

Multimodal Contrastive Learning for Cybersickness Recognition Using Brain Connectivity Graph Representation

Peike Wang , Ming Li , Ziteng Wang , Yong-Jin Liu , and Lili Wang * 

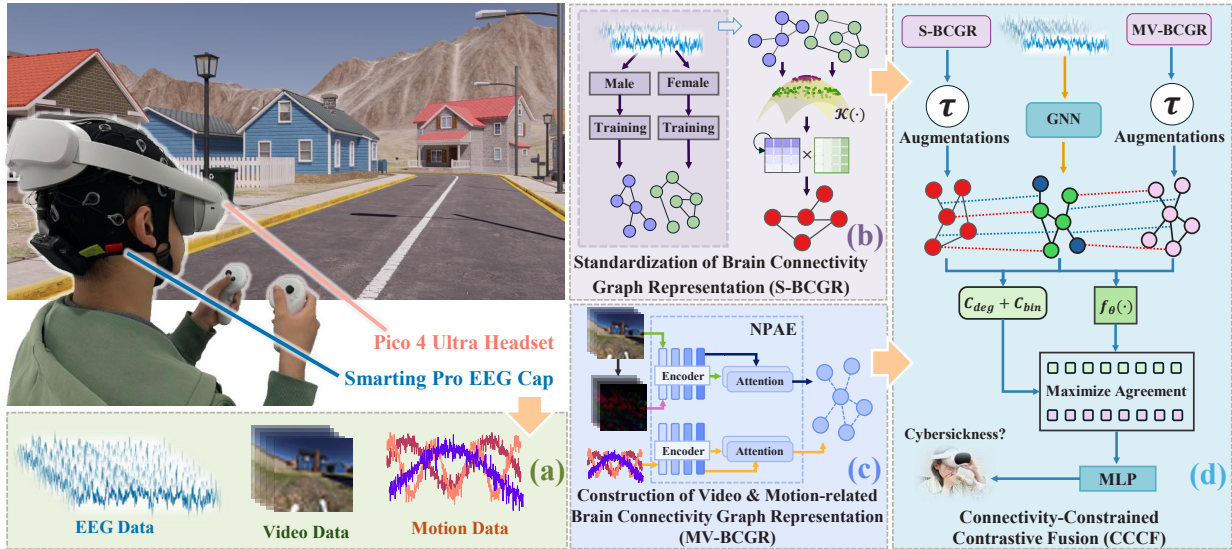


Fig. 1: The proposed method comprises four parts: (a) Multimodal dataset collection and construction, including EEG, video, and motion data; (b) Construction of E-BCGR based on EEG Data, S-BCGR through the proposed SDA; (c) Construction of MV-BCGR based on video and motion data through the proposed NPAAE; and (d) CCCF module, which projects multimodal data into a unified latent space while constraining the connectivity by using S-BCGR.

Abstract—Cybersickness significantly impairs user comfort and immersion in virtual reality (VR). Effective identification of cybersickness leveraging physiological, visual, and motion data is a critical prerequisite for its mitigation. However, current methods primarily employ direct feature fusion across modalities, which often leads to limited accuracy due to inadequate modeling of inter-modal relationships. In this paper, we propose a multimodal contrastive learning method for cybersickness recognition. First, we introduce Brain Connectivity Graph Representation (BCGR), an innovative graph-based representation that captures cybersickness-related connectivity patterns across modalities. We further develop three BCGR instances: E-BCGR, constructed based on EEG signals; MV-BCGR, constructed based on video and motion data; and S-BCGR, obtained through our proposed standardized decomposition algorithm. Then, we propose a connectivity-constrained contrastive fusion module, which aligns E-BCGR and MV-BCGR into a shared latent space via graph contrastive learning while utilizing S-BCGR as a connectivity constraint to enhance representation quality. Moreover, we construct a multimodal cybersickness dataset comprising synchronized EEG, video, and motion data collected in VR environments to promote further research in this domain. Experimental results demonstrate that our method outperforms existing state-of-the-art methods across four critical evaluation metrics: accuracy, sensitivity, specificity, and the area under the curve. Source code: <https://github.com/PEKEW/cybersickness-bcgr>.

Index Terms—Virtual Reality, Cybersickness, Electroencephalography, Brain Connectivity Representation, Contrastive Learning

* Corresponding author

- Peike Wang, Ziteng Wang, and Lili Wang are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, 100191, Beijing, China. E-mails: {pekewang, zitengwang, wanglily}@buaa.edu.cn
- Ming Li and Yong-Jin Liu are with the Department of Computer Science and Technology, Tsinghua University, 100084, Beijing, China. E-mails: {mingli_thu, liuyongjin}@tsinghua.edu.cn

Received 11 April 2025; revised 18 June 2025; accepted 22 July 2025.

Date of publication 6 October 2025; date of current version 3 November 2025.

Digital Object Identifier no. 10.1109/TVCG.2025.3616797

1 INTRODUCTION

In recent years, virtual reality (VR) has achieved substantial advancement and been adopted across diverse domains, including professional training [43], education [34], and medical applications [7]. Despite its widespread adoption, cybersickness remains a critical barrier to user comfort and immersion [27]. The emergence of deep learning has attracted increasing attention in data-driven methods for cybersickness recognition. Video data-based methods are widely adopted due to their direct impact on visual stimuli. Kim et al. [16] employed convolutional autoencoders to detect anomalous motion in VR videos, demonstrating the feasibility of video-content-based cybersickness recognition. Another emerging methodology involves cybersickness recognition through objective physiological monitoring, which offers improved interpretability and has gained increasing adoption in research communities [4].

However, single-modality methods fail to capture the features as-

sociated with cybersickness comprehensively. To overcome such limitations, multimodal fusion has emerged as a promising alternative. Demirel et al. [6] combined electroencephalography (EEG) and head motion data using multitaper spectral estimation and LSTM networks for real-time recognition. More recently, Jeong et al. [15] improved personalization by incorporating attention-based fusion of video, eye movement, head movement, and galvanic skin response. Nevertheless, most fusion strategies still rely on simple feature concatenation or stacking, without effectively modeling inter-modal synergy, resulting in limited recognition performance.

To address these challenges, we propose a multimodal contrastive learning method for cybersickness recognition. Specifically, we introduce a novel Brain Connectivity Graph Representation (BCGR) to model connectivity patterns across multiple modalities. Our selection of EEG, video, and motion modalities is grounded in the neurophysiological mechanisms underlying cybersickness. EEG provides direct access to central neural responses and has been shown to capture characteristic brain activity associated with cybersickness. Video records the actual visual stimuli experienced by participants, which are key triggers of cybersickness under the sensory conflict theory. Motion data quantifies physical movement and approximates vestibular input, allowing the modeling of discrepancies between perceived and actual motion. We construct three specialized instances: E-BCGR, based on EEG signals; MV-BCGR, based on video and motion data; and S-BCGR, constructed using a standardized decomposition algorithm (SDA). To fuse these representations, we develop a connectivity-constrained contrastive fusion (CCCF) module, which aligns E-BCGR and MV-BCGR in a shared latent space under the structural constraint of S-BCGR. To support our method, we construct a new multimodal cybersickness dataset, which includes synchronized EEG, video, and motion data collected in VR.

To validate the effectiveness of our method, we conducted extensive experiments. Results show that our approach outperforms state-of-the-art (SOTA) methods, achieving improvements of ranging from 0.54% to 11.95% in accuracy, 0.11% to 18.04% in specificity, and 0.86% to 18.71% in Area Under the Curve (AUC). The dataset and source code will be made publicly available to support future research. An overview of the proposed method is provided in Fig. 1, highlighting its multimodal and contrastive design for enhanced cybersickness recognition. In summary, the main contributions of this paper are as follows:

1. A multimodal contrastive learning method for cybersickness recognition that aligns modality-specific features through graph-based contrastive learning. The proposed method outperforms the current SOTA methods regarding accuracy, sensitivity, specificity, and AUC.
2. A novel Brain Connectivity Graph Representation, BCGR, comprising three instances: E-BCGR from EEG data, MV-BCGR from video and motion data, and S-BCGR from the standardized decomposition algorithm. BCGR provides a structured and physiologically interpretable representation space for multimodal cybersickness recognition.
3. A Connectivity-Constrained Contrastive Fusion module, CCCF, that projects different modalities into a shared latent space by constraints of the S-BCGR. This design improves cross-modal representation alignment and strengthens the discriminative capacity of the learned representations.
4. A high-quality, synchronized multimodal dataset including EEG, video, and motion data collected concurrently in immersive VR settings, facilitating future cybersickness research.

2 RELATED WORK

In the following section, we briefly discuss several studies that are closely related to our work.

2.1 Cybersickness

According to sensory conflict theory, cybersickness arises from a mismatch between perceived motion (visual) and actual physical movement (vestibular) [27]. Researchers have investigated a range of data modalities and modeling methods, including video data-based recognition,

physiological signal-based recognition, motion data-based recognition, and multimodal methods.

2.2 Video data-based Cybersickness Recognition

Visual elements in VR environments often directly influence motion perception, leading to sensory conflicts. As a result, visual data are widely used to recognize cybersickness. Kim et al. [16] demonstrated the potential of using VR video content alone to recognize cybersickness by employing convolutional autoencoders to extract spatiotemporal motion features. Padmanaban et al. [29] incorporated depth information from VR videos, investigated interactions between depth and velocity, and used aggregated statistical features for recognition. Lee et al. [20] further improved recognition accuracy by extracting key features from VR videos, rather than relying solely on the original video stream. Specifically, they modeled eye movements, velocity, and depth using video saliency, optical flow, and parallax maps. A 3D-CNN model was then employed for cybersickness recognition.

However, most current video-based recognition methods overlook individual variability in cybersickness responses, leading to reduced recognition accuracy in practical scenarios.

2.3 Physiological-based Cybersickness Recognition

Physiological modalities, such as galvanic skin response (GSR) and EEG, can directly reflect changes in physiological state during cybersickness. Prior research has established that cybersickness involves disrupted coordination among multiple brain regions, with EEG providing direct insight into these neural disruptions and showing strong correlations with cybersickness severity [3]. Researchers leverage variations in these physiological data to predict or detect cybersickness. Jeong et al. [14] compared DNN and CNN for EEG-based cybersickness recognition, demonstrating that EEG signals effectively achieve accurate recognition. Liao et al. [24] transformed EEG signals into power spectral density features using Fourier transforms, and applied them with LSTM models. Their results indicated that this approach significantly outperformed traditional models across various metrics. Tasnim et al. [40] integrated multiple physiological signals—including heart rate, galvanic skin response, and eye movement—with individual factors such as gender. They used personalization techniques, found that personalized physiological models notably surpassed non-personalized models in cybersickness recognition accuracy.

Although physiological signals have been widely used for cybersickness recognition, such single physiological modality methods are prone to noise and do not directly reflect users' responses to visual content, which results in lower recognition accuracy.

2.4 Motion-based Cybersickness Recognition

Motion data directly reflects users' motion feedback in virtual environments and is closely related to vestibular conflict. Chang et al. [2] used eye movement data and support vector machines to predict cybersickness based on the subjective vertical mismatch theory. Their results showed that eye movements in VR can serve as important indicators for detecting cybersickness. Feigl et al. [9] explored the relationship between motion parameters and cybersickness by employing position and orientation estimation methods in large-scale VR environments. Their findings demonstrated that motion parameters can be effectively used for cybersickness recognition. Guo et al. [11] investigated the predictive potential of measurable optokinetic after-convulsive nystagmus (OKAN) parameters and experimentally confirmed a correlation between OKAN and susceptibility to cybersickness, identifying OKAN as a possible objective marker for recognizing individual susceptibility.

Although motion-based methods have shown promising results, relying solely on motion parameters fails to capture the full complexity of virtual environments. Features such as orientation, position, and eye movements are often weakly linked with visual and scene-level information, making it difficult to reflect the underlying causes of cybersickness. Consequently, models based exclusively on motion data tend to exhibit limited recognition accuracy.

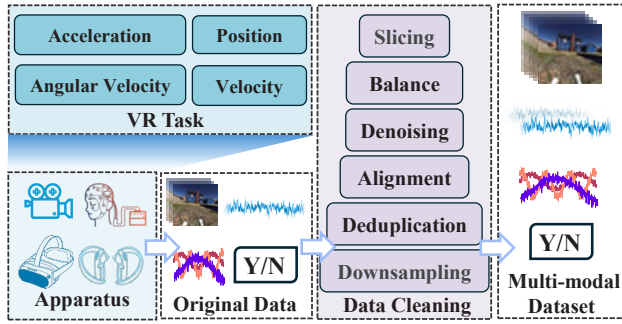


Fig. 2: Pipeline of Dataset Collection and Cleaning.



Fig. 3: One view from the designed virtual environments.

2.5 Multimodal Cybersickness Recognition

Multimodal integration leverages the complementary strengths of different modalities, potentially addressing the noise and limited adaptability associated with single-modality recognitions. By synthesizing the relationships between visual stimuli and physiological responses, multimodal fusion enables more comprehensive and accurate cybersickness recognition. Kim et al. [17] employed a CNN to encode EEG signals into cognitive representations, which were then transferred to video features encoded by a recurrent neural network, demonstrating the approach's effectiveness. Islam et al. [13] independently encoded video, eye movement, and head movement data using 3D-CNN and CNN-LSTM networks, respectively, and integrated them via a deep fusion connectivity layer. Their results confirmed the effectiveness of multimodal fusion, particularly highlighting the predictive value of eye and head movements. Jeong et al. [15] introduced an attention mechanism to integrate temporal features from video content, eye movements, head movements, and galvanic skin responses, emphasizing the importance of both data modality and user-specific characteristics. Chang et al. [33] combined head motion, eye movements, visual complexity, and immersion duration within an LSTM-based architecture, achieving improved recognition accuracy. Similarly, Demirel et al. [6] proposed an LSTM-based model that integrated head-motion data with EEG signals processed through multitaper spectral estimation, enabling real-time cybersickness recognition.

Despite recent advances, many existing fusion methods overlook the intrinsic mechanisms of features from different modal interactions. Most rely on simple concatenation or stacking of features, without effectively modeling the synergy between modalities. In addition, limited consideration of individual differences reduces the model's ability to adapt to diverse users. These limitations collectively hinder the improvement of recognition accuracy in multimodal cybersickness recognition. Compared to other modality combinations, our EEG + video + motion approach offers distinct advantages. For example, EDA + motion methods [23] capture autonomic responses but lack direct neural connectivity information. EEG + motion approaches [6] omit critical visual content. Head/eye-tracking with visual complexity [33] focuses on oculomotor behavior but overlooks broader neural patterns. Our three-modality framework addresses these gaps by integrating neural connectivity (EEG), visual stimuli (video), and movement dynamics (motion) into a unified representation space.

3 MULTIMODAL CYBERSICKNESS DATASET CONSTRUCTION

To validate our methods, we first constructed a multimodal cybersickness dataset comprising EEG signals, video recordings, and motion-related data collected during participants' interactions in immersive virtual environments.

Our dataset collection and construction workflow is illustrated in Fig. 2. Firstly, we designed a virtual task intended to induce cybersickness. Participants were asked to complete the task while three modalities of data were simultaneously collected using the corresponding apparatus. Next, the raw data underwent a rigorous cleaning and

preprocessing procedure, resulting in a high-quality multimodal dataset suitable for cybersickness recognition.

3.1 Virtual Environment Design

We designed an immersive virtual environment using Unity with configurable motion parameters to reliably induce cybersickness and facilitate multimodal data collection. The environment featured an outdoor scene measuring $25 \times 25 \text{ m}$ containing randomly positioned red target spheres, as shown in Figure 3. Participants navigated this environment using joystick buttons to locate 20 target spheres. To induce cybersickness, we manipulated key motion parameters—speed (0.01–5.00 m/s), acceleration (0.01–5.00 m/s²), and steering speed ($\pi/4$ – 2π rad/s)—via a custom-designed UI panel. These parameters were selected based on prior studies demonstrating their effectiveness in triggering cybersickness [31, 39, 41]. We conducted a pilot test ($n = 9$) to refine the parameter ranges, ensuring they reliably induced symptoms while maintaining a sense of immersion.

3.2 Apparatus

The VR environment and interactions were delivered through Pico 4 Ultra head-mounted displays and two handheld controllers. EEG signals were recorded using the Smarting Pro 32 device (mBrainTrain, Serbia) [12, 44], which includes a built-in real-time stream recording tool.

Signals were captured at a sampling rate of 500 Hz, with electrode impedance maintained below 5 k Ω throughout the experiment to ensure high-quality acquisition. The device recorded data locally, while a Bluetooth 5.0 connection to a laptop was used for monitoring purposes during the experiment. Real-time video recordings were captured from the participant's first-person perspective via the Pico developer platform, recorded at 60 FPS in 3840 $\times$ 2160 resolution. These recordings reflect the dynamic visual stimuli experienced during VR navigation and serve as a key modality for analyzing visual-motion relationships in cybersickness recognition. Motion data were logged using Unity's built-in interface at a frequency of 30 Hz. The logged parameters captured participants' movement and orientation dynamics during VR interaction, and included information such as immersion time, position, velocity, acceleration, and rotational motion.

To achieve accurate multimodal synchronization, hardware-triggered markers were used to align EEG, video, and motion data at millisecond-level precision. Data cleaning and post-processing were performed on a workstation equipped with a 3 GHz Intel Core i9-13900K CPU, 64 GB RAM, and an NVIDIA GeForce GTX 4080 GPU. The Pico 4 Ultra HMD rendered VR content at 90 Hz per eye with a binocular field of view of 105°, delivering a smooth and immersive visual experience.

3.3 Participants

The experimental protocol was approved by the Biology and Medical Ethics Committee of the local university and complied with the ethical guidelines outlined in the Declaration of Helsinki. 29 healthy adults (17 males, 12 females), aged 21–26 years (mean age: 23.10 ± 1.49), were recruited through campus announcements. All participants had normal hearing, normal or corrected-to-normal vision, and no history

of neurological or psychiatric disorders. Appointments were scheduled through an online booking system, and all participants signed an informed consent form detailing the EEG signal acquisition procedures.

To minimize EEG artifacts potentially caused by novelty, only individuals with prior VR experience (e.g., playing VR games or watching VR videos) were selected. Participants were screened using the Motion Sickness Susceptibility Questionnaire (MSSQ) [10], and those with scores above the 90th percentile or below the 10th percentile were excluded to ensure data reliability by avoiding extreme susceptibility outliers. To enhance immersion, participants adjusted their height within the virtual environment based on their own vicarious experience prior to the experiment. Upon completion, each participant received a cash remuneration of \$13.83.

3.4 Data Collection Procedure

After completing hardware setup and calibration, participants were instructed to begin the experiment at their convenience. Before the cybersickness induction experiment, each participant completed a 5-minute baseline session in which they remained at rest while viewing a black screen. All experimental data were baseline-corrected to account for individual differences. To initiate the session and synchronize data recording, participants pressed the trigger button on the right-hand controller, then proceeded with the task. Participants were instructed to document the onset of cybersickness symptoms by pressing and holding the right trigger button from the moment symptoms appeared until they subsided, thus accurately capturing the corresponding time interval. Each session lasted approximately 5–7 minutes. Minor variations in duration were due to randomized navigation parameters, the distribution of target objects, and participants' individual route choices within the virtual environment. Participants could withdraw at any point if discomfort occurred. One participant reported severe discomfort during the experiment and was unable to complete the task; their data were excluded from the dataset. Our final analysis includes only participants who completed the session.

The Simulator Sickness Questionnaire (SSQ) [1] and other assessments requiring task interruption were intentionally omitted to maintain ecological validity of the immersive VR experience and reduce EEG recording artifacts caused by mid-session disruptions. Although recent studies have examined continuous severity prediction using EEG, such approaches often depend on cognitively demanding self-assessments or attention shifts, potentially disrupting natural neural responses and compromising EEG signal quality. Some studies employ joystick-based paradigms that allow participants to conveniently report symptoms during predefined paths [6]; however, our experimental design requires users to actively navigate and interact within the VR environment, rather than passively follow scripted trajectories. This setup more closely reflects real-world VR usage, but also makes in-task assessment more intrusive and less compatible with EEG monitoring. In contrast, our study employs an alternative strategy, focusing instead on detecting the presence or absence of cybersickness, thereby ensuring uninterrupted data collection and improved EEG fidelity.

Regarding data recording, video data were saved in '.mp4' format, EEG data in '.xdf', and motion data in '.csv'. Cybersickness reports were synchronized across all modalities using hardware-triggered markers, which served as alignment points during segmentation. For each cybersickness episode, a pair of 8-bit globally unique markers was generated to indicate onset and offset. All data were recorded in real time, and a canonical file-naming scheme was used to ensure correct association across modalities. Following data collection, cleaning procedures were applied to guarantee data quality and maintain temporal alignment across all sources.

3.5 Data Cleaning Procedure

In the raw data processing stage, the data were first screened to remove low-quality and missing data to ensure overall integrity and reliability. Low-quality samples were primarily attributed to signal artifacts, motion-induced distortions, or deviations from expected viewing behavior (e.g., abnormal head postures). Missing data occurred when specific segments or entire modalities were not properly recorded, often

due to temporary software or device-side interruptions. To achieve multimodal synchronization, timestamp alignment was applied, with interpolation performed as needed. Outliers were removed using the interquartile range (IQR) method [42], excluding values falling below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$. Finally, all temporal data were resampled to a uniform rate to ensure consistency across modalities.

EEG data preprocessing was performed using EEGLAB [5]. The data were resampled to 128 Hz and re-referenced to the mastoid electrodes (TP9 and TP10) to reduce reference-related effects. A 0.1–30 Hz band-pass filter was applied to eliminate interference from the monitor's intermediate frequency signals (50–60 Hz) and the VR HMD (90 Hz). High-amplitude artifacts ($\geq 100 \mu V$), unlikely to reflect neural activity, were identified and removed through visual inspection. Independent component analysis [21] was then used to decompose the EEG signals into independent components, enabling artifact rejection based on spatial distribution and spectral characteristics. The cleaned EEG data were aligned with video recordings and segmented into 1-second intervals. Segments with excessively high-frequency noise were excluded, along with their corresponding video segments, to preserve multimodal consistency.

Video data were first segmented into 1-second intervals and aligned with EEG and motion data using hardware-synchronized timestamps. To improve computational efficiency, the frame rate was standardized to 30 FPS, and the resolution was resized to 256×256 . Invalid segments, such as those containing black screens or motion parameter setting interfaces, were manually identified and removed through visual inspection to ensure temporal consistency.

Motion data processing involved extracting features from the log records, including VR immersion time (double), position (vector3), velocity (vector3), acceleration (vector3), and rotational velocity (vector4). The motion data were downsampled to 30 records per second. To standardize the log format, all files were converted to UTF-8 encoding, and labels or special characters were removed. The motion data were then aligned with the video and EEG samples and segmented into 1-second intervals. Invalid log samples, such as those containing missing text, were discarded along with their corresponding EEG and video data to ensure precise temporal alignment across all modalities.

Cybersickness labels (boolean) were extracted from the log records. Unmatched or invalid labels were removed to prevent tagging inaccuracies due to misalignment. The synchronized dataset was then validated to ensure that all multimodal data within each time segment were complete and correctly matched. To address potential class imbalance, stratified sampling was applied during the train-test split to preserve a representative distribution of cybersickness events [38]. The entire preprocessing pipeline was automated using Python and MATLAB scripts.

4 MULTIMODAL CONTRASTIVE LEARNING FOR CYBERSICKNESS RECOGNITION

4.1 BCGR definition

In this study, we define the BCGR as a task-oriented, modality-compatible representation of brain connectivity tailored to cybersickness recognition. Specifically, BCGR is formulated as a symmetric graph (30×30), where each node corresponds to an electrode in the 10–20 system, a standard EEG layout that approximates the spatial topology of brain regions, and each edge encodes the connectivity patterns relevant to cybersickness. It is defined as:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W}) \quad (1)$$

where $\mathcal{V} = v_1, v_2, \dots, v_{30}$ represents the nodes used for graph construction. Edges $\mathcal{E}$ denote the connections between nodes, defined as $\mathcal{E} \subseteq (u, v) \mid u, v \in \mathcal{V}, u \neq v$, and are symmetric, i.e., $(u, v) \in \mathcal{E} \iff (v, u) \in \mathcal{E}$. $\mathcal{W} \in \mathbb{R}^{30 \times 30}$ is the adjacency matrix representing the connection strength associated with cybersickness.

The construction of BCGRs employs diverse data-driven strategies, including neural networks and matrix factorization, tailored to different data sources and design goals. We present three representative

instances. E-BCGR is derived from EEG signals and captures functional connectivity patterns associated with cybersickness. MV-BCGR is constructed from video and motion data using data-driven techniques, with its structure informed by neurophysiological principles of brain connectivity. In contrast, S-BCGR is a standardized representation from multiple E-BCGRs via the SDA, aiming to suppress individual-specific noise while preserving features relevant to cybersickness.

Traditional brain connectivity representation typically captures physiological connectivity patterns derived directly from EEG or fMRI [22]. In contrast, BCGR focuses on recognition task-level representation features through learning or multimodal integration.

4.2 The pipeline of proposed method

As illustrated in Fig. 1, the proposed method consists of four main components, designed to model and integrate multimodal connectivity patterns for cybersickness recognition. (a) First, a multimodal dataset is constructed by collecting EEG, video, and motion data during immersive VR experiences. (b) To extract neural connectivity features, E-BCGR is constructed from EEG signals grouped by gender, capturing individual-specific patterns related to cybersickness. A standardized representation, S-BCGR, is further derived from multiple E-BCGRs using the SDA algorithm to isolate features that are consistent across individuals while suppressing idiosyncratic noise. (c) In parallel, MV-BCGR is generated from video and motion modalities using the Neural Prior Attention Encoder (NPAE), which enhances connectivity-relevant regions informed by neurophysiological priors. (d) Finally, these representations are integrated through the CCCF module, which aligns E-BCGR and MV-BCGR in a shared latent space, guided by the structure of S-BCGR. This fusion yields enhanced multimodal features that support more accurate cybersickness recognition.

4.3 The Construction of E-BCGR and S-BCGR

4.3.1 The construction of E-BCGR

We grouped EEG data by gender and trained two separate Dynamic Graph Convolutional Networks (DGCNNs) to construct gender-specific E-BCGRs. DGCNNs are particularly effective at capturing local spatial structures while allowing the graph topology to adapt dynamically during training [30]. Gender-based grouping was adopted to reduce intra-group variability and improve the consistency of connectivity patterns within each group. This decision is supported by prior research indicating that gender is a significant factor influencing susceptibility to cybersickness [19, 32, 37]. The resulting group-specific E-BCGRs provide a stable and representative basis for the construction of the standardized S-BCGR.

After training, we extracted the edge and node weights from the DGCNNs and reorganized them according to the standard 10-20 EEG electrode system, producing two sets of E-BCGRs. Since the DGCNN produces individualized representations for each sample, we aggregated these across the group by summing along the sample dimension. The final E-BCGR representation for each group (m) is defined as follows:

$$\mathcal{G}_E^{(m)} = (\mathcal{V}, \mathcal{E}, W_E^{(m)}) \quad (2)$$

where $m \in \{\text{male}, \text{female}\}$ denotes the group index, and the aggregated edge weight matrix $W_E^{(m)}$ is obtained by:

$$W_E^{(m)} = \left(\sum_{i=1}^{N_m} E_i^{(m)} + (E_i^{(m)})^\top \right) / 2 \quad (3)$$

where $E_i^{(m)}$ denotes the adjacent matrix of the i -th sample within group m , and N_m is the number of samples in this group.

4.3.2 The construction of S-BCGR

To extract S-BCGR from the group-specific E-BCGRs, we aim to identify common connectivity patterns that are consistent across groups

and likely reflect cybersickness-related components, while disentangling personalized-specific variations. As detailed in Algorithm 1, the algorithm takes as input two E-BCGR edge matrices from distinct groups, m_1 and m_2 , denoted as $W_E^{(m_1)}$ and $W_E^{(m_2)}$. To capture nonlinear relationships in the original space, we first map these matrices into a reproducing kernel Hilbert space using a Gaussian kernel $\Phi(\cdot)$, yielding $K_1 = \Phi(W_E^{(m_1)})$ and $K_2 = \Phi(W_E^{(m_2)})$ (line 1). To preserve physiological plausibility and interpretability,

Algorithm 1 SDA based S-BCGR construction

Require: Edge matrices $W_E^{(m_1)}, W_E^{(m_2)} \in \mathbb{R}^{n \times n}$

- 1: **Compute Kernel Matrices:** $K_1 = \Phi(W_E^{(m_1)})$, $K_2 = \Phi(W_E^{(m_2)})$
- 2: **Initialize:** $W, H_1, H_2 \in \mathbb{R}^{n \times n}$ with non-negative values, $\gamma \leftarrow 1 \times 10^{-3}$, $\varepsilon \leftarrow 1 \times 10^{-3}$
- 3: **while** not converged **do**
- 4: $N \leftarrow K_1 H_1^\top + K_2 H_2^\top$
- 5: $D \leftarrow (H_1 H_1^\top + H_2 H_2^\top) W$
- 6: $W \leftarrow W \cdot \frac{N}{D + \varepsilon}$
- 7: $N_{H_1} \leftarrow W^\top K_1$
- 8: $D_{H_1} \leftarrow W^\top W H_1 + \gamma H_2 H_1^\top H_1$
- 9: $H_1 \leftarrow H_1 \cdot \frac{N_{H_1}}{D_{H_1} + \varepsilon}$
- 10: $N_{H_2} \leftarrow W^\top K_2$
- 11: $D_{H_2} \leftarrow W^\top W H_2 + \gamma H_1 H_2^\top H_2$
- 12: $H_2 \leftarrow H_2 \cdot \frac{N_{H_2}}{D_{H_2} + \varepsilon}$
- 13: Check convergence (maximum iteration)
- 14: **end while**
- 15: **return** $\mathcal{G}_S = (\mathcal{V}, \mathcal{E}, \frac{W + W^\top}{2})$

we retain the original matrix dimensionality ($W, H_1, H_2 \in \mathbb{R}^{30 \times 30}$) and enforce non-negativity constraints. This design ensures that connection strengths remain biologically meaningful, reflecting the non-negative nature of neural connectivity, and that the extracted components remain comparable to the original connectivity structures. The objective function is formulated as follows:

$$\argmin_{W, H_1, H_2} \|K_1 - W H_1\|_F^2 + \|K_2 - W H_2\|_F^2 + \gamma \text{tr}(H_1^\top H_1 H_2^\top H_2) \quad (4)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, γ is the regularization parameter to balance reconstruction quality and sparsity. $\text{tr}(\cdot)$ represents the trace operator. The trace-based canonical term quantifies covariance similarity between personalized components, enforcing feature-space orthogonality between H_1 and H_2 to minimize interference with standard representation extraction.

The iterative optimization process, inspired by the Non-negative Matrix Factorization algorithm, employs multiplicative update rules and consists of the following key steps:

1. Standard component update (lines 4-6): W is updated by aggregating reconstructions, ensuring it captures standardized representations.
2. Personalized component update (lines 7-12): Each H_i is updated through two competing forces. The first encourages alignment with the corresponding group's kernel structure ($W^\top K_i$), while the second penalizes cross-group covariance via the regularization term $\gamma \text{tr}(H_1^\top H_1 H_2^\top H_2)$.
3. Convergence check (line 13): Terminate when the maximum number of iterations is reached ($T = 300$ in this study).

4.4 The Construction of MV-BCGR

To encode video and action modalities into the same graph structure as BCGR for the rational integration of multimodal data while leveraging neurobiological priors, we design the NPAE. The NPAE employs

neural prior knowledge to enhance specific areas' activations through attentional mechanisms, based on how different modalities stimulate different brain areas in evoking cybersickness. Specifically, MV-BCGR is constructed by mapping video and motion information to the adjacency matrix of the BCGR, denoted as W_{MV} , via the NPAE:

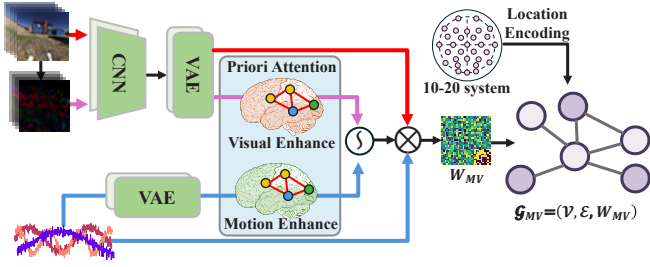


Fig. 4: The detailed architecture of the proposed NPAE. The input is the video data, the optical flow data, and motion data obtained from the video data computation are encoded by the encoder, and priority attention and output MV-BCGR.

$$f_{NPAE} : (D_v, D_o, D_m) \rightarrow W_{MV} \quad (5)$$

$$\mathcal{G}_{MV} = (\mathcal{V}, \mathcal{E}, W_{MV}) \quad (6)$$

where $D_v \in \mathbb{R}^{t \times 3 \times h \times w}$ represents three channel video modality data with dimensions of time (t), height (h), and width (w), $D_o \in \mathbb{R}^{t \times 1 \times h \times w}$ represents optical flow data with dimensions of time (t), height (h), and width (w), $D_m \in \mathbb{R}^{t \times n}$ represents motion modality data with dimensions of time (t) and feature number (n).

The architecture of NPAE is illustrated in Fig. 4. NPAE takes video and motion modality data as input and outputs MV-BCGR through the encoding and attention-guided mapping process. Initial modality data undergoes encoding to extract compact and discriminative representations. Video frames and optical flow are initially processed by a CNN, then further compressed by a variational autoencoder [18]. The extracted optical flow enriches the representation of dynamic visual information, which has been demonstrated to enhance cybersickness recognition performance [8].

To emulate the brain responses to distinct sensory stimuli, we introduce a neuro-guided prior attention module. This design is motivated by neuroscientific evidence that different types of input, such as visual motion or motor-related data, are processed in distinct cortical regions. Accordingly, we implement two modality-specific attention blocks: a visual enhancement block and a motion enhancement block, which amplify features corresponding to visual and motor processing, respectively. The output of each block is then spatially remapped to the 10–20 system layout, producing feature-enhanced representations aligned with the brain topology.

These representations are used to construct the MV-BCGR, where attention weights modulate the edge strengths in the adjacency matrix, allowing the graph to emphasize brain regions most relevant to the corresponding modality. This attention-guided formulation enhances the physiological relevance and task specificity of the resulting connectivity structure. The selection of functionally relevant regions is guided by the 10–20 system topology, as illustrated in Fig. 5, which links electrode positions to underlying cortical areas. Specifically, the central region (red: C3, Cz, C4) corresponds to the motor cortex; the parietal region (green: Cp5, Cp1, Cp2, Cp6, P3, P4, P7, P8, Pz) supports visuospatial processing [35]; and the occipital region (blue: Po3, Po4, O1, O2, Oz) relates to primary visual perception [25, 28].

$$\mathcal{L}_{NPAE} = \mathcal{L}_{MSE} + \lambda \mathcal{L}_1 \quad (7)$$

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=1}^N \|W_{MV}^{(i)} - W_E^{(i)}\|_F^2 \quad (8)$$

$$\mathcal{L}_1 = \frac{1}{N} \sum_{i=1}^N \|W_{MV}^{(i)} - W_E^{(i)}\|_1 \quad (9)$$

where $\mathcal{L}_{MSE}$ denotes the mean squared error loss, which measures the overall difference between $W_{MV}^{(i)}$ and $W_E^{(i)}$. This loss is computed using the Frobenius norm to ensure similar representations across different modalities for the same sample. $\mathcal{L}_1$ represents the sparse regularization term, computed using the element-wise L1 norm, to promote sparsity in the generated edge matrices, highlighting key connections and reducing model complexity and overfitting risks. The regularization parameter λ controls the influence of the sparsity regularization term $\mathcal{L}_1$ on the total loss, $W_{MV}^{(i)}$ is the multimodal edge weight matrix generated by NPAE for the i -th sample using video and motion modalities; $W_E^{(i)}$ denotes the i -th sample; and N is the total number of training samples.

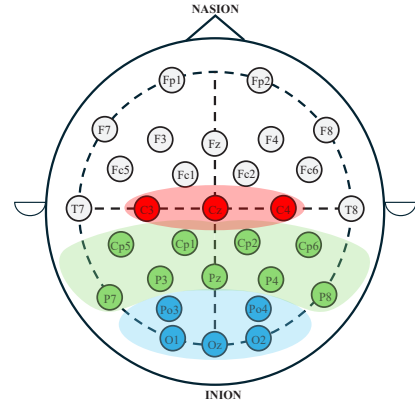


Fig. 5: Brain regions based on the 10–20 system are visualized, with the central area and its corresponding nodes marked in red, parietal lobe regions and nodes marked in green, and occipital lobe regions and nodes marked in blue.

4.5 Connectivity-Constrained Contrastive Fusion

To further integrate features from all modalities and improve cybersickness recognition, we propose the CCCF module. CCCF takes as input three components: S-BCGR, EEG data, and MV-BCGR. We then perform contrastive alignment between MV-BCGR and E-BCGR constructed based on EEG. The output of CCCF is $\mathcal{G}_* = (\mathcal{V}, \mathcal{E}, W_*)$, which integrates all modalities into a unified latent space while preserving connectivity constraints.

CCCF employs contrastive learning to align $\mathcal{G}_{MV}$ and $\mathcal{G}_E$ in the latent share space [45], using InfoNCE as the contrastive loss. We introduce graph augmentations to construct positive and negative pairs. For each batch sample, random augmentations generate variant views of the same connectivity graph to ensure robust representation. Feature pairs from $\mathcal{G}_E$ and $\mathcal{G}_{MV}$ corresponding to the same underlying sample (after augmentation) are treated as positive pairs, while those from different samples form negative pairs.

By maximizing similarity between positive pairs and minimizing it between negative pairs, CCCF effectively reduces modality discrepancies, resulting in the fused graph $\mathcal{G}_*$. To quantify the alignment between modalities, we compute a similarity matrix S based on the adjacency matrices of the EEG and multimodal graphs:

$$S = \frac{W_E \times W_{MV}^T}{\tau} \quad (10)$$

where τ is the temperature parameter. We then apply the cross-entropy in an InfoNCE manner:

$$\mathcal{L}_{NCE} = \frac{1}{2} (\mathcal{L}_{CE}(S, y) + \mathcal{L}_{CE}(S^T, y)) \quad (11)$$

where y is a label vector that differentiates matched and mismatched pairs. After obtaining the fused adjacency W_* , we pass it through a

Table 1: The comparison results with SOTA methods.

Models	Acc $\uparrow$	Pre $\uparrow$	Sen $\uparrow$	Spe $\uparrow$	AUC $\uparrow$	Early Stop Epoch $\downarrow$
Lee et al. [20]	72.22%	72.19%	74.38%	76.91%	73.51%	100
Kim et al.* [17]	79.68%	79.90%	72.33%	79.70%	89.63%	73
Islam et al.* [13]	78.96%	79.79%	78.93%	78.34%	84.18%	82
MAC* [15]	79.88%	81.07%	79.92%	79.21%	85.18%	86
Demirel et al.* [6]	83.34%	84.04%	83.37%	83.39%	84.65%	97
EhancedEEG* [23]	83.63%	87.69%	89.60%	84.07%	89.79%	60
CPNet* [33]	82.03%	74.18%	98.03%	66.14%	91.36%	70
Ours*	84.17%	83.46%	85.09%	84.18%	92.22%	100

*Multimodal fusion methods.

global max-pooling layer on the node dimension to aggregate node-wise features. The pooled features are then fed into an MLP classifier to recognize cybersickness occurrence.

To incorporate $\mathcal{G}_S$ as the connectivity constraint, we follow a two-step procedure at each training iteration. First, we perform contrastive alignment on $\mathcal{G}_E$ and $\mathcal{G}_{MV}$ to obtain $\mathcal{G}_*$ and update model parameters; then, we compute centrality constraints $\mathcal{L}_{deg}$ (Eq. 13) and binary constraints $\mathcal{L}_{bin}$ (Eq. 14) relative to $\mathcal{G}_S$ and use them to further update the network via backpropagation.

$$\mathcal{L}_{deg} = \frac{1}{n} \sum_{i=1}^n |C_i^{\mathcal{G}_*} - C_i^{\mathcal{G}_S}| \quad (12)$$

$$C_i = \frac{\sum_j A_{ij}}{\sum_j \mathbf{I}\{A_{ij} > 0\}} \quad (13)$$

$$\mathcal{L}_{bin} = BCE(\mathbf{I}(\mathcal{G}_* > 0) - \mathbf{I}(\mathcal{G}_S > 0)) \quad (14)$$

n denotes the total number of nodes in the graph; C_i denotes the degree centrality (the sum of connected edge weights of the nodes) of the i th node of the graph; $BCE(\cdot)$ denotes the binary cross-entropy loss function; $\mathbf{I}(\cdot)$ is the indicator function.

All network architectures in our method are trained jointly, sharing a single optimizer. We observed that separating their optimizers or training them in isolation caused unstable convergence. Instead, a unified gradient backpropagation through all losses ensures consistent feature updates and stable end-to-end training. The final loss function integrates the MV-NCGR construction loss $\mathcal{L}_{NPAE}$, InfoNCE-based contrastive loss $\mathcal{L}_{NCE}$, cross entropy loss $\mathcal{L}_{CE}$, and two connectivity constraints $\mathcal{L}_{deg}$ and $\mathcal{L}_{bin}$:

$$\mathcal{L} = \mathcal{L}_{NPAE} + \mathcal{L}_{NCE} + \mathcal{L}_{CE} + \mathcal{L}_{deg} + \mathcal{L}_{bin} \quad (15)$$

4.6 Model evaluation

This study primarily evaluates the occurrence of cybersickness (i.e., presence vs. absence) rather than its severity. Validation is conducted using leave-one-out cross-validation (LOOCV), where each segment serves once as the test sample while the remaining 8,323 segments are used for training. This process is repeated for all segments, ensuring each data point is tested exactly once and training data is maximized in each iteration. LOOCV was selected to make the most of our dataset size. To comprehensively evaluate performance, we employ five metrics: accuracy (Acc), precision (Pre), sensitivity (Sen), specificity (Spe), and area under the curve (AUC). In addition, we evaluate the convergence speed of each model by reporting the early stopping epoch in comparison with other methods.

Accuracy reflects the overall correctness of recognition across both classes. Precision quantifies the model's ability to avoid false positives when predicting cybersickness. Sensitivity measures the ability to correctly identify actual cybersickness episodes, while specificity evaluates the ability to correctly identify non-cybersickness instances. The AUC provides an assessment of the model's discriminative power

between the two classes. Early stopping epoch reflects the number of training iterations required before the model stops improving on the validation set. A smaller early stopping value indicates faster convergence, suggesting higher training efficiency. The metrics are defined as follows:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad (16)$$

$$\text{Pre} = \frac{TP}{TP + FP} \quad (17)$$

$$\text{Sen} = \frac{TP}{TP + FN} \quad (18)$$

$$\text{Spe} = \frac{TN}{TN + FP} \quad (19)$$

$$\text{AUC} = \int_0^1 \text{TPR}(t) d\text{FPR}(t) \quad (20)$$

where TPR is the true positive rate, FPR is the false positive rate, TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative.

5 RESULTS

We implemented and trained our proposed model using PyTorch 2.5.1 on a server running Ubuntu 22.04. Hyperparameters in our method were selected via grid search. Specifically, the regularization factor γ in Algorithm 1 and the coefficient λ in NPAE were both set to 1×10^{-3} , and the temperature parameter τ for contrastive learning was set to 0.5.

Other hyperparameters followed commonly used values in deep learning methods. We employed the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for 100 epochs, using a learning rate of 0.001, dropout rate of 0.5, and batch size of 32. To prevent overfitting and ensure a fair comparison of convergence across different methods, we adopted an early stopping strategy with a patience of 10—training would stop if validation loss did not improve for 10 consecutive epochs. A minimum of 50 epochs was enforced to ensure sufficient training, even when early stopping was triggered.

5.1 Comparison with SOTA Methods

To further assess the effectiveness of our proposed model, we compare its performance against five SOTA cybersickness recognition methods. To accommodate baseline methods with input modalities not directly compatible with our dataset, we adopt an input adaptation strategy. Specifically, we encode additional modality information via an encoder and integrate it as auxiliary input features, without modifying the original network architecture or fusion design. This allows for a fair comparison while respecting the original model structure. Table 1 presents the results across the five key metrics of the recognition task and the early stop epoch. The best-performing metrics are highlighted in bold with a gray background, demonstrating that the proposed method achieves superior overall performance across most evaluation metrics.

In terms of accuracy, our method achieves 84.17%, outperforming the compared methods. The specificity and AUC reach 84.18%

and 92.22%, respectively, demonstrating strong capability in distinguishing both positive and negative samples. Although our method does not achieve the best results in precision and sensitivity (83.46% and 85.09%), it ranks 2nd and 3rd among the 7 compared methods, indicating a stable and well-balanced overall performance.

Regarding convergence behavior, our model was trained for the full number of epochs without triggering early stopping, potentially resulting in more training iterations than models that adopt early stopping strategies. The final evaluation results indicate that this extended training process yields consistent improvements in performance, suggesting that the additional training epochs are beneficial under our model setting.

In summary, our proposed method achieves competitive performance across three of six metrics. The accuracy reaches 84.17%, outperforming all other compared methods. The specificity (84.18%) shows balanced performance in identifying both positive and negative instances. The AUC of 92.22% indicates strong discriminative capability under complex recognition scenarios. Although the precision (83.46%) and sensitivity (85.09%) do not reach the highest values, this represents a modest trade-off within an otherwise balanced performance profile.

5.2 Ablation Study

5.2.1 Ablation experiments of different modalities

Table 2: The ablation results of different modality combinations.

Modality	Acc $\uparrow$	Pre $\uparrow$	Sen $\uparrow$	Spe $\uparrow$	AUC $\uparrow$
EEG	66.01%	65.88%	65.23%	66.15%	72.06%
Video	78.71%	78.08%	79.61%	78.72%	84.16%
Motion	68.94%	73.30%	59.21%	68.91%	78.92%
EEG + Video	76.76%	76.15%	77.65%	76.76%	84.15%
EEG + Motion	79.49%	79.07%	80.00%	79.49%	87.09%
Video + Motion	80.03%	78.75%	73.92%	72.04%	76.40%
All	84.17%	83.46%	85.09%	84.18%	92.22%

In this section, we conduct modality ablation experiments to assess the contribution of data modalities and their combinations. The proposed model is trained using single modalities, pair-modality combinations, and all three modalities. Results are shown in Table 2, with the best-performing values highlighted in bold with a gray background.

The model using all three modalities outperforms all single- and pair-modality settings across all metrics, showing particularly strong AUC (92.22%) and sensitivity (85.09%). These results highlight the effectiveness of multimodal fusion in improving recognition performance. Among the single modalities, video performs best overall, while motion yields higher precision but lower sensitivity than EEG, indicating a tendency to miss certain cases. Pair-modality combinations show progressive gains, with EEG + Motion achieving the best performance (79.49% accuracy, 87.09% AUC) among the paired configurations, suggesting strong complementarity between physiological and kinematic features.

5.2.2 Ablation experiments of different modules

Table 3: The ablation results of the modules.

Module	Acc $\uparrow$	Pre $\uparrow$	Sen $\uparrow$	Spe $\uparrow$	AUC $\uparrow$
w/o S-BCGR	82.03%	81.46%	82.74%	82.34%	90.77%
w/o NPAE	80.08%	79.31%	81.18%	80.08%	85.15%
w/o CCCF	82.62%	80.97%	85.10%	82.27%	92.02%
All	84.17%	83.46%	85.09%	84.18%	92.22%

To assess the contribution of the three main modules in our proposed model, we conducted modality-independent ablation experiments. The results are summarized in Table 3, with the best-performing metrics highlighted in bold with a gray background.

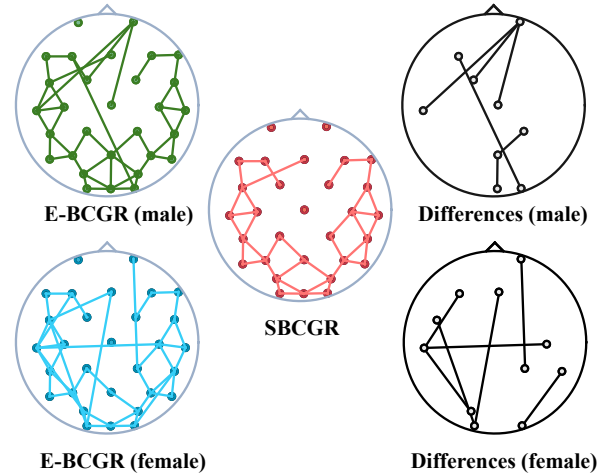


Fig. 6: The E-BCGR grouped by gender, the constructed S-BCGR, and the differences.

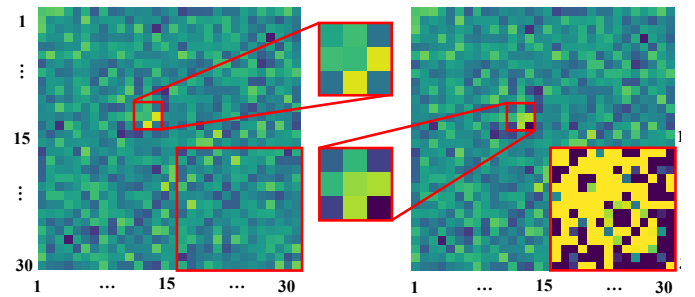


Fig. 7: Visualization of the adjacency matrix of MV-BCGR before and after applying NPAE.

The ablation results demonstrate the impact of each module on cybersickness recognition. The full model achieves the highest performance across all metrics, confirming the necessity of integrating all three modules for optimal accuracy. To gain deeper insights into the specific roles of each module, we conducted visualization analyses of the proposed key methods. As illustrated in Fig. 6, we visualize the E-BCGRs grouped by gender and the S-BCGR generated by our SDA, where connections with edge weights exceeding 0.3 are visualized. Each circular plot represents a Brain Connectivity Graph Representation (BCGR). The left column shows the original E-BCGRs for males (top, in green) and females (bottom, in blue), where the nodes denote electrode positions and the lines represent significant inter-regional connections. The central column presents the S-BCGR, generated by our SDA, shown in red. The right column visualizes the connection differences that were present in each gender-specific E-BCGR but were excluded from the final S-BCGR.

Notably, certain connections are excluded in the final S-BCGR across both groups. For example, connections involving the prefrontal regions—such as those in the frontal pole—are absent. These regions are primarily associated with high-level cognitive processes like the retrieval of non-task-related long-term memory [36], and are not directly implicated in cybersickness processing. Their exclusion thus aligns to highlight cybersickness-relevant and neurophysiologically meaningful patterns. This observation confirms that the proposed SDA successfully filters out irrelevant and individual-specific noise while preserving shared connectivity representations essential for cybersickness recognition.

Figure 7 presents a comparison of the adjacency matrices of MV-BCGR before and after applying the NPAE. The left image shows

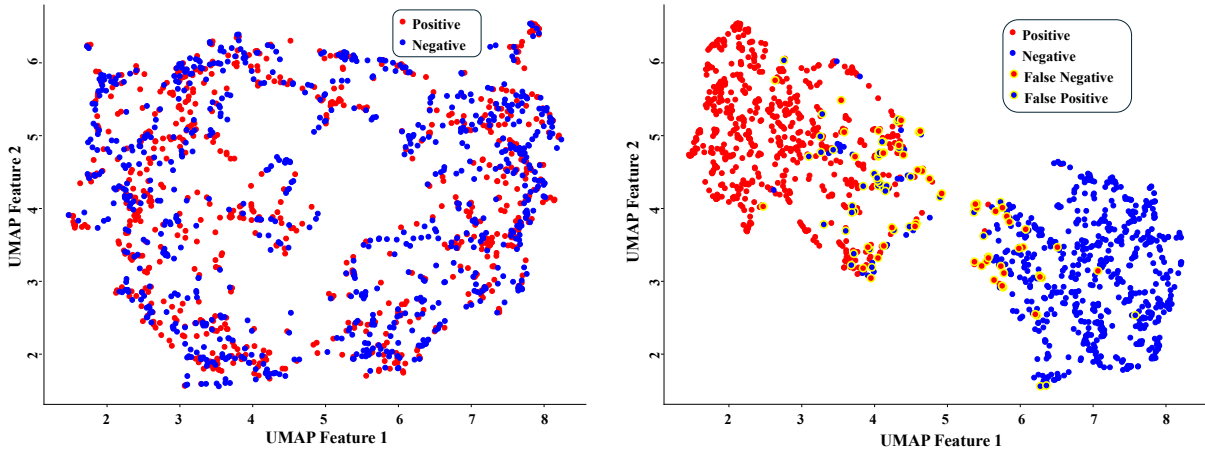


Fig. 8: Comparison between initial features and features obtained from the CCCF.

the connectivity structure prior to NPAE enhancement, while the right reflects the result after enhancement. Within each image, color encodes inter-regional connectivity, representing the strength of connections or learned feature values: deep yellow indicates the strongest connections, followed by green for moderate connections, light blue for weak connections, and deep blue for the weakest. The red-highlighted central area corresponds to motor-related brain regions (indices 12–14, corresponding to the central area), while the bottom-right red-highlighted region reflects connections related to visual processing (indices 16–30, corresponding to the parietal and occipital lobes). The changes observed in both the motor and visual regions clearly demonstrate that NPAE enhances and amplifies connectivity patterns.

Figure 8 shows the two-dimensional feature distribution obtained using UMAP [26], which projects the high-dimensional features into a low-dimensional space for visualization and comparison. The left plot depicts the initial feature distribution, while the right plot displays the features after CCCF. In both plots, red points represent positive samples and blue points represent negative samples. Red points with yellow edges denote false negatives, while blue points with yellow edges denote false positives.

In the initial distribution, positive samples and negative samples are heavily intermixed, with substantial overlap and no clear separation. This suggests that distinguishing between classes in the original feature space is challenging. After applying CCCF, the distribution becomes significantly more structured: positive samples cluster toward the left, while negative samples concentrate on the right, forming a clearer decision boundary. These visualizations confirm that our CCCF module effectively enhances both modality-specific representations and class separability in the feature space, providing a strong foundation for accurate recognition.

6 CONCLUSION, LIMITATIONS, AND FUTURE WORK

In this paper, we proposed a multimodal contrastive learning framework for cybersickness recognition, designed to capture cross-modal connectivity patterns and improve recognition performance through contrastive feature fusion. Our method is built upon the BCGR, instantiated in three instances—E-BCGR, MV-BCGR, and S-BCGR—to enable the integration of EEG, video, and motion data. The proposed CCCF aligns modality-specific representations in a shared latent space, enhancing the consistency and discriminability of the learned features. Experimental results demonstrate that our method outperforms several SOTA methods, achieving an accuracy of 84.17%, sensitivity of 85.09%, specificity of 84.18%, and an AUC of 92.22%. These results show improvements ranging from 0.54% to 11.95% in accuracy, 0.11% to 18.04% in specificity, and 0.86% to 18.71% in AUC. Overall, the findings support the effectiveness of our method in addressing multimodal cybersickness recognition.

Our method still has certain limitations. One such limitation stems from the data collection protocol: to minimize experimental disruptions and reduce artifacts, we avoided collecting subjective severity ratings during the VR experience. As a result, the dataset contains only binary labels (presence or absence of cybersickness), which limits the model’s ability to predict severity levels. Another limitation is the absence of additional physiological modalities due to equipment constraints. While EEG, video, and motion data are informative, the dataset does not include signals such as electrodermal activity (EDA) or eye-tracking, which have been reported to contribute significantly to cybersickness assessment.

In future work, we plan to explore alternative evaluation paradigms that enable the collection of continuous cybersickness labels while minimizing EEG contamination, such as non-intrusive annotation protocols or post-session self-assessments. This would support the development of regression-based models without compromising signal quality. We also aim to incorporate a broader range of physiological signals—including eye-tracking and EDA—to enhance both predictive performance and interpretability. Finally, to improve model robustness and real-world applicability, future studies will consider data collection in more diverse and naturalistic VR environments.

7 ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China through Project 61932003, 62372026, by Beijing Science and Technology Plan Project Z221100007722004, by National Key R&D plan 2019YFC1521102, and by the fundamental research funds for the central universities.

REFERENCES

- [1] Susan Bruck and Paul A Watters. Estimating cybersickness of simulated motion using the simulator sickness questionnaire (ssq): A controlled study. In *2009 sixth international conference on computer graphics, imaging and visualization*, pages 486–488. IEEE, 2009. 4
- [2] Eunhee Chang, Hyun Taek Kim, and Byounghyun Yoo. Predicting cybersickness based on user’s gaze behaviors in hmd-based virtual reality. *Journal of Computational Design and Engineering*, 8(2):728–739, 2021. 2
- [3] Yu-Chieh Chen, Jeng-Ren Duann, Shang-Wen Chuang, Chun-Ling Lin, Li-Wei Ko, Tzzy-Ping Jung, and Chin-Teng Lin. Spatial and temporal eeg dynamics of motion sickness. *NeuroImage*, 49(3):2862–2870, 2010. 2
- [4] Jin Woo Choi, Haram Kwon, Jaehoon Choi, Netiwit Kaongoen, Chaeun Hwang, Minuk Kim, Byung Hyung Kim, and Sungho Jo. Neural applications using immersive virtual reality: A review on eeg studies. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1645–1658, 2023. 1

- [5] Arnaud Delorme and Scott Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004. 4
- [6] Berken Utku Demirel, Adnan Harun Dogan, Juliete Rossie, Max Möbus, and Christian Holz. Beyond subjectivity: Continuous cybersickness detection using eeg-based multitaper spectrum estimation. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–10, 2025. 2, 3, 4, 7
- [7] Alana Dinh, Andrew Lukas Yin, Deborah Estrin, Peter Greenwald, and Alexander Fortenko. Augmented reality in real-time telemedicine and telementoring: scoping review. *JMIR mHealth and uHealth*, 11:e45464, 2023. 1
- [8] Minghan Du, Hui Cui, Yuan Wang, and Henry Been-Lirn Duh. Learning from deep stereoscopic attention for simulator sickness prediction. *IEEE Transactions on Visualization and Computer Graphics*, 29(2):1415–1423, 2021. 6
- [9] Tobias Feigl, Daniel Roth, Stefan Gradl, Markus Wirth, Marc Erich Latoschik, Bjoern M Eskofier, Michael Philippsen, and Christopher Mutschler. Sick moves! motion parameters as indicators of simulator sickness. *IEEE transactions on visualization and computer graphics*, 25(11):3146–3157, 2019. 2
- [10] John F Golding. Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. *Brain research bulletin*, 47(5):507–516, 1998. 4
- [11] Cuiting Guo, Jennifer Ji, and Richard So. Could okan be an objective indicator of the susceptibility to visually induced motion sickness? In *2011 IEEE Virtual Reality Conference*, pages 87–90. IEEE, 2011. 2
- [12] Doli Hazarika, Souptick Chanda, and Cota Navin Gupta. Smartphone-based natural environment electroencephalogram experimentation opportunities and challenges. In *2022 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 370–375. IEEE, 2022. 3
- [13] Rifatul Islam, Kevin Desai, and John Quarles. Cybersickness prediction from integrated hmd's sensors: A multimodal deep fusion approach using eye-tracking and head-tracking data. In *2021 IEEE international symposium on mixed and augmented reality (ISMAR)*, pages 31–40. IEEE, 2021. 3, 7
- [14] Daekyo Jeong, Sangbong Yoo, and Jang Yun. Cybersickness analysis with eeg using deep learning algorithms. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 827–835, 2019. 2
- [15] Dayoung Jeong, Seungwon Paik, YoungTae Noh, and Kyungsik Han. Mac: multimodal, attention-based cybersickness prediction modeling in virtual reality. *Virtual Reality*, 27(3):2315–2330, 2023. 2, 3, 7
- [16] Hak Gu Kim, Wissam J Baddar, Heoun-taek Lim, Hyunwook Jeong, and Yong Man Ro. Measurement of exceptional motion in vr video contents for vr sickness assessment using deep convolutional autoencoder. In *Proceedings of the 23rd ACM symposium on virtual reality software and technology*, pages 1–7, 2017. 1, 2
- [17] Jinwoo Kim, Woojae Kim, Heeseok Oh, Seongmin Lee, and Sanghoon Lee. A deep cybersickness predictor based on brain signal analysis for virtual reality contents. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10580–10589, 2019. 3, 7
- [18] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 6
- [19] Panagiotis Kourtesis, Rayaam Amir, Josie Linnell, Ferran Argelaguet, and Sarah E MacPherson. Cybersickness, cognition, & motor skills: The effects of music, gender, and gaming experience. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2326–2336, 2023. 5
- [20] Tae Min Lee, Jong-Chul Yoon, and In-Kwon Lee. Motion sickness prediction in stereoscopic videos using 3d convolutional neural networks. *IEEE transactions on visualization and computer graphics*, 25(5):1919–1927, 2019. 2, 7
- [21] Te-Won Lee. *Independent Component Analysis*, pages 27–66. Springer US, Boston, MA, 1998. 4
- [22] Kaiming Li, Lei Guo, Jingxin Nie, Gang Li, and Tianming Liu. Review of methods for functional brain connectivity detection using fmri. *Computerized medical imaging and graphics*, 33(2):131–139, 2009. 5
- [23] Ruichen Li, Yuyang Wang, Handi Yin, Jean-Rémy Chardonnet, and Pan Hui. A deep cybersickness predictor through kinematic data with encoded physiological representation. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 1132–1141. IEEE, 2023. 3, 7
- [24] Chung-Yen Liao, Shao-Kuo Tai, Rung-Ching Chen, and Hendry Hendry. Using eeg and deep learning to predict motion sickness under wearing a virtual reality device. *Ieee Access*, 8:126784–126796, 2020. 2
- [25] JC Lynch, VB Mountcastle, WH Talbot, and TC Yin. Parietal lobe mechanisms for directed visual attention. *Journal of neurophysiology*, 40(2):362–389, 1977. 6
- [26] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 9
- [27] Kenneth E Money. Motion sickness. *Physiological reviews*, 50(1):1–39, 1970. 1, 2
- [28] Elisabeth A Murray, Timothy J Bussey, and Lisa M Saksida. Visual perception and memory: a new view of medial temporal lobe function in primates and rodents. *Annu. Rev. Neurosci.*, 30(1):99–122, 2007. 6
- [29] Nitish Padmanaban, Timon Ruban, Vincent Sitzmann, Anthony M Norcia, and Gordon Wetzstein. Towards a machine-learning approach for sickness prediction in 360 stereoscopic videos. *IEEE transactions on visualization and computer graphics*, 24(4):1594–1603, 2018. 2
- [30] Anh Viet Phan, Minh Le Nguyen, Yen Lam Hoang Nguyen, and Lam Thu Bui. Dgcnn: A convolutional neural network over large-scale labeled graphs. *Neural Networks*, 108:533–543, 2018. 5
- [31] Jérémy Plouzeau, Jean-Rémy Chardonnet, and Frédéric Merienne. Using cybersickness indicators to adapt navigation in virtual reality: a pre-study. In *2018 IEEE conference on virtual reality and 3D user interfaces (VR)*, pages 661–662. IEEE, 2018. 3
- [32] Katharina MT Pöhlmann, Gang Li, Mark McGill, Frank Pollick, and Stephen Brewster. Can gender and motion sickness susceptibility predict cybersickness in vr? In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 277–282. IEEE, 2023. 5
- [33] Chang Qi, Ding Ding, Hao Chen, Zheyu Cao, and Wenjie Zhang. Cpnnet: Real-time cybersickness prediction without physiological sensors for cybersickness mitigation. *ACM Transactions on Sensor Networks*, 2025. 3, 7
- [34] Jaziar Radianti, Tim A Majchrzak, Jennifer Fromm, and Isabell Wohlgenannt. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & education*, 147:103778, 2020. 1
- [35] Amna Rehman and Yasir Al Khalili. Neuroanatomy, occipital lobe. 2019. 6
- [36] Jeffrey D Schall, Veit Stuphorn, and Joshua W Brown. Monitoring and control of action by the frontal lobes. *Neuron*, 36(2):309–322, 2002. 8
- [37] Daniel M Shafer, Corey P Carbonara, and Michael F Korpi. Modern virtual reality technology: cybersickness, sense of presence, and gender. *Media Psychology Review*, 11(2):1, 2017. 5
- [38] Ravindra Singh, Naurang Singh Mangat, Ravindra Singh, and Naurang Singh Mangat. Stratified sampling. *Elements of survey sampling*, pages 102–144, 1996. 4
- [39] Richard HY So, Andy Ho, and WT Lo. A metric to quantify virtual scene movement for the study of cybersickness: Definition, implementation, and verification. *Presence*, 10(2):193–215, 2001. 3
- [40] Umama Tasnim, Rifatul Islam, Kevin Desai, and John Quarles. Investigating personalization techniques for improved cybersickness prediction in virtual reality environments. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2
- [41] Nana Tian and Ronan Boulic. Who says you are so sick? an investigation on individual susceptibility to cybersickness triggers using eeg, egg and ecg. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 3
- [42] Xiang Wan, Wenqian Wang, Jiming Liu, and Tiejun Tong. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC medical research methodology*, 14:1–13, 2014. 4
- [43] Biao Xie, Huimin Liu, Rawan Alghofaili, Yongqi Zhang, Yeling Jiang, Flavio Destri Lobo, Changyang Li, Wanwan Li, Haikun Huang, Mesut Akdere, et al. A review on virtual reality skill training applications. *Frontiers in Virtual Reality*, 2:645153, 2021. 1
- [44] Qi Yang and Saleh Kalantari. Real-time continuous uncertainty annotation (rcua) for spatial navigation studies. *arXiv preprint arXiv:2207.14651*, 2022. 3
- [45] Kun Zhang, Zhendong Mao, An-An Liu, and Yongdong Zhang. Unified adaptive relevance distinguishable attention network for image-text matching. *IEEE Transactions on Multimedia*, 25:1320–1332, 2023. 6