Explainable Affective Body Expression Recognition with Multi-Scale Spatiotemporal Encoding and LLM-Based Reasoning

Tao Wang, Haifeng Lu, Jiayi Duan, Tianyu Meng, Rui Mao, *Member, IEEE,* Shuang Liu*, *Senior Member, IEEE,* and Dong Ming*, *Senior Member, IEEE,*

Abstract—Automatic emotion recognition (AER) based on 3D body expressions is a pivotal component of human-centered affective computing. Despite significant progress, existing methods often struggle to jointly capture multi-scale spatiotemporal features and bidirectional long-term dependencies, while lacking explainable reasoning mechanisms. To address these challenges, we propose the Explainable Affective Body Expression Recognition (EABER) framework. Specifically, EABER introduces a Multi-Scale Convolutional Mamba Network (MSCMNet). It employs multi-receptive-field convolutional modeling to learn multi-scale spatiotemporal representation, while a bidirectional state-space architecture is used to capture global bidirectional dependencies. Furthermore, EABER incorporates an LLM-based Emotion–Action Interpreter (EAI-LLM) for final emotion recognition with explainable affective reasoning. The spatiotemporal Serialization with Unified Masking (STSUM) strategy is introduced into the MSCMNet and EAI-LLM models, facilitating cross-dataset joint training while preserving fine-grained spatiotemporal information. Experimental results across four diverse datasets demonstrate that through joint training, the proposed EABER effectively learns shared affective motion patterns, yielding accuracy improvements of up to 7.83%. Moreover, EABER achieves superior performance in emotion recognition compared to state-of-the-art methods, and outperforms general-purpose large models, such as GPT-4o and Gemini 1.5 Pro, in explainable affective reasoning.

Index Terms—Explainable AI, Emotion recognition, 3D Skeleton sequences, Multiscale features, State Space Models (SSM), Large Language Model (LLM).

1 Introduction

Automatic emotion recognition (AER), the core technology enabling machines to perceive, interpret, and respond to human affective states [1], is the foundation for developing socialized human–machine collaborative systems and advancing toward Artificial General Intelligence (AGI) [2], [3]. In affective computing, compared with traditional modalities such as text [4], speech [5], and electroencephalography (EEG) [6], [7], body posture and movement offer distinct advantages in non-intrusive signal acquisition [8]. Human body language is inherently less susceptible to environmental interference or voluntary manipulation, providing a more objective and reliable reflection of an individual's internal affective state [9], [10]. Furthermore, the large visual scale of the human body facilitates non-contact perception at a distance [11]. These attributes ensure that posture-based emotion recognition remains robust in unconstrained, real-world scenarios, establishing it as an indispensable research frontier [12]. Moreover, posture-derived affective cues can serve as non-intrusive behavioral biomarkers for clinical applications, particularly in mental health assessments such as depression [13] and bipolar disorder [14].

Since affective body movements occur inherently in three-dimensional space, 3D skeleton data are widely recognized as one of the most intuitive and effective representations for analysis [15]. Skeleton data precisely encode spatial coordinates and rotational information of joints, providing rich spatiotemporal cues essential for decoding the complex mapping between physical kinematics and affective semantics [16]. With the emergence of cost-effective depth sensors and advanced real-time pose estimation algorithms [17], [18], acquiring high-fidelity 3D skeleton data has become efficient and accessible, further catalyzing research in this domain.

Despite these advancements, skeleton-based emotion recognition still faces several critical bottlenecks. First, existing studies lack the unified architecture that can simultaneously extract multi-scale spatiotemporal features and model bidirectional long-term dynamic dependencies. On the one hand, body expressions exhibit multi-scale representations ranging from transient micro-movements to regional structured posture patterns [19]. Affective states are conveyed through both long-term macroscopic postures (e.g.,

- *This work was supported by the National Key Research and Development Program of China under Grant 2023YFF1203700, the Science Fund for Distinguished Young Scholars of Tianjin under Grant 23JCJQJC00060, the Natural Science Foundation of Tianjin under Grant 24ZXZSSS00330, and the Autonomous Project of Haihe Laboratory of Brain-Computer Interaction and Human-Machine Integration under Grant 25HHNJSS00004. (Tao Wang and Haifeng Lu contributed equally to this work. Corresponding authors: Shuang Liu; Dong Ming.)*
- *Tao Wang, Jiayi Duan, Tianyu Meng, Shuang Liu, and Dong Ming are with the Haihe Laboratory of Brain-Computer Interaction and Human-Machine Integration, the Medical School, and the State Key Laboratory of Advanced Medical Materials and Devices, Tianjin University, Tianjin, 300072, China (e-mail: shuangliu@tju.edu.cn; richardming@tju.edu.cn).*
- *Haifeng Lu is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, China, and the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen, 518172, China*
- *Rui Mao is with the College of Computing and Data Science, Nanyang Technological University, Singapore, 639798, Singapore.*

walking with a lowered head) and short-term subtle actions (e.g., hand tremors) [20]. While prior works highlight the necessity of analysis across diverse temporal (e.g., 0.5 s to 5 s) [21] and spatial [22] scales, efficiently capturing these features simultaneously remains a challenge. On the other hand, accurate decoding of complex emotions relies heavily on aggregating bidirectional long-range context spanning both historical and future dynamics [16]. For instance, an initial arm-raising action can only be disambiguated between an angry strike or a joyful wave by incorporating the subsequent long-range trajectory. This necessitates modeling global dependencies while integrating past context and future evolution. However, prevailing architectures like 3D CNNs, Transformers, and State Space Model (SSM) often struggle to synergize precise multi-scale modeling with effective bidirectional global reasoning. Second, current approaches suffer from a significant lack of interpretability and the challenge of cross-dataset heterogeneity. Conventional deep learning models are often treated as "black boxes" [23], producing emotion labels without explaining how underlying body movements contribute to affective states, thereby undermining trust in human–machine interaction. Although Large Language Model (LLM) shows strong potential for explainable affective reasoning [24], their effective deployment typically depends on large-scale cross-dataset training, which is challenged by heterogeneity in joint definitions and sequence lengths [25]. Moreover, the preservation of fine-grained spatiotemporal information during skeleton encoding and subsequent LLM integration remains a significant challenge.

To address these challenges, we propose the Explainable Affective Body Expression Recognition (EABER) framework. By jointly modeling multi-scale spatiotemporal representations and global bidirectional dependencies, and further leveraging the semantic reasoning capability of LLM, EABER achieves high-accuracy emotion recognition while providing explainable affective reasoning. Specifically, for the first challenge, we employ a posture RGB image construction module to transform temporal actions into translation-invariant texture images. Building upon this, we design the Multi-Scale Convolutional Mamba Network (MSCMNet). It leverages Multi-Scale Convolutional Neural Network (MSCNN) with multi-receptive-field branches to extract features from micro-movements to structured postures, and integrates a Bidirectional Mamba (BMamba) module for efficient modeling of global bidirectional dependencies. To address the second challenge, we extend and optimize the LLM-based Emotion–Action Interpreter (EAI-LLM) proposed in our previous work [26]. Specifically, we replace the original skeleton encoder with MSCMNet to directly extract rich spatiotemporal features, and introduce the spatiotemporal Serialization with Unified Masking (STSUM) strategy to enable efficient joint training across heterogeneous datasets while preserving fine-grained spatiotemporal information. The optimized EAI-LLM enhances the accuracy of emotion recognition while generating more informative and explainable emotional descriptions.

Extensive experiments conducted on four datasets demonstrate that, through joint training, EABER significantly outperforms state-of-the-art methods. Moreover, EABER surpasses general-purpose large models, such as GPT-4o and Gemini 1.5 Pro, in explainable affective description generation, validating its effectiveness in explainable affective reasoning. The main contributions of this work are summarized as follows:

- We propose the EABER framework, which integrates multi-scale spatiotemporal representation learning with LLM-based affective reasoning, thereby achieving high recognition accuracy with interpretability.
- We design MSCMNet, which leverages multi-receptive-field MSCNNs to capture multi-scale features and incorporates a BMamba module to efficiently model long-term bidirectional dependencies.
- We introduce the STSUM strategy, which facilitates efficient cross-dataset joint training while preserving fine-grained spatiotemporal information during representation learning.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed EABER framework. Section 4 introduces the datasets and experimental setups. Section 5 reports the experimental results, and Section 6 concludes the paper.

2 Related Work

2.1 Skeleton-based Affective Body Expression Recognition

AER based on 3D skeleton sequences has evolved into a pivotal research domain, primarily due to its robustness in long-range data acquisition [27]. Extensive research has characterized the spatiotemporal evolution of body movements through diverse deep learning frameworks. Regarding multi-scale feature extraction, Camurri *et al.* [21] observed that macroscopic body expressions (e.g., walking) require sustained, whole-body observation, whereas subtle joint actions (e.g., trembling) necessitate short-term, part-wise modeling. Early methodologies relied heavily on handcrafted features to capture these multi-scale attributes. For instance, Glowinski *et al.* [28] identified emotions by extracting local extrema and peak duration ratios from upper-body movements. Similarly, Fourati *et al.* [29] proposed a comprehensive encoding scheme encompassing over 110 motion features across anatomical, spatial, and postural dimensions.

The subsequent paradigm shift toward deep learning has enabled the automated acquisition of hierarchical representations. Shirian *et al.* [30] introduced a Graph Inception Network (GIN) utilizing multi-branch graph convolutions to capture multi-scale dependencies. Bhattacharya *et al.* [31] developed the STEP framework, which aggregates local joint features into global posture representations via a hierarchical ST-GCN. To handle temporal variations, Beyan *et al.* [19] encoded 3D positions into RGB images representing different intervals, while Wang *et al.* [20] utilized Riemannian networks to capture geometric correlations on SPD manifolds. Despite these advancements, most existing methods either decouple spatial and temporal scales or rely on fixed receptive fields, failing to adaptively integrate transient micro-movements with regional structured patterns.

Beyond multi-scale modeling, accurate decoding of complex emotions necessitates robust bidirectional long-term

dynamic modeling. Early sequence-based efforts, such as those by Sapiński *et al.* [32], employed RNN and LSTM architectures to capture temporal dependencies. Subsequently, Zhang *et al.* [33] proposed an attention-based stacked LSTM (AS-LSTM) to prioritize key affective frames. However, recurrent architectures inherently struggle with vanishing gradients in extended sequences. While Transformers have gained traction due to their superior sequence modeling [34], [35], the quest for computational efficiency has led to the adoption of SSM, such as Mamba [36], as linear-complexity alternatives [37]. Nevertheless, standard SSM is inherently causal, which restricts the aggregation of future contextual information. In affective recognition, this absence of non-causal context is detrimental, as future motion cues are often essential for disambiguating early-stage emotional expressions.

In summary, while progress has been made in multi-scale extraction and long-term modeling, these capabilities remain fragmented. Existing frameworks struggle to balance rich multi-scale representation with efficient bidirectional reasoning. To address this limitation, we propose the EABER framework, which integrates multi-scale feature extraction with a bidirectional state space modeling mechanism for explainable affective body expression recognition.

2.2 Large Language Model

The emergence of LLM, such as ChatGPT [38] and LLaMA [39], has introduced unprecedented reasoning capabilities to the field [40]. Driven by the vision-language pre-training paradigm [41], researchers have begun extending LLM-based semantic understanding to skeleton-based analysis [42], [43], offering a promising route to improve the interpretability of traditional models. Current LLM-based skeleton analysis primarily utilizes contrastive learning to align skeleton and text encoder feature spaces. For example, Qu *et al.* [44] demonstrated the potential of LLM in decoding action semantics. Architectures like ActionCLIP [45] and the GAP framework [46] employ generative prompts to guide the learning of discriminative skeletal features.

However, utilizing LLM for emotion explanation generation is still in its infancy. Although Lu *et al.* [26] pioneered skeleton tokenization to endow models with preliminary interpretability, robust affective understanding remains challenging due to some critical issues. Most existing works prioritize cross-modal alignment of encoder outputs [47], while overlooking the preservation of fine-grained spatiotemporal dynamics within the encoder itself, which is crucial for identifying subtle micro-emotions.

To address these limitations, we incorporate EAI-LLM into the EABER framework. By introducing the STSUM strategy, our approach effectively alleviates the loss of information during cross-modal semantic mapping.

3 Method

3.1 Overall Architecture

This section details the proposed EABER framework. By integrating multi-scale spatiotemporal feature extraction with the semantic reasoning capability of LLM, EABER establishes an effective mapping from low-level physical actions to high-level affective semantics, with the aim of jointly achieving accurate emotion recognition and explainable affective reasoning. The overall architecture of EABER is shown in Fig. 1.

First, the model constructs posture RGB images from input skeleton sequences (Section 3.2), where joint trajectories are encoded into color channels. This transformation converts temporal motion into translation-invariant texture images, providing a standardized input for subsequent deep networks.

Subsequently, the posture RGB images are fed into MSCMNet (Section 3.3). To simultaneously capture multi-scale spatiotemporal features and long-term dynamic dependencies, MSCMNet employs three parallel multi-receptive-field branches, where kernels with different receptive fields introduce human kinematic priors to model patterns ranging from transient micro-movements to regional structured postures. Each branch in MSCMNet is composed of a stack of MSCNN blocks (see Section 3.3.1), which jointly models dependencies from whole-body to part-wise joints. Second, to establish a structured mapping from local convolutional features to global modeling, the framework incorporates the STSUM strategy (see Section 3.3.2), mitigating cross-dataset heterogeneity while preserving joint-level spatial integrity. Serialized features are then processed by BMamba (see Section 3.3.3). Leveraging the linear computational efficiency and selective scanning of SSM, BMamba enables bidirectional global reasoning. Finally, Adaptive Branch Fusion (ABF) (see Section 3.3.4) dynamically integrates complementary representations from different receptive fields, producing a highly compact and dimensionally unified spatiotemporal feature matrix $\bar{\mathbf{H}}$.

To further enhance the reasoning capability of EABER, we incorporate the EAI-LLM module (Section 3.4). Through LoRA-based instruction tuning, the model jointly generates emotion labels and explainable affective reasoning.

3.2 3D Posture Image Construction

To capture complex dynamics, we encode 3D skeletons into posture RGB images, unifying temporal evolution and spatial topology into a compact 2D matrix. This representation enables convolutional networks to model temporal evolution and spatial structure in a unified 2D format. The process is shown in the 3D Posture Image Construction module of Fig. 1.

Formally, for a sequence of T frames with J joints, the 3D coordinates of joint j at frame t are defined as

$$P_j^t = [x_j^t, y_j^t, z_j^t] \in \mathbb{R}^3, \tag{1}$$

where x_j^t, y_j^t, and z_j^t denote the Cartesian coordinates of joint j at time frame t.

To enhance positional consistency, we perform global spatial normalization using a torso-related reference joint of the first frame as the origin:

$$\hat{P}_j^t = P_j^t - P_{\text{ref}}^1 = [x_j^t - x_{\text{ref}}^1,\ y_j^t - y_{\text{ref}}^1,\ z_j^t - z_{\text{ref}}^1], \tag{2}$$

where P_{ref}^1 denotes the 3D position of the reference joint in the first frame. Crucially, the normalization preserves relative posture structure while retaining essential global displacement information.

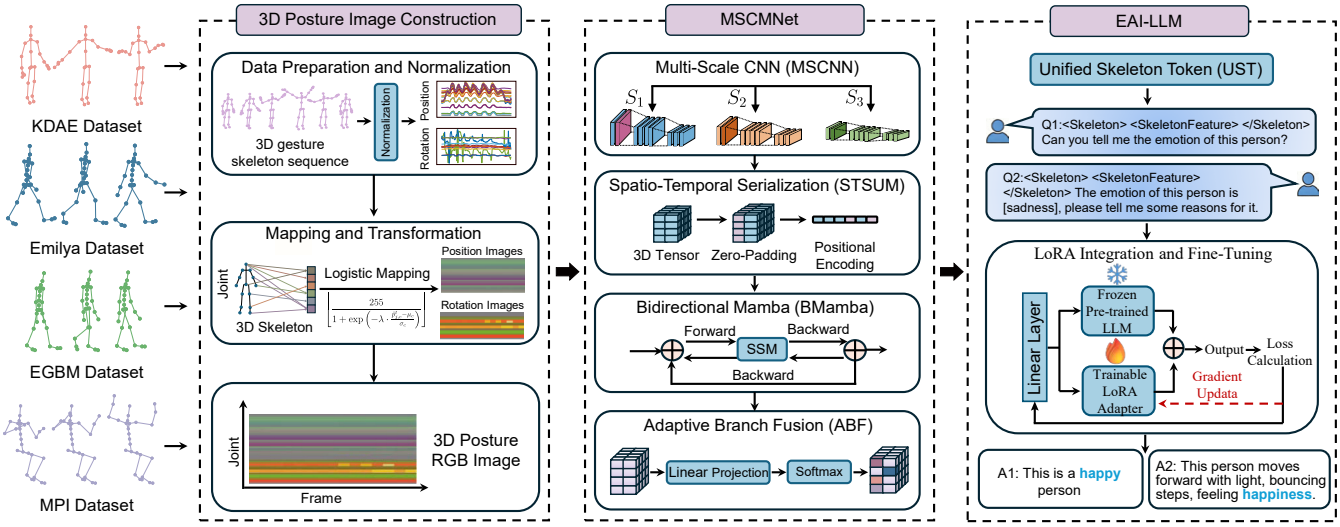

Fig. 1: Overview of the proposed EABER framework. 3D Posture Image Construction transforms skeleton sequences into posture RGB images. MSCMNet extracts multi-scale spatiotemporal representations from constructed posture images and captures global bidirectional dependencies. EAI-LLM interprets the learned representations for emotion recognition with explainable affective reasoning.

Subsequently, we employ a Logistic nonlinear mapping to project normalized coordinates into the RGB pixel value range $[0, 255]$, alleviating the impact of outliers. The mapping is formulated as

$$I_c(j,t) = \left\lfloor \frac{255}{1 + \exp\left(-\lambda \cdot \frac{\hat{p}^t_{j,c} - \mu_c}{\sigma_c}\right)} \right\rfloor, \tag{3}$$

where $c \in \{R, G, B\}$ maps to $\{x, y, z\}$ components, respectively; $\hat{p}^t_{j,c}$ represents the normalized coordinate of joint j at time t in channel c; μ_c and σ_c are the mean and standard deviation of normalized coordinates over all joints and frames in channel c. The parameter λ controls the sensitivity of the mapping.

The resulting pseudo-color image is denoted as $I \in \mathbb{R}^{J \times T \times 3}$, where the RGB channels encode the dynamic variations of the 3D coordinates. In addition to positional information, rotational data are encoded using the same strategy to generate a corresponding rotation posture image. The position and rotation images are then concatenated along the joint dimension, forming the final posture RGB tensor, which serves as the input for subsequent spatiotemporal feature extraction.

3.3 Multi-Scale Convolutional Mamba Network (MSCMNet)

MSCMNet serves as the core spatiotemporal feature extractor, hierarchically integrating a multi-branch MSCNN encoder with a BMamba reasoning module. Specifically, MSCNN captures local multi-scale spatiotemporal patterns, while BMamba models global bidirectional dependencies, together enabling more comprehensive spatiotemporal representation learning. As illustrated in Fig. 2, it is designed to capture multi-scale features and long-term dynamic dependencies through three parallel multi-receptive-field branches, denoted as $\mathcal{S} = \{s_1, s_2, s_3\}$.

The workflow of MSCMNet consists of four collaborative stages. First, to handle variable sequence lengths, it partitions the input into N consecutive subsegments to construct a standardized input interface. In each branch, MSCNN layers perform joint spatiotemporal encoding using kernels with diverse receptive fields, explicitly introducing human kinematic priors to model patterns from transient micro-movements to regional structured postures.

To stabilize training and promote feature interaction, MSCMNet incorporates two connectivity mechanisms: Intra-branch residual connections to prevent degradation and facilitate gradient propagation in deeper convolutional transformations, and cross-receptive-field fusion connections to enable lateral information exchange across parallel branches with different receptive fields, thereby integrating complementary local patterns at multiple scales. Although the convolutional layers excel at extracting local features, their limited receptive fields necessitate the introduction of BMamba to capture bidirectional long-range context. Instead of conventional pooling, the STSUM strategy is employed to preserve physically interpretable joint-level semantics. Finally, the ABF module adaptively merges multi-scale embeddings into a compact, semantically rich representation $\bar{\mathbf{H}}$, which is subsequently aligned for the subsequent EAI-LLM module.

3.3.1 *Multi-Scale Convolutional Neural Network (MSCNN)*

MSCNN adopts a dual-stream architecture (Fig. 3(a)) to simultaneously model global human topology and fine-grained anatomical details. The Coarse-Grained Branch captures whole-body topological dependencies, while the Fine-Grained Branch partitions joints into K non-overlapping local joint subsets $\{\mathcal{G}_k\}_{k=1}^{K}$, each corresponding to a specific

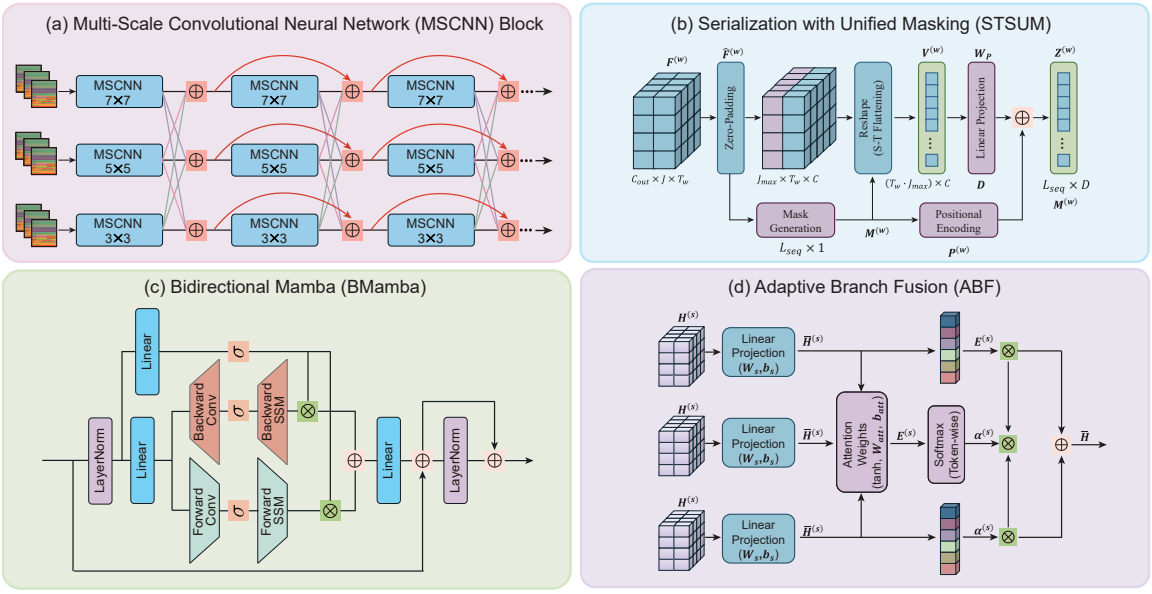

Fig. 2: The structure of the MSCMNet. (a) MSCNN: Extracts hierarchical spatiotemporal features from posture images using parallel convolutional branches with diverse receptive fields ($7 \times 7, 5 \times 5, 3 \times 3$). (b) STSUM: Transforms high-dimensional feature maps into a standardized sequential format to resolve cross-dataset heterogeneity while preserving essential skeletal topology. (c) BMamba: Parallel forward and backward scanning paths utilizing selective SSM and an adaptive gating mechanism enable bidirectional modeling of long-term dynamic dependencies by capturing action initiation and termination logic. (d) ABF: Adaptively integrates multi-scale representations via a token-wise dynamic gating mechanism to generate a compact and semantically rich affective embedding.

local body part or joint cluster (e.g., left upper limb or right lower limb), to extract part-wise semantics.

Formally, the MSCNN module takes the w-th subsegment posture image $I^{(w)} \in \mathbb{R}^{J \times T_w \times 3}$ defined in the previous section as input. To accommodate convolution operators and unify notation, $I^{(w)}$ is first rearranged into a channel-first format $\mathbf{I}^{(w)} \in \mathbb{R}^{C_{in} \times J \times T_w}$ (with $C_{in} = 3$) and normalized using Batch Normalization (BN), formulated as follows:

$$\widetilde{\mathbf{I}}^{(w)} = \mathrm{BN}(\mathbf{I}^{(w)}) \in \mathbb{R}^{C_{in} \times J \times T_w}. \tag{4}$$

Subsequently, within parallel coarse- and fine-grained branches, MSCNN performs multi-level feature integration via coarse-to-fine joint fusion. To be specific, the coarse-grained output is computed as:

$$\mathbf{F}_{\mathrm{CG}}^{(w)} = \sigma\left(\mathbf{W}_{\mathrm{CG}} * \widetilde{\mathbf{I}}^{(w)} + \mathbf{b}_{\mathrm{CG}}\right), \tag{5}$$

where $\mathbf{W}_{\mathrm{CG}}$ and $\mathbf{b}_{\mathrm{CG}}$ denote the learnable parameters of the coarse-grained branch; $\sigma(\cdot)$ is the LeakyReLU activation function. While the fine-grained features are extracted via independent 2D convolutions on joint subsets $\{\mathcal{G}_k\}_{k=1}^{K}$:

$$\mathbf{F}_{\mathrm{FG}}^{(w,k)} = \sigma\left(\mathbf{W}_{\mathrm{FG}}^{(k)} * \widetilde{\mathbf{I}}^{(w,k)} + \mathbf{b}_{\mathrm{FG}}^{(k)}\right). \tag{6}$$

where $\widetilde{\mathbf{I}}^{(w,k)}$ is the feature tensor of the k-th subset given by:

$$\widetilde{\mathbf{I}}^{(w,k)} = \mathcal{S}_k(\widetilde{\mathbf{I}}^{(w)}) \in \mathbb{R}^{C_{in} \times J_k \times T_w}, \tag{7}$$

and $\mathcal{S}_k(\cdot)$ denotes the sub-domain selection operator, $J_k = |\mathcal{G}_k|$ is the number of joints in the subset.

After concatenating all subsets along the joint dimension to form $\mathbf{F}_{\mathrm{FG}}^{(w)}$, a coarse-to-fine joint fusion mechanism injects whole-body priors into local features:

$$\widetilde{\mathbf{F}}_{\mathrm{out}}^{(w)} = \mathbf{F}_{\mathrm{FG}}^{(w)} \oplus \mathbf{F}_{\mathrm{CG}}^{(w)}, \tag{8}$$

where $\oplus$ denotes element-wise addition. The fused feature $\widetilde{\mathbf{F}}_{\mathrm{out}}^{(w)}$ is then processed by a sequence of post-processing operations. As shown in Fig. 3, the final output $\mathbf{F}_{\mathrm{MSCNN}}^{(w)} \in \mathbb{R}^{C_{\mathrm{out}} \times J \times T_w}$ is obtained after average pooling and point-wise convolution, providing an anatomically grounded precursor for temporal modeling, which is defined as

$$\mathbf{F}_{\mathrm{MSCNN}}^{(w)} = \sigma\left(\mathbf{W}_{\mathrm{post}} * \mathrm{AP}\big(\sigma(\widetilde{\mathbf{F}}_{\mathrm{out}}^{(w)})\big) + \mathbf{b}_{\mathrm{post}}\right), \tag{9}$$

where $\mathrm{AP}(\cdot)$ denotes the average pooling operator for reducing spatial resolution, and $\mathbf{W}_{\mathrm{post}}$ is a 1×1 convolution kernel used for inter-channel information interaction and dimensional mapping.

3.3.2 spatiotemporal Serialization with Unified Masking (STSUM)

STSUM transforms high-dimensional convolutional maps into a sequential format compatible with the linear input requirements of BMamba. Crucially, it employs a unified masking mechanism to eliminate cross-dataset heterogeneity while strictly preserving joint topological information. Given the feature tensor of the w-th subsegment $\mathbf{F}^{(w)}$, which is rearranged into a joint-first format $J \times T_w \times C$, zero-padding is applied along the joint dimension to a global maximum J_{max}:

$$\hat{\mathbf{F}}^{(w)} = \mathrm{Pad}(\mathbf{F}^{(w)}) \in \mathbb{R}^{J_{max} \times T_w \times C}. \tag{10}$$

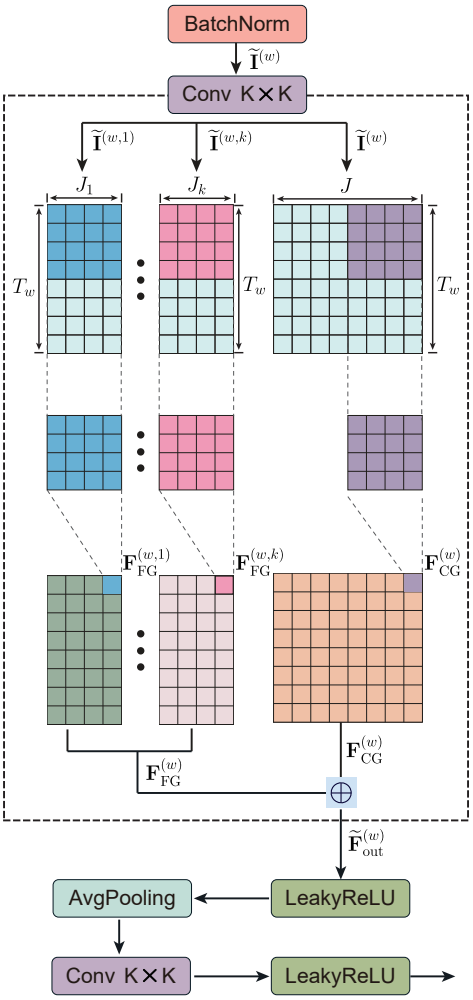

Fig. 3: The structure of the Multi-Scale Convolutional Neural Network (MSCNN) module.

We then perform spatiotemporal flattening into a single sequence $\mathbf{V}^{(w)} \in \mathbb{R}^{(T_w \cdot J_{max}) \times C}$. The resulting sequence is projected to a latent space and augmented with learnable positional encodings $\mathbf{P}^{(w)}$:

$$\mathbf{Z}^{(w)} = \mathbf{V}^{(w)}\mathbf{W}_p + \mathbf{P}^{(w)} \in \mathbb{R}^{L_{seq} \times D}. \tag{11}$$

where $L_{seq} = T_w \cdot J_{max}$ denotes the sequence length and D is the embedding dimension. A binary mask $\mathbf{M}^{(w)} \in \{0,1\}^{L_{seq}}$ is generated simultaneously to filter noise from padding, in which positions corresponding to valid joint features are marked as 1, while padded positions are marked as 0. All subsegment tokens and masks are concatenated in temporal order to form the global sequence $\mathbf{Z}_{full}$ and global mask $\mathbf{M}_{full}$, which are subsequently fed into the BMamba module for bidirectional long-term contextual reasoning.

3.3.3 Bidirectional Mamba (BMamba)

To capture the global long-term dynamic dependencies inherent in affective body expressions, we introduce an SSM-based Bidirectional Mamba (BMamba) module, as shown in Fig. 2(c), which overcomes the unidirectional constraints of standard Mamba by employing parallel forward and backward scanning paths for full-context reasoning. The input $\mathbf{Z}_{full} \in \mathbb{R}^{L \times D}$ (where $L = T \cdot J_{max}$ denotes the global sequence length), is normalized via Layer Normalization (LayerNorm) and then split into a main stream $\mathbf{E}$ and a gating stream $\mathbf{G}$. In each branch, features are processed through Causal Conv1D to extract local neighborhood features and then fed into a directional SSM for global scanning:

$$\begin{aligned} \mathbf{H}_{fwd} &= \mathrm{SSM}_{fwd}\big(\sigma(\mathrm{Conv1D}_{fwd}(\mathbf{E}))\big), \\ \mathbf{H}_{bwd} &= \mathrm{SSM}_{bwd}\big(\sigma(\mathrm{Conv1D}_{bwd}(\mathbf{E}))\big), \end{aligned} \tag{12}$$

where $\sigma(\cdot)$ denotes the SiLU activation function. The forward SSM_{fwd} performs recursive state updates from $t = 1$ to L, capturing the initiation and progression logic of actions (forward dynamics), while the backward SSM_{bwd} scans from $t = L$ to 1, focusing on the termination and convergence states of actions. By jointly modeling both forward progression and backward dependency cues, BMamba provides a more complete characterization of action causality than standard unidirectional Mamba, which only captures one-way temporal evolution.

To effectively integrate bidirectional context and regulate information flow, BMamba introduces an adaptive gating mechanism. After SiLU activation, the gating stream $\mathbf{G}$ serves as a modulation factor and interacts with the outputs of the bidirectional SSM via element-wise multiplication. The modulated features are then aggregated by addition and mapped back to the original dimension through a linear output projection layer. To alleviate gradient vanishing issues in deep networks and enhance feature reuse, a residual connection is introduced at the end of the module. The final bidirectionally gated global long-term representation $\mathbf{H}$ is computed as:

$$\mathbf{Y} = [\mathbf{H}_{fwd} \otimes \sigma(\mathbf{G})] \oplus [\mathbf{H}_{bwd} \otimes \sigma(\mathbf{G})], \tag{13}$$

$$\mathbf{H} = \mathrm{LN}(\mathbf{Z}_{full} + \mathrm{Linear}_{out}(\mathbf{Y})). \tag{14}$$

where $\otimes$ and $\oplus$ denote element-wise multiplication and addition, respectively, LN represents LayerNorm, and Linear_{out} is the output projection layer. This design endows the encoder with robust reasoning over the entire motion sequence while maintaining linear computational complexity.

3.3.4 Adaptive Branch Fusion (ABF)

The ABF module dynamically integrates complementary representations from the multi-receptive-field branches. By discarding pooling in favor of a token-wise dynamic gating mechanism, ABF assigns distinct weights to each spatiotemporal token, ensuring fine-grained complementarity.

Specifically, branch features $\mathbf{H}^{(s)}$ from each receptive field are projected to a shared space $\bar{\mathbf{H}}^{(s)}$. A lightweight generator then computes attention scores $\mathbf{E}^{(s)} = \tanh(\bar{\mathbf{H}}^{(s)}\mathbf{W}_{att} + \mathbf{b}_{att})$, which are normalized via Softmax across branch dimension $\mathcal{S}$ to generate the gating map $\boldsymbol{\alpha}^{(s)}$:

$$\boldsymbol{\alpha}^{(s)} = \exp(\mathbf{E}^{(s)}) / \sum_{s' \in \mathcal{S}} \exp(\mathbf{E}^{(s')}). \tag{15}$$

The final fused representation $\bar{\mathbf{H}}$ is obtained by computing an element-wise product between each branch feature

Emotion Recognition Prompt Template

#Human: <Skeleton> <Skeleton Feature> </Skeleton> Can you tell me the emotion of this person? #Assistant:

Emotion Description Prompt Template

#Human: <Skeleton> <Skeleton Feature> </Skeleton> The emotion of this person is [shame], please tell me some reasons for it. #Assistant:

Fig. 4: Prompt designs for fine-tuning LLM.

and its corresponding gating map, followed by summation across all branches:

$$\bar{\mathbf{H}} = \sum_{s \in \mathcal{S}} \boldsymbol{\alpha}^{(s)} \odot \bar{\mathbf{H}}^{(s)}, \tag{16}$$

which constitutes a semantically rich affective representation, aligned via linear projection for the subsequent EAI-LLM.

3.4 EAI-LLM: Skeleton-Aware Language Model

Although MSCMNet can extract highly discriminative spatiotemporal features from 3D body expressions, conventional classification heads (e.g., fully connected layers) are limited to producing discrete emotion labels without providing explainable affective reasoning. To address this limitation, we adopt the EAI-LLM.

In EAI-LLM, we improve the original skeleton encoder by using MSCMNet. Since MSCMNet has already extracted spatiotemporal features and STSUM has unified heterogeneous skeleton data through masking mechanisms, the MGST and UST modules in the original EAI-LLM are removed. The remaining LLM-based inference components are retained for explainable emotion recognition.

LLMs are primarily designed to process natural-language tokens, whereas skeleton data consist of variable-length spatiotemporal sequences that do not naturally reside in the LLM embedding space. To enable emotion prediction and text-based reasoning from skeleton inputs, we adapt the LLM using Low-Rank Adaptation (LoRA) [48], which introduces lightweight trainable modules while keeping the pre-trained weights fixed. For each sample (skeleton sequence, emotion label, and emotion description), we first extract 768-dimensional spatiotemporal tokens from the skeleton sequence using MSCMNet. Following common practice in multimodal LLM [49], we further employ a linear layer to project these tokens from a 768-dimensional space into a 4096-dimensional space, thereby making them suitable for processing by LLM. Furthermore, we incorporate the contrastive learning-based alignment strategy from EAI-LLM to align skeletal features with linguistic semantics. This strategy effectively anchors the spatiotemporal features from MSCMNet to their corresponding affective meanings, ensuring the semantic compatibility of skeleton tokens with the linguistic space before they are fed into the LLM.

We employ two instruction prompts in the LLaMA-2 dialogue format, one for emotion recognition and the other for explanation generation, as illustrated in Fig. 4. The projected skeleton tokens are inserted into the prompt via a placeholder, and the emotion label is provided as conditioning information for the description task. The skeleton tokens and prompt tokens are concatenated and fed into the LLM. Training is performed with a cross-entropy loss between the predicted tokens and the ground-truth tokens, while updating only the LoRA parameters (Eq. 17).

$$\mathcal{L}_{LoRA} = \mathcal{L}_{ce}(t_p, t_g), \tag{17}$$

where $\mathcal{L}_{ce}(.,.)$ denotes the cross-entropy loss, t_p represents the token sequence (or token distribution) predicted by the LLM, and t_g denotes the corresponding ground-truth target sequence derived from the annotated labels. For more details on the EAI-LLM, please refer to the work of [26].

4 Experiments

4.1 Datasets

The efficiency and generalizability of the proposed method were verified on four publicly available emotional gesture datasets. These datasets were collected using different devices, and participants from various regions were included. A brief description of these four databases is given below.

EGBM [50]: This dataset was captured by a Kinect V2 sensor at a frame rate of 30 Hz. It contains 560 samples performed by 16 Polish professional actors. The dataset includes 7 emotions: happiness, sadness, neutral, anger, disgust, fear, and surprise. Each sample provides 3D coordinates for 25 body joints.

KDAE [51]: This dataset was recorded using the Noitom Perception Neuron (PN) system at 125 Hz. It comprises 1,402 samples from 22 Chinese actors, covering the same 7 emotions as EGBM. The original data contains 72 body markers, from which we selected 24 joints for this study. Specific details can be found in [20].

Emilya [52]: This dataset was also recorded by the Xsens MVN system at 120 Hz. It consists of 8,206 samples performed by 12 actors. The actors performed 8 daily actions (e.g., sitting, walking, lifting) while expressing 8 emotions: anxiety, pride, happiness, sadness, fear, shame, anger, and neutral. Each posture segment contains 3D position data for 28 joints.

MPI [53]: This dataset was recorded by the Xsens MVN motion capture system at a sampling rate of 120 Hz. It contains 1,447 body motion samples performed by 8 actors, representing 11 emotions: anger, fear, happiness, pride, sadness, surprise, relief, disgust, neutral, amusement, and shame. Each posture segment contains 3D coordinates for 28 body joints.

Emotion Description Annotation: The original datasets only provide discrete emotion labels. To support the emotion description task, we manually annotated a subset of the data with fine-grained textual descriptions. Specifically, we labeled 180 samples from Emilya, 120 samples from KDAE, and 150 samples from MPI, resulting in a total of 450 skeleton-text pairs for instruction tuning.

4.2 Implementation Details

All experiments are implemented in PyTorch and conducted on a server with 5 NVIDIA A100 GPUs. The samples shorter

than 2 seconds are discarded. Each sequence is adjusted to 360 frames by padding with previous frames or uniform downsampling. For posture image construction, the Logistic mapping uses $\lambda = 0.01$. In MSCMNet, each 360-frame sequence is split into $N = 5$ subsegments and processed by three parallel branches with kernel sizes 7×7, 5×5, and 3×3 (stride 1, "same" padding). Each branch contains four MSCNN blocks with channel widths 32/64/128/256 and uses $K = 5$ joint subsets, followed by average pooling and a 1×1 convolution. In STSUM, all skeletons are aligned and padded to $J_{\max} = 28$ based on the Emilya 28-joint template, and then fed into a 2-layer BMamba. ABF projects branch features to 768 dimensions and applies token-wise gating to obtain the final 768-dim skeleton tokens for EAI-LLM. For EAI-LLM, we use LLaMA-2-7B [39] and project the 768-dim skeleton token to the 4096-dim LLM space. We apply LoRA and update only the LoRA parameters and projection layers (rank 64, $\alpha = 16$, dropout 0.05), while freezing the pre-trained LLaMA weights. Prompt templates are given in Sec. 3.4.

For Training Strategy, we adopt a two-stage training scheme. First, MSCMNet is pre-trained with SGD for 500 epochs (batch size 64, initial lr 0.01) using early stopping; if the validation loss plateaus for 10 epochs, the learning rate is decayed. Second, we fine-tune MSCMNet together with the EAI-LLM via instruction tuning. Following [26], we perform instruction tuning in the order of emotion description and then emotion recognition to avoid mutual interference: the description task is trained for 10,000 steps (global batch size 16, max lr 1×10^{-5}), and emotion recognition for 800,000 steps (global batch size 64) with the same LoRA settings. For all experiments, we use a 4:1 train/test split.

For evaluation, in the recognition task, we parse the emotion label from the generated text and compute accuracy; outputs with multiple labels or no valid label are counted as Error, while synonyms/morphological variants are treated as correct. In the description task, we report BLEU, ROUGE, and METEOR.

5 Result

5.1 Comparisons for Emotion Recognition

To evaluate the efficacy of EABER in AER, we conducted a comparative analysis with current state-of-the-art (SOTA) skeleton-based methods. To ensure a fair comparison, all baseline models were re-implemented and evaluated using the same training protocols as our framework. As shown in Table 1, EABER achieved superior performance across most benchmarks, with the exception of the EGBM dataset, where it was slightly outperformed by ST-ITE. Specifically, on the KDAE, Emilya, and MPI datasets, EABER improved the classification accuracy by 0.24%, 2.02%, and 3.51%, respectively, compared with the best existing methods. Notably, on the MPI dataset, which features complex categories and subtle micro-motions, EABER achieved an accuracy of 79.93%. This result represents nearly a 10% performance improvement over the EAI-LLM baseline. These results demonstrate that EABER enables high-precision and generalizable cross-dataset emotion recognition by synergizing multi-scale feature fusion with bidirectional temporal modeling and the reasoning power of LLM.

TABLE 1: Comparisons of emotion recognition accuracy (%) with state-of-the-art methods on four datasets. Bold indicates the best performance for each dataset.

Method	EGBM	KDAE	Emilya	MPI
AGCN [54]	22.94	56.58	88.92	46.81
CTR-GCN [55]	63.30	70.46	89.77	63.49
GAP [46]	66.06	67.26	89.16	70.37
EAI-LLM [26]	66.97	71.17	85.44	69.95
Multiscale CNN [19]	78.81	81.46	88.07	73.18
ATSFF [10]	75.26	77.44	85.84	71.01
ST-ITE [20]	**82.60**	85.91	89.61	76.42
EABER (Ours)	81.44	**86.15**	**91.79**	**79.93**

5.2 Comparisons for Explainable Emotion Description

This subsection evaluates the capability of the model to generate interpretable affective descriptions. Since existing multimodal LLM (e.g., GPT-4o, Gemini) do not support direct 3D skeleton input, we visualized the skeleton sequences as videos and tested them using the same prompts. Table 2 presents the quantitative results, where EABER outperformed all existing LLM across all linguistic metrics. These results validate that our MSCMNet and the internal STSUM strategy effectively extract and preserve multi-scale spatiotemporal affective cues. This enables the model to accurately describe key physical features, such as "head drooping" or "shoulder contraction," rather than generating hallucinations.

TABLE 2: Comparison of emotion description capabilities with mainstream LLM. Bold indicates the best performance for each metric.

Method	ROUGE	BLEU	METEOR
GPT-4o	0.1154	0.0615	0.1942
Gemini 1.5 Pro	0.0982	0.0523	0.1765
EAI-LLM	0.1945	0.1108	0.2096
EABER (Ours)	**0.2312**	**0.1487**	**0.2246**

Furthermore, Fig. 5 visualizes the generation results. The descriptions generated by EABER exhibit significant consistency with the input sequences. In contrast, while GPT-4o and Gemini 1.5 Pro generate affective descriptions, their outputs largely reflect the models' internal prior knowledge and often fail to match the specific input sequences. While EAI-LLM produces results similar to ours, it remains inferior in terms of the specific performance metrics detailed in Table 2.

5.3 Evaluation of Cross-Dataset Training Strategies

To verify the effectiveness of collaborative learning from multi-source heterogeneous data, we compared separate training with joint training paradigms. As shown in Table 3, the joint training approach yielded significant performance

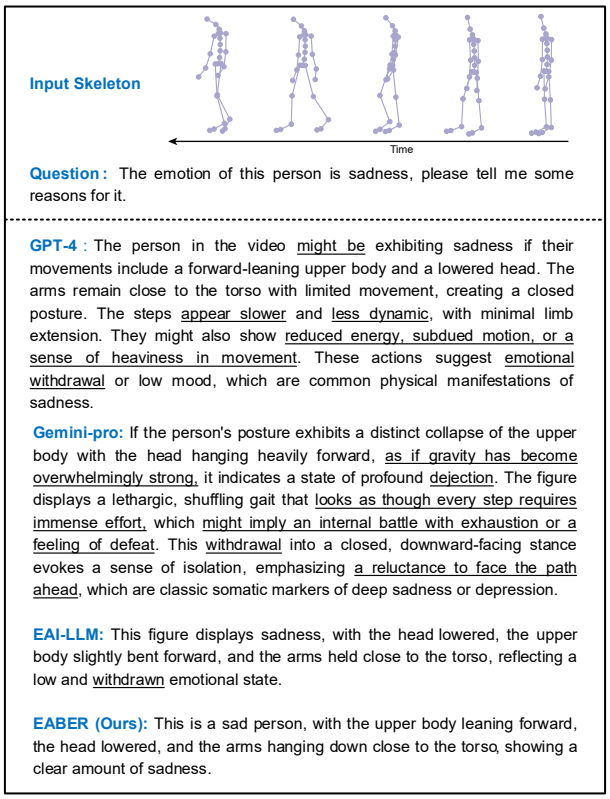

Fig. 5: Examples for emotion description capabilities of EABER on the Emilya dataset. Underlined text indicates that the description is unrelated to the input sequences.

gains across all datasets, with accuracy improvements ranging from 4.32% to 7.82%. These results demonstrate that through joint training, EABER effectively mitigates overfitting by learning diverse motion patterns and acquires shared underlying affective motion laws across domains, thereby enhancing its generalization capability. This improvement may also partly stem from the STSUM strategy, which likely helps capture shared affective information across heterogeneous datasets while retaining complementary dataset-specific characteristics.

TABLE 3: Comparison of recognition accuracy (%) between separate training and joint training across four datasets. Bold indicates the better performance for each dataset.

Dataset	Separate Training	Joint Training
EGBM	73.62	**81.44**
KDAE	81.39	**86.15**
Emilya	87.47	**91.79**
MPI	74.80	**79.93**

To further visualize the optimization effect of joint training on feature learning, we utilized t-SNE to map the latent feature distributions of the KDAE test set. As illustrated in Fig. 6, feature distributions under the separate training paradigm exhibit significant overlap, indicating difficulty in distinguishing similar emotions. In contrast, joint training constructs a more structured feature space with significantly higher intra-class compactness and inter-class separability. This confirms that EABER optimizes the data structure of spatiotemporal features under joint training, building more discriminative affective representations. Consequently, all subsequent experiments utilized the joint training strategy unless otherwise specified.

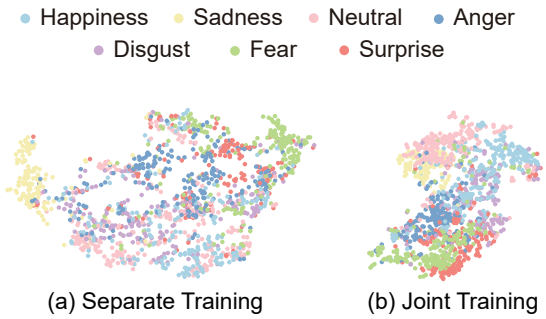

Fig. 6: The t-SNE visualization of learned feature embeddings on the KDAE dataset under (a) separate training and (b) joint training.

5.4 Ablation Studies

5.4.1 *Effectiveness of Multi-Receptive-Field Branches*

MSCMNet extracts multi-scale spatiotemporal features through the fusion of multiple receptive-field branches. To validate this mechanism, we conducted ablation experiments by denoting the three branches as $S_1(3\times3)$, $S_2(5\times5)$, and $S_3(7\times7)$. We tested single-branch, dual-branch, and the full triple-branch configurations, with results shown in Fig. 7. The results indicate that multi-receptive-field combinations generally outperform single-receptive-field configurations, confirming the limitations of a fixed kernel size in balancing transient micro-motions and regional structural patterns. Among dual-branch combinations, the $S_1 + S_3$ configuration performed best, suggesting that the feature span between the smallest and largest receptive fields provides highly discriminative complementary information. Nevertheless, the full $S_1 + S_2 + S_3$ architecture consistently achieved the highest accuracy, improving performance by 1.09%–3.59% over $S_1 + S_3$, which underscores the necessity of the triple-branch synergy in capturing complex affective postures.

5.4.2 *Effectiveness of MSCNN and BMamba Architectures*

To verify the architectural rationale of MSCNN and BMamba, we performed progressive ablation studies on the Coarse-Grained (CG) branch, Fine-Grained (FG) branch, and the bidirectional scanning mechanism. Table 4 summarizes the results.The experiments show that a single-dimensional joint representation is insufficient to capture the full complexity of affective expression. When using Mamba or BMamba, combining both branches (CG + FG) improved average accuracy by 3.19%–9.34% and 3.31%–7.38%, respectively, proving the strong complementarity between global topology and local anatomical details. Furthermore, BMamba consistently outperformed unidirectional Mamba across all configurations, demonstrating its

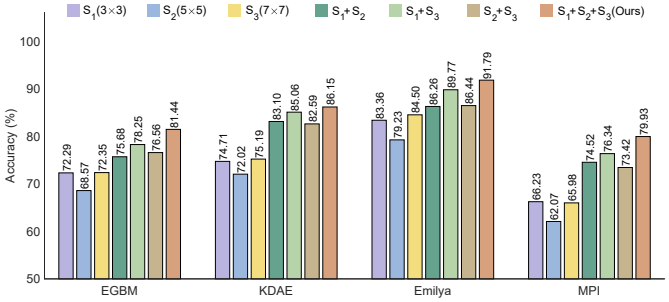

Fig. 7: Comparison of recognition accuracies (%) between different receptive-field branch combinations in MSCMNet on the (a) EGBM dataset, (b) KDAE dataset, (c) Emilya dataset, and (d) MPI dataset.

TABLE 4: Comparison of recognition accuracy (%) between different internal configurations of MSCNN and BMamba across four datasets. Bold indicates the best performance for each dataset.

	EGBM	KDAE	Emilya	MPI
CG + Mamba	67.85	74.70	83.30	69.73
CG + BMamba	74.26	76.37	87.63	74.12
FG + Mamba	72.40	74.24	85.53	71.79
FG + BMamba	77.15	78.77	87.25	76.62
CG + FG + Mamba	78.74	83.58	90.55	74.98
CG + FG + BMamba (Ours)	**81.44**	**86.15**	**91.79**	**79.93**

ability to extract both forward dynamics and backward termination cues for more complete spatiotemporal modeling. Ultimately, the full configuration (CG + FG + BMamba) achieved the highest accuracy across all datasets.As shown in the radar charts in Fig. 8, our proposed configuration provides the most balanced and superior performance across all emotion categories, reinforcing that the synergy between MSCNN's multi-granularity fusion and BMamba's bidirectional modeling enhances both overall precision and robustness.

5.4.3 *Effectiveness of STSUM*

The STSUM strategy was designed to reconstruct joint information while preserving fine-grained topology for the EAI-LLM. We compared STSUM with five spatial aggregation methods: (1) GAP, (2) GMP, (3) SAP, (4) HD-GCN, and (5) GJZM. Evaluation results are shown in Tables 5 and 6. As indicated in Table 5, STSUM yielded the best results on most datasets. Compared to traditional pooling (GAP/GMP), STSUM improved accuracy by 3.82%–7.25%, confirming that mandatory spatial compression leads to the loss of critical local motion details. Even compared to advanced methods like SAP, HD-GCN, and GJZM, STSUM achieved the highest accuracy on three datasets, demonstrating its ability to capture posture dynamics under cross-dataset heterogeneity.

In the explanation generation task (Table 6), GAP and GMP suffered a significant performance drop, supporting our hypothesis that traditional pooling causes Spatial Semantic Collapse. While GJZM slightly outperformed STSUM in BLEU score, STSUM achieved the highest ROUGE and METEOR scores. This indicates that by preserving full joint topology and motion trajectories, STSUM provides high-fidelity inputs for downstream LLM, effectively mitigating the hallucination phenomenon in generative tasks.

TABLE 5: Comparison of recognition accuracy (%) between different spatial aggregation strategies across four datasets. Bold indicates the best performance for each dataset.

	EGBM	KDAE	Emilya	MPI
GAP [56]	75.41	82.33	85.98	73.14
GMP [57]	74.19	81.66	85.74	73.53
SAP [58]	76.75	82.30	86.54	74.11
HD-GCN [59]	78.22	84.53	86.38	77.62
GJZM [60]	**82.02**	84.29	87.71	74.36
STSUM (Ours)	81.44	**86.15**	**91.79**	**79.93**

TABLE 6: Comparison of emotion description capabilities between different spatial aggregation strategies across four datasets. Bold indicates the best performance for each metric.

Method	ROUGE	BLEU	METEOR
GAP [56]	0.0926	0.0851	0.1139
GMP [57]	0.1044	0.0861	0.1294
SAP [58]	0.1839	0.1374	0.1659
HD-GCN [59]	0.1627	0.1248	0.1763
GJZM [60]	0.2216	**0.1572**	0.1931
STSUM (Ours)	**0.2312**	0.1487	**0.2246**

6 Conclusion

This study introduces the Explainable Affective Body Expression Recognition (EABER) framework, providing a unified and interpretable solution for decoding the complex mapping between emotional body motions and high-level affective semantics. By integrating the MSCMNet with the LLM-based EAI-LLM, the proposed framework effectively captures affective cues ranging from transient micro-movements to global long-term dynamic dependencies, achieving explainable emotion recognition. Specifically, the MSCNN module utilizes multi-receptive-field branches to learn multi-scale spatiotemporal representations, while the BMamba module enables efficient bidirectional global reasoning over entire action sequences with linear computational complexity. Furthermore, the introduction of the STSUM strategy successfully resolves the structural discrepancies across heterogeneous datasets, facilitating large-scale joint training and knowledge integration. Extensive experimental results demonstrate that EABER significantly outperforms state-of-the-art methods in emotion recognition

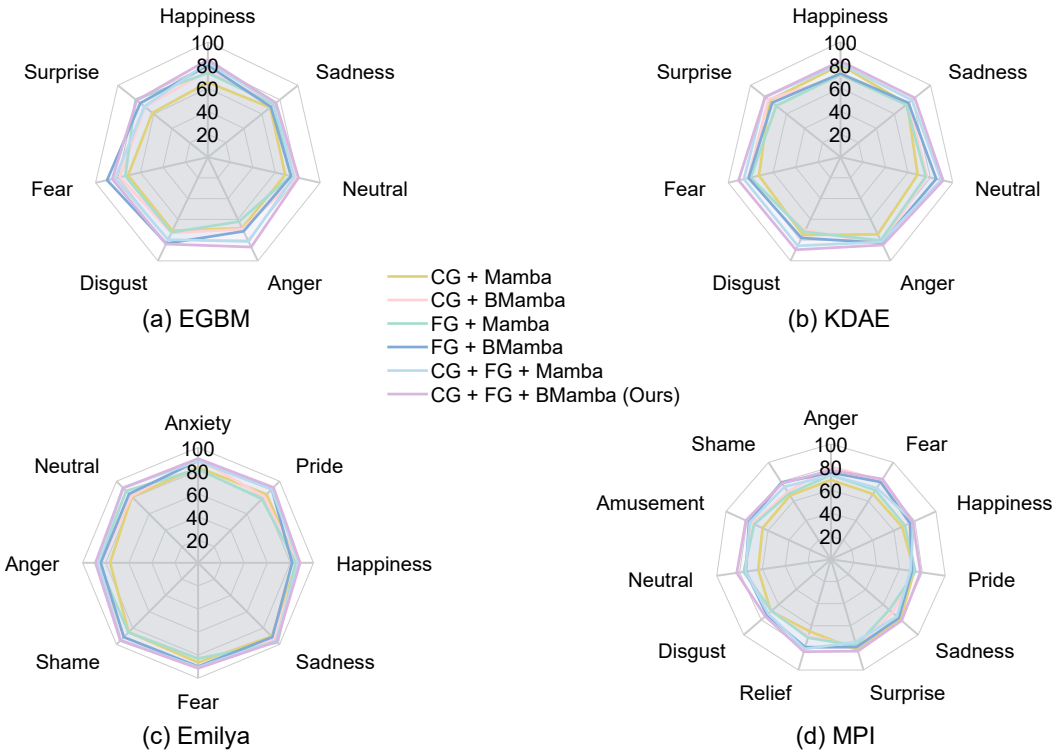

Fig. 8: Comparison of recognition accuracies (%) between different MSCNN and BMamba architectures across various emotion classes on the (a) EGBM dataset, (b) KDAE dataset, (c) Emilya dataset, and (d) MPI dataset.

accuracy and achieves superior performance in generating physically grounded affective descriptions compared to general-purpose models like GPT-4o and Gemini 1.5 Pro.

Although EABER offers explainable affective reasoning from 3D body expressions, future work will focus on improving practicality and scalability. We will reduce the cost of LLM fine-tuning and deployment via knowledge distillation and lightweight parameter-efficient adaptation, and further mitigate the reliance on manually annotated skeleton–text pairs through self-supervised alignment. In the future, we also intend to extend EABER into an explainable affect–disease reasoning framework for clinical healthcare. By leveraging posture-derived behavioral biomarkers, future work will explore the application of our framework to clinically relevant scenarios, including mental health assessments (e.g., depression and bipolar disorder) and neurodegenerative disease monitoring (e.g., Parkinson's disease). In addition, we plan to incorporate complementary clinical modalities, such as questionnaire-based assessments and electronic medical records, to develop a more robust multimodal framework for personalized early detection and longitudinal evaluation.

REFERENCES

[1] X. Xu, J. Chen, C. Fu, and Z. Lyu, "Hypercomplex neural network and cross-modal attention for multi-modal emotion recognition using physiological signals," *IEEE Transactions on Affective Computing*, 2025.

[2] S. K. D'Mello, N. Dowell, and A. Graesser, "Unimodal and multimodal human perceptionof naturalistic non-basic affective statesduring human-computer interactions," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 452–465, 2013.

[3] E. Cambria, R. Mao, M. Chen, Z. Wang, and S.-B. Ho, "Seven pillars for the future of artificial intelligence," *IEEE Intelligent Systems*, vol. 38, no. 6, pp. 62–69, 2023.

[4] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE transactions on affective computing*, vol. 14, no. 3, pp. 1743–1753, 2022.

[5] M. Du, S. Liu, T. Wang, W. Zhang, Y. Ke, L. Chen, and D. Ming, "Depression recognition using a proposed speech chain model fusing speech production and perception features," *Journal of Affective Disorders*, vol. 323, pp. 299–308, 2023.

[6] R. Mao, T. Wang, and E. Cambria, "Decoding metaphors and brain signals in naturalistic contexts: An empirical study based on eeg and metapro," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 47, 2025.

[7] X. Xu, C. Fu, and J. Chen, "Emotion recognition empowered human-computer interaction with domain adaptation network," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 2, pp. 6777–6786, 2024.

[8] T. Wang, J. Sun, J. Chao, S. Zheng, C. Zhao, C. Wu, and H. Peng, "A novel gait analysis method based on the pseudo-velocity model for depression detection," in *2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM)*. IEEE, 2021, pp. 1–6.

[9] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, vol. 12, no. 2, pp. 505–523, 2018.

[10] T. Wang, S. Liu, F. He, M. Du, W. Dai, Y. Ke, and D. Ming, "Affective body expression recognition framework based on temporal and spatial fusion features," *Knowledge-Based Systems*, vol. 308, p. 112744, 2025.

[11] J. Fang, T. Wang, C. Li, X. Hu, E. Ngai, B.-C. Seet, J. Cheng, Y. Guo, and X. Jiang, "Depression prevalence in postgraduate students and its association with gait abnormality," *IEEE Access*, vol. 7, pp. 174 425–174 437, 2019.

[12] Y. Luo, J. Ye, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang, "Arbee: Towards automated recognition of bodily expression of

emotion in the wild," *International journal of computer vision*, vol. 128, no. 1, pp. 1–25, 2020.

[13] T. Wang, C. Li, C. Wu, C. Zhao, J. Sun, H. Peng, X. Hu, and B. Hu, "A gait assessment framework for depression detection using kinect sensors," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3260–3270, 2020.

[14] H. Shou, L. Cui, I. Hickie, D. Lameira, F. Lamers, J. Zhang, C. Crainiceanu, V. Zipunnikov, and K. Merikangas, "Dysregulation of objectively assessed 24-hour motor activity patterns as a potential marker for bipolar i disorder: results of a community-based family study," *Translational psychiatry*, vol. 7, no. 8, pp. e1211–e1211, 2017.

[15] Y. Zhai, G. Jia, Y.-K. Lai, J. Zhang, J. Yang, and D. Tao, "Looking into gait for perceiving emotions via bilateral posture and movement graph convolutional networks," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 1634–1648, 2024.

[16] M.-A. Mahfoudi, A. Meyer, T. Gaudin, A. Buendia, and S. Bouakaz, "Emotion expression in human body posture and movement: A survey on intelligible motion factors, quantification and validation," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2697–2721, 2022.

[17] H. Lu, X. Hu, and B. Hu, "See your emotion from gait using unlabeled skeleton data," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 1826–1834, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/25272

[18] H. Lu, S. Xu, S. Zhao, X. Hu, R. Ma, and B. Hu, "Epic: Emotion perception by spatio-temporal interaction context of gait," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 5, pp. 2592–2601, 2024.

[19] C. Beyan, S. Karumuri, G. Volpe, A. Camurri, and R. Niewiadomski, "Modeling multiple temporal scales of full-body movements for emotion classification," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1070–1081, 2021.

[20] T. Wang, S. Liu, F. He, W. Dai, M. Du, Y. Ke, and D. Ming, "Emotion recognition from full-body motion using multiscale spatio-temporal network," *IEEE Transactions on Affective Computing*, 2023.

[21] A. Camurri, G. Volpe, S. Piana, M. Mancini, R. Niewiadomski, N. Ferrari, and C. Canepa, "The dancer in the eye: towards a multi-layered computational framework of qualities in movement," in *Proceedings of the 3rd International Symposium on Movement and Computing*, 2016, pp. 1–7.

[22] A. Jawaharlalnehru, T. Sambandham, and D. Ravikumar, "Recognizing human emotions through body posture dynamics using deep neural networks," *Engineering Proceedings*, vol. 87, no. 1, p. 49, 2025.

[23] T. Wang, R. Mao, S. Liu, E. Cambria, and D. Ming, "Explainable multi-frequency and multi-region fusion model for affective brain-computer interfaces," *Information Fusion*, vol. 118, p. 102971, 2025.

[24] H. Liu, M. Cheng, X. Wei, F. Dollack, V. Schneider, H. Uchiyama, Y. Kitamura, K. Kiyokawa, and M. Perusquia-Hernandez, "Large language models as perceivers of dynamic full-body expressions of emotion," in *Companion Proceedings of the 27th International Conference on Multimodal Interaction*, 2025, pp. 7–11.

[25] B. Ren, M. Liu, R. Ding, and H. Liu, "A survey on 3d skeleton-based action recognition using learning method," *Cyborg and Bionic Systems*, vol. 5, p. 0100, 2024.

[26] H. Lu, J. Chen, F. Liang, M. Tan, R. Zeng, and X. Hu, "Understanding emotional body expressions via large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 2, 2025, pp. 1447–1455.

[27] H. Lu, J. Chen, Z. Zhang, R. Liu, R. Zeng, and X. Hu, "Emotion recognition from skeleton data: A comprehensive survey," *arXiv preprint arXiv:2507.18026*, 2025.

[28] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for automatic emotion recognition by body gesture analysis," in *2008 IEEE Computer society conference on computer vision and pattern recognition workshops*. IEEE, 2008, pp. 1–6.

[29] N. Fourati, C. Pelachaud, and P. Darmon, "Contribution of temporal and multi-level body cues to emotion classification," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 116–122.

[30] A. Shirian, S. Tripathi, and T. Guha, "Dynamic emotion modeling with learnable graphs and graph inception network," *IEEE Transactions on Multimedia*, vol. 24, pp. 780–790, 2021.

[31] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, and D. Manocha, "Step: Spatial temporal graph convolutional networks for emotion perception from gaits," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 02, 2020, pp. 1342–1350.

[32] T. Sapiński, D. Kamińska, A. Pelikant, and G. Anbarjafari, "Emotion recognition from skeletal movements," *Entropy*, vol. 21, no. 7, p. 646, 2019.

[33] H. Zhang, P. Yi, R. Liu, and D. Zhou, "Emotion recognition from body movements with as-lstm," in *2021 IEEE 7th International Conference on Virtual Reality (ICVR)*. IEEE, 2021, pp. 26–32.

[34] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Computer Vision and Image Understanding*, vol. 208, p. 103219, 2021.

[35] P. V. Paiva, J. J. Ramos, M. L. Gavrilova, and M. A. Carvalho, "Emotion transformer: Attention model for pose-based emotion recognition." in *VISIGRAPP (5: VISAPP)*, 2023, pp. 274–281.

[36] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[37] Z. Zhang, A. Liu, I. Reid, R. Hartley, B. Zhuang, and H. Tang, "Motion mamba: Efficient and long sequence motion generation," in *European Conference on Computer Vision*. Springer, 2024, pp. 265–282.

[38] M. M. Amin, R. Mao, E. Cambria, and B. W. Schuller, "A wide evaluation of chatgpt on affective computing tasks," *IEEE Transactions on Affective Computing*, vol. 15, no. 4, pp. 2204–2212, 2024.

[39] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[40] E. Cambria, R. Mao, A. Hussain, K. Oatley, and G. Hinton, "Artificial intelligence as the fourth decentering revolution: From cosmic, biological, and psychological displacement to cognitive decentering," *Cognitive Computation*, vol. 18, no. 20, pp. 1–13, 2026.

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[42] H. Zhang, M. C. Leong, L. Li, and W. Lin, "Pevl: Pose-enhanced vision-language model for fine-grained human action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 857–18 867.

[43] Y. Chen, T. He, J. Fu, L. Wang, J. Guo, T. Hu, and H. Cheng, "Vision-language meets the skeleton: Progressively distillation with cross-modal knowledge for 3d action representation learning," *IEEE Transactions on Multimedia*, 2024.

[44] H. Qu, Y. Cai, and J. Liu, "Llms are good action recognizers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 395–18 406.

[45] M. Wang, J. Xing, J. Mei, Y. Liu, and Y. Jiang, "Actionclip: Adapting language-image pretrained models for video action recognition," *IEEE transactions on neural networks and learning systems*, 2023.

[46] W. Xiang, C. Li, Y. Zhou, B. Wang, and L. Zhang, "Generative action description prompts for skeleton-based action recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 10 276–10 285.

[47] T. Yan, W. Zeng, Y. Xiao, X. Tong, B. Tan, Z. Fang, Z. Cao, and J. T. Zhou, "Crossglg: Llm guides one-shot skeleton-based 3d action recognition in a cross-level manner," in *European Conference on Computer Vision*. Springer, 2024, pp. 113–131.

[48] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv:2106.09685*, 2021.

[49] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, vol. 36, 2023, pp. 34 892–34 916. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf

[50] T. Sapiński, D. Kamińska, A. Pelikant, C. Ozcinar, E. Avots, and G. Anbarjafari, "Multimodal database of emotional speech, video and gestures," in *International Conference on Pattern Recognition*. Springer, 2018, pp. 153–163.

[51] M. Zhang, L. Yu, K. Zhang, B. Du, B. Zhan, S. Chen, X. Jiang, S. Guo, J. Zhao, Y. Wang *et al.*, "Kinematic dataset of actors expressing emotions," *Scientific data*, vol. 7, no. 1, p. 292, 2020.

[52] N. Fourati and C. Pelachaud, "Perception of emotions and body movement in the emilya database," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 90–101, 2016.

[53] E. Volkova, S. De La Rosa, H. H. Bülthoff, and B. Mohler, "The mpi emotional body expressions database for narrative scenarios," *PloS one*, vol. 9, no. 12, p. e113647, 2014.
[54] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 026–12 035.
[55] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 359–13 368.
[56] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.
[57] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Adasgn: Adapting joint number and model size for efficient skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 413–13 422.
[58] J. Zhang, G. Ye, Z. Tu, Y. Qin, Q. Qin, J. Zhang, and J. Liu, "A spatial attentive and temporal dilated (satd) gcn for skeleton-based action recognition," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 1, pp. 46–55, 2022.
[59] J. Lee, M. Lee, D. Lee, and S. Lee, "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 10 444–10 453.
[60] D. J. Lerch, Z. Zhong, M. Martin, M. Voit, and J. Beyerer, "Unsupervised 3d skeleton-based action recognition using cross-attention with conditioned generation capabilities," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 211–220.

Tao Wang received the Ph.D. degree from Tianjin University, Tianjin, China, and conducted joint doctoral training at Nanyang Technological University, Singapore. He is currently an assistant researcher at the Haihe Laboratory of Brain-Computer Interaction and Human-Machine Integration. His research interests include brain–computer interfaces, affective computing, and large language models (LLM). Contact him at taowang2021@tju.edu.cn.

Haifeng Lu received the Ph.D. degree from Lanzhou University, Lanzhou, China. He is currently a Post-Doctoral Fellow with The University of Hong Kong, Hong Kong, and a Visiting Scholar with Shenzhen MSU-BIT University, Shenzhen, China. He has published several papers in prestigious journals and conferences, such as IEEE TAFFC, IEEE JBHI, and AAAI. His current research interests include affective computing and large language model (LLM). Contact him at luhfhku@hku.hk.

Jiayi Duan received the B.S. degree in the School of Mathematics from Tianjin University, Tianjin, China, in 2021. She is currently pursuing the M.S. degree in Intelligent Medical Engineering in the School of Medicine from Tianjin University, Tianjin, China. Her research interests include affective computing in brain-computer interfaces. Contact her at 2025246046@tju.edu.cn.

Tianyu Meng received the M.S. degree in computer technology from Changchun University of Science and Technology, Changchun, China, in 2025. He is currently pursuing the Ph.D. degree with the Academy of Medical Engineering and Translational Medicine from Tianjin University, Tianjin, China. His research interests focus on affective computing, deep learning, digital twin brain modeling, and cognitive behavior analysis. Contact him at tianyumeng@tju.edu.cn.

Rui Mao (Member, IEEE) is a Research Scientist and Lead Investigator at Nanyang Technological University. He obtained his Ph.D. degree in Computing Science from the University of Aberdeen. His research interest lies at NLP, affective computing, and their applications in finance and cognitive science. He and his funded company (Ruimao Tech) have developed an end-to-end system (MetaPro) for computational metaphor processing. Contact him at rui.mao@ntu.edu.sg.

Shuang Liu received the Ph.D. degree in biomedical engineering from Tianjin University, Tianjin China. She is a chair professor of biomedical engineering at Tianjin University, Tianjin 300072, China. Her research interests include physiological mechanism of emotion, emotion recognition and regulation, and biomarker detection of the depression. Contact her at shuangliu@tju.edu.cn.

Dong Ming received the B. S. and Ph.D. degrees in biomedical engineering with Tianjin University, Tianjin, China, in 1999 and 2004, respectively. During 2005–2006, he was a Visiting Scholar with the Division of Mechanical Engineering and Mechatronics, University of Dundee, Dundee, U.K. In 2006, he joined Tianjin University (TJU) Faculty, College of Precision Instruments and Optoelectronics Engineering and since 2011, has been a Full Professor of biomedical engineering. He is currently the Dean of the Academy of Medical Engineering and Translational Medicine of TJU, the Head of the Neural Engineering and Rehabilitation Laboratory, TJU, and the Chair of IEEE EMBS Tianjin Chapter. His main research interests include neural engineering, rehabilitation engineering, sports science, biomedical instrumentation and signal/image processing, especially in functional electrical stimulation, gait analysis, and brain–computer interface. Contact him at richard-ming@tju.edu.cn.