DHCM-CACL: Dynamic Hierarchical Cross-modal Mamba with Confidence-Adaptive Contrastive Learning for Multimodal Emotion Recognition

Baiqiang Wu, Yang Li*

School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China
{wubaiqiang, liyang}@buaa.edu.cn

Abstract

Multimodal emotion recognition plays a crucial role in enhancing the intelligence of human-computer interaction and emotional understanding. However, conventional approaches face challenges such as scarcity of annotated data, significant modality heterogeneity, and temporal misalignment. To address these issues, we propose DHCM-CACL, a novel self-supervised emotion recognition framework integrating EEG and facial expressions. During the pre-training phase, we propose a Dynamic Hierarchical Cross-modal Mamba module (DHCM), which models long-term dependencies through dynamic state matrices, incorporates forgetting gates for noise suppression, and constructs a hierarchical cross-modal interaction structure, effectively achieving cross-modal temporal alignment and mitigating modality heterogeneity. Subsequently, we propose a Confidence-Adaptive Contrastive Learning module (CACL) that dynamically adjusts sample weights using gated confidence signals derived from DHCM to compute loss, prioritizing reliable samples while suppressing noisy instances through adaptive weighting, thereby enhancing representation reliability and generalization in data-scarce scenarios. During the fine-tuning phase, we integrate a cross-modal attention gating mechanism to reinforce temporal associations and adopt an evidence-aware joint optimization objective, providing probabilistic credibility outputs for emotion prediction. Experimental results on the DEAP and MAHNOB-HCI datasets demonstrate that our approach achieves state-of-the-art performance in emotion classification under both subject-dependent and subject-independent settings.

Introduction

Emotion recognition, as a core technology in human-computer interaction (HCI), has gradually been applied in multiple fields (Saxena, Khanna, and Gupta 2020; Yang et al. 2024), such as intelligent customer service, healthcare, and entertainment. In these application scenarios, emotion recognition can effectively enhance the interaction experience between systems and users, enabling machines to perceive the user's emotional state and respond appropriately.

In recent years, emotion recognition methods based on physiological signals such as EEG, ECG, and EMG have emerged as a new research direction (Chanel et al. 2011; Fan et al. 2023; Bhatlawande et al. 2024). Among these, EEG signals, which reflect brain activity, have the advantages of non-invasiveness, low cost, and high temporal resolution (Li et al. 2022; Jafari et al. 2023; Li et al. 2024), allowing them to directly reflect the physiological basis of emotions.

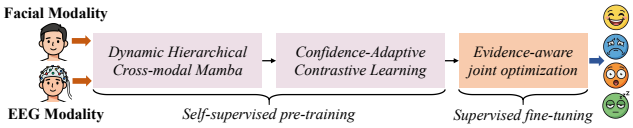

Figure 1: The DHCM-CACL framework.

At the same time, multimodal emotion recognition has gradually become a research hotspot (Pan et al. 2023; Ahmed, Al Aghbari, and Girija 2023; Liu et al. 2024a). Compared to emotion recognition based on a single modality, such as facial expressions (Sun et al. 2023; Zhang et al. 2024; Shao, Li, and Wu 2025), multimodal emotion recognition can more comprehensively express emotional information by integrating data from different modalities.

However, existing multimodal methods still face several key challenges. Firstly, labeled data is scarce. High-quality multimodal emotion annotations require professional equipment and expertise from psychologists, making large-scale supervised training difficult to achieve (Zong, Mac Aodha, and Hospedales 2023). Secondly, modality heterogeneity is significant (Jia et al. 2021). Different modal features exhibit essential distribution differences, and the emotional information they reflect also varies. Furthermore, temporal feature alignment is challenging (Chen et al. 2021). The performance of data sources from different modalities in the time dimension is often inconsistent, and the collection methods and frequencies of different modalities also vary.

To address these issues, we propose a novel self-supervised multimodal emotion recognition approach named DHCM-CACL, as illustrated in Figure 1, which consists of three key components: (1) Dynamic Hierarchical Cross-modal Mamba Module (DHCM): It adaptively models long-term temporal dependencies through dynamically generated state matrices, introduces a forgetting gate mechanism to suppress noise, and constructs a hierarchical cross-modal interaction structure to progressively

*Corresponding author.

fuse cross-modal features. This ensures cross-modal feature alignment and reduces data heterogeneity. (2) Confidence-Adaptive Contrastive Learning Module (CACL): It employs a confidence-adaptive mechanism that dynamically adjusts the contribution of low-confidence samples during loss computation. By suppressing low-confidence samples and emphasizing reliable features, it enhances representation reliability and generalization in data-scarce scenarios. (3) In the fine-tuning phase, we design a multi-head attention interaction mechanism to strengthen temporal correlation. The joint optimization of classification loss, evidence loss, and contrastive consistency loss facilitates probabilistic credibility outputs for emotion prediction.

The proposed framework demonstrates multiple core advantages: self-supervised pre-training minimizes label dependence; hierarchical dynamic modeling with confidence-adaptive contrastive learning addresses temporal alignment and modality heterogeneity; and the joint optimization objective during the fine-tuning phase achieves probabilistic reliability modeling.

Experiments demonstrate that our method achieves state-of-the-art performance on the DEAP and MAHNOB-HCI datasets. In the valence and arousal prediction tasks, the average ACC and F1 scores surpass the baseline, confirming the effectiveness of our approach. In summary, our contributions are as follows:

- We propose DHCM-CACL, a novel self-supervised framework for multimodal emotion recognition, which effectively addresses data scarcity, modality heterogeneity, and temporal misalignment.
- We design DHCM, a dynamic hierarchical cross-modal module that models long-range temporal dependencies while reducing data heterogeneity and achieving cross-modal temporal alignment.
- We design CACL, a confidence-adaptive contrastive learning module that enhances representation reliability and generalization in data-scarce scenarios.
- We develop an evidence-aware joint optimization during the fine-tuning phase to produce credible probability outputs, and our method achieves state-of-the-art performance on the DEAP and MAHNOB-HCI datasets.

Related Work

Multimodal Emotion Recognition. Multimodal emotion recognition enhances perception robustness by fusing heterogeneous features. Early research focused on supervised fusion schemes, such as bilinear pooling with attention mechanisms (Choi, Kim, and Song 2020), graph-attention filtering (Meng et al. 2024), local-global graph representation enhancement (Fu et al. 2025), and dynamic graph structure optimization (Tu et al. 2024). However, two critical limitations persist: 1) heavy reliance on annotated data impedes applications in label-scarce scenarios, and 2) temporal modeling using RNNs (Li et al. 2020) or Transformers (Hazmoune and Bougamouza 2024) fails to resolve cross-modal asynchrony (Yang et al. 2022). These shortcomings motivate the integration of self-supervised learning with efficient temporal architectures.

Self-Supervised Emotion Recognition. Self-supervised learning leverages proxy tasks to exploit unlabeled data, mitigating annotation bottlenecks. While instance discrimination frameworks (e.g., SimCLR (Chen et al. 2020) and CMOET (Wang et al. 2023)) learn general representations, their global contrastive loss overlooks local temporal consistency. S2T (Liu et al. 2024b) designs a self-supervised mechanism aligned with temporal and structural strength, but it cannot efficiently achieve fine-grained cross-modal alignment. In the field of physiological signals, Masked Autoencoders (MAE) have been used for EEG feature pretraining (Mohsenvand, Izadi, and Maes 2020), but single-modal training also limits the potential for cross-modal interaction. These limitations highlight the urgency of developing cross-modal, fine-grained self-supervised temporal frameworks.

Progress of State-Space Models in Cross-Modal Fusion. State space models (SSMs) have emerged as efficient paradigms for long-sequence modeling due to linear complexity. The structured state-space sequence layer (S4) pioneered EEG long-term dependency modeling (Gu, Goel, and Ré 2021), but its static state matrix cannot adapt to dynamic affective fluctuations. Attention computation optimizations (Dao et al. 2022) improve efficiency while retaining quadratic complexity. Recently, selective SSMs (Mamba) achieve context-aware modeling via dynamic state transitions (Gu and Dao 2023), showing promise in multimodal tasks: text-video retrieval (Tang et al. 2025) and audio-visual segmentation (Gong et al. 2025). Nevertheless, existing works neglect modality-differentiated state evolution mechanisms and hierarchical cross-modal interaction, leading to information loss in fine-grained dependencies.

Methodology

In this section, we describe our self-supervised framework named DHCM-CACL. The overall architecture of the proposed approach is shown in Figure 2, which primarily consists of two modules: DHCM and CACL.

The pre-training phase

DHCM. The structure of DHCM is shown in Figure 3. Given input feature matrices $x^{(m)} \in \mathbb{R}^{B\times T\times D_{(m)}}$, where m denotes different modalities, B is the batch size, T represents the number of time windows, and $D_{(m)}$ represents the dimension of the feature for each modality. To mitigate modal heterogeneity and enhance temporal consistency, we design a dual preprocessing mechanism. First, the feature matrices are input to a learnable diffusion layer. This layer adopts a depthwise separable convolution structure, smoothing temporal noise through multi-step diffusion. Its parameters are adaptively optimized during training, avoiding the rigidity of traditional Gaussian filtering.

$$\tilde{x}^{(m)} = \text{DiffusionLayer}(x^{(m)}) \in \mathbb{R}^{B\times T\times D_{(m)}} \tag{1}$$

Subsequently, heterogeneous features are projected into a unified latent space H via linear transformation, with LayerNorm facilitating distributional stability to support robust cross-modal interactions:

$$H_{(m)}^{(0)} = \text{LayerNorm}(W_{(m)}\tilde{x}^{(m)} + b_{(m)}) \tag{2}$$

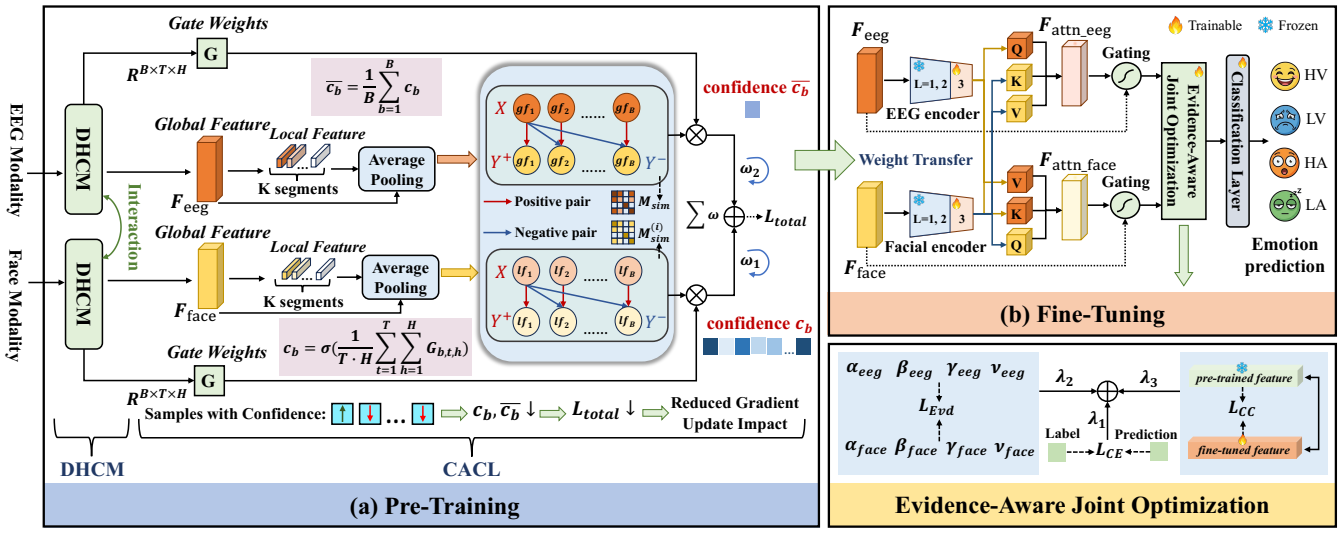

Figure 2: The overall architecture of the proposed DHCM-CACL.

where $W_{(m)} \in \mathbb{R}^{D_{(m)} \times H}$ is learnable projection matrices, and $b_{(m)}$ is bias terms.

After preprocessing, we proceed to dynamic state-space modeling. To capture the temporal dynamics of emotional expression, we innovatively design a Selective State Transition mechanism. Dynamic state transition is represented as:

$$X_{\text{state}} = W_{\text{state}} H_t \tag{3}$$

$$A_t = \text{Reshape}(\text{MLP}(H_t)) \in \mathbb{R}^{d_s \times d_s} \tag{4}$$

$$f_t = \sigma(W_f H_t) \tag{5}$$

$$h_t = f_t \odot (A_t h_{t-1}) + (1 - f_t) \odot X_{\text{state},t} \tag{6}$$

where X_{state} denotes the state projection; A_t is the dynamic matrix; f_t is the forget gate; h_t represents the state update; $d_s = 32$ is the state dimension. This design overcomes the fixed state transition pattern of traditional RNN or LSTM, allowing A_t to dynamically adapt to input features for context-aware temporal modeling.

We introduced hierarchical cross-modal interaction in DHCM. After DynamicSSM processing, multi-level modality fusion is achieved through a gated cross mechanism:

$$S_{(m)}^{(l)} = \text{DynamicSSM}(H_{(m)}^{(l)}) \tag{7}$$

$$H_{(m)}^{\text{enh}} = W_{\text{rec}} S_{(m)}^{(l)} \tag{8}$$

$$G = \sigma\left(W_g [H_E^{(l)}; H_F^{(l)}]\right) \tag{9}$$

$$H_E^{(l+1)} = H_E^{(l)} + G \odot H_F^{\text{enh}} \tag{10}$$

$$H_F^{(l+1)} = H_F^{(l)} + G \odot H_E^{\text{enh}} \tag{11}$$

where $S_{(m)}^{(l)}$ is the state space feature; G is the gating signal; $H_E^{(l+1)}$ and $H_F^{(l+1)}$ represent the fused features. Ultimately, it produces enhanced features F_{eeg}, F_{face}, and gating weights G, providing high-quality representations for the subsequent CACL module. Here, $F_{(m)} = \text{LayerNorm}(H_{(m)}^{(L)})$.

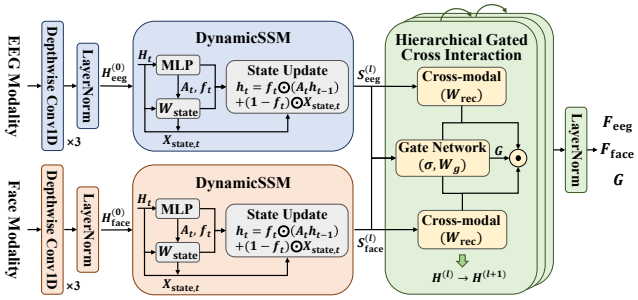

Figure 3: The structure of DHCM.

DHCM achieves dynamic alignment of cross-modal temporal features while reducing modality heterogeneity.

CACL. To enhance sensitivity to affect-critical segments, we first implement local segment decoupling which partitions temporal features into K segments.

$$F_{(m)}^{(i)} = \text{chunk}\left(F_{(m)}, K, \text{dim} = 1\right)[i] \tag{12}$$

Thereafter, temporal average pooling is performed on each segment's features to preserve salient affective patterns while filtering transient noise.

$$S_{(m)}^{(i)} = \frac{1}{T/K} \sum_{t=1}^{T/K} F_{(m),t}^{(i)} \tag{13}$$

Subsequently, the cosine similarity between the EEG segment $S_{eeg}^{(i)}$ and the facial segment $S_{face}^{(i)}$ is computed to quantify their cross-modal semantic alignment. The similarity matrix is expressed as:

$$M_{\text{sim}}^{(i)} = \frac{S_{\text{eeg}}^{(i)} (S_{\text{face}}^{(i)})^T}{\tau} \tag{14}$$

We devised a multi-granularity contrastive loss to achieve dual-alignment of both granular modeling for affect-critical

segments and cross-modal global semantics. Expliitycitly, this encompasses segment-level contrastive loss and cross-modal global contrastive loss.

(1) Segment-level Contrastive Loss. The contrastive loss for each segment is defined as:

$$\mathcal{L}_{\text{seg}}^{(i)} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{\delta=0}^{1}\log\frac{\exp M_{\text{sim},bb}^{(i)}}{\sum_{k\neq b}^{B}\exp M_{\text{sim},\delta,bk}^{(i)}} \tag{15}$$

where $M_{\text{sim},bb}^{(i)}$ denotes positive sample pairs, and other similarity matrices denote negative sample pairs.

By enhancing segment similarity constraints, the model is forced to learn fine-grained cross-modal alignment within segments, improving sensitivity to key emotional segments.

(2) Cross-modal Global Contrastive Loss. By reinforcing sample-level global similarity constraints, this approach enforces cross-modal semantic alignment and enhances the model's macro-generalization capability for emotion representation. The global contrastive loss is expressed as:

$$\mathcal{L}_{\text{global}} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{\delta=0}^{1}\log\frac{\exp M_{\text{sim},bb}}{\sum_{k\neq b}^{B}\exp M_{\text{sim},\delta,bk}} \tag{16}$$

Following the acquisition of multi-grained contrastive losses, we implement a confidence-adaptive mechanism. Sample-level confidence c_b are derived from gating weights G output by DHCM, suppressing low-confidence samples and emphasizing reliable features.

$$c_b = \sigma\left(\frac{1}{T\cdot H}\sum_{t=1}^{T}\sum_{h=1}^{H}G_{b,t,h}\right) \tag{17}$$

Subsequently, confidence-adaptive weighting is applied to both the segment-level contrastive loss and global contrastive loss, dynamically down-weighting samples with low confidence to enhance feature robustness.

$$\mathcal{L}_{\text{total}} = w_1\sum_{i=1}^{K}c_b\cdot\mathcal{L}_{\text{seg}}^{(i)} + w_2\cdot\bar{c_b}\cdot\mathcal{L}_{\text{global}} \tag{18}$$

where w_1 and w_2 are initialized to 1.0 and optimize autonomously during training. This adaptive weight allocation strengthens multi-granularity feature learning.

The fine-tuning phase

To effectively transfer the generic multimodal representations learned during the self-supervised pre-training phase to downstream emotion recognition tasks, we adopt a layer-wise partial freezing fine-tuning strategy and further introduce cross-modal attention gating together with an evidence-aware joint optimization objective.

We freeze the first two layers of DHCM. Meanwhile, we unfreeze the last layer of DHCM and jointly optimize it with the newly initialized task-specific components.

Subsequently, multi-head attention computes cross-modal interactions, and the attention outputs are integrated via gating mechanism. For instance, considering the EEG modality, the calculation is formulated as follows:

$$attn_{\text{eeg}} = \text{MHA}(Q = F_{\text{eeg}}, K = F_{\text{face}}, V = F_{\text{face}}) \tag{19}$$

$$G_{\text{eeg}} = \sigma\left(W_g\left[F_{\text{eeg}}; attn_{\text{eeg}}\right]\right) \tag{20}$$

$$F_{\text{eeg}}^{\text{align}} = F_{\text{eeg}} + G_{\text{eeg}}\odot attn_{\text{eeg}} \tag{21}$$

where W_g is a learnable parameter.

Dirichlet distribution parameters are then generated for each modality.

$$\gamma_{\text{eeg}} = \text{ReLU}\left(W_\gamma F_{\text{eeg}}^{\text{align}}\right) \tag{22}$$

$$v_{\text{eeg}} = \text{Softplus}\left(W_v F_{\text{eeg}}^{\text{align}}\right) + \epsilon \tag{23}$$

$$\alpha_{\text{eeg}} = \text{Softplus}\left(W_\alpha F_{\text{eeg}}^{\text{align}}\right) + 1.0 \tag{24}$$

$$\beta_{\text{eeg}} = \text{Softplus}\left(W_\beta F_{\text{eeg}}^{\text{align}}\right) \tag{25}$$

where γ denotes the point estimate, v represents the uncertainty, and α, β are the shape and scale parameters of the Gamma distribution. The facial modality parameters are computed identically. By modeling modality-level uncertainty through probability distributions, we address the inability of traditional classification models to quantify predictive uncertainty.

We design an evidence-aware joint optimization objective aiming to guarantee classification accuracy, quantify predictive uncertainty, while preserving latent alignment. This joint optimization enhances model generalizability and decision reliability in data-scarce scenarios by constructing probabilistically trustworthy outputs through a triple-objective loss.

(1) Classification Loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C}y_c\log\left(\text{softmax}(\text{logit})_c\right) \tag{26}$$

(2) Evidence Loss:

$$\mathcal{L}_{\text{Evd}} = \sum_{i=1}^{B}\left((y_i - \frac{\gamma_i}{S_i - 1})^2 + \frac{v_i}{S_i(S_i-1)}\right) + \lambda\left\|\frac{1}{\beta+\epsilon} - 0\right\|^2 \tag{27}$$

where $S = \alpha + v$, and λ = 0.1 controls the strength of the uncertainty penalty.

(3) Contrastive Consistency Loss:

$$\mathcal{L}_{\text{Contrast}} = -\log\frac{\exp\left(\sin\left(z_p, z_f\right)/\tau\right)}{\sum_{j\neq i}\exp\left(\sin\left(z_p, z_j\right)/\tau\right)} \tag{28}$$

where z_p is the pre-trained feature, z_f is the fine-tuned feature, and τ = 0.07 is the temperature parameter.

The total loss comprises three components:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{CE}} + \lambda_2\mathcal{L}_{\text{Evd}} + \lambda_3\mathcal{L}_{\text{Contrast}} \tag{29}$$

where $\mathcal{L}_{\text{CE}}$ ensures classification accuracy, $\mathcal{L}_{\text{Evd}}$ constrains predictive uncertainty, and $\mathcal{L}_{\text{Contrast}}$ preserves consistency with pre-trained features.

This joint optimization framework simultaneously satisfies the triple objectives of classification accuracy, uncertainty quantification, and representation consistency, significantly improving generalization in data-scarce scenarios.

Method	Year	Valence(Independent)		Arousal(Independent)		Valence(Dependent)		Arousal(Dependent)	
		ACC/Std(%)	F1/Std(%)	ACC/Std(%)	F1/Std(%)	ACC/Std(%)	F1/Std(%)	ACC/Std(%)	F1/Std(%)
VigilanceNet	2022	64.42/10.46	65.75/11.98	59.69/11.76	69.81/13.04	89.77/05.26	90.95/05.12	90.63/04.40	92.39/05.93
SGMC	2023	68.98/10.12	66.37/12.54	69.02/10.34	71.41/14.88	90.23/05.12	86.35/05.60	88.54/05.33	91.77/06.05
CAFNet	2023	68.75/08.42	67.26/10.57	66.41/11.84	70.59/13.76	90.41/06.24	89.46/05.08	90.25/04.87	91.19/06.08
MAET	2023	65.34/09.18	65.39/11.74	67.10/09.55	71.26/13.75	91.66/07.25	92.87/05.09	91.17/05.31	92.08/05.73
EEG2Rep	2024	69.58/09.74	66.95/09.71	69.86/09.14	73.42/12.13	93.29/04.78	90.18/06.82	94.05/03.85	88.12/08.94
DMMR	2024	72.48/07.32	68.97/07.58	70.37/08.72	72.51/09.40	92.36/05.13	88.54/04.28	91.20/05.37	91.97/06.72
PhysioSync	2025	65.01/07.58	63.61/06.48	65.33/08.56	64.25/08.55	**96.25/02.19**	94.13/03.78	94.78/04.07	94.34/03.51
Ours	**2025**	**72.89/06.38**	**71.83/07.97**	**71.71/08.86**	**74.78/10.24**	95.21/03.86	**94.45/03.91**	**94.87/03.12**	**94.52/03.36**

Table 1: The subject-independent and subject-dependent classification results on the DEAP dataset.

Method	Year	Valence(Independent)		Arousal(Independent)		Valence(Dependent)		Arousal(Dependent)	
		ACC/Std(%)	F1/Std(%)	ACC/Std(%)	F1/Std(%)	ACC/Std(%)	F1/Std(%)	ACC/Std(%)	F1/Std(%)
VigilanceNet	2022	69.96/11.25	64.27/18.10	69.48/12.27	65.74/17.31	92.25/02.86	92.69/03.26	93.17/04.84	91.81/05.33
SGMC	2023	72.93/10.87	67.62/14.57	71.96/09.42	68.77/18.45	92.06/04.12	91.22/05.45	90.26/07.11	89.72/09.24
CAFNet	2023	70.87/10.65	68.20/16.79	69.83/12.41	68.65/19.10	91.68/04.24	90.26/03.12	91.75/05.38	88.46/09.42
MAET	2023	70.57/11.33	64.62/17.76	70.01/10.89	67.23/15.12	92.87/02.35	90.70/06.08	91.22/07.43	89.81/08.39
EEG2Rep	2024	**75.52/08.83**	71.83/15.69	73.25/11.95	70.12/17.46	**94.89/02.95**	93.06/04.28	94.05/05.12	92.43/06.51
DMMR	2024	74.53/10.87	69.32/12.15	73.56/11.08	71.24/16.18	92.39/05.10	92.76/03.11	93.74/02.96	91.86/06.82
PhysioSync	2025	66.32/10.87	64.39/17.15	65.28/12.26	63.15/18.31	94.15/04.82	95.03/02.20	93.78/04.13	92.64/06.73
Ours	**2025**	75.12/09.43	**73.45/14.26**	**73.78/11.16**	**72.29/15.20**	94.63/02.08	**95.26/02.55**	**95.34/03.89**	**93.32/05.84**

Table 2: The subject-independent and subject-dependent classification results on the MAHNOB-HCI dataset.

Experiments

Dataset and Data Preprocessing. We conduct experiments on two public datasets: DEAP (Koelstra et al. 2011) and MAHNOB-HCI (Soleymani et al. 2011).

For the DEAP dataset, the experiment selects 18 participants with complete facial video recordings; for the MAHNOB-HCI dataset, 24 participants with complete facial video recordings are selected. Both datasets apply identical data preprocessing methods. For the EEG modality, the Welch power spectral estimation algorithm are employed to extract PSD from consecutive 3-second non-overlapping time windows, utilizing five frequency bands. For the facial expression modality, 17 facial action units (AUs) are extracted from facial expressions using the OpenFace toolkit. Facial video sequences are divided into consecutive 3-second non-overlapping time windows, and facial expression feature vectors are generated by computing the mean AU intensity within each time window.

Implementation Details. In the experiments, we use the K-fold cross-validation to divide the training set and the test set. We explore two classification modes: subject-independent classification and subject-dependent classification. All experiments are conducted on an NVIDIA L40S GPU using PyTorch. We employ the Adam optimizer for training with a learning rate of 1e-4. The pretraining phase runs for 200 epochs, followed by a 40-epoch fine-tuning phase where we apply exponential learning rate decay. Hyperparameters are determined through grid search: the loss weighting parameters λ_1, λ_2, and λ_3 are searched over the set $\{0.1, 0.2, ..., 1.0\}$, ultimately converging to $\lambda_1 = 1$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.9$. The temperature parameter τ is set to 0.07, which is a common value in contrastive learning frameworks. K is set to the number of subjects with complete data recordings (DEAP: 18, MAHNOB-HCI: 24) to explicitly model and leverage individual differences. For the DHCM module, the d_state dimension is set to 32, the number of multi-head attention heads is 9, and the feed-forward

Modality	Valence		Arousal	
	ACC/Std(%)	F1/Std(%)	ACC/Std(%)	F1/Std(%)
(i) E	66.82/08.31	64.26/12.90	66.18/11.52	70.19/14.75
(i) F	65.53/08.82	65.74/11.13	65.29/12.90	69.97/16.11
(i) E + F	**72.89/06.38**	**71.83/07.97**	**71.71/08.86**	**74.78/10.24**
(ii) E	92.85/04.34	91.73/05.91	91.76/06.22	91.14/05.01
(ii) F	91.39/05.30	92.04/07.17	89.35/05.68	90.55/06.65
(ii) E + F	**95.21/03.86**	**94.45/03.91**	**94.87/03.12**	**94.52/03.36**

Table 3: The influence of different modalities. E represents the EEG modality; F represents the facial expression modality; (i) indicates subject-independent classification; (ii) indicates subject-dependent classification.

Method	Valence-ACC(%)	Arousal-ACC(%)
w/o DHCM	67.15/07.12	66.03/09.85
w/o CACL	69.82/08.03	68.57/10.22
w/o EDL	71.05/07.41	70.16/09.13
w/o SSL	64.27/11.35	63.76/12.87
Ours	**72.89/06.38**	**71.71/08.86**

Table 4: Core module ablation study on the DEAP dataset.

Module Variant	Valence-ACC	Arousal-ACC
Static State Matrix	70.15/07.21	69.82/09.05
No Forget Gate	69.24/07.83	68.93/09.62
Random State Matrix	67.58/08.47	66.41/10.31
Full DHCM	**72.89/06.38**	**71.71/08.86**

Table 5: Ablation study on dynamic state transition mechanisms in DHCM.

network hidden layer dimension is 1024. We primarily adopt ReLU and SoftPlus as the activation functions.

Baseline. We select several state-of-the-art methods from recent years as baselines, covering two aspects: multimodal methods and self-supervised methods. The multimodal methods include VigilanceNet (Cheng et al. 2022), CAFNet (Zhu et al. 2023) and MAET (Jiang et al. 2023). The self-supervised methods include SGMC (Kan et al. 2023), EEG2Rep (Mohammadi Foumani et al. 2024), DMMR (Wang, Zhang, and Tang 2024), and PhysioSync (Cui et al. 2025).

Experimental Results. We employ accuracy and F1-score as primary metrics for evaluating model performance. To ensure a fair comparison, all baselines are retested on two datasets in the same experiment environment, using their publicly available source code. As shown in Table 1 and Table 2, our DHCM-CACL model achieves SOTA performance on both the DEAP and MAHNOB-HCI datasets, demonstrating robust prediction capabilities in both subject-independent and subject-dependent settings. On the DEAP dataset, subject-independent valence prediction attains an accuracy of 72.89% and an F1-score of 71.83%, and arousal prediction achieves an accuracy of 71.71% and an F1-score of 74.78%, outperforming all baseline methods. On the MAHNOB-HCI dataset, although the subject-independent valence accuracy is marginally lower than that of EEG2Rep, the F1-score is significantly higher, and the arousal prediction exceeds all baseline methods. For subject-dependent classification tasks on both datasets, our model demonstrates marginal improvements over the multimodal and self-supervised baselines, achieving performance close to the upper bound.

As shown in Table 3, multimodal fusion significantly outperforms single-modal baselines on the DEAP dataset. In the subject-independent settings, the fusion model achieved 72.89% ACC and 71.83% F1-score for valence recognition, demonstrating over 6% improvement in both ACC and F1-score compared with unimodal approaches. The fusion

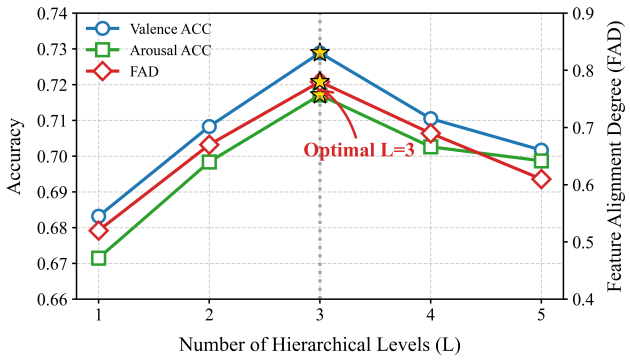

Figure 4: The impact of hierarchical levels (L) of DHCM on multimodal emotion recognition performance.

model also demonstrates substantial gains in arousal recognition tasks. These results indicate that the DHCM module effectively mitigates modality heterogeneity through dynamic state-space modeling and hierarchical cross-modal interaction. In the subject-dependent settings, the fusion model approaches the performance upper limit with 95.21% ACC and 94.45% F1, while the standard deviation decreases to 3.86%. This outcome further substantiates that the confidence-adaptive mechanism of CACL enhances model performance by suppressing low-confidence samples.

Component-wise ablation studies under subject-independent settings demonstrate the critical advantages of each proposed module over conventional alternatives. Table 4 presents the results of the ablation experiments. Replacing the DHCM with standard transformer results in a significant performance degradation. Replacing the CACL with standard contrastive learning also reduces model performance. Replacing the evidence layer with a standard softmax classifier yields a slight performance reduction. Crucially, removing self-supervised pre-training catastrophically impairs emotion recognition performance, demonstrating that self-supervised learning is essential for overcoming label scarcity and enhancing model generalization capabilities.

Table 5 systematically evaluates core innovations in the DHCM's dynamic state transition mechanism. The full DHCM achieves optimal performance, while all ablated variants exhibit statistically significant degradation, validating the necessity of dynamic hierarchical modeling. The substitution with the static state matrix (fixed A_t) reduces valence accuracy by 2.74% and arousal accuracy by 1.89%, confirming that dynamic state transitions are essential for adapting to affective temporal dynamics. The absence of the forget gate (fixed $f_t = 1$) degrades model performance, which is attributed to a weakened noise suppression capability due to the loss of selective forgetting. The substitution with the random state matrices results in significant performance degradation due to their inability to effectively propagate hierarchical cross-modal interactions.

As shown in Figure 4, the number of levels L in the DHCM significantly affects cross-modal emotion recognition performance. We introduce the feature alignment degree (FAD) to quantify the semantic consistency of cross-

Method	Annotation Ratio	ACC/Std(%)	F1/Std(%)
SGMC	50%	61.47/13.51	58.23/15.74
	20%	56.26/16.75	55.16/19.79
EEG2Rep	50%	65.07/10.48	63.55/12.62
	20%	61.31/12.92	58.34/13.46
DMMR	50%	64.35/09.08	61.03/09.35
	20%	59.07/11.26	57.75/10.81
PhysioSync	50%	58.17/09.12	55.68/09.63
	20%	52.30/11.28	50.76/13.05
Ours	50%	**69.46/08.74**	**69.14/08.91**
	20%	**65.72/11.63**	**65.15/10.16**

Table 6: Performance comparison at different annotation ratios on the DEAP dataset when fine-tuning, using valence emotion recognition under the subject-independent setting as an example.

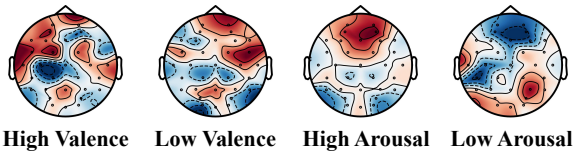

Figure 5: Saliency maps corresponding to different emotions on the DEAP dataset.

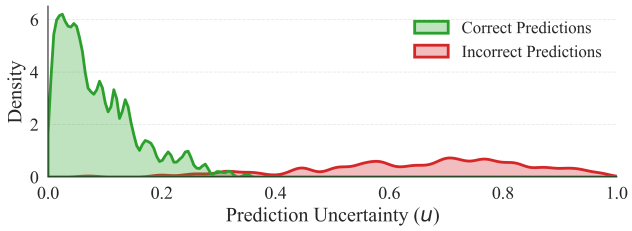

Figure 6: Uncertainty distribution on the DEAP dataset.

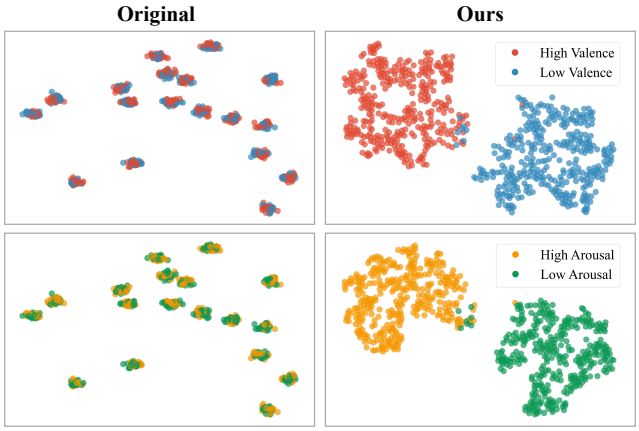

Figure 7: The t-SNE visualization in 2D embedding space of features on the DEAP dataset.

modal features in the latent space, calculated as FAD $= \frac{1}{T}\sum_{t=1}^{T}\frac{F_{EEG}^{(t)}\cdot F_{Face}^{(t)}}{\|F_{EEG}^{(t)}\|\cdot\|F_{Face}^{(t)}\|}$. The model achieves peak performance at $L = 3$, with maximum accuracy on valence and arousal classification and an optimal FAD of 0.78. This indicates strong synergy in cross-modal emotional representations, demonstrating that this hierarchical structure effectively balances the exploitation of modal complementarity with noise suppression.

As illustrated in Table 6, we compare the performance variations of our model with four self-supervised baselines under 50% and 20% annotation ratios. The results demonstrate that when fine-tuned on the DEAP dataset with low annotation ratios, DHCM-CACL significantly outperforms existing self-supervised approaches while exhibiting the smallest performance degradation, thereby validating its robust generalization capability in data-scarce scenarios.

Figure 5 illustrates the activation of brain regions associated with different emotions on the DEAP dataset. It can be generally observed that the prefrontal, temporal, and parietal regions are strongly associated with emotion recognition, which is largely consistent with previous research (Peng et al. 2022; Ma et al. 2023; Cui et al. 2023). This finding confirms that our DHCM-CACL can effectively capture the spatiotemporal dynamics of EEG and identify key neural correlates related to different emotions.

Figure 6 demonstrates that correctly classified samples cluster in low-uncertainty regions, contrasting with the high uncertainty of misclassified instances. This distinct separation demonstrates that our Evidence-Aware Joint Optimization effectively equips the model with "risk awareness", allowing it to signal potential errors through high uncertainty values (u), rather than making overconfident mistakes.

Furthermore, we employ t-SNE to visualize features learned by the proposed DHCM-CACL, as shown in Figure 7. The visualization demonstrates high intra-class compactness and distinct inter-class boundaries among samples, validating the effectiveness and superiority of our approach in emotion classification tasks.

Conclusion

In this work, we propose DHCM-CACL, a novel self-supervised framework for multimodal emotion recognition. The Dynamic Hierarchical Cross-modal Mamba module (DHCM) ensures cross-modal feature alignment and reduces data heterogeneity. The Confidence-Adaptive Contrastive Learning module (CACL) enhances representation reliability and generalization in data-scarce scenarios. During the fine-tuning phase, evidence-aware optimization provides probabilistic credibility outputs. Experiments on the DEAP and MAHNOB-HCI datasets demonstrate SOTA performance in both subject-dependent and subject-independent settings. Our approach addresses the critical challenges of label scarcity and cross-modal asynchrony, establishing a new paradigm for reliable affective computing.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2024YFC3606900, 2024YFC3606903), the National Natural Science Foundation of China (62325301, U24B20186), and Zhejiang Provincial Natural Science Foundation of China (LZ23F030001).

References

Ahmed, N.; Al Aghbari, Z.; and Girija, S. 2023. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 17: 200171.

Bhatlawande, S.; Shilaskar, S.; Pramanik, S.; and Sole, S. 2024. Multimodal emotion recognition based on the fusion of vision, EEG, ECG, and EMG signals. *International journal of electrical and computer engineering systems*, 15(1): 41–58.

Chanel, G.; Rebetez, C.; Bétrancourt, M.; and Pun, T. 2011. Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(6): 1052–1063.

Chen, B.; Cao, Q.; Hou, M.; Zhang, Z.; Lu, G.; and Zhang, D. 2021. Multimodal emotion recognition with temporal and semantic consistency. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3592–3603.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Cheng, X.; Wei, W.; Du, C.; Qiu, S.; Tian, S.; Ma, X.; and He, H. 2022. VigilanceNet: Decouple intra-and inter-modality learning for multimodal vigilance estimation in RSVP-based BCI. In *Proceedings of the 30th ACM international conference on multimedia*, 209–217.

Choi, D. Y.; Kim, D.-H.; and Song, B. C. 2020. Multimodal attention network for continuous-time emotion recognition using video and EEG signals. *IEEE Access*, 8: 203814–203826.

Cui, K.; Li, J.; Liu, Y.; Zhang, X.; Hu, Z.; and Wang, M. 2025. PhysioSync: Temporal and Cross-Modal Contrastive Learning Inspired by Physiological Synchronization for EEG-Based Emotion Recognition. *IEEE Transactions on Computational Social Systems*, 1–14.

Cui, W.; Ma, Y.; Ren, J.; Liu, J.; Ma, G.; Liu, H.; and Li, Y. 2023. Personalized functional connectivity based spatio-temporal aggregated attention network for MCI identification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31: 2257–2267.

Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; and Ré, C. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359.

Fan, T.; Qiu, S.; Wang, Z.; Zhao, H.; Jiang, J.; Wang, Y.; Xu, J.; Sun, T.; and Jiang, N. 2023. A new deep convolutional neural network incorporating attentional mechanisms for ECG emotion recognition. *Computers in Biology and Medicine*, 159: 106938.

Fu, C.; Qian, F.; Su, K.; Su, Y.; Wang, Z.; Shi, J.; Liu, Z.; Liu, C.; and Ishi, C. T. 2025. HiMul-LGG: A hierarchical decision fusion-based local–global graph neural network for multimodal emotion recognition in conversation. *Neural Networks*, 181: 106764.

Gong, S.; Zhuge, Y.; Zhang, L.; Wang, Y.; Zhang, P.; Wang, L.; and Lu, H. 2025. Avs-mamba: Exploring temporal and multi-modal mamba for audio-visual segmentation. *IEEE Transactions on Multimedia*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Hazmoune, S.; and Bougamouza, F. 2024. Using transformers for multimodal emotion recognition: Taxonomies and state of the art review. *Engineering Applications of Artificial Intelligence*, 133: 108339.

Jafari, M.; Shoeibi, A.; Khodatars, M.; Bagherzadeh, S.; Shalbaf, A.; García, D. L.; Gorriz, J. M.; and Acharya, U. R. 2023. Emotion recognition in EEG signals using deep learning methods: A review. *Computers in Biology and Medicine*, 165: 107450.

Jia, Z.; Lin, Y.; Wang, J.; Feng, Z.; Xie, X.; and Chen, C. 2021. HetEmotionNet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition. In *Proceedings of the 29th ACM international conference on multimedia*, 1047–1056.

Jiang, W.-B.; Liu, X.-H.; Zheng, W.-L.; and Lu, B.-L. 2023. Multimodal adaptive emotion transformer with flexible modality inputs on a novel dataset with continuous labels. In *proceedings of the 31st ACM international conference on multimedia*, 5975–5984.

Kan, H.; Yu, J.; Huang, J.; Liu, Z.; Wang, H.; and Zhou, H. 2023. Self-supervised group meiosis contrastive learning for EEG-based emotion recognition: H. Kan et al. *Applied Intelligence*, 53(22): 27207–27225.

Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; and Patras, I. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1): 18–31.

Li, C.; Bao, Z.; Li, L.; and Zhao, Z. 2020. Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition. *Information Processing & Management*, 57(3): 102185.

Li, X.; Chen, C. P.; Chen, B.; and Zhang, T. 2024. Gusa: Graph-based unsupervised subdomain adaptation for cross-subject EEG emotion recognition. *IEEE Transactions on Affective Computing*, 15(3): 1451–1462.

Li, X.; Zhang, Y.; Tiwari, P.; Song, D.; Hu, B.; Yang, M.; Zhao, Z.; Kumar, N.; and Marttinen, P. 2022. EEG based emotion recognition: A tutorial and review. *ACM Computing Surveys*, 55(4): 1–57.

Liu, H.; Lou, T.; Zhang, Y.; Wu, Y.; Xiao, Y.; Jensen, C. S.; and Zhang, D. 2024a. EEG-based multimodal emotion recognition: A machine learning perspective. *IEEE Transactions on Instrumentation and Measurement*, 73: 1–29.

Liu, M.; Liang, K.; Zhao, Y.; Tu, W.; Zhou, S.; Gan, X.; Liu, X.; and He, K. 2024b. Self-supervised temporal graph learning with temporal and structural intensity alignment. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4): 6355–6367.

Ma, Y.; Cui, W.; Liu, J.; Guo, Y.; Chen, H.; and Li, Y. 2023. A multi-graph cross-attention-based region-aware feature fusion network using multi-template for brain disorder diagnosis. *IEEE Transactions on Medical Imaging*, 43(3): 1045–1059.

Meng, T.; Zhang, F.; Shou, Y.; Shao, H.; Ai, W.; and Li, K. 2024. Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Mohammadi Foumani, N.; Mackellar, G.; Ghane, S.; Irtza, S.; Nguyen, N.; and Salehi, M. 2024. Eeg2rep: enhancing self-supervised eeg representation through informative masked inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5544–5555.

Mohsenvand, M. N.; Izadi, M. R.; and Maes, P. 2020. Contrastive representation learning for electroencephalogram classification. In *Machine learning for health*, 238–253. PMLR.

Pan, B.; Hirota, K.; Jia, Z.; and Dai, Y. 2023. A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing*, 561: 126866.

Peng, Y.; Liu, H.; Li, J.; Huang, J.; Lu, B.-L.; and Kong, W. 2022. Cross-session emotion recognition by joint label-common and label-specific EEG features exploration. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31: 759–768.

Saxena, A.; Khanna, A.; and Gupta, D. 2020. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1): 53–79.

Shao, Y.; Li, Y.; and Wu, B. 2025. A Directional Attention Fusion and Multi-Head Spatial-Channel Attention Network for Facial Expression Recognition. *Journal of Machine Learning and Information Security*, 1(1): 7.

Soleymani, M.; Lichtenauer, J.; Pun, T.; and Pantic, M. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1): 42–55.

Sun, M.; Cui, W.; Zhang, Y.; Yu, S.; Liao, X.; Hu, B.; and Li, Y. 2023. Attention-rectified and texture-enhanced cross-attention transformer feature fusion network for facial expression recognition. *IEEE Transactions on Industrial Informatics*, 19(12): 11823–11832.

Tang, H.; Cao, M.; Huang, J.; Liu, R.; Jin, P.; Li, G.; and Liang, X. 2025. Muse: Mamba is efficient multi-scale learner for text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7238–7246.

Tu, G.; Xie, T.; Liang, B.; Wang, H.; and Xu, R. 2024. Adaptive graph learning for multimodal conversational emotion detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19089–19097.

Wang, Y.; Han, Y.; Wang, H.; and Zhang, X. 2023. Contrast everything: A hierarchical contrastive framework for medical time-series. *Advances in Neural Information Processing Systems*, 36: 55694–55717.

Wang, Y.; Zhang, B.; and Tang, Y. 2024. DMMR: Cross-subject domain generalization for EEG-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 628–636.

Yang, D.; Kuang, H.; Huang, S.; and Zhang, L. 2022. Learning modality-specific and-agnostic representations for asynchronous multimodal language sequences. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1708–1717.

Yang, Y.; Wang, Z.; Tao, W.; Liu, X.; Jia, Z.; Wang, B.; and Wan, F. 2024. Spectral-spatial attention alignment for multi-source domain adaptation in EEG-based emotion recognition. *IEEE Transactions on Affective Computing*, 15(4): 2012–2024.

Zhang, Y.; Yao, Y.; Liu, X.; Qin, L.; Wang, W.; and Deng, W. 2024. Open-set facial expression recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 646–654.

Zhu, Q.; Zheng, C.; Zhang, Z.; Shao, W.; and Zhang, D. 2023. Dynamic confidence-aware multi-modal emotion recognition. *IEEE Transactions on Affective Computing*, 15(3): 1358–1370.

Zong, Y.; Mac Aodha, O.; and Hospedales, T. 2023. Self-supervised multimodal learning: A survey. *arXiv preprint arXiv:2304.01008*.