HB-IGCL: HEART-BRAIN INSTANCE-GLOBAL CONTRASTIVE LEARNING FOR EMOTION RECOGNITION USING LONG-DURATION ECG AND EEG

Anonymous authors
Paper under double-blind review

ABSTRACT

Emotion recognition is a key topic in affective computing, where electrocardiogram (ECG) and electroencephalogram (EEG) provide complementary peripheral and central physiological views. For long-duration recordings, a common practice is to divide each recording into short segments and train segment representations with the recording-level emotion label. This practice is convenient but conceptually restrictive: it assumes instance-level emotion supervision that is not actually observed, while emotional physiological responses may vary substantially across time. In addition, many ECG-EEG methods process the two modalities independently or combine them only through late fusion, underusing the natural correspondence between synchronized cardiac and neural observations. To address these issues, we propose **H**eart-**B**rain **I**nstance-**G**lobal **C**ontrastive **L**earning (HB-IGCL). Rather than treating every segment as an emotion-labeled sample, HB-IGCL learns whether a local physiological instance is compatible with the paired global context from the other modality within the same synchronized recording. ECG-to-EEG and EEG-to-ECG objectives provide bidirectional local-to-global contextual alignment, and a spatial-guided alignment module (SGAM) further relates lead-level spatial structure to global temporal context. We evaluate HB-IGCL in subject-independent experiments on AMIGOS, DREAMER, and MPED. Results show that the proposed framework improves representation learning across multiple encoders and datasets.

1 INTRODUCTION

Emotion recognition (ER) is a key research area in affective computing, human-computer interaction, and intelligent healthcare. In this context, physiological signals serve as a reliable foundation for ER. Specifically, EEG captures central nervous system activity, and ECG reflects cardiac rhythms and autonomic regulation. These two modalities are linked via the heart-brain axis, where cortical and subcortical systems modulate cardiac activity through autonomic pathways, while cardiovascular afferent signals contribute to interoception and emotional processing Thayer & Lane (2000); Critchley & Harrison (2013); Candia-Rivera et al. (2025). By offering complementary perspectives on emotional responses from central neural and peripheral physiological levels, research integrating EEG and ECG has emerged as a prominent frontier in affective computing.

Nevertheless, ER using long-duration physiological signals faces three interconnected challenges (Fig. 1). First, masked reconstruction, a widely used self-supervised learning approach for physiological signals Sarkar & Etemad (2022); Ghallab et al. (2025), can obscure sparse temporal evidence and make representation learning sensitive to which parts of a long recording are masked. Second, dividing continuous recordings into segments and directly assigning the recording-level label to every segment introduces an unobserved instance-label assumption: the dataset provides only recording-level emotion labels, not labels for each local physiological instance. Third, current fusion strategies typically treat ECG and EEG as parallel inputs without explicitly using the synchronized recording-level correspondence between local cardiac observations and global neural context, or vice versa.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

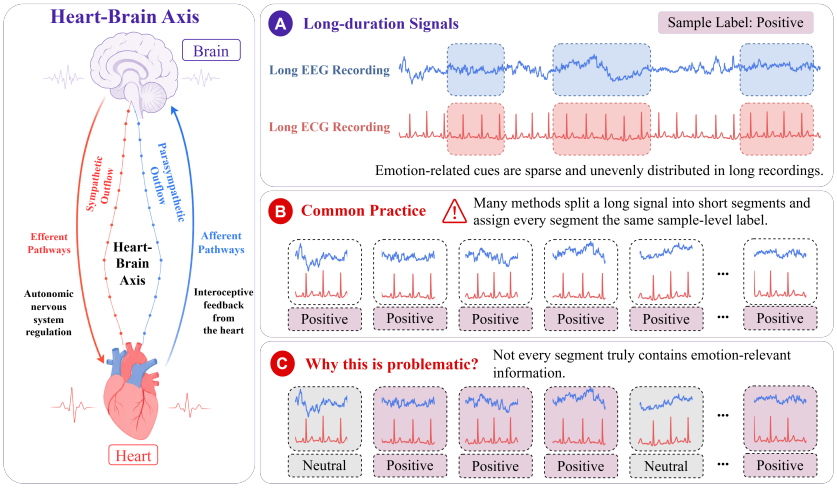


Figure 1: Motivation for long-duration ECG-EEG ER.

We propose Heart-Brain Instance-Global Contrastive Learning (HB-IGCL), a framework that learns directly from raw long-duration ECG-EEG recordings without imposing instance-level emotion labels. Each synchronized recording is treated as a bag of non-overlapping instances. For cross-modal contextual alignment, ECG-to-EEG (C2E) contrasts each ECG instance with the global EEG representation from the same recording, while EEG-to-ECG (E2C) contrasts each EEG instance with the global ECG representation from the same recording. Thus, a positive pair denotes a naturally paired local-to-global context within the same physiological episode, rather than claiming that every instance independently expresses the recording-level emotion. A spatial-guided alignment module (SGAM) further relates lead-level ECG-EEG structure to the global temporal context. Our contributions are threefold:

- We formulate non-masked, bidirectional instance-global contrastive pretraining for long ECG-EEG recordings, replacing unobserved instance-level emotion supervision with synchronized local-to-global cross-modal correspondence.
- Temporally, we learn bidirectional ECG-EEG contextual compatibility between local instances and paired global representations. Spatially, we design SGAM to align lead-level spatial features with global temporal context, yielding complementary spatiotemporal representations.
- Extensive experiments indicate that HB-IGCL improves representation learning across datasets, encoders, and downstream evaluation settings.

2 Related Work

ECG-based ER ECG provides noninvasive access to cardiac and autonomic responses. Existing ECG-based ER models use hypergraph transfer learning Ye et al. (2022), attention-enhanced CNNs Fan et al. (2023), global trajectory graphs Li et al. (2024), temporal convolutions Sweeney-Fanelli & Imtiaz (2024), or phase-space convolutional attention Yu et al. (2026). These methods improve cardiac representations, but do not use synchronized neural activity to contextualize local ECG responses.

EEG-based ER EEG contains rich temporal, spatial, and spectral information, but is affected by high dimensionality and cross-subject variability. Representative solutions combine channel-temporal attention with CNN-BiLSTM Jiang et al. (2022), multi-scale dynamic convolution with gated Transformer and temporal convolution Cheng et al. (2024), multi-band convolutional attention Tang et al. (2025), or attention-based channel selection Valderrama & Sheoran (2025). These methods remain brain-only models.

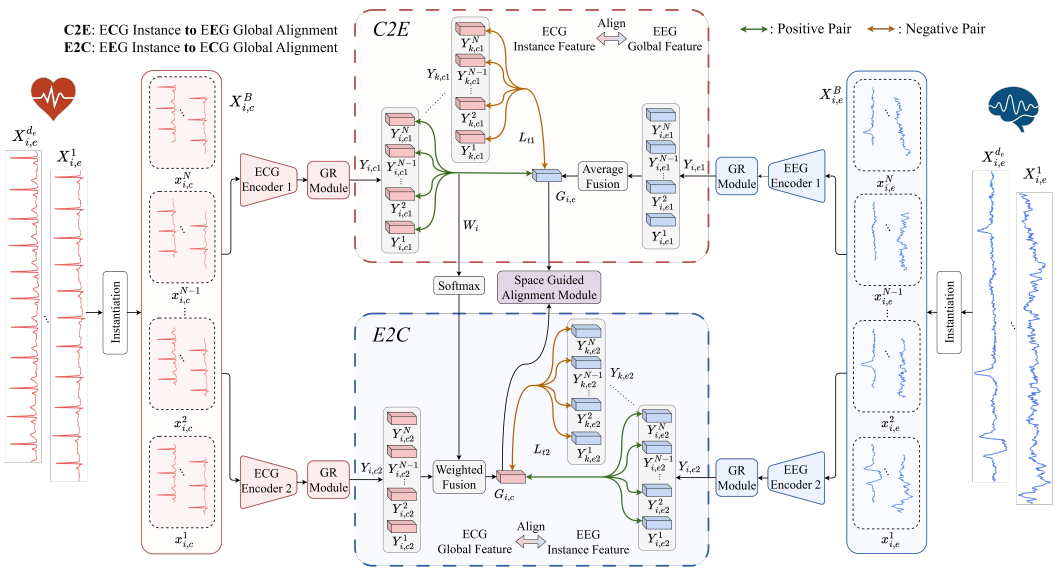


Figure 2: Overview of the proposed HB-IGCL framework.

Joint ECG-EEG ER Joint modeling exploits complementary cardiac and neural views. Prior work uses pretrained EEG features to supervise ECG representations Wu et al. (2025), engineered EEG-ECG connectivity Sedehi et al. (2025), or independently pretrained modality encoders followed by downstream concatenation Ghallab et al. (2025). Their interaction is thus expressed through one-way supervision, engineered connectivity, or late fusion.

3 METHOD

For bidirectional collaborative representation of ECG and EEG, this paper proposes the HB-IGCL framework, illustrated in Fig. 2. The key idea is to use the naturally synchronized ECG-EEG recording as supervision: a local instance from one modality is encouraged to match the paired global context from the other modality within the same recording, while contexts from other recordings in the batch serve as negatives. HB-IGCL consists of the following components: instantiation, instance encoder, a Gate and Refinement (GR) module, C2E, E2C, and SGAM.

Instantiation Let i index a synchronized ECG-EEG recording, obtained from the same subject under the identical experimental condition. We denote the ECG and EEG signals of recording i as $X_{i,c} = \{X^u_{i,c} \in \mathbb{R}^{1\times L} \mid u = 1, \ldots, d_c\}$ and $X_{i,e} = \{X^v_{i,e} \in \mathbb{R}^{1\times L} \mid v = 1, \ldots, d_e\}$, where u and v index the ECG and EEG leads, d_c and d_e denote the corresponding numbers of leads, and L is the temporal length of the signals. Each recording is divided into $N = \lceil L/T \rceil$ non-overlapping instances of length T, yielding the following instance sets for ECG and EEG:

$$X^B_{i,c} = \{x^j_{i,c} \in \mathbb{R}^{d_c\times T} \mid j = 1, \ldots, N\}, \tag{1}$$

$$X^B_{i,e} = \{x^j_{i,e} \in \mathbb{R}^{d_e\times T} \mid j = 1, \ldots, N\}. \tag{2}$$

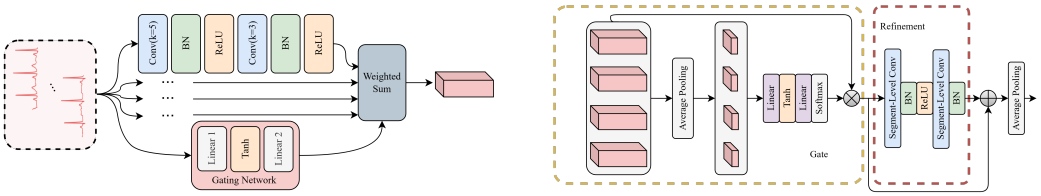
Here, the superscript B denotes the segmented bag, and j indexes the temporal instances within each bag.

Instance encoder HB-IGCL consists of four structurally identical but parameter-independent encoders (see Fig. 2). The default encoder used is our proposed Adaptive Multi-branch Instance Encoder (AMIE), which can be replaced by any other instance encoder. As shown in Fig. 3(a), AMIE comprises four parallel branches and an input-dependent gating network. Each branch operates with a different receptive field to capture complementary features within an instance. The gating network determines the importance weights of the four branches, and their outputs are fused via weighted summation to generate the final instance representation. This design enables the contribution of

each branch to adapt to the input content dynamically. Additional encoder variants are reported in Appendix E.

GR Long-duration recordings contain temporally heterogeneous physiological patterns, so instances may contribute differently to the final recording-level representation. To model this heterogeneity, we use the GR module (Fig. 3(b)), which integrates bag-level contextual information into each instance representation. A gating mechanism assigns contextual contribution scores to individual instances, weighting and aggregating their features accordingly, thereby calibrating instance representations with bag-level context.

However, the gating weight is a scalar applied to the whole instance, which only adjusts its overall contribution and cannot reorganize inter-channel or cross-instance relationships after weighting. To overcome this limitation, we further introduce a refinement layer to reconstruct and optimize the weighted instance representations. Finally, the outputs are passed through an average pooling layer to obtain the representation for each instance.



(a) AMIE encoder

(b) GR module

Figure 3: Architecture of the AMIE encoder and GR module.

C2E The long-duration ECG is processed through instantiation module, ECG Encoder 1, and GR module to produce ECG instance-level representations that incorporate global contextual information: $Y_{i,c1} = [Y^1_{i,c1}, \ldots, Y^N_{i,c1}]$. Similarly, the long-duration EEG is processed through instantiation module, EEG Encoder 1, and GR module to obtain EEG instance-level representations with global contextual information: $Y_{i,e1} = [Y^1_{i,e1}, \ldots, Y^N_{i,e1}]$, where $Y^j_{i,c1}, Y^j_{i,e1} \in \mathbb{R}^D$.

The EEG instance-level representation $Y_{i,e1}$ is averaged via pooling to yield its global representation: $G_{i,e} = \frac{1}{N}\sum_{j=1}^{N} Y^j_{i,e1}$.

To establish local-to-global cross-modal contextual correspondence, a positive pair is formed as $(Y^j_{i,c1}, G_{i,e})$. This pair means that the ECG instance and the EEG global representation come from the same synchronized recording, not that the ECG instance itself is assigned the recording-level emotion label. Negative pairs are formed as $(Y^j_{i,c1}, G_{k,e})$ for $i \neq k$, where $G_{k,e}$ is the EEG global representation from the k-th recording within the same training batch. Based on this setup, the following loss function is defined:

$$\mathcal{L}_{t1} = -\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\log\frac{\exp(w^j_{i,i}/\tau)}{\sum_{k=1}^{M}\exp(w^j_{i,k}/\tau)} \tag{3}$$

where $w^j_{i,k} = \mathrm{Cosine}(Y^j_{i,c1}, G_{k,e})$, M denotes the batch size, and $\tau > 0$ is the temperature parameter.

E2C Similarly, the long-duration ECG/EEG signal is processed through instantiation module, ECG/EEG Encoder 2, and GR module to obtain ECG/EEG instance-level representations: $Y_{i,c2} = [Y^1_{i,c2}, \ldots, Y^N_{i,c2}]$ and $Y_{i,e2} = [Y^1_{i,e2}, \ldots, Y^N_{i,e2}]$, where $Y^j_{i,c2}, Y^j_{i,e2} \in \mathbb{R}^D$.

The ECG instance representations $Y_{i,c2}$ are then aggregated into a global representation $G_{i,c}$ as follows:

$$G_{i,c} = \sum_{j=1}^{N} softmax(w^j_{i,i})Y^j_{i,c2} \tag{4}$$

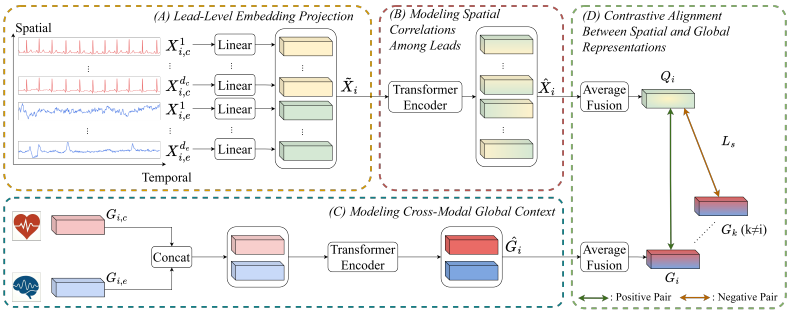


Figure 4: Architecture of the spatial-guided alignment module.

Symmetrically, E2C learns whether an EEG instance is compatible with the paired global ECG context. A positive pair is formed as $(Y^j_{i,e2}, G_{i,c})$, and negative pairs are formed as $(Y^j_{i,e2}, G_{k,c})$ for $i \neq k$, where $G_{k,c}$ is the ECG global representation from the k-th recording within the same training batch. Based on this setup, the following loss function is defined:

$$\mathcal{L}_{t2} = -\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\log\frac{\exp(s^j_{i,i}/\tau)}{\sum_{k=1}^{M}\exp(s^j_{i,k}/\tau)} \tag{5}$$

where $s^j_{i,k} = \text{Cosine}(Y^j_{i,e2}, G_{k,c})$.

SGAM The instance-global alignment focuses on temporal information but does not explicitly model the relationships between ECG and EEG leads. This is complemented by the SGAM (shown as Fig. 4), which introduces lead-level structural modeling. First, each lead of the long-duration ECG and EEG is linearly projected into a D-dimensional embedding. These embeddings are then concatenated to obtain $\widetilde{X}_i \in \mathbb{R}^{(d_c+d_e)\times D}$. A Transformer encoder Vaswani et al. (2017) is applied to $\widetilde{X}_i$ for contextual modeling, producing $\widehat{X}_i \in \mathbb{R}^{(d_c+d_e)\times D}$ to capture both intra-modal and cross-modal spatial relationships. Finally, average fusion is performed over the lead dimension to derive the global spatial representation $Q_i \in \mathbb{R}^D$.

Simultaneously, the global representations of ECG and EEG, $G_{i,c}$ and $G_{i,e}$, are concatenated and fed into another Transformer encoder. This module models their global spatial interaction and outputs $\widehat{G}_i \in \mathbb{R}^{2\times D}$. Average fusion over the modality dimension is then applied to obtain the temporal context representation $G_i \in \mathbb{R}^D$.

The spatial and temporal pathways are integrated via contrastive alignment. Here, (Q_i, G_i) forms a positive pair, while Q_i and the temporal representations of other recordings within the batch form negative pairs. The corresponding SGAM loss is defined as:

$$\mathcal{L}_s = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp(Z_{i,i}/\tau)}{\sum_{k=1}^{M}\exp(Z_{i,k}/\tau)} \tag{6}$$

where $Z_{i,k} = \text{Cosine}(Q_i, G_k)$.

Finally, the overall pretraining objective of HB-IGCL is formulated as:

$$\mathcal{L} = \beta\left[\alpha\mathcal{L}_{t1} + (1-\alpha)\mathcal{L}_{t2}\right] + (1-\beta)\mathcal{L}_s \tag{7}$$

where $\alpha \in [0,1]$, $\beta \in [0,1]$, and τ is a shared contrastive temperature hyperparameter.

Downstream fine-tuning To transfer the instance-representation capability and cross-modal alignment learned during pretraining to the downstream ER task, a downstream task pipeline is designed as illustrated in Fig. 5.

Specifically, the long-duration ECG signal is first split into instances, encoded, subsequently passed through global average pooling (GAP), then concatenated to obtain its comprehensive representation vector $Y_{i,c}$. Similarly, the long-duration EEG signal is processed to derive its comprehensive

representation vector $Y_{i,e}$. Subsequently, cross-attention is applied, where $Y_{i,c}$ serves as the Query and $Y_{i,e}$ serves as both the Key and Value. This allows the ECG features to adaptively aggregate complementary emotional information from the EEG modality. Finally, a linear classification head is used for the final emotion classification.

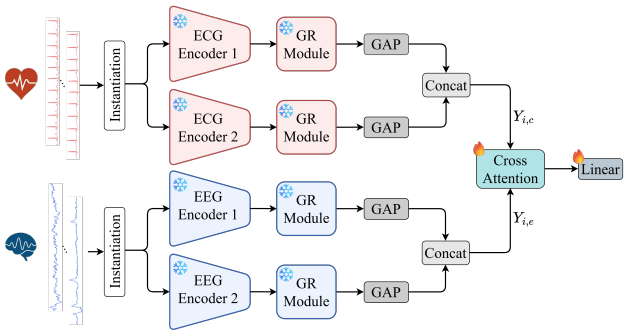


Figure 5: Pipeline of the downstream task.

4 Experiments

4.1 Datasets

We evaluate HB-IGCL on three publicly available multimodal emotion datasets with synchronized ECG and EEG signals: AMIGOS Miranda-Correa et al. (2021), DREAMER Katsigiannis & Ramzan (2018), and MPED Song et al. (2019). Following prior ECG ER studies Xu et al. (2025); Zhong et al. (2025), we formulate AMIGOS and DREAMER as four-class classification tasks based on the four valence-arousal quadrants. For MPED, we adopt the three-class classification scheme specified in its original protocol. All datasets are divided into mutually exclusive training, validation, and test subject partitions to ensure subject-independent evaluation. Detailed dataset statistics and splitting configurations are provided in the Appendix A.

4.2 Experimental Settings

Each ECG and EEG lead was independently normalized using min-max scaling. As recording durations varied across experimental conditions, all recordings were right-padded with zeros to match the longest recording in each dataset. This resulted in aligned lengths of 19,886, 50,432, and 73,750 samples for AMIGOS, DREAMER, and MPED, respectively. When the aligned length was not divisible by the instance length, the final instance was also zero-padded. To prevent zero-padded instances from affecting the model, a validity mask is applied to the encoders, GR modules, and loss computation so that invalid instances make no contribution.

All models are optimized using AdamW with an initial learning rate of 10^{-4} and a weight decay of 0.01. Each configuration was repeated across five random seeds: 1111, 2222, 3333, 4444, and 5555. All experiments were run on an NVIDIA GeForce RTX 4090, and performance is reported as the mean $\pm$ standard deviation of F1 and accuracy (Acc).

4.3 Results

In this work, we evaluate five encoders, namely ModernTCN Luo & Wang (2024), PatchTST Nie et al. (2023), PatchTST+, OS-CNN Tang et al. (2022), and AMIE, under four pre-training frameworks: LMM Ghallab et al. (2025), EMM (a variant of LMM), CCM Wu et al. (2025), and HB-IGCL. PatchTST+ denotes an adapted variant of PatchTST for the present task. The ECG and EEG branches use the same encoder architecture but are parameterized independently. Following pre-training, all encoders are evaluated using the same downstream transfer protocol. Architectural details of PatchTST+ and implementation details of EMM, LMM, and CCM are provided in

Table 1: The comparison results on AMIGOS, DREAMER, and MPED.

Framework	ModernTCN F1 (%)	ModernTCN Acc(%)	PatchTST F1(%)	PatchTST Acc(%)	PatchTST+ F1(%)	PatchTST+ Acc(%)	OS-CNN F1(%)	OS-CNN Acc(%)	AMIE F1(%)	AMIE Acc(%)
Dataset: AMIGOS										
EMM	30.76$_{\pm 6.55}$	35.64$_{\pm 4.61}$	28.48$_{\pm 4.71}$	32.31$_{\pm 4.82}$	29.77$_{\pm 3.09}$	33.59$_{\pm 4.54}$	22.34$_{\pm 3.76}$	31.28$_{\pm 3.20}$	23.77$_{\pm 4.81}$	31.54$_{\pm 4.41}$
LMM	35.67$_{\pm 2.91}$	37.44$_{\pm 2.05}$	32.51$_{\pm 3.91}$	39.49$_{\pm 5.15}$	27.57$_{\pm 2.14}$	34.87$_{\pm 2.97}$	30.97$_{\pm 3.89}$	35.13$_{\pm 3.10}$	62.45$_{\pm 1.43}$	65.38$_{\pm 1.62}$
CCM	24.38$_{\pm 4.82}$	32.31$_{\pm 4.82}$	30.62$_{\pm 2.96}$	39.74$_{\pm 3.03}$	29.05$_{\pm 3.24}$	**39.49$_{\pm 5.28}$**	29.37$_{\pm 3.72}$	37.18$_{\pm 4.59}$	56.07$_{\pm 2.99}$	57.69$_{\pm 2.81}$
HB-IGCL	**46.86$_{\pm 1.95}$**	**48.20$_{\pm 2.64}$**	**38.19$_{\pm 4.52}$**	**42.56$_{\pm 4.83}$**	**43.24$_{\pm 3.78}$**	37.95$_{\pm 4.63}$	**40.46$_{\pm 7.06}$**	**43.33$_{\pm 3.75}$**	**66.23$_{\pm 1.71}$**	**68.21$_{\pm 1.49}$**
Dataset: DREAMER										
EMM	25.54$_{\pm 2.75}$	43.33$_{\pm 6.71}$	23.38$_{\pm 2.76}$	37.50$_{\pm 4.21}$	21.69$_{\pm 4.38}$	33.61$_{\pm 5.30}$	26.13$_{\pm 2.51}$	41.11$_{\pm 6.61}$	27.50$_{\pm 4.66}$	35.55$_{\pm 8.03}$
LMM	30.29$_{\pm 4.45}$	**50.83$_{\pm 1.14}$**	25.06$_{\pm 3.61}$	43.88$_{\pm 1.41}$	23.31$_{\pm 3.60}$	39.72$_{\pm 3.24}$	28.83$_{\pm 6.20}$	47.78$_{\pm 3.89}$	39.39$_{\pm 5.37}$	55.83$_{\pm 4.76}$
CCM	25.80$_{\pm 2.53}$	49.17$_{\pm 3.24}$	19.32$_{\pm 0.81}$	44.17$_{\pm 1.62}$	20.48$_{\pm 2.87}$	**45.55$_{\pm 0.06}$**	24.38$_{\pm 2.28}$	47.50$_{\pm 3.09}$	46.36$_{\pm 6.74}$	56.39$_{\pm 4.18}$
HB-IGCL	**31.73$_{\pm 5.05}$**	50.00$_{\pm 7.03}$	**28.99$_{\pm 2.02}$**	**44.44$_{\pm 3.29}$**	**26.99$_{\pm 5.08}$**	40.83$_{\pm 8.27}$	**31.34$_{\pm 1.45}$**	**48.89$_{\pm 2.55}$**	**50.53$_{\pm 2.89}$**	**63.61$_{\pm 2.04}$**
Dataset: MPED										
EMM	28.93$_{\pm 3.21}$	56.19$_{\pm 2.43}$	31.16$_{\pm 4.24}$	48.33$_{\pm 8.17}$	32.17$_{\pm 3.51}$	57.38$_{\pm 4.84}$	30.26$_{\pm 2.76}$	54.76$_{\pm 3.10}$	37.44$_{\pm 2.97}$	55.95$_{\pm 4.94}$
LMM	33.04$_{\pm 3.63}$	58.09$_{\pm 1.90}$	34.64$_{\pm 2.28}$	58.33$_{\pm 1.99}$	32.88$_{\pm 2.72}$	55.24$_{\pm 2.33}$	35.59$_{\pm 1.83}$	58.33$_{\pm 1.06}$	81.21$_{\pm 2.14}$	85.00$_{\pm 1.61}$
CCM	25.91$_{\pm 2.21}$	57.85$_{\pm 0.95}$	33.04$_{\pm 4.08}$	**59.52$_{\pm 1.68}$**	35.15$_{\pm 2.49}$	58.81$_{\pm 2.21}$	34.73$_{\pm 1.97}$	57.85$_{\pm 1.61}$	82.89$_{\pm 8.15}$	84.76$_{\pm 8.23}$
HB-IGCL	**35.47$_{\pm 7.49}$**	**59.04$_{\pm 1.78}$**	**36.84$_{\pm 1.69}$**	58.09$_{\pm 1.90}$	**44.88$_{\pm 2.43}$**	**62.62$_{\pm 3.16}$**	**39.77$_{\pm 4.20}$**	**59.04$_{\pm 5.81}$**	**84.66$_{\pm 2.04}$**	**85.71$_{\pm 2.00}$**

Table 2: Ablation results of training objectives and fine-tuning strategies. Best results in bold.

Variant	AMIGOS F1(%)	AMIGOS Acc(%)	DREAMER F1(%)	DREAMER Acc(%)	MPED F1(%)	MPED Acc(%)
Only $\mathcal{L}_{t1}$	62.13$_{\pm 2.49}$	64.10$_{\pm 2.93}$	43.92$_{\pm 4.78}$	61.94$_{\pm 3.24}$	84.23$_{\pm 1.04}$	85.71$_{\pm 0.75}$
Only $\mathcal{L}_{t2}$	57.31$_{\pm 3.25}$	60.77$_{\pm 2.38}$	46.98$_{\pm 0.89}$	63.61$_{\pm 1.36}$	84.52$_{\pm 2.16}$	**86.43$_{\pm 1.78}$**
w/o $\mathcal{L}_s$	60.32$_{\pm 1.26}$	62.05$_{\pm 1.31}$	48.89$_{\pm 4.28}$	62.22$_{\pm 2.70}$	83.85$_{\pm 1.88}$	85.47$_{\pm 1.75}$
Weighted EEG	59.68$_{\pm 6.74}$	64.36$_{\pm 3.48}$	48.97$_{\pm 5.78}$	63.61$_{\pm 2.84}$	83.80$_{\pm 2.05}$	85.47$_{\pm 1.94}$
HB-IGCL	**66.23$_{\pm 1.71}$**	**68.21$_{\pm 1.49}$**	**50.53$_{\pm 2.89}$**	**63.61$_{\pm 2.04}$**	**84.66$_{\pm 2.04}$**	85.71$_{\pm 2.00}$
Rand. Init.	56.31$_{\pm 6.11}$	60.25$_{\pm 3.34}$	46.75$_{\pm 2.03}$	62.50$_{\pm 0.88}$	85.86$_{\pm 6.09}$	87.62$_{\pm 5.25}$
Full-FT	61.90$_{\pm 2.63}$	63.84$_{\pm 2.05}$	46.24$_{\pm 3.27}$	61.39$_{\pm 1.84}$	**89.51$_{\pm 1.85}$**	**90.24$_{\pm 1.39}$**
Frozen-FT	**66.23$_{\pm 1.71}$**	**68.21$_{\pm 1.49}$**	**50.53$_{\pm 2.89}$**	**63.61$_{\pm 2.04}$**	84.66$_{\pm 2.04}$	85.71$_{\pm 2.00}$

Appendix B. Further adaptation details and analyses of LMM, EMM, and CCM are provided in Appendices F-H.

Table 1 summarizes the results for all pretraining framework-encoder combinations. For a given encoder, the highest performance across pre-training frameworks is highlighted in bold. For a given pre-training framework, the best F1 score is indicated in red text, while the highest Acc is marked with a gray background.

Performance comparison of pre-training frameworks A vertical comparison in Table 1 shows that HB-IGCL obtains the best F1 scores for all encoders on all three datasets, with competitive or best Acc in most cases. This indicates that the proposed local-to-global cross-modal objective is not tied to a specific backbone.

Performance comparison of encoders A horizontal comparison under the same pretraining framework shows that AMIE provides the strongest overall results in most settings. Together, the vertical and horizontal comparisons indicate that HB-IGCL and AMIE contribute complementary benefits.

Instance score analysis To examine the learned cross-modal contextual compatibility scores, we ranked ECG and EEG instances by their C2E scores $w^j_{i,i}$ and E2C scores $s^j_{i,i}$ on AMIGOS, then removed the same fraction of instances from both modalities from 10% to 50% in 10% increments. The results are shown in Fig.6.

Removing high-scoring instances generally causes larger degradation than removing low-scoring ones, indicating that the scores are associated with downstream utility. They are not ground-truth emotional relevance labels, but local-to-global cross-modal compatibility estimates. Fluctuations at large removal ratios are expected because removing many instances distorts the original temporal structure.

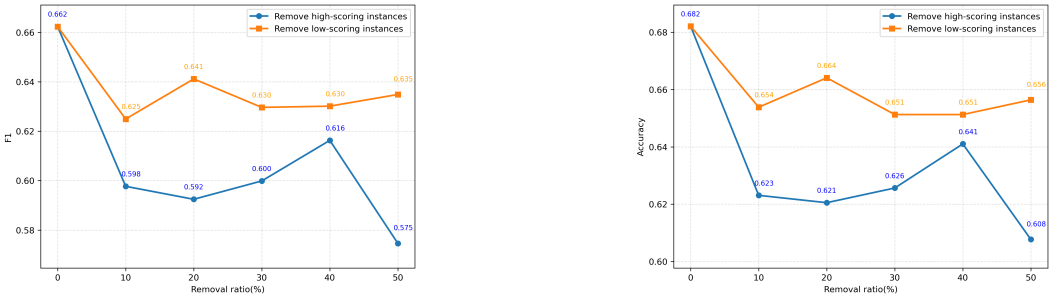


Figure 6: Experimental results for instance removing on the AMIGOS dataset: F1 (left) and Acc (right).

Table 3: Ablation results of modalities and representation levels. Best results in bold.

Variant	AMIGOS F1(%)	AMIGOS Acc(%)	DREAMER F1(%)	DREAMER Acc(%)	MPED F1(%)	MPED Acc(%)
ECG-Ins	$25.33_{\pm 6.26}$	$35.90_{\pm 6.39}$	$29.29_{\pm 9.52}$	$48.89_{\pm 9.27}$	$34.28_{\pm 7.00}$	$52.14_{\pm 8.43}$
ECG-Glo	$21.74_{\pm 6.26}$	$28.46_{\pm 6.20}$	$37.88_{\pm 3.93}$	$55.56_{\pm 4.56}$	$29.89_{\pm 5.59}$	$52.86_{\pm 7.32}$
EEG-Ins	$24.93_{\pm 5.68}$	$33.08_{\pm 5.46}$	$36.28_{\pm 7.02}$	$55.55_{\pm 5.83}$	$31.99_{\pm 3.91}$	$51.67_{\pm 6.71}$
EEG-Glo	$26.86_{\pm 6.04}$	$36.92_{\pm 7.96}$	$25.03_{\pm 8.57}$	$43.89_{\pm 10.99}$	$34.90_{\pm 7.27}$	$55.00_{\pm 7.24}$
ECG-Full	$23.32_{\pm 7.75}$	$34.87_{\pm 9.15}$	$39.90_{\pm 3.22}$	$56.95_{\pm 4.12}$	$35.35_{\pm 4.95}$	$60.95_{\pm 1.39}$
EEG-Full	$28.34_{\pm 5.68}$	$42.82_{\pm 8.33}$	$42.17_{\pm 3.36}$	$60.56_{\pm 3.99}$	$39.20_{\pm 4.45}$	$60.00_{\pm 4.62}$
ECG-EEG-Ins (Concat)	$22.64_{\pm 7.54}$	$31.79_{\pm 8.90}$	$33.24_{\pm 9.95}$	$49.72_{\pm 11.43}$	$39.12_{\pm 1.67}$	$60.95_{\pm 2.05}$
ECG-EEG-Ins (CA)	$50.31_{\pm 1.69}$	$58.97_{\pm 8.13}$	$46.69_{\pm 1.56}$	$\mathbf{64.45_{\pm 1.42}}$	$82.31_{\pm 1.91}$	$85.71_{\pm 1.51}$
ECG-EEG-Glo (Concat)	$25.62_{\pm 6.58}$	$38.72_{\pm 10.17}$	$38.99_{\pm 4.79}$	$58.61_{\pm 3.87}$	$35.88_{\pm 5.96}$	$59.28_{\pm 1.90}$
ECG-EEG-Glo (CA)	$51.13_{\pm 7.73}$	$58.98_{\pm 2.92}$	$44.99_{\pm 0.84}$	$62.22_{\pm 0.56}$	$83.71_{\pm 3.26}$	$\mathbf{85.95_{\pm 2.96}}$
ECG-EEG-All (Concat)	$27.45_{\pm 6.23}$	$38.46_{\pm 10.60}$	$43.07_{\pm 1.79}$	$61.39_{\pm 2.04}$	$42.24_{\pm 7.25}$	$63.81_{\pm 5.67}$
ECG-EEG-All (CA)	$\mathbf{66.23_{\pm 1.71}}$	$\mathbf{68.21_{\pm 1.49}}$	$\mathbf{50.53_{\pm 2.89}}$	$63.61_{\pm 2.04}$	$\mathbf{84.66_{\pm 2.04}}$	$85.71_{\pm 2.00}$

Table 4: Ablation results of GR on DREAMER. Best results in bold.

Variant	AMIE F1(%)	AMIE Acc(%)	ModernTCN F1(%)	ModernTCN Acc(%)	PatchTST F1(%)	PatchTST Acc(%)	PatchTST+ F1(%)	PatchTST+ Acc(%)	OS-CNN F1(%)	OS-CNN Acc(%)
w/o GR	$25.04_{\pm 1.46}$	$43.61_{\pm 2.99}$	$\mathbf{26.04_{\pm 1.97}}$	$46.39_{\pm 2.08}$	$11.93_{\pm 2.30}$	$21.94_{\pm 3.87}$	$11.22_{\pm 4.14}$	$21.39_{\pm 5.31}$	$\mathbf{25.02_{\pm 0.70}}$	$44.16_{\pm 1.36}$
w/o Refine	$23.24_{\pm 1.97}$	$39.72_{\pm 2.99}$	$25.94_{\pm 1.59}$	$\mathbf{46.67_{\pm 2.57}}$	$11.50_{\pm 1.98}$	$20.83_{\pm 2.32}$	$11.81_{\pm 3.08}$	$21.39_{\pm 3.58}$	$24.99_{\pm 1.82}$	$42.78_{\pm 3.77}$
GR → TE	$23.20_{\pm 2.13}$	$40.00_{\pm 3.87}$	$24.91_{\pm 1.40}$	$45.28_{\pm 2.08}$	$13.32_{\pm 7.91}$	$22.78_{\pm 8.17}$	$15.06_{\pm 1.09}$	$25.83_{\pm 3.79}$	$22.91_{\pm 2.03}$	$40.00_{\pm 4.06}$
GR	$\mathbf{42.47_{\pm 1.16}}$	$\mathbf{60.55_{\pm 1.89}}$	$24.80_{\pm 2.58}$	$46.39_{\pm 4.18}$	$\mathbf{15.92_{\pm 2.83}}$	$\mathbf{32.22_{\pm 11.33}}$	$\mathbf{18.55_{\pm 3.65}}$	$\mathbf{39.16_{\pm 8.58}}$	$24.84_{\pm 2.48}$	$\mathbf{44.45_{\pm 3.51}}$

4.4 Ablation Studies

Impact of training objectives To assess the impact of different training objectives, we design four ablation variants: **Only** $\mathcal{L}_{t1}$: the overall loss is $\mathcal{L}_{t1}$; **Only** $\mathcal{L}_{t2}$: the overall loss is $\mathcal{L}_{t2}$; **w/o** $\mathcal{L}_s$: the loss term $\mathcal{L}_s$ is removed; **Weighted EEG**: the weighting mechanism is transferred from ECG to EEG. The results are shown in Table 2.

Using only $\mathcal{L}_{t1}$ or $\mathcal{L}_{t2}$ is generally weaker than the full objective, suggesting that C2E and E2C provide complementary contextual supervision. Removing $\mathcal{L}_s$ also reduces performance, while the Weighted EEG variant underperforms the default ECG-weighted aggregation. A sensitivity analysis of the default loss function hyperparameters $(\alpha = 0.5, \beta = 0.8)$ is provided in the Appendix C.

Impact of fine-tuning strategies To further examine how fine-tuning strategies affect downstream performance, we compare three approaches: **Rand. Init.**: the model is trained from scratch using downstream labels, without pre-trained weights; **Full-FT**: all parameters are updated after loading HB-IGCL pre-trained weights; **Frozen-FT**: pre-trained weights are frozen, and only the interaction attention module and the classification head are optimized. The results are shown in Table 2.

Pretraining generally improves over Rand. Init. on AMIGOS and DREAMER, and Frozen-FT performs best on these two small-sample datasets. On MPED, Full-FT achieves the best result, suggesting that updating the pretrained encoder can further adapt the representation to its three-class label space.

Impact of modalities and representation levels To assess modality type, representation level, and fusion strategy, we compare **Ins**, **Glo**, and **Full**, denoting instance-level, global-level, and concatenated representations. ECG-side and EEG-side single-modality variants are defined accordingly. Cross-modal variants fuse ECG and EEG features at instance, global, or full levels using concatenation (Concat) or cross-attention (CA); HB-IGCL corresponds to ECG-EEG-Full (CA).

Table 3 shows that multimodal variants generally outperform single-modality ones, and cross-attention is usually stronger than direct concatenation. This supports adaptive ECG-EEG interaction across both instance-level and global-level representations.

Impact of GR To validate the gate layer and refinement block in GR, we evaluate three variants on DREAMER: **w/o GR**, **w/o Refine**, and **GR** $\rightarrow$ **TE**, where GR is replaced by a Transformer encoder over average-pooled instance representations. Results in Table 4 show that GR delivers the best overall performance. For AMIE, w/o Refine is weaker than w/o GR, indicating that the gate layer alone provides only coarse instance weighting and requires refinement for stable representation learning.

For PatchTST and PatchTST+, GR $\rightarrow$ TE can be stronger than simpler variants, but the full GR design gives the strongest overall result with AMIE, the default encoder of HB-IGCL.

Impact of instance length We vary instance length T while fixing all other pretraining and fine-tuning configurations. Full results are provided in Appendix D. The best overall trade-off is achieved at $T = 4$ seconds: shorter instances may lack physiological context, whereas longer ones reduce temporal units and can mix heterogeneous patterns.

5 Conclusion

We presented HB-IGCL for subject-independent ER from long-duration ECG-EEG recordings. HB-IGCL learns bidirectional local-to-global cross-modal contextual compatibility within synchronized recordings, while SGAM relates lead-level spatial structure to global temporal context. Experiments across three datasets and multiple encoders indicate that HB-IGCL improves ECG-EEG representation learning under recording-level supervision.

6 Limitations

This study has three limitations. First, the analysis uses controlled video-elicitation datasets with limited subject numbers, and the fixed 4-second instance may not generalize to unconstrained recordings or other affective states. Second, HB-IGCL learns statistical cross-modal contextual compatibility rather than causal physiological mechanisms. Third, without ground-truth instance-level annotations, instance-removal analysis provides only indirect evidence about downstream utility. Future work should explore adaptive segmentation, causal validation, missing-modality robustness, and externally annotated local events.

AI use statement

Generative AI tools were used to assist with language polishing, LaTeX format conversion, and editorial organization. The scientific ideas, experimental design, implementation, analyses, and claims were specified and reviewed by the authors. All AI-assisted text and artifacts were manually checked, and the authors take responsibility for the final content of this submission.

Ethics statement

This work uses publicly available ECG-EEG emotion-recognition datasets collected in controlled experimental settings. We do not release new human-subject data. The potential application of physiological emotion recognition raises privacy and misuse risks; therefore, models based on such signals should be deployed only with informed consent, appropriate data protection, and clear limits on decision-making authority.

Reproducibility statement

We provide the model definitions, training scripts, data-processing descriptions, subject-independent splits, hyperparameter settings, and additional ablations in the appendix and supplementary code. All reported configurations are repeated with five random seeds, and performance is reported as mean and standard deviation.

References

Diego Candia-Rivera, Luca Faes, Fabrizio De Vico Fallani, and Mario Chavez. Measures and models of brain-heart interactions. *IEEE Reviews in Biomedical Engineering*, pp. 1–17, 2025. doi: 10.1109/RBME.2025.3529363.

Zhuoling Cheng, Xuekui Bu, Qingnan Wang, Tao Yang, and Jihui Tu. EEG-based emotion recognition using multi-scale dynamic CNN and gated transformer. *Scientific Reports*, 14:31319, 2024. doi: 10.1038/s41598-024-82705-z.

Hugo D. Critchley and Neil A. Harrison. Visceral influences on brain and behavior. *Neuron*, 77(4): 624–638, 2013. doi: 10.1016/j.neuron.2013.02.008.

Tianqi Fan, Sen Qiu, Zhelong Wang, Hongyu Zhao, Junhan Jiang, Yongzhen Wang, Junnan Xu, Tao Sun, and Nan Jiang. A new deep convolutional neural network incorporating attentional mechanisms for ECG emotion recognition. *Computers in Biology and Medicine*, 159:106938, 2023. doi: 10.1016/j.compbiomed.2023.106938.

Youssef Ghallab, Omar Iraqy, Mohamed Kandil, Mohamed Ashraf, Saadeldine Eletter, Morougue Ghazal, Ayman Khalafallah, and Nagwa El-Makky. Leveraging foundational models and simple fusion for multi-modal physiological signal analysis. In *NeurIPS 2025 Workshop on Foundation Models for the Brain and Body*, 2025.

Lei Jiang, Panote Siriaraya, Dongeun Choi, Fangmeng Zeng, and Noriaki Kuwahara. Electroencephalogram signals emotion recognition based on convolutional neural network-recurrent neural network framework with channel-temporal attention mechanism for older adults. *Frontiers in Aging Neuroscience*, 14:945024, 2022. doi: 10.3389/fnagi.2022.945024.

Stamos Katsigiannis and Naeem Ramzan. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1):98–107, 2018. doi: 10.1109/JBHI.2017.2688239.

Ya Li, Runxi Tan, Tianxin Lin, Qing Liu, Chang-Dong Wang, and Min Chen. ER-GET: Emotion recognition based on global ECG trajectory. *IEEE Journal of Biomedical and Health Informatics*, 28(9):5201–5213, 2024. doi: 10.1109/JBHI.2024.3403188.

Donghao Luo and Xue Wang. Moderntcn: A modern pure convolution structure for general time series analysis. In *ICLR*, 2024.

Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. AMIGOS: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12(2):479–493, 2021. doi: 10.1109/TAFFC.2018.2884461.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Pritam Sarkar and Ali Etemad. Self-supervised ECG representation learning for emotion recognition. *IEEE Transactions on Affective Computing*, 13(3):1541–1554, 2022. doi: 10.1109/TAFFC.2020.3014842.

Javid Farhadi Sedehi, Nader Jafarnia Dabanloo, Keivan Maghooli, and Ali Sheikhani. Develop an emotion recognition system using jointly connectivity between electroencephalogram and electrocardiogram signals. *Heliyon*, 11:e41767, 2025. doi: 10.1016/j.heliyon.2025.e41767.

Tengfei Song, Wenming Zheng, Cheng Lu, Yuan Zong, Xilei Zhang, and Zhen Cui. MPED: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access*, 7: 12177–12191, 2019. doi: 10.1109/ACCESS.2019.2891579.

Timothy C. Sweeney-Fanelli and Masudul H. Imtiaz. ECG-based automated emotion recognition using temporal convolution neural networks. *IEEE Sensors Journal*, 24(18):29039–29046, 2024. doi: 10.1109/JSEN.2024.3434479.

Wei Tang, Linhui Fan, Xuefen Lin, and Yifan Gu. EEG emotion recognition based on efficient-capsule network with convolutional attention. *Biomedical Signal Processing and Control*, 103: 107473, 2025. doi: 10.1016/j.bspc.2024.107473.

Wensi Tang, Guodong Long, Lu Liu, Tianyi Zhou, Michael Blumenstein, and Jing Jiang. Omni-scale CNNs: A simple and effective kernel size configuration for time series classification. In *International Conference on Learning Representations*, 2022.

Julian F. Thayer and Richard D. Lane. A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders*, 61(3):201–216, 2000. doi: 10.1016/S0165-0327(00)00338-4.

Camilo E. Valderrama and Anshul Sheoran. Identifying relevant EEG channels for subject-independent emotion recognition using attention network layers. *Frontiers in Psychiatry*, 16: 1494369, 2025. doi: 10.3389/fpsyt.2025.1494369.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Yi Wu, Yuhang Chen, Jiahao Cui, Jiaji Liu, Lin Liang, and Shuai Li. Cross-modal contrastive learning for emotion recognition: Aligning ECG with EEG-derived features. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, Lecture Notes in Computer Science, pp. 69–79. Springer, 2025. doi: 10.1007/978-3-032-05141-7_7.

Tianze Xu, Xunde Dong, Yang Yang, Chen Xu, and Xiuling Liu. Hisan: A hierarchical segment attention network for ecg-based emotion recognition. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 7129–7136. IEEE, 2025.

Yalan Ye, Tongjie Pan, Qianhe Meng, Jingjing Li, and Li Lu. Online ECG emotion recognition for unknown subjects via hypergraph-based transfer learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 3666–3672, 2022. doi: 10.24963/ijcai.2022/509.

Sung-Nien Yu, Chia-Wei Cheng, and Yu Ping Chang. Emotion recognition using electrocardiogram trajectory variation in attention networks. *Engineering Proceedings*, 120(1):17, 2026. doi: 10.3390/engproc2025120017.

Chengxin Zhong, Xunde Dong, Yang Yang, Xiuling Liu, and Chen Xu. Emonet-ecg: A lightweight inverted pyramid architecture for joint valence-arousal prediction using raw ecg. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4535–4540. IEEE, 2025.

A Datasets and Subject-Independent Splits

Table 5 distinguishes the characteristics of the original datasets from the data configuration used in this study. AMIGOS contains both short and long recordings, while DREAMER contains recordings lasting approximately one to six and a half minutes. Their original ECG signals are sampled at 256 Hz and their EEG signals at 128 Hz; ECG is therefore resampled to 128 Hz for synchronized modeling. MPED contains 28 trials per participant, with recording durations ranging from 122 to 295 seconds. Its ECG and EEG are originally sampled at 250 and 1,000 Hz, respectively, and EEG is resampled to 250 Hz.

AMIGOS and DREAMER natively provide the same 14 EEG leads: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4. MPED provides a denser EEG montage. To control the

Table 5: Dataset characteristics and subject-independent splits. The upper block describes the source datasets, whereas the lower block reports the data used in this study. Sampling-rate and lead entries are ordered as ECG/EEG. AMIGOS and DREAMER natively contain 14 EEG leads; MPED uses the same selected 14 leads. No subject occurs in more than one subset.

Dataset	Subjects	Raw recordings	Duration (s)	Raw sampling ECG/EEG (Hz)	Original labels or ratings
AMIGOS	40	769	51-150; 846-1415	256/128	V/A: 1-9
DREAMER	23	414	67-394	256/128	V/A: 1-5
MPED	23	644	122-295	250/1000	Seven discrete classes; self-report: 1-9

Dataset	Task	Used recordings	Input rate (Hz)	Used leads ECG/EEG	Subject split (train/val/test)	Recording split (train/val/test)
AMIGOS	4-class	629	128	2/14	30/5/5	469/80/80
DREAMER	4-class	414	128	2/14	17/2/4	306/36/72
MPED	3-class	644	250	1/14	17/3/3	476/84/84

computational cost of modeling its long recordings, we retain the same 14 leads and exclude the remaining EEG channels. Thus, the 14-lead entries for AMIGOS and DREAMER represent their original configurations, whereas the MPED entry represents the selected configuration used in this study.

AMIGOS and DREAMER are formulated as four-class valence-arousal quadrant classification. Dataset-specific median valence and arousal ratings separate low and high values, yielding the HVHA, HVLA, LVLA, and LVHA classes. For MPED, Joy and Funny are grouped as positive, Neutrality is retained as neutral, and Anger, Sadness, Fear, and Disgust are grouped as negative.

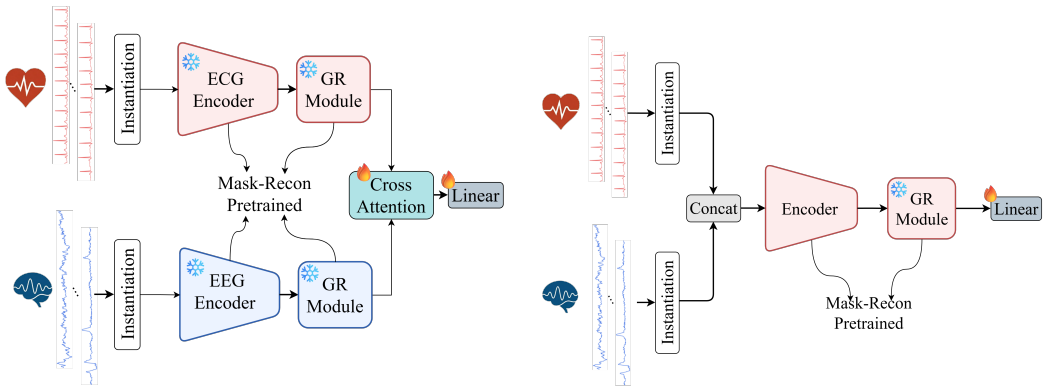
Subjects are assigned to mutually exclusive training, validation, and test subsets. The split is fixed before model training and reused for every framework, encoder, ablation, and random seed. No subject occurs in more than one subset. This protocol prevents subject-specific patterns from being shared between training and evaluation. The processed MPED data provide fixed train/validation/test subsets, which are used unchanged.

B Encoder and Baseline Implementations

Common instance-encoder interface All five encoders are evaluated within the same HB-IGCL pipeline. A long recording is first divided into non-overlapping 4-s instances. The selected encoder processes each instance and produces a 128-channel latent feature map, which is then passed to the common Gate and Refinement (GR) module. Average pooling after GR produces a 128-dimensional representation for each instance. The four ECG/EEG branches use the same selected encoder architecture but do not share parameters. Consequently, changing an encoder does not alter the GR module, the pretraining objectives, or the downstream fusion and classification modules.

AMIE uses four parallel convolutional experts. Each expert contains two convolution-normalization-activation blocks, and an input-dependent gate forms a weighted sum of their outputs. ModernTCN is applied within each instance and its output is projected to the common 128-channel interface. PatchTST uses channel-independent patch encoding, and OS-CNN uses parallel convolutions with multiple receptive fields. Their outputs are likewise mapped to the same interface before GR, ensuring that the framework comparison changes only the local instance encoder.

PatchTST+ PatchTST+ is our task-adapted variant of PatchTST. Standard PatchTST preserves channel independence during temporal encoding but concatenates channel embeddings before its prediction head. With 14 EEG leads and a 128-dimensional embedding, this creates a 1,792-dimensional vector before fusion. PatchTST+ replaces this concatenation with learnable channel aggregation. After channel-independent Transformer encoding, patch features are averaged within each channel. A shared two-layer gating network with a Tanh activation assigns a score to every channel, and a softmax normalizes these scores across channels. The final representation is



(a) LMM **(b) EMM**

Figure 7: Downstream fine-tuning pipelines of (a) LMM and (b) EMM. Snowflake and flame symbols denote frozen and trainable modules, respectively.

the weighted sum of channel embeddings and therefore remains 128-dimensional regardless of the number of leads.

For the 4-s instances used in the main comparison, the patch length and stride are 128 samples on AMIGOS and DREAMER and 250 samples on MPED, yielding four non-overlapping patches per instance. PatchTST+ uses one Transformer layer, two attention heads, a feed-forward dimension of 512, and dropout 0.2. The aggregated representation is projected to the same feature interface used by the other encoders before entering GR. Thus, PatchTST+ retains the patch-based temporal modeling of PatchTST while avoiding dimensional growth caused by channel concatenation.

Pretraining baseline adaptations The source baselines were not designed for synchronized, variably long ECG-EEG recordings represented as bags of instances. We therefore preserve each learning paradigm and make only the structural changes required to accept the common instance-encoder output.

LMM LMM retains its source late-masking paradigm. ECG and EEG are first encoded separately as sequences of 4-s instance representations. Masking and reconstruction are then applied to the high-level representations before late multimodal fusion. The source foundation encoder is replaced with each of the five instance encoders so that LMM can be compared under the same representation interface. The final late-mask ratio is 0.1.

EMM EMM is a controlled early-fusion variant of LMM. It first combines the ECG and EEG representations and then applies the same masking and reconstruction objective. All settings other than fusion timing and the selected mask ratio are shared with LMM. The final early-mask ratio is 0.1. EMM therefore isolates whether fusion before masking is beneficial for long synchronized recordings.

Downstream transfer Figure 7 illustrates how the pretrained LMM and EMM representations are transferred to downstream classification. LMM retains separate ECG and EEG instance-encoding paths and fuses their outputs with a trainable cross-attention module before classification. EMM concatenates the instantiated ECG and EEG signals first; the resulting joint input is then processed by a shared multimodal encoder and GR module before classification. Modules marked as frozen preserve their pretrained parameters, whereas the fusion and classification components marked as trainable are optimized on the downstream task.

CCM CCM retains its source cross-modal contrastive paradigm. A patch in the source method corresponds to one 4-s instance in our setting, while a source sequence corresponds to a complete recording. The source method constructs an augmented view by shuffling a small number of patches within a short sample. Applying this operation to our recordings would shuffle many chronological

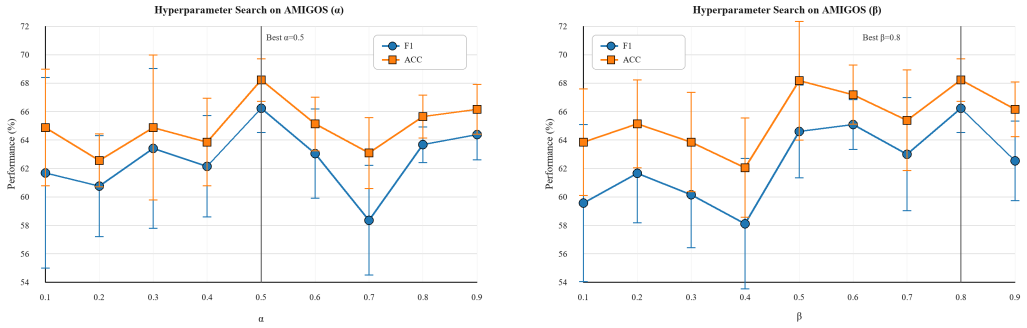


Figure 8: Sensitivity of macro-F1 and Acc to the loss weights on AMIGOS. Left: α is varied with $\beta = 0.8$. Right: β is varied with $\alpha = 0.5$. Curves show the mean over five runs and error bars show the standard deviation.

instances and disrupt long-range order. We therefore remove instance shuffling while retaining cross-modal contrast and masking. The baseline-specific mask ratio and loss-balance coefficient are set to $m = 0.1$ and $a = 0.3$ in the main comparison; the full sensitivity analysis is reported in Appendix H.

C Loss-Weight Sensitivity Analysis

The overall pretraining objective is

$$\mathcal{L} = \beta\left[\alpha\mathcal{L}_{t1} + (1-\alpha)\mathcal{L}_{t2}\right] + (1-\beta)\mathcal{L}_{s}. \tag{8}$$

Here, α balances the two directional instance-global objectives, whereas β balances their combined contribution against the spatial-guided alignment objective. We conduct a one-at-a-time sensitivity analysis on AMIGOS while keeping all other pretraining and downstream settings unchanged. We first fix $\beta = 0.8$ and vary α from 0.1 to 0.9; we then fix $\alpha = 0.5$ and vary β over the same range.

As shown in Fig. 8, $\alpha = 0.5$ achieves the best joint result, with 66.23% F1 and 68.21% Acc. This setting gives equal nominal weight to $\mathcal{L}_{t1}$ and $\mathcal{L}_{t2}$ within the bidirectional term. Moving toward either directional extreme generally reduces F1 and can increase variability, indicating that C2E and E2C provide complementary supervision rather than one direction consistently dominating the other.

With α fixed at 0.5, $\beta = 0.8$ again gives the strongest combined performance. Smaller values shift more weight toward $\mathcal{L}_s$ and weaken the instance-global objectives, whereas values near 1 suppress the contribution of lead-level spatial alignment. The results therefore support using bidirectional instance-global alignment as the principal objective while retaining a nonzero spatial constraint. We consequently use $\alpha = 0.5$ and $\beta = 0.8$ in the main experiments.

D Instance-Length Ablation

Fig. 9 reports the effect of varying the instance length while keeping the remaining pretraining and fine-tuning settings fixed. Across AMIGOS, DREAMER, and MPED, the 4-second setting provides the best overall trade-off between local physiological context and the number of temporal instances available for local-to-global interaction.

E Additional Encoder Variants

The upper block of Table 6 fixes the ECG branch to AMIE and replaces only the EEG local encoder. PureConvNet and the 4-s Hybrid Net yield similar results, whereas extending the Hybrid Net input from 4 to 14 s reduces both metrics. Under the present sample size and optimization setting, simply increasing the temporal span of one encoder does not improve recognition and may dilute local affective cues.

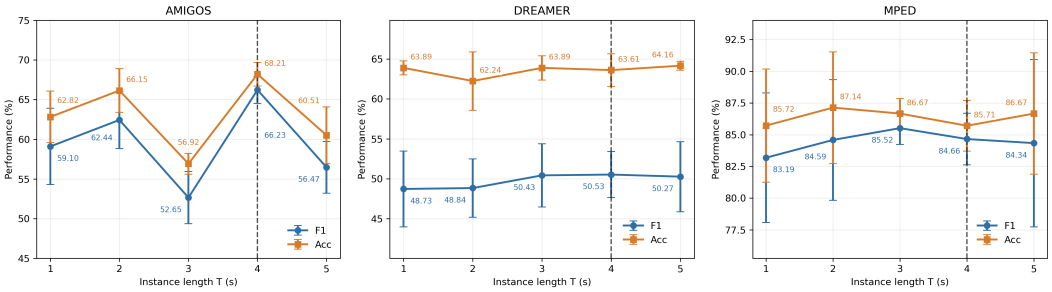


Figure 9: Ablation results of instance length.

Table 6: Additional encoder comparisons on AMIGOS (%).

ECG encoder	EEG encoder	F1	Acc
EEG encoder replacement			
AMIE	PureConvNet	$57.85_{\pm 2.50}$	$60.77_{\pm 2.08}$
AMIE	Hybrid Net (4 s)	$57.37_{\pm 4.22}$	$60.26_{\pm 3.34}$
AMIE	Hybrid Net (14 s)	$49.72_{\pm 2.73}$	$50.51_{\pm 2.76}$
Cross-encoder pairing			
AMIE	AMIE	$\mathbf{66.23_{\pm 1.71}}$	$\mathbf{68.21_{\pm 1.49}}$
PatchTST+	PatchTST+	$55.96_{\pm 0.37}$	$57.43_{\pm 1.25}$
AMIE	PatchTST+	$63.67_{\pm 2.08}$	$64.87_{\pm 1.31}$
PatchTST+	AMIE	$46.34_{\pm 2.53}$	$46.16_{\pm 2.69}$

The lower block cross-combines AMIE and PatchTST+ in the ECG and EEG branches. Using AMIE for both modalities gives the strongest result. The two heterogeneous configurations are asymmetric: AMIE on ECG with PatchTST+ on EEG remains relatively competitive, whereas exchanging their modal assignments causes a substantial decline. This result suggests that the ECG-side local encoder is particularly sensitive under the current E2C aggregation design. Taken together, neither the unilateral EEG replacement nor the cross-encoder combinations outperform the homogeneous AMIE configuration.

Table 7: Criss-Cross Transformer adaptation on AMIGOS (%). The 1-s rows follow the source LMM setting, whereas the 4-s rows use our setting.

Instance	Mask	Layers	Update	F1	Acc
1 s	0.5	12	Frozen	$23.25_{\pm 4.88}$	$29.75_{\pm 2.74}$
1 s	0.5	6	Frozen	$21.84_{\pm 3.08}$	$31.03_{\pm 3.66}$
1 s	0.5	3	Frozen	$25.60_{\pm 2.59}$	$34.36_{\pm 3.38}$
1 s	0.5	1	Frozen	$23.95_{\pm 3.85}$	$34.10_{\pm 3.50}$
4 s	0.5	1	Frozen	$31.80_{\pm 6.03}$	$36.67_{\pm 3.20}$
4 s	0.5	2	Frozen	$\mathbf{34.60_{\pm 3.38}}$	$37.18_{\pm 2.43}$
4 s	0.5	3	Frozen	$31.69_{\pm 3.22}$	$38.21_{\pm 2.20}$
4 s	0.5	6	Frozen	$33.96_{\pm 5.44}$	$37.95_{\pm 4.02}$
4 s	0.7	2	Frozen	$30.34_{\pm 4.07}$	$35.64_{\pm 0.96}$
4 s	0.3	2	Frozen	$26.91_{\pm 3.70}$	$\mathbf{38.97_{\pm 4.62}}$
1 s	0.5	12	Full	$19.40_{\pm 1.19}$	$29.75_{\pm 3.18}$
1 s	0.5	6	Full	$22.07_{\pm 3.48}$	$30.77_{\pm 4.80}$
1 s	0.5	3	Full	$24.65_{\pm 2.74}$	$30.26_{\pm 3.68}$
1 s	0.5	1	Full	$24.18_{\pm 3.76}$	$32.82_{\pm 4.18}$
4 s	0.5	1	Full	$31.25_{\pm 4.12}$	$35.38_{\pm 4.02}$
4 s	0.5	2	Full	$30.28_{\pm 3.72}$	$36.15_{\pm 2.20}$
4 s	0.5	3	Full	$27.58_{\pm 4.26}$	$37.18_{\pm 4.51}$
4 s	0.5	6	Full	$23.50_{\pm 2.33}$	$31.54_{\pm 1.54}$
4 s	0.7	2	Full	$25.78_{\pm 1.78}$	$34.87_{\pm 4.47}$
4 s	0.3	2	Full	$30.31_{\pm 5.66}$	$36.41_{\pm 3.30}$

F LMM ADAPTATION

The source LMM framework employs the Criss-Cross Transformer, a Transformer-based encoder, with 1-s signal segments as its original input setting. Transformer-based encoders can be difficult to learn reliably from the small subject-independent datasets considered in this study. Directly retaining this encoder could therefore confound the evaluation of the masking paradigm with a backbone that is poorly matched to our small-sample setting. We consequently search instance duration, mask ratio, Transformer depth, and downstream update scope on AMIGOS. In Table 7, the 1-s rows follow the source LMM setting, whereas the 4-s rows use the instance duration adopted by HB-IGCL. Frozen updates only downstream parameters, while Full updates both the encoder and downstream model.

Four-second instances consistently outperform the original 1-s setting. Increasing Transformer depth does not produce monotonic gains. With frozen transfer, two layers at mask ratio 0.5 give the highest F1, while two layers at mask ratio 0.3 give the highest Acc. Full updating is generally less stable, and the six-layer model deteriorates markedly. These results indicate that deeper token-based encoders can be difficult to optimize in the long-record, small-sample setting and justify comparing the masking paradigm with multiple instance encoders rather than retaining a single source backbone.

We additionally evaluate BIOT under the same protocol. It obtains $25.66_{\pm 0.55}$ F1 and $30.77_{\pm 1.40}$ Acc, further showing that a token-based biosignal encoder does not transfer automatically to the present long-record, small-sample setting.

Table 8: Mask-position and mask-ratio search on AMIGOS (%).

Position	Ratio	F1	Acc
Early	0.1	$23.77_{\pm 1.30}$	$31.54_{\pm 3.30}$
Early	0.3	$23.54_{\pm 4.81}$	$30.26_{\pm 4.41}$
Early	0.5	$21.18_{\pm 3.04}$	$29.49_{\pm 1.62}$
Early	0.7	$19.33_{\pm 2.27}$	$29.23_{\pm 1.49}$
Late	0.1	$\mathbf{62.45_{\pm 1.43}}$	$\mathbf{65.38_{\pm 1.62}}$
Late	0.3	$58.75_{\pm 2.32}$	$62.82_{\pm 2.15}$
Late	0.5	$56.38_{\pm 2.74}$	$59.23_{\pm 2.62}$
Late	0.7	$56.06_{\pm 2.69}$	$60.00_{\pm 2.49}$

Table 9: CCM mask-ratio search on AMIGOS (%) with $a = 0.5$.

m	F1	Acc
0.1	$\mathbf{54.15_{\pm 2.49}}$	$\mathbf{56.67_{\pm 2.97}}$
0.3	$50.59_{\pm 7.34}$	$53.08_{\pm 7.14}$
0.5	$53.38_{\pm 7.06}$	$54.61_{\pm 6.36}$
0.7	$50.07_{\pm 2.93}$	$52.57_{\pm 2.29}$

Table 10: CCM loss-balance coefficient search on AMIGOS (%) with $m = 0.1$.

a	F1	Acc
0.1	$51.61_{\pm 1.84}$	$55.90_{\pm 3.59}$
0.2	$50.50_{\pm 1.72}$	$53.59_{\pm 3.08}$
0.3	$\mathbf{56.07_{\pm 2.99}}$	$\mathbf{57.69_{\pm 2.81}}$
0.4	$45.63_{\pm 5.55}$	$48.97_{\pm 5.46}$
0.5	$54.15_{\pm 2.49}$	$56.67_{\pm 2.97}$
0.6	$47.74_{\pm 5.31}$	$51.03_{\pm 4.96}$
0.7	$54.78_{\pm 4.32}$	$57.44_{\pm 4.96}$
0.8	$49.81_{\pm 6.59}$	$54.10_{\pm 5.22}$
0.9	$54.04_{\pm 3.34}$	$56.92_{\pm 1.74}$

G MASK POSITION AND RATIO

All experiments in this appendix use AMIE encoders for both the ECG and EEG branches.

Table 8 isolates the effects of masking location and strength on AMIGOS. Early masking perturbs the joint input before modality-specific representations are sufficiently formed, whereas late masking first preserves modality-specific encoding and then masks high-level representations. Late masking outperforms early masking at every tested ratio, with its best result at 0.1. Its performance generally decreases as the ratio increases, suggesting that strong masking can remove a disproportionate share of the sparse affective evidence in long recordings. This experiment supports the late-masking paradigm retained by LMM and explains the weaker AMIE result obtained by the early-fusion EMM variant.

H CCM ADAPTATION AND SENSITIVITY

All experiments in this appendix use AMIE encoders for both the ECG and EEG branches.

The adapted CCM baseline removes instance shuffling because reordering instances would destroy the chronology of a long recording. We retain cross-modal contrast and masking, and search the remaining mask ratio m and loss-balance coefficient a on AMIGOS. Tables 9 and 10 report the complete sweeps.

After fixing $m = 0.1$, $a = 0.3$ achieves the highest F1 and Acc and is therefore used in the main comparison. The response to a is non-monotonic, and several settings have relatively large standard deviations, indicating that the adapted baseline is sensitive to this coefficient. Reporting the complete sweep makes this sensitivity explicit rather than presenting only the selected configuration.