



PRECYSE: Predicting Cybersickness using Transformer for Multimodal Time-Series Sensor Data

DAYOUNG JEONG, Hanyang University, Republic of Korea

KYUNGSIK HAN*, Hanyang University, Republic of Korea

Cybersickness, a factor that hinders user immersion in VR, has been the subject of ongoing attempts to predict it using AI. Previous studies have used CNN and LSTM for prediction models and used attention mechanisms and XAI for data analysis, yet none explored a transformer that can better reflect the spatial and temporal characteristics of the data, beneficial for enhancing prediction and feature importance analysis. In this paper, we propose cybersickness prediction models using multimodal time-series sensor data (i.e., eye movement, head movement, and physiological signals) based on a *transformer* algorithm, considering *sensor data pre-processing* and *multimodal data fusion methods*. We constructed the MSCVR dataset consisting of normalized sensor data, spectrogram formatted sensor data, and cybersickness levels collected from 45 participants through a user study. We proposed two methods for embedding multimodal time-series sensor data into the transformer: modality-specific spatial and temporal transformer encoders for normalized sensor data (MS-STTN) and modality-specific spatial-temporal transformer encoder for spectrogram (MS-STTS). MS-STTN yielded the highest performance in the ablation study and the comparison of the existing models. Furthermore, by analyzing the importance of data features, we determined their relevance to cybersickness over time, especially the salience of eye movement features. Our results and insights derived from multimodal time-series sensor data and the transformer model provide a comprehensive understanding of cybersickness and its association with sensor data. Our MSCVR dataset and code are publicly available: <https://github.com/dayoung-jeong/PRECYSE.git>.

CCS Concepts: • **Computing methodologies** → **Modeling methodologies**; • **Human-centered computing** → *Virtual reality*.

Additional Key Words and Phrases: Virtual reality, Cybersickness, Transformer, Multimodal time-series sensor data

ACM Reference Format:

Dayoung Jeong and Kyungsik Han. 2024. PRECYSE: Predicting Cybersickness using Transformer for Multimodal Time-Series Sensor Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 42 (June 2024), 24 pages. <https://doi.org/10.1145/3659594>

1 INTRODUCTION

With the continuous development of virtual reality (VR), many domains have started to show interest in adopting VR. Along with the adoption of VR technology, it is also important to provide users with a high level of immersion and a satisfying experience. However, cybersickness is one of the known factors that hinder user experience. Cybersickness can primarily be explained by sensory conflict theories, such as sensory rearrangement theory and subjective vertical theory [10, 71]. According to these theories, discrepancies between the movement perceived by the vestibular system and the visual input seen on the screen, or inconsistencies between predicted movement based on past experiences and actual movement can lead to symptoms such as dizziness and nausea.

*Corresponding author

Authors' Contact Information: Dayoung Jeong, dayoungjeong@hanyang.ac.kr, Hanyang University, Republic of Korea; Kyungsik Han, kyungsikhan@hanyang.ac.kr, Hanyang University, Republic of Korea.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike International 4.0 License.

© 2024 Copyright held by the owner/author(s).

ACM 2474-9567/2024/6-ART42

<https://doi.org/10.1145/3659594>

Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., Vol. 8, No. 2, Article 42. Publication date: June 2024.

VR users interact with a virtual environment (VE) through various sensory inputs (e.g., vision, hearing) and perceive and move or act within the VE as if it were real [8, 25, 29]. To precisely understand the user's state during these VR experiences, it is important to learn and integrate the interactions between different sensory signals. Cybersickness is one of the main states that VR users can experience, and it is important to theoretically validate and utilize sensor data related to it. Previous studies have focused on the simple use of data that can be collected in VEs.

Because understanding and handling cybersickness is critical to VR user experience and wider VR adoption, many studies have shown great interest in this symptom. Based on our literature review, research on cybersickness can be broadly divided into three categories. The first is the identification of the cause of cybersickness. Studies in this category typically focus on identifying the specific characteristics of signals occurring during the VR experience that trigger cybersickness [14, 21, 26, 44, 62, 67, 75]. The second is the mitigation of cybersickness, where studies investigate technical interventions (e.g., field of view (FOV) restrictions, rotation blurring, and the use of a virtual nose) to directly address cybersickness perception in the VE [11, 24, 58, 90]. The third is the prediction of cybersickness using artificial intelligence (AI) techniques through sensor signals generated from a standard head-mounted display (HMD) or wearable devices. Various AI models (e.g., convolutional neural network (CNN), deep neural network (DNN), recurrent neural network (RNN), and long short-term memory (LSTM)) have been applied and shown to be effective in predicting cybersickness [34–37, 39, 40, 43, 50, 51, 54, 55]. Most studies present models that mainly use CNN and LSTM. This is because CNN specializes in learning data features, while LSTM specializes in handling time series data. In addition, explainable AI (XAI) methods, such as Shapley Additive Explanations (SHAP) [59] and Local Interpretable Model-Agnostic Explanations (LIME) [73], have been used to show the influence of input data features on the prediction value, and attention mechanisms have been applied to CNN- and LSTM-based models to interpret feature importance through attention weights. Our work extends the efforts of prediction modeling and XAI in previous studies, as we propose a transformer-based model.

A transformer was originally proposed in the field of natural language processing (NLP) [86] and has advantages in terms of computational cost and performance compared to CNN and LSTM [12, 15, 86], and in capturing feature importance for time series data. As such, a transformer is particularly useful for learning video or sensor data with temporal characteristics, making it widely used in many domains, such as digital healthcare, automated driving, and urban planning [23, 32, 60, 68, 79]. Since we use data with multimodal and temporal characteristics in our study, the architecture of the transformer model varies in terms of sensor data pre-processing and multimodal data fusion methods. In this sense, the investigation of these aspects is important and expected to provide researchers with valuable insights into the construction of a transformer-based cybersickness prediction model.

In this paper, we present a transformer-based cybersickness prediction model and, in particular, consider two key points on data in the construction of the transformer model: *sensor data pre-processing* and *multimodal data fusion methods*. We used sensor data types that are relevant to cybersickness theories [71, 74, 85], including eye movement, head movement, and physiological signals, and explored methods to model the interaction characteristics among multimodal sensor data. Sensor data inherently have time-series characteristics, and we defined the interaction characteristics between sensor data features at specific times as spatial characteristics. We proposed a transformer-based model considering an input data format that could reflect spatial characteristics. We employed normalized data and spectrogram as the input data format to embed multimodal time-series sensor data into the transformer. The spectrogram is a widely used method in previous studies to reflect both spatial and temporal characteristics, using normalized data in its processing: modality-specific spatial and temporal transformer encoders for normalized sensor data (**MS-STTN**) and a modality-specific spatial-temporal transformer encoder for spectrogram (**MS-STTS**).

We recruited 45 participants for a user study and constructed a **MSCVR** (multimodal sensor data for cybersickness in VR) dataset, consisting of normalized sensor data, spectrogram formatted sensor data, and cybersickness levels. Our study results showed that MS-STTN achieved the highest performance, with a mean absolute error

(MAE) of 0.12 and a root mean square error (RMSE) of 0.25. Based on MS-STTN, we further analyzed the feature importance using the attention weights of the transformer encoder. The data features of gaze direction z , gaze origin x , head position x , and electrodermal activity (EDA) demonstrated high correlations with cybersickness across all timestamps. In addition, we found that all data features were influential to varying degrees.

In summary, this paper makes the following contributions:

- We constructed the multimodal sensor data for cybersickness in VR (MSCVR) dataset, which contains multimodal time-series sensor data (eye movement, head movement, and physiological signals) in the form of normalized data and spectrograms. Each data sample is presented with a cybersickness level. We have provided accessible links to the MSCVR dataset and code for use by the research community.
- We provided comprehensive analysis results for two embedding methods that can be considered in terms of sensor data pre-processing and multimodal data fusion methods: MS-STTN and MS-STTS.
- We analyzed the relationships between cybersickness and sensor data with temporal characteristics by extracting feature importance through the attention weights of the transformer.

2 RELATED WORK

2.1 Correlation between cybersickness and sensor signals

We investigate the correlation between cybersickness and sensor signals based on representative theories and studies to identify the cause of cybersickness in order to make better use of highly correlated sensor signals. Cybersickness can typically be explained by three theories: the sensory conflict theory, the postural instability theory, and the poison hypothesis [16, 56, 72, 85].

The sensory conflict theory is one of the well-known cybersickness theories that postulates that cybersickness occurs when information about body orientation provided by the vestibular system and information about movement seen by the eyes do not match [10, 30, 63, 64, 71]. The postural instability theory claims that cybersickness is induced when posture is maintained in an unstable manner for a prolonged period of time [74, 81]. Meanwhile, the poison hypothesis is used to explain the signs and symptoms of cybersickness. It states that when we ingest a toxin, the body induces vomiting to remove harmful substances, and sensory conflict serves as a warning that the toxin has entered the body [45, 61, 85].

Chang et al. [14] demonstrated the sensory conflict theory based on eye movement. When tracking users' eye movement in VR roller coaster content, cybersickness was less pronounced when the user's actual movement and the movement in the VR content were similar. Jung et al. [44] hypothesized that if the vibrations of the floor in the real world minimally match the characteristics of VR content, the magnitude of sensory conflict would be reduced, resulting in less cybersickness. To prove this hypothesis, they collected data from three features—heart rate (HR), EDA, and pupil diameter—and responses to the simulator sickness questionnaire (SSQ), a questionnaire about the cybersickness they felt after completing the VR experience [46]. When comparing the data between the experimental group (with floor vibrations) and the control group (without floor vibrations), the experimental group showed fewer cybersickness symptoms in terms of oculomotor responses in the SSQ and signal changes in the EDA. Dong et al. [21] reported that although drivers (directly controlling a driving simulation) tended to exhibit more head and torso movements than passengers (watching the driving screen), passengers experienced more severe motion sickness than drivers. This was attributed to the fact that the drivers' visual inputs and body movements were congruent, while the passengers' visual inputs were not congruent with their body movements. Risi et al. [75] designed an experiment in which participants actively controlled their movements during the VR experience and had limited postural restraints during passive movement. The participants showed high levels of cybersickness for postures that involved more extensive movement vectors within the VE. Nalivaiko et al. [62] focused on sweating as one of the symptoms of motion sickness. EDA is a sensitive approach to quantitatively evaluating sweating, and the researchers mentioned that cybersickness triggers defensive hypothermia, which

Table 1. Summary of recent research on cybersickness prediction modeling based on deep learning. “VR exposure” refers to the type of VR content (Active means participants move the virtual environment and Passive means participants watch VR videos in a fixed position). “Sensor” refers to the modality used in the study (V: VR video content, E: Eye movement, H: Head movement, and P: Physiological signal). Note that the four types of modalities are not exclusive, but are identified based on our literature review. “Predicted” means the target (dependent) variable; CS: cybersickness.

Literature	Year	VR exposure	Sensor data	# of Participants	Model	Predicted	Performance
Jin et al. [43]	2018	Active	V/H	24	LSTM	SSQ score	0.86 (R^2 score)
Jeong et al. [41]	2019	Passive	P	25	CNN/DNN	CS level (2 classes)	0.98/0.98 (Accuracy)
Lee et al. [57]	2019	Passive	V/P	20	ConvLSTM	SSQ score	7.34 (RMSE)
Kim et al. [50]	2019	Passive	V/P	202	CNN-RNN	CS level (5 classes)	0.89 (Accuracy)
Islam et al. [37]	2020	Passive	P	23	CNN-LSTM	CS level (3 classes)	0.98 (Accuracy)
Islam et al. [35]	2021	Active	V/E/H	26	CNN-LSTM	CS level (4 classes)	0.87 (Accuracy)
Islam et al. [34]	2021	Passive	P	23	DNN	CS level (0-10)	2.61 (RMSE)
Wang et al. [88]	2022	Passive	E/H	20	LSTM	SSQ score	0.83 (Accuracy)
Qu et al. [70]	2022	Active	P	9	LSTM-Attention	CS level (2 classes)	0.96 (Accuracy)
Islam et al. [36]	2022	Passive	E/H/P	23	DNN	CS level (0-10)	0.49 (RMSE)
Jeong et al. [39]	2022	Passive	E/H/P	27	CNN-Attention -LSTM-Attention	CS level (4 classes)	0.84 (Accuracy)
Kundu et al. [55]	2023	Passive	E/H/P	30	LSTM/GRU/MLP	CS level (0-10)	0.36/0.41/0.57 (RMSE)
Jeong et al. [40]	2023	Passive	V/E/H/P	27	CNN-Attention -LSTM-Attention	CS level (4 classes)	0.87 (Accuracy)

acts to cool the sufferer’s body. Gavvani et al. [26] collected physiological signals—HR, EDA, breathing rate (BR), sweating, and electrocardiography (ECG)—and nausea levels from riding content. They found that each physiological signal was correlated with cybersickness. In particular, changes in sweating and EDA were reported to be indicators of the onset of cybersickness.

In summary, the postural theory showed a strong association between cybersickness and eye and head movements, while the poison hypothesis revealed a high correlation between cybersickness and physiological signals (e.g., EDA, HR, sweating). The sensory conflict theory showed that both body movement and physiological signals relate to cybersickness. Therefore, in Section 2.2, we review research on modeling cybersickness prediction using three modalities: eye movement, head movement, and physiological signals.

2.2 Recent studies on cybersickness prediction models

Recently, many studies have used deep learning models to predict cybersickness (Table 1). We review studies that have considered eye and head movements and physiological signals as modalities for modeling cybersickness.

Jeong et al. [41] extracted cybersickness-related features through electroencephalography (EEG) analysis and demonstrated the effectiveness of a CNN/DNN model with an accuracy of 98%/98%, respectively. In their work, the EEG spectrogram was used as a data feature to classify cybersickness. Islam et al. [37] proposed a CNN-LSTM model using physiological data (EDA, HR, BR, and heart rate variability (HRV)) from 31 participants in a VR roller

coaster simulation and achieved a mean accuracy of 97.44%. Lee et al. [57] presented a deep learning framework of a ConvLSTM structure, to be used for individual cybersickness prediction using visual and physiological features (e.g., EEG, EDA). The model result was an RMSE of 7.34 when both types of features were applied and an RMSE of 9.56 when only the physiological signals were applied. Islam et al. [34] presented Cybersense, a framework that collects physiological signals and transmits information after a DNN model predicts the cybersickness level, ranging between 0 and 10. The DNN model was trained on physiological signals collected from 23 participants, and an RMSE was 2.61. Wang et al. [88] used the magnitude of each eye acceleration calculated from 13 raw eye movement data collected from the HTC VIVE Pro Eye and the magnitude of character movement acceleration in VR as training data. Both Model A and Model B used LSTM to predict SSQ scores. The result for Model A was based on its performance on test data during the training process, while the result for Model B reflected its real-time prediction performance when embedded in VR content. When comparing the two models, Model B showed better performance, and they reported that it is important to reflect the user's time-series data collected in real-time to a model. Qu et al. [70] built a real-time cybersickness monitoring system using physiological signals (EDA, ECG). The LSTM attention model, trained with data collected from nine people, showed an accuracy of 0.96 in classifying the level of cybersickness. Jin et al. [43] built the CNN and LSTM-RNN models with head movement and video data. They found that the best performance for predicting cybersickness was obtained by the LSTM-RNN. Kim et al. [50] proposed a CNN-LSTM model to train the cognitive representation of the EEG spectrogram and the time-series characteristics of VR videos. The proposed model demonstrated a performance improvement of over 30%, achieving an accuracy of 87%, compared to the CNN model. Islam et al. [35] predicted cybersickness by collecting eye movement, head movement, and video data. The proposed model used 3D-CNN for learning video data, and CNN-LSTM for learning eye and head movement data. When learning with all data modalities, the accuracy was 0.52, while the model with eye and head movement data showed a high accuracy of 0.87. Jeong et al. [39] predicted cybersickness using eye movement, head movement, and EDA. The CNN-LSTM architectures were proposed, and the highest accuracy was achieved when both models were equipped with an attention mechanism. Islam et al. [36] proposed a multimodal deep fusion network, called DeepTCN, using eye movement, head movement, and physiological signals to predict cybersickness 30 to 60 seconds in advance. The model achieved an RMSE of 0.49. They also showed that fusing eye movement, EDA, and HR was effective in predicting cybersickness. Kundu et al. [55] proposed a cybersickness prediction model based on LSTM, gated recurrent unit (GRU), and multi-layer perceptron (MLP) using eye movement, head movement, and physiological signals. The LSTM model achieved an accuracy of 94% in classification and an RMSE of 0.30 in regression. In addition, the highest correlation was found with eye movement when XAI methods were used to extract data features that influence cybersickness. Jeong et al. [40] predicted cybersickness using video view, eye movement, head movement, and EDA. They proposed a CNN-LSTM model integrated with an attention mechanism. The model trained with all data modalities achieved an accuracy of 82%. Furthermore, when interpreting the importance of features based on attention weights, eye movement showed the highest correlation with cybersickness.

Based on our literature review, we found that most cybersickness prediction models have been developed based on CNN and LSTM. Although there are studies using attention mechanisms and XAI methods to investigate the correlation between cybersickness and data features, the results did not reflect the temporal aspects of the model's cybersickness prediction. Thus, in this work, we used a transformer algorithm that allows parallel processing during training, resulting in lower computational cost than CNN and LSTM. The transformer also allows for the interpretation of the importance of data features for each modality over time. Research in cybersickness modeling has not yet employed transformers or discussed key considerations for developing transformer-based models.

2.3 Transformer models trained with sensor data

The advent of transformers has seen their use and expansion into many domains. There is also a growing interest in using transformers in modeling studies based on sensor data. Dirgova et al. [20] collected accelerometer and gyroscope data from smartphones and applied them to a transformer for human activity recognition (HAR). The features for each data point were embedded in the transformer as a single token. Zhang et al. [93] applied a transformer to WiFi network-based trajectory data to analyze the movement patterns of indoor users. The trajectory data were embedded in the transformer as raw time-series data. Similarly, Ijaz et al. [33] used accelerometer data to build a vision transformer (ViT)-based model of nursing activity. Zhang et al. [94] used ConvTransformer for sensor-based HAR. By proposing an inertial measurement unit (IMU) block for multi-modal fusion of sensors, ConvTransformer can better learn the characteristics of sensor data.

In addition, previous studies have considered methods that transform sensor data values instead of directly using them in a raw data format. For example, Tian et al. [83] transformed accelerometer, gyroscope, and magnitude data generated by smartphones into three-channel spectrogram images, using continuous wavelet transform (CWT), which were then applied to ViT. Zhang et al. [95] converted radiomyography (RMG) data, which detect muscle activity, into spectrograms using short-time Fourier transform (STFT) and CWT before applying them to ViT.

Although previous studies have proposed various methods to train transformer models using sensor signals (raw or spectrogram format), they have not presented or discussed a comparative analysis of different pre-processing methods of input data in terms of model training and performance.

Our research objective goes beyond the use of transformers for cybersickness prediction. We also aim to propose and analyze methods for embedding multimodal time-series sensor data into a transformer in the context of cybersickness, and provide insights into the construction of model architectures based on data format. Another goal is to investigate and present the temporal characteristics of sensor signals from different modalities in cybersickness learning through the lens of the feature importance of the sensor signals over time.

3 MSCVR DATASET

There are many ways to embed multimodal sensor data into a transformer as an input. In this work, we constructed a *multimodal sensor data for cybersickness in VR* (MSCVR) dataset through a user study to facilitate model training in multiple ways.

3.1 VR video selection

Following the user study design in previous studies, we prepared a virtual environment for data collection [34, 41, 42, 49–51, 54, 57, 66]; the participants were asked to sit down and watch the VR videos.

Although cybersickness can vary from person to person, it may or may not cause cybersickness depending on the content. Collecting cybersickness levels using only a few types of content could lead to data bias due to a lack of data representation, which in turn could lead to problems such as model overfitting. Therefore, in this work, we prepared different content topics for data collection. We selected 10 topics, including [cycling, driving, flat rides, flying, motorbikes, planes, roller coasters, sailing, space travel, and water slides], as shown in Figure 1. Each topic has two videos, having a total of 20 videos for the user study. Each video was played for 45 seconds. The video content was 360 degrees to provide the user with an immersive experience [4, 40, 77] and was set to 4K resolution and 30 frames per second (FPS).

3.2 Equipment

We used the HTC VIVE Pro Eye HMD to collect eye and head movements and the Empatica E4 wristband to collect physiological signals (EDA, BVP, and skin temperature (SKT)). As explained in Section 2.1, the most

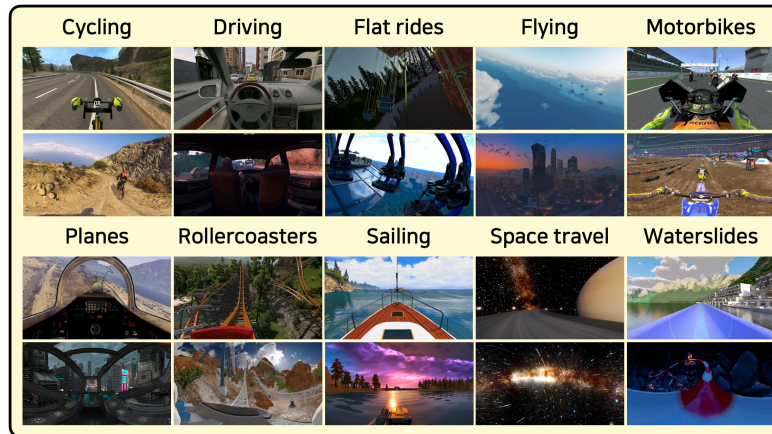


Fig. 1. The screenshots of the VR videos used in the user study.

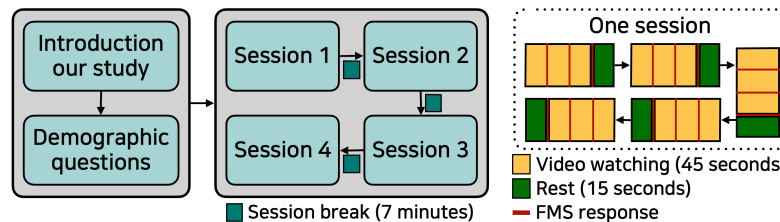


Fig. 2. The procedure of the user study for data collection.

commonly used psychological signals are EDA, HR, and sweating. In the E4 wristband, since HR is an averaged value calculated based on BVP, it is appropriate to use the BVP value provided in real-time, while sweating is closely related to body temperature. Therefore, we used EDA, BVP, and SKT as our physiological signals. We ran the VR videos on a Unity 3D over a Windows 10 PC with Intel Core i7 and GeForce RTX 2070 GPU.

3.3 Study procedure

Our user study was approved by the Institutional Review Board (IRB) at the authors' university. We recruited 45 participants via university bulletin boards, local community websites, and word of mouth. The age of the participants ranged from 20 to 36 years (mean = 24.4, SD = 3.3), and 24 of them were male, with a balanced gender distribution. We collected the responses to the Motion Sickness Susceptibility Questionnaire (MSSQ) [27], which measures susceptibility to motion sickness in everyday life. The average MSSQ of the participants was 10.0 (SD = 8.71), which means that our participants can be defined as average people, and this criterion is similar to that mentioned in previous studies [9, 27, 40].

Figure 2 illustrates the procedure of the user study for data collection. We first explained the goal and procedure of our study, and obtained informed consent from each participant. Participants were asked to answer demographic questions (age, gender, and MSSQ), and wore the HMD and the E4 wristband while seated, and were given sufficient time to familiarize themselves with the devices. We adjusted the participants' visual focus on the VR content and ensured that the sound level was appropriate for the participants to hear before the experiment began. The study consisted of four sessions. One session consisted of watching five VR videos, and the topics of the videos

Table 2. Data features collected in the user study.

Device	Collected data features
HMD	[From both eyes integrated]
	- combined gaze direction (x,y,z)
	[From each of right and left eyes]
	- eye openness
	- gaze direction/origin (x,y,z)
	- pupil diameter
	- pupil position (x,y)
E4 wristband	[Single feature]
	- head position (x,y,z)
	- head rotation (roll, pitch, and yaw)
	- electrodermal activity (EDA)
E4 wristband	- blood volume pulse (BVP)
	- skin temperature (SKT)

did not overlap. To measure participants' cybersickness level in real-time, we used the Fast Motion Sickness scale (FMS). The FMS is a verbal rating scale to which users subjectively respond regarding their cybersickness level, ranging from 0 (no sickness) to 20 (severe sickness) [47]. In addition, previous studies have demonstrated the similarity between quantitative indicators (i.e., SSQ) that evaluate cybersickness and FMS, suggesting the validity of its use [37, 48]. Participants responded verbally to the FMS every 15 seconds while watching the VR videos. Note that the time interval at which the FMS was collected varied widely across studies, ranging from 10 to 30 seconds [2, 3, 7, 35]. Frequent collection of cybersickness levels may undermine the user experience, and it may be difficult to track instantaneous changes in cybersickness when collection time intervals are long. In reviewing the existing literature, we found no established standard for the optimal time interval for requesting FMS responses. In our work, we chose a 15-second interval to collect data frequently enough without compromising the user experience. All participants reported that they were not bothered by speaking the FMS while experiencing VR. We gave them a seven-minute break at the end of each session. If participants requested a longer break, we gave them as much time as they needed.

3.4 Data processing

During the experiment, we collected sensor data as shown in Table 2. Eye movement data consist of 23 features; combined gaze direction (x,y,z), right eye openness, right eye gaze direction (x,y,z), right eye gaze origin (x,y,z), right eye pupil diameter, right eye pupil position (x,y), left eye openness, left eye gaze direction (x,y,z), left eye gaze origin (x,y,z), left eye pupil diameter, and left eye pupil position (x,y). Head movement data consist of six features; position (x,y,z) and rotation (roll, pitch, and yaw). Physiological signal data consist of three features; EDA, BVP, and SKT.

Figure 3 shows the pre-processing of sensor data before they are used as an input for a transformer model. We collected raw data at 90Hz, the highest refresh rate currently supported by the HMD. We used the E4 streaming server to collect physiological signals from the E4 wristband at 90Hz. We established TCP/IP communication between the E4 streaming server and Unity3D, which allowed us to request the same data frequency as the HMD. We noted that sensor data often show data redundancy when collected [78, 92]. To make the data more reliable before passing them to the transformer model, we downsampled the data from 90Hz to 25Hz, removing successively duplicated values. Although the sensor signal was reconstructed to 25Hz in the context of cybersickness prediction modeling, it is likely that various sensors such as body and brain signals generate

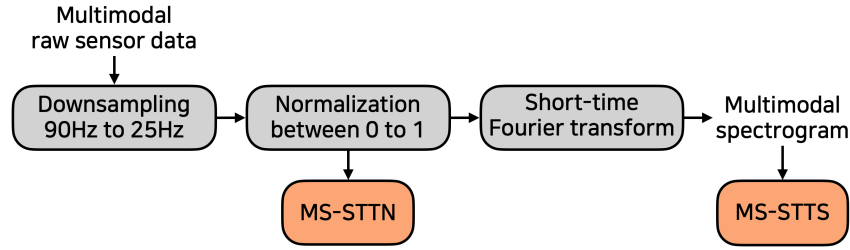


Fig. 3. The data preprocessing procedure for use as transformer input data.

signals with different frequencies (Hz). Therefore, when embedding sensor data into a model for a particular domain, finding the most appropriate frequency for optimal performance is also considered an important task. The data were then normalized to values between 0 and 1 to train the model in a more reliable way [38, 87].

We used the participants' responses to the FMS every 15 seconds as the ground truth. We grouped the sensor data for each 45-second video into three 15-second segments. As a result, our data took the form of 375 data points (25Hz×15 seconds) paired with the corresponding FMS value. In addition, we used the Short-time Fourier transform (STFT), a method that divides the entire data into segments and performs the Fourier transform on each segment [31, 69, 91], to convert the normalized sensor data into spectrograms. The spectrograms are in the form of 2D images with time information on the x-axis and the magnitude of the frequency components in the signals is displayed on the y-axis.

Overall, the MSCVR dataset consists of multimodal sensor data (eye movement, head movement, and physiological signals) in the form of normalized sensor data, spectrograms, and cybersickness levels.

4 METHOD FOR EMBEDDING SENSOR DATA IN TRANSFORMER

In this section, we first define the problem that justifies the objectives of our experiment, and then explain two proposed methods for embedding multimodal time-series sensor data to be used in a transformer.

4.1 Problem definition

Given multimodal sensor data $\{X_1, X_2, \dots, X_M\}$, where M represents the number of modalities and $X_M = \{x_1, x_2, \dots, x_f\}$, where f is the number of features for each modality. In our data, $M = 3$ (eye movement ($e = 23$), head movement ($h = 6$), and physiological signals ($p = 3$)). If the format of data is normalized sensor data, $N = \{n_1, n_2, \dots, n_t\}$ ($t = 1, 2, \dots, 375$) with one data point $n_t = \{f_1, f_2, \dots, f_i\}$ ($i = e + h + p$). If the format of data is spectrogram, $S = \{s_1, s_2, \dots, s_i\}$ and each feature is a 2D image.

In all embedding methods, data are passed to the linear patch embedding layer and the positional embedding layer, before being used as an input to a transformer encoder. The linear patch embedding layer maps the input data to D dimensions because all layers of the transformer use a constant vector size D . The positional embedding layer is used to reflect the positional information of the data and to remember the order of the input vectors in the transformer. A class token (CLS), a learnable embedding, is prepended to the sequence of embedded patches.

The transformer encoder we used follows the architecture of Vaswani et al. [86]. A transformer encoder consists of the multi-head self-attention, the multi-layer perceptron (MLP), and the Layer Normalization. The MLP contains two layers with a Gaussian Error Linear Unit (GELU) and the head of the MLP is a CLS token that serves as a representation vector with important features for the data in predicting cybersickness after passing through a transformer. Previous studies used a total of 12 transformer encoders [6, 19, 22, 84], and we used the same number of encoders.

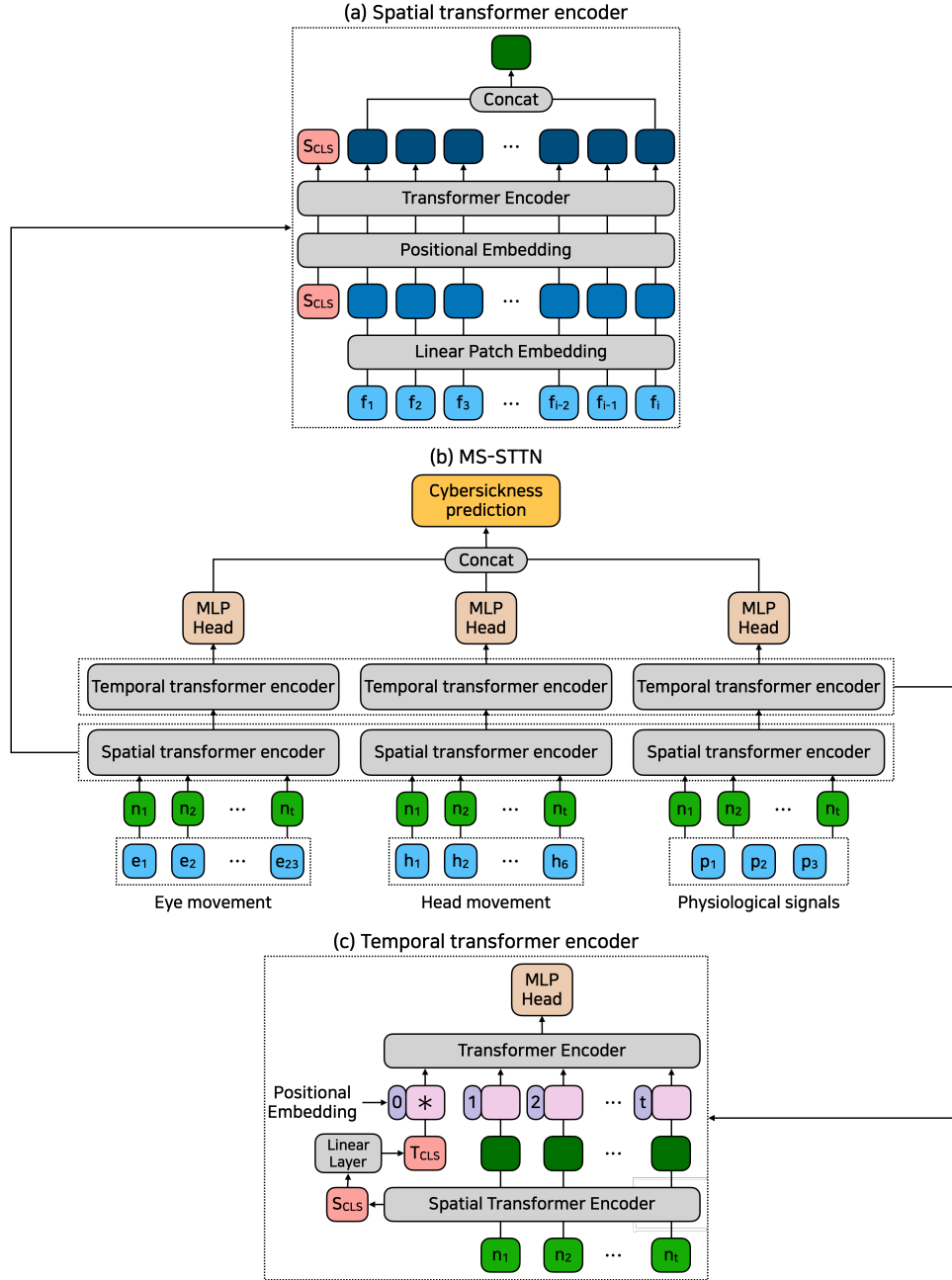


Fig. 4. The architecture of the modality-specific spatial and temporal transformer encoders for normalized sensor data (MS-STTN). (a) The spatial transformer encoder (STE) learns interaction characteristics between sensor signals. (b) The MS-STTN learns spatial and temporal characteristics for each modality. (c) The temporal transformer encoder (TTE) learns temporal characteristics of sensor data.

4.2 Using normalized sensor data

As mentioned in Section 1, we designed a spatial transformer encoder that learns by reflecting spatial characteristics when using normalized sensor data. In addition, we have implemented a separate temporal transformer encoder to reflect the inherent time-series characteristics of sensor data.

4.2.1 Spatial and temporal transformer encoders. The spatial transformer encoder (STE; illustrated in a dotted box in Figure 4-(a)) learns a meaningful relationship between tokens extracted from the same temporal index (i.e., a data point), similar to the idea of a CNN which learns information about local connectivity. The temporal transformer encoder (TTE) learns temporal characteristics based on a representation vector that has characteristics of the relationship between data features of one data point extracted by an STE (Figure 4-(c)). Before being passed to the linear patch embedding layer, the input data take the form of $N = \{n_1, \dots, n_t\}$. Each data point is represented as $n_t = \{e_1, \dots, e_{23}, h_1, \dots, h_6, p_1, p_2, p_3\}$ (where e means a data feature of eye movement, h means a data feature of head movement, and p means a data feature of physiological signals), which is the same as the notion of patches in a ViT [22]. The STE is applied to each data point n_t . After passing through the STE, vectors for each data feature are extracted. After concatenating these vectors, a new patch is generated and used as an input patch for a TTE. Similarly, all S_{CLS} of the spatial transformer CLS token corresponding to each data feature are concatenated before going through the linear layer, making sure that one S_{CLS} has the information of the spatial characteristics for all data features of n_t . S_{CLS} is passed to the linear layer to match the dimension of the CLS token of the TTE T_{CLS} . Thus, T_{CLS} , a head of MLP that passes through the twelve transformer encoders, reflects the data's spatial and temporal characteristics.

4.2.2 MS-STTN. Given the diverse data features within each modality and the distinct characteristics of each data feature, constructing separate models for each modality allows for more precise learning of the relationships between their respective data features. By subsequently merging the models of the modalities later before making predictions, there is the potential for better understanding and learning of the representation for each modality, potentially leading to higher model performance. The modality-specific spatial-temporal transformer for normalized sensor data (MS-STTN) is a model that learns the spatial and temporal characteristics of each modality by applying STE and TTE to each modality (Figure 4-(b)). The input data format of each modality is $N_e = \{n_1, n_2, \dots, n_t\}$ ($n_t = \{e_1, e_2, \dots, e_{23}\}$) for eye movement, $N_h = \{n_1, n_2, \dots, n_t\}$ ($n_t = \{h_1, h_2, \dots, h_6\}$) for head movement, and $N_p = \{n_1, n_2, \dots, n_t\}$ ($n_t = \{p_1, p_2, p_3\}$) for physiological signals.

The E_{CLS} , H_{CLS} , and P_{CLS} are extracted from the STE and TTE of eye movement, head movement, and physiological signals, respectively. An embedding vector is extracted by a simple fusion (i.e., concatenation) of E_{CLS} , H_{CLS} , and P_{CLS} , and the level of cybersickness is predicted by passing through a linear function.

4.3 Using spectrogram

If the input sensor data are in the form of a spectrogram, ViT can be applied by dividing the image into patches for embedding. However, this method is only applicable to a single modality. To embed spectrograms from multiple modalities into the transformer, a different method is required. We treated multimodal sensor data based on spectrograms by extending the idea of ViT. A spectrogram contains both spatial and temporal characteristics that represent changes in sensor signals within time windows. Thus, we considered each spectrogram of a data feature as a patch and embedded the patch in a transformer.

4.3.1 Spatial-temporal transformer encoder. The spatial-temporal transformer encoder for spectrogram (STTE) applies the STFT to 1D raw sensor data, transforming it into 2D image format in the form of a spectrogram, which is then used as the input data (Figure 5-(a)); Each patch represents the spectrogram of each data feature for 15 seconds, resulting in the same number of 32 patches in our MSCVR dataset. The input vector takes the form of $S = \{s_1, \dots, s_t\}$. To handle 2D images, we reshaped the images S into a sequence of flattened 2D patches S_L .

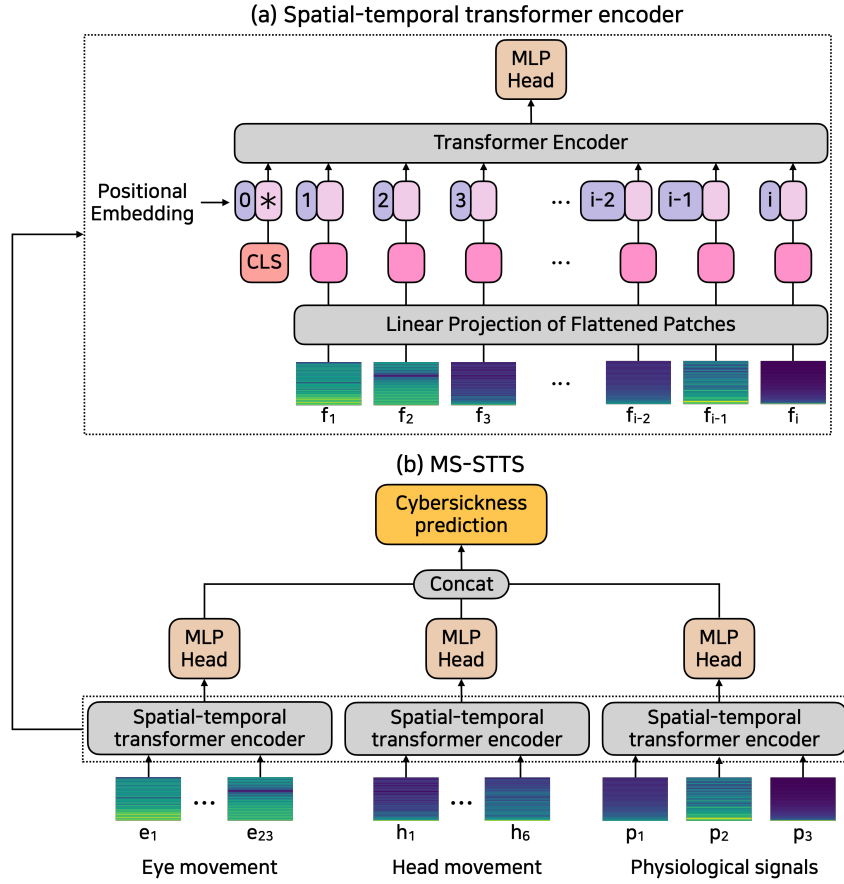


Fig. 5. The architecture of the modality-specific spatial-temporal transformer encoder for spectrogram (**MS-STTS**). (a) The spatial-temporal transformer encoder (**STTE**) learns both interaction and temporal characteristics. (b) The MS-STTS learns spatial and temporal characteristics for each modality.

$\in \mathbb{R}^{(P^2 \cdot C) \times 32}$, where C is the number of channels, (P, P) is the resolution of each patch. After passing through the positional embedding layer and the 12 transformer encoders, the CLS of the extracted MLP head reflects both spatial and temporal characteristics and passes through a linear function for cybersickness prediction.

4.3.2 MS-STTS. MS-STTS is identical to MS-STTN in terms of the model structure, but uses spectrogram as the data format (Figure 5-(b)). MS-STTS is expected to learn specific spatial and temporal characteristics for each data modality using a separate STTE for each modality. The input data format of STTE of each modality is $S_e = \{s_1, s_2, \dots, s_{23}\}$ for eye movement, $S_h = \{s_1, s_2, \dots, s_6\}$ for head movement, and $S_p = \{s_1, s_2, s_3\}$ for physiological signals. The E_{CLS} , H_{CLS} , and P_{CLS} are extracted from the STTE of each data modality and fused (concatenated) to predict cybersickness.

5 RESULTS

All model training and evaluation experiments were performed on a PC equipped with an Intel Core i7-12700 with 24GB RAM and an NVIDIA GeForce RTX 3080 GPU. Data handling and model development were implemented using PyTorch.

Our model is a regression model and takes as input multimodal sensor data corresponding to a duration of 15 seconds and predicts the cybersickness levels. Therefore, we use mean absolute error (MAE) and root mean square error (RMSE), two commonly used metrics for regression problems to evaluate the performance of different models. MAE is the average absolute error between predicted and actual values (Equation 1). RMSE is the square root of the average square error between predicted and actual values (Equation 2).

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

where n is the number of samples, y_i is the actual value, and $\hat{y}_i$ is the predicted value of the model. A model with MAE and RMSE values close to zero is considered to have better performance.

We measured the characteristics of cybersickness changes in 15 seconds. The data were randomly split into the training and the test sets at a ratio of 8:2. We performed a five-fold cross-validation using the training set. The model was trained with a learning rate of 0.001 and up to 500 epochs.

Table 3. The performance of models by data format and embedding method. “Normalized” refers to the normalized sensor data. TTE and STE+TTE methods are based on modality integration, and the spatial transformer encoder is not applied in the TTE method.

Data format	Embedding method	MAE	RMSE
Normalized	TTE	0.38	0.56
Normalized	STE+TTE	0.25	0.50
Normalized	MS-STTN	0.12	0.25
Spectrogram	STTE	0.50	0.70
Spectrogram	MS-STTS	0.37	0.55

5.1 Ablation study

We compared the performance of the models according to data format and embedding method as summarized in Table 3. MS-STTN yielded the best performance with an MAE of 0.12 and an RMSE of 0.25. MS-STTN performed better than integrated STE and TTE (STE+TTE), and MS-STTS performed better than STTE. These results suggest that the method of learning the characteristics of each data modality through independent encoders and then fusing them for the final prediction of cybersickness was effective, regardless of the data format. The result of the performance of STE+TTE compared to that of TTE also highlighted the role of spatial characterization in cybersickness prediction. It appears that our proposed spatial transformer encoder worked reasonably well. This suggests the importance of learning both spatial and temporal features from multimodal time-series sensor data. On the other hand, STTE performed worse than TTE. According to our experimental results, it seems that using normalized sensor data instead of spectrograms is more efficient in the context of cybersickness prediction. We will discuss this finding later in Section 6.1.

Table 4. Characteristics of the datasets used in our evaluation.

Dataset	Modality	Subjects	Size	Cybersickness level
Physiological dataset [37]	Physiological signals (EDA, BR, HR, HRV)	23	14.7K	restructured FMS (3 classes)
MSCVR (ours)	Eye movement	45	2.7K	0-20 FMS
	Head movement			
	Physiological signals (EDA, BVP, SKT)			

5.2 Direct performance comparison of cybersickness prediction models

We conducted a performance evaluation of publicly available cybersickness prediction models using both our MSCVR dataset and the publicly available physiological dataset.

5.2.1 Physiological datasets. The physiological dataset includes the following physiological signals—EDA, BR, HR, and HRV—from 23 participants who experienced a VR roller coaster simulation [37]. The following is the data pre-processing procedure according to the documentation. Data were collected during five minutes of VR relaxation content (stationary roller coaster car) and 15 minutes of fast roller coaster simulation. EDA was sampled at 10Hz, while HR, HRV, and BR were sampled at 1Hz. All modalities were downsampled to 1Hz based on the lowest sampling rate and normalized by min-max. Metrics including running average, rate of change, minimum, and maximum were calculated for the previous three seconds for each modality after the start of the VR relaxation content. Participants verbally responded to the level of cybersickness in the FMS every minute during the roller coaster simulation. The ground-truth labels were reconstructed from the participants' FMS responses into three classes (low: 0, moderate: 1, and acute: 2), and all 60 data points corresponding to one minute were assigned the same label.

Consequently, the physiological dataset is a single modality and the total number of data samples is 14,774, and our MSCVR dataset is multi-modality and the total number of data samples is 2,700. In both datasets, training data and test data were split in a ratio of 8:2.

Table 5. Direct performance comparison based on the Physiological and MSCVR datasets. Each model uses a specific backbone architecture, with the first three using a CNN-LSTM, while our MS-STTN model is based on a transformer architecture. “Year” refers to the year in which the model was proposed.

Model	Year	Backbone	Physiological dataset		MSCVR (ours)	
			MAE	RMSE	MAE	RMSE
Islam et al. [37]	2020	CNN-LSTM	0.33	0.22	0.54	0.38
Islam et al. [35]	2021	CNN-LSTM	0.46	0.28	0.50	0.62
MAC [40]	2023	CNN-LSTM	0.15	0.11	0.17	0.28
MS-STTN (ours)	2024	Transformer	0.13	0.16	0.12	0.25

5.2.2 Evaluation. Table 5 shows the performance of four models on two datasets. MS-STTN yielded the best performance in most of the evaluation matrices on both datasets, highlighting its robustness in understanding complex patterns in cybersickness data. However, in the RMSE metric of the physiological dataset, the MAC model outperformed MS-STTN by 0.05. This can be interpreted as an influence of the dataset. The physiological dataset contains four data features calculated over three seconds for each modality, resulting in signals with less variability and more consistent patterns. The spatial transformer encoder in MS-STTN is designed to learn the precise signal patterns of each modality; thus, it may not effectively capture pattern characteristics in data with low signal variability. Consequently, the RMSE performance, which penalizes larger errors more heavily,

appears to be somewhat reduced in MS-STTN. On the other hand, in the RMSE for the MSCVR dataset, MS-STTN showed a slightly higher performance than MAC by 0.03, indicating the effectiveness of our model on multimodal time series data. It is possible that MSCVR, with its multiple modalities, broader range of VR content, and larger participant population, contributed to the performance of the model.

Table 6. Indirect performance comparison of the regression models for predicting cybersickness based on sensor data. “Modality” refers to the sensor data used for model training. “-” means not reported in the study.

Model	Modality	MAE	RMSE
Lee et al. [57]	Physiological signal	-	7.34
Islam et al. [34]	Physiological signal	-	2.61
DeepTCN [36]	Eye movement	-	0.49
	Head movement		
	Physiological signal		
Kundu et al. [55]	LSTM Eye movement	0.21	0.36
	GRU Head movement	0.24	0.41
	MLP Physiological signal	0.40	0.57
MS-STTN (ours)	Eye movement	0.12	0.25
	Head movement		
	Physiological signal		

5.3 Indirect performance comparison of cybersickness prediction models

Cybersickness prediction modeling studies, as shown in Section 5.2, are somewhat limited due to a lack of publicly available code and datasets. Given this limitation, we had to conduct an indirect comparison of existing studies. Specifically, we focused on cybersickness prediction models that use regression and multimodal data (eye movement, head movement, and physiological signals) as summarized in Table 6. Overall, the model performed significantly better when using multimodal data than when using data from single modality data. In particular, MS-STTN showed reasonably good performance compared to the deep learning models presented in previous studies. In addition, the ranges of cybersickness levels used in the previous studies [36, 55] were different. Only our study considered the original range of the FMS from 0 to 20, while the previous studies shortened the range from 0 to 10 without any specific reason. Even in the longer range of the FMS, our model achieved lower MAE and RMSE, which further implies the higher or more generalizable performance of our model.

Table 7. The performance of our MS-STTN model by gender. The sample size was not large; thus, normality was tested using the Shapiro-Wilk test, and it was confirmed only for males.

Gender	Avg. of FMS (SD)	Shapiro-Wilk test	MS-STTN (ours)	
			MAE	RMSE
Male	2.91 (2.82)	W=0.76, p=5.98	0.84	0.77
Female	2.22 (2.07)	W=0.75, p <0.001	0.47	0.54

5.4 The performance of our model by user attribute

We conducted further analysis based on user attributes (Table 7). In this paper, we could not take age into account, because our participants were mostly in their twenties. For gender, since the two gender groups were independent and the female did not follow a normal distribution, we used the Mann-Whitney U test, a non-parametric test method. While our study did not show a significant difference (U=264.5, p=0.34), Nevertheless, we noted that

many studies have reported that female participants are more likely to experience higher levels of cybersickness in VR [16, 76, 80, 89]. Therefore, we additionally evaluated the performance of our model by gender.

As a result, the model using the female participant's data shows a higher performance of 0.37 and 0.23 in MAE and RMSE, respectively, than the model using the male participant's data. This can be seen as affecting the performance of the cybersickness prediction model, and there is a need to continuously consider gender when applying the model to content with a specific purpose in VR. In addition, the performance of the gender-specific model is slightly lower than when using a gender-integrated dataset. Since MSCVR consists of data from 24 males and 21 females, separating by gender and training the model reduced the amount of training data by half, resulting in a decrease in performance.

5.5 The feature importance through attention weight

As highlighted earlier, examining the importance of features during model training is important to understand which features influence the learning of cybersickness and to determine how to better design the model architecture and what data types should be used. To do this, we chose MS-STTN, which showed the highest performance in Section 5.1, extracted the importance of data features that had a high impact on the prediction of cybersickness, and analyzed the importance of each data feature over time.

The transformer encoder contains multi-head self-attention, which performs self-attention in parallel as many as the number of heads (we used six heads). For each input patch, its similarity to the cybersickness level can be extracted as a weight. We thus calculated the total attention weight by averaging the values of the heads over a patch.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \quad (3)$$

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, \dots, head_h)W^O \\ head_h &= Attention(QW_h^Q, KW_h^K, VW_h^V) \end{aligned} \quad (4)$$

where Q , K , and V is query, key, and value of self-attention and h is the number of heads. W_h^Q , W_h^K , and W_h^V are the weights of the $head_h$ for Q , K , and V , respectively. W^O is the extra parameter matrix to project the concatenation of the head outputs to the output space. $W_h^Q \in \mathbb{R}^{d_{model} \times d_q}$, $W_h^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_h^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_{model} \times d}$. d_{model} is the dimension of the output passed through the position embedding layer and d is d_{model}/h , which equals to d_q , d_k and d_v .

5.5.1 Eye movement. Figures 6-(a), (b), and (c) show the importance of eye movement data features for both eyes integrated, right eye, and left eye. For both eyes integrated, the combined gaze direction z generally showed the highest feature importance. The combined gaze direction y influenced cybersickness in the middle phase.

When comparing the right and left eyes, patterns of feature importance for each data feature were quite different. The gaze origin x and gaze origin z of the right eye and the gaze direction z and gaze origin x of the left eye showed high feature importance in most phases. These data features seem to have a strong influence on cybersickness. The gaze direction y , pupil diameter, and pupil position x were found to be influential in the middle phase. The gaze origin y showed an impact during the early and late phases. Lastly, the combined gaze direction x , eye openness, and gaze direction x show a relatively low feature importance in all time phases.

5.5.2 Head movement. Figure 6-(e) shows the importance of the head movement data features. We observed that head position x showed a very strong correlation with cybersickness in all time phases and head rotation *roll* showed a high influence in the middle and late phases. The head rotation *roll* refers to rotation around the

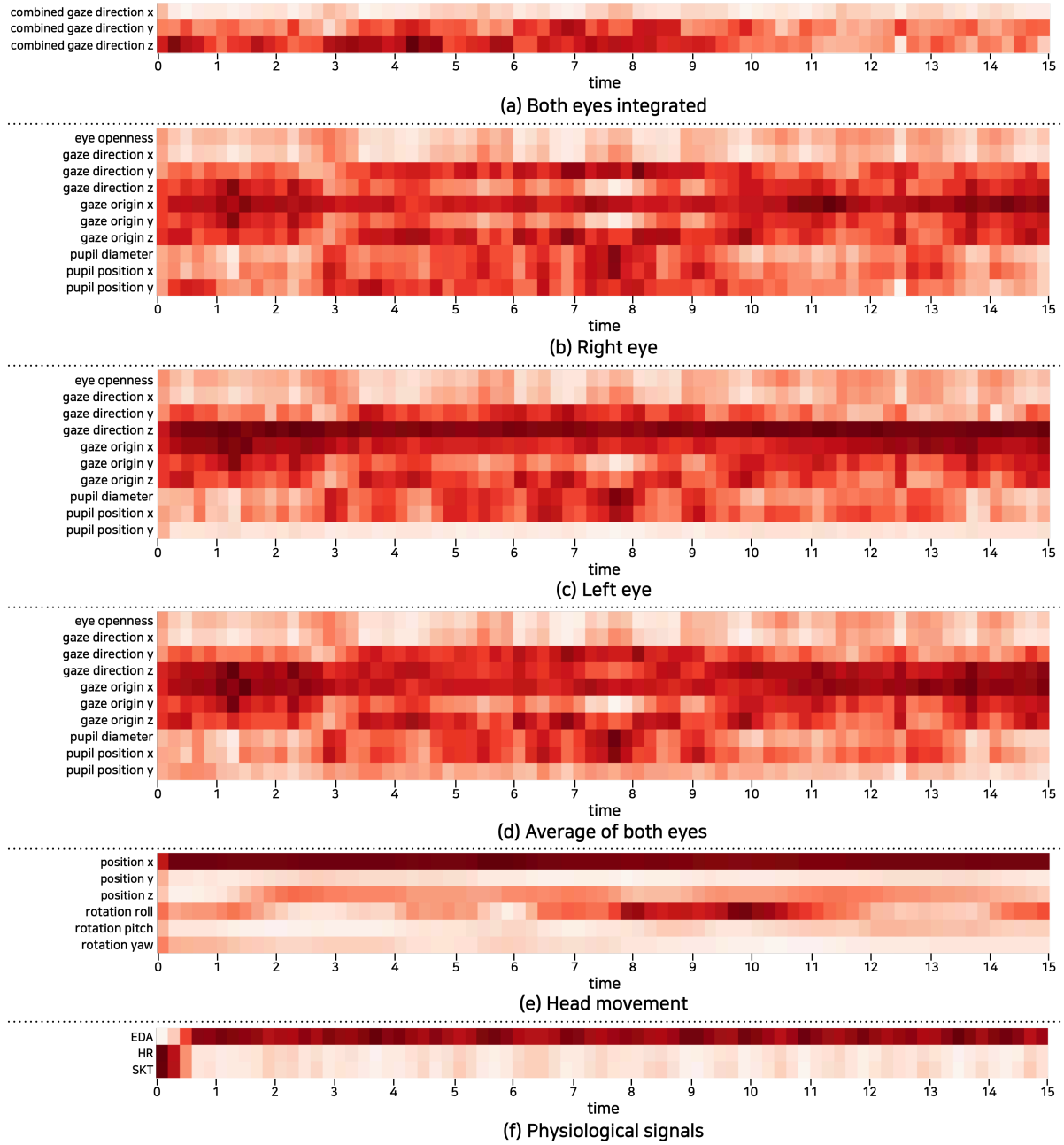


Fig. 6. Heatmaps of attention weights from the spatial transformer encoders of eye movement, head movement, and physiological signals. The correlation ranges from 0 to 1, where a darker red color denotes a stronger correlation with cybersickness.

x -axis, including nodding forward or tilting backward. It appears that the head position x signifies the left-right head movement of the user. As the VR videos used in our user study were 360 degrees, many participants made such head movements to explore their surroundings while experiencing the content. According to previous studies [5, 67], such active head movements can cause postural instability and display lag, potentially leading to cybersickness.

5.5.3 Physiological signals. Figure 6-(f) illustrates the feature importance of physiological signals. EDA showed a high correlation with cybersickness in all time phases. Previous studies have shown that EDA serves well as a quantitative indicator for assessing cybersickness levels [44, 62], and our results align with these findings. Interestingly, both HR and SKT showed high correlations with cybersickness in the very early phase but the values became smaller later on. On the other hand, EDA consistently showed a strong influence on cybersickness.

In summary, the following data features showed high correlations with cybersickness in most time phases: right eye gaze origin x , right eye gaze origin z , left eye gaze direction z , and left eye gaze origin x for eye movement, head position x for head movement, and EDA for physiological signals. As the importance of the data features is high, they may have influenced the level of cybersickness during the model training process. However, all data features are still influential to varying degrees. While it is possible to build a model using only important features, the model performance is likely to vary depending on the study context, users, etc. This means that taking advantage of various data features is likely to help build a model with good performance.

6 DISCUSSION

6.1 Input data embedding methods

Our experimental results showed little influence of spectrograms on cybersickness prediction (Section 5.1). Table 3 revealed that the modality-specific transformer is effective for both data formats, but MS-STTS showed small performance improvements compared to TTE, and STE+TTE had the lowest performance. To construct the spectrogram image, we converted the normalized sensor data into a spectrogram using STFT based on previous studies. While STFT has a great advantage in analyzing signals that have characteristics that change with time, its disadvantage is that the frequency resolution decreases due to the trade-off relationship between signals and time. Another possible reason relates to the size of the time window, which may have affected the frequency resolution. Since we collected a cybersickness level every 15 seconds, a window size with a different time length could affect the performance of the model. In addition, the lower performance of the spectrogram compared to the normalized data may not always be the case in other VR tasks or contexts. Yet, in this work, we tried to find the best conditions for cybersickness prediction among different embedding methods based on three modalities and a transformer. Thus, we believe that the use of a spectrogram method may be effective in other contexts that involve different data, virtual environments, and users.

6.2 Explainability of feature importance over time

In Section 5.5, we analyzed the importance of the data features over time for each data modality. Data features that showed a high correlation with cybersickness across all time phases included gaze direction z and gaze origin x in eye movement data, position x in head movement data, and EDA in physiological signals. In previous studies that investigated the correlation between cybersickness and data features through feature importance, Kundu et al. [53–55] extracted feature importance through XAI methods and reported that eye movement in deep learning and EDA and HR in machine learning were highly correlated with cybersickness. The deep learning model was trained using only eye and head movement data with high correlations, which showed an improvement in model performance. Jeong et al. [40] extracted feature importance using an attention mechanism, and eye movement data showed a strong effect on cybersickness. Combining the results of previous studies with ours, we confirmed

a strong correlation between cybersickness and eye movement data, which could be considered in different model structures and datasets.

Our experimental results also provide an opportunity to consider the importance of other data features in cybersickness in relation to time. For example, a pupil diameter, which is known to be highly correlated with cybersickness in many studies [13, 52, 65], was only activated in the middle phase in our study. This can be interpreted as participants experiencing cybersickness during specific time periods, thus it may be necessary to look closely at each time phase and compare differences in the value of the pupil diameter, other data features, and possibly even VR content that may increase or decrease the level of cybersickness. It would also be important to investigate the significance of data features over time and to build a cybersickness prediction model based on accurate and precise interpretations.

6.3 Limitations and future work

While our study results provide many important insights, some limitations can be considered for future work.

First, our study results may not be generalizable, even though we trained the transformer models and extracted the results using the MSCVR dataset collected from 45 participants. This number of participants in our study was more than that in other previous studies, which mostly had a range between 9 and 31 [14, 18, 35, 37, 40–43, 49, 55, 57, 70]. Our study also used diverse types of VR content in the data collection phase. Thus, we believe that our study results were comprehensive but model development with more participants and samples would be necessary.

Second, the ways in which data were preprocessed and prediction models were constructed could certainly be varied. Data can be formatted differently, and different model architectures can be considered depending on the domains and target variables. We also acknowledge that some of the data features that showed high correlations in our study may not be applicable in other domains. However, to the best of our knowledge, our study is the first to present cybersickness prediction using a transformer and considering spatial and temporal characteristics of sensor signals in VR. Our study provides meaningful insights into the handling of multimodal time-series sensor data in the development of a transformer-based prediction model. One of the methods considered in our study can be used to interpret multimodal time-series sensor data in other domains, and meaningful data analysis can be expected. We aim to develop a more complex model architecture using recent advanced methods, such as obtaining encoded sensor data from pre-trained models based on large-scale related datasets for better data representation, applying contrastive learning to learn the relationships between positive and negative pairs [1, 17].

Third, the length of the VR content may affect the performance of the model as well as the importance of the data features. Our results for feature importance over time were extracted based on the FMS every 15 seconds, but 15 seconds may not be an optimal time frame. It is certainly difficult to find the best time frame, and as described previously, other studies have used different time lengths [2, 7, 35]. Considering the more frequent collection of the FMS may be better in terms of using more data for modeling. However, this may degrade the user experience, because the user has to say the FMS score more frequently. Apart from this issue, regular collection of FMS can help with qualitative evaluation. For example, by interviewing users every 15 seconds (or any length of time depending on the researcher) based on the cybersickness result and the feature importance, the users can reflect on their VR experience and check whether they actually experienced significant cybersickness during the time period identified as highly important. This qualitative analysis can help complement the quantitative results obtained from the attention weights and provide a more comprehensive understanding of the user experience. Therefore, it would be important to determine an appropriate length of time for both researchers and users.

Lastly, it is important to confirm the practical applicability of cybersickness prediction models. We built a model using sensor data collected in a somewhat passive virtual environment, similar to previous cybersickness

studies. However, VR interactions can also take place in a more active virtual environment, such as walking, running, and other body movements. Thus, it would also be important to confirm whether our prediction model can work effectively in other types of virtual environments. An accurate understanding of the user's state is likely to help researchers design better user experiences in the context of VR [28, 82]. Additionally, addressing the challenge of reducing cybersickness after detection can be considered. While previous studies have proposed methods such as reducing the field of view or altering content rendering [11, 24], the integration of these methods with a cybersickness prediction model in a practical scenario remains untested, indicating a gap between model development and application. Even with the model that performs well on given training and test datasets, its performance may not be guaranteed in a practical environment, where more diverse data will be generated by users. Therefore, additional research on these aspects should be conducted to better apply the model to the practical level.

7 CONCLUSION

Research on mitigating cybersickness is actively being conducted as a way to support a satisfying VR user experience. In this work, we proposed transformer-based cybersickness prediction models in terms of *sensor data pre-processing* and *multimodal data fusion*. We adopted three sensor modalities highly related to cybersickness, based on previous theories, and constructed the MSCVR dataset through the user study. Two methods—MS-STTN and MS-STTS—for embedding multimodal time-series sensor data into a transformer were considered. The experimental results showed that MS-STTN performed the best (MAE = 0.12, RMSE = 0.25) in the ablation study and significantly performed well compared to other existing methods in previous studies. The analysis of the importance of data features revealed different degrees of feature importance over time, but many eye movement features (especially gaze direction and origin) showed strong effects on cybersickness, confirming the findings of previous studies. Finally, we discussed the applicability of the embedding methods and the explainability between cybersickness and data features. We hope that our work will be useful to researchers who wish to use multimodal time-series sensor data, and prediction models in the context of cybersickness or other aspects that are important for VR user experience.

ACKNOWLEDGMENTS

This research was supported by the National Research Foundation of Korea (NRF) (2021M3A9E4080780) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) (2020-0-01373).

REFERENCES

- [1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems* 34 (2021), 24206–24221.
- [2] Samuel Ang, Amanda Fernandez, Michael Rushforth, and John Quarles. 2023. You Make Me Sick! The Effect of Stairs on Presence, Cybersickness, and Perception of Embodied Conversational Agents. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 561–570.
- [3] Samuel Ang and John Quarles. 2022. You're in for a Bumpy Ride! Uneven Terrain Increases Cybersickness While Navigating with Head Mounted Displays. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 428–435.
- [4] Muhammad Shahid Anwar, Jing Wang, Wahab Khan, Asad Ullah, Sadique Ahmad, and Zesong Fei. 2020. Subjective QoE of 360-degree virtual reality videos and machine learning predictions. *IEEE Access* 8 (2020), 148084–148099.
- [5] Benjamin Arcioni, Stephen Palmisano, Deborah Apthorp, and Juno Kim. 2019. Postural stability predicts the likelihood of cybersickness in active HMD-based virtual reality. *Displays* 58 (2019), 3–11.
- [6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6836–6846.
- [7] Iqra Arshad, Paulo De Mello, Martin Ender, Jason D McEwen, and Elisa R Ferré. 2021. Reducing Cybersickness in 360-Degree virtual reality. *Multisensory Research* 35, 2 (2021), 203–219.

- [8] Mahdi Azmandian, Mark Hancock, Hrvoje Benko, Eyal Ofek, and Andrew D Wilson. 2016. Haptic retargeting: Dynamic repurposing of passive haptics for enhanced virtual reality experiences. In *Proceedings of the 2016 chi conference on human factors in computing systems*. 1968–1979.
- [9] Paulo Bala, Dina Dionísio, Valentina Nisi, and Nuno Nunes. 2018. Visually induced motion sickness in 360° videos: Comparing and combining visual optimization techniques. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 244–249.
- [10] Willem Bles, Jelte E Bos, Bernd De Graaf, Eric Groen, and Alexander H Wertheim. 1998. Motion sickness: only one provocative conflict? *Brain research bulletin* 47, 5 (1998), 481–487.
- [11] Pulkit Budhiraja, Mark Roman Miller, Abhishek K Modi, and David Forsyth. 2017. Rotation blurring: use of artificial blurring to reduce cybersickness in virtual reality first person shooters. *arXiv preprint arXiv:1710.02599* (2017).
- [12] Yudong Cao, Yifei Ding, Minping Jia, and Rushuai Tian. 2021. A novel temporal convolutional network with residual self-attention mechanism for remaining useful life prediction of rolling bearings. *Reliability Engineering & System Safety* 215 (2021), 107813.
- [13] Berk Cebeci, Ufuk Celikcan, and Tolga K Capin. 2019. A comprehensive study of the affective and physiological responses induced by dynamic virtual reality environments. *Computer Animation and Virtual Worlds* 30, 3-4 (2019), e1893.
- [14] Eunhee Chang, Hyun Taek Kim, and Byounghyun Yoo. 2021. Predicting cybersickness based on user’s gaze behaviors in HMD-based virtual reality. *Journal of Computational Design and Engineering* 8, 2 (2021), 728–739.
- [15] Fang-Yi Chao, Cagri Ozcinar, and Aljosa Smolic. 2021. Transformer-based Long-Term Viewport Prediction in 360° Video: Scanpath is All You Need.. In *MMSP*. 1–6.
- [16] Simon Davis, Keith Nesbitt, and Eugene Nalivaiko. 2014. A systematic review of cybersickness. In *Proceedings of the 2014 conference on interactive entertainment*. 1–9.
- [17] Shohreh Deldari, Hao Xue, Aaqib Saeed, Daniel V Smith, and Flora D Salim. 2022. Cocoa: Cross modality contrastive learning for sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–28.
- [18] Mark S Dennison, A Zachary Wisti, and Michael D’Zmura. 2016. Use of physiological signals to predict cybersickness. *Displays* 44 (2016), 42–52.
- [19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [20] Iveta Dirgová Luptáková, Martin Kubovčík, and Jiří Pospíchal. 2022. Wearable sensor-based human activity recognition with transformer model. *Sensors* 22, 5 (2022), 1911.
- [21] Xiao Dong, Ken Yoshida, and Thomas A Stoffregen. 2011. Control of a virtual vehicle influences postural activity and motion sickness. *Journal of Experimental Psychology: Applied* 17, 2 (2011), 128.
- [22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [23] Runyu Fan, Jun Li, Weijing Song, Wei Han, Jining Yan, and Lizhe Wang. 2022. Urban informal settlements classification via a transformer-based spatial-temporal fusion network using multimodal remote sensing and time-series human activity data. *International Journal of Applied Earth Observation and Geoinformation* 111 (2022), 102831.
- [24] Ajoy S Fernandes and Steven K Feiner. 2016. Combating VR sickness through subtle dynamic field-of-view modification. In *2016 IEEE symposium on 3D user interfaces (3DUI)*. IEEE, 201–210.
- [25] Maria Gallagher and Elisa Raffaella Ferrè. 2018. Cybersickness: a multisensory integration perspective. *Multisensory research* 31, 7 (2018), 645–674.
- [26] Alireza Mazloumi Gavgani, Keith V Nesbitt, Karen L Blackmore, and Eugene Nalivaiko. 2017. Profiling subjective symptoms and autonomic changes associated with cybersickness. *Autonomic Neuroscience* 203 (2017), 41–50.
- [27] John F Golding. 1998. Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. *Brain research bulletin* 47, 5 (1998), 507–516.
- [28] Kunal Gupta, Sam WT Chan, Yun Suen Pai, Nicholas Strachan, John Su, Alexander Sumich, Suranga Nanayakkara, and Mark Billinghurst. 2022. Total VREcall: Using Biosignals to Recognize Emotional Autobiographical Memory in Virtual Reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–21.
- [29] Bruno Herbelin, Roy Salomon, Andrea Serino, and Olaf Blanke. 2016. *Neural mechanisms of bodily self-consciousness and the experience of presence in virtual reality*. Technical Report. De Gruyter.
- [30] Lawrence J Hettinger, Kevin S Berbaum, Robert S Kennedy, William P Dunlap, and Margaret D Nolan. 1990. Vection and simulator sickness. *Military psychology* 2, 3 (1990), 171–181.
- [31] Franz Hlawatsch, Gloria Faye Boudreaux-Bartels, et al. 1992. Linear and quadratic time-frequency signal representations. *IEEE signal processing magazine* 9, 2 (1992), 21–67.
- [32] Zhongxu Hu, Yiran Zhang, Yang Xing, Yifan Zhao, Dongpu Cao, and Chen Lv. 2022. Toward human-centered automated driving: A novel spatiotemporal vision transformer-enabled head tracker. *IEEE Vehicular Technology Magazine* 17, 4 (2022), 57–64.

- [33] Momal Ijaz, Renato Diaz, and Chen Chen. 2022. Multimodal transformer for nursing activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2065–2074.
- [34] Rifatul Islam, Samuel Ang, and John Quarles. 2021. Cybersense: A closed-loop framework to detect cybersickness severity and adaptively apply reduction techniques. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 148–155.
- [35] Rifatul Islam, Kevin Desai, and John Quarles. 2021. 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)*. IEEE, 31–40.
- [36] Rifatul Islam, Kevin Desai, and John Quarles. 2022. Towards Forecasting the Onset of Cybersickness by Fusing Physiological, Head-tracking and Eye-tracking with Multimodal Deep Fusion Network. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 121–130.
- [37] Rifatul Islam, Yonggun Lee, Mehrad Jaloli, Imtiaz Muhammad, Dakai Zhu, Paul Rad, Yufei Huang, and John Quarles. 2020. Automatic detection and prediction of cybersickness severity using deep neural networks from user's physiological signals. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 400–411.
- [38] Arash Jahangiri and Hesham A Rakha. 2015. Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. *IEEE transactions on intelligent transportation systems* 16, 5 (2015), 2406–2417.
- [39] Dayoung Jeong and Kyungsik Han. 2022. Leveraging multimodal sensory information in cybersickness prediction. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*. 1–2.
- [40] Dayoung Jeong, Seungwon Paik, YoungTae Noh, and Kyungsik Han. 2023. MAC: multimodal, attention-based cybersickness prediction modeling in virtual reality. *Virtual Reality* (2023), 1–16.
- [41] Daekyo Jeong, Sangbong Yoo, and Jang Yun. 2019. Cybersickness analysis with eeg using deep learning algorithms. In *2019 IEEE conference on virtual reality and 3D user interfaces (VR)*. IEEE, 827–835.
- [42] Hyunwook Jeong, Hak Gu Kim, and Yong Man Ro. 2017. Visual comfort assessment of stereoscopic images using deep visual and disparity features based on human attention. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 715–719.
- [43] Weina Jin, Jianyu Fan, Diane Gromala, and Philippe Pasquier. 2018. Automatic prediction of cybersickness for virtual reality games. In *2018 IEEE Games, Entertainment, Media Conference (GEM)*. IEEE, 1–9.
- [44] Sungchul Jung, Richard Li, Ryan McKee, Mary C Whitton, and Robert W Lindeman. 2021. Floor-vibration vr: mitigating cybersickness using whole-body tactile stimuli in highly realistic vehicle driving experiences. *IEEE Transactions on Visualization & Computer Graphics* 27, 05 (2021), 2669–2680.
- [45] RS Kennedy, A Graybiel, RC McDonough, and D Beckwith. 1968. Symptomatology under storm conditions in the North Atlantic in control subjects and in persons with bilateral labyrinthine defects. *Acta oto-laryngologica* 66, 1-6 (1968), 533–540.
- [46] Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology* 3, 3 (1993), 203–220.
- [47] Behrang Keshavarz and Heiko Hecht. 2011. Validating an efficient method to quantify motion sickness. *Human factors* 53, 4 (2011), 415–426.
- [48] Hyewon Kim, Dong Jun Kim, Won Ho Chung, Kyung-Ah Park, James DK Kim, Dowan Kim, Kiwon Kim, and Hong Jin Jeon. 2021. Clinical predictors of cybersickness in virtual reality (VR) among highly stressed people. *Scientific reports* 11, 1 (2021), 12139.
- [49] Hak Gu Kim, Wissam J Baddar, Heoun-taek Lim, Hyunwook Jeong, and Yong Man Ro. 2017. 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*. 1–7.
- [50] Jinwoo Kim, Woojae Kim, Heeseok Oh, Seongmin Lee, and Sanghoon Lee. 2019. A deep cybersickness predictor based on brain signal analysis for virtual reality contents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10580–10589.
- [51] Jinwoo Kim, Heeseok Oh, Woojae Kim, Seonghwa Choi, Wookho Son, and Sanghoon Lee. 2020. A deep motion sickness predictor induced by visual stimuli in virtual reality. *IEEE Transactions on Neural Networks and Learning Systems* (2020).
- [52] Panagiotis Kourtesis, Josie Linnell, Rayaam Amir, Ferran Argelaguet, and Sarah E MacPherson. 2023. Cybersickness in virtual reality questionnaire (csq-vr): A validation and comparison against ssq and vrsq. In *Virtual Worlds*, Vol. 2. MDPI, 16–35.
- [53] Ripan Kumar Kundu, Osama Yahia Elsaid, Prasad Calyam, and Khaza Anuarul Hoque. 2023. VR-LENS: Super Learning-based Cybersickness Detection and Explainable AI-Guided Deployment in Virtual Reality. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 819–834.
- [54] Ripan Kumar Kundu, Rifatul Islam, Prasad Calyam, and Khaza Anuarul Hoque. 2022. TruVR: Trustworthy Cybersickness Detection using Explainable Machine Learning. *arXiv preprint arXiv:2209.05257* (2022).
- [55] Ripan Kumar Kundu, Rifatul Islam, John Quarles, and Khaza Anuarul Hoque. 2023. LiteVR: Interpretable and Lightweight Cybersickness Detection using Explainable AI. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 609–619.
- [56] Joseph J LaViola Jr. 2000. A discussion of cybersickness in virtual environments. *ACM Sigchi Bulletin* 32, 1 (2000), 47–56.

- [57] Sangmin Lee, Seongyeop Kim, Hak Gu Kim, Min Seob Kim, Seokho Yun, Bumseok Jeong, and Yong Man Ro. 2019. Physiological fusion net: Quantifying individual vr sickness with content stimulus and physiological response. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 440–444.
- [58] Yun-Xuan Lin, Rohith Venkatakrishnan, Roshan Venkatakrishnan, Elham Ebrahimi, Wen-Chieh Lin, and Sabarish V Babu. 2020. How the presence and size of static peripheral blur affects cybersickness in virtual reality. *ACM Transactions on Applied Perception (TAP)* 17, 4 (2020), 1–18.
- [59] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
- [60] Arnab Kumar Mondal, Arnab Bhattacharjee, Parag Singla, and AP Prathosh. 2021. xViTCOS: explainable vision transformer based COVID-19 screening using radiography. *IEEE Journal of Translational Engineering in Health and Medicine* 10 (2021), 1–10.
- [61] KE Money and BS Cheung. 1983. Another function of the inner ear: facilitation of the emetic response to poisons. *Aviation, space, and environmental medicine* (1983).
- [62] Eugene Nalivaiko, John A Rudd, and Richard HY So. 2014. Motion sickness, nausea and thermoregulation: The “toxic” hypothesis. *Temperature* 1, 3 (2014), 164–171.
- [63] Charles M Oman. 1982. A heuristic mathematical model for the dynamics of sensory conflict and motion sickness. *Acta Oto-Laryngologica* 94, sup392 (1982), 4–44.
- [64] Charles M Oman. 1990. Motion sickness: a synthesis and evaluation of the sensory conflict theory. *Canadian journal of physiology and pharmacology* 68, 2 (1990), 294–303.
- [65] Hiroshi Oyamada, Atsuhiko Iijima, Akira Tanaka, Kazuhiko Ukai, Haruo Toda, Norihiro Sugita, Makoto Yoshizawa, and Takehiko Bando. 2007. A pilot study on pupillary and cardiovascular changes induced by stereoscopic video movies. *Journal of neuroengineering and rehabilitation* 4 (2007), 1–7.
- [66] Nitish Padmanaban, Timon Ruban, Vincent Sitzmann, Anthony M Norcia, and Gordon Wetzstein. 2018. Towards a machine-learning approach for sickness prediction in 360 stereoscopic videos. *IEEE transactions on visualization and computer graphics* 24, 4 (2018), 1594–1603.
- [67] Stephen Palmisano, Robert S Allison, Joel Teixeira, and Juno Kim. 2022. Differences in virtual and physical head orientation predict sickness during active head-mounted display-based virtual reality. *Virtual Reality* (2022), 1–21.
- [68] Sangjoon Park, Gwanghyun Kim, Jeongsol Kim, Boah Kim, and Jong Chul Ye. 2021. Federated split task-agnostic vision transformer for COVID-19 CXR diagnosis. *Advances in Neural Information Processing Systems* 34 (2021), 24617–24630.
- [69] Michael Portnoff. 1980. Time-frequency representation of digital signals and systems based on short-time Fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28, 1 (1980), 55–69.
- [70] Chenxin Qu, Xiaoping Che, Siqi Ma, and Shuqin Zhu. 2022. Bio-physiological-signals-based VR cybersickness detection. *CCF Transactions on Pervasive Computing and Interaction* (2022), 1–17.
- [71] James T Reason. 1978. Motion sickness adaptation: a neural mismatch model. *Journal of the Royal Society of Medicine* 71, 11 (1978), 819–829.
- [72] Lisa Rebenitsch and Charles Owen. 2016. Review on cybersickness in applications and visual displays. *Virtual Reality* 20, 2 (2016), 101–125.
- [73] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why should i trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
- [74] Gary E Riccio and Thomas A Stoffregen. 1991. An ecological theory of motion sickness and postural instability. *Ecological psychology* 3, 3 (1991), 195–240.
- [75] Dante Risi and Stephen Palmisano. 2019. Effects of postural stability, active control, exposure duration and repeated exposures on HMD induced cybersickness. *Displays* 60 (2019), 9–17.
- [76] Daniel M Shafer, Corey P Carbonara, and Michael F Korpi. 2017. Modern virtual reality technology: cybersickness, sense of presence, and gender. *Media Psychology Review* 11, 2 (2017), 1.
- [77] Muhammad Shahid Anwar, Jing Wang, Sadique Ahmad, Asad Ullah, Wahab Khan, and Zesong Fei. 2020. Evaluating the factors affecting QoE of 360-degree videos and cybersickness levels predictions in virtual reality. *Electronics* 9, 9 (2020), 1530.
- [78] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. 2022. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*. Springer, 461–478.
- [79] Debaditya Shome, T Kar, Sachi Nandan Mohanty, Prayag Tiwari, Khan Muhammad, Abdullah AlTameem, Yazhou Zhang, and Abdul Khader Jilani Saudagar. 2021. Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. *International Journal of Environmental Research and Public Health* 18, 21 (2021), 11086.
- [80] Kay Stanney, Cali Fidopiastis, and Linda Foster. 2020. Virtual reality is sexist: but it does not have to be. *Frontiers in Robotics and AI* 7 (2020), 4.

- [81] Thomas A Stoffregen and L James Smart Jr. 1998. Postural instability precedes motion sickness. *Brain research bulletin* 47, 5 (1998), 437–448.
- [82] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jitrapol Intarasisrisawat, Maxine Glancy, and Chee Siang Ang. 2021. Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 4 (2021), 1–20.
- [83] Ye Tian, Dulmini Hettiarachchi, and Shunsuke Kamijo. 2022. Transportation mode detection combining CNN and vision transformer with sensors recalibration using smartphone built-in sensors. *Sensors* 22, 17 (2022), 6453.
- [84] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR, 10347–10357.
- [85] Michel Treisman. 1977. Motion sickness: an evolutionary hypothesis. *Science* 197, 4302 (1977), 493–495.
- [86] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [87] Chenxing Wang, Haiyong Luo, Fang Zhao, and Yanjun Qin. 2020. Combining residual and LSTM recurrent networks for transportation mode detection using multimodal sensors integrated in smartphones. *IEEE Transactions on Intelligent Transportation Systems* 22, 9 (2020), 5473–5485.
- [88] Jialin Wang, Hai-Ning Liang, Diego Monteiro, Wenge Xu, and Jimin Xiao. 2022. Real-time prediction of simulator sickness in virtual reality games. *IEEE Transactions on Games* (2022).
- [89] Séamas Weech, Sophie Kenny, and Michael Barnett-Cowan. 2019. Presence and cybersickness in virtual reality are negatively related: a review. *Frontiers in psychology* 10 (2019), 158.
- [90] Carolin Wienrich, Christine Katharina Weidner, Celina Schatto, David Obremski, and Johann Habakuk Israel. 2018. A virtual nose as a rest-frame-the impact on simulator sickness and game experience. In *2018 10th international conference on virtual worlds and games for serious applications (VS-Games)*. IEEE, 1–8.
- [91] Birsan Yazici and Gerald B Kliman. 1999. An adaptive statistical time-frequency method for detection of broken bars and bearing faults in motors using stator current. *IEEE Transactions on Industry Applications* 35, 2 (1999), 442–452.
- [92] Yunkai Yu, Zhihong Yang, Peiyao Li, Zhicheng Yang, and Yuyang You. 2019. Work-in-progress: On the feasibility of lightweight scheme of real-time atrial fibrillation detection using deep learning. In *2019 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, 552–555.
- [93] Qi Zhang, Hengshu Zhu, Peng Wang, Enhong Chen, and Hui Xiong. 2023. Hierarchical Wi-Fi Trajectory Embedding for Indoor User Mobility Pattern Analysis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–21.
- [94] Ye Zhang, Longguang Wang, Huiling Chen, Aosheng Tian, Shilin Zhou, and Yulan Guo. 2022. IF-ConvTransformer: A framework for human activity recognition using IMU fusion and ConvTransformer. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–26.
- [95] Zijing Zhang and Edwin C Kan. 2022. Novel Muscle Monitoring by Radiomyography (RMG) and Application to Hand Gesture Recognition. *arXiv preprint arXiv:2211.03767* (2022).