60%. Notably, theta and gamma EEG frequencies were identified as robust indicators of internal attention across analyses, underscoring their potential as reliable markers for attention allocation detection. Furthermore, our deep learning exploration revealed that extending EEG channel selection enhances classification performance, affirming the value of comprehensive physiological data in attention research. Our results highlight the effectiveness of combining multiple physiological signals and advancing the understanding of attention mechanisms in immersive environments, offering modeling approaches for developing intelligent and context-aware VR systems.

6.2 Enhanced Classification with Multimodality

Our analysis confirms and extends the understanding that multimodal classifiers exhibit superior performance over their unimodal counterparts, a notion consistent with the broader findings in the

field. This superiority is evident across subject-dependent and -independent settings, underscoring the robustness of integrating EEG and eye tracking data for attention state classification. Interestingly, we observed a reversal in the relative importance of pupil diameter and EEG features between subject-dependent and -independent classifications, with EEG features showing greater discriminatory power in the former.

6.2.1 Differential Sensitivity of Pupil Diameter and EEG Features. The observation that pupil diameter is a more sensitive indicator in subject-independent classification, while EEG features take precedence in subject-dependent scenarios, underscores the complexity of physiological responses to attentional states. This discrepancy raises important considerations for designing attention-aware systems, particularly in selecting the most appropriate physiological markers based on the use context. For instance, applications requiring user-specific customization might benefit more from EEG-based indicators. At the same time, those aimed at a broader audience could leverage pupil diameter as a more universally applicable marker.

6.2.2 EEG Power Correlations with Internal Attention. We conducted binary classification of external and internal attentional states. Our analysis of the LDA feature weights reveals that in both subject-dependent and subject-independent settings, the coefficients associated with theta and gamma PSD features are predominantly positive. In contrast, the coefficients of alpha, delta, and beta features are predominantly negative. This indicates that internal attention correlates with increased power in theta and gamma bands and decreased power in alpha, delta, and beta bands, while external attention shows the opposite pattern. This association of internal attention with specific EEG power bands indicates that attention allocation can differentially impact EEG frequencies.

The significant increase in theta and gamma power observed in our study aligns with their known roles in cognitive control [68], working memory [47]. Theta power's increase replicates previous results as an EEG correlate in internal tasks such as problem-solving [25] or memory recall [17] within VR environments. This increase in theta activity suggests its utility as a neural marker, potentially guiding the development of VR systems that dynamically respond to the user's cognitive engagement. Similarly, the association of gamma power with internal attention underscores its role in integrating cognitive processes across various brain regions, indicating the intricate nature of internal focus maintenance. Contrary to traditional perspectives linking alpha power decrease primarily to external attentional shifts [23, 42], our findings suggest a more intricate role. The observed decrease in alpha power during internal attention states implies its function as a sensory gating mechanism, modulating the balance between internal and external focuses by filtering out irrelevant sensory information. The reductions in delta and beta powers replicate previous work on their role when disengaging from external stimuli and facilitating transitions towards internally directed cognitive states [34, 39].

6.2.3 Tailoring Adaptive Systems to Individual Differences. The distinct roles of EEG and pupil data in subject-dependent and -independent classifications reveal the variability in attentional processes among individuals. For instance, the pronounced discriminatory power of EEG features in a subject-dependent context suggests that individuals' not-stationary and highly variable EEG patterns persist as robust markers of attention allocation. Our results replicate previous work in AR settings [73], demonstrating substantial success with person-dependent models, reinforcing the potential for such personalized approaches to accommodate the EEG signal not-stationarity and individual differences during attention fluctuations. Here, person-dependent classification enables the creation of adaptive systems finely tuned to an individual's physiological and cognitive responses. This user-tailored approach ensures a higher degree of precision in detecting and adapting to shifts in attention, potentially leading to a more responsive VR interaction for the single user [1, 29, 64].

6.3 Real-time Adaptation using Attention States

Attention detection using linear and non-linear modeling must consider how such approaches are feasible for real-time adaptation based on attention states in online VR systems.

6.3.1 Real-time Data Segmentation and Processing. Our methodology involves segmenting EEG and eye tracking data into windows of 4, 8, and 11 seconds, yielding distinct datasets. For real-time applications, this segmentation must occur on the fly, necessitating efficient algorithms that can quickly partition and process incoming data without significant latency, which could disrupt the VR experience.

6.3.2 Computational Complexity. Both our subject-dependent and -independent LDA classification approaches involve complex computations, including stratified random permutation cross-validation and grid search for hyperparameter optimization. These processes are computationally intensive, especially considering the need for real-time feedback in VR. The deep learning approach further compounds this complexity by transforming time series into two-dimensional images for classification, demanding substantial computational power for real-time analysis. LDA models and deep learning architectures should be simplified to reduce computational load without significantly sacrificing accuracy. Techniques such as model pruning [63], quantization [48], and knowledge distillation [78] may offer a solution.

6.4 Limitations and Future Work

Our study on attention states in virtual reality offers new insights but encounters limitations, highlighting areas for further research. These include refining EEG electrode placement, standardizing experimental design, and exploring advanced imaging techniques for data analysis. Addressing these challenges will enhance our comprehension of VR attention mechanisms and the practicality of BCIs.

6.4.1 Electrode Detection. Our study raises important questions about the optimal placement and number of EEG electrodes for detecting attention states, balancing BCI usability with accuracy. Liu and Sourina [49] showed that recognizing emotions could be

achieved with just four electrodes, highlighting a trade-off between system simplicity and detection precision. This balance is crucial, especially when considering real-time processing demands against the need for accurate attention detection. The decision to use a minimal or comprehensive electrode setup depends largely on the application context: research settings might tolerate more complex setups for greater accuracy, while consumer applications favor ease of use and quick responses. Moreover, their approach is based on a person-dependent approach, which might not replicate effectiveness in a person-independent manner. This discussion points to further exploring electrode configurations and alternative brain imaging techniques to enhance BCI effectiveness across different uses.

6.4.2 Design of Experiments. Our experimental design necessitated executing two distinct types of tasks during the external and internal blocks. The inherent differences between these two types of tasks may have introduced uncontrolled, confounding variables into our experiments. For example, eye measures might have been moderated by the task type. We advocate for using two more similar tasks in future studies to eliminate uncontrolled variables.

Another limitations of our study design was the randomization of blocks for each participant using a simple shuffle method without manually imposed counterbalancing. While a sufficiently large sample size could potentially mitigate the noise introduced by this approach, our sample size of 24 participants may not be large enough to nullify these effects. Consequently, our results may suffer from uncontrolled order effects. Also, the prolonged duration of the experiment session and VR exposure (1.5 hours, which is relatively long for a VR study) may have led to participant exhaustion. Future studies should implement proper counterbalancing techniques (e.g., Latin square) to control for these effects, and we recommend careful consideration of experiment duration and participant workload in future studies.

6.4.3 Time Series Imaging. In our deep classification approach, we only worked with Markov Transition Fields and Gramian Angular Summation Fields (GASF). Whether adding GADF could yield improved results remains still open. Further, it is worth investigating whether different imaging techniques are better suited for specific modalities.

7 CONCLUSION

In this study, we explored attention detection within virtual reality environments using modern machine learning techniques. By classifying attention into external and internal states, our objective was to identify the user's cognitive state based on passive EEG and eye tracking measurements. Overall, our findings confirm the predictive power of specific EEG and eye features proposed by existing research, while we also present discrepancies and introduce new perspectives. Our observations reveal that by combining multimodal features from pupillometric and EEG sensors, our classification models consistently outperform the variants that consider only unimodal data. This aligns with our understanding of multimodal learning and encourages further research in this direction. With the vision of achieving subject-independent attention detection in mind, which is critical for the promotion and commercialization of future training-free brain–computer interfaces, we conducted extensive studies on that matter. As expected, we found that multimodality consistently yielded superior results compared to unimodalities. This result contributes to the expanding body of knowledge in the field of physiological computing and brain–computer interfacing and holds practical implications for the future development of user-adaptive systems.

8 OPEN SCIENCE

We encourage readers to reproduce and extend our results and analysis methods. Therefore, our experimental setup, collected datasets, and analysis scripts are available at https://osf.io/jsk37/.

ACKNOWLEDGMENTS

Francesco Chiossi was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) with Project ID 251654672 TRR 161. This work has been funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 952026 (https://www.humane-ai.eu/).

REFERENCES

[1] Reza Abiri, Soheil Borhani, Eric W Sellers, Yang Jiang, and Xiaopeng Zhao. 2019. A comprehensive review of EEG-based brain–computer interface paradigms. *Journal of Neural Engineering* 16, 1 (jan 2019), 011001. https://doi.org/10.1088/1741-2552/aaf12e

[2] Susan Aliakbaryhosseinabadi, Ernest Nlandu Kamavuako, Ning Jiang, Dario Farina, and Natalie Mrachacz-Kersting. 2017. Classification of EEG signals to identify variations in attention during motor task execution. *Journal of Neuroscience Methods* 284 (2017), 27–34. https://doi.org/10.1016/j.jneumeth.2017.04.008

[3] Mitra Alirezaei and Sepideh Hajipour Sardouie. 2017. Detection of Human Attention Using EEG Signals. In *2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME)*. IEEE, New York, NY, USA, 1–5. https://doi.org/10.1109/ICBME.2017.8430244

[4] Allison P. Anderson, Michael D. Mayer, Abigail M. Fellows, Devin R. Cowan, Mark T. Hegel, and Jay C. Buckey. 2017. Relaxation with Immersive Natural Scenes Presented Using Virtual Reality. *Aerospace Medicine and Human Performance* 88, 6 (2017), 520–526. https://doi.org/10.3357/AMHP.4747.2017

[5] Sonja Annerer-Walcher, Simon M Ceh, Felix Putze, Marvin Kampen, Christof Körner, and Mathias Benedek. 2021. How reliably do eye parameters indicate internal versus external attentional focus? *Cognitive Science* 45, 4 (2021), e12977. https://doi.org/10.1111/cogs.12977

[6] Sonja Annerer-Walcher, Simon M. Ceh, Felix Putze, Marvin Kampen, Christof Körner, and Mathias Benedek. 2021. How Reliably Do Eye Parameters Indicate Internal Versus External Attentional Focus? *Cognitive Science* 45, 4 (2021), e12977. https://doi.org/10.1111/cogs.12977

[7] Aurélien Appriou, Andrzej Cichocki, and Fabien Lotte. 2018. Towards Robust Neuroadaptive HCI: Exploring Modern Machine Learning Methods to Estimate Mental Workload From EEG Signals. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3170427.3188617

[8] Christopher Baker and Stephen H. Fairclough. 2022. Chapter 9 - Adaptive virtual reality. In *Current Research in Neuroadaptive Technology*, Stephen H. Fairclough and Thorsten O. Zander (Eds.). Academic Press, Academic Press, 159–176. https://doi.org/10.1016/B978-0-12-821413-8.00014-2

[9] Anthony Bell and Terrence Sejnow. 1995. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation* 7 (1995). https://doi.org/10.1162/neco.1995.7.6.1129

[10] Mathias Benedek, Rainer J. Schickel, Emanuel Jauk, Andreas Fink, and Aljoscha C. Neubauer. 2014. Alpha power increases in right parietal cortex reflects focused internal attention. *Neuropsychologia* 56 (2014), 393–400. https://doi.org/10.1016/j.neuropsychologia.2014.02.010

[11] Mathias Benedek, Robert Stoiser, Sonja Walcher, and Christof Körner. 2017. Eye Behavior Associated with Internally versus Externally Directed Cognition. *Frontiers in Psychology* 8 (2017). https://doi.org/10.3389/fpsyg.2017.01092

[12] Sergi Bermúdez i Badia, Luis Velez Quintero, Mónica S. Cameirão, Alice Chirico, Stefano Triberti, Pietro Cipresso, and Andrea Gaggioli. 2019. Toward Emotionally Adaptive Virtual Reality for Mental Health Applications. *IEEE Journal of Biomedical and Health Informatics* 23, 5 (2019), 1877–1887. https:

//doi.org/10.1109/JBHI.2018.2878846
[13] Simon Majed Ceh, Sonja Annerer-Walcher, Christof Körner, Christian Rominger, Silvia Erika Kober, Andreas Fink, and Mathias Benedek. 2020. Neurophysiological indicators of internal attention: An electroencephalography–eye-tracking coregistration study. *Brain and behavior* 10, 10 (2020), e01790. https://doi.org/10.1002/brb3.1790
[14] Francesco Chiossi, Thomas Kosch, Luca Menghini, Steeven Villa, and Sven Mayer. 2023. SensCon: Embedding Physiological Sensing into Virtual Reality Controllers. *Proc. ACM Hum.-Comput. Interact.* 7, MHCI, Article 223 (sep 2023), 32 pages. https://doi.org/10.1145/3604270
[15] Francesco Chiossi and Sven Mayer. 2023. How Can Mixed Reality Benefit From Physiologically-Adaptive Systems? Challenges and Opportunities for Human Factors Applications. https://doi.org/10.48550/arXiv.2303.17978
[16] Francesco Chiossi, Changkun Ou, Carolina Gerhardt, Felix Putze, and Sven Mayer. 2023. Designing and Evaluating an Adaptive Virtual Reality System using EEG Frequencies to Balance Internal and External Attention States. arXiv:2311.10447 [cs.HC]
[17] Francesco Chiossi, Changkun Ou, and Sven Mayer. 2023. Exploring Physiological Correlates of Visual Complexity Adaptation: Insights from EDA, ECG, and EEG Data for Adaptation Evaluation in VR Adaptive Systems. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 118, 7 pages. https://doi.org/10.1145/3544549.3585624
[18] Francesco Chiossi, Changkun Ou, and Sven Mayer. 2024. Optimizing Visual Complexity for Physiologically-Adaptive VR Systems: Evaluating a Multimodal Dataset using EDA, ECG and EEG Features. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces* (Arenzano, Genoa, Italy) *(AVI '24)*. Association for Computing Machinery, New York, NY, USA, Article 25, 9 pages. https://doi.org/10.1145/3656650.3656657
[19] Francesco Chiossi, Robin Welsch, Steeven Villa, Lewis Chuang, and Sven Mayer. 2022. Virtual Reality Adaptation Using Electrodermal Activity to Support the User Experience. *Big Data and Cognitive Computing* 6, 2 (2022). https://doi.org/10.3390/bdcc6020055
[20] Francesco Chiossi, Robin Welsch, Steeven Villa, Lewis L. Chuang, and Sven Mayer. 2022. Designing a Physiological Loop for the Adaptation of Virtual Human Characters in a Social VR Scenario. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, New York, NY, USA, 578–579. https://doi.org/10.1109/VRW55335.2022.00140
[21] Francesco Chiossi, Johannes Zagermann, Jakob Karolus, Nils Rodrigues, Priscilla Balestrucci, Daniel Weiskopf, Benedikt Ehinger, Tiare Feuchtner, Harald Reiterer, Lewis L. Chuang, Marc Ernst, Andreas Bulling, Sven Mayer, and Albrecht Schmidt. 2022. Adapting visualizations and interfaces to the user. *it - Information Technology* 64, 4-5 (2022), 133–143. https://doi.org/10.1515/itit-2022-0035
[22] Marvin M. Chun, Julie D. Golomb, and Nicholas B. Turk-Browne. 2011. A Taxonomy of External and Internal Attention. *Annual Review of Psychology* 62, Volume 62, 2011 (2011), 73–101. https://doi.org/10.1146/annurev.psych.093008.100427
[23] Giorgia Cona, Francesco Chiossi, Silvia Di Tomasso, Giovanni Pellegrino, Francesco Piccione, Patrizia Bisiacchi, and Giorgio Arcara. 2020. Theta and alpha oscillations as signatures of internal and external attention to delayed intentions: A magnetoencephalography (MEG) study. *NeuroImage* 205 (2020), 116295. https://doi.org/10.1016/j.neuroimage.2019.116295
[24] Andrew W. Corcoran, Phillip M. Alday, Matthias Schlesewsky, and Ina Bornkessel-Schlesewsky. 2018. Toward a reliable, automated method of individual alpha frequency (IAF) quantification. *Psychophysiology* 55, 7 (2018), e13064. https://doi.org/10.1111/psyp.13064
[25] Raimundo da Silva Soares, Kevin L. Ramirez-Chavez, Altona Tufanoglu, Candida Barreto, João Ricardo Sato, and Hasan Ayaz. 2024. Cognitive Effort during Visuospatial Problem Solving in Physical Real World, on Computer Screen, and in Virtual Reality. *Sensors* 24, 3 (2024). https://doi.org/10.3390/s24030977
[26] F. Darvas, R. Scherer, J.G. Ojemann, R.P. Rao, K.J. Miller, and L.B. Sorensen. 2010. High gamma mapping using EEG. *NeuroImage* 49, 1 (2010), 930–938. https://doi.org/10.1016/j.neuroimage.2009.08.041
[27] Arindam Dey, Thammathip Piumsomboon, Youngho Lee, and Mark Billinghurst. 2017. Effects of Sharing Physiological States of Players in a Collaborative Virtual Reality Gameplay. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4045–4056. https://doi.org/10.1145/3025453.3026028
[28] Dennis Dietz, Carl Oechsner, Changkun Ou, Francesco Chiossi, Fabio Sarto, Sven Mayer, and Andreas Butz. 2022. Walk This Beam: Impact of Different Balance Assistance Strategies and Height Exposure on Performance and Physiological Arousal in VR. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology* (Tsukuba, Japan) *(VRST '22)*. Association for Computing Machinery, New York, NY, USA, Article 32, 12 pages. https://doi.org/10.1145/3562939.3567818
[29] Henry W. Dong, Caitlin Mills, Robert T. Knight, and Julia W. Y. Kam. 2021. Detection of mind wandering using EEG: Within and across individuals. *PLOS ONE* 16, 5 (05 2021), 1–18. https://doi.org/10.1371/journal.pone.0251490
[30] Andrew T. Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. The Index of Pupillary Activity: Measuring Cognitive Load vis-à-vis Task Difficulty with Pupil Oscillation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173856
[31] Stephen H. Fairclough. 2008. Fundamentals of physiological computing. *Interacting with Computers* 21, 1-2 (11 2008), 133–145. https://doi.org/10.1016/j.intcom.2008.10.011
[32] Johann Faouzi and Hicham Janati. 2020. pyts: A Python Package for Time Series Classification. *Journal of Machine Learning Research* 21 (2020).
[33] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. 2013. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience* 7 (2013). https://doi.org/10.3389/fnins.2013.00267
[34] Thalia Harmony. 2013. The functional significance of delta oscillations in cognitive processing. *Frontiers in Integrative Neuroscience* 7 (2013). https://doi.org/10.3389/fnint.2013.00083
[35] Thalía Harmony, Thalía Fernández, Juan Silva, Jorge Bernal, Lourdes Díaz-Comas, Alfonso Reyes, Erzsébet Marosi, Mario Rodríguez, and Miguel Rodríguez. 1996. EEG delta activity: an indicator of attention to internal processing during performance of mental tasks. *International Journal of Psychophysiology* 24, 1 (1996), 161–171. https://doi.org/10.1016/S0167-8760(96)00053-0 New Advances in EEG and cognition.
[36] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9
[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New York, NY, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90
[38] Wijnand IJsselsteijn, Yvonne de Kort, and Karolien Poels. 2013. *The Game Experience Questionnaire*. Technische Universiteit Eindhoven, Eindhoven, NB, The Netherlands.
[39] Yuying Jiang, Haoran Zhang, and Shan Yu. 2021. Changes in delta and theta oscillations in the brain indicate dynamic switching of attention between internal and external processing. In *4th International Conference on Biometric Engineering and Applications* (Taiyuan, China) *(ICBEA '21)*. Association for Computing Machinery, New York, NY, USA, 25–31. https://doi.org/10.1145/3476779.3476784
[40] Peter Kiefer, Ioannis Giannopoulos, Andrew Duchowski, and Martin Raubal. 2016. Measuring Cognitive Load for Map Tasks Through Pupil Diameter. In *Geographic Information Science*, Jennifer A. Miller, David O'Sullivan, and Nancy Wiegand (Eds.). Springer International Publishing, Cham, 323–337.
[41] Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/arXiv.1412.6980
[42] Wolfgang Klimesch. 2012. Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences* 16, 12 (2012), 606–617. https://doi.org/10.1016/j.tics.2012.10.007
[43] Effie L.-C. Law, Florian Brühlmann, and Elisa D. Mekler. 2018. Systematic Review and Validation of the Game Experience Questionnaire (GEQ) - Implications for Citation and Reporting Practice. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (Melbourne, VIC, Australia) *(CHI PLAY '18)*. Association for Computing Machinery, New York, NY, USA, 257–270. https://doi.org/10.1145/3242671.3242683
[44] Te-Won Lee, Mark Girolami, and Terrence Sejnowski. 1999. Independent Component Analysis Using an Extended Infomax Algorithm for Mixed Subgaussian and Supergaussian Sources. *Neural Computation* 11 (1999). https://doi.org/10.1162/089976699300016719
[45] Adam Li, Jacob Feitelberg, Anand Prakash Saini, Richard Höchenberger, and Mathieu Scheltienne. 2022. MNE-ICALabel: Automatically annotating ICA components with ICLabel in Python. *Journal of Open Source Software* 7, 76 (2022), 4484. https://doi.org/10.21105/joss.04484
[46] Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. 2021. The Perils and Pitfalls of Block Design for EEG Classification Experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (2021), 316–333. https://doi.org/10.1109/TPAMI.2020.2973153
[47] John Lisman. 2010. Working memory: the importance of theta and gamma oscillations. *Current Biology* 20, 11 (2010), R490–R492.
[48] Minjie Liu, Mingming Zhou, Tao Zhang, and Naixue Xiong. 2020. Semi-supervised learning quantization algorithm with deep features for motor imagery EEG Recognition in smart healthcare application. *Applied Soft Computing* 89 (2020), 106071. https://doi.org/10.1016/j.asoc.2020.106071

[49] Yisi Liu and Olga Sourina. 2014. *Real-Time Subject-Dependent EEG-Based Emotion Recognition Algorithm.* Springer Berlin Heidelberg, Berlin, Heidelberg, 199–223. https://doi.org/10.1007/978-3-662-43790-2_11

[50] F Lotte, L Bougrain, A Cichocki, M Clerc, M Congedo, A Rakotomamonjy, and F Yger. 2018. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of Neural Engineering* 15, 3 (apr 2018), 031005. https://doi.org/10.1088/1741-2552/aab2f2

[51] Elisa Magosso, Francesca De Crescenzio, Giulia Ricci, Sergio Piastra, and Mauro Ursino. 2019. EEG alpha power is modulated by attentional changes during cognitive tasks and virtual reality immersion. *Comp. Intelligence and Neuroscience* 2019 (2019). https://doi.org/10.1155/2019/7051079

[52] Kathryn M. McMillan, Angela R. Laird, Suzanne T. Witt, and M. Elizabeth Meyerand. 2007. Self-paced working memory: Validation of verbal variations of the n-back paradigm. *Brain Research* 1139 (2007), 133–142. https://doi.org/10.1016/j.brainres.2006.12.058

[53] Chukwuemeka Nwagu, Alaa AlSlaity, and Rita Orji. 2023. EEG-Based Brain-Computer Interactions in Immersive Virtual and Augmented Reality: A Systematic Review. *Proc. ACM Hum.-Comput. Interact.* 7, EICS, Article 174 (jun 2023), 33 pages. https://doi.org/10.1145/3593226

[54] Oskar Palinko, Andrew L. Kun, Alexander Shyrokov, and Peter Heeman. 2010. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications* (Austin, Texas) *(ETRA '10)*. Association for Computing Machinery, New York, NY, USA, 141–144. https://doi.org/10.1145/1743666.1743701

[55] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA.

[56] Rebecca Patient, Fawaz Ghali, Hoshang Kolivand, William Hurst, and Nigel John. 2021. Application of Virtual Reality and Electrodermal Activity for the Detection of Cognitive Impairments. In *2021 14th International Conference on Developments in eSystems Engineering (DeSE)*. IEEE, New York, NY, USA, 156–161. https://doi.org/10.1109/DeSE54285.2021.9719442

[57] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011).

[58] Peter Putman, Bart Verkuil, Elsa Arias-Garcia, Ioanna Pantazi, and Charlotte van Schie. 2014. EEG theta/beta ratio as a potential biomarker for attentional control and resilience against deleterious effects of stress on attention. *Cognitive, Affective, and Behavioral Neuroscience* 14 (2014). https://doi.org/10.3758/s13415-013-0238-7

[59] Felix Putze, Maximilian Scherer, and Tanja Schultz. 2016. Starring into the Void? Classifying Internal vs. External Attention from EEG. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (Gothenburg, Sweden) *(NordiCHI '16)*. ACM, New York, NY, USA, Article 47, 4 pages. https://doi.org/10.1145/2971485.2971555

[60] Giulia Ricci, Francesca De Crescenzio, Sandhya Santhosh, Elisa Magosso, and Mauro Ursino. 2022. Relationship between electroencephalographic data and comfort perception captured in a Virtual Reality design environment of an aircraft cabin. *Scientific Reports* 12, 1 (2022), 10938.

[61] James B Rowe, Ivan Toni, Oliver Josephs, Richard SJ Frackowiak, and Richard E Passingham. 2000. The prefrontal cortex: response selection or maintenance within working memory? *Science* 288, 5471 (2000), 1656–1660. https://doi.org/10.1126/science.288.5471.1656

[62] Mikko Salminen, Simo Järvelä, Antti Ruonala, Ville J Harjunen, Juho Hamari, Giulio Jacucci, and Niklas Ravaja. 2019. Evoking physiological synchrony and empathy using social VR with biofeedback. *IEEE Transactions on Affective Computing* 13, 2 (2019), 746–755.

[63] Claudia Sannelli, Mikio Braun, and Klaus-Robert Müller. 2009. Improving BCI performance by task-related trial pruning. *Neural Networks* 22, 9 (2009), 1295–1304.

[64] Christina Schneegass, Max L Wilson, Jwan Shaban, Jasmin Niess, Francesco Chiossi, Teodora Mitrevska, and Paweł W Woźniak. 2024. Broadening the mind: how emerging neurotechnology is reshaping HCI and interactive system design. *i-com* 0 (2024). https://doi.org/10.1515/icom-2024-0007

[65] Mansi Sharma, Shuang Chen, Philipp Müller, Maurice Rekrut, and Antonio Krüger. 2023. Implicit Search Intent Recognition using EEG and Eye Tracking: Novel Dataset and Cross-User Prediction. In *Proceedings of the 25th International Conference on Multimodal Interaction* (Paris, France) *(ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 345–354. https://doi.org/10.1145/3577190.3614166

[66] Mansi Sharma, Maurice Rekrut, Jan Alexandersson, and Antonio Krüger. 2023. Towards Improving EEG-Based Intent Recognition in Visual Search Tasks. In *Neural Information Processing: 29th International Conference, ICONIP 2022, Virtual Event, November 22–26, 2022, Proceedings, Part III* (New Delhi, India). Springer-Verlag, Berlin, Heidelberg, 604–615. https://doi.org/10.1007/978-3-031-30111-7_51

[67] Sokkeang Try, Kriengsak Panuwatwanich, Ganchai Tanapornraweekit, and Manop Kaewmoracharoen. 2021. Virtual reality application to aid civil engineering laboratory course: A multicriteria comparative study. *Computer Applications in Engineering Education* 29, 6 (2021), 1771–1792. https://doi.org/10.1002/cae.22422

[68] Tom Verguts. 2017. Binding by random bursts: A computational model of cognitive control. *Journal of cognitive neuroscience* 29, 6 (2017), 1103–1118.

[69] Lisa-Marie Vortmann, Simon Ceh, and Felix Putze. 2022. Multimodal EEG and Eye Tracking Feature Fusion Approaches for Attention Classification in Hybrid BCIs. *Frontiers in Computer Science* 4 (2022). https://doi.org/10.3389/fcomp.2022.780580

[70] Lisa-Marie Vortmann, Jannes Knychalla, Sonja Annerer-Walcher, Mathias Benedek, and Felix Putze. 2021. Imaging Time Series of Eye Tracking Data to Classify Attentional States. *Frontiers in Neuroscience* 15 (2021). https://doi.org/10.3389/fnins.2021.664490

[71] Lisa-Marie Vortmann, Felix Kroll, and Felix Putze. 2019. EEG-Based Classification of Internally- and Externally-Directed Attention in an Augmented Reality Paradigm. *Frontiers in Human Neuroscience* 13 (2019). https://doi.org/10.3389/fnhum.2019.00348

[72] Lisa-Marie Vortmann and Felix Putze. 2021. Combining Implicit and Explicit Feature Extraction for Eye Tracking: Attention Classification Using a Heterogeneous Input. *Sensors* 21, 24 (2021). https://doi.org/10.3390/s21248205

[73] Lisa-Marie Vortmann and Felix Putze. 2021. Exploration of Person-Independent BCIs for Internal and External Attention-Detection in Augmented Reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 80 (jun 2021), 27 pages. https://doi.org/10.1145/3463507

[74] Lisa-Marie Vortmann, Moritz Schult, Mathias Benedek, Sonja Walcher, and Felix Putze. 2019. Real-Time Multimodal Classification of Internal and External Attention. In *Adjunct of the 2019 International Conference on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, Article 14, 7 pages. https://doi.org/10.1145/3351529.3360658

[75] Zhiguang Wang and Tim Oates. 2015. Imaging Time-Series to Improve Classification and Imputation. https://doi.org/10.48550/arXiv.1506.00327

[76] P. Welch. 1967. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics* 15, 2 (1967), 70–73. https://doi.org/10.1109/TAU.1967.1161901

[77] Mahdi Zarour, Hamdi Ben Abdessalem, and Claude Frasson. 2023. Distraction Detection and Monitoring Using Eye Tracking in Virtual Reality. In *Augmented Intelligence and Intelligent Tutoring Systems*, Claude Frasson, Phivos Mylonas, and Christos Troussas (Eds.). Springer Nature Switzerland, Cham, 491–503.

[78] Guangyi Zhang and Ali Etemad. 2023. Distilling EEG representations via capsules for affective computing. *Pattern Recognition Letters* 171 (2023), 99–105.

[79] Rongxiang Zhang. 2020. Virtual Reality Games based on Brain Computer Interface. In *2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*. IEEE, New York, NY, USA, 227–230. https://doi.org/10.1109/ICHCI51889.2020.00056