Scientific Research Publishing

Journal of Computer and Communications, 2026, 14(7), 89-109
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

State-Aware Cross-Modal Contrastive Learning for Multimodal Vigilance Estimation

Wenfei Shu, Weidong Li

School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai, China
Email: 243391548@st.usst.edu.cn

How to cite this paper: Shu, W.F. and Li, W.D. (2026) State-Aware Cross-Modal Contrastive Learning for Multimodal Vigilance Estimation. *Journal of Computer and Communications*, **14**, 89-109.
https://doi.org/10.4236/jcc.2026.147005

Received: June 24, 2026
Accepted: July 20, 2026
Published: July 23, 2026

Abstract

To enhance multimodal representation learning for continuous vigilance estimation, this paper proposes a State-Aware Cross-Modal Contrastive Learning (SA-CL) framework for EEG-EOG collaborative modeling. In continuous vigilance regression tasks, samples from different temporal segments may correspond to similar vigilance states, while conventional alignment strategies tend to treat them as strictly negative pairs, thereby undermining the semantic consistency of cross-modal representations. To address this issue, a state-aware constraint is introduced into the alignment process between EEG and EOG representations, where negative sample relationships are selectively filtered or adaptively weighted according to inter-sample state similarity, thus alleviating the unreasonable repulsion between samples with similar vigilance states. Based on this strategy, an EEG-EOG multimodal collaborative framework is constructed. Specifically, the EEG branch extracts the spatial, topological, and temporal dynamic features of brain signals, while the EOG branch encodes eye-movement behavioral information. The complementary representations from both modalities are then fused to perform continuous vigilance regression prediction. Experimental results on the SEED-VIG dataset demonstrate that the proposed method effectively improves multimodal vigilance estimation performance while preserving the semantic consistency of cross-modal features.

Keywords

Electroencephalography (EEG), Electrooculography (EOG), Multimodal Learning, Vigilance Estimation, Contrastive Learning

1. Introduction

Vigilance refers to an individual's ability to sustain attention and responsiveness

during continuous tasks. Its variation is closely related to driving safety, fatigue monitoring, intelligent human-computer interaction, and brain-computer interfaces (BCIs). Vigilance decline often leads to slower reactions, more judgment errors, and reduced task performance. Therefore, continuous and accurate vigilance estimation is of considerable theoretical and practical importance.

Existing methods mainly rely on behavioral cues, visual information, and physiological signals. Physiological signals can more directly reflect internal cognitive states and may enable earlier detection of vigilance decline than external behavioral features [1]. In particular, EEG reflects neural activity and cognitive changes, whereas EOG captures fatigue-related eye movements, such as blinking, fixation, and saccades. Their complementary properties make EEG and EOG suitable for joint continuous vigilance estimation [2].

Early EEG-EOG-based studies mainly used traditional machine learning methods. Zheng and Lu combined EEG and forehead EOG features with SVR, CCRF, and CCNF to model temporal vigilance dynamics [3]. Huo *et al.* applied graph-regularized extreme learning machines to fused EEG and forehead EOG features for driving fatigue detection [4]. Although these methods demonstrated the feasibility of multimodal physiological modeling, they relied heavily on handcrafted features and prior knowledge, limiting their ability to capture complex nonlinear relationships and cross-modal interactions.

Deep learning has shifted research from shallow fusion toward high-level representation learning. LSTM networks have been used to model temporal dependencies in vigilance transitions [5], while capsule-attention models and deep neural networks with subnetwork neurons have improved EEG-EOG representation learning in different latent spaces [6] [7]. Recent studies further indicate that temporal convolution, spatial attention, and self-supervised learning can enhance discriminative spatiotemporal feature extraction from multimodal physiological signals [8].

However, independently extracting EEG and EOG features cannot fully exploit their complementary information. EEG mainly reflects neural and frequency-band-related changes, whereas EOG is more sensitive to eye-movement behaviors and visual fatigue. Conventional feature-level and decision-level fusion methods typically concatenate features or integrate predictions [9] [10], but often treat fusion as a static process and may overlook deeper cross-modal dependencies.

Attention-based multimodal learning has been introduced to better model such dependencies. Multimodal Transformers can adaptively weight different modalities and capture cross-modal relationships in unaligned sequences [11]. Channel-attention and Transformer-based methods have also been applied to EEG-EOG vigilance estimation [12], while decoupled intra- and inter-modality learning explicitly models modality-specific and complementary information [13]. Nevertheless, most of these methods emphasize feature interaction and adaptive weighting without explicitly constraining representation consistency across vigilance states.

Contrastive learning improves representation consistency by bringing seman-

tically similar samples closer and separating dissimilar samples. It has achieved success in computer vision [14] [15], natural language processing [16], speech representation learning [17], and image-text alignment [18]. In physiological signal analysis, modality-pairwise contrastive learning has been used to align EEG and EOG samples from the same temporal segment [19], and hierarchical contrastive learning has explored local and global alignment for vigilance detection [20].

However, conventional contrastive learning usually treats all unmatched samples as negative pairs. This assumption is inappropriate for continuous regression tasks because labels evolve smoothly and samples from different temporal segments may represent similar cognitive states. Treating all unmatched pairs as negatives can introduce false negatives and disrupt the semantic structure of the feature space [21] [22]. This issue is particularly relevant to continuous EEG-EOG vigilance estimation, where adjacent or state-similar segments are inherently correlated.

To address this problem, this study proposes a State-Aware EEG-EOG Cross-Modal Contrastive Learning (SA-CL) framework for continuous vigilance estimation. In addition to aligning EEG and EOG samples from the same temporal segment, SA-CL screens or reweights unmatched samples according to vigilance-state similarity. This strategy reduces unreasonable repulsion between state-similar samples and promotes cross-modal representations that better reflect the continuous structure of vigilance states. Based on this mechanism, an EEG-EOG collaborative framework is constructed for continuous vigilance regression.

The main contributions are summarized as follows. First, a State-Aware Cross-Modal Contrastive Learning strategy is proposed to alleviate false negatives caused by the uniform negative-pair assumption in continuous vigilance regression. Second, an EEG-EOG collaborative framework is developed to jointly learn EEG spatiotemporal features, EOG eye-movement representations, and cross-modal complementary information. Third, EEG and EOG features are fused for regression prediction to improve continuous vigilance estimation. Finally, extensive experiments on a public dataset demonstrate the effectiveness and robustness of the proposed method.

2. Methods

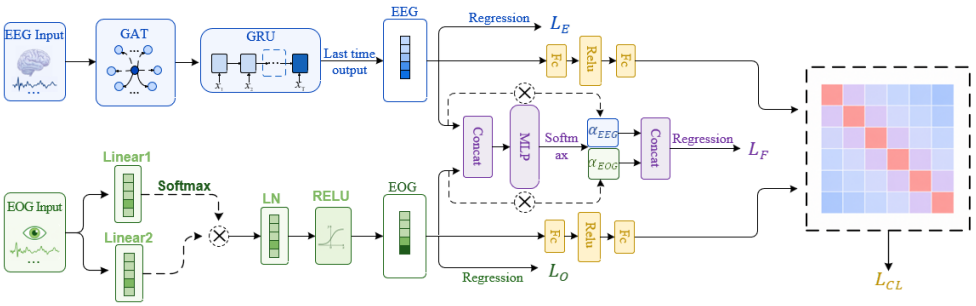

Figure 1. Overall structure of the proposed framework.

This section introduces the EEG and EOG feature extraction methods, as well as the proposed multimodal continuous vigilance regression framework, which consists of EEG-EOG collaborative modeling, state-aware cross-modal contrastive learning, and fusion-based regression prediction. The overall network architecture is illustrated in Figure 1.

2.1. EEG and EOG Feature Presentation

1) EEG Features:

Differential Entropy (DE) is an extension of information entropy for continuous random variables and is commonly used to measure the complexity of EEG signals. Since DE can effectively reflect changes in brain activity states and has demonstrated good stability and discriminative capability in tasks such as emotion recognition and fatigue detection [23], this study adopts DE as the frequency-domain feature representation for EEG signals. The DE is defined as follows:

$$\mathrm{DE} = -\int_{-\infty}^{+\infty} f(x)\log\left(f(x)\right)\mathrm{d}x. \tag{1}$$

where $f(x)$ denotes the probability density function of the continuous random variable x. For EEG signals approximately following a Gaussian distribution $\mathcal{N}\left(\mu,\sigma^2\right)$, the DE can be further simplified as follows:

$$\mathrm{DE} = \frac{1}{2}\log\left(2\pi\mathrm{e}\sigma^2\right), \tag{2}$$

where μ and σ represent the mean and standard deviation of the signal, respectively.

In this study, differential entropy (DE) features are extracted from five classical EEG frequency bands, namely δ (1 - 3 Hz), θ (4 - 7 Hz), α (8 - 13 Hz), β (13 - 30 Hz), and γ (31 - 50 Hz), to characterize the complexity of EEG signals. The detailed EEG segmentation strategy, sliding-window configuration, and multi-time-step feature representation are described in Section 2.1.

This representation not only preserves multi-band and spatial electrode information, but also retains local temporal evolution characteristics within each segment, thereby providing rich inputs for subsequent EEG temporal modeling.

2) EOG Features:

Electrooculography (EOG) signals are mainly used to record potential changes caused by eye movements and are generally divided into vertical electrooculography (VOG) and horizontal electrooculography (HOG) components. Among them, VOG mainly reflects vertical eye movement activities and is more sensitive to blinking, eye closure, and eyelid opening/closing behaviors. In contrast, HOG mainly reflects horizontal eye movements and can characterize behaviors such as horizontal saccades and gaze shifts. Since fatigue or vigilance decline is often accompanied by increased blink frequency, prolonged eye closure duration, and altered eye movement patterns, both VOG and HOG can characterize fatigue-related eye movement features from different perspectives, thereby providing important complementary information for vigilance state recognition.

In this study, following the method proposed by Zheng and Lu [3], a total of 36 eye-movement-related features associated with saccades, blinking, and fixation behaviors were extracted from each 8 s EOG segment, as listed in Table 1.

Table 1. Statistical EOG features used for vigilance estimation.

Category	Features
Blink	blink rate (max/min/mean)
	blink amplitude (max/min/mean)
	blink rate variance (mean/max)
	blink amplitude variance (mean/max)
	blink amplitude power (mean/max)
	blink number
Saccade	saccade rate (max/min/mean)
	saccade amplitude (max/min/mean)
	saccade rate variance (mean/max)
	saccade amplitude variance (mean/max)
	saccade amplitude power (mean/max)
	saccade number
Fixation	blink duration (max/min/mean)
	blink duration variance (mean/max)
	saccade duration (max/min/mean)
	saccade duration variance (mean/max)

2.2. EEG Encoder

In this study, each 8 s EEG segment is treated as a basic sample and corresponds to one segment-level vigilance target. The EEG segment has the same segment-level vigilance label as the temporally aligned 8 s EOG segment, and the detailed PERCLOS-based label computation is described in Section 3.1. To preserve local temporal dynamics within each segment, an overlapping sliding-window strategy is adopted for DE feature extraction. Specifically, each 8 s EEG segment is divided into 2 s short-time windows with a stride of 1 s, corresponding to a 1 s overlap between adjacent windows. As a result, each 8 s segment contains $T = 7$ internal EEG time steps. For the i-th sample at time step t, the differential entropy (DE) features extracted from five frequency bands (δ, θ, α, β, and γ) over $N = 17$ EEG electrodes are represented as:

$$\boldsymbol{x}_{i,t} \in \mathbb{R}^{5\times N},\ i = 1,\cdots,M;\ t = 1,\cdots,T. \tag{3}$$

Accordingly, the multi-temporal EEG feature sequence of the entire 8 s segment can be formulated as:

$$\boldsymbol{X}_i = \left[\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \cdots, \boldsymbol{x}_{i,T}\right] \in \mathbb{R}^{T\times 5\times N}. \tag{4}$$

To capture the spatial relationships among EEG electrodes, each electrode is regarded as a graph node at each time step, and a Graph Attention Network (GAT) [24] is employed for spatial feature learning. In this study, the EEG graph is constructed as a fully connected graph with self-loops. Specifically, for the $N=17$ EEG electrodes, the node set is defined as $\mathcal{V}=\{1,2,\cdots,N\}$, and an edge is established between every pair of electrodes. Therefore, the neighbor set of node m is given by $\mathcal{N}(m)=\mathcal{V}$. This design allows the GAT layer to adaptively learn the contribution of each electrode to any other electrode through attention weights, without imposing a manually predefined spatial adjacency structure. For the input feature $\boldsymbol{x}_{i,t}$, the feature vector of the n-th node is denoted as $\boldsymbol{h}_n \in \mathbb{R}^5$. GAT first projects the node features into a higher-dimensional representation space through a learnable linear transformation:

$$\boldsymbol{z}_n = \boldsymbol{W}\boldsymbol{h}_n, \tag{5}$$

where $\boldsymbol{W}$ denotes the learnable projection matrix and $\boldsymbol{z}_n$ represents the transformed node embedding.

Subsequently, the attention score between two connected nodes n and m is computed as follows:

$$e_{nm} = \phi\left(\boldsymbol{\alpha}_s^{\top}\boldsymbol{z}_n + \boldsymbol{\alpha}_t^{\top}\boldsymbol{z}_m\right), \tag{6}$$

where $\boldsymbol{\alpha}_s$ and $\boldsymbol{\alpha}_t$ are learnable attention vectors corresponding to the source and target nodes, respectively, and $\phi(\cdot)$ denotes the LeakyReLU activation function.

The attention coefficients are then normalized using the softmax operation:

$$\alpha_{nm} = \frac{\exp\left(e_{nm}\right)}{\sum_{k\in\mathcal{N}(m)}\exp\left(e_{km}\right)}, \tag{7}$$

where $\mathcal{N}(m)$ denotes the neighbor set of node m. The normalized coefficient α_{nm} reflects the contribution of node n when updating node m.

Based on the learned attention weights, neighboring node features are aggregated to obtain the updated node representation:

$$\boldsymbol{h}'_m = \sum_{n\in\mathcal{N}(m)} \alpha_{nm}\boldsymbol{z}_n, \tag{8}$$

where $\boldsymbol{h}'_m$ denotes the updated feature representation of node m. This process enables the model to adaptively capture the spatial dependencies among EEG electrodes.

To further enhance the spatial representation capability, multi-head attention is adopted to jointly learn electrode relationships from different representation subspaces. In this work, four attention heads are employed for spatial feature extraction.

After GAT processing, the spatial representation of the i-th EEG segment at time step t is denoted as $\boldsymbol{g}_{i,t}$. The spatial feature sequence of the entire segment can therefore be expressed as follows:

$$\boldsymbol{G}_i = \left[\boldsymbol{g}_{i,1}, \boldsymbol{g}_{i,2}, \cdots, \boldsymbol{g}_{i,T}\right]. \tag{9}$$

Although GAT effectively models spatial correlations among electrodes within each time step, it cannot explicitly capture temporal dependencies across different subwindows. Therefore, a Gated Recurrent Unit (GRU) [25] is further introduced to model the temporal dynamics of the spatial feature sequence $\boldsymbol{G}_i$. Let $\boldsymbol{s}_{i,t}$ denote the hidden state of the GRU at time step t. The hidden state of the last time step is adopted as the global EEG representation of the corresponding 8 s segment:

$$\boldsymbol{f}_i^e = \boldsymbol{s}_{i,T}. \tag{10}$$

Finally, $\boldsymbol{f}_i^e$ serves as the global EEG representation of the i-th EEG segment and is subsequently utilized for cross-modal contrastive learning and multimodal fusion regression.

2.3. EOG Encoder

In the previous subsection, the extracted EOG feature representation was denoted as $\boldsymbol{x}_i^o \in \mathbb{R}^{36}$, where 36 represents the number of handcrafted features extracted from each 8 s temporal segment. EOG signals often exhibit significant coupling relationships among different physiological processes. For example, the amplitude and velocity of blinking, the energy and duration of saccades, as well as the stability and deviation energy of fixation behaviors, usually exhibit multiplicative second-order relationships rather than simple linear correlations. Therefore, relying solely on first-order linear transformations is insufficient to effectively characterize the intrinsic structure of EOG signals, motivating the introduction of mechanisms capable of modeling higher-order feature interactions.

Based on this consideration, an EOG encoder is employed to structurally enhance the 36-dimensional feature representation of each temporal segment. Specifically, for the input feature vector of the i-th 8 s segment, $\boldsymbol{x}_i^o \in \mathbb{R}^{36}$, the feature vector is first projected into two latent spaces:

$$\boldsymbol{u}_i = \boldsymbol{U}^\top \boldsymbol{x}_i^o, \ \boldsymbol{v}_i = \boldsymbol{V}^\top \boldsymbol{x}_i^o, \tag{11}$$

where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{36\times R}$ are learnable projection matrices, and R denotes the latent feature dimension. The introduction of latent spaces enables the model to automatically learn physiologically meaningful directional combinations and capture global correlations among the original handcrafted features.

Subsequently, one projected branch is treated as a gating branch and mapped into the interval $(0,1)$ through the sigmoid activation function to modulate the features of the other branch in an element-wise manner:

$$\boldsymbol{z}_i = \boldsymbol{u}_i \odot \sigma(\boldsymbol{v}_i), \ \boldsymbol{z}_i \in \mathbb{R}^R, \tag{12}$$

where $\odot$ denotes element-wise multiplication and $\sigma(\cdot)$ represents the sigmoid activation function.

To stabilize the feature distribution and further enhance nonlinear representation capability, layer normalization and nonlinear activation are subsequently applied to the gated interaction representation, yielding the enhanced high-level EOG representation:

$$\tilde{z}_i = \mathrm{LN}(z_i), \quad f_i^o = \phi(\tilde{z}_i), \tag{13}$$

where $\mathrm{LN}(\cdot)$ denotes Layer Normalization and $\phi(\cdot)$ represents the LeakyReLU activation function. Finally, f_i^o is regarded as the global EOG representation of the i-th 8 s temporal segment and is further utilized for subsequent cross-modal contrastive learning and fusion-based regression prediction.

2.4. Fusion Model

After obtaining the global EEG representation f_i^e and the global EOG representation f_i^o, a modality-attention-based fusion strategy is adopted to adaptively weight the two modalities. Specifically, the EEG and EOG features are first concatenated into a joint representation:

$$h_i = \left[f_i^e ; f_i^o \right]. \tag{14}$$

The joint representation is then fed into a modality attention layer to generate the modality score vector:

$$s_i = W_a h_i + b_a, \tag{15}$$

where W_a and b_a are learnable parameters.

Subsequently, the normalized modality weights are obtained through the softmax function:

$$\alpha_i = \mathrm{softmax}(s_i) = \left[\alpha_i^e, \alpha_i^o \right]. \tag{16}$$

The obtained weights are further used to adaptively reweight the EEG and EOG representations:

$$\tilde{f}_i^e = \alpha_i^e f_i^e, \quad \tilde{f}_i^o = \alpha_i^o f_i^o. \tag{17}$$

Finally, the weighted modality representations are concatenated to form the fused feature representation:

$$f_i^F = \left[\tilde{f}_i^e ; \tilde{f}_i^o \right], \tag{18}$$

which is subsequently fed into the fusion regression head to generate the final continuous vigilance prediction. This design enables the model to adaptively adjust the importance of different modalities according to each sample while preserving the complementary information between EEG and EOG signals.

2.5. State-Aware Cross-Modal Contrastive Learning

After obtaining the segment-level high-level representations from the EEG and EOG branches, a State-Aware Cross-Modal Contrastive Learning (SA-CL) strategy is introduced to enhance semantic consistency between the two modalities in the representation space.

Traditional cross-modal contrastive learning regards EEG-EOG samples from the same temporal segment as positive sample pairs, while samples from different temporal segments are treated as negative sample pairs. By pulling positive pairs closer and pushing negative pairs farther apart, modality alignment can be achieved.

However, in continuous vigilance regression tasks, samples from different temporal segments may still correspond to highly similar vigilance states. Blindly treating such samples as negative pairs may introduce false negatives, thereby damaging the continuity of the cross-modal representation space.

Assume that the EEG representation and EOG representation of the i-th 8 s segment are denoted as $\boldsymbol{f}_i^e$ and $\boldsymbol{f}_i^o$, respectively, and the corresponding vigilance label is y_i. Let the batch size be B.

Positive sample pairs: EEG-EOG pairs from the same temporal segment are defined as:

$$\left(\boldsymbol{f}_i^e, \boldsymbol{f}_i^o\right). \tag{19}$$

Negative sample pairs: samples from different temporal segments are defined as:

$$\left(\boldsymbol{f}_i^e, \boldsymbol{f}_j^o\right),\ \left(\boldsymbol{f}_i^o, \boldsymbol{f}_j^e\right),\ i \neq j. \tag{20}$$

1) Cosine Similarity:

The cross-modal cosine similarity is defined as:

$$\mathrm{sim}\left(\boldsymbol{f}_i^e, \boldsymbol{f}_j^o\right) = \frac{\left(\boldsymbol{f}_i^e\right)^\top \boldsymbol{f}_j^o}{\left\|\boldsymbol{f}_i^e\right\|\left\|\boldsymbol{f}_j^o\right\|},\ \mathrm{sim}\left(\boldsymbol{f}_i^o, \boldsymbol{f}_j^e\right) = \frac{\left(\boldsymbol{f}_i^o\right)^\top \boldsymbol{f}_j^e}{\left\|\boldsymbol{f}_i^o\right\|\left\|\boldsymbol{f}_j^e\right\|}. \tag{21}$$

The state similarity between samples is further defined according to the label difference:

$$s_{ij} = \exp\left(-\frac{\left(y_i - y_j\right)^2}{\sigma^2}\right), \tag{22}$$

where σ is a scale parameter controlling the decay rate of similarity with respect to label differences.

Accordingly, the negative sample weights are defined as:

$$w_{ij} = 1 - s_{ij},\ w_{ji} = 1 - s_{ji}. \tag{23}$$

2) Contrastive Loss Formulation:

For the EEG-to-EOG direction, the contrastive loss is formulated as:

$$\mathcal{L}_{E\to O} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\mathrm{sim}\left(\boldsymbol{f}_i^e, \boldsymbol{f}_i^o\right)/\tau\right)}{\exp\left(\mathrm{sim}\left(\boldsymbol{f}_i^e, \boldsymbol{f}_i^o\right)/\tau\right) + \sum_{\substack{j=1\\ j\neq i}}^{B} w_{ij}\exp\left(\mathrm{sim}\left(\boldsymbol{f}_i^e, \boldsymbol{f}_j^o\right)/\tau\right)}. \tag{24}$$

For the EOG-to-EEG direction, the contrastive loss is formulated as:

$$\mathcal{L}_{O\to E} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\mathrm{sim}\left(\boldsymbol{f}_i^o, \boldsymbol{f}_i^e\right)/\tau\right)}{\exp\left(\mathrm{sim}\left(\boldsymbol{f}_i^o, \boldsymbol{f}_i^e\right)/\tau\right) + \sum_{\substack{j=1\\ j\neq i}}^{B} w_{ji}\exp\left(\mathrm{sim}\left(\boldsymbol{f}_i^o, \boldsymbol{f}_j^e\right)/\tau\right)}. \tag{25}$$

where τ is the temperature scalar.

3) State-Aware Cross-Modal Contrastive Loss:

The final SA-CL objective is defined as:

$$\mathcal{L}_{CL} = \frac{1}{2}\left(\mathcal{L}_{E\to O} + \mathcal{L}_{O\to E}\right). \tag{26}$$

Through the above design, SA-CL not only preserves the representation consistency between EEG and EOG from the same temporal segment, but also alleviates the influence of falsely treating semantically similar non-paired samples as negative pairs. As a result, the proposed method learns cross-modal high-level representations that better conform to the continuous semantic structure of vigilance states.

2.6. Regression Model and Overall Objective

To ensure that the features extracted by the EEG and EOG encoders effectively contribute to vigilance estimation, we introduce three supervised regression branches: an EEG-only branch, an EOG-only branch, and a fusion branch. The regression process for the EEG branch is formulated as:

$$\hat{y}_i^e = \tanh\left(\text{Linear}_e\left(\boldsymbol{f}_i^e\right)\right), \tag{27}$$

$$\mathcal{L}_E = \text{MSE}\left(\hat{y}_i^e, y_i\right), \tag{28}$$

where $\hat{y}_i^e$ denotes the continuous vigilance predicted from the EEG representation $\boldsymbol{f}_i^e$, y_i is the ground-truth vigilance label for the i-th sample, and $\mathcal{L}_E$ represents the mean squared error (MSE) loss for the EEG branch.

Similarly, the regression process for the EOG branch is defined as:

$$\hat{y}_i^o = \tanh\left(\text{Linear}_o\left(\boldsymbol{f}_i^o\right)\right), \tag{29}$$

$$\mathcal{L}_O = \text{MSE}\left(\hat{y}_i^o, y_i\right), \tag{30}$$

where $\hat{y}_i^o$ is the vigilance predicted from the EOG representation $\boldsymbol{f}_i^o$, and $\mathcal{L}_O$ is the corresponding MSE loss.

For the fusion branch, the modality attention mechanism is first used to obtain the fused representation $\boldsymbol{f}_i^F$, which is then fed into a regression layer:

$$\hat{y}_i^F = \tanh\left(\text{Linear}_F\left(\boldsymbol{f}_i^F\right)\right), \tag{31}$$

$$\mathcal{L}_F = \text{MSE}\left(\hat{y}_i^F, y_i\right), \tag{32}$$

where $\hat{y}_i^F$ is the vigilance predicted from the fused representation, and $\mathcal{L}_F$ is its MSE loss.

Finally, the overall loss function combines the regression losses and the previously defined state-aware cross-modal contrastive loss $\mathcal{L}_{CL}$:

$$\mathcal{L} = \mathcal{L}_E + \mathcal{L}_O + \mathcal{L}_F + \mathcal{L}_{CL}. \tag{33}$$

Finally, the overall loss function combines the EEG regression loss, EOG regression loss, fusion regression loss, and the state-aware cross-modal contrastive loss:

$$\mathcal{L} = \lambda_E\mathcal{L}_E + \lambda_O\mathcal{L}_O + \lambda_F\mathcal{L}_F + \lambda_{CL}\mathcal{L}_{CL}. \tag{34}$$

Here, λ_E, λ_O, λ_F, and λ_{CL} denote the weighting coefficients of the EEG,

EOG, fusion, and contrastive losses, respectively. In this study, they are set to $\lambda_E = \lambda_O = \lambda_F = \lambda_{CL} = 1$ in all experiments. This equal-weight setting avoids introducing additional tuning hyperparameters and allows the supervised regression objectives and the cross-modal alignment constraint to be jointly optimized.

3. Results

3.1. Dataset

Experiments and analyses were conducted on the SEED-VIG dataset [3], which was provided by the Brain-like Computing and Machine Intelligence (BCMI) Center of Shanghai Jiao Tong University. The dataset was collected using a simulated driving system designed for vigilance estimation. It contains EEG, EOG, and eye-tracking data collected in a simulated driving environment. During the experiment, a four-lane highway scene was displayed on a large LCD screen in front of the subjects, and the vehicle movements were controlled using a steering wheel and gas pedal. The road was primarily straight and monotonous to induce fatigue more easily. Most experiments were conducted in the early afternoon after lunch, and each experimental session lasted approximately 2 h.

During data acquisition, EEG and EOG signals were recorded using the Neuroscan system. Specifically, the EEG features used in this study were extracted from 17 channels mainly distributed over the temporal, parietal, and occipital regions. The temporal-related channels included FT7, FT8, T7, T8, TP7, and TP8, while the parietal/occipital-related channels included CP1, CP2, P1, PZ, P2, PO3, POZ, PO4, O1, OZ, and O2. Accordingly, the EEG input dimensionality in the EEG branch is set to $N = 17$ electrodes. In addition, 36-dimensional eye-movement-related EOG features were extracted for each temporal sample to characterize eye activity variations. The vigilance labels were calculated using the PERCLOS indicator derived from SMI eye-tracking glasses, and the continuous PERCLOS values range from 0 to 1. The electrode placements of EEG and EOG are illustrated in Figure 2.

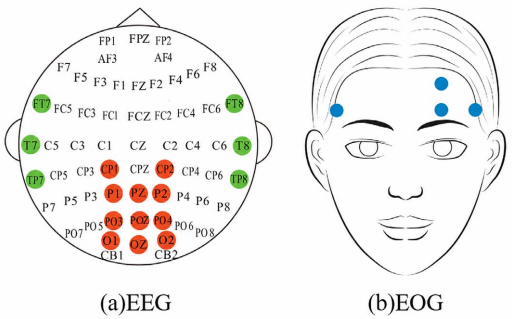

Figure 2. Electrode placements for the EEG and EOG setups.

Furthermore, the SEED-VIG dataset adopted a relatively rigorous fatigue annotation strategy based on eye-tracking techniques. During the experiments, SMI eye-tracking glasses were used to record eye-movement information. The eye

states were categorized into fixation, blinking, and saccade behaviors, and the duration of each state was simultaneously recorded. Among these states, fixation and saccade behaviors generally indicate relatively alert or normal driving conditions, whereas blinking and prolonged eye closure are closely associated with fatigue or vigilance decline. It should be noted that the SMI eye-tracking glasses could not directly detect slow eye closure or prolonged eye closure states, namely the CLOS state. Therefore, the dataset combines blinking and CLOS information to compute the PERCLOS index during annotation. PERCLOS represents the proportion of time in which the subject's eyes remain in closed or fatigue-related states within a given temporal window and is widely used as an important continuous indicator for driver fatigue and vigilance variation:

$$\mathrm{PERCLOS} = \frac{T_{\mathrm{blink}} + T_{\mathrm{CLOS}}}{T_{\mathrm{blink}} + T_{\mathrm{fixation}} + T_{\mathrm{saccade}} + T_{\mathrm{CLOS}}}. \tag{35}$$

For each training sample, the vigilance target y_i is assigned based on the PERCLOS value computed over the same 8 s interval as the temporally aligned EEG and EOG segments. Specifically, the i-th sample corresponds to the interval $\left[8(i-1), 8i\right)$ s. Within this interval, the durations of blink, fixation, saccade, and CLOS states are accumulated and substituted into Equation (35). The EOG features are extracted from the same 8 s interval, while the corresponding EEG input is the multi-time-step DE sequence constructed from the $T = 7$ internal subwindows defined in Section 2.1. Therefore, the sample-level EEG sequence, EOG feature vector, and PERCLOS label are temporally aligned at the 8 s segment level. The outer 8 s samples are non-overlapping; overlap exists only among the internal EEG DE sliding windows.

3.2. Evaluation Metrics

Root Mean Square Error (RMSE) and Correlation Coefficient (CORR) are two widely used evaluation metrics for continuous regression models. RMSE measures the difference between the predicted values and the ground-truth labels, which is defined as follows:

$$\mathrm{RMSE}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}, \tag{36}$$

where $\boldsymbol{Y} = \left\{y_i\right\}_{i=1}^{N}$ denotes the ground-truth labels, $\hat{\boldsymbol{Y}} = \left\{\hat{y}_i\right\}_{i=1}^{N}$ denotes the predicted values, and N is the number of samples.

Since RMSE alone cannot provide structural relationship information between predictions and ground truth, the Correlation Coefficient (CORR) is further adopted to complement RMSE. CORR evaluates the linear relationship between predicted values and ground-truth labels, thereby reflecting the consistency of their variation trends. The Pearson correlation coefficient is defined as follows:

$$\mathrm{CORR}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = \frac{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)\left(\hat{y}_i - \bar{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}\sqrt{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)^2}}, \tag{37}$$

where $\bar{y}$ and $\bar{\hat{y}}$ denote the mean values of $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$, respectively.

A cross-validation strategy was adopted to evaluate the model performance. Specifically, the data of each subject were divided into five subsets according to temporal order. In each round, four subsets were used for training, and the remaining subset was used for testing.

However, CORR is sensitive to short temporal segments and is more suitable for long-term evaluation. Therefore, the predicted values and ground-truth labels from the five folds were concatenated together, and the overall CORR and RMSE were computed as the final evaluation metrics. In general, a more accurate model is expected to achieve a higher CORR and a lower RMSE.

3.3. Implementation Details and Statistical Analysis

For reproducibility, the main hyperparameters were fixed as follows unless otherwise stated: the EOG latent dimension was set to $R = 64$; the GAT employed four attention heads with a 32-dimensional projection per head; the GRU hidden size was 128; the modality-attention projection dimension was 128; the contrastive projection dimension was 64; the temperature in the contrastive loss was set to $\tau = 0.07$; and the vigilance-state similarity scale was set to $\sigma = 0.05$. The model was optimized using Adam with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}. The batch size was 64, and each fold was trained for 100 epochs.

For paired model comparisons conducted on the same 21 subjects and the same temporal splits, statistical significance was assessed via two-sided Wilcoxon signed-rank tests on subject-level CORR and RMSE values. The significance level was set at $p < 0.05$, and the Holm-Bonferroni correction was applied when multiple paired comparisons were performed. For comparisons with previously published methods where subject-level predictions were unavailable, we report descriptive average-performance comparisons rather than paired significance tests.

3.4. Performance Comparison

To evaluate the effectiveness of the proposed model, comparisons were conducted with several state-of-the-art methods, including Support Vector Regression (SVR), Continuous Conditional Neural Fields (CCNF), Continuous Conditional Random Fields (CCRF), Graph-Regularized Extreme Learning Machine (GELM), Long Short-Term Memory (LSTM), and Deep Neural Networks with Subnetwork Neurons (DNNSN). In addition, single-modality EEG-only and EOG-only models were adopted as baseline methods.

Table 2 presents the performance comparison between the proposed method, existing advanced approaches, and the two single-modality baseline models. The reported CORR and RMSE values are the average results over all 21 subjects. The proposed method achieves the highest average CORR and a competitive RMSE, supporting the effectiveness of the EEG-EOG collaborative modeling framework and the state-aware cross-modal contrastive learning strategy for continuous vigilance estimation. For the in-house baselines evaluated under identical data splits,

paired statistical comparisons were further conducted on the subject-level CORR and RMSE values. Compared with the EEG-only baseline, the proposed method achieved a significant improvement in CORR ($p = 0.008$) and a significant reduction in RMSE ($p = 0.021$). Compared with the EOG-only baseline, the proposed method also significantly improved CORR ($p < 0.001$) and reduced RMSE ($p = 0.014$). These results indicate that the performance gains of the proposed model are not only reflected in the average metrics, but are also statistically reliable across subjects.

Table 2. Performance comparison with existing methods.

Ref.	Method	CORR	RMSE
[3]	SVR	0.830	0.100
[4]	GELM	0.808	0.072
[3]	CCRF	0.840	0.100
[3]	CCNF	0.845	0.095
[5]	LSTM	0.830 ± 0.101	**0.081 ± 0.014**
[7]	DNNSN	0.850	0.090
Baseline	EOG-only	0.776 ± 0.183	0.112 ± 0.050
Baseline	EEG-only	0.866 ± 0.107	0.107 ± 0.045
Proposed	Multimodal	**0.920 ± 0.051**	0.083 ± 0.032

3.5. Ablation Experiments

To verify the effectiveness of the proposed State-Aware Cross-Modal Contrastive Learning (SA-CL) strategy for continuous vigilance estimation, ablation experiments were conducted in this study. In all experiments, the compared models adopted the same backbone network architecture, input features, data partition strategy, and major training parameters. The only difference was whether the state-aware cross-modal contrastive learning constraint was introduced. Specifically, the baseline model without state-aware contrastive learning was compared with the complete model equipped with SA-CL to analyze the impact of this module on multimodal representation learning and regression prediction performance.

Figure 3 presents the comparison results of CORR and RMSE between the baseline model and SA-CL on 10 representative subjects, while the statistical test was conducted on all 21 subjects using the paired protocol described above. From the CORR results, introducing state-aware cross-modal contrastive learning improves the correlation performance for most subjects, indicating that SA-CL enhances the consistency between the predicted vigilance trends and the ground-truth vigilance variations. From the RMSE results, SA-CL achieves lower prediction errors for most subjects, demonstrating that the proposed strategy not only improves trend consistency but also enhances the numerical accuracy of continuous vigilance estimation.

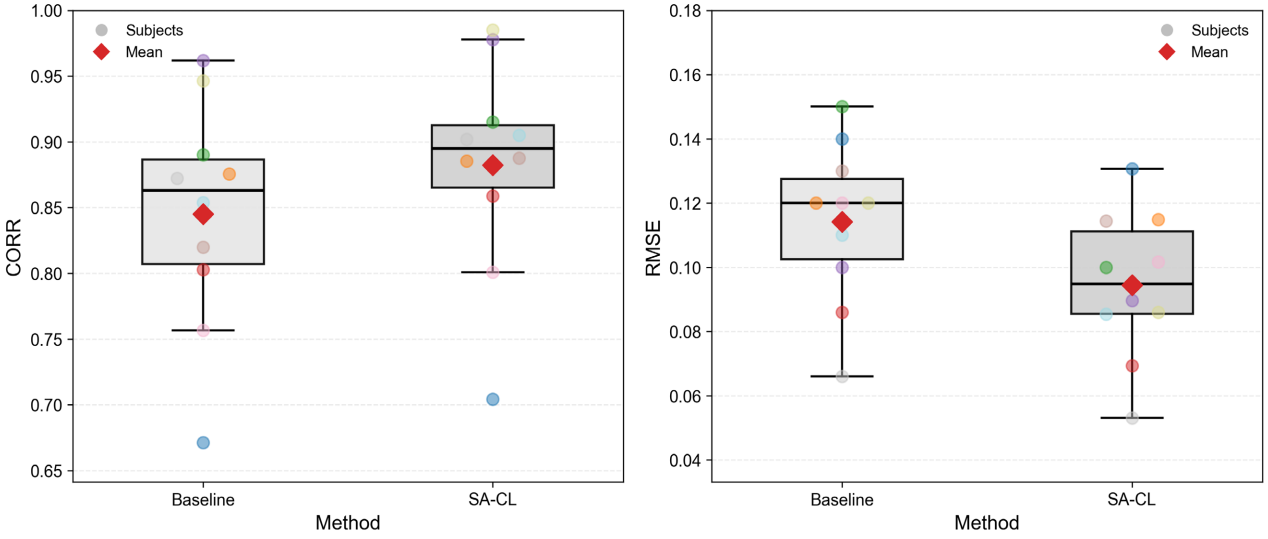

Figure 3. CORR and RMSE of the fusion model with and without SA-CL on 10 representative subjects.

To quantitatively verify the effectiveness of SA-CL, paired statistical comparisons were conducted between the baseline model without SA-CL and the complete model with SA-CL using the subject-level results of all 21 subjects. The complete model achieved a higher average CORR than the baseline model, and the improvement was statistically significant ($p = 0.006$). In terms of RMSE, the complete model also reduced the prediction error compared with the baseline model, with the difference reaching statistical significance ($p = 0.018$). These results demonstrate that the proposed SA-CL module provides a statistically reliable improvement for continuous vigilance estimation.

Further analysis reveals that the baseline model mainly relies on EEG and EOG feature extraction together with fusion-based regression for vigilance estimation. Although such a framework can exploit the complementary information between the two modalities, it lacks explicit constraints on cross-modal high-level representation consistency, resulting in insufficient representation alignment. In contrast, SA-CL introduces vigilance-state similarity among samples to adaptively weight the negative effects of non-paired samples, thereby alleviating the problem of incorrectly pushing apart semantically similar samples. Consequently, the model can learn cross-modal representations that better conform to the continuous semantic structure of vigilance states and provide a more effective feature foundation for subsequent fusion regression.

To further investigate the influence of SA-CL on EEG and EOG feature representations, t-SNE visualization was performed on the high-level features of a representative subject (Subject ID: 7), as illustrated in Figure 4. Vigilance states were divided into three levels, namely alert, fatigue, and drowsiness, with thresholds of 0.35 and 0.7. Compared with the baseline model, the EEG features (green) and EOG features (orange) learned by SA-CL exhibit significantly more compact clustering in the representation space. Point clouds from different modalities under

the same vigilance state are distributed much closer to each other, indicating better cross-modal alignment. Meanwhile, smoother continuous transitions between different vigilance states can also be observed. In contrast, the feature distribution of the baseline model appears relatively scattered, and the boundaries among vigilance states are less distinguishable. These observations demonstrate that SA-CL not only enhances semantic alignment between EEG and EOG but also improves the discriminative capability of multimodal representations while preserving the intrinsic continuous structure of vigilance states.

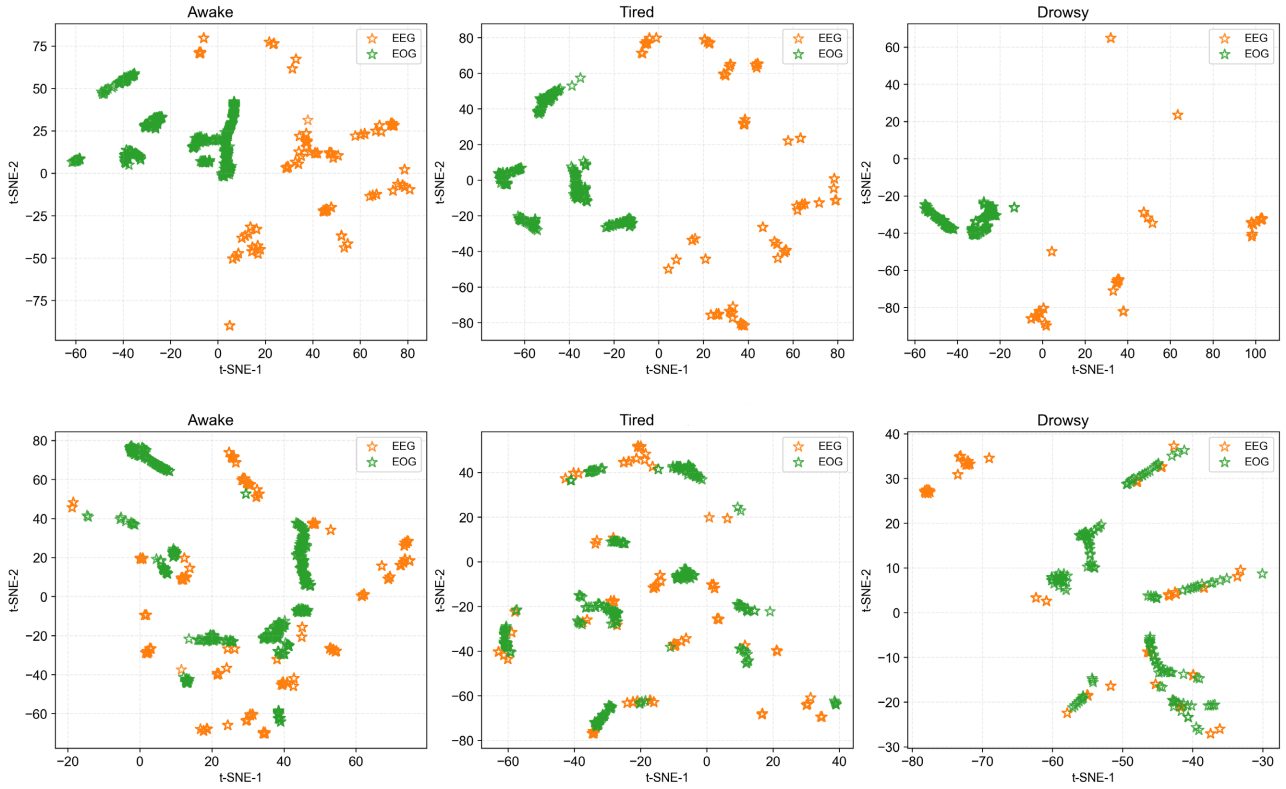

Figure 4. Without SA-CL (1st row) vs. with SA-CL (2nd row): t-SNE results of EEG and EOG high-level features extracted from the trained encoders.

In addition, the continuous-label feature visualizations of EEG and EOG representations for both the baseline model and SA-CL are illustrated in **Figure 5**, where different colors denote different vigilance levels. From **Figure 4** and **Figure 5**, it can be concluded that SA-CL effectively aligns features from different modalities while maintaining the separability of different vigilance states. Specifically, EEG and EOG features exhibit highly overlapping distributions in the representation space and form continuous transitions along with changes in fatigue labels. These results further indicate that SA-CL not only strengthens semantic alignment across modalities, but also preserves the intrinsic structure of continuous vigilance states, thereby improving the discriminability and consistency of multimodal representations for continuous vigilance estimation tasks.

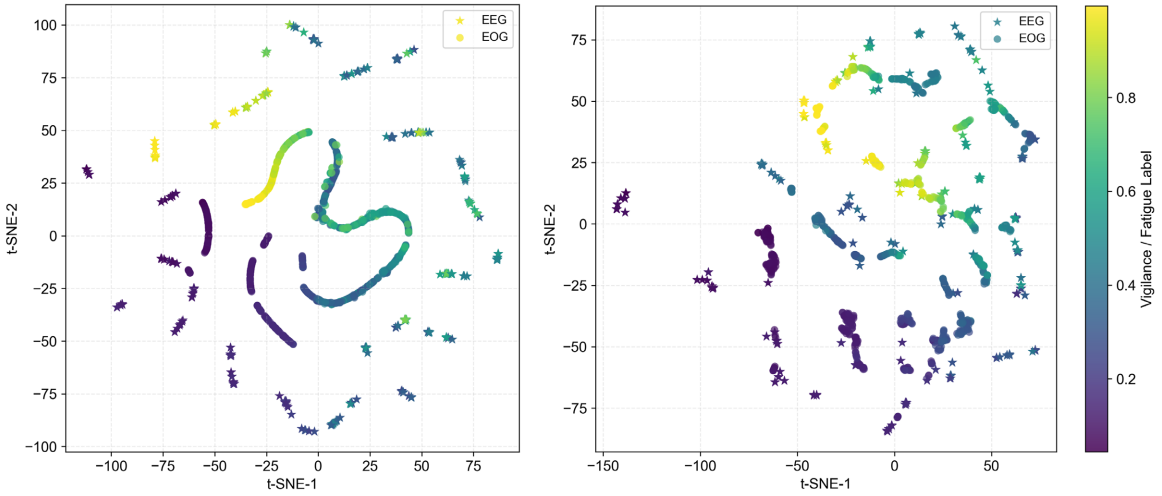

Figure 5. SNE results of EEG and EOG high-level features extracted from the trained encoders without SA-CL and with SA-CL under continuous vigilance labels.

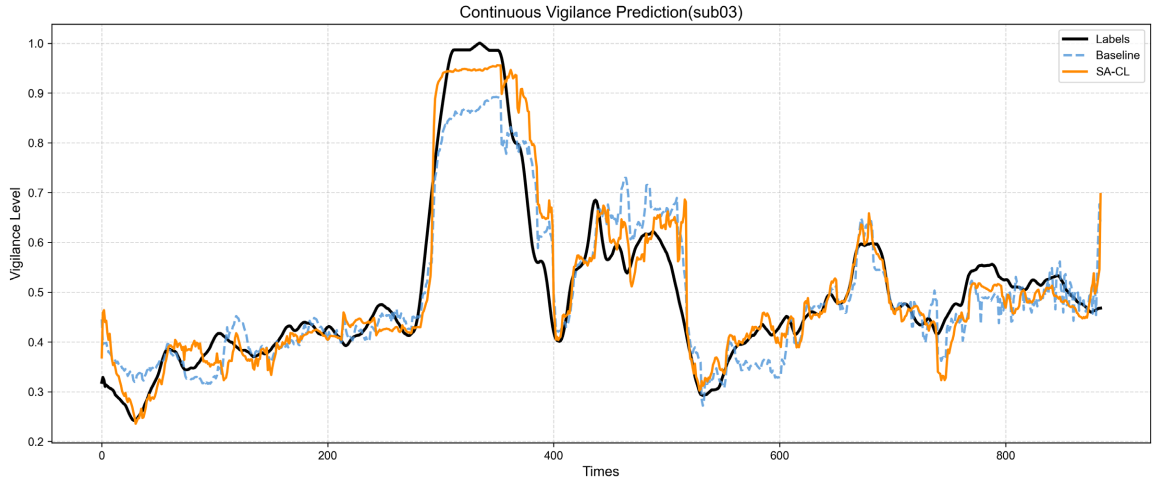

Figure 6. Continuous vigilance estimation curve for Subject 3.

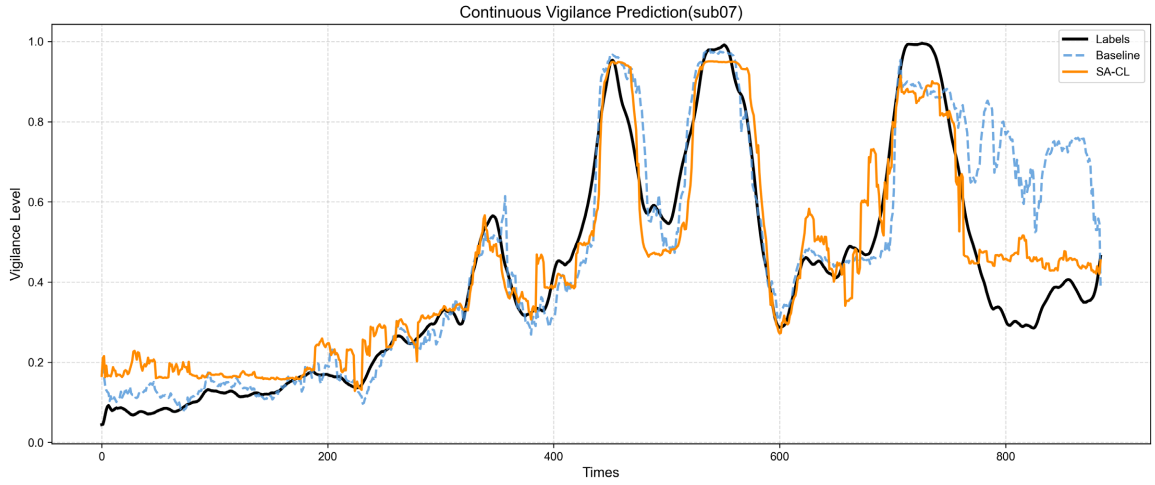

Figure 7. Continuous vigilance estimation curve for Subject 7.

To further verify whether the vigilance variations predicted by SA-CL are consistent with the ground-truth changes, the continuous PERCLOS label curves and the corresponding prediction curves of the proposed method for two representative subjects (Subject IDs: 3 and 7) are plotted in **Figure 6** and **Figure 7**. It can be observed that the proposed method can accurately predict the PERCLOS values, and the predicted curves exhibit highly consistent trends with the ground-truth vigilance variation curves.

4. Discussion

In recent years, physiological signals have been widely used for driver vigilance and cognitive state estimation. Foong *et al.* [26] used EEG frequency-domain features for driver fatigue detection, while Ji *et al.* [27] extracted EOG-related features such as blink frequency and eye-closure duration to identify drowsiness. Although these single-modality methods are effective for local feature extraction, they cannot fully exploit complementary information across modalities or capture the dynamic changes of continuous vigilance states.

To improve multimodal representation, previous studies have fused EEG with other physiological or behavioral signals. For example, Shahbakhti *et al.* [28] combined EEG and facial expression features for fatigue detection, and Zheng *et al.* [29] proposed an attention-based multimodal fusion framework for EEG, EOG, and related signals. However, most existing methods focus on feature-level fusion and lack explicit cross-modal alignment constraints, which may lead to inconsistent feature distributions and unclear state boundaries.

To address this issue, this study proposes State-Aware Cross-Modal Contrastive Learning (SA-CL). SA-CL aligns high-level EEG and EOG representations by maximizing the similarity of multimodal features from the same temporal segment while reducing modality discrepancies. This explicit alignment improves cross-modal consistency and maps features into a normalized representation space, helping the model better distinguish different vigilance states and capture smooth state transitions.

Experimental results demonstrate the effectiveness of SA-CL. Compared with single-modality models and multimodal baselines, SA-CL achieves the best average CORR and competitive RMSE performance under the same subject-wise evaluation protocol. In addition, t-SNE visualizations show compact EEG-EOG feature distributions and continuous transitions in the representation space, indicating that SA-CL can effectively model the dynamic evolution of vigilance states.

Despite its effectiveness, several limitations remain. First, the SEED-VIG dataset includes 23 experiments from 21 subjects in a simulated driving environment, so the robustness and generalization of SA-CL still need validation on larger-scale and real-world datasets. Second, although SA-CL improves prediction performance, the physiological interpretability of its learned representations remains limited. Future work should further explore interpretable mechanisms to better support practical driver monitoring applications.

5. Conclusions

In this paper, a State-Aware Cross-Modal Contrastive Learning (SA-CL) strategy was proposed for continuous driver vigilance estimation. The proposed method aims to improve the accuracy and consistency of continuous vigilance prediction by enhancing the semantic alignment between multimodal EEG and EOG representations. Experimental results demonstrate that SA-CL achieves superior CORR and competitive RMSE performance on the SEED-VIG dataset compared with both baseline models and existing state-of-the-art methods. In addition, the proposed method can accurately capture vigilance variation trends in continuous PERCLOS prediction tasks.

Further analyses reveal that SA-CL not only effectively aligns high-level representations across modalities but also preserves the intrinsic structure of continuous vigilance states, thereby improving the discriminability and continuity consistency of multimodal representations. Ablation studies and t-SNE visualizations further verify the effectiveness of the proposed strategy in feature alignment and continuous state modeling.

Overall, this work provides an effective new framework for multimodal continuous state estimation and offers valuable insights for applications such as driver fatigue monitoring, cognitive state assessment, and brain-computer interfaces. Future work will focus on expanding the dataset scale, integrating additional physiological and behavioral modalities, and exploring the deployment of the proposed framework in real-time online environments to further improve its generalization capability and practical applicability.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Khessiba, S., Blaiech, A.G., Ben Khalifa, K., Ben Abdallah, A. and Bedoui, M.H. (2021) Innovative Deep Learning Models for EEG-Based Vigilance Detection. *Neural Computing and Applications*, **33**, 6921-6937. https://doi.org/10.1007/s00521-020-05467-5

[2] Pan, J., Cai, X., Mo, D., Yu, Y. and Li, Y. (2023) Residual Attention Capsule Network for Multimodal EEG- and EOG-Based Driver Vigilance Estimation. *IEEE Transactions on Instrumentation and Measurement*, **72**, 1-12. https://doi.org/10.1109/tim.2023.3307756

[3] Zheng, W.L. and Lu, B.L. (2017) A Multimodal Approach to Estimating Vigilance Using EEG and Forehead EOG. *Journal of Neural Engineering*, **14**, Article 026017. https://doi.org/10.1088/1741-2552/aa5a98

[4] Huo, X.Q., Zheng, W.L. and Lu, B.L. (2016) Driving Fatigue Detection with Fusion of EEG and Forehead EOG. 2016 *International Joint Conference on Neural Networks* (*IJCNN*), Vancouver, 24-29 July 2016, 897-904. https://doi.org/10.1109/ijcnn.2016.7727294

[5] Zhang, N., Zheng, W.L., Liu, W. and Lu, B.L. (2016) Continuous Vigilance Estimation Using LSTM Neural Networks. In: *Lecture Notes in Computer Science*, Springer, 530-537. https://doi.org/10.1007/978-3-319-46672-9_59

[6] Zhang, G. and Etemad, A. (2021) Capsule Attention for Multimodal EEG-EOG Representation Learning with Application to Driver Vigilance Estimation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **29**, 1138-1149. https://doi.org/10.1109/tnsre.2021.3089594

[7] Wu, W., Wu, Q.M.J., Sun, W., Yang, Y., Yuan, X., Zheng, W., *et al.* (2021) A Regression Method with Subnetwork Neurons for Vigilance Estimation Using EOG and EEG. *IEEE Transactions on Cognitive and Developmental Systems*, **13**, 209-222. https://doi.org/10.1109/tcds.2018.2889223

[8] Gao, P., Liu, T., Liu, J.W., Lu, B.L. and Zheng, W.L. (2024) Multimodal Multi-View Spectral-Spatial-Temporal Masked Autoencoder for Self-Supervised Emotion Recognition. 2024 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Seoul, 14-19 April 2024, 1926-1930. https://doi.org/10.1109/icassp48485.2024.10447194

[9] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A.Y. (2011) Multimodal Deep Learning. *Proceedings of the* 28*th International Conference on Machine Learning* (*ICML*), Bellevue, 28 June 2011, 689-696.

[10] Atrey, P.K., Hossain, M.A., El Saddik, A. and Kankanhalli, M.S. (2010) Multimodal Fusion for Multimedia Analysis: A Survey. *Multimedia Systems*, **16**, 345-379. https://doi.org/10.1007/s00530-010-0182-0

[11] Tsai, Y.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L. and Salakhutdinov, R. (2019) Multimodal Transformer for Unaligned Multimodal Language Sequences. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, July 2019, 6558-6569. https://doi.org/10.18653/v1/p19-1656

[12] Pan, J. and Lu, D. (2024) A Deep Channel Attention Transformer for Multimodal EEG-EOG-Based Vigilance Estimation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, **46**, 3235-3241.

[13] Cheng, X., Wei, W., Du, C., Qiu, S., Tian, S., Ma, X., *et al.* (2022). VigilanceNet: Decouple Intra- and Inter-Modality Learning for Multimodal Vigilance Estimation in RSVP-Based BCI. *Proceedings of the* 30*th ACM International Conference on Multimedia*, Lisboa, 10-14 October 2022, 209-217. https://doi.org/10.1145/3503161.3548367

[14] Chen, T., Kornblith, S., Norouzi, M. and Hinton, G. (2020) A Simple Framework for Contrastive Learning of Visual Representations. *Proceedings of the* 37*th International Conference on Machine Learning* (*ICML*), Online, 13-18 July 2020, 1597-1607.

[15] He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R. (2020) Momentum Contrast for Unsupervised Visual Representation Learning. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 9726-9735. https://doi.org/10.1109/cvpr42600.2020.00975

[16] Gao, T., Yao, X. and Chen, D. (2021) SimCSE: Simple Contrastive Learning of Sentence Embeddings. *Proceedings of the* 2021 *Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, November 2021, 6894-6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

[17] Baevski, A., Zhou, H., Mohamed, A. and Auli, M. (2020) wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. 34*th Conference on Neural Information Processing Systems*, Vancouver, 6-12 December 2020, 1-12.

[18] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., *et al.* (2021) Learning Transferable Visual Models from Natural Language Supervision. *Proceedings of the* 38*th International Conference on Machine Learning* (*ICML*), Online, 18-24 July 2021, 8748-8763.

[19] Zhang, M., Luo, Z., Xie, L., Liu, T., Yan, Y., Yao, D., *et al.* (2024) Multimodal Vigilance

Estimation with Modality-Pairwise Contrastive Loss. *IEEE Transactions on Biomedical Engineering*, **71**, 1139-1150. https://doi.org/10.1109/tbme.2023.3328942

[20] Liu, Y. and Zhang, G. (2026) MHCL: Multi-Modal Hierarchical Contrastive Learning for Physiological Signal-Based Vigilance Detection. *IEEE Journal of Biomedical and Health Informatics*, 1-13. https://doi.org/10.1109/jbhi.2026.3695698

[21] Zha, K., Cao, P., Son, J., Yang, Y. and Katabi, D. (2023) Rank-N-Contrast: Learning Continuous Representations for Regression. *Advances in Neural Information Processing Systems* 36, New Orleans, 10-16 December 2023, 17882-17903. https://doi.org/10.52202/075280-0786

[22] Jin, X., Wang, J., Liu, L. and Lin, Y. (2023) Time-Series Contrastive Learning against False Negatives and Class Imbalance.

[23] Shi, L.C., Jiao, Y.Y. and Lu, B.L. (2013) Differential Entropy Feature for EEG-Based Vigilance Estimation. 2013 35*th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (*EMBC*), Osaka, 3-7 July 2013, 6627-6630. https://doi.org/10.1109/embc.2013.6611075

[24] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P. and Bengio, Y. (2018) Graph Attention Networks. *Proceedings of the International Conference on Learning Representations* (*ICLR*), Vancouver, 30 April-3 May 2018.

[25] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., *et al.* (2014) Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, October 2014, 1724-1734. https://doi.org/10.3115/v1/d14-1179

[26] Foong, R., Ang, K.K., Quek, C., Guan, C. and Phyo Wai, A.A. (2015) An Analysis on Driver Drowsiness Based on Reaction Time and EEG Band Power. 2015 37*th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (*EMBC*), Milan, 25-29 August 2015, 7982-7985. https://doi.org/10.1109/embc.2015.7320244

[27] Ji, Q., Zhu, Z. and Lan, P. (2004) Real-Time Nonintrusive Monitoring and Prediction of Driver Fatigue. *IEEE Transactions on Vehicular Technology*, **53**, 1052-1068. https://doi.org/10.1109/tvt.2004.830974

[28] Shahbakhti, M., Beiramvand, M., Nasiri, E., Far, S.M., Chen, W., Solé-Casals, J., *et al.* (2023) Fusion of EEG and Eye Blink Analysis for Detection of Driver Fatigue. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **31**, 2037-2046. https://doi.org/10.1109/tnsre.2023.3267114

[29] Zheng, W.L., Liu, W., Lu, Y., Lu, B.L. and Cichocki, A. (2019) EmotionMeter: A Multimodal Framework for Recognizing Human Emotions. *IEEE Transactions on Cybernetics*, **49**, 1110-1122. https://doi.org/10.1109/tcyb.2018.2797176