SAGE: Subject-Adaptive Graph-Based Modeling with Decision-Level Calibration for Stress and Cognitive Workload Monitoring

Chenpei Xie*, Yiting Wei*, Mostafa Haghi, Nima TaheriNejad

***Abstract*— In real-world driving scenarios, reliable monitoring of drivers' stress and cognitive workload is critical for driving safety. Both tasks can be characterized using multimodal physiological signals; however, due to pronounced inter-subject variability, decision instability becomes particularly prominent under limited calibration data. This paper proposes a subject-adaptive graph-based modeling with decision-level calibration (SAGE) framework, which supports both driver stress and cognitive workload estimation under a unified modeling and decision paradigm. SAGE learns shared physiological representations with cross-subject generalization capability and extends subject adaptation from the representation level to the decision level. By incorporating sample-quality-aware calibration and threshold adaptation, SAGE enables reliable personalized adaptation using only a small amount of labeled calibration data. Under a strict leave-one-subject-out (LOSO) evaluation protocol, we validate the proposed method on two datasets with distinct feature distribution characteristics. Experimental results show that SAGE achieves a classification accuracy of 80.6% for stress recognition on the AffectiveROAD and 80.5% for cognitive workload recognition on the hciLab, significantly outperforming representative existing methods. Under random-split settings, SAGE further achieves classification accuracies of 91.8% on AffectiveROAD and 89.6% on hciLab Driving. These results suggest that, under conditions of subject heterogeneity and unconstrained signal quality, combining multimodal physiological representations with decision-level adaptive mechanisms provides an effective approach for stable driver state modeling. Furthermore, validation on the SWELL Knowledge Work (SWELL-KW) dataset suggests that the proposed framework captures transferable physiological representations that generalize across different tasks and real-world scenarios.**

***Index Terms*— Driver state monitoring, subject adaptation, decision-level calibration, multimodal physiological signals, stress recognition, cognitive workload estimation**

I. Introduction

ROAD traffic crashes remain a major global public health challenge, causing approximately 1.19 million deaths each year and resulting in 20–50 million non-fatal injuries [1]. In most cases, crashes are not caused by mechanical failures but arise from human errors when the driver's internal state no longer matches the demands of the driving task [2]. Psychological stress and cognitive workload are particularly critical for drivers because they directly influence attention, information processing, and vehicle control behavior [3].

To address this issue, many driver monitoring systems attempt to estimate the driver's cognitive–affective state. Performance-based indicators, such as lane-keeping quality, steering variability, or reaction times, typically change only after vehicle control has already been affected [4]. Vision-based methods infer driver state from facial expressions, eye movements, and body posture; however, their performance is easily affected by occlusion, illumination conditions, and camera setup. Although recent advances in low-light enhancement and diffusion-based image restoration have improved visual quality under degraded imaging conditions, reliable performance across diverse real-world environments remains challenging [5], [6]. Self-report scales and secondary-task probes provide more direct access to subjective experience but are difficult to use continuously in safety-critical settings [7]. These limitations have driven increasing interest in objective physiological monitoring, which can reflect autonomic and central nervous system activity and reveal subtle changes in internal states before they manifest in overt behavior [8].

At the physiological level, psychological stress and cognitive workload are driven by partially overlapping but functionally distinct central regulatory mechanisms. Psychological stress primarily activates the hypothalamic–pituitary–adrenal (HPA) axis and the sympathetic branch of the autonomic nervous system, enabling rapid adaptation to acute situations [9]. In contrast, cognitive workload arises from sustained attentional and executive control demands, involving prolonged engagement of cortical control networks, particularly the fronto-parietal systems [10]. These central regulatory mechanisms are reflected through multiple peripheral physiological pathways. Electrodermal activity (EDA) primarily indexes sympathetic arousal [11], while blood volume pulse (BVP) and heart rate (HR) capture cardiovascular responses to autonomic modulation [12]. In addition, accelerometry (ACC) data provides information about behavioral responses associated with postural adjustments and subtle motor movements [13]. As these signals capture complementary aspects of physiological and behavioral responses, relying on any single modality provides only a limited characterization of the internal state. Effectively

* Chenpei Xie and Yiting Wei contributed equally to this work. Corresponding author: Yiting Wei (email: yiting.wei@uni-heidelberg.de)
All Authors are with the Heidelberg University, Heidelberg, 69120.

We would like to thank Hector Stiftung for partially funding this work.

integrating information from multiple sources is therefore essential for obtaining a more comprehensive representation of stress and cognitive workload [14]. This underscores the importance of learning shared representations from complementary physiological signals to obtain a more comprehensive characterization of stress and cognitive workload [15]. Graph-based modeling provides a natural and effective way to capture complex and dynamic dependencies among physiological modalities, enabling explicit cross-modal interactions [16].

In this paper, we propose SAGE, a graph-based framework for multimodal physiological driver state modeling. The core contributions of this work are as follows: **a)** A subject-adaptive graph-based physiological modeling framework for cross-subject stress and cognitive workload estimation. **b)** A dynamic-adjacency heterogeneous physiological graph network that learns sample-specific cross-modal physiological interactions through adaptive graph construction and message passing. **c)** A decision-level personalized calibration mechanism that enables data-efficient and stable subject adaptation without high-dimensional parameter fine-tuning. **d)** A unified, task-agnostic framework that supports the simultaneous estimation of stress and cognitive workload across different driver-state monitoring tasks, while also generalizing to stress assessment in office knowledge-work environments.

II. Related Work

Recent attention-based models have demonstrated strong capability in capturing long-range dependencies and enhancing feature representation learning, and cross-domain generalization [17], [18]. However, effectively modeling the complex cross-modal and temporal dependencies of multimodal physiological signals remains challenging in driver-state monitoring scenarios [19]. At the subject-adaptation level, existing methods often rely on subject-specific parameter tuning and substantial individual data, resulting in unstable adaptation to unseen drivers or low-sample settings [20]. In addition, stress and cognitive workload are often modeled independently, leading to task-specific architectures and feature designs that limit reuse across driver-state monitoring tasks [21]. Furthermore, physiological signal heterogeneity and the lack of explicit reliability and uncertainty modeling further undermine decision robustness in real-world driving [22]. Healey and Picard [23] demonstrated the feasibility of using electrocardiogram (ECG), electromyography (EMG), and skin conductance signals for stress classification in real-world driving, but their approach relies on handcrafted features and offline processing, making it difficult to model temporal dynamics and cross-signal dependencies. The representation and fusion of multimodal physiological signals have become a research focus. Lee *et al.* [24] introduced multimodal Convolutional Neural Networks (CNNs) to learn dynamic features from short-term physiological signals; however, implicit modality fusion can lead to performance degradation when modality quality fluctuates. Chen *et al.* [25] combined feature selection and transfer learning to improve cross-subject performance, but their method depends on large-scale, high-quality data and is limited under small-sample or noisy conditions. Existing studies have demonstrated the value of multimodal physiological signals for driver state monitoring, yet challenges remain in modeling cross-modal dependencies, achieving stable personalization, and ensuring robust decision-making. Therefore, a unified framework that effectively integrates multimodal physiological information is still needed.

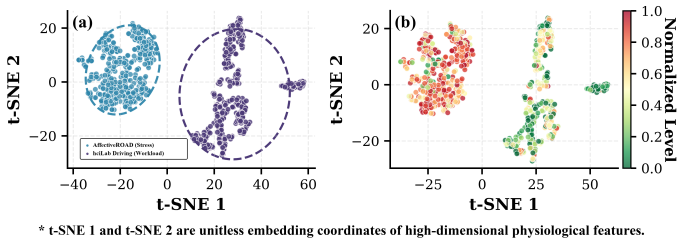

Fig. 1. 2D t-SNE visualization of the AffectiveROAD and hciLab Driving datasets: (a) Dataset distribution with 95% confidence ellipses, (b) Distribution colored by normalized stress/workload level.

III. Materials and Methods

A. Datasets

1) AffectiveROAD Dataset (Stress): The AffectiveROAD dataset was designed to capture driver stress in real-world scenarios using synchronized physiologicaland audiovisual recordings [26]. It comprises 13 driving experiments conducted by 9 unique participants along a 31 km metropolitan route, with an average session duration of 86 minutes, including scheduled rest periods. Physiological signals were collected using two Empatica E4 wristbands and a Zephyr BioHarness 3.0 chest belt. The E4 devices recorded electrodermal activity (EDA) and skin temperature at 4 Hz, HR at 1 Hz, and 3-axis hand movement at 36 Hz, while the BioHarness provided respiration-related and cardiac measurements at 1 Hz. Driver stress was annotated as a continuous subjective variable in the range $[0, 1]$. Stress labels were initially generated in real time during driving and subsequently validated by the driver through synchronized post-drive video review.

2) hciLab Driving Dataset (Cognitive Workload): The hciLab Driving dataset is a large-scale benchmark for assessing driver workload under naturalistic conditions [27]. It contains approximately 2.5 million samples collected from 10 participants driving their own vehicles along a 23.6 km route. The route includes heterogeneous traffic contexts, such as 30 km/h and 50 km/h urban segments, highways, and tunnels. Physiological data were recorded using a Nexus 4 biofeedback system. ECG signals were sampled at 1024 Hz, while derived and peripheral measures, including HR, heart rate variability (HRV), skin conductance response (SCR) and skin temperature (TEMP), were processed at 128 Hz. Ground-truth workload annotations were obtained retrospectively. After each drive, participants provided time-aligned workload ratings on a 0–128 scale while reviewing synchronized trip videos.

3) Feature Space and Domain Comparison: To analyze the distribution differences between stress and cognitive workload, we project the physiological features from the AffectiveROAD and hciLab datasets into a shared two-dimensional t-SNE

space. As shown in Fig. 1(a), samples from the two datasets form clearly separated regions, indicating substantial distribution shifts between the two tasks. After dataset-wise min–max normalization, Fig. 1(b) shows that both states exhibit continuous distributions, suggesting that the physiological feature space preserves information about internal state variations. These observations indicate that stress and cognitive workload exhibit both continuous state structures and significant task-specific distribution differences, motivating the need for a unified framework capable of handling distribution shifts while maintaining stable decision behavior.

B. Data Preprocessing

1) Signal Preprocessing: Dataset-specific preprocessing was performed prior to window construction. For the AffectiveROAD dataset, excessively short, and temporally discontinuous recordings were removed. The EDA, TEMP, BVP, HR, and ACC signals were retained, and percentile clipping was applied to BVP. For the hciLab Driving dataset, records with invalid timestamps were removed and missing values were imputed via linear interpolation. The same five physiological signals as those used in the AffectiveROAD dataset were adopted. The SCR signal, treated as the EDA channel, was filtered using a fourth-order zero-phase Butterworth low-pass filter (2.0 Hz cutoff frequency) to remove high-frequency noise, followed by Savitzky–Golay smoothing (approximately 2.5 s window, second-order polynomial) to reduce local random fluctuations while preserving signal peak characteristics, and three-level db4 wavelet denoising to further suppress non-stationary noise and motion artifacts. The TEMP signal was smoothed using a Savitzky–Golay filter (approximately 3 s window, second-order polynomial), and the ECG signal was filtered using a fourth-order zero-phase Butterworth low-pass filter (40 Hz). The HR and ACC signals were retained in their original form. All signal channels were clipped to the 1st–99th percentile range prior to window construction.

2) Handcrafted Feature Extraction: For each window analysis, we extract generic statistical descriptors from all available physiological channels, together with domain-specific EDA phasic SCR, HRV features, and temperature features when applicable. All handcrafted features are summarized in Table I.

3) Alignment and Segmentation: To enable multimodal modeling without imposing resampling, we retain each signal at its native sampling rate after dataset-specific preprocessing and align modalities at the timestamp and window-boundary level. Windows are retained only when the required modalities provide valid samples within the same time interval. Windows containing NaN values, empty segments, or insufficient modality coverage are discarded. Aligned signals are segmented into overlapping fixed-duration windows that serve as the basic modeling units. For both AffectiveROAD and hciLab Driving, we use a 15 s window with a 7.5 s step, corresponding to 50% overlap. Each modality contributes its available windows in the 15 s interval according to its own sampling rate. A 15 s window length was adopted because it preserves the completeness of multimodal physiological information while meeting real-time monitoring requirements, thereby providing a favorable trade-off between feature representativeness and temporal responsiveness [28]. A 50% overlap was adopted to balance temporal continuity, sample coverage, and computational efficiency. This setting mitigates window-boundary effects, preserves transitional physiological dynamics, and increases the number of effective training samples without additional data collection. To avoid data leakage introduced by the 50% window overlap, the continuous signal of the left-out subject was first partitioned into two non-overlapping segments, with 20% allocated to the calibration set and 80% allocated to the test set. Window segmentation was subsequently performed independently within each set.

TABLE I
HANDCRAFTED FEATURES EXTRACTED FROM BOTH DATASETS.

Feature type	Graph node	Dataset	Source	Extracted features
Statistical	Statistical nodes: *Source*_stat	AR / HCI	EDA, TEMP, BVP / ECG, HR, ACC	Mean, std, min, max, skew, kurt
EDA phasic (SCR)	Derived nodes: EDA_SCR	AR / HCI	EDA	No. of peaks, mean/max amplitude, rise, recovery (+ sum amplitude for HCI)
HRV	Derived nodes: IBI_HRV	AR	BVP	SDNN, RMSSD, pNN50, SD1, SD2, CVNN, IBI
		HCI	HRV_LF	Mean, std, min, max, skew, kurt
TEMP dynamic	Derived nodes: TEMP_dyn	AR / HCI	TEMP	Per-window temperature change (ΔT)

AR refers to the AffectiveROAD dataset; HCI refers to the hciLab Driving dataset. BVP is available in the AR dataset, while ECG originates from the HCI dataset. Each signal in the *Source* column is processed independently, and the corresponding features listed in the *Extracted features* column are extracted.

4) Labeling method: Window labels are derived from subject-specific extreme ratings to account for inter-subject scale differences. For each subject, we first compute the mean rating for all windows in the training set and then calculate the quantiles based on these mean ratings to assign labels. Specifically, the mean ratings of all windows in the training set are used to determine subject-specific quantiles, with the 20th and 80th percentiles (P20/P80) used for both AffectiveROAD and hciLab Driving. Windows with mean ratings below the lower quantile are assigned a negative label 0, and windows with mean ratings above the upper quantile are assigned a positive label 1, while mid-range windows are discarded. Quantile-based thresholds are preferred over fixed cutoffs as they normalize subject-dependent baselines and yield more consistent class definitions across individuals. Importantly, all quantile statistics are computed solely on the training windows (i.e., thresholds are determined within the training portion of each subject's split) to prevent information leakage, and ambiguous mid-range windows are excluded to reduce label noise near the decision boundary.

5) Standardization: We apply z-score standardization to each node feature vector: $x_{\text{norm}} = \frac{x-\mu}{\sigma+\epsilon}$, where μ and σ denote the mean and standard deviation estimated for the corresponding node feature from the training subjects, and $\epsilon = 10^{-6}$ prevents division by zero. Standardization parameters are estimated from the training subjects.

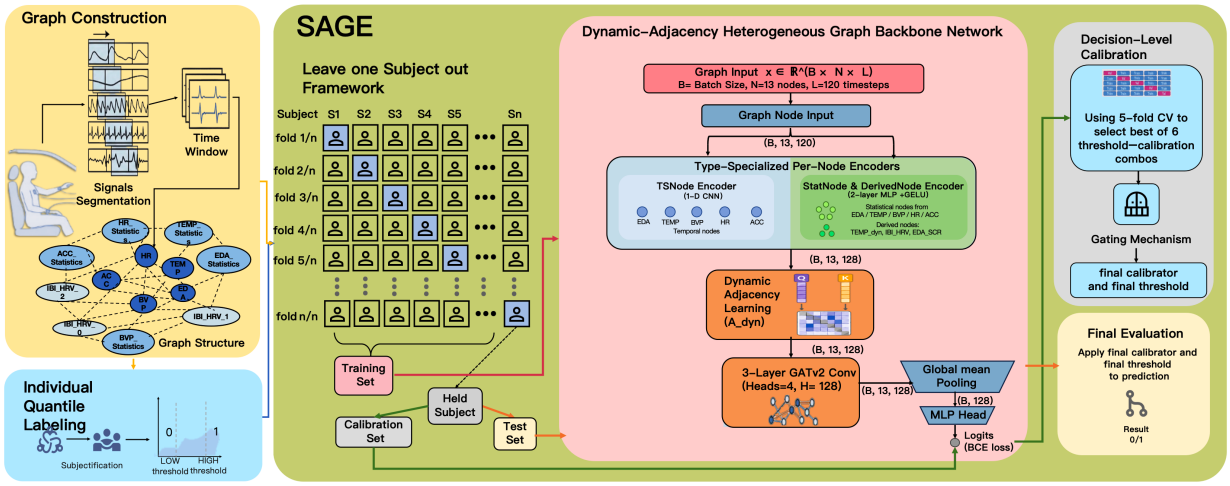

Fig. 2. Overall framework of the proposed subject-adaptive graph-based modeling with decision-level calibration method.

C. Model Architecture

1) Graph Construction: Driver state recognition benefits from jointly leveraging multiple physiological modalities whose interactions are physiologically coupled. Physiological modalities exhibit structured dependencies mediated by shared autonomic mechanisms. Conventional multi-channel fusion encodes each modality independently and concatenates features, failing to explicitly model these structured priors. Therefore, graph representations provide a natural way to model cross-modal physiological interactions rather than treating modalities as independent features. We construct each window as a heterogeneous physiological graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the topology encodes modality-level physiological couplings.

a) Node set definition: The node set $\mathcal{V}$ contains N=13 nodes organized into three categories that capture complementary physiological levels. In our framework, each node represents a physiological representation, whose input may consist of either a temporal signal segment or a feature vector. The first category consists of five raw signal nodes (EDA_ts, TEMP_ts, BVP_ts, HR_ts, and ACC_ts), each temporal node is constructed from the complete time-series segment of the corresponding physiological modality collected within the analysis window. The second category consists of five statistical nodes (EDA_stat, TEMP_stat, BVP_stat, HR_stat, and ACC_stat). Each statistical node is constructed from a modality-specific statistical feature vector computed within the analysis window, capturing the distributional characteristics of the corresponding physiological signal. The detailed features are summarized in Table I. The third category consists of three derived nodes (TEMP_dyn, IBI_HRV, EDA_SCR). Each derived node is constructed from a modality-specific feature vector obtained through domain-specific physiological signal processing and is used as the node input. The detailed feature definitions are summarized in Table I.

b) Edge definition: Each edge represents a potential dependency between two physiological representations. Unlike conventional physiological graphs that rely on manually predefined connectivity patterns, the proposed graph adopts a dynamic topology, where edge connections are generated from the learnable similarities among node representations within each analysis window. This enables the graph structure to adapt to subject-specific and time-varying physiological interactions. These dependencies may arise from physiological mechanisms such as autonomic regulation, cardiovascular coupling, thermoregulation, and cross-modal behavioral responses. By dynamically modeling inter-node relationships, the graph captures evolving physiological coupling patterns across subjects and analysis windows rather than relying on static connectivity priors. To maintain sparsity and interpretability, each node is connected only to a limited number of its most relevant neighbors. Self-loops are not regarded as physiological relationships and are introduced only as computational connections when required. The detailed dynamic edge generation process is described in Section III-C.2.b.

2) Dynamic-Adjacency Heterogeneous Physiological Graph Network: After constructing the heterogeneous physiological graph, we employ a dynamic-adjacency heterogeneous physiological graph network, as illustrated in Fig. 2. The network first models modality-specific temporal dynamics through node-specific encoders and subsequently learns cross-modal physiological interactions via dynamic graph construction and graph-attention message passing. By explicitly separating temporal modeling from cross-modal fusion, the network preserves modality-specific temporal characteristics while effectively capturing dependencies across physiological modalities.

a) Type-Specialized Per-Node Encoders: For each temporal node, a 1D-convolutional neural network (CNN) encoder captures local within-window patterns. Adaptive pooling subsequently maps the variable-length per-modality sequences to a fixed H-dimensional embedding. For each statistical and derived node, a node-specific multilayer perceptron (MLP) encoder projects its compact descriptor vector into the same H-dimensional latent space; statistical and derived nodes share the MLP encoder architecture but use independently parameterized instances, maintaining their distinction as node categories in the graph. Keeping 1D-CNN and MLP as two parallel encoder families is necessary because raw temporal waveforms and pre-computed descriptors encode information

at fundamentally different abstraction levels and should not be forced to share a single encoder. For each node input $\mathbf{x}_i$, the corresponding encoder produces a node embedding: $\mathbf{h}_i = f_i(\mathbf{x}_i)$, $\mathbf{h}_i \in \mathbb{R}^H$, where $f_i(\cdot)$ denotes the node-specific encoder associated with node i, and $\mathbf{h}_i$ is the resulting H-dimensional node embedding. The resulting 13 node embeddings are stacked to form the graph embedding tensor: $\mathbf{X}_\mathcal{G} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N] \in \mathbb{R}^{B\times N\times H} = \mathbb{R}^{B\times 13\times 128}$, where B denotes the batch size, $N = 13$ denotes the number of graph nodes, and H denotes the hidden embedding dimension. The graph embedding tensor $\mathbf{X}_\mathcal{G}$ serves as the input to the subsequent dynamic graph module.

b) Dynamic edge generation: After node-specific encoding, all physiological nodes are represented in the same hidden space. Given the encoded node embeddings $\mathbf{X}_\mathcal{G} \in \mathbb{R}^{B\times N\times H}$, query and key projections are computed as $\mathbf{Q} = \mathbf{X}_\mathcal{G}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}_\mathcal{G}\mathbf{W}_K$. The dense adjacency scores are obtained by scaled dot-product similarity: $\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{H}}\right)$, where the softmax is applied over candidate source nodes. We then apply top-k sparsification to retain the strongest $k = 4$ neighbors for each node. The resulting active edge set $\mathcal{E}_{\mathrm{dyn}}$ is used by the subsequent graph attention network v2 (GATv2) layers for attention-based message passing.

c) Graph-attention message passing: Cross-modal dependencies are modeled after node encoding. The sparse dynamic graph is passed to a three-layer GATv2 backbone. For each layer, attention is computed only over active edges selected by the dynamic adjacency module, allowing the model to assign different importance to different physiological relations in each window. Residual connections and layer normalization are used after each graph-attention layer to stabilize training:

$$\mathbf{H}^{(l+1)} = \mathrm{LayerNorm}\left(\mathrm{GELU}\left(\mathrm{GATv2}^{(l)}(\mathbf{H}^{(l)}, \mathcal{E}_{\mathrm{dyn}})\right)\right) + \mathbf{H}^{(l)} \tag{1}$$

Stacking three graph-attention layers enables multi-hop cross-modal message passing while preserving a sparse and interpretable dynamic relation structure. After three graph-attention layers, the final node representations $\mathbf{H}^{(L)}$ are used for graph-level pooling.

d) Graph pooling and classification head: After graph-attention propagation, node embeddings are aggregated into a window-level graph representation by global mean pooling over the 13 nodes: $\mathbf{g} = \mathrm{MeanPool}\big(\mathbf{H}^{(L)}\big) \in \mathbb{R}^H$. The graph representation is then passed to a shallow MLP head to produce one scalar logit per window: $\ell = f_{\mathrm{cls}}(\mathbf{g})$. The logit ℓ is used as the base neural prediction score before subject adaptation, ensemble stacking, and decision-level calibration.

3) Decision-Level Calibration and Configuration Gating: Raw model outputs are often poorly calibrated, meaning that predicted probabilities do not necessarily reflect the true likelihood of correct predictions. Since discriminative training primarily optimizes class separability rather than probability estimation accuracy, models tend to produce overconfident predictions. Under subject-specific distribution shifts and class imbalance, a fixed global threshold (e.g., 0.5) may therefore fail to achieve optimal decision performance. To address this issue, we propose a decision-level calibration and configuration gating framework, as shown in Fig. 3, which leverages a small amount of calibration data to transform model outputs into calibrated probabilities and automatically select an appropriate decision configuration for the target subject. The framework explicitly accounts for three key factors in cross-subject physiological learning: distribution shifts, calibration uncertainty, and class-prior shifts. Therefore, it integrates three complementary mechanisms: subject-specific posterior estimation, implemented through Section III-C.3.a; prior-aware threshold adaptation, implemented through Section III-C.3.b; and sample-quality-aware calibrator selection, implemented through Section III-C.3.c. Together, these components improve the stability and reliability of predictions for unseen subjects. From a statistical perspective, the proposed gating mechanism implements an explicit bias–variance tradeoff. When calibration evidence is limited, the system favors more conservative configurations to reduce estimation variance. As calibration information becomes more reliable, it progressively adopts more adaptive configurations to reduce prediction bias and enhance personalization.

a) Calibrator design: The calibrator establishes a mapping from model score to probability. Temperature scaling learns a single parameter T to adjust mapping sharpness with minimal variance while preserving ranking; isotonic regression is a nonparametric monotonic piecewise mapping $p = g(\ell)$ that corrects nonlinear miscalibration but requires sufficient balanced samples. We retain both as candidates alongside a no-calibration option (raw sigmoid $p = \sigma(\ell)$) and let subsequent mechanisms perform data-driven selection.

b) Threshold strategy design: Given calibrated probability $\hat{p}$, we perform binary prediction $\hat{y} = \mathbb{I}[\hat{p} \geq t]$. We design two complementary threshold selection strategies: the aggressive strategy and the conservative strategy. The aggressive strategy performs fine-grained grid search, selecting the best threshold over $t \in [0, 1]$ to maximize the balanced objective $J(t) = 0.5 \cdot \mathrm{F1}(t) + 0.5 \cdot \mathrm{acc}(t)$, while ensuring a precision floor to avoid threshold collapse under sparse labels. The conservative strategy selects the best threshold from a predefined candidate threshold library, ensuring higher robustness under weak calibration evidence.

c) Configuration gating mechanism: Combining three calibrators (temperature, isotonic, none) with two threshold strategies yields six candidate decision configurations. Reliably selecting the optimal one from limited calibration data is challenging: aggressive data-driven selection may overfit to particular splits, while excessive conservatism cannot exploit personalization opportunities. We therefore design a dual-path gating mechanism. The aggressive path uses 5-fold cross-validation (CV) within the calibration set to evaluate all six configurations, selecting the one with the highest mean CV accuracy. The conservative path uses a preset robust configuration (isotonic or temperature scaling with the fixed threshold library). The gate decides which path to trust based on four reliability indicators: sample size n (estimation reliability), class proportion π (label informativeness), calibration-set Area Under the ROC Curve (AUROC) (subject-level discriminability), and CV accuracy standard deviation (selection stability). When information quality is insufficient, the gate directly adopts the conservative configuration to avoid overfitting;

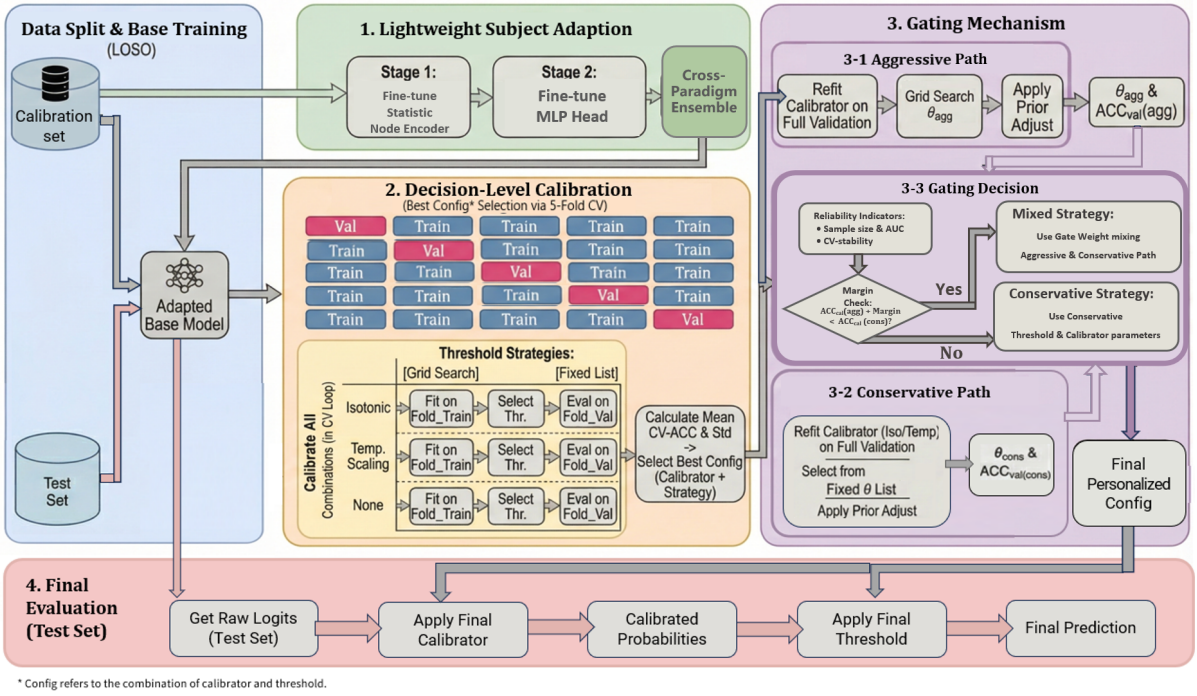

Fig. 3. Decision-level calibration and gating mechanism pipeline.

when it is sufficient but the conservative path outperforms the aggressive path on the calibration set, the gate again favors the conservative configuration to avoid unnecessary performance loss. Otherwise, the gate blends thresholds from both paths: $t^\star = \gamma\, t_{\text{aggr}} + (1-\gamma)\, t_{\text{cons}}$, where $\gamma = \text{clip}(2|\pi - 0.5|, 0, 1)$. This blending favors the conservative threshold under near-balanced class proportions and shifts toward the data-driven threshold as imbalance grows, since stronger imbalance indicates clearer subject-specific class structure that warrants more aggressive personalization.

In contrast to conventional threshold moving methods, our framework's main advantages can be summarized as follows: 1) it jointly evaluates multiple decision configurations formed by different combinations of calibrators and threshold-selection strategies; 2) it incorporates a reliability-aware gating mechanism to quantify the trustworthiness of calibration information; and 3) it automatically reverts to more conservative decision configurations when calibration evidence is limited or unreliable. Together, these mechanisms enable more robust personalized decision-making across subjects under limited calibration data. Existing feature-level adaptation and domain adaptation methods primarily focus on learning transferable representations to mitigate distribution mismatch. Our proposed framework explicitly models the reliability of subject-specific adaptation at the decision level through calibration-aware configuration selection, thereby addressing not only how adaptation is achieved but also when the adapted decision can be trusted.

4) SAGE End-to-End Procedure:

a) Data split and base training: We evaluate cross-subject generalization using the LOSO protocol. In each iteration, all subjects except the held-out subject are used for training. Data from the held-out subject are further split into a calibration set (20%) and an independent test set. No test samples or labels are accessed during model adaptation, configuration selection, or hyperparameter tuning. The test set remains isolated throughout training and calibration and is used only for final evaluation, preventing information leakage from the target subject. On the training set, we employ stratified 3-fold CV for model selection. In each fold, one split serves as validation and the remaining two as training. Each fold initializes the model parameters randomly and trains independently until convergence or early stopping based on validation F1 score. Among the three trained parameter sets, we select the one achieving highest validation F1 score as the base model. Training minimizes binary cross-entropy loss with AdamW optimization: $\mathcal{L} = -\frac{1}{M}\sum_{m=1}^{M}[y_m \log p_m + (1-y_m)\log(1-p_m)]$, where $p_m = \sigma(\ell_m)$ and $\sigma(\cdot)$ is the sigmoid function. This procedure yields a base model encoding cross-subject general physiological-state mappings, without any held-out-subject-specific adaptation.

b) Lightweight subject adaptation: Directly applying the base model to a new subject suffers from individual baseline differences and session shifts. Calibration-set fine-tuning can mitigate this, but excessive tuning overfits on limited samples. Our strategy updates only parameters most directly related to the decision boundary and session bias, keeping the backbone feature extractor frozen. Adaptation is performed in two stages. First, only the statistical-node encoders are updated on the calibration set with a relatively large learning rate, while all other parameters remain frozen, enabling adaptation to subject-specific shifts in distributional physiological descriptors. Second, the statistical-node encoders are frozen and only the MLP classification head is updated using a smaller learning rate and early stopping to refine the decision boundary. Throughout both stages, the temporal-signal encoders, dynamic adjacency

module, GATv2 layers, normalization layers, and other backbone components remain frozen, preserving population-level physiological representations and reducing overfitting under limited calibration data.

c) Decision configuration selection: After adaptation, four classical learners are introduced alongside the adapted graph model to produce five prediction scores (logits). A logistic-regression stacking strategy then fuses these into a single ensemble score, which feeds the configuration-selection procedure. Test-set logits are generated by the same frozen pipeline, while test labels remain hidden throughout. Configuration selection proceeds entirely on the calibration set in three steps. First, 5-fold CV evaluates all six candidate configurations: each fits its calibrator and selects its threshold on fold-train and is scored on fold-val; the configuration with the highest mean accuracy across folds becomes the aggressive-path candidate. Second, because CV produces per-fold parameters that cannot be directly merged, both paths are refitted on the full calibration set to obtain their final calibrators, thresholds, and calibration-set accuracies. Third, the gate determines the final configuration based on reliability indicators and margin comparison: under weak evidence or when the conservative path outperforms the aggressive on the calibration set, both its calibrator and threshold are adopted; otherwise the aggressive-path calibrator is kept and the two thresholds are blended, yielding the final calibrator and threshold $t^\star$.

d) Test evaluation: The test set remains isolated throughout base training, adaptation, calibrator fitting, threshold search, and gating decisions. Final evaluation proceeds as follows: test-set logits (already obtained during adaptation) are transformed by the final calibrator into calibrated probabilities, calibrated probabilities are compared against $t^\star$ to yield binary predictions, and predictions are compared with ground-truth test labels to compute evaluation metrics. This protocol ensures unbiased evaluation: test labels are used only at the final metric-computation step, with all preceding decisions based solely on the calibration set.

D. Performance Evaluation

We report complementary metrics to characterize classification performance from different perspectives, including Accuracy, F1 score, AUROC, Area Under the Precision-Recall Curve (AUPRC), Precision, and Recall. It is worth noting that for LOSO evaluation, metrics are first computed within each held-out subject and then averaged across subjects, giving each subject equal weight. Accordingly, the reported F1 score is the mean of subject-wise F1 scores and is not necessarily equal to the harmonic mean of the averaged precision and recall. Statistical significance is assessed using the Wilcoxon signed-rank test on per-subject accuracy, paired by subject; the test direction and reference method are specified in the corresponding table footnote.

IV. Results and Discussion

A. Overall Performance

Deployment-Oriented Cross-Subject Performance: Under the strict LOSO protocol, SAGE achieves stable performance on both the AffectiveROAD and hciLab Driving datasets, as shown in Table II. Compared with existing related work and baselines, SAGE demonstrates the best performance to date under the strict LOSO evaluation protocol. These results indicate that SAGE is capable of reliable cross-subject generalization under realistic deployment constraints.

Benefits of Dynamic Physiological Interaction Modeling: Compared with representative static graph variants, the proposed dynamic graph consistently achieves the best performance across datasets, highlighting the importance of modeling sample-specific physiological interactions, as shown in Table VI. Graph visualization shows that the most influential connections mainly occur across physiological modalities rather than within individual signals. These results suggest that physiological states are better characterized by cross-modal interactions, which are effectively captured by the proposed framework for robust physiological-state estimation.

The Importance of Decision-Level Adaptation: As shown in Tables VIII and X, the performance gains of SAGE under the strict LOSO protocol primarily arise from decision-level adaptation. Decision-level calibration and reliability-aware gating consistently improve both dataset-level and subject-level performance across AffectiveROAD and hciLab Driving. Beyond improving average accuracy, these mechanisms reduce performance disparities across subjects, leading to more stable and reliable predictions under realistic cross-subject deployment conditions.

Consistent Performance Gains Across Tasks and Scenarios: Despite differences in label characteristics and physiological mechanisms between stress recognition and cognitive workload tasks, SAGE achieves consistent and stable performance improvements across both tasks, as shown in Table II. Experiments on the SWELL-KW office knowledge-work stress dataset further validate the effectiveness and generalization capability of the proposed method under a different application scenario, as shown in Table XI. These results indicate that the proposed method exhibits good cross-task applicability, cross-scenario generalization capability and reuse potential.

Robustness and Real-World Deployability: The proposed framework exhibits strong robustness and real-world deployability. Stable performance under modality-masking and noise-perturbation analyses suggests that the learned physiological representations remain reliable under sensing imperfections (Table VII). Combined with its lightweight architecture and millisecond-level inference latency, these properties make the framework well suited for real-time physiological monitoring applications (Table XII).

B. LOSO Evaluation and Random-Split Comparison

To assess cross-subject generalization under realistic deployment conditions, we evaluated SAGE on the AffectiveROAD and hciLab datasets using the strict LOSO protocol. As shown in Table II, under this setting, SAGE achieved an Accuracy of 80.6%, an F1 score of 0.82, an AUROC of 0.872, and an AUPRC of 0.917 on the AffectiveROAD stress-recognition task, with Precision and Recall reaching 0.754 and 0.946, respectively. On the hciLab Driving cognitive

TABLE II
COMPARISON WITH EXISTING METHODS ON BOTH DATASETS.

Method	Accuracy (%)	F1	AUROC	AUPRC	Precision	Recall	Protocol
AffectiveROAD Dataset							
SVM single-task [29]	70.0	0.72	–	–	–	–	10-fold CV
MT-MKL [29]	83.0	0.87	–	–	–	–	10-fold CV
TPCFL [30]	73.0	–	–	–	–	–	LOOCV
Random Forest [31]	74.1	–	–	–	–	–	LOSO
Dense ANN 2L [32]	83.1	0.85	–	–	–	–	random split
ResNet Conv1D [32]	71.5	0.67	–	–	–	–	random split
CNN	52.5	0.61	–	–	–	–	LOSO
LSTM	56.0	0.68	–	–	–	–	LOSO
SAGE (ours)	**91.8**	**0.91**	**0.978**	0.979	0.878	0.974	**random split**
SAGE (ours)	**80.6**	**0.82**	**0.872**	0.917	0.754	0.946	**LOSO**
hciLab Driving Dataset							
SVM single-task [29]	64.0	0.58	–	–	–	–	10-fold CV
Multi-task RBF [29]	71.0	0.48	–	–	–	–	10-fold CV
CNN	60.7	0.22	–	–	–	–	LOSO
LSTM	62.9	0.30	–	–	–	–	LOSO
SAGE (ours)	**89.6**	**0.88**	**0.929**	0.893	0.835	0.935	**random split**
SAGE (ours)	**80.5**	**0.77**	**0.862**	0.835	0.758	0.802	**LOSO**

TABLE III
STATISTICAL SUMMARY OF ACCURACY ACROSS DIFFERENT EVALUATION PROTOCOLS.

Dataset	Protocol	Mean (%)	Std	CV (%)	95% CI	Wilcoxon p
AffectiveROAD	No Calib.	65.70	8.72	13.28	[60.23, 70.79]	—
	Random Split	91.83	10.48	11.41	[84.66, 97.73]	0.008^{**}
	LOSO (Ours)	80.63	8.33	10.33	[75.51, 85.82]	0.004^{**}
hciLab Driving	No Calib.	56.91	8.58	15.08	[52.24, 62.08]	—
	Random Split	89.57	8.02	8.95	[84.38, 94.42]	0.004^{**}
	LOSO (Ours)	80.46	8.21	10.20	[75.56, 85.35]	0.002^{**}

p values are from Wilcoxon signed-rank tests on per-subject Accuracy against the no-calibration LOSO baseline within each dataset, paired by subject, two-sided. $^{**}p < 0.01$, $^{*}p < 0.05$; — denotes the reference row.

TABLE IV
PER-SUBJECT ACCURACY UNDER LOSO AND RANDOM-SPLIT.

AffectiveROAD			*hciLab Driving*		
Subject	**LOSO (%)**	**Random split (%)**	**Subject**	**LOSO (%)**	**Random split (%)**
D1	77.28	96.00	P1	76.19	100.00
D2	67.09	100.00	P2	71.63	75.00
S_Drv3	81.89	—	P3	79.56	93.75
S_Drv4	74.91	75.00	P4	65.86	86.67
S_Drv5	93.59	81.82	P5	80.00	85.71
S_Drv7	73.27	81.82	P6	82.96	90.00
D8	88.69	100.00	P7	86.00	100.00
S_Drv9	86.72	100.00	P8	89.56	91.67
S_Drv11	82.21	100.00	P9	79.17	83.33
			P10	93.68	—
Mean	**80.63**	**91.83**	**Mean**	**80.46**	**89.57**

"—" in the random-split column denotes a degenerate subject-level split in which the test subset contains only one class; such cases are excluded from the random-split mean.

TABLE V
COMPARISON WITH REPRESENTATIVE BASELINE ON BOTH DATASETS.

Dataset	Method	Accuracy (%)	F1	AUROC	AUPRC	Precision	Recall	Wilcoxon p
Affective-ROAD	CNN+LSTM	71.2	0.74	0.761	0.732	0.676	0.834	0.0020^{**}
	GCN	73.8	0.77	0.798	0.774	0.697	0.876	0.0039^{**}
	DANN	70.3	0.74	0.774	0.759	0.662	0.856	0.0020^{**}
	Transformer	71.0	0.75	0.782	0.764	0.671	0.860	0.0020^{**}
	SAGE (ours)	**80.6**	**0.82**	**0.872**	**0.917**	**0.754**	**0.946**	—
hciLab Driving	CNN+LSTM	68.5	0.66	0.718	0.696	0.666	0.705	0.0010^{***}
	GCN	74.6	0.73	0.790	0.745	0.681	0.799	0.0029^{**}
	DANN	78.4	0.75	0.839	0.796	0.726	**0.810**	0.2158
	Transformer	69.3	0.67	0.716	0.693	0.685	0.712	0.0010^{***}
	SAGE (ours)	**80.5**	**0.77**	**0.862**	**0.835**	**0.758**	0.802	—

P-values were obtained from Wilcoxon signed-rank tests on per-subject Accuracy. ($n = 9$ for AffectiveROAD, $n = 10$ for hciLab Driving). $^{**}p < 0.01$, $^{*}p < 0.05$; — denotes the proposed reference row.

TABLE VI
GRAPH CONSTRUCTION ABLATION ACROSS TWO DATASETS.

Dataset	Graph	Accuracy (%)	F1	AUROC
AffectiveROAD	Static-Uniform	78.12	0.8076	0.8446
	Static-Correlation	78.52	0.8104	0.8454
	Static-MI	79.08	0.8137	0.8507
	Dynamic (ours)	**80.63**	**0.8218**	**0.8720**
hciLab Driving	Static-Uniform	73.54	0.7295	0.7877
	Static-Correlation	74.72	0.7263	0.8061
	Static-MI	78.02	**0.7813**	0.8286
	Dynamic (ours)	**80.46**	0.7722	**0.8619**

Static-Uniform, Static-Correlation, and Static-MI use fixed edge weights from uniform, Pearson correlation, and mutual information, respectively.

workload recognition task, SAGE attained an Accuracy of 80.5%, an F1 score of 0.77, an AUROC of 0.862, and an AUPRC of 0.835, with corresponding Precision and Recall values of 0.758 and 0.802, respectively. Meanwhile, under the random-split protocol, SAGE achieves the best performance on both datasets, attaining accuracies of 91.8% and 89.6% on AffectiveROAD and hciLab, respectively, substantially outperforming all competing methods.

To further analyze the performance of SAGE under different evaluation protocols, we compared the subject-level performance distributions obtained under the no-calibration baseline, random-split, and LOSO settings, as summarized in Table III. The no-calibration baseline refers to applying the trained model to the held-out subject without using any subject-specific calibration samples. From a statistical perspective, both random-split and LOSO yielded significant improvements over the no-calibration baseline on both datasets (Wilcoxon signed-rank test, all $p < 0.01$). Moreover, the standard deviations under LOSO were comparable to, or even lower than, those under random-split, indicating stable performance across different subjects. The bootstrap 95% confidence intervals further suggest that the observed performance gains were consistently achieved across the majority of participants rather than being driven by a small number of high-performing subjects. Table IV further reports the per-subject accuracies under the LOSO and random-split protocols. Although some performance variability remains across subjects, SAGE achieves satisfactory predictive performance for the majority of participants under the LOSO setting, without exhibiting any obvious failure cases. Importantly, the overall performance is not driven by a small number of high-performing individuals, but remains relatively consistent across different subjects. These results further indicate that the proposed framework can effectively accommodate inter-subject physiological variability while maintaining stable performance on previously unseen individuals. To sum up, while random-split yields higher average performance, the competitive and stable results achieved by SAGE under the stricter LOSO protocol demonstrate its ability to leverage limited target-subject data for personalized adaptation without sacrificing cross-subject generalization.

C. Comparison with Existing Methods and Baselines

Table II summarizes the performance comparison between SAGE and representative existing methods on both datasets. It should be noted that most prior studies were evaluated under 10-fold CV or random splits. Such evaluation settings may be affected by temporal correlations or subject-level data leakage, potentially leading to overly optimistic performance estimates. In contrast, all results reported for SAGE are obtained under the more rigorous LOSO protocol. Despite the substantially more challenging evaluation setting, SAGE achieves competitive results on both datasets. Under the same LOSO evaluation protocol, SAGE achieved the best overall performance across all evaluation metrics on both datasets. Under the random-split protocol, SAGE achieves accuracies of 91.8% and 89.6% on AffectiveROAD and hciLab, respectively, attaining the best overall performance. This suggests that, when subject-level distribution shifts between the training and test sets are reduced, the proposed framework can fully exploit its multimodal physiological-state modeling capability and achieve superior recognition performance.

Further, we selected several representative state-of-the-art methods for comparison. To ensure a fair evaluation, all models were assessed under the same LOSO protocol, using the same decision-level calibration pipeline and the same calibration-set ratio. Therefore, the only varying factor in this experiment was the modeling architecture itself. As shown in Table V, SAGE achieved the best overall performance on both datasets. These results indicate that, even when all methods benefit from the same calibration and adaptation mechanisms, the proposed dynamic physiological graph representation is able to learn more discriminative cross-modal physiological features, thereby providing additional performance gains. Wilcoxon signed-rank tests further validated these findings. On the AffectiveROAD dataset, SAGE achieved statistically significant improvements over all competing methods ($p < 0.01$ for all comparisons). On the hciLab Driving dataset, all comparisons except Domain-Adversarial Neural Network (DANN) also reached statistical significance ($p < 0.01$). Notably, although SAGE achieved the best performance in terms of Accuracy, F1 score, and AUROC, its improvement over DANN did not reach statistical significance ($p = 0.2158$). One possible explanation is that DANN explicitly performs domain adaptation by learning domain-invariant representations, thereby mitigating a substantial portion of the cross-subject distribution shift. Consequently, compared with conventional discriminative models that do not explicitly account for subject variability, the additional performance gains achieved by the proposed framework over DANN are naturally more limited. This effect is particularly evident in cognitive workload recognition, where physiological responses exhibit greater inter-individual variability and heterogeneity across subjects. As a result, the relative advantages of different adaptation-based methods may not be consistently manifested across all subjects, leading to increased performance variance and reduced statistical power for detecting performance differences among strong competing approaches. Taken together, these findings highlight a remaining challenge beyond distribution alignment: modeling persistent subject-specific physiological characteristics that are not fully captured by short-term calibration and adaptation mechanisms. Therefore, future work may explore incorporating long-term subject-specific physiological characteristics and richer contextual information to better account for inter-individual variability, thereby further enhancing the robustness and generalizability of adaptation-based cognitive workload recognition systems.

TABLE VII
FEATURE-GROUP ABLATION AND ROBUSTNESS TEST ON BOTH DATASETS UNDER THE LOSO PROTOCOL.

Dataset	Condition	AUROC	Accuracy (%)	ΔAUROC
AffectiveROAD	Complete (full)	0.8720	80.63	—
	Mask EDA	0.8543	78.20	−0.0177
	Mask TEMP	0.8387	76.91	−0.0333
	Mask BVP	0.8689	80.17	−0.0031
	Mask HR	0.8634	79.91	−0.0086
	Mask ACC	0.8145	74.92	−0.0575
	Mask HRV	0.8676	79.87	−0.0044
	Noise (SNR = 10 dB)	0.8601	79.76	−0.0119
	Noise (SNR = −1 dB)	0.8588	78.19	−0.0132
	Noise (SNR = −2 dB)	0.8457	76.76	−0.0263
	Noise (SNR = −3 dB)	0.8408	76.64	−0.0312
	Noise (SNR = −4 dB)	0.8246	75.90	−0.0474
	Noise (SNR = −5 dB)	0.8221	75.09	−0.0499
hciLab Driving	Complete (full)	0.8619	80.46	—
	Mask EDA	0.8206	78.02	−0.0413
	Mask BVP	0.8291	76.86	−0.0328
	Mask TEMP	0.8373	77.41	−0.0246
	Mask HR	0.8570	80.35	−0.0049
	Mask HRV	0.8545	78.71	−0.0074
	Mask ACC	0.8533	80.73	−0.0086
	Noise (SNR = 10 dB)	0.8518	78.86	−0.0101
	Noise (SNR = −1 dB)	0.7795	72.73	−0.0824
	Noise (SNR = −2 dB)	0.7702	72.69	−0.0917
	Noise (SNR = −3 dB)	0.7653	71.14	−0.0966
	Noise (SNR = −4 dB)	0.7549	70.08	−0.1070
	Noise (SNR = −5 dB)	0.7543	70.43	−0.1076

ΔAUROC is relative to the complete model.

D. Dynamic Graph Validation and Interpretation

1) Validation of Dynamic Graph Construction: To validate the effectiveness of the proposed dynamic graph strategy, we compared dynamic graph with three static graph variants, as summarized in Table VI. For a fair comparison, all methods shared the same network architecture, training procedure, and calibration pipeline, with graph construction being the only varying factor. Specifically, uniform adopts a fully connected fixed topology, correlation constructs a fixed adjacency matrix using Pearson correlation coefficients, and Mutual Information (MI) models node relationships using normalized mutual information. In contrast, dynamic graph generates sample-specific graph structures through a learnable adjacency mechanism, allowing interactions to adapt to individual physiological states and temporal variations. As shown in Table VI, dynamic graph achieves the best accuracy on both datasets. Compared with uniform, it improves accuracy by 2.86% and 6.92% on AffectiveROAD and hciLab, respectively, while also maintaining consistent advantages over both correlation and MI. Although correlation and MI incorporate data-driven physiological relationships and achieve improvements over the uniform baseline, their graph structures remain fixed across all samples and therefore cannot capture the dynamic physiological interac-

tions that vary across subjects and time windows. These results suggest that relying solely on global statistical relationships is insufficient to fully characterize complex cross-modal physiological coupling patterns, whereas the proposed sample-specific dynamic graph can more effectively model dynamic physiological connectivity, leading to improved personalized physiological state estimation performance.

2) Analysis of Learned Physiological Interactions: To analyze the cross-modal physiological interactions learned by SAGE, we further visualized the dynamic edge activation rates generated by the graph constructor and the corresponding mask-aware graph attention network (GAT) attention weights, as shown in Fig. 4. The activation-rate maps (Fig. 4(a) and (c)) quantify how frequently a directed edge is selected by the dynamic graph generator, whereas the attention maps (Fig. 4(b) and (d)) reflect the relative importance assigned to an edge once it is activated during message passing. As shown in Fig. 4(a) and Fig. 4(b), extensive cross-modal connections are observed under both high- and low-stress conditions in the AffectiveROAD dataset. However, edges involving EDA_SCR, EDA_stat, inter-beat interval (IBI)_HRV, HR_stat, and TEMP_stat exhibit higher activation rates and attention weights under high stress. For example, connections such as EDA_SCR $\rightarrow$ ACC_ts, IBI_HRV $\rightarrow$ ACC_ts, and BVP_stat $\rightarrow$ TEMP_stat are consistently strengthened. These patterns suggest that the model places greater emphasis on the coordinated interactions among autonomic nervous system activity, cardiovascular regulation, and movement- and temperature-related responses, rather than relying on any individual physiological signal. A similar trend can be observed in the hciLab Driving dataset. Under high cognitive workload, enhanced interactions are primarily concentrated among EDA_stat, ECG_stat, HR_stat, ACC_stat, TEMP_stat. In particular, connections such as EDA_stat $\rightarrow$ TEMP_stat, EDA_stat $\rightarrow$ ECG_stat, ECG_stat $\rightarrow$ ACC_stat, and HR_stat $\rightarrow$ ACC_stat receive higher attention weights, indicating that the model increasingly relies on the dynamic coupling among electrodermal activity, cardiovascular responses, and movement-related indicators as cognitive workload increases. Remarkably, for both stress recognition and cognitive workload estimation, the most strongly weighted connections consistently involve multiple physiological modalities rather than a single node or sensor. This observation is consistent with the modality-masking results in Table VII, where removing any individual modality does not lead to substantial performance degradation. Together, these findings suggest that SAGE learns meaningful cross-modal physiological dependencies rather than relying on a specific signal source.

E. Decision-Level Calibration Analysis

1) Component-Wise Ablation: This section analyzes the contribution of the key decision-level components in SAGE through incremental ablation experiments under the strict LOSO evaluation protocol. As summarized in Table VIII, the introduction of decision-level calibration yields substantial performance gains over the baseline LOSO pipeline on both datasets. On AffectiveROAD, the accuracy improves from 65.7% to 76.6%, while the AUROC increases from 0.724 to 0.797. Similar improvements are observed on hciLab Driving, where the accuracy rises from 56.9% to 75.3% and the AUROC increases from 0.636 to 0.783. These results suggest that explicit calibration of model outputs effectively mitigates cross-subject distribution shifts and enhances decision reliability under strict LOSO conditions. Further incorporating the sample-quality-aware gating mechanism leads to additional improvements on both datasets, resulting in the final SAGE performance of 80.6% accuracy and 0.872 AUROC on AffectiveROAD, and 80.5% accuracy and 0.862 AUROC on hciLab Driving. The performance gain on hciLab Driving is particularly pronounced, indicating that calibration reliability becomes increasingly important in scenarios characterized by greater subject variability and higher label ambiguity. Overall, the consistent improvements observed across both datasets demonstrate that decision-level calibration and reliability-aware configuration selection are the primary contributors to the performance gains of SAGE.

TABLE VIII
INCREMENTAL (BUILD-UP) ABLATION STUDY ON AFFECTIVEROAD AND HCILAB DRIVING DATASETS UNDER THE LOSO PROTOCOL.

Method	AffectiveROAD				hciLab Driving			
	Acc.(%)	Δ_{Prev}	Δ_{Base}	AUROC	Acc.(%)	Δ_{Prev}	Δ_{Base}	AUROC
Base	65.7	–	–	0.724	56.9	–	–	0.636
+ Cal	76.6	+10.9	+10.9	0.797	75.3	+18.4	+18.4	0.783
+ Cal + Gate (SAGE)	**80.6**	+4.0	+14.9	**0.872**	**80.5**	+5.2	+23.6	**0.862**

Acc. denotes classification accuracy. Δ_{Prev} and Δ_{Base} represent the performance gain relative to the previous stage and the baseline model, respectively.

TABLE IX
SENSITIVITY ANALYSIS OF DIFFERENT CALIBRATION RATIOS ON AFFECTIVEROAD AND HCILAB DRIVING DATASETS.

Dataset	Calibration Ratio	AUROC	F1	Accuracy (%)
AffectiveROAD	0%	0.7238	0.6608	65.70
	10%	0.8532	0.8165	79.27
	15%	0.8673	0.8281	81.15
	20%	0.8720	0.8218	80.63
	30%	0.8742	0.8298	80.86
	40%	0.8741	0.8287	81.39
	50%	0.8815	0.8313	81.59
hciLab Driving	0%	0.6357	0.6269	56.91
	10%	0.8373	0.7678	76.08
	15%	0.8420	0.7662	77.92
	20%	0.8619	0.7722	80.46
	30%	0.8710	0.8114	80.65
	40%	0.8546	0.7959	80.05
	50%	0.8668	0.8036	80.84

The 0% entry represents results obtained without target-driver calibration.

2) Calibration Data Efficiency Analysis: To evaluate the sensitivity of SAGE to the amount of target-subject calibration data, we further analyzed the impact of different calibration ratios on model performance, as summarized in Table IX. In summary, introducing only a small amount of target-subject data leads to substantial performance improvements. For example, the accuracy increases from 65.7% to 79.3% on AffectiveROAD and from 56.9% to 76.1% on hciLab when the calibration ratio increases from 0% to 10%, demonstrating the high data efficiency of the proposed personalized adaptation mechanism. From a practical deployment perspective, the

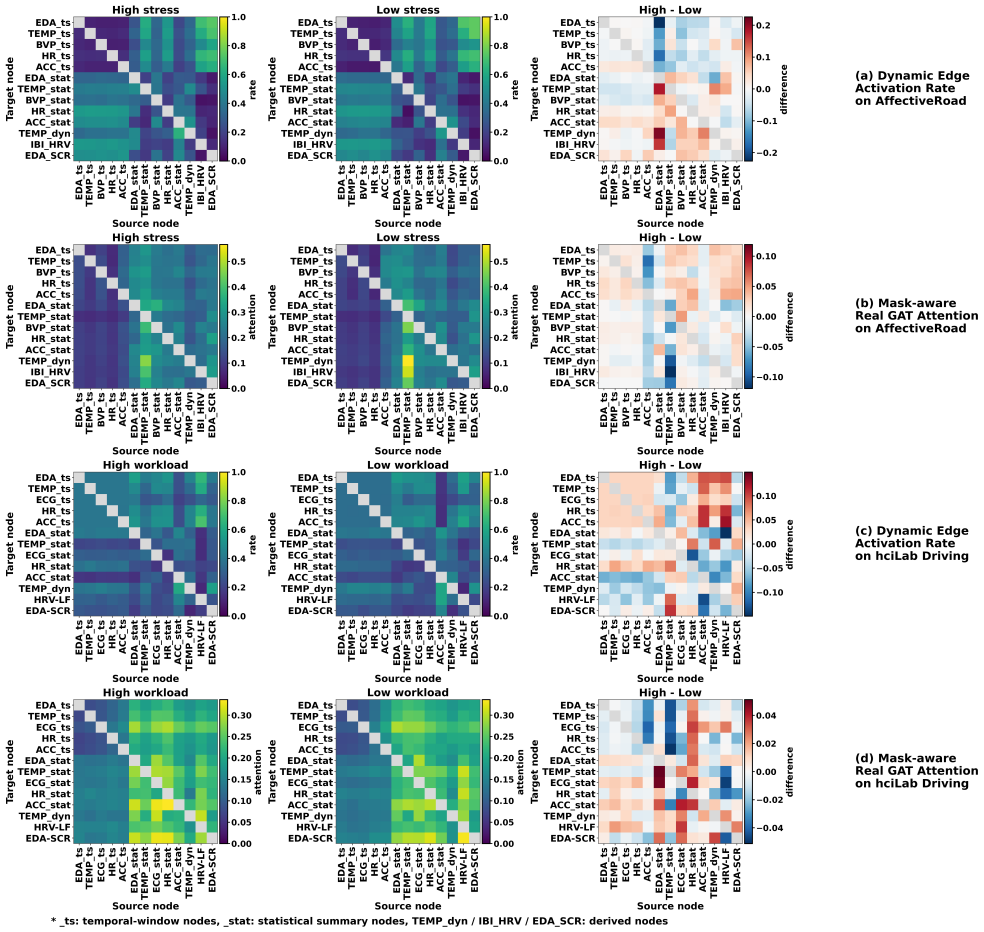

Fig. 4. Cross-dataset comparison of dynamic edge activation and mask-aware GAT attention patterns on AffectiveROAD and hciLab Driving.

objective is not to maximize performance by using as much calibration data as possible, but rather to achieve strong and stable predictive performance while minimizing the calibration burden on new users. The results indicate that a calibration ratio of 20% is sufficient to consistently achieve strong accuracy, F1 score, and AUROC performance across both datasets. Considering that further increases in calibration data provide only limited performance gains while requiring substantially more user-specific data, 20% was selected as the default calibration ratio in this study. This choice is also consistent with the 15%–30% calibration range commonly adopted in previous driving-related physiological monitoring studies [23], [33]. From an application perspective, a 20% calibration ratio corresponds to only approximately 5–6 minutes of subject-specific data in a typical 30-minute driving task, yet is sufficient for SAGE to achieve the performance reported in this study. Even with only 10% calibration data (approximately 3 minutes), substantial performance gains can still be obtained. Both datasets exhibit a similar trend: model performance improves rapidly at low calibration ratios and gradually stabilizes within the 10%–20% range, while further increasing the amount of calibration data yields only marginal gains. These results indicate that SAGE is highly robust to calibration-set size, effectively alleviates calibration data scarcity, and maintains reliable adaptation across different calibration conditions.

TABLE X
SUBJECT-WISE LOSO ACCURACY AND ABLATION VARIANTS.

AffectiveROAD				*hciLab Driving*			
Subject	**Base (%)**	**+Cal (%)**	**SAGE (%)**	**Subject**	**Base (%)**	**+Cal (%)**	**SAGE (%)**
D1	69.9	73.2	**77.3**	P1	63.3	73.5	**76.2**
D2	59.4	62.7	**67.1**	P2	55.7	66.2	**71.6**
S_Drv3	65.5	76.5	**81.9**	P3	47.5	**81.3**	79.6
S_Drv4	52.4	68.3	**74.9**	P4	45.7	60.9	**65.9**
S_Drv5	74.5	88.8	**93.6**	P5	62.6	77.3	**80.0**
S_Drv7	63.1	69.8	**73.3**	P6	70.0	73.1	**83.0**
D8	59.7	84.9	**88.7**	P7	68.3	80.8	**86.0**
S_Drv9	65.0	80.3	**86.7**	P8	51.9	81.4	**89.6**
S_Drv11	81.6	**84.7**	82.2	P9	52.3	71.0	**79.2**
				P10	51.9	87.3	**93.7**
Mean	**65.7**	**76.6**	**80.6**	**Mean**	**56.9**	**75.3**	**80.5**

In AR, D1, D2, and D8 denote drivers with multiple driving sessions, whereas the prefix "S_" denotes single-session drivers; HCI participants P1–P10 are all single-session.

F. Performance Heterogeneity across Subjects

We analyze subject-level performance to characterize inter-subject heterogeneity under realistic cross-subject deployment conditions. Table X summarizes the subject-wise LOSO results on both datasets. On the AffectiveROAD dataset, subject-level accuracies obtained by the baseline model range from 52.4% to 81.6%, indicating substantial variability across drivers. Introducing decision-level calibration consistently improves performance for nearly all subjects, increasing the mean accu-

TABLE XI
COMPARISON WITH EXISTING METHODS OF SWELL-KW.

Method	Accuracy (%)	F1	AUROC	AUPRC	Precision	Recall	Protocol
SVM [34]	58.9	–	–	–	–	–	LOSO
CNN+Transformer [35]	58.1	0.588	–	–	–	–	LOSO
AttX VGG (II) [36]	65.2	0.623	–	–	–	–	LOSO
CNN+Transformer with fine-tuning [35]	71.6	0.742	–	–	–	–	LOSO
SAGE (ours)	**97.47**	**0.979**	**0.993**	**0.997**	**0.972**	**0.987**	**LOSO**

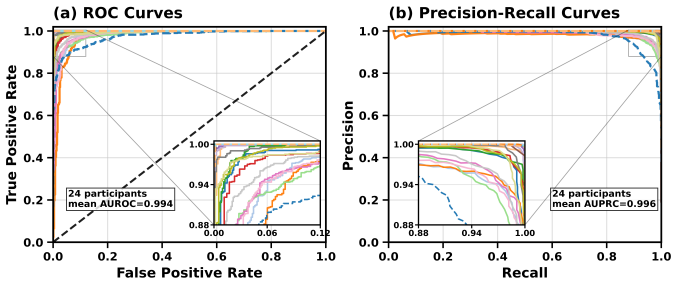

Fig. 5. Per-subject ROC and precision–recall curves on SWELL-KW under the LOSO protocol.

racy from 65.7% to 76.6%. The complete SAGE framework further improves the mean accuracy to 80.6% and yields performance gains for the vast majority of subjects, demonstrating that the proposed decision-centric adaptation strategy effectively mitigates subject-specific distribution shifts. A similar trend is observed on the hciLab Driving dataset. While baseline performance varies considerably across participants, decision-level calibration provides substantial improvements for most subjects, increasing the average accuracy from 56.9% to 75.3%. After incorporating the complete SAGE framework, the mean accuracy further increases to 80.5%. Notably, the largest improvements are generally observed on lower-performing subjects, suggesting that the proposed framework is particularly effective in reducing performance disparities caused by subject heterogeneity. Taken together, the subject-level analyses across both datasets support two key observations. First, substantial inter-subject variability remains a fundamental challenge for physiological-state monitoring under strict LOSO evaluation. Second, the proposed decision-level calibration and adaptation mechanisms not only improve average performance but also consistently reduce subject-level performance disparities, leading to more stable and reliable predictions across diverse individuals.

G. Cross-Scenario Generalization Evaluation

1) Validation on an Office Knowledge-Work Scenario: To further evaluate the generalization capability of SAGE across different application scenarios, we conducted additional experiments on the SWELL-KW dataset [34]. SWELL-KW is a stress-recognition dataset collected from 25 subjects in an office knowledge-work environment, representing a substantially different task context and application scenario from the driving datasets used in previous experiments. One subject was excluded due to considerable missing data, leaving 24 subjects for LOSO evaluation. Only the ECG and skin conductance

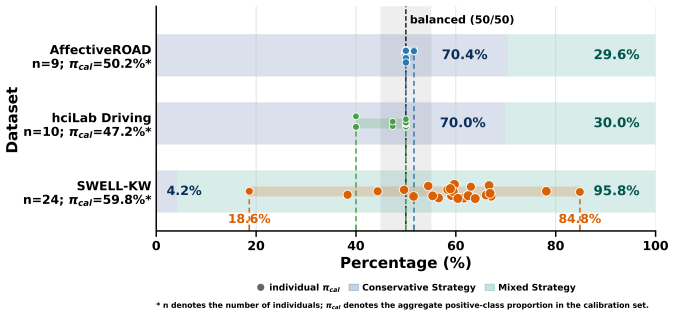

Fig. 6. Calibration-set class-prior variability and gating-strategy activation across the three datasets.

(SC) were used for analysis, because they are the only physiological modalities provided by the dataset. ECG signals were band-pass filtered at 0.5–40 Hz and resampled to 256 Hz, while SC signals were low-pass filtered at 5 Hz, resampled to 32 Hz, and further decomposed into tonic and phasic components. Apart from these preprocessing steps, all other modeling settings, including segment length and calibration ratio, were kept consistent with those used for the AffectiveROAD and hciLab Driving datasets (refer to Section III-B). For each analysis window, handcrafted features were extracted, including statistical features (Hamming-weighted mean, standard deviation, skewness, kurtosis, minimum, maximum, range, and interquartile range), EDA-derived SCR features (number of peaks, mean peak amplitude, maximum peak amplitude, and peak rate), and ECG-derived HRV features (mean heart rate, standard deviation of NN intervals (SDNN), root mean square of successive differences (RMSSD), percentage of successive NN intervals differing by more than 50 ms (pNN50), Low Frequency band of HRV (0.04–0.15 Hz) (LF) power, and High Frequency band of HRV (0.15–0.40 Hz) (HF) power).

For a fair comparison, Table XI includes only representative published methods evaluated under the same LOSO protocol. SAGE achieved the best performance across all reported metrics, attaining an accuracy of 97.47%, an F1 score of 0.979, an AUROC of 0.993, and an AUPRC of 0.997, substantially outperforming the strongest previously reported LOSO result (71.6% accuracy and 0.742 F1 score). Subject-level ROC and precision–recall curves in Fig. 5 further show that most subjects exhibit near-ideal classification characteristics. Overall, these findings indicate that the stress-related physiological representations learned by SAGE are highly discriminative and transferable across scenarios, further demonstrating its strong cross-scenario generalization capability.

2) Gating Activation Analysis Under Class-Prior Variability: The performance gain achieved by SAGE on SWELL-KW was considerably larger than that observed on the driving datasets. To further investigate this phenomenon, Fig. 6 summarizes the calibration-set positive-class proportions under the LOSO protocol. While AffectiveROAD and hciLab Driving remained close to class balance ($r_{cal} = 50.2\%$ and 47.2%, respectively), this observation is consistent with the subject-specific quantile-based labeling strategy described in Section III-B.4. SWELL-KW exhibited substantially greater inter-subject variability, with r_{cal} ranging from 18.6% to 84.8%. In contrast to

the driving datasets, where labels are defined using subject-specific quantiles of continuously varying states, SWELL-KW employs protocol-defined stress conditions, resulting in substantially greater inter-subject variability in class priors. Such heterogeneous class priors make fixed calibration strategies more susceptible to subject-specific bias. Notably, the activation rate of the mixed strategy containing the aggressive path reached 95.8% in SWELL-KW, whereas the corresponding activation rates were only approximately 30% in AffectiveROAD and hciLab Driving. This observation suggests that, under substantial class-prior variability, the proposed gating mechanism more frequently selects personalized calibration configurations to compensate for inter-subject distribution differences. Combined with the largest performance gain achieved on SWELL-KW, these results further support the effectiveness of the sample-quality-aware decision-level calibration mechanism under pronounced class-prior heterogeneity.

H. Robustness and Deployment Considerations

1) Robustness to Modality Loss and Signal Degradation: To evaluate the robustness of SAGE to modality loss and signal degradation, we conducted modality-masking and noise-perturbation experiments (Table VII). Modality masking simultaneously removes all nodes (including temporal, statistical and derived nodes) associated with a physiological modality to simulate sensor failure. Noise perturbation injects additive Gaussian noise into all node embeddings at signal-to-noise ratio (SNR) of 10 dB, -1 dB, -2 dB, -3 dB, -4 dB, and -5 dB to evaluate the robustness of the model under noise conditions ranging from moderate to extremely severe. Although removing an individual modality reduced performance, the overall degradation remained limited: From the perspective of ΔAUROC, ACC and TEMP had the largest impact on AffectiveROAD, whereas EDA and BVP were most influential on hciLab Driving. Even with an entire modality removed, the model maintained relatively high AUROC and accuracy values, indicating that SAGE effectively exploits complementary information across modalities rather than relying on a single physiological source. Under noise perturbation, the model performance slightly degraded as the SNR decreased from 10 dB to -5 dB. On the AffectiveROAD dataset, the AUROC under the most severe noise condition (SNR = -5 dB) reached 0.8221, corresponding to a decrease of 0.0499 relative to the complete model. On the hciLab Driving dataset, the AUROC decreased to 0.7543, corresponding to a reduction of 0.1076 relative to the complete model, while the accuracy exhibited a similar downward trend. Despite the fact that the noise power exceeded the signal power of the node embeddings under negative SNR conditions, SAGE maintained satisfactory predictive performance without catastrophic degradation. Notably, the classification accuracies achieved on both datasets under the most severe noise condition remained higher than those reported by existing methods in Table II under the same LOSO protocol. Taken together, SAGE maintains stable performance under sensor loss and signal degradation. This robustness can be attributed to the distributed nature of the graph representation, as well as the ability of dynamic graph construction and graph-attention propagation to adaptively adjust node interactions and information weights.

TABLE XII
COMPUTATIONAL COMPLEXITY AND INFERENCE-TIME ANALYSIS OF SAGE.

Item	Model parameters	FP32 size	FLOPs/win	Latency/win	RTF
Value	0.914 M	3.49 MB	4.21 M	2.20 ± 0.06 ms	1.46×10^{-4} ($\sim$6828$\times$)

2) Deployment Feasibility and Computational Efficiency: We further analyzed the deployment requirements and computational complexity of SAGE. Practical deployment in real-world scenarios requires transforming the offline analysis pipeline into an online continuous monitoring pipeline, satisfying the requirements for real-time responsiveness, computational efficiency, and long-term operation. Specifically, the deployment platform should be equipped with the multimodal physiological sensors required by the proposed framework to continuously acquire the corresponding physiological signals. The incoming data should then undergo the same data processing pipeline described in Section III-B, followed by continuous window-based inference using SAGE. In practical deployment, additional noise, motion artifacts, and contextual factors that might induce physiological responses similar to the target states should also be considered. To satisfy real-time requirements, each decision window should be processed immediately after data acquisition while avoiding excessive system latency caused by high model complexity. Furthermore, the model should maintain low computational complexity, a compact model size, and low memory consumption. In addition, it should provide a stable computational pipeline to support long-term continuous operation in application scenarios requiring prolonged monitoring.

As shown in Table XII, SAGE contains only 0.91 M trainable parameters and occupies 3.49 MB in floating point (FP32) format, resulting in low storage and memory requirements that are well suited for resource-constrained embedded and mobile platforms. In addition, SAGE requires only 4.21 MFLOPs for each 15-s decision window, indicating a low computational burden that facilitates deployment on edge devices and embedded processors. To further evaluate its real-time deployment capability, CPU inference was performed on an Apple M4 CPU with 16 GB unified memory (batch size = 1). The average inference latency for a single 15-s decision window was 2.20 ± 0.06 ms, corresponding to a real-time factor (RTF) of 1.46×10^{-4}, which accounts for only approximately 0.015% of the available decision interval. The results obtained on a general-purpose CPU demonstrate that SAGE satisfies the computational requirements of the representative deployment scenarios discussed above. In summary, SAGE features a lightweight model architecture, low computational resource requirements, and millisecond-level inference latency, providing a solid foundation for its deployment in real-world continuous physiological monitoring applications from the model perspective. Nevertheless, practical deployment of SAGE will still require further platform adaptation, system integration, and application-specific validation.

V. Conclusion

This paper addresses the challenge of decision stability in driver-state monitoring under realistic cross-subject deployment conditions and proposes SAGE, a unified framework for driver stress recognition and cognitive workload estimation. SAGE extends subject adaptation from the representation level to the decision level, enabling data-efficient personalized decision-making on top of shared physiological representations. By combining dynamic physiological graph modeling with decision-level calibration and configuration gating, the framework achieves reliable cross-subject generalization without extensive subject-specific retraining. Experimental results under the strict LOSO protocol across multiple datasets demonstrate that SAGE achieves stable and competitive performance under heterogeneous physiological monitoring conditions. Validation on the SWELL-KW dataset further demonstrates the transferability of the learned physiological representations and the strong cross-scenario generalization capability of the proposed framework. Future work will explore online adaptation, continual calibration, and long-term subject-specific physiological modeling to further improve robustness in real-world deployments. To sum up, this study suggests that, under pronounced subject heterogeneity and limited calibration data, representation-level modeling alone is insufficient for reliable generalization, whereas explicit decision-level adaptation provides an effective mechanism for personalized decision-making. Beyond driver monitoring, the proposed decision-centric adaptation framework offers a unified approach for AI-enabled modeling of cognitive and affective states and provides methodological insights for translational cognitive neuroscience and mental-health research.

References

[1] World Health Organization. Road traffic injuries. WHO health topic page; cites approx. 1.19 million annual deaths and 20–50 million non-fatal injuries.

[2] J. M. Lyznicki *et al.* Sleepiness, driving, and motor vehicle crashes. *Jama*, 279(23):1908–1913, 1998.

[3] S.-H. Park *et al.* Dynamic multi-biosignal fusion for detecting the mental states of drivers and passengers in vehicles. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[4] L. Hu *et al.* The challenges of driving mode switching in automated vehicles: A review. *IEEE Transactions on Vehicular Technology*, 73(2):1777–1791, 2023.

[5] K. Jiang *et al.* Degrade is upgrade: Learning degradation for low-light image enhancement. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1078–1086, 2022.

[6] Y. Xiao *et al.* Ediffsr: An efficient diffusion probabilistic model for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2023.

[7] F. Chen *et al.* Multimodal behavior and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(4):1–36, 2013.

[8] G. J. Norman *et al.* Wearable ans monitoring in real life: A critical review of context-sensitive interpretation and implications for psychophysiology. *Autonomic Neuroscience*, pp. 103364, 2025.

[9] Y. M. Ulrich-Lai and J. P. Herman. Neural regulation of endocrine and autonomic stress responses. *Nature reviews neuroscience*, 10(6):397–409, 2009.

[10] D. Kahneman. *Attention and Effort*. Prentice-Hall, 1973.

[11] W. Boucsein. *Electrodermal activity*. Springer science & business media, 2012.

[12] G. G. Berntson *et al.* Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology*, 34(6):623–648, 1997.

[13] J. Dieffenderfer *et al.* Low-power wearable systems for continuous monitoring of environment and health for chronic respiratory disease. *IEEE journal of biomedical and health informatics*, 20(5):1251–1264, 2016.

[14] M. Esmaeili *et al.* Hierarchical attention mechanism combined with deep neural networks for accurate semantic segmentation of dental structures in panoramic radiographs. *Scientific Reports*, 15(1):38725, 2025.

[15] Z. Khodaverdian *et al.* A unified multi-task learning framework for automated assessment of left ventricular structure and its systolic function from echocardiography. *Scientific Reports*, 2025.

[16] T. Zheng *et al.* Graphmamba: Whole slide image classification meets graph-driven selective state space model. *Pattern Recognition*, 167:111768, 2025.

[17] H. Sadr *et al.* Web-based ai application for enhanced dental disease diagnosis using advanced object detection integrated with transformer-based attention mechanism. *Oral Radiology*, pp. 1–12, 2026.

[18] M. Nazari *et al.* Enhancing cardiac function assessment: developing and validating a domain adaptive framework for automating the segmentation of echocardiogram videos. *Computerized medical imaging and graphics*, pp. 102627, 2025.

[19] S. Aravinth *et al.* Dynamic cross-domain transfer learning for driver fatigue monitoring: multi-modal sensor fusion with adaptive real-time personalizations. *Scientific Reports*, 15(1):15840, 2025.

[20] H. Zeng *et al.* An eeg-based transfer learning method for cross-subject fatigue mental state prediction. *Sensors*, 21(7):2369, 2021.

[21] Q. Meteier *et al.* Classification of drivers' workload using physiological signals in conditional automation. *Frontiers in psychology*, 12:596038, 2021.

[22] Z. Jiang *et al.* Multimodal mental health digital biomarker analysis from remote interviews using facial, vocal, linguistic, and cardiovascular patterns. *IEEE journal of biomedical and health informatics*, 28(3):1680–1691, 2024.

[23] J. Healey and R. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166, 2005.

[24] J. Lee *et al.* Driving stress detection using multimodal convolutional neural networks with nonlinear representation of short-term physiological signals. *Sensors*, 21(7):2381, 2021.

[25] L.-l. Chen *et al.* Cross-subject driver status detection from physiological signals based on hybrid feature selection and transfer learning. *Expert Systems with Applications*, 137:266–280, 2019.

[26] N. E. Haouij *et al.* Affectiveroad system and database to assess driver's attention. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pp. 800–803, 2018.

[27] S. Schneegass *et al.* A data set of real world driving to assess driver workload. In *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications*, pp. 150–157, 2013.

[28] M. Hosseini *et al.* A multimodal stress detection dataset with facial expressions and physiological signals. *Scientific Data*, 12(1):1844, 2025.

[29] D. Lopez-Martinez *et al.* Detection of real-world driving-induced affective state using physiological signals and multi-view multi-task machine learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 356–361. IEEE, 2019.

[30] H. Rafi *et al.* Tree-based personalized clustered federated learning: A driver stress monitoring through physiological data case study. *IEEE Internet of Things Journal*, 2024.

[31] P. Siirtola and J. Röning. Comparison of regression and classification models for user-independent and personal stress detection. *Sensors*, 20(16):4402, 2020.

[32] D. Jaiswal *et al.* Tinystressnet: On-device stress assessment with wearable sensors on edge devices. In *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pp. 166–171. IEEE, 2024.

[33] M. Choi *et al.* Wearable device-based system to monitor a driver's stress, fatigue, and drowsiness. *IEEE Transactions on Instrumentation and Measurement*, 67(3):634–645, 2017.

[34] S. Koldijk *et al.* The swell knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th international conference on multimodal interaction*, pp. 291–298, 2014.

[35] B. Behinaein *et al.* A transformer architecture for stress detection from ecg. In *Proceedings of the 2021 ACM international symposium on wearable computers*, pp. 132–134, 2021.

[36] A. Bhatti *et al.* Attx: Attentive cross-connections for fusion of wearable signals in emotion recognition. *ACM Trans. Comput. Healthcare*, 5(3), September 2024.