CyberGATE: Multiscale State-Space Fusion for Ocular and Vestibular Cybersickness Forecasting

Jianing Zhang[1,*]
[1]Brunel London School, North China University of Technology, Beijing, China
Corresponding author: Jianing Zhang (email: 2379536@brunel.ac.uk).

ABSTRACT Cybersickness is still a major usability and safety issue for virtual reality because discomfort may emerge gradually during exposure or appear suddenly when scene motion changes. Questionnaire-based evaluation is difficult to use as a continuous monitoring tool. This study presents CyberGATE, a multimodal time-series framework that estimates cybersickness from eye-tracking and head-motion signals collected by an HMD. The framework separates long-context representation and short-event modeling into two coordinated branches. The global branch uses multi-view embedding, Mamba sequence modeling, and KAN-based sparse experts to learn persistent temporal patterns, whereas the local branch applies decomposition-aware patch encoding and window attention to detect transient disturbances. Experiments on two public datasets show that CyberGATE reduces prediction error compared with recurrent, Transformer, and existing multimodal baselines. Component removal and sensitivity analyses further confirm that global dynamics, local responses, and adaptive routing all improve performance. These results indicate that ordinary HMD behavioral signals can support objective, real-time, and individualized cybersickness warning in practical VR applications.

INDEX TERMS Cybersickness, Virtual Reality, Multimodal Sequence Modeling, Mamba, Local Attention, Gated Fusion

I. INTRODUCTION

Cybersickness remains one of the key barriers to comfortable immersive interaction. With VR systems moving into entertainment, training, education, medical rehabilitation, and industrial simulation, the system must be able to recognize discomfort early and respond before the user experience deteriorates. Conventional assessment relies heavily on post-exposure questionnaires such as the SSQ [1]. These scales are valuable for clinical and experimental reporting, but they are delayed, subjective, and unsuitable for online monitoring in dynamic virtual scenes.

Eye trackers and inertial sensors embedded in modern head-mounted displays provide a more direct source of behavioral evidence. Gaze behavior and head movement describe different parts of the ocular-vestibular response. Existing methods often concatenate them immediately or handle them as interchangeable feature channels [2]. Such processing may blur useful modality differences. Rapid gaze events, fixation instability, and saccade changes usually reflect near-instant visual discomfort, while head kinematics contain broader information about navigation, posture, and motion inconsistency.

A second problem is the limited temporal flexibility of many prediction models [3]. A single encoder, simple pooling, or fixed fusion can represent the overall sequence trend, but may be insensitive to abrupt local fluctuations. In adaptive VR, symptoms can change within a short time window, so a useful model should represent both slowly accumulated exposure effects and brief signal disturbances.

Recent studies have introduced behavioral and physiological signals to reduce dependence on subjective scales. Eye movement, head trajectory, EEG, ECG, EDA, and other measurements have been used to infer symptom severity, while deep fusion and Transformer-based models have improved time-series representation [2,4-8]. Even so, many methods still emphasize either whole-sequence aggregation or local variation alone. In addition, their fusion weights are commonly fixed after training. This makes it hard to distinguish gradual discomfort accumulation from sudden visual-vestibular conflict across different users, scenes, and interaction modes.

This work proposes CyberGATE to address these limitations. The model is an adaptive gated framework for multiscale cybersickness prediction. Its global branch combines multi-view embedding, Mamba state-space modeling [9], and KAN-based mixture-of-experts learning to capture long-range modality-aware patterns [10]. Its local branch uses decomposition-guided patch transformers and local attention to capture short-lived disturbances [11]. A routing module then assigns sample-dependent weights to the two branches, allowing the model to adjust the relative importance of global context and local dynamics.

The main contributions are as follows:

First, CyberGATE introduces a dual-branch temporal architecture that couples a Mamba-MoE global encoder with a local-attention patch branch for continuous cybersickness estimation.

Second, the model keeps eye and head information in complementary long-range and short-range pathways, so modality-specific cues and multiscale temporal evidence are retained before final fusion.

Third, comparative experiments, ablation tests, and parameter analyses on two public datasets verify the effectiveness of the proposed design.

II. RELATED WORK

Cybersickness refers to discomfort caused by immersive virtual environments. Nausea, dizziness, disorientation, postural instability, visual fatigue, and headache are commonly reported symptoms [12]. The mechanism is not fully settled. Sensory conflict theory explains the problem as inconsistency between visual and vestibular inputs [13], while postural instability theory highlights the difficulty of maintaining a stable body state during exposure [14]. Individual differences are also important. Age, gender, previous motion-sickness experience, and familiarity with VR can change susceptibility [15]. Among subjective assessment tools, the Simulator Sickness Questionnaire remains one of the most widely used instruments [1].

Hardware and content conditions further influence discomfort. Display latency, refresh rate, field of view, optical flow, camera movement, and locomotion strategy can all change sickness intensity. Early mitigation work emphasized lower latency and higher frame rate [16], and later studies compared navigation modes such as teleportation and continuous movement. Accurate symptom measurement is therefore the basis for risk analysis, model training, and real-time intervention.

Subjective scales were the earliest and most common measurement approach. The SSQ uses self-reported ratings to summarize discomfort [1], and the VIMSSQ extends this idea for visually induced motion sickness by combining related questionnaire tools [17]. These methods are convenient, but they depend on user memory and cooperation. They also cannot provide dense labels throughout a VR session.

Objective physiological sensing has been explored to overcome these limits. EGG, ECG, EEG, heart-rate variation, and blink-related features have been examined as discomfort indicators [4,18]. However, cybersickness is dynamic and user-dependent, so a single signal type often cannot describe the full process or generalize reliably across experimental conditions.

Recent work therefore turns toward multimodal temporal modeling, especially the combination of gaze and head-motion features. Prior physiological studies provided useful ideas for dynamic representation [19]. MS-STTN used Transformer modules to model spatiotemporal patterns from eye and head data [2,20], improving prediction accuracy but still relying on synchronized collection and relatively sparse labels. Adaptive cross-modal attention used in other domains also suggests that learnable modality weighting can be beneficial [5]. The joint modeling of long-context trends and short transient responses, however, remains insufficient.

More broadly, cybersickness prediction is moving toward HMD-compatible behavioral sensing. Some methods combine body movement with machine learning but do not explicitly represent long-range dynamics [3]. Others improve cross-scene robustness through self-supervised learning but do not fully exploit the temporal complementarity between gaze and head motion [21].

In summary, the research trajectory has shifted from questionnaire scoring to physiological measurement, multimodal learning, deep temporal modeling, and cross-user adaptation. The remaining gaps lie in early modality interaction, unified global-local modeling, and flexible fusion.

Real-time deployment studies further require a balance among accuracy, interpretability, computational cost, and early-warning ability [22-26]. These requirements motivate the lightweight but adaptive structure developed in this paper.

III. METHODOLOGY

CyberGATE predicts cybersickness by extracting complementary temporal information from synchronized eye and head signals. Figure 1 presents the overall structure. One branch focuses on long-range sequence context, and the other focuses on local short-duration responses. This design is based on the fact that cybersickness can be produced by both cumulative exposure and brief sensory disturbances [6,7]. The global branch learns modality-aware temporal dependencies through multi-view embedding, Mamba encoding, and KAN-based expert routing. The local branch processes decomposed patches to identify short-term variations. Finally, a fusion router combines the two representations for sample-specific prediction.

The input stage adopts a dual-path cross-attention embedding module to allow eye and head features to interact before deeper encoding. Each modality is first projected into a latent representation. The paired modality then provides complementary keys and values for cross-attention, so visual and vestibular information can be aligned at an early stage. Equation (1) defines the cross-modal attention computation used by the embedding module.

$$A = \mathrm{Softmax}\left(\frac{QK^T}{d_k}\right), \quad X_{\mathrm{cross}} = A \cdot V \tag{1}$$

Equation (2) gives the updated token representation after the

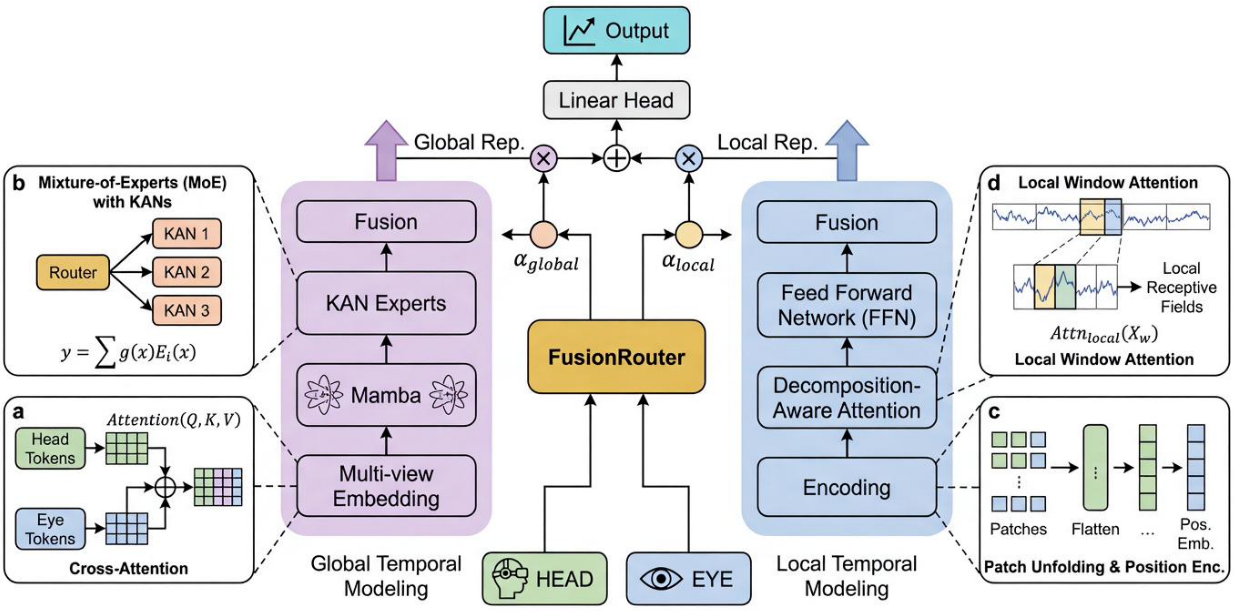

FIGURE 1. Overview of CyberGATE.

Note: The framework uses a FusionRouter to combine global state-space features and local patch-attention features. The major blocks include multi-view embedding, KAN-MoE experts, decomposition-aware encoding, and local window attention.

cross-modal interaction.

$$X = X_{\text{token}} + X_{\text{cross}} \tag{2}$$

This embedding design preserves modality-specific information while allowing the two channels to exchange evidence. Such early interaction helps represent visual-vestibular inconsistency, which is closely related to cybersickness [27].

For long-horizon temporal modeling, Mamba is used as the backbone of the global branch [9]. Unlike recurrent models that update hidden states sequentially and are less parallelizable [28], the state-space formulation can process wearable signal sequences efficiently. Its state update is expressed as

$$\begin{aligned} h(t) &= Ah(t-1) + Bx(t), \\ y(t) &= Ch(t). \end{aligned} \tag{3}$$

In this expression, $h(t)$ is the hidden state, $x(t)$ is the input at time t, and A, B, and C are trainable parameters. Global average pooling then converts the encoded sequence into a fixed-length vector that summarizes the whole observation window.

The model also includes a sparse expert module to handle large differences among users and motion patterns. The experts are implemented with Kolmogorov-Arnold Networks rather than ordinary MLPs [10]. Their spline-based functions are suitable for smooth continuous trajectories and provide better interpretability for behavioral signals [29]. A router selects the top-k experts for each sample and assigns normalized weights to their outputs.

$$f_{\text{global}} = \sum_{i \in \text{Top-}k} w_i \cdot f_i(z) \tag{4}$$

Because only a subset of experts is activated, different experts can specialize in different response patterns while the whole model still keeps generalization ability.

The local branch is introduced to capture short events that may disappear after global pooling. It uses a Decomposition-Aware Patch Transformer. A learnable decomposer separates the input sequence into a low-frequency trend and a high-frequency residual [11]. These two components are processed by independent patch encoders. With patch size p and stride s, the trend component is segmented as follows.

$$X_{\text{patch}} = \text{Unfold}(T) \in \mathbb{R}^{B \times C \times p \times N} \tag{5}$$

The patch number depends on the sequence length, patch size, and stride. After segmentation, patches are flattened, projected into D-dimensional tokens, and enhanced with learnable positional encodings, as shown in Figure 1(c). Local attention is then applied to the patch tokens. Compared with full-sequence attention, windowed attention reduces computation and focuses more directly on abrupt temporal changes [8,30,31]. Its local self-attention is formulated as [11]

$$\text{Attention}(Q_w, K_w, V_w) = \text{softmax}\left(\frac{Q_w K_w^T}{d_k}\right) V_w \tag{6}$$

Here, the query, key, and value are limited to the current window. The attention-enhanced trend and residual outputs are added and then pooled in two stages.

$$H_{\text{local}} = \text{AvgPool}_{\text{channel}}\big(\text{AvgPool}_p\left(T_{\text{trend}} + T_{\text{residual}}\right)\big) \tag{7}$$

The pooled local vector compactly represents short-range temporal behavior. The two branches are integrated by an adaptive fusion module. Eye and head features are first combined inside each branch. A lightweight router then generates two weights for the global and local outputs, with the weights constrained to sum to one. The final representation is

$$H_{\text{final}} = \alpha_{\text{global}} \cdot H_{\text{global}} + \alpha_{\text{local}} \cdot H_{\text{local}} \tag{8}$$

With this routing strategy, the network can rely more on long-term context or on local fluctuations depending on the observed signal pattern. The fused vector is fed into a linear regression head to produce the cybersickness score.

$$\hat{y} = W_f f_{\text{final}} + b \tag{9}$$

The regression head contains a learnable weight and bias. The main optimization target is mean squared error,

$$L_{\text{MSE}} = \frac{1}{B}\sum_{i=1}^{B}(\hat{y}_i - y_i)^2 \tag{10}$$

and a KAN regularization item is added to improve smoothness and control complexity,

$$L_{\text{KAN}} = \sum_{i=1}^{N_{\text{KAN}}} R_{\text{KAN}}^{(i)} \tag{11}$$

so the total training loss is obtained by combining the two parts.

$$L_{\text{total}} = L_{\text{MSE}} + \lambda \cdot L_{\text{KAN}} \tag{12}$$

The coefficient of the regularization term is set to 1e-3 in the experiments.

IV. EXPERIMENTS AND RESULTS

A. DATASETS

Two public datasets are used in the experiments: MSCVR from the PRE-CYSE project and the VR Cybersickness dataset from ETH Zurich [2,32]. They differ in task setting, label design, and available signals, but both were collected with the HTC VIVE Pro Eye device. To make the comparison consistent, this study keeps only the common eye-tracking and head-motion features and excludes EEG, EDA, BVP, and other physiological channels. This setting tests the predictive ability of HMD-native measurements.

1) MSCVR Dataset

MSCVR contains 2,700 labeled samples collected from 45 participants between 20 and 36 years old. The mean age is 24.4 years, with 24 male and 21 female participants. The average MSSQ score is 10.0, and the standard deviation is 8.71, indicating a normal range of cybersickness susceptibility. Participants watched twenty 45-second panoramic VR videos from ten themes. Eye and head signals were recorded during viewing, and Fast Motion Sickness scores were reported every 15 seconds. Therefore, each video produced three labeled segments.

TABLE 1. Modalities and extracted features

Modality	Extracted features
Eye movement	[From each of the left and right eyes] eye openness; gaze direction (x, y, z); gaze origin (x, y, z); pupil diameter
Head transform	Head position (x, y, z); head rotation (yaw, pitch, roll)

2) VR Cybersickness Dataset

The VR Cybersickness dataset includes 16 participants aged from 26 to 37 years, including 10 males and 6 females. During a VR cycling task, participants continuously indicated discomfort with a joystick. Although EEG is emphasized in the original dataset, this paper extracts the same 22 eye and head features for a unified input format. All participants had prior VR experience and no severe motion-sickness history.

The two datasets together provide stable, high-frequency HMD measurements that cover gaze movement and head-pose variation. A shared feature and preprocessing scheme makes it possible to evaluate one model under comparable conditions.

B. PREPROCESSING

The two datasets were first handled independently and then converted to the same sampling rate, feature dimension, and target scale. The MSCVR processing procedure served as the reference. For MSCVR, the 90 Hz recordings were resampled to 25 Hz to remove redundancy and reduce noise. Duplicate records were discarded. Each 45-second video was split into three 15-second windows, so each sample contained 375 time steps. The input contained 22 features, including 16 gaze-related features and 6 head-motion features, as listed in Table 1. Min-Max normalization scaled all features to [0,1], and FMS scores on the 0-20 range were used as labels.

For the VR Cybersickness dataset, the same 22-dimensional feature set was retained and resampled to 25 Hz. Joystick labels originally ranging from 0 to 1 were linearly mapped to the 1-20 interval to match the MSCVR label range. After these steps, both datasets shared consistent input and output formats, allowing the same training and testing pipeline to be applied.

C. EXPERIMENTAL SETTINGS

Training was performed with Adam. The initial learning rate was 1e-4, and cosine annealing was used for 128 epochs. Each iteration cleared gradients, executed the forward pass, calculated the loss, and updated parameters. The loss contained the regression term and the KAN regularization term in Eq. (12). Gradient accumulation was used when the mini-batch was smaller than the target batch size. To reduce target imbalance, samples were weighted by the inverse frequency of their label bins. The model with the best validation score was saved.

The final architecture integrates Multi-View Embedding, MambaMoE, KAN experts, and DAPatchTrans. This combination is intended to capture both global context and short local changes. The evaluation protocol follows the original dataset designs: five-fold cross-validation for MSCVR and leave-one-subject-out testing for VR Cybersickness. These settings provide a fair comparison and examine cross-subject robustness.

D. MODEL COMPLEXITY AND EFFICIENCY ANALYSIS

Computational feasibility is important for online cybersickness warning. Therefore, CyberGATE is compared with LSTM, Bi-LSTM, and Transformer baselines under identical input shapes: 128 x 16 for eye features and 128 x 6 for head features, with a batch size of 128. The comparison uses trainable parameter count, FLOPs, and inference latency on CPU and GPU.

As shown in Figure 2, Transformer has the largest model size, with 1.13M trainable parameters. This is much larger than LSTM and Bi-LSTM, mainly because multi-head attention and feed-forward layers introduce substantial capacity. Transformer also reaches 2.25M FLOPs, far above the recurrent baselines. The main reason is the quadratic dependence of self-attention on sequence length.

Latency follows a similar pattern. On CPU, LSTM achieves the shortest batch latency, while Transformer is the slowest. Bi-LSTM lies between them because the bidirectional structure increases recurrent computation. GPU acceleration reduces absolute latency for all models, but the ranking remains similar. Thus, the architecture itself remains a key factor in real-time efficiency.

For Mamba-based models, the FLOPs reported by the thop profiler should be interpreted as approximate lower bounds because custom selective state-space kernels are not fully traced. The reported value mainly reflects operations in ordinary linear layers. Even under this limitation, the results show that Transformer models can be computationally expensive for latency-sensitive VR scenarios. CyberGATE is designed to reduce this pressure by combining linear-time state-space modeling with localized attention.

E. EVALUATION METRICS

Two regression metrics are used: MAE and RMSE. MAE measures the average absolute prediction error. RMSE gives larger penalties to large deviations and is therefore more sensitive to outlier predictions.

F. COMPARISON WITH CURRENT METHODS

Ten representative methods are selected for comparison. For reproducible baselines, all models use the same bimodal input, including 16 eye features and 6 head-motion features. Preprocessing, splitting, and evaluation are kept consistent as far as possible.

TABLE 2. Performance comparison on the MSCVR dataset

Work	Method	MAE	RMSE
DeepLSTM [33]	LSTM	0.79	1.26
Informer [34]	Transformer	0.68	1.14
Mamba [9]	State-space model	0.65	1.15
PhysioDNN [35]	CNN-LSTM	0.54	0.38
HMDPrediction [36]	CNN-LSTM	0.20	0.62
MAC [37]	CNN-LSTM	0.17	0.28
MS-STTN [2]	Transformer	0.12	0.25
AutoDetectionCS [38]	cGAN-LSTM	0.56	1.05
CyberGATE (ours)	Mamba-Transformer	0.08	0.19

1) DeepLSTM [33]: a ten-layer LSTM autoencoder with encoder-decoder stacking, dropout, and a fully connected reconstruction head.
2) Informer [34]: a long-sequence Transformer variant that uses ProbSparse attention, distillation, and generative decoding.
3) Mamba [9]: a state-space baseline that processes projected features in separate modules and concatenates their outputs for final prediction.
4) PhysioDNN [35]: a CNN-LSTM network combining time-distributed convolution, bidirectional recurrent modeling, and a dense output layer.
5) HMDPrediction [36]: a 3D-CNN and temporal CNN-LSTM model for spatiotemporal feature extraction and sickness-score estimation.
6) MAC [37]: a multimodal attention model using convolution-attention and Bi-LSTM-attention modules.
7) MS-STTN [2]: a dual-stream Transformer with separate spatial and temporal encoders for eye and head modalities.
8) AutoDetectionCS [38]: a Bi-LSTM model supported by conditional-GAN augmentation and temporal convolution.
9) KinematicModel [39]: a deep network integrating physiological encoding, LSTM temporal processing, and a linear prediction head.
10) BeyondSubjectivity [32]: a compact two-stream model that fuses multiscale spatiotemporal information for symptom grading.

Tables 2 and 3 summarize the results on MSCVR and VR Cybersickness. Reproduced methods are tested under the same conditions, while non-reproduced results follow the original reports. On MSCVR, CyberGATE achieves the lowest MAE and RMSE. It outperforms strong baselines such as MS-STTN and AutoDetectionCS. Paired t-tests and Wilcoxon signed-rank tests confirm significant improvements over representative competitors at $p < 0.01$.

On the VR Cybersickness dataset, CyberGATE also obtains the best MAE and a competitive RMSE. It improves over Mamba and AutoDetectionCS in the main comparison. Although BeyondSubjectivity and KinematicModel report lower RMSE values, their MAE values are much larger,

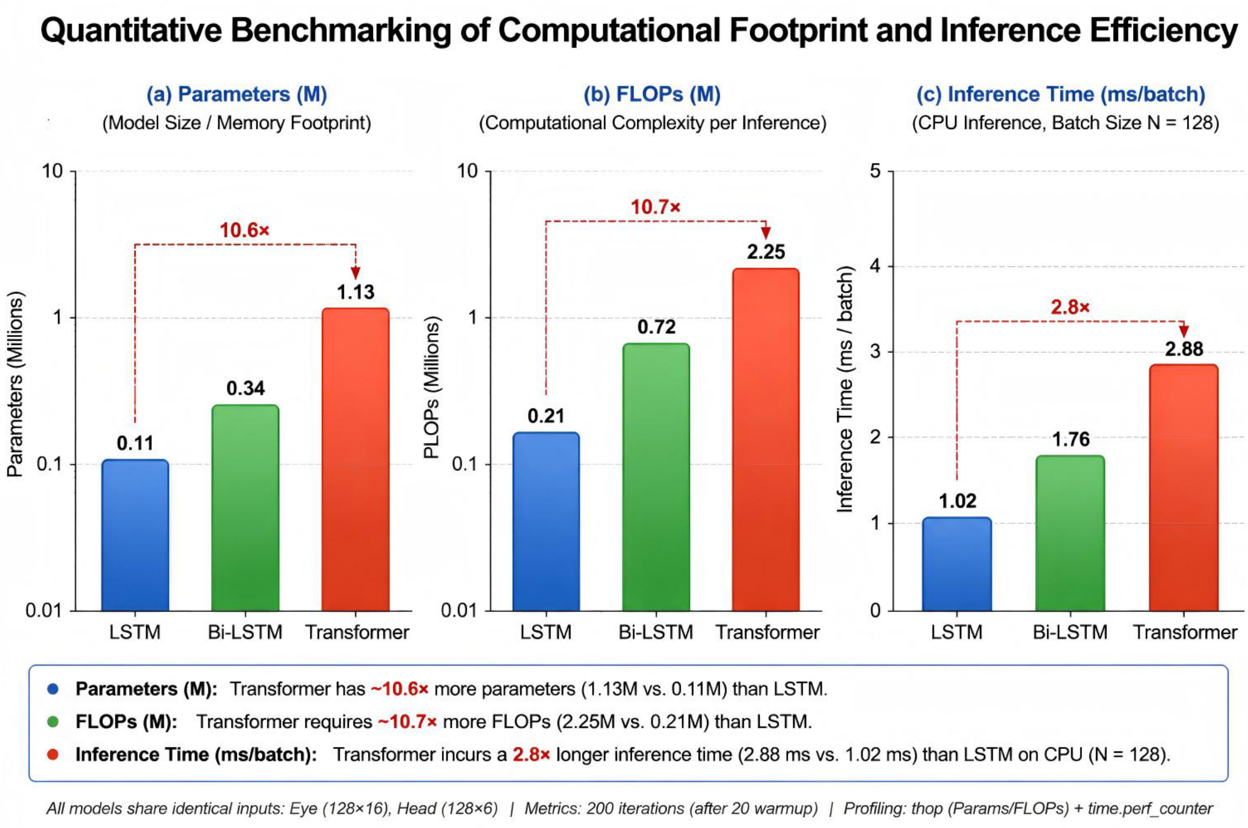

FIGURE 2. Comparison of model size, computational cost, and inference latency.

Note: The panels compare parameters, FLOPs, and CPU latency using the same batch size. Transformer requires more computation than recurrent baselines.

TABLE 3. Performance comparison on the VR Cybersickness dataset

Work	Method	MAE	RMSE
DeepLSTM [33]	LSTM	0.80	1.10
Informer [34]	Transformer	0.95	1.31
Mamba [9]	State-space model	0.70	1.07
KinematicModel [39]	Conv-LSTM	0.86	0.40
BeyondSubjectivity [32]	Conv-LSTM	0.96	0.44
AutoDetectionCS [38]	cGAN-LSTM	0.61	0.97
CyberGATE (ours)	Mamba-Transformer	0.52	0.79

showing weaker overall regression stability. Statistical testing again supports significant improvement over the principal baselines.

Overall, the results show that CyberGATE provides a strong balance between accuracy and generalization across the two datasets.

G. VISUALIZATION OF ADAPTIVE FUSION AND EXPERT ROUTING

The results in Section 4.6 show the overall advantage of CyberGATE, but they do not reveal which internal mechanisms are responsible. The model is based on two assumptions. First, a learnable gate should select different proportions of global state-space information and local spectral-temporal information for different samples. Second, sparse expert routing should allow heterogeneous eye and head patterns to be modeled without forcing every sample through one shared mapping. To test these assumptions, a one-factor-at-a-time sensitivity study is conducted on five hyperparameters of the fusion gate and MoE router.

All other settings follow Section 4.3. Each variant is trained and evaluated with the same data split, so the comparison is less affected by random partitioning. Table 4 lists the sensitivity results, and Figure 3–Figure 5 visualize routing and fusion behavior.

Figure 3 analyzes the MoE router. Validation accuracy first rises when the number of experts increases from two to four, then declines at six experts and partially recovers at eight. This suggests that the router needs enough experts for specialization but also enough samples per expert for stable fitting. If there are too few experts, heterogeneous gaze and head patterns are mixed into the same function. If there are too many, each expert receives limited data and KAN spline fitting becomes less reliable. The top-k results change only slightly, but sparse routing performs marginally better, indicating that the experts are not simply redundant.

Figure 4 examines the adaptive fusion gate. The window size of the local SST-LWT branch barely changes performance across tested values, which indicates that the router

TABLE 4. Sensitivity analysis of adaptive fusion and expert routing in CyberGATE. Bold values denote the default setting, and delta indicates the sweep range in percentage points

Module	Hyper-parameter	Value	Val. Acc. (%)	Val. macro-F1 (%)
Expert routing	N_e ($\Delta = 0.60/0.73$)	2	98.88	98.63
Expert routing	N_e ($\Delta = 0.60/0.73$)	**4**	**99.24**	**99.08**
Expert routing	N_e ($\Delta = 0.60/0.73$)	6	98.65	98.35
Expert routing	N_e ($\Delta = 0.60/0.73$)	8	98.83	98.57
Top-k sparsity	k ($\Delta = 0.15/0.18$)	1	99.42	99.30
Top-k sparsity	k ($\Delta = 0.15/0.18$)	**2**	**99.41**	**99.28**
Top-k sparsity	k ($\Delta = 0.15/0.18$)	3	99.37	99.23
Top-k sparsity	k ($\Delta = 0.15/0.18$)	4	99.28	99.12
SSM state dim.	d_{state} ($\Delta = 0.64/0.79$)	4	99.30	99.15
SSM state dim.	d_{state} ($\Delta = 0.64/0.79$)	8	99.28	99.12
SSM state dim.	d_{state} ($\Delta = 0.64/0.79$)	**16**	**99.37**	**99.23**
SSM state dim.	d_{state} ($\Delta = 0.64/0.79$)	32	98.73	98.45
Local window	w_ℓ ($\Delta = 0.02/0.03$)	3	99.39	99.25
Local window	w_ℓ ($\Delta = 0.02/0.03$)	5	99.37	99.23
Local window	w_ℓ ($\Delta = 0.02/0.03$)	**7**	**99.37**	**99.23**
Local window	w_ℓ ($\Delta = 0.02/0.03$)	9	99.39	99.26
Local window	w_ℓ ($\Delta = 0.02/0.03$)	11	99.37	99.24
Backbone width	d_{model} ($\Delta = 2.70/3.31$)	128	96.62	95.86
Backbone width	d_{model} ($\Delta = 2.70/3.31$)	**256**	**99.32**	**99.17**
Backbone width	d_{model} ($\Delta = 2.70/3.31$)	512	99.03	98.82

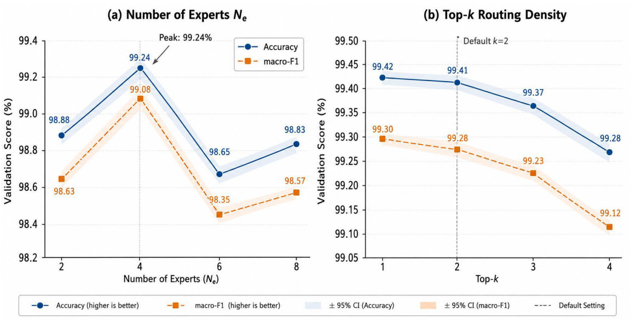

FIGURE 3. Expert-routing landscape.

Note: The expert-count curve shows an inverted-U pattern, while denser top-k routing slightly reduces validation performance, implying that expert specialization is beneficial.

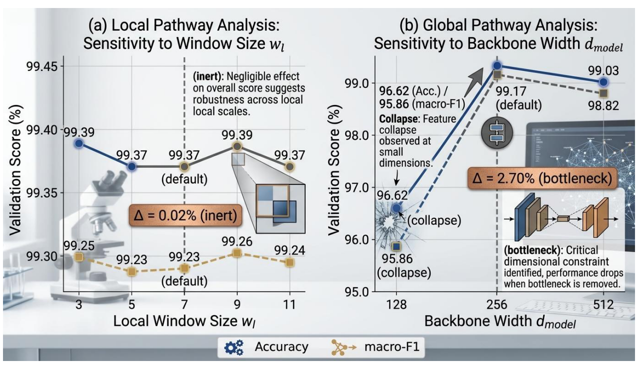

FIGURE 4. Asymmetric sensitivity of the adaptive fusion gate.

Note: Local-window size has little effect, whereas too small a backbone width clearly hurts performance. FusionRouter can absorb local-window uncertainty but cannot compensate for insufficient global representation.

can rescale local evidence and reduce dependence on a precise receptive-field setting. Backbone width behaves differently. When the width is reduced, accuracy drops sharply, revealing a representation bottleneck. A larger width brings

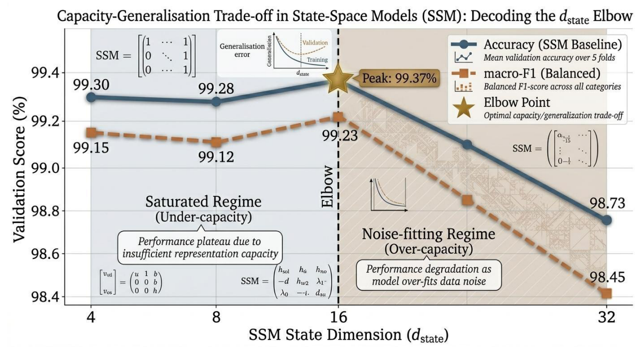

FIGURE 5. Capacity-generalization elbow of the SSM state dimension.

Note: Accuracy saturates at moderate state dimensions and then drops at the largest setting, suggesting a transition from useful capacity to noise fitting.

only a slight regression, likely because the fixed training budget cannot fully support the increased capacity. This contrast shows that adaptive fusion handles local uncertainty but cannot recover information that the upstream encoder never extracts.

The last sweep considers the SSM state dimension. Performance stays almost unchanged at moderate dimensions and decreases when the dimension is increased to 32. This pattern implies that the hidden state only needs enough capacity to capture the intrinsic temporal structure of cybersickness. Beyond that point, extra dimensions may model noise rather than signal. The selected dimension therefore corresponds to a capacity-generalization elbow.

Three observations can be drawn from these visualizations. Router hyperparameters and SSM state dimension show limited variation, indicating that CyberGATE is not highly fragile to tuning. The local-window parameter is almost inactive, consistent with the expected compensatory role of adaptive gating. Backbone width is the only clearly sensitive factor, because later reweighting cannot replace missing upstream representation. These results support the robustness of CyberGATE along the dimensions that adaptive fusion is designed to handle.

H. ABLATION STUDY

TABLE 5. Ablation study results on the MSCVR and VR Cybersickness datasets

Model variants	MSCVR MAE	MSCVR RMSE	VR Cybersickness MAE	VR Cybersickness RMSE
w/o Head features	0.33	0.90	0.74	1.10
w/o Eye features	0.89	1.38	0.80	1.15
w/o Multi-view embedding	1.02	1.18	1.13	1.28
w/o Mamba backbone	0.17	0.29	0.63	0.98
w/o KAN experts	0.37	0.84	0.87	1.28
w/o Global path	1.21	1.53	1.21	1.57
w/o Local path	0.18	0.37	0.90	1.26
w/o Adaptive fusion	0.52	1.05	0.71	1.09
Full CyberGATE	0.08	0.19	0.52	0.79

Table 5 gives the ablation results on both datasets. Each major component contributes to the final performance, since removing any module increases prediction error.

The global branch has the strongest effect on long-range temporal modeling. Without it, MAE and RMSE increase noticeably. Removing Multi-View Embedding or Mamba also weakens performance, showing that early modality interaction and global sequence encoding are necessary for describing cybersickness trends.

The local branch improves sensitivity to short disturbances. This is especially clear for the VR Cybersickness dataset, where visual and vestibular changes can be abrupt. On MSCVR, removing the local branch or Mamba causes a smaller decline, probably because the video content is more stable and global patterns dominate.

The ablation results demonstrate that the two branches provide different but complementary information. Their integration enables the model to describe both persistent discomfort trends and transient response changes.

V. CONCLUSION

This paper develops CyberGATE, a multimodal cybersickness prediction model based on adaptive multiscale gated fusion. The model uses eye and head signals to estimate symptom intensity continuously. The global branch learns long-range temporal patterns through multi-view embedding and Mamba-MoE modeling, while the local branch captures short changes with decomposition-aware patch encoding and local attention. Experiments on MSCVR and VR Cybersickness show that CyberGATE performs better than representative baselines in most evaluation settings. Ablation and LOSO analyses further confirm the importance of both branches and indicate good cross-subject generalization. These findings suggest that CyberGATE can serve as an effective technical basis for real-time cybersickness monitoring in immersive VR.

The study also has limitations. The two datasets contain limited participant diversity, so further tests are needed across broader ages, devices, interaction styles, and VR contents. Eye and head features are convenient for HMD deployment, but excluding physiological sensors may miss some hidden discomfort states. In addition, the present method is trained offline, whereas practical VR systems may face distribution drift and personalized symptom thresholds. Future work will explore personal calibration, uncertainty-aware prediction, and online updating for long-term use. Lightweight physiological sensing, real-time scene descriptors, and adaptive intervention policies can also be integrated so that prediction results directly support dynamic content adjustment and user comfort optimization.

FUNDING

This research was supported by the Beijing High-level Overseas Talent Returning Grant, the NCUT Undergraduate Educational Reform Project, and Undergraduate Training Programs for Innovation and Entrepreneurship.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

DATA SHARING AGREEMENT

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

[1] R. S. Kennedy, N. E. Lane, K. S. Berbaum, et al., "Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness," The International Journal of Aviation Psychology, vol. 3, no. 3, pp. 203–220, 1993.

[2] D. Jeong and K. Han, "PRECYSE: Predicting Cybersickness using Transformer for Multimodal Time-Series Sensor Data," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 8, no. 2, pp. 1–24, 2024.

[3] B. Keshavarz, K. Peck, S. Rezaei, et al., "Detecting and predicting visually induced motion sickness with physiological measures in combination with machine learning techniques," International Journal of Psychophysiology, vol. 176, pp. 14–26, 2022.

[4] N. Tian and R. Boulic, "Who says you are so sick? An investigation on individual susceptibility to cybersickness triggers using EEG, EGG and ECG," IEEE Transactions on Visualization and Computer Graphics, vol. 30, no. 5, pp. 2379–2389, 2024.

[5] B. Fu, W. Chu, C. Gu, et al., "Cross-modal guiding neural network for multimodal emotion recognition from EEG and eye movement signals," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 10, pp. 5865–5876, 2024.

[6] J. Sameri, H. Coenegracht, S. Van Damme, et al., "Physiology-driven cybersickness detection in virtual reality: a machine learning and explainable AI approach," Virtual Reality, vol. 28, no. 4, Art. no. 174, 2024.

[7] E. Chang, H. T. Kim, and B. Yoo, "Predicting cybersickness based on user's gaze behaviors in HMD-based virtual reality," Journal of Computational Design and Engineering, vol. 8, no. 2, pp. 728–739, 2021.

[8] Y. Nam, U. Hong, H. Chung, et al., "Eye movement patterns reflecting cybersickness: evidence from different experience modes of a virtual reality game," Cyberpsychology, Behavior, and Social Networking, vol. 25, no. 2, pp. 135–139, 2022.

[9] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," arXiv preprint arXiv:2312.00752, 2024.

[10] Z. Liu, Y. Wang, S. Vaidya, et al., "KAN: Kolmogorov-Arnold Networks," arXiv preprint arXiv:2404.19756, 2024.

[11] X. Xu, C. Chen, Y. Liang, et al., "SST: Multi-scale hybrid Mamba-Transformer experts for long-short range time series forecasting," arXiv preprint arXiv:2404.14757, 2024.

[12] H. Oh, W. Son, J. Park, et al., "Cybersickness and its severity arising from virtual reality content: a comprehensive study," Sensors, vol. 22, no. 4, Art. no. 1314, 2022.

[13] J. T. Reason and J. J. Brand, Motion Sickness. London: Academic Press, 1975.
[14] G. E. Riccio and T. A. Stoffregen, "An ecological theory of motion sickness and postural instability," Ecological Psychology, vol. 3, no. 3, pp. 195–240, 1991.
[15] P. Kourtesis, A. Papadopoulou, and P. Roussos, "Cybersickness in virtual reality: the role of individual differences, effects on cognitive functions, motor skills, and intensity differences during and after immersion," Virtual Worlds, vol. 3, no. 1, Art. no. 4, 2024.
[16] R. Hussain, M. Chessa, and F. Solari, "Mitigating cybersickness in virtual reality systems through foveated depth-of-field blur," Sensors, vol. 21, no. 12, Art. no. 4006, 2021.
[17] I. Lukacova, B. Keshavarz, and J. F. Golding, "Measuring the susceptibility to visually induced motion sickness and its relationship with vertigo, dizziness, migraine, syncope and personality traits," Experimental Brain Research, vol. 241, pp. 1381–1391, 2023.
[18] Z. T. Gu, D. Ding, Y. T. Chen, et al., "Evaluation and mitigation methods for virtual reality motion sickness," Journal of Zhejiang University (Science Edition), vol. 52, no. 1, pp. 37–46, 2025.
[19] E. Krokos and A. Varshney, "Quantifying VR cybersickness using EEG," Virtual Reality, vol. 26, no. 1, pp. 77–89, 2022.
[20] G. Zerveas, S. Jayaraman, D. Patel, et al., "A Transformer-based framework for multivariate time series representation learning," in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021, pp. 2114–2124.
[21] M. Pang, H. Wang, J. Huang, et al., "Multi-scale masked autoencoders for cross-session emotion recognition," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 32, pp. 1637–1646, 2024.
[22] R. K. Kundu, R. Islam, J. Quarles, et al., "LiteVR: interpretable and lightweight cybersickness detection using explainable AI," arXiv preprint arXiv:2302.03037, 2023.
[23] Y. Zhu, T. Li, and Y. Wang, "Real-time cross-modal cybersickness prediction in virtual reality," arXiv preprint arXiv:2501.01212, 2025.
[24] Y. Choi, D. Jeong, B. Kim, et al., "Early prediction of cybersickness in virtual reality using multimodal sensor data," in Proceedings of the ACM Symposium on Applied Perception, 2024, pp. 1–10.
[25] Y. Long, X. Zhang, H. Li, et al., "Toward accurate cybersickness prediction in virtual reality," Sensors, vol. 25, no. 18, Art. no. 5606, 2025.
[26] A. N. Ramaseri-Chandra, J. M. Bogle, and S. M. Peterson, "Exploring the feasibility of head-tracking data for cybersickness prediction in virtual reality," Electronics, vol. 14, no. 3, Art. no. 502, 2025.
[27] S. S. Yeo, S. Y. Park, and S. H. Yun, "Investigating cortical activity during cybersickness by fNIRS," Scientific Reports, vol. 14, no. 1, Art. no. 8093, 2024.
[28] R. Islam, S. Ang, and J. Quarles, "Towards forecasting the onset of cybersickness by fusing physiological, head-tracking and eye-tracking with multimodal deep fusion network," in Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, 2022, pp. 121–130.
[29] L. Chiparova and V. Popov, "Kolmogorov-Arnold Networks for system identification of first- and second-order dynamic systems," Engineering Proceedings, vol. 100, no. 1, Art. no. 100059, 2025.
[30] Q. Yu, Y. Xia, Y. Bai, et al., "Glance-and-gaze vision transformer," arXiv preprint arXiv:2106.02277, 2021.
[31] J. Zhang, Y. Wang, X. Liu, et al., "How different text display patterns affect cybersickness in augmented reality environments," Scientific Reports, vol. 14, Art. no. 12679, 2024.
[32] B. U. Demirel, A. H. Dogan, J. Rossie, et al., "Beyond subjectivity: continuous cybersickness detection using EEG-based multitaper spectrum estimation," IEEE Transactions on Visualization and Computer Graphics, vol. 31, no. 5, pp. 2525–2534, 2025.
[33] Y. Wang, J. R. Chardonnet, and F. Merienne, "VR sickness prediction for navigation in immersive virtual environments using a deep long short-term memory model," in Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces, 2019, pp. 1874–1881.
[34] H. Zhou, S. Zhang, J. Peng, et al., "Informer: beyond efficient Transformer for long sequence time-series forecasting," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 12, pp. 11106–11115, 2021.
[35] M. Yalcin, A. Halbig, M. Fischbach, et al., "Automatic cybersickness detection by deep learning of augmented physiological data from off-the-shelf consumer-grade sensors," Frontiers in Virtual Reality, vol. 5, Art. no. 1364207, 2024.
[36] R. Islam, K. Desai, and J. Quarles, "Cybersickness prediction from integrated HMD's sensors: a multimodal deep fusion approach using eye-tracking and head-tracking data," in Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, 2021, pp. 31–40.
[37] D. Jeong, S. Paik, Y. Noh, et al., "MAC: multimodal, attention-based cybersickness prediction modeling in virtual reality," Virtual Reality, vol. 27, no. 3, pp. 2315–2330, 2023.
[38] M. Yalcin, A. Halbig, M. Fischbach, et al., "Automatic cybersickness detection by deep learning of augmented physiological data from off-the-shelf consumer-grade sensors," Frontiers in Virtual Reality, vol. 5, Art. no. 1364207, 2024.
[39] R. Li, Y. Wang, H. Yin, et al., "A deep cybersickness predictor through kinematic data with encoded physiological representation," in Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, 2023, pp. 1132–1141.