SAIN: A Hierarchical Dynamics-Aware EEG-fNIRS Framework for Continuous Affect Estimation

Feng-Qi Cui
Hefei University of Technology, & University of Science and Technology of China
Hefei, China

Jinyang Huang*
Hefei University of Technology, & Hefei Joy Embodied Intelligence Technology Co., Ltd
Hefei, China

Jiacheng Zhao
Hefei University of Technology
Hefei, China

Sirui Zhao
University of Science and Technology of China
Hefei, China

Yi Yang
Hefei University of Technology
Hefei, China

Dan Guo
Hefei University of Technology
Hefei, China

Xiaokang Zhou
Kansai University
Osaka, Japan

Zhun Zhong
Hefei University of Technology
Hefei, China

Fusang Zhang
Beihang University
Beijing, China

Jianwei Lu
Shanghai University of Traditional Chinese Medicine
Shanghai, China

Meng Li
Hefei University of Technology
Hefei, China

Bing-Kun Bao
Hefei University of Technology
Hefei, China

Abstract

Continuous affect estimation from electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) aims to recover time-varying Valence and Arousal from heterogeneous physiological signals. Their differences in channel organisation, response latency, temporal statistics, and noise characteristics make homogeneous fusion and single-timescale modelling insufficient. We propose the *Structured Affective Interaction Network* (SAIN), a hierarchical framework comprising Structured-Axial Physiological Encoder (SAPE), Adaptive Cross-Modal Interaction (ACI), and Dual-Rate Evolution (DRE). SAPE independently preserves modality-specific channel relations while integrating time-indexed and trial-level channel-feature context. ACI explicitly models feature-wise selection, modality discrepancy, and coordinated responses to produce a fused physiological representation and a direct physiological estimate. DRE then uses a training-derived Group Affective Trajectory (GAT) as coarse temporal context and combines dimension-wise adaptive gain, fast correction, and a retention-controlled slow recurrent state. Experiments demonstrate that SAIN outperforms reproduced baselines and independently trained unimodal variants. Modality controls support the complementary integration of EEG and fNIRS, channel-masking analyses reveal spatially structured physiological dependence, and temporal ablations establish the distinct contributions of fast correction and slow recurrent accumulation. These results jointly support the structured encoding, adaptive interaction, and dual-rate evolution underlying SAIN.

*Corresponding Author. E-mail: hjy@hfut.edu.cn

MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3838613

CCS Concepts

• **Human-centered computing**;

Keywords

Continuous affect estimation, EEG, fNIRS, multimodal interaction, temporal modelling

ACM Reference Format:
Feng-Qi Cui, Jinyang Huang, Jiacheng Zhao, Sirui Zhao, Yi Yang, Dan Guo, Xiaokang Zhou, Zhun Zhong, Fusang Zhang, Jianwei Lu, Meng Li, and Bing-Kun Bao. 2026. SAIN: A Hierarchical Dynamics-Aware EEG-fNIRS Framework for Continuous Affect Estimation. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3767308.3838613

Our team (HFUT-Beyond Vision) achieved **1st place** in Track 4 of the 2026 Multimodal Emotional Recognition (MER) Challenge, as reported on the official competition website: https://www.codabench.org/competitions/15992/.

1 Introduction

Continuous affect estimation predicts time-varying trajectories of Valence and Arousal rather than assigning a single emotion label to an entire recording. This formulation captures how affect emerges, persists, and changes throughout a stimulus, making it relevant to adaptive media [45], human-computer interaction [2], affect-aware education, and physiological monitoring [21, 30, 31]. Compared with recording-level recognition [3–6, 36, 46–48], continuous estimation is more demanding because the model must preserve meaningful local variations while suppressing physiological noise and annotation fluctuations that may otherwise produce unstable frame-wise predictions.

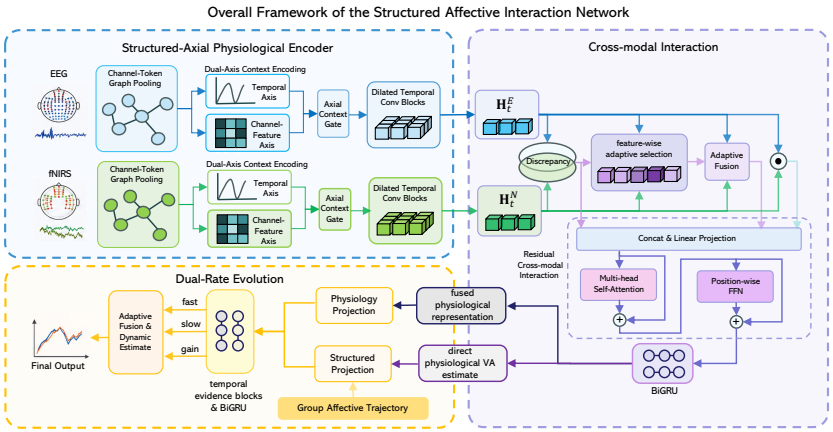

Figure 1: SAIN architecture. SAPE independently encodes EEG and fNIRS through channel-token relations, dual-axis context, and dilated temporal blocks. ACI produces a fused physiological representation and a direct physiological VA estimate, which DRE refines through fast and slow temporal pathways.

Table 1: Comparison with reproduced baselines.

Method	Context-Augmented			Context-Free
	MAE	Val.	Aro.	MAE
SAIN	**0.0998**	0.0883	**0.1113**	–
EEGNet [20]	0.1092	0.0853	0.1331	0.2235
ASAC-Net [8]	0.1094	**0.0844**	0.1344	0.1874
TeCh [40]	0.1145	0.0897	0.1393	0.1600
MulT [38]	0.1300	0.1045	0.1556	0.2346

Table 2: Modality controls.

Physiological	MAE	Val.	Aro.
EEG only	0.1047	0.0928	0.1166
fNIRS only	0.1052	0.0921	0.1182
EEG + fNIRS	**0.0998**	**0.0883**	**0.1113**

Among physiological modalities, electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) provide complementary observations of affective dynamics. EEG records rapid electrophysiological activity with high temporal resolution, whereas fNIRS captures smoother and delayed haemodynamic responses associated with neurovascular activity. Continuous regression from their simultaneous measurements nevertheless remains challenging because the two modalities differ substantially in channel organisation, response latency, temporal statistics, and artefact characteristics [7, 24, 32, 41]. Their affective evidence may therefore become informative at different temporal stages, such that direct frame-level alignment or homogeneous feature processing can suppress useful modality-specific information.

Existing studies have improved physiological affect modelling through stronger EEG representations, haemodynamic feature learning, and multimodal fusion [8, 20, 23]. However, these advances mainly strengthen individual components, while the joint modelling of modality-specific channel structure, time-varying EEG-fNIRS interaction, and multi-timescale affective dynamics remains insufficient for continuous regression. These challenges are closely coupled. First, EEG electrodes and fNIRS channels follow different sensing organisations. Flattening the channel dimension too early discards intra-modal relational structure, whereas imposing a shared topology ignores their distinct acquisition geometries and measurement mechanisms. Second, the contributions of EEG and fNIRS may agree during some periods and diverge during others because their reliability, response delay, and physiological sensitivity vary over time. Static concatenation cannot explicitly distinguish agreement from discrepancy, while a fixed modality weight assumes temporally stationary reliability. Third, rapid electrophysiological fluctuations, gradual haemodynamic responses, and persistent affective trends coexist within the same trial. A single temporal update mechanism may either oversmooth transient variations or overreact to local noise, making it difficult to preserve short-term changes while maintaining long-term continuity.

To address these challenges, we propose the *Structured Affective Interaction Network* (**SAIN**), illustrated in Figure 1. SAIN organises continuous affect estimation into three hierarchical stages: Structured-Axial Physiological Encoder (**SAPE**), Adaptive Cross-Modal Interaction (**ACI**), and Dual-Rate Evolution (**DRE**). First, SAPE learns modality-specific channel relations and multi-scale temporal context to preserve the distinct structural and dynamic characteristics of EEG and fNIRS, addressing information loss caused by premature homogeneous processing. Second, ACI combines adaptive feature selection with explicit modelling of cross-modal discrepancy and coordination to construct complementary physiological evidence, addressing the time-varying reliability and asynchronous responses of the two modalities. Third, DRE integrates a training-derived Group Affective Trajectory (**GAT**) with fast correction and slow recurrent accumulation to model transient and persistent affective changes, addressing the conflict between local sensitivity and long-term continuity.

Through this hierarchy, SAIN progressively transforms heterogeneous physiological measurements into modality-specific structural representations, complementary multimodal evidence, and continuous affective trajectories. Experiments show that SAIN outperforms reproduced baselines and independently trained unimodal models. Modality controls verify the complementarity of EEG and fNIRS, temporal-path ablations demonstrate the distinct roles of fast correction and slow recurrent accumulation, and visual analyses reveal structured channel dependence and sustained temporal improvement. Our main contributions are summarised as follows:

- We propose **SAIN**, a hierarchical framework that unifies Structured-Axial Physiological Encoding, Adaptive Cross-Modal Interaction, and Dual-Rate Evolution for continuous EEG-fNIRS Valence-Arousal estimation.
- We develop SAPE and ACI to preserve modality-specific structure and model adaptive cross-modal coordination, together with DRE to integrate group-level temporal context,

Table 3: Controlled analysis of temporal evolution.

Configuration	MAE	Val.	Aro.
SAIN w/o fast pathway	0.1012	**0.0882**	0.1142
SAIN w/o slow pathway	0.1016	0.0885	0.1147
SAIN (full)	**0.0998**	0.0883	**0.1113**

fast correction, and slow recurrent accumulation for multi-timescale affective evolution.

- Experiments show that SAIN outperforms reproduced baselines and independently trained unimodal models. Modality controls, temporal ablations, and behavioural analyses further support EEG-fNIRS complementarity, structured physiological dependence, and the distinct contributions of fast and slow temporal pathways.

2 Related Work

2.1 Physiological signal modelling

Dimensional affect models represent emotional states using continuous Valence and Arousal, requiring the recovery of an evolving trajectory rather than a single label for each recording [19, 35]. This formulation places greater emphasis on temporal consistency, local transitions, and response latency, because errors may arise not only from incorrect affective magnitude but also from delayed or over-smoothed predictions. EEG is widely used in affective computing because of its high temporal resolution and sensitivity to rapid electrophysiological changes. Compact convolutional networks extract spatial-temporal features directly from multichannel inputs [20], while graph-based architectures represent inter-electrode dependencies through predefined or learnable topologies [14]. Emotion-oriented models further combine spatial, spectral, and temporal information to capture complementary views of neural activity [49]. Recent foundation models extend this direction through contrastive, masked, and autoregressive pre-training over large EEG corpora [17, 39, 42]. These approaches improve transferable representations and reduce reliance on task-specific supervision, but most are designed for unimodal classification or recording-level prediction rather than multimodal continuous trajectories [13, 16].

In contrast, fNIRS measures cortical haemodynamic activity through changes in oxygenated and deoxygenated haemoglobin. Owing to neurovascular coupling, fNIRS responses are generally smoother and more delayed than EEG, and are affected by different physiological artefacts and noise sources. Existing methods therefore focus on haemoglobin dynamics, cross-channel connectivity, and robust temporal encoding [7, 32, 41]. The slower haemodynamic response can provide a more stable description of sustained affective states, whereas EEG is more sensitive to rapid local changes. Simultaneous measurement studies show that the two modalities are correlated but neither equivalent nor strictly synchronous [34, 37]. Their relationship is therefore better understood as complementary observations of related physiological processes rather than two interchangeable views of the same signal. A continuous EEG-fNIRS model should consequently preserve their individual channel structures and temporal behaviour before learning shared or complementary relations in a higher-level representation space.

2.2 Multimodal physiological interaction

Multimodal physiological signals can be combined at the input, representation, or decision level. Input-level fusion enables early interaction but is sensitive to differences in sampling rate, feature scale, channel topology, and temporal alignment. Decision-level fusion reduces the need for low-level alignment, but limits the exchange of information before prediction. Representation-level fusion provides a more flexible compromise by preserving modality-specific encoders while allowing higher-level physiological relations to be learned [11]. Hybrid EEG-fNIRS studies have demonstrated the complementarity of electrophysiological and haemodynamic responses through feature-level and decision-level fusion [9, 18]; however, response latency, modality-specific noise, normalisation, and fusion-stage selection remain persistent challenges [1].

Recent methods increasingly move beyond direct concatenation [12, 15, 27–29]. ASAC-Net learns alignment and complementary fusion between EEG and fNIRS [8], while cross-signal generation and general physiological alignment methods seek shared structures without assuming identical responses [24, 43]. In broader multimodal learning, directional attention and shared-specific decomposition have been used to distinguish common evidence from modality-dependent information [22, 38, 44]. These strategies improve cross-modal interaction, but most of them focus primarily on alignment or representation fusion. They do not explicitly account for the possibility that EEG and fNIRS may agree during some periods, diverge during others, and contribute differently across latent dimensions.

Evaluation studies further show that apparent multimodal gains depend on temporal alignment, retained windows, missing modalities, and comparison protocols [25, 26]. Fusion quality therefore depends not only on the interaction operator itself, but also on whether structural and temporal heterogeneity is preserved before interaction. For continuous affect estimation, this issue is particularly important because rapid electrophysiological fluctuations, delayed haemodynamic responses, and persistent affective trends coexist within the same sequence. Existing approaches rarely coordinate modality-specific channel structure, time-varying agreement and discrepancy, and transient versus persistent affective evolution within a unified framework. SAIN addresses this gap through the unified hierarchy of SAPE, ACI, and DRE.

3 Method

SAIN comprises three modules for continuous EEG-fNIRS affect estimation: Structured-Axial Physiological Encoder (SAPE), Adaptive Cross-Modal Interaction (ACI), and Dual-Rate Evolution (DRE). SAPE preserves modality-specific channel organisation and dynamics, ACI constructs complementary physiological evidence by modelling agreement and discrepancy, and DRE coordinates transient fluctuations with persistent affective trends.

3.1 Structured-Axial Physiological Encoder

The Structured-Axial Physiological Encoder (SAPE) processes EEG and fNIRS through structurally identical but parameter-independent

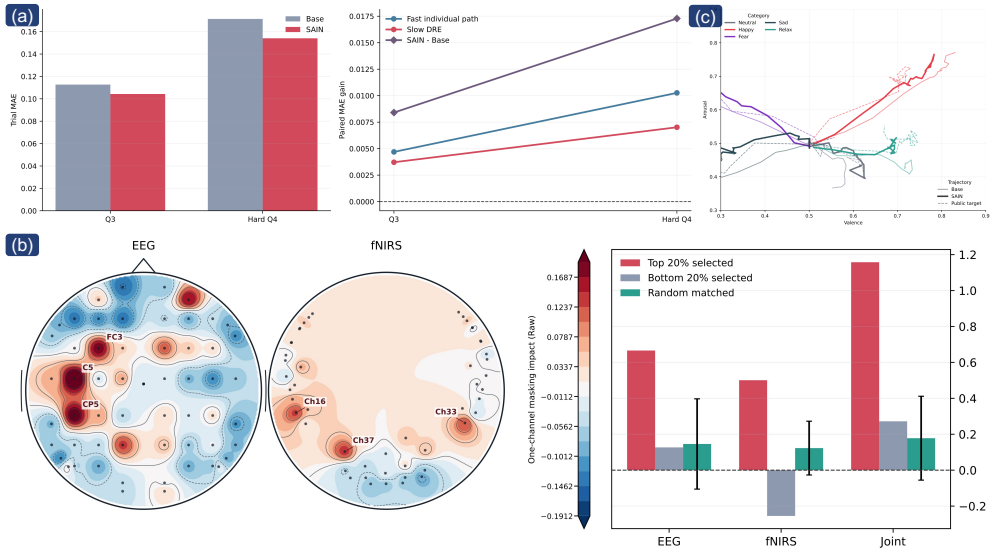

Figure 2: Behavioural and structural analysis. (a) Difficult-trial performance, (b) channel masking, and (c) category-level VA trajectories.

branches. Each branch first models latent relations among modality-specific channels and then captures complementary temporal and channel-feature contexts. The resulting representations are adaptively integrated and further processed by multi-scale dilated temporal blocks, preserving the structural and dynamic characteristics of each modality before cross-modal interaction.

3.1.1 Learnable channel-token relations. For modality $m \in \{E, N\}$, each channel is projected into a latent token and augmented with a learnable channel-position representation. Pairwise similarities between the position embeddings form a data-driven relational matrix, enabling information exchange among channel tokens without imposing a predefined topology. Depthwise temporal convolution is subsequently applied within each token, and attentive, mean, and max pooling are used to summarise the relational features. This process is expressed as

$$\mathbf{H}_m^G = \mathcal{R}_m(\mathbf{X}_m). \tag{1}$$

The resulting representation retains channel identity while capturing latent intra-modal dependencies. Relational propagation exchanges information across channel tokens, whereas attentive, average, and maximum pooling respectively emphasise salient, distributed, and high-response patterns. Their combination therefore preserves structured channel information while producing a compact modality representation.

3.1.2 Dual-axis context modelling. Channel relations alone are insufficient to capture both time-indexed dependencies and trial-level global context. The temporal axis therefore processes the flattened channel-feature representation at each time step, while the channel-feature axis resamples each trial to a fixed length and treats each channel-feature pair as a token. The two complementary contexts are formulated as

$$\mathbf{H}_m^T = \mathcal{A}_T(\mathbf{X}_m), \tag{2}$$

$$\mathbf{h}_m^{CF} = \mathcal{A}_{CF}(\mathbf{X}_m). \tag{3}$$

The trial-level summary is then broadcast along the temporal dimension and combined with the temporal-axis representation through a projection layer,

$$\mathbf{H}_m^A = P_m\left(\mathbf{H}_m^T + \mathbf{1}\,\mathbf{h}_m^{CF\top}\right). \tag{4}$$

A bounded gate adaptively controls the contribution of the axial context,

$$\widetilde{\mathbf{H}}_m = \mathbf{H}_m^G + \eta_m \tanh\left(W_m[\mathbf{H}_m^G; \mathbf{H}_m^A]\right) \odot \mathbf{H}_m^A. \tag{5}$$

Here, η_m constrains the strength of contextual injection. Because the gate is jointly conditioned on the relational and axial representations, trial-level context refines local channel relations only when supported by the current modality state. Multi-scale gated dilated temporal blocks subsequently enlarge the temporal receptive field without premature cross-modal mixing, producing $\mathbf{H}_t^E$ and $\mathbf{H}_t^N$. This processing order preserves modality-specific dynamics before EEG and fNIRS enter the cross-modal interaction stage.

3.2 Adaptive Cross-Modal Interaction

The Adaptive Cross-Modal Interaction (ACI) module models both complementary and divergent information between the modality-specific representations. It first computes their discrepancy as

$$\mathbf{D}_t = \mathbf{H}_t^E - \mathbf{H}_t^N. \tag{6}$$

A feature-wise gate then estimates the relative contribution of EEG and fNIRS at each time step and latent dimension,

$$\mathbf{M}_t = \sigma\left(F_g[\mathbf{H}_t^E; \mathbf{H}_t^N; \mathbf{D}_t]\right), \tag{7}$$

$$\mathbf{H}_t^F = \mathbf{M}_t \odot \mathbf{H}_t^E + (1 - \mathbf{M}_t) \odot \mathbf{H}_t^N. \tag{8}$$

The adaptively fused state is further refined through a residual interaction that combines the current fusion result, the cross-modal discrepancy, and the element-wise coordinated response,

$$\widetilde{\mathbf{H}}_t^F = \mathbf{H}_t^F + F_u\left[\mathbf{H}_t^F; \mathbf{D}_t; \mathbf{H}_t^E \odot \mathbf{H}_t^N\right]. \tag{9}$$

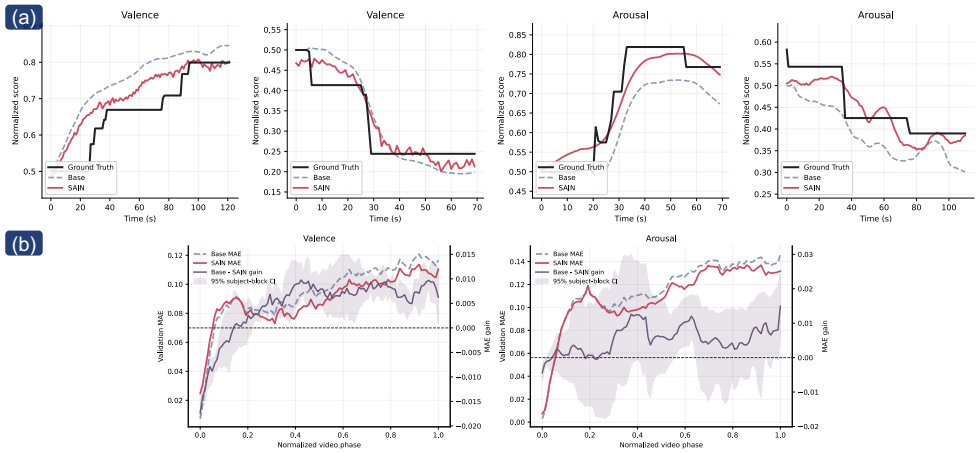

Figure 3: Temporal analysis. (a) Representative predictions and (b) phase-wise MAE with paired improvement.

These components respectively represent adaptive modality selection, cross-modal divergence, and coordinated physiological evidence. The feature-wise gate allows modality contributions to vary across time and latent dimensions, while the residual interaction preserves explicit information about agreement and discrepancy. A two-layer bidirectional GRU subsequently contextualises the updated sequence and produces the fused physiological representation $\mathbf{h}_t^{\text{phys}}$. A shared readout head maps this representation to the direct physiological estimate,

$$\mathbf{y}_t^{\text{phys}} \in \mathbb{R}^2. \tag{10}$$

3.3 Dual-Rate Evolution

The direct physiological estimate reflects the current multimodal evidence, but local perturbations may introduce unstable fluctuations. The Dual-Rate Evolution (DRE) module therefore constructs structured temporal evidence and refines the affective trajectory through complementary fast and slow pathways.

3.3.1 Structured temporal evidence. The Group Affective Trajectory (GAT) describes the population-level affective state together with its dynamics, uncertainty, and effective support. It is represented as

$$\mathcal{G}_{v,t} = \left\{\mathbf{g}_{v,t}, \dot{\mathbf{g}}_{v,t}, \ddot{\mathbf{g}}_{v,t}, \Sigma_{v,t}, c_{v,t}\right\}. \tag{11}$$

Here, $\mathbf{g}_{v,t}$ denotes the group-level Valence-Arousal state, $\dot{\mathbf{g}}_{v,t}$ and $\ddot{\mathbf{g}}_{v,t}$ describe its first- and second-order temporal variation, and $\Sigma_{v,t}$ and $c_{v,t}$ represent uncertainty and effective support. State filtering and backward smoothing are applied to preserve trajectory continuity. The deviation of the direct physiological estimate from the corresponding group state is defined as

$$\mathbf{d}_t = \mathbf{y}_t^{\text{phys}} - \mathbf{g}_{v,t}. \tag{12}$$

DRE jointly encodes the fused physiological representation, the direct estimate and its local variation, the group trajectory state, and the normalised trial phase ϕ_t,

$$\mathbf{e}_t = H_e\left(\mathbf{h}_t^{\text{phys}}, \mathbf{d}_t, \dot{\mathbf{y}}_t^{\text{phys}}, \mathbf{g}_{v,t}, \dot{\mathbf{g}}_{v,t}, \ddot{\mathbf{g}}_{v,t}, \Sigma_{v,t}, c_{v,t}, \phi_t\right). \tag{13}$$

The resulting evidence vector combines current multimodal physiology, short-term variation in the direct estimate, deviation from the group trajectory, trajectory dynamics, uncertainty, and trial phase. Two temporal evidence blocks capture local patterns, after which a bidirectional GRU aggregates contextual information for fast-slow evolution.

3.3.2 Adaptive fast-slow update. DRE first predicts a dimension-wise adaptive gain,

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_0 \odot \exp(r_\alpha \tanh H_\alpha(\mathbf{e}_t)), \tag{14}$$

which allows Valence and Arousal to evolve at different scales. The fast pathway then generates an instantaneous correction,

$$\mathbf{f}_t = \mathbf{B}_f \tanh H_f(\mathbf{e}_t). \tag{15}$$

In parallel, the slow pathway recursively updates its internal state using an adaptive retention rate,

$$\rho_t = \rho_{\min} + (\rho_{\max} - \rho_{\min})\sigma\big(H_\rho(\mathbf{e}_t)\big), \tag{16}$$

$$\mathbf{u}_t = \rho_t \mathbf{u}_{t-1} + (1-\rho_t)\tanh H_s(\mathbf{e}_t), \tag{17}$$

$$\mathbf{s}_t = \mathbf{B}_s \mathbf{u}_t. \tag{18}$$

A larger ρ_t preserves more of the accumulated state, whereas a smaller value allows the slow pathway to incorporate more current evidence. The retention mechanism therefore adapts the effective temporal memory instead of imposing a fixed smoothing horizon. The final dynamic estimate is computed as

$$\mathbf{y}_{s,v,t}^{\text{dyn}} = \mathbf{g}_{v,t} + \boldsymbol{\alpha}_{s,v,t} \odot \left(\mathbf{y}_{s,v,t}^{\text{phys}} - \mathbf{g}_{v,t}\right) + \mathbf{f}_{s,v,t} + \mathbf{s}_{s,v,t}. \tag{19}$$

The group trajectory supplies coarse temporal context, while the adaptive deviation term preserves trial-specific physiological evidence. The fast pathway responds to local changes, and the slow pathway accumulates persistent evidence over time. Their combination enables DRE to retain short-term sensitivity without sacrificing long-term continuity.

4 Experiments

4.1 Experimental Setup

The physiological backbone is pre-trained on public EEG and simultaneous EEG-fNIRS datasets [10, 33] and subsequently fine-tuned for continuous affect estimation. The training partition is used for model optimisation, GAT construction, and all training-derived temporal statistics, while the validation partition is used for checkpoint selection and limited hyperparameter tuning. Valence and Arousal

labels are normalised to $[0, 1]$, and performance is evaluated using mean absolute error (MAE). All baselines are retrained on the same downstream split following the architectures, objectives, and optimisation settings of their public implementations. In Table 1, Context-Augmented denotes each pipeline using its corresponding training-derived temporal information, whereas Context-Free denotes the physiological pipeline without such information.

4.2 Performance Analysis

Table 1 compares SAIN with the reproduced baselines. Incorporating training-derived temporal context consistently improves aggregate performance across the evaluated pipelines, demonstrating that physiological representations and temporal regularity provide complementary information for continuous affect estimation. Under the Context-Augmented setting, SAIN achieves the strongest overall and Arousal performance, while several baselines remain competitive on Valence. This pattern indicates that the two affective dimensions benefit differently from physiological representation learning and temporal refinement.

Table 2 examines the contribution of each physiological modality. The EEG-only and fNIRS-only models achieve comparable overall performance but exhibit complementary dimension-wise behaviour. EEG performs better on Arousal, whereas fNIRS is slightly stronger on Valence. Combining the two modalities improves both dimensions, showing that ACI integrates complementary electrophysiological and haemodynamic evidence rather than relying on a single dominant modality.

Table 3 further isolates the contributions of the two DRE pathways. Removing either pathway degrades overall performance, confirming that fast correction and slow recurrent accumulation capture distinct temporal properties. The larger decline after removing the slow pathway highlights the importance of persistent evidence, whereas the fast pathway provides additional sensitivity to local variations. This contribution is particularly relevant to Arousal, which exhibits stronger temporal variation than Valence. Together, these results show that SAIN improves continuous affect estimation through complementary multimodal interaction and multi-timescale temporal evolution.

4.3 Behavioural and Temporal Analysis

Figure 2(a) stratifies validation trials according to the trial-level error of the base model. SAIN reduces error in both difficult intervals, with a larger improvement on the hardest trials. Both DRE pathways contribute positively, indicating that fast correction and slow recurrent accumulation remain effective when rapid local variation coexists with sustained deviation from the group trajectory. The stronger improvement under these conditions supports the role of DRE in adapting population-level temporal context to trial-specific physiological dynamics rather than applying a uniform temporal correction.

Figure 2(b) examines sensor dependence through counterfactual channel masking. The spatially non-uniform error changes for both EEG and fNIRS indicate that SAPE retains informative channel relations instead of reducing each modality to an unstructured global representation. EEG sensitivity is concentrated around left fronto-central, central, and centro-parietal channels, including FC3, C5, and CP5, broadly agreeing with affective EEG findings across frontal, central, and parietal regions [49]. fNIRS sensitivity forms clusters across prefrontal and temporal regions, consistent with distributed haemodynamic involvement reported in previous studies [7, 32, 41]. Masking top-ranked channels causes greater degradation than masking low-ranked or randomly selected channels, supporting the effectiveness of SAPE in preserving structured modality-specific evidence. Joint masking produces the largest error increase, further indicating that ACI relies on complementary information from both modalities rather than substituting one physiological stream for the other.

Figure 2(c) projects category-level internal states into the Valence-Arousal space using a linear probe fitted on the training partition. SAIN produces trajectories that more closely follow the target geometry for Neutral, Happy, and Relax, with clearer category-specific directions than the base representations. This improved organisation is consistent with ACI combining adaptive modality selection, cross-modal discrepancy, and coordinated responses into a more affectively discriminative fused representation. Fear and Sad remain less separable, reflecting the greater overlap and temporal variability of negative affective states.

Figure 3(a) presents representative Valence and Arousal trajectories. SAIN follows rises, falls, and plateaus more closely than the base model, particularly around transition intervals and sustained deviations. The recovery of sharper local changes is consistent with the instantaneous fast pathway, while the improved tracking of prolonged rises and plateaus reflects the accumulated state of the slow pathway. Their combination allows DRE to recover trial-specific variation without sacrificing temporal continuity.

Figure 3(b) compares phase-wise performance after sequences of different lengths are mapped to a normalised trial phase. SAIN maintains lower average error across most phases, showing that the benefit of DRE extends throughout the trajectory rather than being confined to isolated transitions. The relatively stable improvements in the middle and later phases support progressive evidence accumulation in the slow recurrent state, whereas the gains around changing regions are consistent with the local responsiveness of the fast pathway. These phase-dependent behaviours agree with the intended separation between transient correction and persistent affective evolution.

5 Conclusion

We proposed SAIN, a hierarchical framework comprising SAPE, ACI, and DRE for continuous EEG-fNIRS affect estimation. SAPE preserves modality-specific physiological structure, ACI integrates complementary electrophysiological and haemodynamic evidence, and DRE coordinates fast local correction with slow recurrent accumulation around GAT. SAIN outperforms reproduced baselines and unimodal controls, while visual analyses reveal stronger gains on difficult trials, structured channel dependence, improved affective geometry, and sustained temporal improvement. These findings support the unified modelling of physiological structure, cross-modal interaction, and multi-timescale dynamics for continuous affect estimation. Future work will explore subject-aware cross-modal alignment and adaptive temporal modelling to better accommodate individual variability and modality-specific response delays.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 62302145), the Major Scientific and Technological Project of Anhui Provincial Science and Technology Innovation Platform (Grant No. 202305a12020012), the Fundamental Research Funds for the Central Universities (JZ2025HGTB0225).

References

[1] Sangtae Ahn and Sung C. Jun. 2017. Multi-Modal Integration of EEG–fNIRS for Brain–Computer Interfaces: Current Limitations and Future Directions. *Frontiers in Human Neuroscience* 11 (2017), 503. doi:10.3389/fnhum.2017.00503

[2] Feng-Qi Cui, Jinyang Huang, Sirui Zhao, Kun Li, Zhi Liu, Meng Li, Ziyu Jia, Dan Guo, and Meng Wang. 2026. PersoMoni: A Comprehensive Video-Based Benchmark Dataset for Fine-Grained Personality Assessment with 15 Trait Dimensions. *IEEE Transactions on Affective Computing* (2026).

[3] Feng-Qi Cui, Jinyang Huang, Sirui Zhao, Xinyu Li, Xin Yan, Ziyu Jia, and Xiaokang Zhou. 2026. Robust Low-Rank Sparse Framework for Video-Based Affective Computing. *IEEE Transactions on Consumer Electronics* (2026). doi:10.1109/TCE.2026.3697969

[4] Feng-Qi Cui, Zhen Lin, Xinlong Rao, Anyang Tong, Shiyao Li, Fei Wang, Changlin Chen, and Bin Liu. 2025. MICACL: Multi-Instance Category-Aware Contrastive Learning for Long-Tailed Dynamic Facial Expression Recognition. In *2025 IEEE International Symposium on Parallel and Distributed Processing with Applications*. IEEE, 601–608.

[5] Feng-Qi Cui, Anyang Tong, Jinyang Huang, Jie Zhang, Dan Guo, Zhi Liu, and Meng Wang. 2025. Learning from Heterogeneity: Generalizing Dynamic Facial Expression Recognition via Distributionally Robust Optimization. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 5587–5596. doi:10.1145/3746027.3755036

[6] Feng-Qi Cui, Anyang Tong, Jinyang Huang, Jie Zhang, Meng Li, Xin Yan, Linsheng Huang, Dan Guo, and Meng Wang. 2026. Towards Trustworthy Dynamic Facial Expression Recognition via Information Bottleneck Modeling. *IEEE Transactions on Information Forensics and Security* (2026).

[7] Condell Eastmond, Aseem Subedi, Suvranu De, and Xavier Intes. 2022. Deep Learning in fNIRS: A Review. *Neurophotonics* 9, 4 (2022), 041411. doi:10.1117/1.NPh.9.4.041411

[8] Kaining Fang, Jing Qu, Zixing Ding, Junhang Ding, and Lingguo Bu. 2026. ASAC-Net: A Novel Multimodal Alignment-Complementary Fusion Framework for EEG-fNIRS Emotion Recognition. *Information Fusion* 134 (2026), 104329. doi:10.1016/j.inffus.2026.104329

[9] Siamac Fazli, Jan Mehnert, Jens Steinbrink, Gabriel Curio, Arno Villringer, Klaus-Robert Müller, and Benjamin Blankertz. 2012. Enhanced Performance by a Hybrid NIRS–EEG Brain–Computer Interface. *NeuroImage* 59, 1 (2012), 519–529. doi:10.1016/j.neuroimage.2011.07.084

[10] Lufeng Feng, Baomin Xu, Haoran Zhang, Bihai Lin, Zuxuan Deng, Sidi Tao, Chenyu Liu, Shifan Jia, Li Duan, and Ziyu Jia. 2026. A Multimodal fNIRS–EEG Dataset for Unilateral Limb Motor Imagery. *arXiv preprint arXiv:2602.04299* (2026). https://arxiv.org/abs/2602.04299

[11] Ziyu Jia, Xiyang Cai, and Zehui Jiao. 2022. Multi-Modal Physiological Signals Based Squeeze-and-Excitation Network with Domain Adversarial Learning for Sleep Staging. *IEEE Sensors Journal* 22, 4 (2022), 3464–3471. doi:10.1109/JSEN.2022.3140383

[12] Ziyu Jia, Tingyu Du, Zhengyu Tian, Hongkai Li, Yong Zhang, and Chenyu Liu. 2025. A Multimodal BiMamba Network with Test-Time Adaptation for Emotion Recognition Based on Physiological Signals. In *Advances in Neural Information Processing Systems*, Vol. 38. https://proceedings.neurips.cc/paper_files/paper/2025/hash/d616a353c711f11c722e3f28d2d9e956-Abstract-Conference.html

[13] Ziyu Jia, Heng Liang, Yucheng Liu, Haichao Wang, and Tianzi Jiang. 2025. DistillSleepNet: Heterogeneous Multi-Level Knowledge Distillation via Teacher Assistant for Sleep Staging. *IEEE Transactions on Big Data* 11, 3 (2025), 1273–1284. doi:10.1109/TBDATA.2024.3453763

[14] Ziyu Jia, Youfang Lin, Jing Wang, Xiaojun Ning, Yuanlai He, Ronghao Zhou, Yuhan Zhou, and Li-wei H. Lehman. 2021. Multi-View Spatial-Temporal Graph Convolutional Networks with Domain Generalization for Sleep Stage Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29 (2021), 1977–1986. doi:10.1109/TNSRE.2021.3110665

[15] Ziyu Jia, Yucheng Liu, Haichao Wang, and Tianzi Jiang. 2025. Cross-Modal Knowledge Distillation for Enhanced Unimodal Emotion Recognition. *IEEE Transactions on Affective Computing* (2025), 1–13. doi:10.1109/TAFFC.2025.3583594

[16] Ziyu Jia, Haichao Wang, Yucheng Liu, and Tianzi Jiang. 2024. Mutual Distillation Extracting Spatial-Temporal Knowledge for Lightweight Multi-Channel Sleep Stage Classification. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1279–1289. doi:10.1145/3637528.3671981

[17] Weibang Jiang, Liming Zhao, and Bao-Liang Lu. 2024. LaBraM: Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *International Conference on Learning Representations*.

[18] Muhammad Umer Khan and Mustafa A. H. Hasan. 2020. Hybrid EEG–fNIRS BCI Fusion Using Multi-Resolution Singular Value Decomposition (MSVD). *Frontiers in Human Neuroscience* 14 (2020), 599802. doi:10.3389/fnhum.2020.599802

[19] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A Database for Emotion Analysis Using Physiological Signals. *IEEE Transactions on Affective Computing* 3, 1 (2012), 18–31. doi:10.1109/T-AFFC.2011.15

[20] Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. 2018. EEGNet: A Compact Convolutional Neural Network for EEG-Based Brain–Computer Interfaces. *Journal of Neural Engineering* 15, 5 (2018), 056013. doi:10.1088/1741-2552/aace8c

[21] Mingzheng Li, Xiao Sun, Xinke Wang, Feng-Qi Cui, and Xun Yang. 2026. CausalSymptom: Learning Causal Disentangled Representation for Depression Severity Estimation on Transcribed Clinical Interviews. *IEEE Transactions on Affective Computing* 17, 1 (2026), 277–289. doi:10.1109/TAFFC.2025.3613376

[22] Xinhui Li, Hao Chen, Minchao Wu, Qingwei Song, Fan Li, Jinyang Huang, Feng-Qi Cui, and Zhao Lv. 2026. MS^{2}FL: Modality-Shared and Modality-Specific Feature Learning for Multimodal Emotion Recognition. *ACM Transactions on Autonomous and Adaptive Systems* (2026).

[23] Xinyu Li, Jinyang Huang, Meng Wang, Peng Zhao, Feng-Qi Cui, Yuanhao Feng, Zheng Gong, Feiyu Han, and Fusang Zhang. 2026. Hiding Identity, Preserving Respiration: Semantic-Decoupled Privacy Protection for mmWave Radar Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 10, 3, Article 125 (Sept. 2026), 28 pages. doi:10.1145/3832038

[24] Yisheng Li and Shuqiang Wang. 2024. SCDM: Unified Representation Learning for EEG-to-fNIRS Cross-Modal Generation in MI-BCIs. *arXiv preprint arXiv:2407.04736* (2024).

[25] Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, Ziyang Ma, Xiaojiang Peng, Xie Chen, Ya Li, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2025. MER 2025: When Affective Computing Meets Large Language Models. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 13837–13842. doi:10.1145/3746027.3762007

[26] Zheng Lian, Xiaojiang Peng, Kele Xu, Ziyu Jia, Xinyi Che, Zebang Cheng, Fei Ma, Laizhong Cui, Yazhou Zhang, Xin Liu, Liang Yang, Jia Li, Fan Zhang, Liumeng Xue, Erik Cambria, Guoying Zhao, Bjorn W. Schuller, and Jianhua Tao. 2026. MER 2026: From Discriminative Emotion Recognition to Generative Emotion Understanding. *arXiv preprint arXiv:2604.19417* (2026). doi:10.48550/arXiv.2604.19417

[27] Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, Bin Liu, Rui Liu, Shan Liang, Ya Li, Jiangyan Yi, and Jianhua Tao. 2025. OV-MER: Towards Open-Vocabulary Multimodal Emotion Recognition. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*. PMLR, 37015–37050.

[28] Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, Zhuofan Wen, Siyuan Zhang, Shun Chen, Mingyu Xu, Ke Xu, Kang Chen, Lan Chen, Shan Liang, Ya Li, Jiangyan Yi, Bin Liu, and Jianhua Tao. 2023. Explainable Multimodal Emotion Recognition. *arXiv preprint arXiv:2306.15401* (2023). doi:10.48550/arXiv.2306.15401

[29] Zheng Lian, Licai Sun, Lan Chen, Haoyu Chen, Zebang Cheng, Fan Zhang, Ziyu Jia, Ziyang Ma, Fei Ma, Xiaojiang Peng, and Jianhua Tao. 2026. EmoPrefer: Can Large Language Models Understand Human Emotion Preferences?. In *International Conference on Learning Representations*. https://openreview.net/forum?id=EhA4znYsuG

[30] Fangyuan Liu, Sirui Zhao, Kang Yin, Tong Xu, and Enhong Chen. 2025. DepFormer: A Unified Framework with Bimodal Collaborative Transformer for Depression Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 13930–13936.

[31] Fangyuan Liu, Sirui Zhao, Zeyu Zhang, Jinyang Huang, Feng-Qi Cui, Bin Luo, Tong Xu, Meng Li, and Enhong Chen. 2026. Towards Trustworthy Depression Estimation via Disentangled Evidential Learning. *arXiv preprint arXiv:2604.16579* (2026). doi:10.48550/arXiv.2604.16579

[32] Pablo Ortega and Aldo Faisal. 2021. HemCNN: Deep Learning Enables Decoding of fNIRS Cortical Signals in Hand Grip Motor Tasks. *arXiv preprint arXiv:2103.05338* (2021).

[33] Milan Rybář, Riccardo Poli, and Ian Daly. 2025. Simultaneous EEG and fNIRS Recordings for Semantic Decoding of Imagined Animals and Tools. *Scientific Data* 12 (2025), 613. doi:10.1038/s41597-025-04967-0

[34] Jaeyoung Shin, Alexander von Lühmann, Do-Won Kim, Jan Mehnert, Han-Jeong Hwang, and Klaus-Robert Müller. 2018. Simultaneous Acquisition of EEG and NIRS during Cognitive Tasks for an Open Access Dataset. *Scientific Data* 5 (2018), 180003. doi:10.1038/sdata.2018.3

[35] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing* 3, 1 (2012), 42–55. doi:10.1109/T-AFFC.2011.25

[36] Hongkai Sui, Sirui Zhao, Xianglong Shi, Yiming Zhang, Jinyang Huang, Feng-Qi Cui, and Tong Xu. 2026. MaMoE: Leveraging Masked Autoencoders and Mixture-of-Experts for Identity-Free Micro-Action Recognition. In *2026 IEEE International Conference on Systems, Man, and Cybernetics*.

[37] Yanjia Sun, Hasan Ayaz, and Ali N. Akansu. 2020. Multimodal Affective State Assessment Using fNIRS + EEG and Spontaneous Facial Expression. *Brain Sciences* 10, 2 (2020), 85. doi:10.3390/brainsci10020085

[38] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6558–6569. doi:10.18653/v1/P19-1656

[39] Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. 2023. BIOT: Cross-Data Biosignal Learning in the Wild. In *Advances in Neural Information Processing Systems*, Vol. 36.

[40] Guoqi Yu, Juncheng Wang, Chen Yang, Jing Qin, Angelica I. Aviles-Rivero, and Shujun Wang. 2026. Decentralized Attention Fails Centralized Signals: Rethinking Transformers for Medical Time Series. In *International Conference on Learning Representations*. https://arxiv.org/abs/2602.18473

[41] Meryem A. Yücel, Alexander von Lühmann, Felix Scholkmann, Judit Gervain, Ippeita Dan, Hasan Ayaz, David Boas, Robert J. Cooper, Joseph Culver, Clare E. Elwell, et al. 2021. Best Practices for fNIRS Publications. *Neurophotonics* 8, 1 (2021), 012101. doi:10.1117/1.NPh.8.1.012101

[42] Tongtian Yue, Shuning Xue, Xuange Gao, Yepeng Tang, Longteng Guo, Jie Jiang, and Jing Liu. 2024. EEGPT: Unleashing the Potential of EEG Generalist Foundation Model by Autoregressive Pre-training. *arXiv preprint arXiv:2410.19779* (2024).

[43] Daoze Zhang, Zhizhang Yuan, Junru Chen, Kerui Chen, and Yang Yang. 2024. Brant-X: A Unified Physiological Signal Alignment Framework. *arXiv preprint arXiv:2409.00122* (2024).

[44] Sirui Zhao, Yu Bai, Jinyang Huang, Fangyuan Liu, Feng-Qi Cui, Xinqi Chen, Guo Cheng, Tong Xu, and Enhong Chen. 2026. MEC^2-TT: Multimodal Emotion Consistency Correction and Trajectory Tracking for Empathetic Dialogue Generation. In *Proceedings of the 34th ACM International Conference on Multimedia*.

[45] Sirui Zhao, Yubo Huang, Xinchen Yao, Zhengye Zhang, Jinyang Huang, Feng-Qi Cui, Shiwei Wu, and Enhong Chen. 2026. DynaForcing: Overcoming Dynamic Collapse in Self-Forcing Distillation for Streaming Avatar Generation. In *Proceedings of the 34th ACM International Conference on Multimedia*.

[46] Sirui Zhao, Huaying Tang, Shifeng Liu, Yangsong Zhang, Hao Wang, Tong Xu, Enhong Chen, and Cuntai Guan. 2022. ME-PLAN: A Deep Prototypical Learning with Local Attention Network for Dynamic Micro-Expression Recognition. *Neural Networks* 153 (2022), 427–443.

[47] Sirui Zhao, Huaying Tang, Xinglong Mao, Shifeng Liu, Yiming Zhang, Hao Wang, Tong Xu, and Enhong Chen. 2024. DFME: A New Benchmark for Dynamic Facial Micro-Expression Recognition. *IEEE Transactions on Affective Computing* 15, 3 (2024), 1371–1386.

[48] Sirui Zhao, Hanqing Tao, Yangsong Zhang, Tong Xu, Kun Zhang, Zhongkai Hao, and Enhong Chen. 2021. A Two-Stage 3D CNN-Based Learning Method for Spontaneous Micro-Expression Recognition. *Neurocomputing* 448 (2021), 276–289.

[49] Wei-Long Zheng and Bao-Liang Lu. 2015. Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development* 7, 3 (2015), 162–175. doi:10.1109/TAMD.2015.2431497