Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

VRS-LLM: Personalized Prediction of Virtual Reality Sickness Using Large Language Models Based on Multimodal Time Series Data

Yoonseon Choi
Hanyang University
Dayoung Jeong
Hanyang University
Bogoan Kim
Chungbuk National University
Kyungsik Han
kyungsikhan@hanyang.ac.kr

Hanyang University

Research Article

Keywords: Virtual reality sickness, Personalized prediction, Large language model, Dynamic time warping, Multimodal time series data

Posted Date: June 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9919190/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

VRS-LLM: Personalized Prediction of Virtual Reality Sickness Using Large Language Models Based on Multimodal Time Series Data

Yoonseon Choi[1], Dayoung Jeong[2], Bogoan Kim[3,4], Kyungsik Han[1,2*]

[1*]Department of Data Science, Hanyang University, 222 Wangsimni-ro, Seoul, 04763, Republic of Korea.
[2]Department of Artificial Intelligence, Hanyang University, 222 Wangsimni-ro, Seoul, 04763, Republic of Korea.
[3]School of Information and Communication Engineering, Chungbuk National University, 1 Chungdae-ro, Seowon-gu, Cheongju, 28644, Republic of Korea.
[4]Cancer Research Center, Chungbuk National Univerity, 1 Chungdae-ro, Seowon-gu, Cheongju, 28644, Republic of Korea.

*Corresponding author(s). E-mail(s): kyungsikhan@hanyang.ac.kr;
Contributing authors: yoonseonchoi@hanyang.ac.kr; dayoungjeong@hanyang.ac.kr; bogoan@cbnu.ac.kr;

Abstract

Virtual reality sickness (VRS) remains a significant barrier to sustained user engagement, necessitating robust predictive modeling. Conventional approaches often rely on multiple sensors for data diversity, yet struggle to incorporate rich contextual information and face challenges in user-specific training with limited data. We present *VRS-LLM*, a framework that leverages the reasoning capabilities of large language models (LLMs) to predict VRS using head-mounted display (HMD) data. Our approach effectively extracts temporal features, integrates multifaceted context, and facilitates personalized prediction from individual-level data. *VRS-LLM* consists of two modules: *Modality Matcher*, which quantifies inter-modality relationships via Dynamic Time Warping (DTW), and *Context Matcher*, which encodes contextual factors into three structured types. Experiment results show superior performance (MSE 0.06, MAE 0.13) over baselines. Ablation studies validate the critical contributions of two modules. This work

highlights the potential of LLM-based user state prediction through effective data representations and prompt engineering using multimodal time series data.

Keywords: Virtual reality sickness, Personalized prediction, Large language model, Dynamic time warping, Multimodal time series data

1 Introduction

Virtual reality sickness (VRS) remains a critical impediment to user experience in immersive environments, characterized by symptoms including nausea, dizziness, and disorientation. Mitigating these effects necessitates proactive VRS prediction to enable the timely application of intervention strategies, such as modifying the field of view (Wu and Suma Rosenberg 2022) or integrating haptic feedback (Jung et al. 2021). Recent advancements in VR hardware and wearable sensors have facilitated the collection of diverse physiological and behavioral signals, driving research into deep learning-based predictive models (Islam et al. 2021; Jeong et al. 2023; Jeong and Han 2024; Yang et al. 2025). However, the relationship between sensor signals and VRS is inherently user-dependent, governed by individual differences in susceptibility and behavioral patterns (Kim et al. 2025b; Tian et al. 2022). Accordingly, there is a growing scholarly consensus on the necessity of personalized VRS prediction models derived from multimodal data (Lee et al. 2021; Li et al. 2023).

Despite these efforts, three fundamental limitations persist in current multimodal personalized VRS modeling. First, reliance on high-dimensional modalities introduces practical deployment constraints and may degrade user experience during data collection. While increasing the number of modalities—such as electroencephalography (EEG), electrodermal activity (EDA), heart rate (HR) and data (e.g., video, eye movements, and head movements) collected from head-mounted displays (HMDs)—typically enhances performance (Jeong et al. 2023; Islam et al. 2022), the requirement for additional wearable sensors increases system complexity, reduces ecological validity and user experience (Poria et al. 2017; Oloyede and Hancke 2016; Kamińska et al. 2022; Kim et al. 2022). Second, existing approaches often fail to adequately integrate contextual information within multimodal streams (Qu et al. 2022; Kim et al. 2020; Liao et al. 2020). Effectively interpreting time-varying data requires understanding the interactions and environmental conditions underlying specific sessions (Xu et al. 2024; Liu et al. 2024; Kim et al. 2023; Ge et al. 2024). While some studies have used additional modalities (e.g., demographics or sentiment indices) to provide semantic grounding (Jin et al. 2020; Rebenitsch and Owen 2021; Gao et al. 2023), these methods are often domain-dependent and lack generalizability. Third, existing deep learning architectures struggle with personalization due to the data-intensive nature of individual-level training. Real-world constraints on user-specific data volume and quality often lead to overfitting, hindering the development of robust, stable personalized models.

To address these limitations, we propose a three-fold approach. First, we utilized a streamlined set of three modalities (i.e., video, eye movement, and head movement) collected from HMDs, standard devices for VR, which are directly implicated in VRS induction (Caserman et al. 2021; Chang et al. 2021; Palmisano et al. 2020). Based on

these inputs, we designed a model architecture that effectively captures complex inter-modal dependencies. Central to our methodology is the systematic representation of multimodal temporal relationships, employing dynamic time warping (DTW) (Müller 2007) to quantify inter-modality synchronization. Second, we demonstrate that contextual information, often underrepresented in prior approaches, can be effectively synthesized by leveraging large language model (LLMs) pre-trained on extensive textual data and capable of understanding contexts. Third, we demonstrate the feasibility of personalized prediction without model retraining by leveraging LLMs in a few-shot setting with user-specific data.

We propose *VRS-LLM*, a personalized VRS prediction model that integrates DTW and LLMs to analyze multimodal time-series data derived from HMDs. *VRS-LLM* consists of two primary modules. The first module, *Modality Matcher*, employs DTW to quantitatively evaluate synchronization levels and pattern similarities across modality pairs. By applying DTW to three modalities—video, eye movement, and head movement—the module extracts nuanced temporal correlations between each time-series pair. The second module, *Context Matcher*, uses a structured prompt architecture to provide the LLM with contextual grounding for HMD data interpretation. The prompt systematically integrates (1) basic information e.g., number of participants, VRS rating scales) related to VRS measurement, (2) semantic scene descriptions of the VR content, and (3) data feature specifications detailing the input characteristics and the DTW application methodology. To validate the effectiveness and generalizability of our approach, we evaluated *VRS-LLM* across multiple pre-trained LLMs of varying scales and architectures, including GPT, LLaMA, and Gemini families. *VRS-LLM* achieved the greatest personalized prediction in a few-shot configuration (MSE = 0.06, MAE = 0.13) and demonstrated consistent performance across LLM families. In addition, ablation studies confirmed the necessity of each component. Specifically, the data representation with DTW outperformed that with traditional metrics (e.g., Euclidean distance, correlation, and cosine similarity), validating the choice of DTW for capturing temporal dynamics. Incremental evaluation of the *Context Matcher* also revealed that the addition of contextual information layers led to progressive performance gains. *VRS-LLM* offers a structured paradigm for designing data representations and prompts that facilitate the analysis of inter-modality interactions within multimodal time series. This architecture holds broad potential for various user-state prediction tasks where understanding temporal relationships across modalities is important, such as stress detection or sleep stage classification (Kim et al. 2025a).

In summary, the contributions of this paper are as follows:

- We propose *VRS-LLM*, a personalized VRS prediction framework that effectively models complex inter-modality interactions using exclusively HMD-derived time-series data.
- To facilitate the integration of multimodal time-series data with LLMs, we introduce a DTW-based input representation and a structured prompt that incorporates multifaceted contextual information.
- We demonstrate that pre-trained LLMs to facilitate accurate personalized VRS prediction without requiring user-specific model training.

2 Related Work

2.1 Data Representation Techniques for Time Series

Extensive research on time series analysis has used multivariate or multimodal data to integrate diverse information sources, yielding more accurate and robust predictions. In the context of VRS prediction, temporal discrepancies frequently arise between signal pairs (e.g., video–eye, eye–head) even when captured by a single HMD device. Consequently, it is imperative to develop models capable of accounting for such misalignments. Various data representation techniques have been proposed to effectively characterize the relationships across time series sequences or modalities.

A foundational approach involves using Euclidean distance to measure point-to-point similarity at identical time indices. For example, Kalpakis et al. (Kalpakis et al. 2001) used this metric to quantify time series similarity for subsequent clustering tasks. Other methods use Pearson correlation or cosine similarity to capture global dependencies. Zhou et al. (Zhou et al. 2020) measured synchrony between EEG signals using correlation to quantify the linear relationships for epileptic seizure data analysis. Dong et al. (Dong et al. 2006) applied cosine similarity to detect anomalies by comparing current patterns against normative reference signals. However, these methods primarily assume linear relationships and fixed temporal alignments, limiting their efficacy in capturing nonlinear dynamics and the stochastic temporal variations inherent in human behavioral data.

To address these limitations, one of the most widely used techniques is dynamic time warping (DTW), which allows temporal warping between sequences of varying lengths. This capability facilitates the alignment of nonlinear patterns and the derivation of quantitative similarity measures that are robust to phase shifts. Xia et al. (Xia et al. 2024) employed DTW to compute similarities between multivariate financial time series data and constructed a dynamic hypergraph for volatile stock price prediction. Similarly, Masnad et al. (Masnad et al. 2019) employed DTW in human activity recognition to smartphone sensor data (gyroscope and accelerometer) by comparing reference and test samples.

As demonstrated, various data representation techniques have been employed to effectively model the relationships between time series data. Within VR environments, modalities collected via HMDs often exhibit variable sampling rates and asynchronous patterns, necessitating alignment strategies that account for nonlinear temporal variations. In this study, we design a DTW-based *Modality Matcher* to accurately capture the relationships between HMD-based multimodal data without requiring auxiliary sensors, thereby facilitating high-precision VRS prediction through enhanced relational modeling.

2.2 Leveraging LLMs for Time Series Data

As LLMs continue to demonstrate versatility across diverse domains, their potential for time series forecasting has received significant research interest. Recent studies have attempted to interpret the contextual nuances embedded in temporal data by

using the pre-trained knowledge and reasoning capabilities of of LLMs, leading to the emergence of methodological streams.

Research leveraging the representational capacity of pre-trained models often employs attention-based mechanisms to align temporal information with linguistic embedding spaces. For example, Zhou et al. (Zhou et al. 2023) introduced GPT4TS, which uses parameters from a pre-trained GPT-2 (Radford et al. 2019) to compute attention over time series data for downstream tasks such as imputation, classification, anomaly detection, and prediction. Expanding on this, Jin et al. (Jin et al. 2023) developed Time-LLM, which aligns time series data with the pre-trained word embeddings of LLaMA 2 7B (Touvron et al. 2023) and reinforces contextual understanding through prompts that incorporate domain knowledge, statistical information, and task instructions. Liu et al. (Liu et al. 2024) proposed TimeCMA, using a cross-modal alignment mechanism to combine variable-wise encoded embeddings with GPT-2 tokens. While these alignment-centric models effectively harness pre-trained representations, they often require significant computational overhead and face efficiency constraints during fine-tuning.

In contrast, prompt-based approaches have been proposed to convert time series data into natural language input. Xue and Salim (Xue and Salim 2023) proposed PromptCast, a prompt-based model for time series forecasting. The input prompt consists of a context part that contains past time series data along with corresponding descriptions, and a question part that requests predictions for future values. The output prompt is designed to generate natural language responses using a predefined template. Chen et al. (Chen et al. 2024) proposed Sensor2Text, a model designed to support dialogue generation about a user's daily activities based on wearable sensor data (eye, muscle, and body tracking). Data from each modality is processed through a Transformer encoder and a Querying Transformer to generate "sensor tokens," which serve as embedding vectors for the LLM input. These tokens are then included in the prompt along with an instruction specifying the LLM's task. Englhardt et al. (Englhardt et al. 2024) designed a Chain-of-Thought (CoT) (Wei et al. 2022)-based prompt to infer depression status and symptoms. Their prompt includes task and dataset descriptions, as well as domain knowledge based on the DSM-V (EDITION 1980), and employs a step-by-step questioning structure to help the LLM understand various formats of raw sensor data (e.g., CSV, tabular, and markdown).

Our work extends these prompt-based approaches by introducing a preprocessing pipeline tailored for heterogeneous HMD data—including video, eye movement, and head movement signals—along with prompt construction strategies. These components enable LLMs to better interpret the characteristics of the input data for the VRS prediction task. Specifically, our study proposes a *Context Matcher* that constructs a hierarchical prompt structure—compromising domain-level background (Basic Information), content-level variation (Scene Information), and feature-level structure (Feature Information)—to facilitate the LLM's capacity for causal reasoning, specifically optimized for the complexities of VRS prediction.

2.3 Prediction Models for VRS

Physiological signals (e.g., EDA, EEG) and movement data (e.g., eye and head movements) are frequently utilized in VRS prediction tasks, as they are directly linked to the vestibular and ocular disturbances that induce sickness. Consequently, recent studies have actively explored deep learning-based modeling approaches designed to map these high-dimensional signals to VRS prediction.

Existing research focused on establishing the efficacy of diverse neural network architectures. Islam et al. (Islam et al. 2021) developed a deep neural network (DNN) model using four physiological signals (HR, breathing rate (BR), heart rate variability (HRV), galvanic skin response (GSR)), achieving an RMSE of 2.47. In a follow-up study (Islam et al. 2021), they subsequently proposed a CNN-LSTM-based model using only HMD data, which improved the RMSE to 0.51. Jeong and Han (Jeong and Han 2024) proposed a Transformer-based model, MS-STTN, which integrated multimodal data (i.e., EDA, HR, skin temperature (SKT), and eye/head movement data) to achieve an RMSE of 0.25. Other work has also focused on classification. For example, Jeong et al. (Jeong et al. 2023) developed an attention-based classification model that classify the severity of VRS with an accuracy of 87% using multimodal data (i.e., video, eye/head movement data, EDA). As VRS symptoms can vary depending on individual-specific characteristics such as susceptibility to motion sickness and prior experience with VR, the need for personalized prediction has also been emphasized (Ramaseri-Chandra and Reza 2024; Wang et al. 2021). However, most deep learning-based models have limited capacity to capture these characteristics, primarily due to user overlap between the training and test datasets and the need for a large number of participants to learn individual-level biases.

To mitigate these limitations, researchers have explored personalization through demographic integration (e.g., gender, age), subjective background data (e.g., VR experience, susceptibility to motion sickness), or specialized training regimes such as transfer learning or sample weighting. Jin et al. (Jin et al. 2018) developed a support vector regression (SVR) model that combines questionnaire-based user profiles with head-movement data, achieving an MSE of 0.01. Tasnim et al. (Tasnim et al. 2024) developed multiple adaptation strategies—including early shaping and demographic grouping—achieving an RMSE of 0.21. Lee et al. (Lee et al. 2021) implemented a personalized model by using physiological signals along with video data, designing a feature extractor that accounts for individual physiological variability.

Despite these advancements, existing personalization methods often rely on broad group-level heuristics or require computationally expensive retraining and fine-tuning stages. Our work addresses these limitations by leveraging the in-context learning capabilities of pre-trained LLMs. By providing the model with a limited set of user-specific examples via few-shot learning, our approach facilitates flexible, individual-level inference without the need for model retraining. This approach simplifies the personalization pipeline while maintaining high sensitivity to the variations in user response patterns.

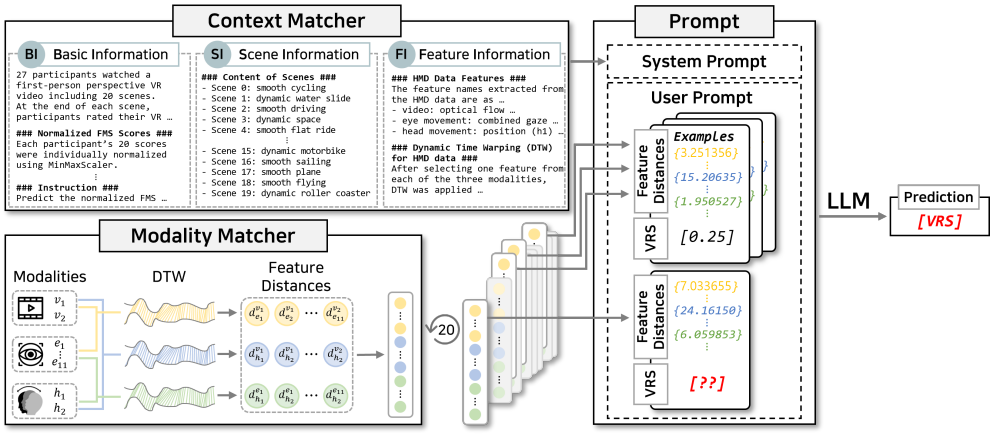


Fig. 1 Architecture of VRS-LLM, consisting of (a) the Contextual Matcher module and (b) the Modality Matcher module. (a) incorporates contextual information of the input attributes into the prompt, and (b) synchronizes time variations across modalities using DTW. Each module provides information for prompt generation, and the LLM performs the prediction of VRS.

3 Method

In this section, we present the methodology for developing *VRS-LLM*. The overall architecture is illustrated in Fig. 1. To predict VRS, we process multimodal time series data collected from an HMD—including VR video, eye movements, and head movements—for ingestion into an LLM. To effectively represent these data, we propose two key modules: (1) *Modality Matcher* and (2) *Context Matcher*. *Modality Matcher* utilizes the DTW algorithm to quantify temporal misalignment across modalities. *Context Matcher* constructs a structured prompt to provide the LLM with necessary semantic grounding. By integrating these strategies, *VRS-LLM* facilitates high-precision, LLM-based VRS prediction.

Table 1 Features and dimensions of the data collected from the HMD sensors.

Modalities	Features	Dimension (axes)
Eye movement	combined gaze direction	3D (x, y, z)
	right / left eye openness	1D
	right / left gaze direction	3D (x, y, z)
	right / left gaze origin	3D (x, y, z)
	right / left pupil diameter	1D
	right / left pupil position	2D (x, y)
Head movement	head position	3D (x, y, z)
	head rotation	3D $(roll, pitch, yaw)$

3.1 Dataset

We used the publicly-available dataset (Jeong et al. 2023), which consists of HMD-derived signals and corresponding subjective sickness scores. Data were collected from 27 participants using an HTC VIVE Pro Eye during sessions where each participant viewed 20 VR videos from a seated position. The dataset comprises three primary modalities—(1) VR video, (2) eye movements, and (3) head movements—along with Fast Motion Sickness Scale (FMS) scores that reflect perceived VRS levels. Each video is 40 to 45 seconds long, recorded in a 360-degree first-person perspective, and presented at 30 frames per second (fps). These videos encompass 10 thematic categories (i.e., cycling, water slides, driving, space travel, sailing, flat rides, motorbikes, flying, planes, and roller coasters), systematically designed to induce varying VRS intensities through "smooth" or "dynamic" motion profiles. Eye and head movement data were recorded as vector values using the HMD's built-in sensors while participants watched the videos. A summary of the features extracted from the eye and head movement data is presented in Table 1. Following each session, participants verbally reported their FMS scores on a scale of 0 ("no sickness at all") to 20 ("frank sickness").

3.2 Data Preprocessing

3.2.1 VR video

To quantify the dynamic variations within the video, such as object motion or global screen transitions, we applied an optical flow algorithm (Farnebäck 2003). This algorithm converts successive video frames to grayscale and computes pixel-wise motion vectors along the x- and y-axes to characterize inter-frame displacement. From these pixel-level vectors, we derived the average motion along the x-axis (dx) and y-axis (dy) for each frame. For a frame of a height h and width w, the motion vector at pixel (i, j) is represented as $flow[i, j, 0]$ for the x-axis and $flow[i, j, 1]$ for the y-axis. The average motion values, dx and dy, are calculated as follows:

$$dx = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} flow[i, j, 0], \quad dy = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} flow[i, j, 1] \tag{1}$$

3.2.2 Eye and Head Movements

Eye and head movement data were preprocessed to extract temporal variations between time steps. As shown in Table 1, the data were recorded as vector values and stored in CSV format. Each column in the dataset represents a univariate time series corresponding to a single axis. Given that user movements involve integrated multi-axial rotations and translations rather than isolated single-axis maneuvers, independent analysis of each axis may fail to capture holistic behavioral patterns. To address this, we derived a unified feature representing the magnitude of the vector displacement (Δv) by computing the Euclidean distance between consecutive time steps, defined as follows:

$$1D: \quad \Delta v = |x_{t+1} - x_t| \tag{2}$$

$$2D: \quad \Delta v = \sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2} \tag{3}$$

$$3D: \quad \Delta v = \sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2 + (z_{t+1} - z_t)^2} \tag{4}$$

, where (x_t, y_t, z_t) represent the position at time step t, while $(x_{t+1}, y_{t+1}, z_{t+1})$ denote the position at the next time step.

This procedure quantitatively captures both the magnitude of the vector changes and the temporal continuity of movement patterns. In addition, this approach facilitates significant dimensionality reduction, decreasing the number of eye movement features from 23 to 11, and the number of head movement features from 6 to 2. Despite this reduction, the fundamental variability of the signals was preserved, yielding a compact yet highly informative representation of participant movement data.

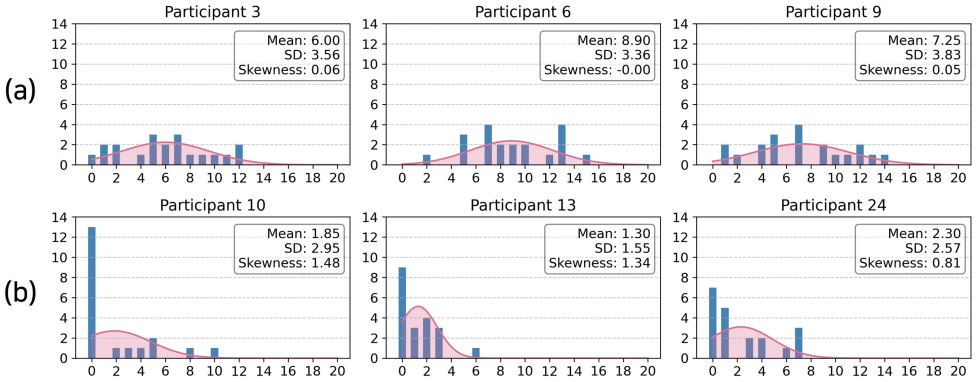


Fig. 2 Distributions of FMS scores from some participants. The x-axis indicates the FMS scale scores (0-20), and the y-axis indicates the response frequency for each score. Each subplot corresponds to one participant. Row (a) presents relatively uniform or symmetric distributions, while row (b) presents skewed distributions. Descriptive statistics (mean, standard deviation (SD), and skewness) are displayed in each subplot. Subplots for all 27 participants are provided in Appendix A.

3.2.3 Normalization for FMS scores

Min-max scaling was applied to normalize the FMS scores to a range between 0 and 1 for each participant. Analysis of the frequency distribution, mean, standard deviation, and skewness of the reported FMS scores revealed that most participants exhibited asymmetric distributions (Fig. 2-(b)). The VRS levels were not uniformly distributed across the full FMS scale (0-20), but were concentrated within participant-specific intervals, suggesting that perceived VRS levels are relative to each individual. Using raw FMS scores without accounting for these distributional characteristics could cause the model to generate predictions outside of a participant's observed response range. To mitigate this, we used normalized FMS scores.

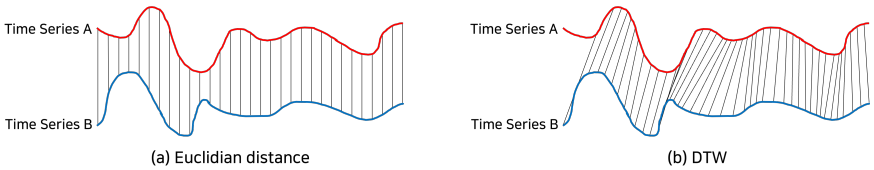


Fig. 3 Illustration of DTW applied to two time series. (a) shows two unaligned sequences with different temporal dynamics. (b) shows how DTW non-linearly adjusts the time axis to align the sequences, minimizing the distance and increasing the comparability of the patterns.

3.3 Modality Matcher

The *Modality Matcher* module utilizes the DTW algorithm to compare and synchronize temporal variations across different modalities. DTW accommodates nonlinear temporal distortions, facilitating optimal alignment and similarity computation between time series characterized by varying lengths and sampling rates (Müller 2007; Hartmann and Link 2010). This capability is particularly suitable for addressing modality mismatches in HMD data resulting from heterogeneous sampling rates and data acquisition mechanisms. Specifically, HMD data are recorded as video, while eye and head movement data are collected using sensors, leading to discrepancies in sampling rates and data lengths. *Modality Matcher* compensates for these offsets using DTW to provide a quantitative measure of pattern similarity.

Fig. 3 illustrates the conceptual operation of the DTW algorithm. Conventional metrics for assessing time series relationships include Euclidean distance via direct alignment (Fig. 3-(a)), and correlation or cosine similarity for summarizing global dependencies. However, these methods typically fail to account for temporal distortions. In contrast, DTW (Fig. 3-(b)) dynamically adjusts the temporal axis to derive an optimal alignment path that minimizes pattern differences between the two sequences. DTW computes the similarity between two time series sequences, $A = \{a_1, \ldots, a_i, \ldots, a_n\}$, and $B = \{b_1, \ldots, b_j, \ldots, b_m\}$, using a cumulative distance matrix D. Here, i and j denote the indices of sequences A and B with lengths n and m, respectively. $D(i, j)$ represents the cumulative minimum distance up to point (i, j) and is recursively defined as follows:

$$D(i,j) = d(a_i, b_j) + \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\} \tag{5}$$

, where $d(a_i, b_j)$ denotes the Euclidean distance between points a_i and b_j. The matrix D is initialized as follows:

$$D(1,1) = d(a_1, b_1) \tag{6}$$

$$D(i,1) = \sum_{k=1}^{i} d(a_k, b_1) \quad \text{for} \quad i = 2, \ldots, n \tag{7}$$

$$D(1,j) = \sum_{k=1}^{j} d(a_1, b_k) \quad \text{for} \quad j = 2, \ldots, m \tag{8}$$

, where k is the index used in the summation for computing the cumulative distance.

Finally, the similarity between the two time series sequences is derived by backtracking the shortest path from the final value of the matrix $D(n, m)$. This enables the quantitative evaluation of nonlinear alignments. Since the DTW algorithm is inherently bivariate, we selected one representative feature from each of the three modalities (video, eye and head movements) to compute DTW distances for all modality pair combinations: [eye-head], [eye-video], and [head-video] (Fig. 1-(b)). To ensure that the LLM effectively distinguishes between disparate signal sources, we excluded feature combinations within the same modality from the DTW computation. As a result, the DTW algorithm generates a single distance value for each pair, where a larger feature distance signifies a greater degree of pattern mismatch between the respective modalities. This approach provides a quantifiable metric of complex cross-modal interactions, which facilitates the LLM's capacity to interpret HMD data and its causal relationship with VRS.

3.4 Context Matcher

The *Context Matcher* module utilizes a structured prompt architecture to facilitate the contextual interpretation of time series input data by the LLM (Fig. 1-(a)). It consists of three types of information—Basic Information (BI), Scene Information (SI), and Feature Information (FI)—which provide domain-level context, content-level variation, and feature-level dynamics, respectively (see Appendix B for details).

Basic Information (BI) provides domain information (e.g., experimental configuration, VRS measurement protocols) to establish the broader context of the prediction task. While time series data generally follow a standardized format (i.e., sensor signals recorded sequentially in temporal order), interpretive criteria and contextual relevance vary significantly across domains in which the data were collected (Wang et al. 2024). By specifying experimental parameters—including participant count, content variety, input modalities, and target metrics—the BI module enables the LLM to align sensor values with the specific experimental environment rather than relying solely on its generalized prior knowledge.

Scene Information (SI) delivers thematic metadata for the specific content used within the experimental domain, allowing the LLM to account for semantic differences during inference. Even within a unified domain, user responses and sensor signals can fluctuate according to the topic or motion characteristics of the stimulus (Guna et al. 2019). SI provides the thematic topic of each content instance along with a corresponding label (smooth or dynamic) that describes its motion profiles. In scenarios lacking thematic variation or employing a single content type, this component may be simplified or omitted.

Feature Information (FI) ensures that the LLMs accurately maps input values to their respective modality pairs. In this study, we provided DTW-based distance values derived from three modalities (video, eye and head movements) as primary inputs. To ensure that the LLM focuses on relational inference rather than redundant computation, we specify only the feature labels and their pairings, explicitly indicating that these values represent DTW distances rather than raw data. FI enables the LLM to evaluate the relative magnitudes of input values and perform precise inference by considering the quantitative interactions between modalities.

In summary, *Context Matcher* ensures that numerical inputs are interpreted within the specific experimental and structural framework of the data. By providing these multifaceted perspectives, BI, SI, and FI facilitate a consistent and semantically grounded interpretation of sensor signals, an approach extensible to various time series prediction tasks.

4 Results

To evaluate the performance of *VRS-LLM*, we conducted experiments from four perspectives: (1) a comparative analysis against conventional time series forecasting and existing VRS prediction models, including an evaluation of personalized prediction stability across varying few-shot configurations, (2) validation of DTW-based inter-modality representation in *Modality Matcher*, and (3) an ablation study on the contextual layers of *Context Matcher*. To enable the LLM's processing of HMD-based input data, we structured the feature distances extracted by *Modality Matcher* on a batch basis and converted them into a JSON format to optimize parallel processing. The final dataset consists of 540 data points, covering full combinations of 27 participants and 20 videos. Each data point contains two types of input attributes (see Appendix C for details). The first type is metadata attributes, including the data point index (0-539), participant number, and video number, which uniquely identify each data point. The second type is feature distance attributes, including feature distance values derived from three modality pairings: [eye-head], [eye-video], and [head-video].

We used three LLM families—GPT (Achiam et al. 2023; Singh et al. 2025), LLaMA (Grattafiori et al. 2024), and Gemini (Comanici et al. 2025)—as the backbone of *VRS-LLM*, evaluating a total of eight LLMs: GPT-3.5 Turbo, GPT-4o mini, GPT-5 mini, LLaMA 3.1 8B Instruct, LLaMA 3.1 70B Instruct, Gemini 2.0 Flash-Lite, Gemini 2.5 Flash, and Gemini 2.5 Pro. Since the normalized FMS is a continuous variable, we used mean squared error (MSE) and mean absolute error (MAE) for evaluation:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2, \quad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \tag{9}$$

n denotes the number of data points, and i is the index of a data point. y_i is the ground truth value and $\hat{y_i}$ is the predicted value for the i-th data point.

Experiments were executed in zero- and few-shot (1 to 10-shot) settings. We compared generalized prediction—where few-shot examples were sampled from the global dataset—against our personalized prediction approach, which used examples exclusively from the specific user being evaluated. This comparison was designed to isolate the performance gains attributable to individual-level conditioning. To ensure empirical reliability, all experiments were repeated five times, and the reported results represent the mean performance across trials.

Table 2 Performance comparison of *VRS-LLM* with baseline models. The baselines include two categories: time series forecasting models and existing VRS prediction models. VRS-LLM is evaluated under generalized prediction and personalized prediction. The best results are highlighted in **bold**.

Models (Year)	**Backbone**	**MSE**	**MAE**
Time Series Forecasting Models			
PatchTST (Nie 2022)	Transformer-encoder	0.31	0.25
DLinear (Zeng et al. 2023)	Linear	0.26	0.22
Time-LLM (Jin et al. 2023)	GPT-2	0.87	0.63
TimeGPT (Garza et al. 2023)	Transformer	0.19	0.36
TimesFM (Das et al. 2024)	Transformer-decoder	0.37	0.51
Chronos (Ansari et al. 2024)	T5 (Raffel et al. 2020)	0.17	0.34
Existing VRS Prediction Models			
Islam et al. (Islam et al. 2021)	LSTM	0.10	0.27
Islam et al. (Islam et al. 2021)	CNN-LSTM	0.10	0.26
MAC (Jeong et al. 2023)	CNN-LSTM	1.34	0.22
MS-STTN (Jeong and Han 2024)	Transformer	1.22	0.21
VRS-LLM			
Generalized Prediction	GPT-3.5 Turbo	0.11	0.28
	GPT-4o mini	0.10	0.26
	GPT-5 mini	0.12	0.27
	LLaMA 3.1 8B Instruct	0.12	0.30
	LLaMA 3.1 70B Instruct	0.10	0.27
	Gemini 2.0 Flash-Lite	0.14	0.31
	Gemini 2.5 Flash	0.18	0.33
	Gemini 2.5 Pro	0.13	0.28
Personalized Prediction (ours)	GPT-3.5 Turbo	0.10	0.17
	GPT-4o mini	**0.06**	0.14
	GPT-5 mini	0.07	0.14
	LLaMA 3.1 8B Instruct	0.10	0.19
	LLaMA 3.1 70B Instruct	**0.06**	0.14
	Gemini 2.0 Flash-Lite	0.18	0.35
	Gemini 2.5 Flash	0.19	0.34
	Gemini 2.5 Pro	**0.06**	**0.13**

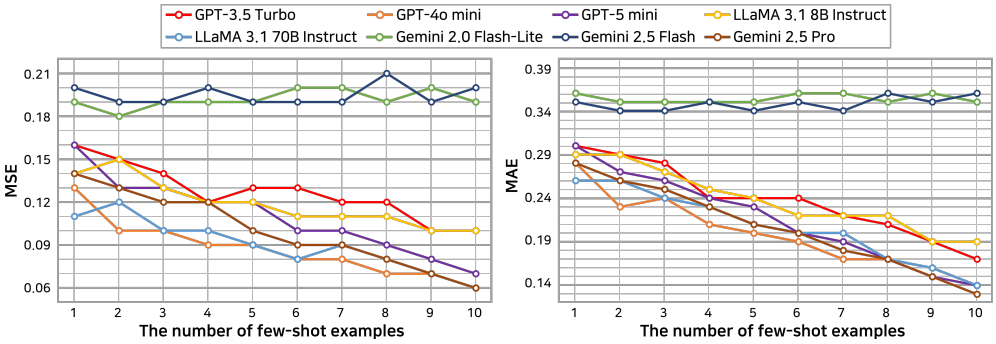


Fig. 4 Performance trends as the number of few-shot examples increases. Each graph reports MSE and MAE, showing how performance varies from 1-shot to 10-shot across LLM backbones.

4.1 Evaluation of VRS-LLM Performance

To evaluate the effectiveness of the *VRS-LLM* architecture, we compared it with baselines from two perspectives (Table 2). First, we chose time series forecasting models designed for general-purpose time series domain (e.g., weather, electricity), representing standard temporal modeling approaches. Second, we included deep learning models specifically engineered for the VRS domain. Since all baselines utilize raw time series data, we used optical flow and magnitude-based HMD data before applying *Modality Matcher*. Unlike the baselines, *VRS-LLM* relies on the reasoning capability of pre-trained LLMs without a training process. To ensure a fair comparison with baselines whose performance is optimized through training, we report the best performance for each LLM under both generalized and personalized settings, independent of the specific few-shot count.

We first assessed *VRS-LLM* within a generalized prediction framework. In this setting, *VRS-LLM* significantly outperformed the time series forecasting models, indicating that our proposed architecture more effectively captured the nuances of HMD data. In addition, *VRS-LLM* in generalized prediction achieved performance comparable to existing VRS prediction models across all tested LLMs, demonstrating the feasibility of using LLMs for VRS prediction.

We then evaluated the performance in a personalized prediction setting. It achieved the best performance on Gemini 2.5 Pro (MSE of 0.06, MAE of 0.13), showing significant improvements over both the baselines and generalized *VRS-LLM* configuration.

Building on the effectiveness of personalized prediction, we further examined the relationship between predictive accuracy and the volume of few-shot examples, incrementally increasing the count from 1 to 10 shots (Fig. 4). As a result, most LLMs exhibited a general downward trend in both MSE and MAE as the number of shots increased, typically reaching optimal performance at the 10-shot mark. This suggests that additional few-shot examples contribute to capturing user-specific data patterns and distributions, enabling more accurate personalized prediction without additional model retraining.

Notably, smaller models within the Gemini family (i.e., Gemini 2.0 Flash-Lite and Gemini 2.5 Flash) exhibited divergent trends. These models performed poorly in personalized settings relative to their generalized counterparts and failed to show significant marginal gains beyond the 2-shot threshold. These findings indicate that in the VRS domain, performance is highly sensitive to model scale and architecture, underscoring the necessity of selecting a robust backbone LLM—contingent upon specific domain characteristics and input data properties—and determining an optimized few-shots to ensure predictive stability.

Table 3 Performance comparison of VRS-LLM using Modality Matcher based on different metrics to represent relationships between time series. DTW is used in our study, while Euclidean distance, correlation, and cosine similarity are considered as baselines. The best results are highlighted in **bold**.

Method	DTW (ours)		Euclidean Distance		Correlation		Cosine Similarity	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
GPT-3.5 Turbo	0.10	0.17	0.11	0.18	0.10	0.18	0.11	0.19
GPT-4o mini	**0.06**	0.14	**0.06**	0.14	0.08	0.16	0.07	0.15
GPT-5 mini	0.07	0.14	0.07	0.14	0.07	0.15	0.07	0.15
LLaMA 3.1 8B Instruct	0.10	0.19	0.17	0.33	0.16	0.32	0.16	0.32
LLaMA 3.1 70B Instruct	**0.06**	0.14	0.07	0.17	0.07	0.16	0.07	0.16
Gemini 2.0 Flash-Lite	0.19	0.35	0.20	0.35	0.19	0.35	0.19	0.35
Gemini 2.5 Flash	0.20	0.36	0.21	0.36	0.20	0.35	0.21	0.36
Gemini 2.5 Pro	**0.06**	**0.13**	0.07	0.14	0.07	0.15	**0.06**	0.14

4.2 Effectiveness of DTW in the Modality Matcher

The *Modality Matcher* module was developed to use DTW as an input representation for time series data within the LLM framework. To evaluate the effectiveness of DTW, we compared the performance of *VRS-LLM* against three representative similarity metrics: Euclidean distance, correlation, and cosine similarity (Table 3). These metrics were applied to the same feature pairs as those used in DTW computations. To isolate the contribution of the DTW-based representation, all experiments were conducted in a 10-shot setting. In addition, within the FI component of the *Context Matcher*, the term "DTW" was replaced with "Euclidean distance", "correlation", or "cosine similarity," depending on the similarity metric used in each experiment.

Experimental results showed that the DTW-based *VRS-LLM* achieved higher or at least comparable performance across all evaluated LLMs compared to the other metrics. Euclidean distance incorporates point-wise relationships at the same time steps, while correlation and cosine similarity consider only global temporal relationships

of the time series. In contrast, DTW facilitates temporal warping to accommodate inter-modal asynchrony. This capability allows for a more precise characterization of cross-modal correspondences, enhancing the inference capabilities of *VRS-LLM* in the presence of nonlinear temporal distortions.

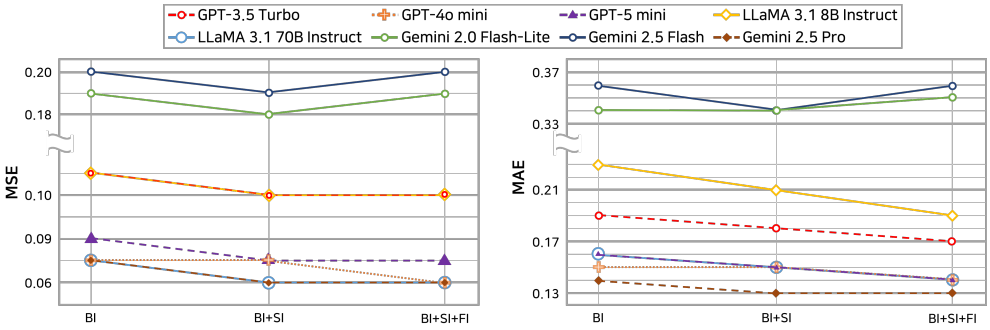


Fig. 5 Impact of additional information types on performance. Each graph reports MSE and MAE, illustrating how performance changes as BI, SI, and FI are sequentially added to the Context Matcher across LLM backbones. Different marker and line styles are used where the performance values of some LLMs overlap.

4.3 Validity Evaluation of Context Matcher Design

To examine the impact of the three information types in the *Context Matcher* (BI, SI, and FI) on VRS prediction, we conducted a stepwise ablation analysis across three settings: (1) BI only, (2) BI+SI, and (3) BI+SI+FI (*VRS-LLM*), as summarized in Fig. 5. BI, which established the fundamental task context, served as the baseline for all settings. SI and FI were incrementally integrated to align the prompt with the content-level characteristics and feature-level structure of the input attributes. SI incorporates topic-related metadata for each VR video, while FI details the inter-modality feature distance relationships. This experimental design facilitates an assessment of the isolated effect of each information type on prediction performance. To ensure consistency, all evaluations were conducted in a 10-shot configuration.

Experimental results showed that, except for the smaller models in the Gemini family, *VRS-LLM* exhibited a consistent downward trend in both MSE and MAE as information types were incrementally added. The inclusion of SI yielded clear performance gains, while the addition of FI either further improved or maintained performance. These findings suggest that the majority of the evaluated LLM families effectively utilized SI and FI as meaningful contextual anchors to interpret DTW-based inputs for VRS prediction. Conversely, the smaller Gemini-family models exhibited divergent behavior, demonstrating negligible performance improvements despite the sequential addition of BI, SI, and FI. While this variability across model scales, our findings confirm that the *Context Matcher* architecture helps achieve competitive performance, underscoring its potential utility in future VRS prediction research using LLMs.

5 Discussion

5.1 Considerations for Efficient VRS-LLM

In this study, we used LLMs to incorporate contextual information from HMD data. While *Modality Matcher* condenses raw data into feature distances, providing numerical representations to an LLM increases token consumption, incurring higher financial costs and computational overhead. Given that effective VRS mitigation relies on early prediction, maintaining low inference latency is critical. Consequently, an efficient implementation strategy should balance the cost-performance trade-off through two architectural considerations.

First, the adoption of small LLMs (SLMs) offers a viable path for resource-constrained environments. As shown in Table 2, SLMs exhibited only marginal differences in performance compared to their larger counterparts. Notably, GPT-4o mini, which achieved the best performance among the SLMs, showed an MSE identical to that of the best-performing model, Gemini 2.5 Pro, with an MAE difference of only 0.01. Since API costs and latency typically scale with parameter count, SLMs represent a practical solution for stable performance. Furthermore, SLMs can be deployed in on-premises or "in-house" environments, eliminating the dependency on external API calls and ensuring data privacy. However, our results also revealed that smaller models within the Gemini family yielded significantly higher error rates, suggesting that model efficacy is highly sensitive to the specific application domain. This implies that the selection of a backbone model should be guided by rigorous empirical evaluation within the target domain, rather than relying exclusively on general-purpose benchmarks or parameter count.

Second, determining an optimal few-shot configuration is a critical consideration for efficient implementation. As illustrated in Figure 5, the performance of *VRS-LLM* generally improved as the number of samples increased across the majority of evaluated LLMs. However, the inclusion of additional few-shot examples scales the input token count linearly, which exacerbates financial costs and may introduce prohibitive inference latency in real-time applications. Consequently, the few-shot setting should be strategically selected by balancing target performance against practical architectural constraints, rather than simply maximizing the shot count. For example, researchers might establish a shot threshold that satisfies a predefined accuracy requirement or identifies the maximum performance achievable within a specific resource.

5.2 Suitability of Applying DTW to HMD Data

In Section 4.2, we demonstrated the effectiveness of applying DTW to *Modality Matcher*. While we observed that other similarity metrics (i.e., Euclidean distance, correlation, cosine similarity) achieved comparable performance, it would be a mistake to conclude that the choice of metric is negligible. Despite the competitive baseline, DTW consistently yielded the superior performance across the majority of evaluated LLMs. This is because DTW specifically addresses the structural discrepancies inherent in HMD data, such as heterogeneous sampling rates and latent offsets, which simpler metrics tend to overlook.

Furthermore, while our experimental dataset was captured in a controlled environment, real-world deployment often involve severe temporal asynchrony and signal noise due to hardware jitter or background system processes. In these unconstrained settings, metrics that fail to account for nonlinear temporal alignment struggle to capture true inter-modal correspondences. Therefore, we determined that DTW is a more robust and resilient choice for multimodal time series representation, ensuring stability against the uncertainties of real-world data. Beyond the VRS domain, we anticipate that DTW can be effectively applied to other user-state prediction tasks that rely on heterogeneous sensors, effectively mitigating the challenges of asynchronous and nonlinear temporal alignment between disparate modalities.

5.3 Limitations and Future Work

Despite the effectiveness of *VRS-LLM* in leveraging structured multimodal context, two limitations warrant further investigation.

First, *VRS-LLM* relies on static, manually engineered prompts. While *Context Matcher* facilitates effective data interpretation, this prompt-centric architecture may constrain the model's ability to capture fine-grained, latent variations inherent in the raw data. To incorporate more objective and precise contextual information, future iterations should aim to reduce this dependency. For example, the architecture could be extended by utilizing SLMs for cost efficiency and applying parameter-efficient fine-tuning (PEFT) to specialize the model on HMD-specific datasets. Such an approach would enable more autonomous, domain-adaptive reasoning and diminish the necessity for extensive manual prompt engineering.

Second, although *VRS-LLM* showed superior predictive performance, its practical utility in real-time environments has not been validated. Since this study focused on demonstrating the potential of LLM-based user state prediction, future work should further investigate whether *VRS-LLM* can operate effectively beyond controlled experimental settings, such as in real usage scenarios and across more diverse user groups. For example, the practical applicability of *VRS-LLM* could be examined through user studies in which the model is embedded in VR content. Furthermore, real-world application settings may involve greater heterogeneity in user behavior and more variable data quality. In contrast, this study applied DTW to time series derived from the publicly available dataset, which may not sufficiently capture the less stable data characteristics that could arise in real-world application settings. Therefore, further comprehensive analyses are needed to examine whether DTW continues to maintain performance under these conditions.

6 Conclusion

The primary objective of this study was to evaluate the effectiveness of LLMs in predicting user state and to establish robust strategies for data preprocessing and prompt engineering. We introduced *VRS-LLM*, a novel framework that leverages DTW and LLMs to predict VRS using HMD-based multimodal time series data (video, eye movement, and head movement). By implementing *Modality Matcher* to effectively capture inter-modality relationships and *Context Matcher* to provide three types of semantic

grounding (BI, SI, and FI), we enabled the LLM to interpret numerical time series data within a structured experimental context. Our results demonstrate that *VRS-LLM* achieved superior performance—yielding an MSE of 0.06 and an MAE of 0.13 in the 10-shot setting—outperforming time series forecasting models and specialized VRS prediction baselines. Moreover, our ablation studies validated the effectiveness of DTW-based representations and the incremental benefits of contextual prompts. These findings offer significant insights into the integration of time series data with LLMs, underscoring the necessity of selecting an appropriate backbone architecture and optimizing few-shot samples for operational efficiency. In summary, this work offers practical guidance for researchers seeking to utilize the reasoning capabilities of LLMs for high-fidelity user-state prediction in multimodal environments.

Declarations

- **Funding**: This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University)), and the Korea Institute of Energy Technology Evaluation and Planning(KETEP) and the Ministry of Climate, Energy & Environment(MCEE) of the Republic of Korea (No. RS-2026-25530097).

- **Competing interests**: The authors declare that they have no competing interests.
- **Ethics approval and consent to participate**: Not applicable. This study was conducted using a publicly available, anonymized dataset.
- **Data availability**: The dataset analyzed during the current study is publicly available as described in Jeong et al. (2023).
- **Code availability**: The code generated during the current study is available from the corresponding author on reasonable request.
- **Author contributions**: All authors contributed to conceptualization, methodology development, project administration, data curation, formal analysis, investigation, validation, visualization, original draft preparation, and manuscript review and editing. K. H. acquired the funding.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

Ansari, A.F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S.S., Arango, S.P., Kapoor, S., et al.: Chronos: Learning the language of time series. arXiv preprint arXiv:2403.07815 (2024)

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)

Chen, W., Cheng, J., Wang, L., Zhao, W., Matusik, W.: Sensor2text: Enabling natural language interactions for daily activity tracking using wearable sensors. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **8**(4), 1–26 (2024)

Caserman, P., Garcia-Agundez, A., Gámez Zerban, A., Göbel, S.: Cybersickness in current-generation virtual reality head-mounted displays: systematic review and outlook. Virtual Reality **25**(4), 1153–1170 (2021)

Chang, E., Kim, H.T., Yoo, B.: Predicting cybersickness based on user's gaze behaviors in hmd-based virtual reality. Journal of Computational Design and Engineering **8**(2), 728–739 (2021)

Das, A., Kong, W., Sen, R., Zhou, Y.: A decoder-only foundation model for time-series forecasting. In: Forty-first International Conference on Machine Learning (2024)

Dong, Y., Sun, Z., Jia, H.: A cosine similarity-based negative selection algorithm for time series novelty detection. Mechanical Systems and Signal Processing **20**(6), 1461–1472 (2006)

EDITION, F.: Diagnostic and statistical manual of mental disorders. American psychiatric association, Washington, DC, 205–224 (1980)

Englhardt, Z., Ma, C., Morris, M.E., Chang, C.-C., Xu, X.O., Qin, L., McDuff, D., Liu, X., Patel, S., Iyer, V.: From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **8**(2), 1–25 (2024)

Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13, pp. 363–370 (2003). Springer

Garza, A., Challu, C., Mergenthaler-Canseco, M.: Timegpt-1. arXiv preprint arXiv:2310.03589 (2023)

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)

Guna, J., Geršak, G., Humar, I., Song, J., Drnovšek, J., Pogačnik, M.: Influence of video content type on users' virtual reality sickness perception and physiological response. Future Generation Computer Systems **91**, 263–276 (2019)

Ge, W., Mou, G., Agu, E.O., Lee, K.: Deep heterogeneous contrastive hyper-graph learning for in-the-wild context-aware human activity recognition. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **7**(4), 1–23

(2024)

Gao, C., Marsic, I., Sarcevic, A., Gestrich-Thompson, W., Burd, R.S.: Real-time context-aware multimodal network for activity and activity-stage recognition from team communication in dynamic clinical settings. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies **7**(1), 1–28 (2023)

Hartmann, B., Link, N.: Gesture recognition with inertial sensors and optimized dtw prototypes. In: 2010 IEEE International Conference on Systems, Man and Cybernetics, pp. 2102–2109 (2010). IEEE

Islam, R., Ang, S., Quarles, J.: 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), pp. 148–155 (2021). IEEE

Islam, R., Desai, K., Quarles, J.: 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), pp. 31–40 (2021). IEEE

Islam, R., Desai, K., Quarles, J.: 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), pp. 121–130 (2022). IEEE

Jin, W., Fan, J., Gromala, D., Pasquier, P.: Automatic prediction of cybersickness for virtual reality games. In: 2018 IEEE Games, Entertainment, Media Conference (GEM), pp. 1–9 (2018). IEEE

Jeong, D., Han, K.: Precyse: Predicting cybersickness using transformer for multimodal time-series sensor data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **8**(2), 1–24 (2024)

Jung, S., Li, C., McKee, R., Whitton, M.C., Lindeman, R.W.: Floor-vibration vr: mitigating cybersickness using whole-body tactile stimuli in highly realistic vehicle driving experiences. IEEE Transactions on Visualization and Computer Graphics **27**(5), 2669–2680 (2021)

Jeong, D., Paik, S., Noh, Y., Han, K.: Mac: multimodal, attention-based cybersickness prediction modeling in virtual reality. Virtual Reality **27**(3), 2315–2330 (2023)

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J.Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., et al.: Time-llm: Time series forecasting by reprogramming large language models. arXiv preprint arXiv:2310.01728 (2023)

Jin, Z., Yang, Y., Liu, Y.: Stock closing price prediction based on sentiment analysis

and lstm. Neural Computing and Applications **32**, 9713–9729 (2020)

Kalpakis, K., Gada, D., Puttagunta, V.: Distance measures for effective clustering of arima time-series. In: Proceedings 2001 IEEE International Conference on Data Mining, pp. 273–280 (2001). IEEE

Kim, B., Jeong, D., Choi, Y., Choi, Y., Kim, H., Han, K.: Enhancing mindfulness-based cognitive therapy in a virtual reality: A prospective interventional study. Scientific Reports **15**(1), 29534 (2025)

Kim, B., Jeong, D., Choi, Y., Kim, H., Han, K.: VR-ACT: transforming clinical evidence of acceptance and commitment therapy through virtual reality for subthreshold depression. Virtual Reality (2025)

Kim, B., Jeong, D., Jeong, M., Noh, T., Kim, S.-I., Kim, T., Jang, S.-Y., Yoo, H.J., Kim, J.G., Hong, H., *et al.*: VISTA: User-centered vr training system for effectively deriving characteristics of people with autism spectrum disorder. In: Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology, pp. 1–12 (2022)

Kim, B., Jeong, D., Kim, J.G., Hong, H., Han, K.: V-DAT (virtual reality data analysis tool): Supporting self-awareness for autistic people from multimodal vr sensor data. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pp. 1–13 (2023)

Kim, J., Oh, H., Kim, W., Choi, S., Son, W., Lee, S.: A deep motion sickness predictor induced by visual stimuli in virtual reality. IEEE Transactions on Neural Networks and Learning Systems **33**(2), 554–566 (2020)

Kamińska, D., Zwoliński, G., Laska-Leśniewicz, A.: Usability testing of virtual reality applications—the pilot study. Sensors **22**(4), 1342 (2022)

Lee, S., Kim, S., Kim, H.G., Ro, Y.M.: Assessing individual vr sickness through deep feature fusion of vr video and physiological response. IEEE Transactions on Circuits and Systems for Video Technology **32**(5), 2895–2907 (2021)

Liao, C.-Y., Tai, S.-K., Chen, R.-C., Hendry, H.: Using eeg and deep learning to predict motion sickness under wearing a virtual reality device. Ieee Access **8**, 126784–126796 (2020)

Li, R., Wang, Y., Yin, H., Chardonnet, J.-R., Hui, P.: A deep cybersickness predictor through kinematic data with encoded physiological representation. In: 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 1132–1141 (2023). IEEE

Liu, C., Xu, Q., Miao, H., Yang, S., Zhang, L., Long, C., Li, Z., Zhao, R.: Timecma: Towards llm-empowered time series forecasting via cross-modality alignment. arXiv

preprint arXiv:2406.01638 (2024)

Liu, H., Xu, S., Zhao, Z., Kong, L., Prabhakar Kamarthi, H., Sasanur, A., Sharma, M., Cui, J., Wen, Q., Zhang, C., *et al.*: Time-mmd: Multi-domain multimodal dataset for time series analysis. Advances in Neural Information Processing Systems **37**, 77888–77933 (2024)

Masnad, M., MukitHasan, G., Iftekhar, K.M., Rahman, M.S.: Human activity recognition using dtw algorithm. In: 2019 IEEE Region 10 Symposium (TENSYMP), pp. 39–43 (2019). IEEE

Müller, M.: Dynamic time warping. Information retrieval for music and motion, 69–84 (2007)

Nie, Y.: A time series is worth 64words: Long-term forecasting with transformers. arXiv preprint arXiv:2211.14730 (2022)

Oloyede, M.O., Hancke, G.P.: Unimodal and multimodal biometric sensing systems: a review. IEEE access **4**, 7532–7555 (2016)

Palmisano, S., Allison, R.S., Kim, J.: Cybersickness in head-mounted displays is caused by differences in the user's virtual and physical head pose. Frontiers in Virtual Reality **1**, 587698 (2020)

Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information fusion **37**, 98–125 (2017)

Qu, C., Che, X., Ma, S., Zhu, S.: Bio-physiological-signals-based vr cybersickness detection. CCF Transactions on Pervasive Computing and Interaction **4**(3), 268–284 (2022)

Ramaseri-Chandra, A.N., Reza, H.: Predicting cybersickness using machine learning and demographic data in virtual reality. Electronics **13**(7), 1313 (2024)

Rebenitsch, L., Owen, C.: Estimating cybersickness from virtual reality applications. Virtual Reality **25**(1), 165–174 (2021)

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)

Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)

Tasnim, U., Islam, R., Desai, K., Quarles, J.: Investigating personalization techniques for improved cybersickness prediction in virtual reality environments. IEEE Transactions on Visualization and Computer Graphics (2024)

Tian, N., Lopes, P., Boulic, R.: A review of cybersickness in head-mounted displays: raising attention to individual susceptibility. Virtual Reality **26**(4), 1409–1441 (2022)

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

Wang, J., Bai, G., Cheng, W., Chen, Z., Zhao, L., Chen, H.: Pond: Multi-source time series domain adaptation with information-aware prompt tuning. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3140–3151 (2024)

Wang, Y., Chardonnet, J.-R., Merienne, F., Ovtcharova, J.: Using fuzzy logic to involve individual differences for predicting cybersickness during vr navigation. In: 2021 IEEE Virtual Reality and 3D User Interfaces (VR), pp. 373–381 (2021). IEEE

Wu, F., Suma Rosenberg, E.: Adaptive field-of-view restriction: Limiting optical flow to mitigate cybersickness in virtual reality. In: Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology, pp. 1–11 (2022)

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

Xia, H., Ao, H., Li, L., Liu, Y., Liu, S., Ye, G., Chai, H.: Ci-sthpan: Pre-trained attention network for stock selection with channel-independent spatio-temporal hypergraph. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 9187–9195 (2024)

Xu, B., Bikakis, A., Onah, D., Vlachidis, A., Dickens, L.: Context helps: Integrating context information with videos in a graph-based har framework. In: International Conference on Neural-Symbolic Learning and Reasoning, pp. 3–28 (2024). Springer

Xue, H., Salim, F.D.: Promptcast: A new prompt-based learning paradigm for time series forecasting. IEEE Transactions on Knowledge and Data Engineering (2023)

Yang, A.H.X., Galán-Augé, C., Kasabov, N.K., Cakmak, Y.O.: Machine learning-guided high-definition transcranial direct current stimulation prevents cybersickness. Virtual reality **29**(3), 94 (2025)

Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? In: Proceedings of the AAAI Conference on Artificial Intelligence, vol.

37, pp. 11121–11128 (2023)

Zhou, T., Mei, Z., Zhu, X., Huang, Z.: Synchrony detection of epileptic eeg signals based on attention and pearson's correlation coefficient. In: 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 531–535 (2020). IEEE

Zhou, T., Niu, P., Sun, L., Jin, R.: One fits all: Power general time series analysis by pretrained lm. Advances in neural information processing systems **36**, 43322–43355 (2023)

A Individual Distributions of FMS Scores



Fig. 6 Distributions of FMS scores for all 27 participants. The x-axis indicates the FMS scale scores (0-20), and the y-axis indicates the response frequency for each score. Descriptive statistics (mean, standard deviation (SD), and skewness) are shown in each subplot.

B Details of Context Matcher

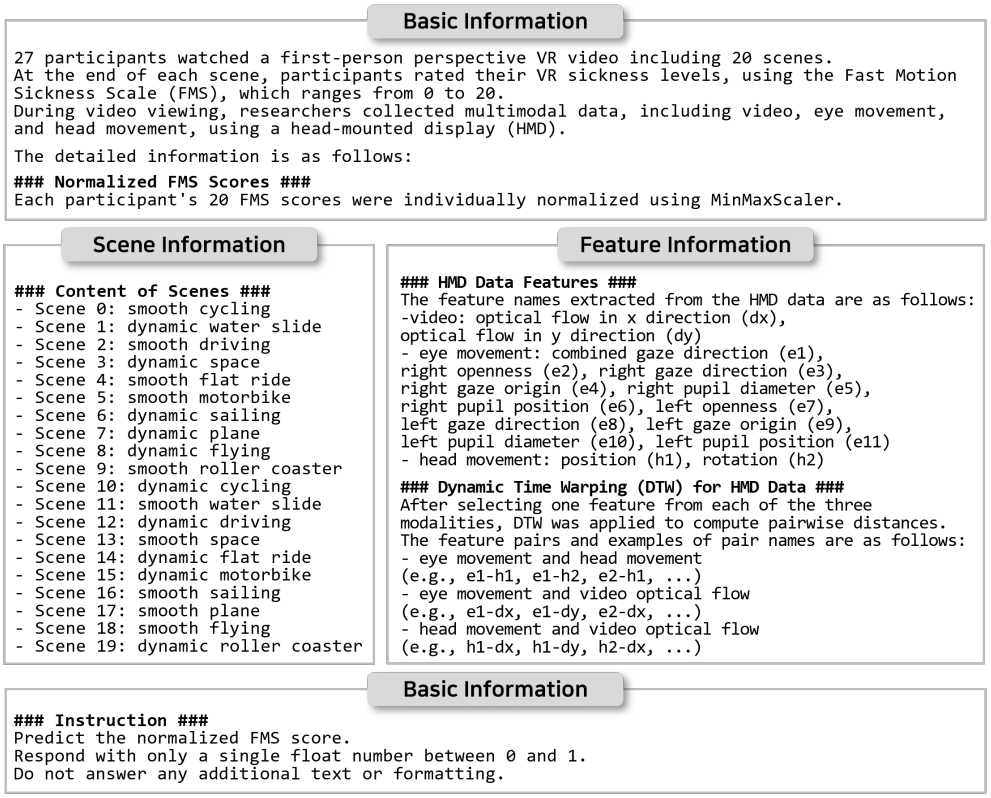


Fig. 7 Full information of Context Matcher.

C Input Attributes for LLMs

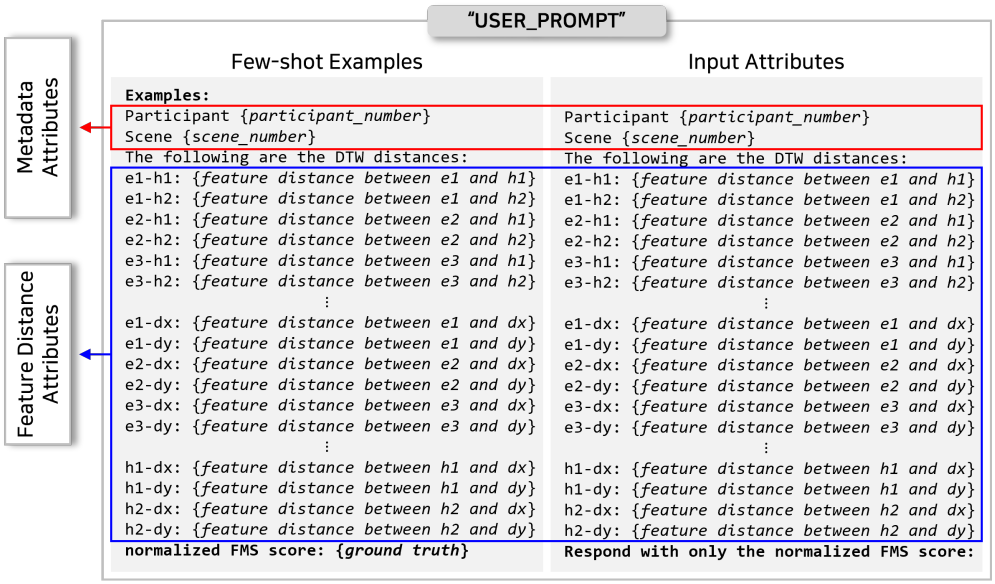


Fig. 8 Details of few-shot examples and input attributes included in "USER_PROMPT". 'ei' and 'hj' denote feature aliases defined in Context Matcher for magnitude-based eye and head movement data, respectively ($i = 1, 2, ..., 11, j = 1, 2$). 'dx' and 'dy' denote optical flow along the $x-$ and $y-$ axes, respectively.