



Design and Performance Evaluation of an English Oral Teaching Feedback System Integrating Visual Emotion Recognition and Pose Estimation

Xiaoli Li 

Faculty of Foreign Languages, North Henan Medical University, Xinxiang 453003, China

Corresponding Author Email: xllsq121207@163.com

Copyright: ©2026 The author. This article is published by IIETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.430410>

ABSTRACT

Received: 26 March 2026

Revised: 3 July 2026

Accepted: 18 July 2026

Available online: 31 August 2026

Keywords:

visual emotion recognition, multimodal fusion, pose estimation, attention mechanism, explainable artificial intelligence, intelligent education

Monitoring learners' emotional states in English oral teaching is of great significance for personalized instructional intervention. However, existing visual emotion recognition methods suffer from notable limitations in three aspects: cross-modal spatial alignment, dynamic confidence awareness, and interpretable teaching feedback generation. This paper proposes an oral teaching feedback system that integrates visual emotion recognition and pose estimation, addressing the aforementioned issues from three perspectives: feature interaction, fusion strategy, and system output. At the feature interaction level, a pose-guided facial visual attention mechanism is designed. It utilizes skeletal keypoint coordinates to generate spatial attention masks, dynamically guiding the facial feature extraction network to focus on facial regions that are semantically relevant to body context, thereby achieving fine-grained alignment of cross-modal information in the image feature space. At the fusion strategy level, a dynamic confidence-weighted cross-modal adaptive fusion strategy is proposed. It evaluates the real-time confidence of each modality from two dimensions—data quality and feature discriminability—and implements data-driven dynamic weight allocation through a gating mechanism, enabling the system to maintain robust performance under disturbances such as illumination changes, head rotations, and occlusions. At the system output level, a spatiotemporal joint interpretability feedback generation mechanism is proposed. By combining Gradient-weighted Class Activation Mapping (Grad-CAM) spatial attribution with temporal analysis of emotion confidence, and mapping them through semantic templates, the system generates visually diagnostic reports with instructional guidance. Experiments on a self-built oral teaching emotion dataset show that the proposed system achieves an emotion recognition accuracy of 86.4% and a feedback usability score of 4.6. Ablation studies verify the effective contribution of each module, and robustness experiments demonstrate that the system maintains a high-performance retention rate under extreme conditions such as modality absence. This paper provides a solution that combines theoretical innovation with practical deployment value for the in-depth application of computer vision technology in educational scenarios.

1. INTRODUCTION

Monitoring learners' emotional states in English oral teaching is of great significance for personalized instructional intervention [1]. Learners' emotional states during oral expression, such as confidence, nervousness, focus, or anxiety, directly affect the quality of their language output and learning outcomes [2]. Traditional emotion assessment relies on teachers' subjective observation or post-hoc questionnaires, making it difficult to achieve real-time, objective, and large-scale dynamic monitoring [3]. With the rapid development of computer vision technology, automated emotion recognition based on visual cues provides a feasible technical path for this need [4]. Facial expressions provide the richest fine-grained cues about emotional states [5], while body pose reflects learners' overall emotional state and engagement level [6]. Combining the two organically is expected to break through the bottlenecks in accuracy and robustness of single-modal

emotion recognition [7]. However, applying visual emotion recognition technology to oral teaching scenarios still faces several key problems that need to be urgently addressed [8].

Existing facial-pose emotion recognition methods generally lack fine-grained spatial alignment at the feature interaction level [9]. Most methods extract features separately in two independent branches and then directly perform feature concatenation or simple attention fusion [10]. The facial expression feature extraction process lacks spatial guidance from the pose modality, causing the network to be unable to perceive the current body context. When learners are in different poses, the same facial micro-expression may carry completely different emotional semantics; facial feature extraction without pose context is highly prone to semantic ambiguity [11]. Although some studies have explored cross-modal attention mechanisms, attention interaction usually occurs at the global feature level and fails to sink the structured spatial information of pose into the spatial attention level of

facial feature maps. The insufficient precision of cross-modal information alignment directly limits the discriminability of fused features [12]. At the fusion strategy level, existing methods mostly adopt fixed-weight feature concatenation or static attention mechanisms, ignoring the reliability differences of each modality at different moments [13]. In real oral teaching scenarios, factors such as illumination changes, head rotations, and body occlusions cause dynamic fluctuations in data quality of facial expression and pose modalities [14]. When a learner lowers their head to read notes, facial expression information is greatly attenuated, and pose information should occupy a higher weight at this time [15]; whereas when the learner faces the camera directly, the discriminability of facial expressions is far stronger than that of pose [16]. Existing methods lack the ability to perceive real-time modality confidence and perform adaptive fusion; fixed-weight fusion strategies struggle to maintain optimal performance in dynamic scenarios [17]. In addition, there is a significant disconnect between the output form of emotion recognition systems and the actual needs of teaching scenarios [18]. Most systems only output emotion classification labels and cannot explain which visual cues the model's judgment is based on, let alone translate the judgment results into actionable feedback with instructional guidance. Although some studies have attempted to introduce explainable AI techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) into emotion recognition [19], the visualization results usually remain at the technical demonstration level, lacking semantic transformation oriented to teaching scenarios and temporal context association, and thus fail to truly serve teaching decisions [20].

To address the above three shortcomings, this paper proposes an English oral teaching feedback system that integrates visual emotion recognition and pose estimation. At the feature interaction level, a pose-guided facial visual attention mechanism is proposed, which utilizes skeletal keypoint coordinates obtained from pose estimation to generate spatial attention masks, dynamically guiding the facial feature extraction network to focus on facial regions that are semantically relevant to the current body context, thereby achieving fine-grained alignment and enhancement of cross-modal information in the image feature space. At the fusion strategy level, a dynamic confidence-weighted cross-modal adaptive fusion strategy is proposed, which estimates the real-time confidence of each modality from two dimensions: data quality and feature discriminability, and implements data-adaptive dynamic fusion weight allocation through a gating mechanism, improving the system's robustness in complex real-world scenarios. At the system output level, a spatiotemporal joint interpretability feedback generation mechanism oriented to oral teaching is proposed, which combines Grad-CAM spatial attribution with temporal analysis of emotion confidence to generate visually diagnostic reports with instructional guidance, upgrading the system output from classification labels to interpretable teaching feedback. The three modules are organically connected with feature alignment, adaptive fusion, and interpretable feedback as the technical route, forming an end-to-end emotion recognition and feedback system for oral teaching scenarios.

The remainder of this paper is organized as follows: Section 2 introduces the overall framework and technical methods of the proposed system in detail; Section 3 describes the experimental setup, dataset construction, comparative experiments, and ablation study results; Section 4 discusses

the limitations of the method and future directions; Section 5 concludes the paper.

2. PROPOSED METHOD

2.1 Problem formulation and overall system framework

This paper aims to address the problem of visual emotion recognition and teaching feedback generation for learners in English oral teaching scenarios. Let the input video sequence be $V = \{I_t \in \mathbb{R}^{H \times W \times 3}\}_{t=1}^T$, where T is the total number of frames, H and W are the frame height and width, respectively. The system needs to synchronously extract facial expression features and body pose features from each frame I_t , and after multimodal fusion, output the emotion category probability distribution $y_t \in \mathbb{R}^C$, where $C = 5$ corresponds to five emotion states: confidence, nervousness, focus, fatigue, and neutral. Additionally, structured teaching feedback is generated based on the spatiotemporal information of the entire sequence. The core challenges of this task lie in three aspects: spatial alignment of cross-modal features, adaptive perception of modality reliability in dynamic scenarios, and effective transformation of recognition results into teaching semantics. These three challenges correspond to the design motivations of the three core modules of the system, respectively.

The overall system framework consists of three sequentially cascaded modules. Module 1 is the pose-guided facial visual attention mechanism, which takes the skeletal keypoint coordinates obtained from pose estimation as spatial priors to generate attention masks and applies them to the facial feature maps, enabling the facial feature extraction network to perceive the current body context and output pose-enhanced facial emotion features $f_{expr} \in \mathbb{R}^d$, while the pose encoder outputs pose emotion features $f_{pose} \in \mathbb{R}^d$, and in this paper, $d = 256$. Module 2 is the dynamic confidence-weighted cross-modal adaptive fusion strategy, which estimates the real-time confidence c_{expr} and c_{pose} of the two modalities from the two dimensions of data quality and feature discriminability, respectively, and dynamically computes fusion weights through a gating mechanism to output the fused feature $f_{fusion} \in \mathbb{R}^d$. Module 3 is the spatiotemporal joint interpretability feedback generation mechanism oriented to oral teaching, which, based on the fused feature sequence, performs joint modeling of Grad-CAM spatial attribution and temporal analysis of emotion confidence, and generates structured teaching feedback F via semantic template mapping. The three modules are organically connected with feature alignment, adaptive fusion, and interpretable feedback as the technical main line. All modules adopt an end-to-end joint training strategy. The overall loss function L_{total} is composed of a weighted combination of emotion classification loss and feedback consistency loss; see Section 2.6 for details.

2.2 Visual feature extraction module

The facial feature extraction branch starts from the input frame I_t . First, an improved RetinaFace detector is used to extract the facial region, obtaining a facial image $I_{face}^{(t)} \in \mathbb{R}^{224 \times 224 \times 3}$ cropped and resized to 224×224 . The backbone network for facial feature extraction is ResNet-50 pre-trained on AffectNet, and the feature map output from its

fourth residual stage is $F_{face}^{(l)} \in \mathbb{R}^{2048 \times 7 \times 7}$. Although single-scale deep features possess strong semantic abstraction capabilities, their spatial resolution is low, making it difficult to preserve the textural details of subtle facial regions. To this end, this paper introduces a multi-scale feature aggregation strategy: the $1024 \times 14 \times 14$ feature map output from the third residual stage is concatenated with the fourth-stage feature via a skip connection, and then reduced in dimension by a 1×1 convolution, ultimately obtaining the multi-scale facial feature map $F_{face}^{(l)} \in \mathbb{R}^{C_f \times H_f \times W_f}$, where $C_f = 256$, $H_f = 7$, and $W_f = 7$.

This multi-scale feature map provides a feature basis with more complete spatial structure preservation for the subsequent pose-guided attention weighting. Synchronously, 468 facial landmark coordinates $L_{face} = \{(u_l, v_l, r_l)\}_{l=1}^{468}$ are extracted, where r_l is the visibility confidence; these landmarks will be used for the data quality assessment in modality confidence estimation in Section 2.3. Figure 1 illustrates the network architecture of the dual-branch visual feature extraction module.

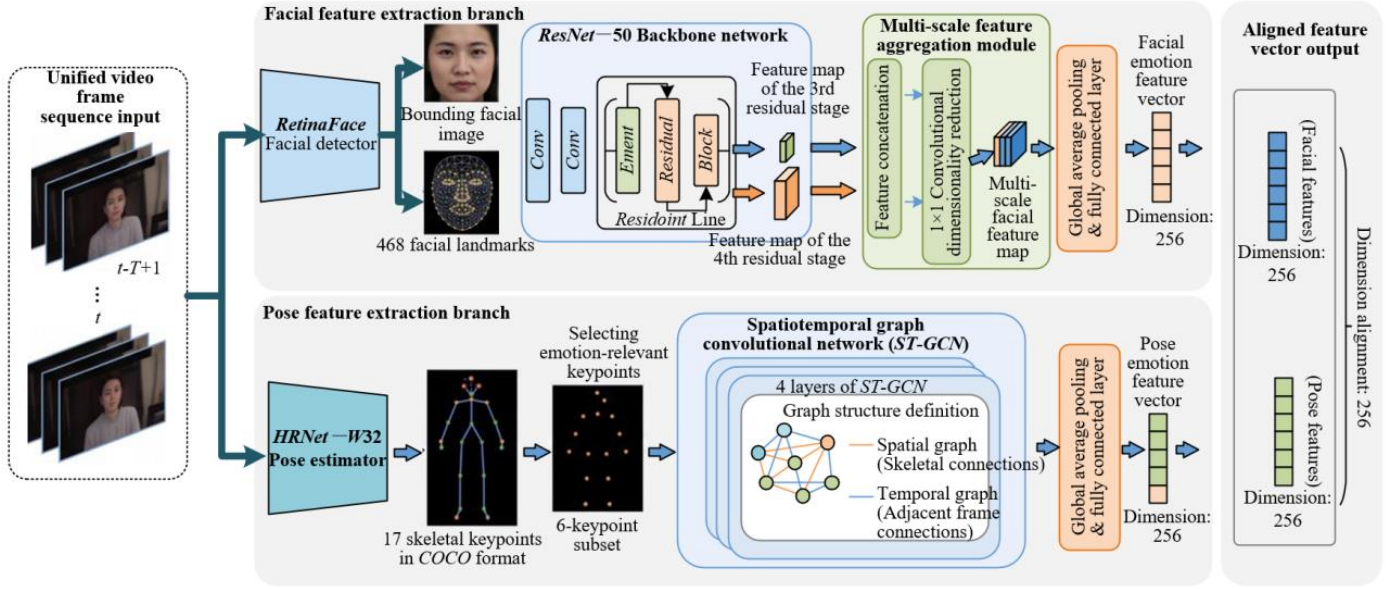


Figure 1. Network architecture diagram of the dual-branch visual feature extraction module

The pose feature extraction branch adopts the HRNet-W32 architecture, with the original frame I_t as input, and outputs 17 skeletal keypoint coordinates in COCO format $P_t = \{p_k = (x_k, y_k, s_k)\}_{k=1}^{17}$, where $s_k \in [0, 1]$ is the detection confidence. Not all 17 keypoints contribute equally to emotion expression, so this paper selects $K' = 6$ keypoints that are highly correlated with emotion expression to form the pose subset $P_{pose} = \{p_{head}, p_{neck}, p_{l-shoulder}, p_{r-shoulder}, p_{l-hip}, p_{r-hip}\}$, aiming to remove redundant nodes weakly associated with emotional semantics, such as limb extremities, while reducing the computational overhead of subsequent spatiotemporal graph encoding. To capture the spatial structural features and temporal dynamic features of the pose, a Spatiotemporal Graph Convolutional Network (ST-GCN) is employed to encode the pose sequence. The spatial graph structure is defined as $G = (V, E)$, where the node set V contains K' keypoints, and the edge set E is defined based on natural human skeletal connections. The graph convolution operation of the l -th layer of ST-GCN is:

$$H^{(l+1)} = \sigma \left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)} \right) \quad (1)$$

where, $\tilde{A} = A + I_{K'}$ is the adjacency matrix with added self-loops, $I_{K'}$ is the $K' \times K'$ identity matrix, the introduction of self-loops allows the graph convolution to retain the node's original features while aggregating neighborhood information; $\tilde{D}$ is the degree matrix of $\tilde{A}$, used for symmetric normalization of the adjacency matrix to prevent nodes with large degrees from

dominating the feature update during aggregation; $H^{(l)} \in \mathbb{R}^{K' \times d_l}$ is the node feature matrix of the l -th layer, and d_l is the feature dimension of the current layer; $W^{(l)} \in \mathbb{R}^{d_l \times d_{l+1}}$ is the learnable weight matrix, and σ is the ReLU activation function. In the temporal dimension, a $1 \times \tau$ temporal convolution kernel ($\tau = 5$) is used to capture the pose evolution patterns between adjacent frames, enabling the network to perceive the dynamic change trend of pose within a short-time window.

After 4 layers of ST-GCN encoding, global average pooling is performed on the node dimension to aggregate the features of each node into a fixed-length global pose representation, resulting in the pose emotion feature vector $f_{pose}^{(t)} \in \mathbb{R}^{256}$. The facial feature branch is mapped via global average pooling and a fully connected layer from the multi-scale feature map $f_{face}^{(t)}$, similarly outputting a facial emotion feature $F_{expr}^{(t)}$ with a dimension of 256. The output dimensions of the two branches are kept consistent, providing a prerequisite for feature space alignment for subsequent cross-modal attention guidance and dynamic fusion. The alignment of facial and pose features in dimension is not merely an engineering choice, but the foundation for ensuring that the pose-guided attention mask can directly act on the spatial facial feature map and that the gating fusion mechanism can compare the contributions of the two modalities in a unified feature space.

2.3 Pose-guided facial visual attention mechanism

The quality of facial expression feature extraction directly depends on whether the spatial regions attended to by the network on the feature map are semantically relevant to

emotions. Traditional facial feature extraction networks learn attention distributions in a data-driven manner, lacking explicit spatial guidance from body context. When the facial image itself suffers from uneven illumination or partial occlusion, it is difficult for the network to autonomously determine which regions are more critical for emotion judgment in the current pose context. This paper designs a pose-guided facial visual attention mechanism that transforms

the structured spatial information obtained from pose estimation into an attention prior on the facial feature map, enabling the facial feature extraction process to perceive the current body pose context. This module comprises three steps: spatial coordinate mapping, heatmap generation, and attention weighting. Figure 2 illustrates the principle of the pose-guided facial visual attention mechanism.

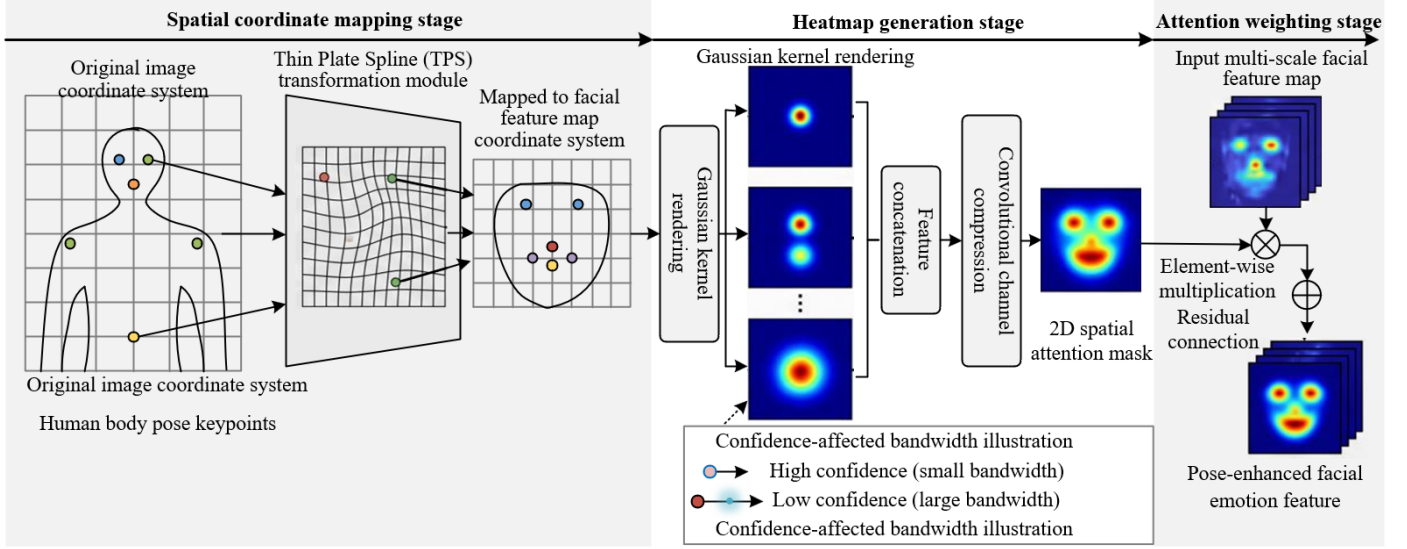


Figure 2. Schematic diagram of the pose-guided facial visual attention mechanism

Spatial coordinate mapping addresses the cross-coordinate system alignment problem between pose keypoints and the facial feature map. The pose keypoints $P_{pose}^{(i)}$ are mapped from the original image coordinate system to the facial feature map coordinate system. Due to the affine deformation of the facial region in the image caused by head pose changes, linear transformations struggle to accurately describe this non-rigid mapping relationship. Therefore, this paper employs Thin Plate Spline (TPS) transformation for nonlinear mapping. The TPS transformation parameters Θ_{TPS} are solved jointly using the head-related points p_{head} and p_{neck} from the pose keypoints and the four corner points of the facial detection box:

$$\Theta_{TPS} = \arg \min_{\Theta} \sum_{q \in Q} \|q - TPS(q; \Theta)\|_2^2 + \lambda \cdot \text{bending energy} \quad (2)$$

where, Q is the control point set, formed by the head pose keypoints and the corner points of the facial detection box; λ is the regularization coefficient used to control the smoothness of the transformation to prevent overfitting. After TPS mapping, the corresponding coordinate of each pose keypoint p_k in the facial feature map space is $(\tilde{x}_k, \tilde{y}_k) = TPS(p_k; \Theta_{TPS}^*)$. This mapping unifies the spatial structure information of the pose from the original image coordinate system into the facial feature map space, providing a prerequisite for spatial alignment for the subsequent attention mask generation.

A 2D Gaussian heatmap is generated for each mapped keypoint, constructing a spatial attention distribution centered at the keypoint location:

$$G_k(i, j) = \exp\left(-\frac{(i - \tilde{x}_k)^2 + (j - \tilde{y}_k)^2}{2\sigma_k^2}\right), \quad i \in [1, H_f], j \in [1, W_f] \quad (3)$$

where, the Gaussian kernel bandwidth σ_k is adaptively adjusted based on the detection confidence of the keypoint s_k : $\sigma_k = \sigma_0 \cdot (1 + \beta \cdot (1 - s_k))$. σ_0 is the base bandwidth, set to 2.5 in this paper, and β is the adjustment coefficient, set to 0.6. The physical rationale behind this design is that keypoints with lower detection confidence exhibit greater spatial position uncertainty; hence, a larger heatmap diffusion radius is assigned to cover a broader potential facial region. The K' heatmaps are concatenated along the channel dimension into $G \in \mathbb{R}^{K' \times H_f \times W_f}$, which is then compressed to 1 channel via a 1×1 convolution, followed by a Sigmoid activation to obtain the attention mask:

$$A_{pose} = \sigma(\text{Conv}_{1 \times 1}(G)), A_{pose} \in [0, 1]^{H_f \times W_f} \quad (4)$$

This mask possesses a clear semantic interpretation in the spatial dimension: high-response regions indicate the facial areas that the feature extraction network should focus on given the current pose context. The mask is multiplied element-wise with the facial feature map, and a residual connection is introduced to preserve the original features, preventing potential deviations in the pose prior from causing the loss of effective information:

$$F_{face}^{guided} = F_{face} \odot \text{UpSample}(A_{pose}) + F_{face} \quad (5)$$

where, $\text{UpSample}()$ uses bilinear interpolation to upsample A_{pose} to the spatial resolution of F_{face} . The introduction of the residual connection is based on the following consideration: while spatial guidance from the pose prior is beneficial in most cases, when pose estimation suffers severe deviations, the attention mask may mislead the facial feature network to focus on incorrect regions. The residual connection provides a data-

driven pathway for the network to bypass this misleading information. F_{face}^{guided} is subsequently mapped via global average pooling and a fully connected layer to obtain the pose-guided enhanced facial emotion feature $f_{expr}^{(t)} \in \mathbb{R}^{256}$. At this point, the facial feature extraction process incorporates structured spatial information from the pose modality, achieving fine-grained alignment of cross-modal information in the image feature space.

2.4 Dynamic confidence-weighted cross-modal adaptive fusion strategy

The Pose-Guided Facial Attention (PGFA) module achieves fine-grained alignment of cross-modal features at the spatial

level, but the relative importance of the two modalities during fusion still relies on a fixed weight allocation strategy. In oral teaching scenarios, factors such as sudden illumination changes, head rotations, and body occlusions alternately affect the data quality of the facial and pose modalities. A fixed-weight fusion strategy struggles to maintain optimal performance under dynamic conditions. This paper designs a dynamic confidence-weighted cross-modal adaptive fusion strategy, which evaluates the real-time confidence of each modality from two dimensions—data quality and feature discriminability—and implements data-driven dynamic weight allocation through a gating mechanism. Figure 3 presents the architecture of the dynamic confidence-weighted cross-modal adaptive fusion module.

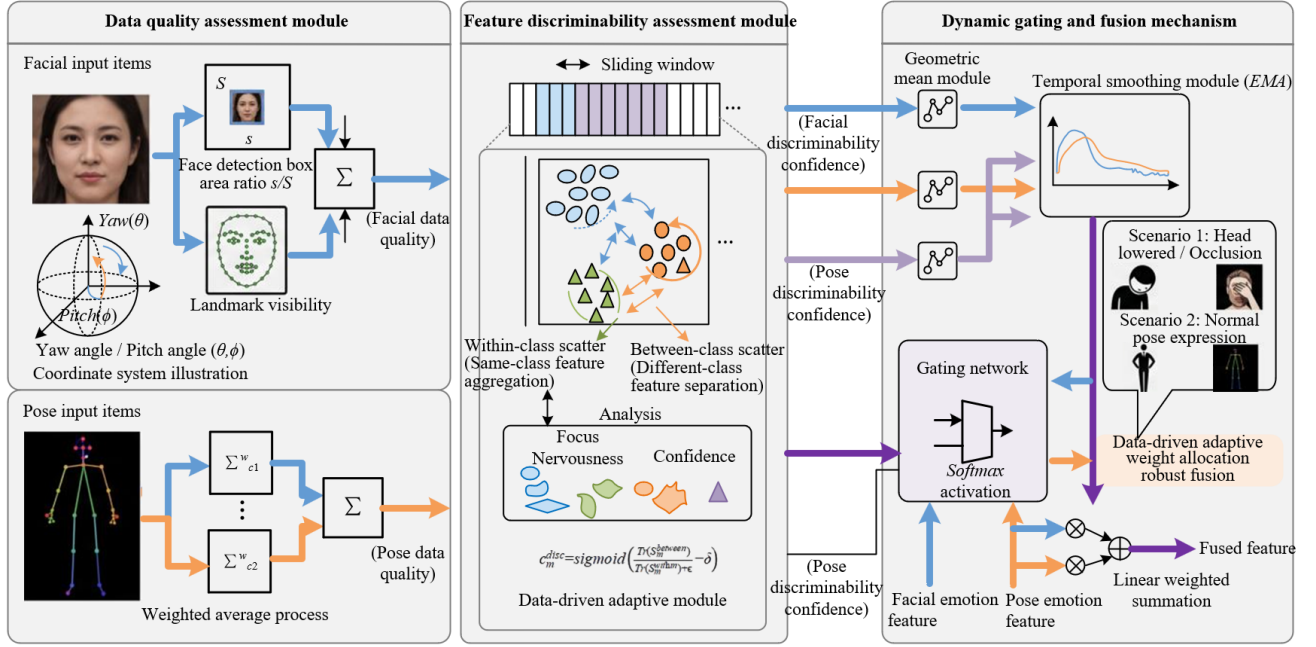


Figure 3. Dynamic confidence-weighted cross-modal adaptive fusion module

Data quality confidence reflects the reliability of the input signal. For the facial modality, data quality is determined jointly by three aspects: facial detection box area, head orientation, and landmark visibility:

$$c_{expr}^{data} = \lambda_1 \cdot \frac{area(face)}{H \cdot W} + \lambda_2 \cdot \cos(\theta_{yaw}) \cdot \cos(\theta_{pitch}) + \lambda_3 \cdot \frac{1}{468} \sum_{l=1}^{468} r_l \quad (6)$$

where, $area(face)$ is the area of the facial detection box; $H \cdot W$ is the total image area, their ratio reflects the proportion of the face in the frame; θ_{yaw} and θ_{pitch} are the head yaw and pitch angles, respectively, solved via the Perspective-n-Point (PnP) algorithm using the 468 facial landmarks. Their cosine product approaches 1 when the head is facing forward and approaches 0 when the head is turned sideways or lowered; r_l is the visibility confidence of the l -th landmark. The weights $\lambda_1, \lambda_2, \lambda_3$ are set to 0.3, 0.4, and 0.3, respectively, highlighting the dominant influence of head orientation on facial expression recognition. For the pose modality, the pose data quality confidence is the weighted average of the detection confidence of the 17 skeletal keypoints:

$$c_{pose}^{data} = \sum_{k=1}^{17} \omega_k \cdot s_k, \sum_{k=1}^{17} \omega_k = 1, \omega_k > 0 \quad (7)$$

where, s_k is the detection confidence of the k -th keypoint, and ω_k is the weight corresponding to each keypoint. Points related to the shoulders and head are assigned higher weights to emphasize their relevance to emotion expression. The data quality confidence only reflects the reliability at the input signal level, but high data quality does not guarantee that the feature is discriminative in the current classification task. Therefore, a supplementary assessment from the feature discriminability dimension is required. The feature discriminability confidence is dynamically updated during the training process based on the ratio of between-class scatter to within-class scatter, calculated using a sliding window approach. Let the feature set of modality m in the current mini-batch be $\{f_m^{(n)}\}_{n=1}^N$, with corresponding class labels $\{y^{(n)}\}_{n=1}^N$. The between-class scatter matrix $S_m^{between}$ and within-class scatter matrix S_m^{within} are defined as:

$$S_m^{between} = \sum_{c=1}^C N_c \cdot (\mu_c - \mu)(\mu_c - \mu)^\top, S_m^{within} = \sum_{c=1}^C \sum_{n: y^{(n)}=c} (f_m^{(n)} - \mu_c)(f_m^{(n)} - \mu_c)^\top \quad (8)$$

where, μ_c is the feature mean of class c , μ is the global feature mean, and N_c is the number of samples in class c . The larger the between-class scatter and the smaller the within-class scatter, the more effective the feature is for the current class discrimination task. The feature discriminability confidence maps the scatter ratio to the $[0,1]$ interval via the Sigmoid function:

$$c_m^{disc} = \text{sigmoid}\left(\frac{\text{Tr}(S_m^{between})}{\text{Tr}(S_m^{within}) + \epsilon} - \delta\right) \quad (9)$$

where $\text{Tr}()$ is the trace of the matrix, $\delta = 0.5$ is the bias threshold, and $\epsilon = 10^{-8}$ is a tiny constant to prevent division by zero.

The comprehensive confidence c_m is the geometric mean of the data quality confidence and the feature discriminability confidence:

$$c_m = \sqrt{c_m^{data} \cdot c_m^{disc}}, m \in \{expr, pose\} \quad (10)$$

The geometric mean is adopted instead of the arithmetic mean because the geometric mean is more sensitive to low values in either dimension. The comprehensive confidence approaches 1 only when both confidence values are high, which aligns with the design expectation that the overall confidence should be significantly suppressed if either data quality or discriminability is lacking. The calculation of fusion weights also needs to consider the temporal evolution trend of confidence to avoid severe inter-frame oscillations. This paper introduces exponential moving average for temporal smoothing, defining the smoothed confidence $\tilde{c}_m^{(t)}$ as:

$$\tilde{c}_m^{(t)} = \alpha \cdot c_m^{(t)} + (1 - \alpha) \cdot \tilde{c}_m^{(t-1)} \quad (11)$$

The smoothing coefficient $\alpha = 0.3$ and the initial value $\tilde{c}_m^{(0)} = 0.5$ are chosen, making the fusion weights evolve smoothly over time. The smoothed confidence is input into the gating network to compute the dynamic fusion weights:

$$g^{(t)} = \text{Softmax}\left(W_g \cdot [\tilde{c}_{expr}^{(t)}, \tilde{c}_{pose}^{(t)}]^\top + b_g\right) \quad (12)$$

where, $W_g \in \mathbb{R}^{2 \times 2}$ and $b_g \in \mathbb{R}^2$ are learnable parameters. The Softmax function ensures $g_{expr}^{(t)} + g_{pose}^{(t)} = 1$ and that each component is non-negative. The fused feature is the weighted sum of the two modal features:

$$f_{fusion}^{(t)} = g_{expr}^{(t)} \cdot f_{expr}^{(t)} + g_{pose}^{(t)} \cdot f_{pose}^{(t)} \quad (13)$$

The advantage of this linear weighted form is that the gradient can be backpropagated end-to-end to the feature extraction networks and the gating network of both modalities. To further enhance the model's robustness when either modality fails, during the training stage, the gating output $g^{(t)}$ is replaced with a uniform distribution $[0.5, 0.5]$ with a probability of 0.1. This random perturbation regularization forces both the pose encoder and facial encoder to possess the ability to perform classification independently even when the other modality is missing, thereby improving the robustness of modality complementarity.

2.5 Spatiotemporal joint interpretability feedback generation mechanism for oral teaching

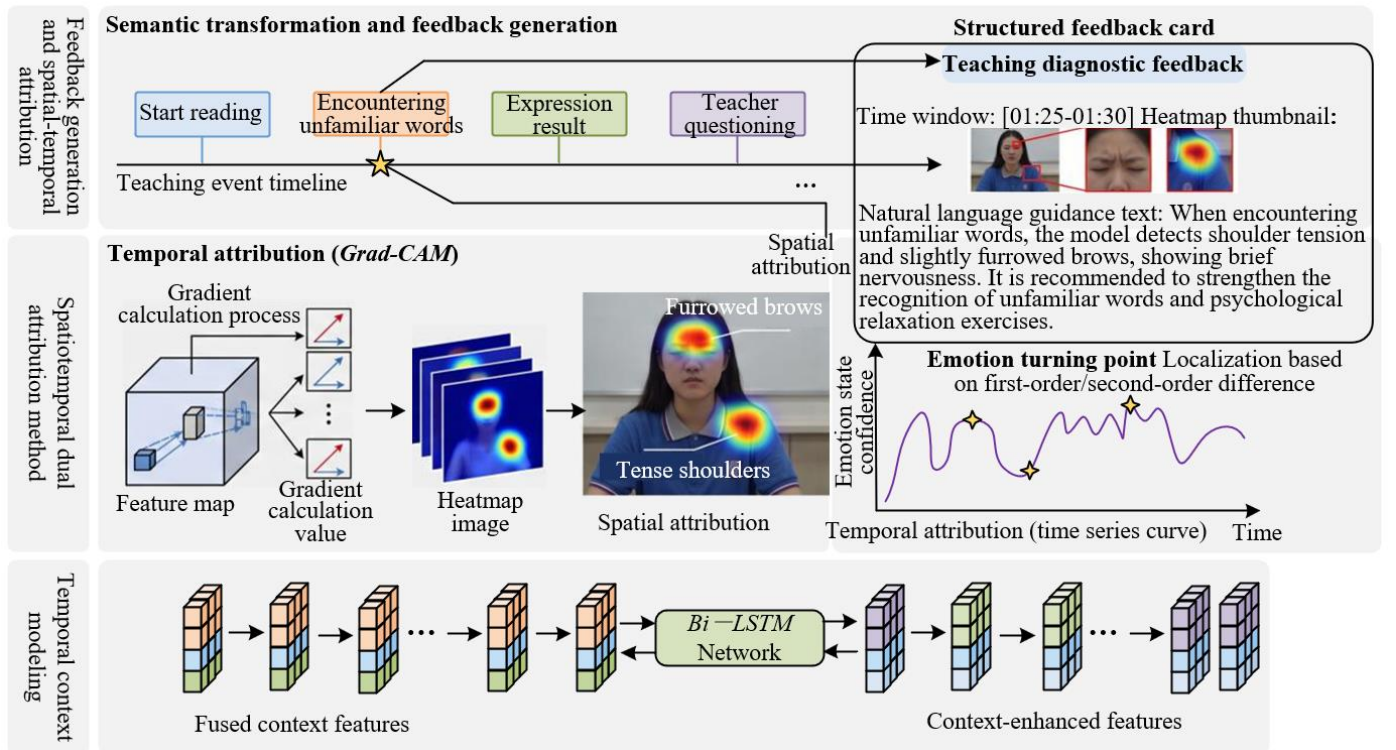


Figure 4. Visualization of spatiotemporal joint interpretability feedback generation for oral teaching
Note: Bi-LSTM: Bidirectional Long Short-Term Memory.

The first two modules complete the extraction and adaptive fusion of emotion features. The fused feature $f_{fusion}^{(t)}$ passes through a two-layer fully connected network and Softmax activation to obtain the emotion probability distribution $y^{(t)} = \text{Softmax}(W_{cls} f_{fusion}^{(t)} + b_{cls})$, and the final emotion category is $\hat{y}^{(t)} = \arg \max_c y^{(t)}_c$. However, single-frame emotion classification results fluctuate temporally, and isolated emotion labels themselves lack teaching guidance value, making it difficult to directly serve teaching decisions. To this end, this paper designs a spatiotemporal joint interpretability feedback generation mechanism, which transforms emotion recognition results into structured feedback with spatiotemporal localization and teaching semantics. Figure 4 presents the visualization principle of the spatiotemporal joint interpretability feedback generation for oral teaching. First, a bidirectional long short-term memory network (BiLSTM) is used to model the temporal dynamics of the fused feature sequence, obtaining context-enhanced features:

$$h_t^{bi} = [\overrightarrow{LSTM}(f_{fusion}^{(t)}); \overleftarrow{LSTM}(f_{fusion}^{(t)})] \quad (14)$$

where, $\overrightarrow{LSTM}$ and $\overleftarrow{LSTM}$ model the feature sequence from the forward and backward directions, respectively. The concatenated vector h_t^{bi} of their outputs integrates past and future contextual information. This design enables the final emotion prediction to be based on both the current frame's fused feature and the preceding and succeeding temporal context, effectively suppressing accidental fluctuations in single-frame predictions. For temporal boundary frames, a forward or backward filling strategy is adopted to complete the context window, ensuring that predictions at any position in the sequence have sufficient temporal information support.

To explain the visual evidence relied upon by the model's decisions, this paper performs spatial attribution mapping on the emotion prediction scores. For the prediction score $y_c^{(t)}$ of the target emotion category c at time t , the gradients with respect to the k -th facial feature map channel $A_{face}^k \in \mathbb{R}^{H_f \times W_f}$ and the k' -th pose graph convolution layer feature map $A_{pose}^{k'}$ are calculated separately. The facial Grad-CAM weight $\alpha_k^{c,t}$ is obtained by globally average pooling the gradients over all spatial positions of the feature map:

$$\alpha_k^{c,t} = \frac{1}{Z_f} \sum_{i=1}^{H_f} \sum_{j=1}^{W_f} \frac{\partial y_c^{(t)}}{\partial (A_{face}^k)_{ij}} \quad (15)$$

where, $Z_f = H_f \times W_f$ is the total number of spatial positions. The facial attribution heatmap is obtained by linearly weighting the channel feature maps according to the weights and applying ReLU activation:

$$L_{face}^{c,t} = \text{ReLU} \left(\sum_k \alpha_k^{c,t} A_{face}^k \right) \quad (16)$$

Similarly, pose attribution is obtained by calculating gradients with respect to the node features of the graph convolution layer, yielding $L_{pose}^{c,t} \in \mathbb{R}^K$, which indicates the skeletal keypoints that contribute most to the current decision. The attribution heatmap answers, in the spatial dimension, which facial regions and which pose keypoints the model focuses on. However, frame-by-frame independent attribution is susceptible to gradient noise interference; heatmaps of adjacent frames may exhibit discontinuous jumps in spatial

distribution, affecting the coherence of subsequent feedback. This paper applies temporal Gaussian filter smoothing to the heatmap sequence, giving the attribution results smooth evolution characteristics in the temporal dimension.

Attribution in the temporal dimension reveals the dynamic evolution pattern of emotional states during oral expression. Define the emotion state confidence sequence as $s_c(t) = y_c^{(t)}$. To locate key moments when the emotion state changes significantly, its first-order difference and second-order difference are calculated:

$$\Delta s_c(t) = s_c(t) - s_c(t-1), \Delta^2 s_c(t) = \Delta s_c(t) - \Delta s_c(t-1) \quad (17)$$

The emotion turning point t^* is located at the moment when the second-order difference takes a local maximum, i.e., the inflection point of the confidence curve, which corresponds to the critical point where the emotion state accelerates upward or downward. Further explanation requires establishing an association between emotion dynamics and teaching events. Teaching event nodes $E = \{e_1, e_2, \dots, e_M\}$ related to oral teaching tasks (including starting reading, encountering unfamiliar words, teacher questioning, ending expression, etc.) are aligned with the video timeline. For each turning point t^* , calculate its temporal distance from the nearest event node $\tau = \min_{e \in E} |t^* - t_e|$, where $\tau < \Delta T$ and $\Delta T = 2$. If and seconds, then this turning point is attributed to the corresponding event-driven cause, giving the emotion change a semantic explanation at the teaching level.

The attribution results and event associations ultimately need to be transformed into actionable feedback for learners. This paper constructs a semantic template library T , where each template consists of a trigger condition C , feedback text R , and visual annotation V_{is} . The trigger condition C is a multimodal logical expression, for example, $C = \{AU4 \text{ activation} \wedge \text{shoulder forward} \wedge \Delta s_{conf}(t) < -0.2\}$, which integrates three aspects: facial action units, pose spatial features, and emotion temporal changes. For the input video sequence, the template library is traversed to match trigger conditions, and the template with the highest confidence is selected for semantic filling. To avoid feedback content generated in adjacent windows being too dense in time, the longest valid subsequence algorithm is used to select M' key time windows in the sequence, with at least 1 second time interval between windows. The final structured feedback is output as:

$$F = \{(t_{start}^{(m)}, t_{end}^{(m)}, L_{face}^{(m)}, L_{pose}^{(m)}, R^{(m)})\}_{m=1}^{M'} \quad (18)$$

where, $t_{start}^{(m)}$ and $t_{end}^{(m)}$ mark the time window corresponding to the feedback, $L_{face}^{(m)}$ and $L_{pose}^{(m)}$ are the corresponding spatiotemporal attribution heatmaps, and $L^{(m)}$ is the natural language feedback text. This structured format covers three dimensions simultaneously: temporal localization, visual evidence, and teaching semantics, elevating the system output from isolated classification labels to teaching feedback with a complete explanation chain.

2.6 Loss function and training strategy

The overall loss function consists of three parts: emotion classification loss, confidence consistency loss, and regularization loss, combined as $L_{total} = L_{cls} + \lambda_1 L_{conf} + \lambda_2 L_{reg}$. The emotion classification loss adopts Focal Loss to alleviate

the problem of imbalanced emotion category distribution in oral teaching scenarios, where neutral and focused samples are significantly more numerous than nervous and fatigued samples:

$$L_{cls} = -\frac{1}{T} \sum_{t=1}^T \sum_{c=1}^C \alpha_c (1-y_c^{(t)})^\gamma \log(y_c^{(t)}) \quad (19)$$

where, α_c is the class-balanced weight, determined by normalizing the inverse of each class's sample frequency, used to assign higher loss weights to minority classes; $\gamma = 2$ is the focusing parameter, controlling the loss decay rate for easy-to-classify samples, so that the model focuses optimization on hard-to-classify samples; $y_c^{(t)}$ is the model's predicted probability of frame t belonging to class c . The confidence consistency loss imposes a constraint on the positive correlation between the gating fusion weights and the unimodal classification confidence:

$$L_{conf} = \frac{1}{T} \sum_{t=1}^T \sum_m \|g_m^{(t)} - \text{confidence}_m^{(t)}\|_2^2 \quad (20)$$

where, $g_m^{(t)}$ is the gating fusion weight of modality m at time t , and $\text{confidence}_m^{(t)}$ is the maximum Softmax probability output when only the unimodal feature of that modality is input to the classifier. The purpose of this loss term is to make the fusion weights assigned by the gating network match the classification reliability of each modality. When a modality's feature has strong discriminability at the current moment, its unimodal classification confidence is high, and the gating network should correspondingly increase the fusion weight of that modality, and vice versa. This explicit positive correlation constraint provides direct supervision from feature discriminability to the gating network's decision process, preventing the gating weights from relying solely on data quality confidence while ignoring the actual effectiveness of features in the classification task. At the same time, it provides a clear optimization direction for the gating network in the early training stage. The regularization loss L_{reg} imposes a norm constraint l_2 on the weight matrix of the gating network W_g , with a coefficient of 1×10^{-4} , to suppress the excessive growth of gating parameters toward extreme values during training.

Training adopts a two-stage strategy to avoid coupled oscillations in the early stage of multi-module joint optimization. In the first stage, all parameters of the gating fusion module are frozen, and only the facial feature extractor and pose feature extractor are trained under supervision, with the optimization objective being L_{cls} , for 20 epochs. This stage enables the feature extractors of both modalities to establish good unimodal classification capabilities in their respective feature spaces, providing reliable unimodal classification confidence $\text{confidence}_m^{(t)}$ for subsequent confidence estimation. In the second stage, all module parameters are unfrozen, and end-to-end joint fine-tuning is performed with L_{total} as the optimization objective for 30 epochs. The optimizer uses Adam with an initial learning rate of 1×10^{-4} , decaying to 0.5 times the current value every 10 epochs, and the batch size is set to 32. The weight coefficients λ_1 and λ_2 are set to 0.5 and 1×10^{-4} , respectively. This combination was determined through grid search on the validation set to keep the magnitudes of each loss term similar in the early training stage.

All experiments were completed under the PyTorch framework using a single NVIDIA A100 40GB GPU.

3. EXPERIMENTS

3.1 Experimental setup

Since existing public emotion recognition datasets such as FERV39k, Emotic, and CAER are primarily oriented toward general scenarios, lacking annotations for oral teaching scenarios and not simultaneously containing dense frame-level annotations for both facial and full-body poses, this paper constructs the SPEECH-EMO dataset. The dataset was collected from English oral classes in 5 middle schools, comprising oral expression videos from 120 learners (aged 14–17, male-to-female ratio 1:1), with a total duration of approximately 45 hours. Each learner completed a 3–5 minute oral task in a natural teaching environment, including passage reading, free expression, and situational dialogue. The video resolution was 1920×1080 at 30 fps. The emotion categories were independently annotated by 3 graduate students majoring in psychology using a majority voting system. A total of five categories were annotated: confidence, nervousness, focus, fatigue, and neutral, with sample proportions of 18.2%, 15.7%, 31.4%, 10.3%, and 24.4%, respectively. The dataset was divided into training, validation, and test sets in a 7:1.5:1.5 ratio. The division ensured that videos from the same learner appeared in only one set to avoid data leakage.

Evaluation metrics used accuracy and weighted F1-score as the main quantitative indicators. In addition, a feedback usability score was introduced, where 5 experts with over 5 years of middle school English teaching experience rated the system-generated feedback on a 1–5 scale across three dimensions: Accuracy reflects the consistency between the feedback description and the actual behavior in the video; Actionability reflects whether the feedback includes clear teaching improvement suggestions; Instructional Value reflects whether the feedback provides substantial help for optimizing the teaching process. The final feedback usability score is the average of the three dimensions. Comparison methods include the Face-only and Pose-only unimodal baselines, Early Fusion and Late Fusion, KGSE-ER (IEEE TIP 2023), SG-TE (CVPR 2024), and MLDF (IEEE TMM 2024). All comparison methods were retrained on the SPEECH-EMO dataset, with hyperparameters set to the optimal settings from the original papers.

3.2 Overall performance comparison

Figure 5 shows the overall performance comparison of different methods on the SPEECH-EMO test set. The results show that the proposed full system achieves an accuracy of 86.4%, which is 4.8 and 5.9 percentage points higher than the current optimal methods Multi-level Feature Decomposition and Fusion (MLDF) and Spatial Guidance and Temporal Enhancement (SG-TE), respectively. The improvement trend of the weighted F1-score is consistent with accuracy, indicating that the performance gain stably improves all categories rather than coming only from the majority classes. The Pose-only accuracy is only 61.7%, significantly lower than the Face-only accuracy of 68.3%, indicating that facial expression remains the most important information source for emotion recognition in oral teaching scenarios, which also

verifies the rationality of the asymmetric fusion design that takes the face as the primary modality and pose as the auxiliary. In terms of feedback usability, the proposed method achieves a Feedback Usability Score (FUS) of 4.6, far exceeding MLDF's 4.0. The main reason is that the Spatio-Temporal Explainable Feedback Generation (STFEG) module can locate specific action details and time points, while comparison methods only output category labels or simple heatmap overlays, lacking teaching semantic transformation.

Figure 6 presents the fine-grained performance across different emotion categories. The accuracy of the fatigue class on Face-only and Pose-only is 63.1% and 55.6%, respectively, both far below other categories, indicating that single-modality coverage of discriminative cues for fatigue is insufficient. The proposed method, through the PGFA mechanism, enables the network to actively shift attention to the eye region when learning fatigue, while pose features provide structured evidence of overall slackness, increasing the fatigue class accuracy to 82.3%, an increase of 18.7 percentage points. The nervousness class performs worst on Pose-only, at only 59.4%, but the proposed confidence fusion mechanism dynamically increases the pose modality weight when facial feature discriminability is insufficient, achieving a nervousness class accuracy of 84.9%, demonstrating the value of the Dynamic Confidence-Weighted cross-modal Fusion (DCWF) module.

To verify the emotion feature extraction and cross-modal collaboration capabilities of the proposed method when visual information is complete and learners are in a natural expression state, a confidence expression sample with a forward gaze and relaxed posture was selected for qualitative analysis. As shown in Figure 7(a), the facial feature extraction module can stably locate fine-grained regions such as the eyes, eyebrows, nose, and mouth. The pose estimation result accurately maintains the structural relationship between the head, shoulders, and arms, where the head orientation is stable, shoulders are naturally open, and upper limb posture is coordinated, forming a visual representation consistent with the confidence state. The dynamic fusion stage further integrates facial and pose information, with the facial modality weight reaching 0.66, higher than the pose modality's 0.34, indicating that under front view and sufficient facial information conditions, the system can actively increase the contribution of facial features with stronger emotion discriminability, while retaining pose information as global behavioral context supplement. The final fusion result stably determines confidence, showing that the proposed method can not only obtain a consistent representation between local expression and overall body structure, but also form a reasonable asymmetric modality allocation according to current visual quality, providing a reliable basis for accurate recognition and targeted feedback of positive expression states in English oral teaching.

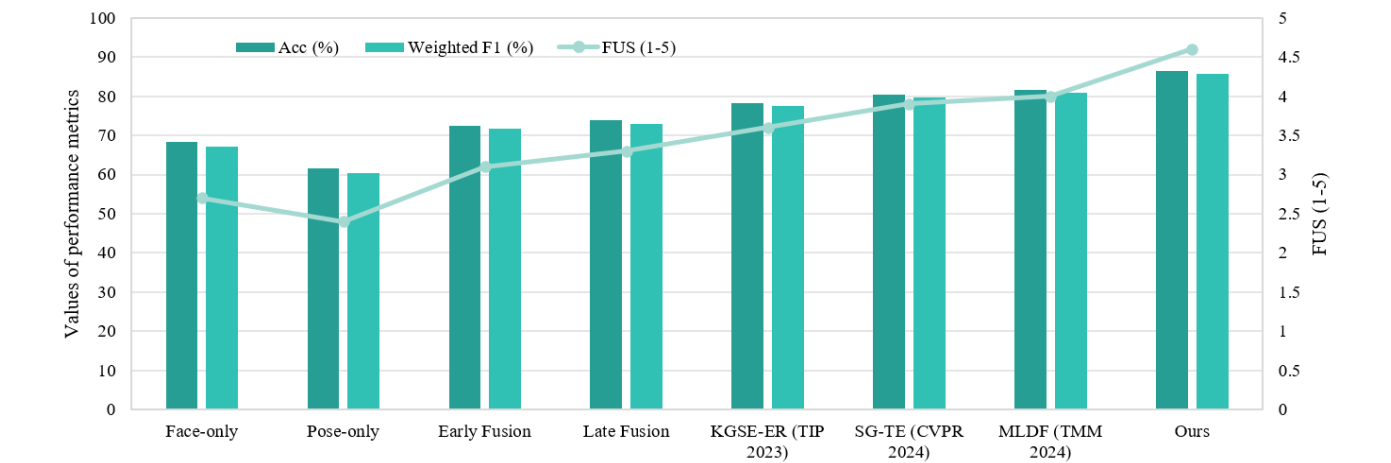


Figure 5. Overall performance comparison of different methods on the SPEECH-EMO test set
Note: Acc: Accuracy; MLDF: Multi-Level Feature Decomposition and Fusion; SG-TE: Spatial Guidance and Temporal Enhancement.

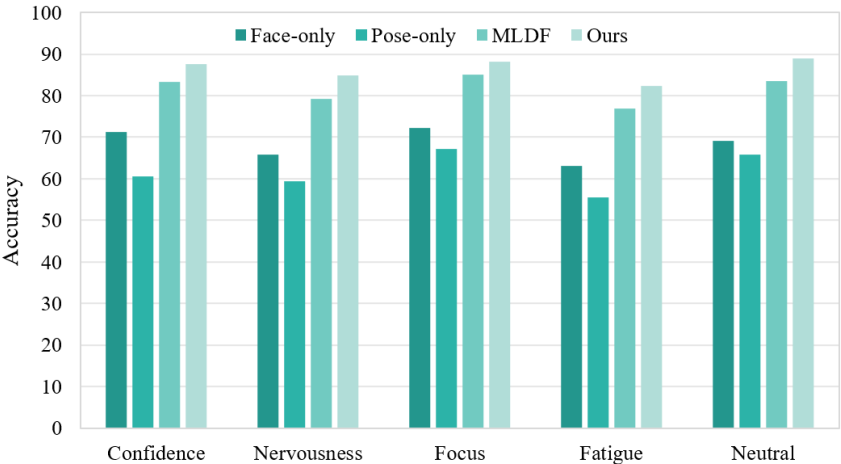
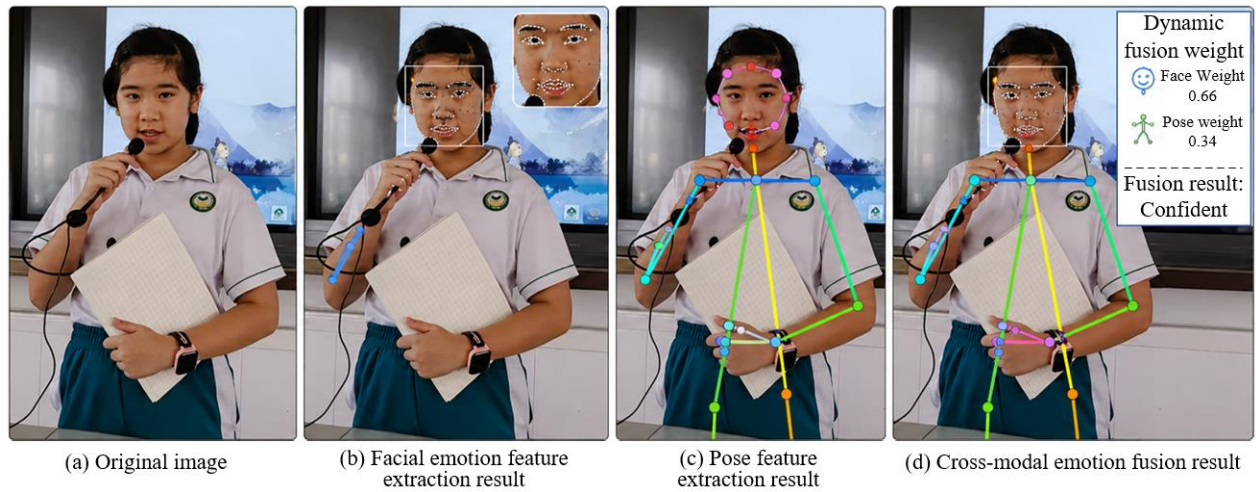
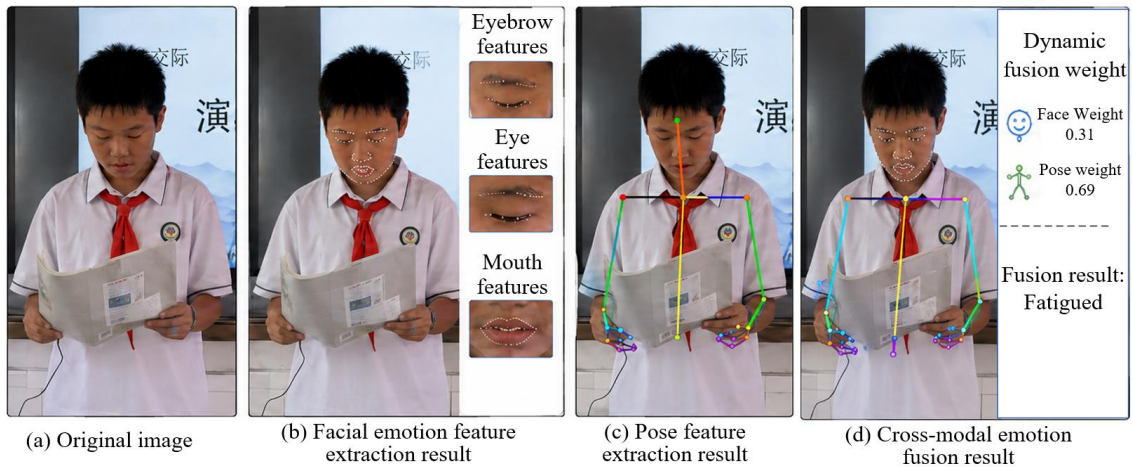


Figure 6. Comparison of classification accuracy for each category (%)
Note: MLDF: Multi-Level Feature Decomposition and Fusion.



(a) Facial-dominant collaborative recognition effect in confidence expression scenario



(b) Pose-compensated recognition effect in fatigue and head-lowered scenario

Figure 7. Implementation effect of English oral teaching emotion analysis integrating visual emotion recognition and pose estimation

To verify whether the proposed method can still maintain reliable emotion judgment capability when head lowering leads to degraded facial information quality, a fatigue expression sample with lowered head, drooping eyelids, and relaxed shoulders was selected for qualitative analysis. In Figure 7(b), although the facial feature module can still extract local information such as eye closure degree and mouth state, due to the learner's continuous head lowering, the frontal facial texture and landmark visibility are significantly weaker than in the first two scenarios. In contrast, the pose branch can stably capture structural clues with high fatigue discrimination value, such as head drooping, neck-shoulder relationship changes, and shoulders relaxation. Different from the first two samples, the dynamic fusion stage reduces the facial weight to 0.31 and significantly increases the pose weight to 0.69, forming an adaptive transition from facial dominance to pose dominance, and correctly outputs the fatigue state. This phenomenon shows that the proposed method can utilize body pose to complete effective information compensation when facial visual evidence degrades, avoiding single-modality failure caused by head lowering, rotation, or partial occlusion. It qualitatively verifies the adaptability of the dynamic confidence fusion mechanism to complex classroom visual conditions, and also provides important support for the system to continuously generate stable and credible emotion state assessment and teaching feedback in actual oral teaching

environments. The robustness experiment of the paper also shows that under head-lowered conditions, the performance of the facial single modality drops significantly, while the fusion system can still maintain a high-performance retention rate, consistent with the pose compensation phenomenon presented in Figure 7.

3.3 Ablation study

Figure 8 presents the ablation contribution of each innovative module. The baseline method has an accuracy of 74.3% and a FUS score of only 2.9, indicating that simple multimodal concatenation has limited performance in oral teaching scenarios. Adding PGFA alone increases the accuracy by 5.5 percentage points, demonstrating that pose-guided spatial attention helps the facial feature network focus more precisely on emotion-relevant regions. Adding DCWF alone results in a 6.2 percentage point improvement, indicating that data-adaptive dynamic fusion is better equipped to handle modal quality fluctuations than fixed concatenation. When PGFA and DCWF work together, the accuracy reaches 85.1%, which is only 1.3 percentage points lower than the complete model, suggesting that the innovations at the feature extraction and fusion levels have largely resolved the accuracy issue of emotion recognition. The FUS score of the complete model jumps from 3.8 to 4.6; this 0.8-point improvement is solely

attributed to the STEFG module, confirming that transforming recognition results into understandable and actionable teaching language is the key link determining the system's practical value.

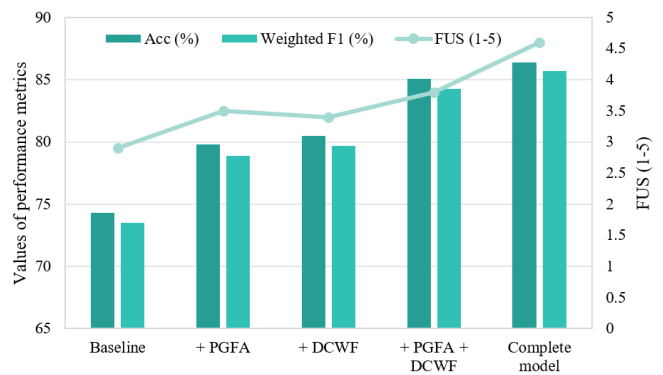


Figure 8. Ablation contribution of each innovative module
Note: Acc: Accuracy; FUS: Feedback Usability Score; PGFA: Pose-Guided Facial Attention; DCWF: Dynamic Confidence-Weighted Cross-modal Fusion.

Table 1. Sensitivity analysis of the Gaussian kernel bandwidth and adaptive adjustment coefficient for the Pose-Guided Facial Attention (PGFA) module (Accuracy (%))

β	$\sigma_0 = 1.5$ $\sigma_0 = 1.5$	$\sigma_0 = 2.5$ $\sigma_0 = 2.5$	$\sigma_0 = 3.5$ $\sigma_0 = 3.5$	$\sigma_0 = 5.0$ $\sigma_0 = 5.0$
0	82.7	83.9	83.2	82.1
0.3	83.5	85.2	84.6	83
0.6	84	86.4	85.7	83.9
1	83.8	85.9	85.3	84.1

Table 1 examines the impact of the Gaussian heatmap bandwidth parameters in the PGFA module on performance. With a fixed bandwidth, the optimal yields an accuracy of 83.9%. If the bandwidth is too small, the attention fails to cover the actual facial region when keypoint positions deviate; if the bandwidth is too large, the attention mask becomes overly smooth, losing spatial guidance precision. After introducing the adaptive adjustment mechanism, the performance under all configurations improves, because keypoints with lower detection confidence do exhibit greater positional uncertainty. The highest accuracy of 86.4% is

achieved when $\sigma = 2.5$ and $\beta = 0.6$. At this point, the heatmap maintains sharp focus on high-confidence keypoints while appropriately diffusing for low-confidence keypoints, achieving the optimal balance between spatial guidance precision and robustness.

3.4 Robustness experiments

Table 2 evaluates the system's robustness under common interferences in real teaching scenarios. The facial single modality's retention rate under illumination changes is only 76.7% and 71.6%, and drops to 60.3% and 52.1% in head rotation 60° and head-lowered scenarios, exposing the vulnerability of pure facial emotion recognition in practical applications. In contrast, the pose single modality maintains a retention rate above 90% under all interference conditions, because pose estimation relies on the structured contour of the full-body skeleton, which is naturally insensitive to illumination and facial occlusion. The proposed full model still maintains an accuracy of 73.6% under the most severe head-lowered scenario, with a retention rate of 85.2%, far superior to MLDF's 57.9% and 71.0%. Even at head rotation 60°, the proposed accuracy still reaches 74.8%, 33.6 percentage points higher than Face-only. This advantage stems from the DCWF module: when facial confidence drops significantly, the gating network automatically adjusts the fusion weight from about 0.65 to 0.20–0.30, realizing an intelligent switch to pose-modality dominance.

Table 3 simulates extreme modality-missing situations. Late Fusion's accuracy drops sharply to 58.4% when the face is missing; the fixed average weight causes invalid input to directly contaminate the fused features. The proposed method achieves an accuracy of 78.3% when the face is missing, a decrease of only 8.1 percentage points from normal conditions. This benefits from the DCWF module: when the facial detection box area approaches 0, the confidence is reduced to near 0, and the gating weight allows the pose modality to fully take over the decision. Meanwhile, the random perturbation regularization during training forces the pose encoder to possess the ability to perform classification independently. When the pose is missing, the proposed accuracy is 80.1%, with a smaller decrease than in the facial-missing case, consistent with the fact that the Face-only baseline is higher than the Pose-only baseline.

Table 2. Accuracy (%) of each method under different interference conditions (Robustness retention rate relative to normal conditions in parentheses)

Interference Condition	Face-only	Pose-only	MLDF	Ours
Normal condition	68.3	61.7	81.6	86.4 (100%)
Low illumination	52.4 (76.7%)	57.3 (92.9%)	69.8 (85.5%)	80.2 (92.8%)
Side lighting	48.9 (71.6%)	55.8 (90.4%)	67.2 (82.4%)	78.5 (90.9%)
Head rotation 30°	55.6 (81.4%)	58.9 (95.5%)	73.4 (89.9%)	82.1 (95.0%)
Head rotation 60°	41.2 (60.3%)	56.3 (91.2%)	62.5 (76.6%)	74.8 (86.6%)
Hand occlusion of face	49.7 (72.8%)	60.1 (97.4%)	68.3 (83.7%)	79.3 (91.8%)
Head lowered	35.6 (52.1%)	58.4 (94.7%)	57.9 (71.0%)	73.6 (85.2%)

Note: MLDF: Multi-Level Feature Decomposition and Fusion.

Table 3. Performance comparison under extreme modality-missing conditions (Accuracy (%) / F1-score (%))

Test Condition	Late Fusion	MLDF	Ours
Normal	73.8 / 72.9	81.6 / 80.9	86.4 / 85.7
Facial detection completely failed	58.4 / 57.1	65.2 / 64.0	78.3 / 77.2
Pose detection completely failed	66.7 / 65.8	72.4 / 71.5	80.1 / 79.4

Note: MLDF: Multi-Level Feature Decomposition and Fusion.

3.5 Interpretability evaluation

Table 4 employs the removal diagnostic paradigm to evaluate attribution faithfulness. When randomly removing 50% of the regions, the confidence drops by only 15.1%, indicating that random masking has limited impact on the model output. When removing the facial heatmap highlighted regions, the confidence drops by 38.6%, demonstrating that the facial attention guided by PGFA indeed focuses on local

regions that make substantial contributions to emotion discrimination. The confidence drop is 31.7% when removing the pose heatmap highlighted keypoints, lower than that of the face, because pose features are dispersed across multiple nodes after graph convolution. When jointly removing high-contribution regions from both modalities, the confidence drops by 53.8%, indicating that the attributions of the two modalities are complementary, and the joint attribution covers a broader range of decision evidence.

Table 4. Evaluation of Gradient-weighted Class Activation Mapping (Grad-CAM) attribution faithfulness (confidence drop after removing high-contribution regions; a larger drop indicates more accurate attribution)

Removal Ratio	Random Removal	Remove Facial Heatmap Highlighted Regions	Remove Pose Heatmap Highlighted Keypoints	Remove Joint Heatmap Highlighted Regions
10%	3.20%	8.70%	7.10%	13.50%
30%	9.80%	22.30%	18.50%	35.20%
50%	15.10%	38.60%	31.70%	53.80%

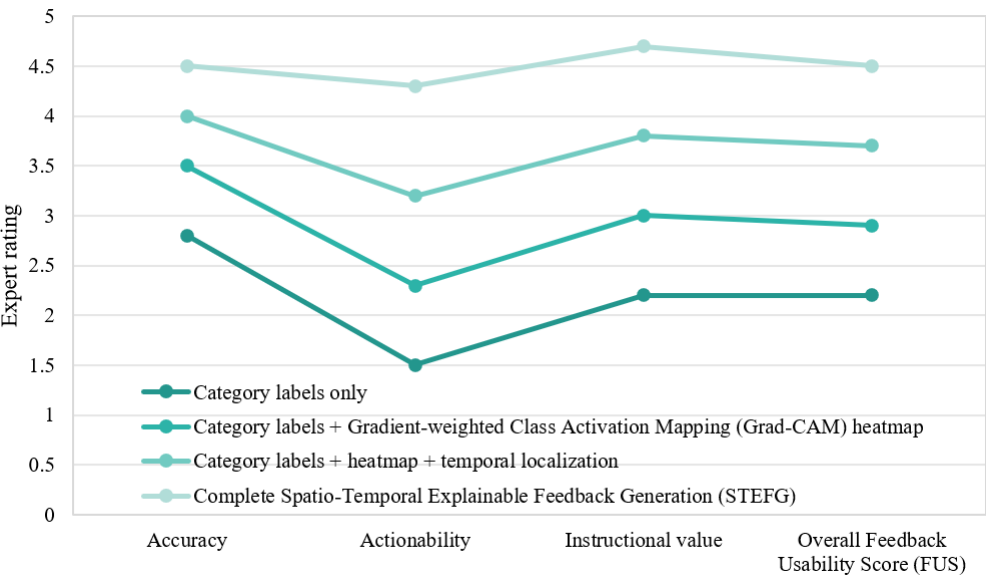


Figure 9. Expert ratings on the quality of teaching-oriented feedback (breakdown of feedback usability score dimensions)

Figure 9 presents expert ratings for different feedback granularities. When only outputting category labels, the actionability score is only 1.5, and experts consider it impossible to conduct teaching interventions. After adding Grad-CAM heatmaps, the accuracy score increases to 3.5, but the actionability remains very low. After further adding temporal localization, the instructional value reaches 3.8. The complete STEFG introduces semantic template filling, causing the actionability to surge to 4.3 and the comprehensive FUS to reach 4.5. Experts believe that the feedback text can be directly used for lesson preparation and teaching improvement. The increase in the accuracy dimension from 4.0 to 4.5 is relatively small, indicating that the core contribution of semantic template filling lies in translating attribution results into teaching language, thereby amplifying the system's practical value.

3.6 Computational efficiency analysis

Table 5 reports the inference efficiency of each method on a single NVIDIA A100 GPU. The total single-frame processing time of the proposed method is 63.9 ms, approximately 15.6 FPS. Although slower than single-modal

methods and MLDF, considering the offline processing scenario of oral teaching videos, this efficiency fully meets practical application requirements. The feature extraction stage is the main time-consuming source, with HRNet pose estimation taking 19.7 ms and ResNet-50 facial feature extraction taking 28.3 ms. The feedback generation stage of the proposed method takes 8.5 ms, mainly used for temporal smoothing filtering and template matching search; this additional overhead is acceptable in offline scenarios. If deployment to real-time interactive scenarios is required, the total time consumption can be further compressed through TensorRT quantization and frame rate downsampling.

Table 5. Comparison of inference efficiency of each method (single-frame processing time in ms)

Stage/Method	Face-only	Pose-only	MLDF	Ours
Feature extraction	28.3	19.7	48.2	49.1
Fusion and classification	2.1	1.8	5.7	6.3
Feedback generation	—	—	4.2	8.5
Total	30.4	21.5	58.1	63.9

Note: MLDF: Multi-Level Feature Decomposition and Fusion.

4. DISCUSSION AND LIMITATIONS

The three modules proposed in this paper form a complete chain from feature alignment and adaptive fusion to interpretable feedback in terms of technical route. However, the actual effectiveness of each module is limited by several preconditions and implementation constraints. The semantic template library in the STEFG module currently contains 120 templates, which already covers the vast majority of common emotion-behavior patterns in oral teaching scenarios. Nevertheless, it is essentially an enumerative design within a finite state space. When learners exhibit emotional expressions that are not preset in the template library—such as limb stiffness caused by excessive anxiety or rare nervous smiles—the system will degrade to a basic mode that only outputs category labels and attribution heatmaps, and cannot generate matching semantic feedback text. The root of this limitation lies in the fact that the feedback generation stage adopts a matching-based rather than generation-based strategy. A possible direction for future improvement is to introduce a large language model, taking the attribution results as multimodal context input and having the large language model generate free-text feedback suggestions. This solution can not only break through the ceiling of the template library size, but also enable the feedback text to possess stronger semantic diversity and contextual adaptability. However, its cost is that it requires additional resolution of the balance between the inference latency of the large language model and content controllability in teaching scenarios.

The effectiveness of the PGFA module is highly dependent on the detection accuracy and spatial stability of pose estimation. Although the currently adopted HRNet-W32 performs excellently on standard pose estimation benchmarks, in the field deployment of oral teaching scenarios, the positioning accuracy of shoulder and torso keypoints noticeably decreases when learners wear loose clothing. The fundamental reason for this phenomenon is that the training data of HRNet mainly consists of images of people in tight or regular clothing. There is a distribution shift between the visual appearance prior learned by the model and the contour features under loose clothing. The positional deviation of shoulder keypoints is transmitted to the facial feature map space via TPS spatial mapping, directly affecting the spatial precision of the attention mask. In extreme cases, it may even mislead the facial feature extraction network to focus on edge regions irrelevant to emotional semantics. This problem reveals an inherent contradiction in the pose-guided attention mechanism: the original intention of this mechanism is to use pose information to compensate for the deficiencies of facial features, but when the pose information itself is unreliable, the introduced spatial prior becomes a source of noise instead. A feasible path to solve this problem is to introduce pose uncertainty estimation, using the detection confidence of each keypoint not only for the adaptive adjustment of heatmap bandwidth, but further transmitting positional uncertainty to the generation process of the attention mask in the form of a probability distribution, enabling the network to automatically reduce the intensity of spatial guidance when pose estimation is unreliable.

The design of the current system takes single-person video input as the basic assumption, and this constraint limits the applicability of the method in real classroom scenarios. In teaching activities such as group discussions and peer assessments, multiple learners may appear in the same frame,

and the problem of individual association confusion becomes the main obstacle to system deployment. The skeletal keypoints output by the pose estimation branch in multi-person scenarios lack explicit modeling of identity attribution. The shoulder and head keypoints of different individuals are spatially interleaved, making it difficult for the PGFA module to determine which subject's pose information should be used to guide facial attention. If directly extended to multi-person scenarios, the system framework needs to be redesigned: introducing a multi-target tracking mechanism at the perception level to achieve temporal association of individual identities; constructing an individual matching strategy at the attention guidance level based on the spatial subordination between facial detection boxes and pose keypoints; and maintaining independent emotion state sequences for each individual and generating individualized diagnostic reports at the feedback generation level. The above extensions not only involve the reconstruction of the perception front-end, but also require the dataset annotation paradigm to evolve from single-subject emotion labels to multi-subject interactive emotion annotation. Its complexity exceeds the scope of the current methodology and can be explored in depth as an independent follow-up research direction.

5. CONCLUSION

This paper addresses three core problems of visual emotion recognition in English oral teaching scenarios: insufficient cross-modal feature spatial alignment, lack of dynamic confidence awareness in fusion strategies, and lack of interpretable teaching feedback in system output. Correspondingly, a pose-guided facial visual attention mechanism, a dynamic confidence-weighted cross-modal adaptive fusion strategy, and a spatiotemporal joint interpretability feedback generation mechanism oriented to oral teaching are proposed. The three modules are organically connected with feature alignment, adaptive fusion, and interpretable feedback as the technical main line, forming an end-to-end emotion recognition and feedback system for oral teaching scenarios. The core design philosophy of this system is to elevate pose information from a traditional auxiliary modality that merely serves as classification features to a structured prior that can actively guide facial feature extraction and dynamically regulate fusion weights, enabling multimodal fusion to shift from fixed concatenation to adaptive collaboration that perceives data quality and task requirements.

Extensive experiments on the self-built SPEECH-EMO dataset show that the complete system achieves high recognition accuracy and feedback usability scores. Ablation studies verify the independent effectiveness of each module. Robustness experiments demonstrate that the system maintains a high-performance retention rate under interferences such as illumination changes, head rotations, occlusions, and modality absence. Interpretability evaluation verifies the credibility and teaching practicality of the feedback through attribution faithfulness tests and expert ratings. Current method still has limitations in the coverage breadth of the template library, multi-person scenario expansion, and pose estimation accuracy under loose clothing. Future research will focus on three directions: free-text feedback generation based on large language models, system reconstruction under a multi-target tracking framework, and uncertainty-aware robust attention mechanisms, to promote

the system's transition from laboratory verification to normalized deployment in real classrooms.

REFERENCES

- [1] Bielak, J. (2022). To what extent are foreign language anxiety and foreign language enjoyment related to L2 fluency? An investigation of task-specific emotions and breakdown and speed fluency in an oral task. *Language Teaching Research*, 29(3): 911-941. <https://doi.org/10.1177/13621688221079319>
- [2] Solhi, M. (2024). The impact of EFL learners' negative emotional orientations on (Un)willingness to communicate in in-person and online L2 learning contexts. *Journal of Psycholinguistic Research*, 53(2): 1-25. <https://doi.org/10.1007/s10936-024-10071-y>
- [3] Tang, X., Gong, Y., Xiao, Y., Xiong, J., Bao, L. (2024). Facial expression recognition for probing students' emotional engagement in science learning. *Journal of Science Education and Technology*, 34(1): 13-30. <https://doi.org/10.1007/s10956-024-10143-7>
- [4] Tonguç, G., Ozkara, B.O. (2020). Automatic recognition of student emotions from facial expressions during a lecture. *Computers & Education*, 148: 103797. <https://doi.org/10.1016/j.compedu.2019.103797>
- [5] Leong, S.C., Tang, Y.M., Lai, C.H., Lee, C. (2023). Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. *Computer Science Review*, 48: 100545. <https://doi.org/10.1016/j.cosrev.2023.100545>
- [6] Martinez-Martin, E., Fernández-Caballero, A. (2025). Improved human emotion recognition from body and hand pose landmarks on the GEMEP dataset using machine learning. *Expert Systems with Applications*, 269: 126427. <https://doi.org/10.1016/j.eswa.2025.126427>
- [7] Wei, J., Hu, G., Yang, X., Luu, A.T., Dong, Y. (2023). Learning facial expression and body gesture visual information for video emotion recognition. *Expert Systems with Applications*, 237: 121419. <https://doi.org/10.1016/j.eswa.2023.121419>
- [8] Khediri, N., Ben Ammar, M., Kherallah, M. (2023). A real-time multimodal intelligent tutoring emotion recognition system (MITERS). *Multimedia Tools and Applications*, 83(19): 57759-57783. <https://doi.org/10.1007/s11042-023-16424-4>
- [9] Singh, N., Kapoor, R. (2023). Multi-modal expression detection (MED): A cutting-edge review of current trends, challenges and solutions. *Engineering Applications of Artificial Intelligence*, 125: 106661. <https://doi.org/10.1016/j.engappai.2023.106661>
- [10] Chen, S., Tang, J., Zhu, L., Kong, W. (2022). A multi-stage dynamical fusion network for multimodal emotion recognition. *Cognitive Neurodynamics*, 17(3): 671-680. <https://doi.org/10.1007/s11571-022-09851-w>
- [11] Yan, J., Li, P., Du, C., et al. (2024). Multimodal emotion recognition based on facial expressions, speech, and body gestures. *Electronics*, 13(18): 3756. <https://doi.org/10.3390/electronics13183756>
- [12] Xu, M., Shi, T., Zhang, H., Liu, Z., He, X. (2025). A hierarchical cross-modal spatial fusion network for multimodal emotion recognition. *IEEE Transactions on Artificial Intelligence*, 6(5): 1429-1438. <https://doi.org/10.1109/tai.2024.3523250>
- [13] Kalateh, S., Estrada-Jimenez, L.A., Nikghadam-Hojjati, S., Barata, J. (2024). A systematic review on multimodal emotion recognition: Building blocks, current state, applications, and challenges. *IEEE Access*, 12: 103976-104019. <https://doi.org/10.1109/access.2024.3430850>
- [14] Shou, Z., Huang, Y., Li, D., et al. (2024). A student facial expression recognition model based on multi-scale and deep fine-grained feature attention enhancement. *Sensors*, 24(20): 6748. <https://doi.org/10.3390/s24206748>
- [15] Kang, B., Wang, S., Wang, Z., et al. (2025). Progressive masking oriented self-taught learning for occluded facial expression recognition. *IEEE Transactions on Affective Computing*, 16(3): 1277-1289. <https://doi.org/10.1109/taffc.2025.3544677>
- [16] Chen, N., Kok, V.J., Chan, C.S. (2024). Enhancing facial expression recognition under data uncertainty based on embedding proximity. *IEEE Access*, 12: 85324-85337. <https://doi.org/10.1109/access.2024.3415154>
- [17] Zhu, Q., Zheng, C., Zhang, Z., Shao, W., Zhang, D. (2023). Dynamic confidence-aware multi-modal emotion recognition. *IEEE Transactions on Affective Computing*, 15(3): 1358-1370. <https://doi.org/10.1109/taffc.2023.3340924>
- [18] Johnson, D.S., Hakobyan, O., Paletschek, J., Drimalla, H. (2024). Explainable AI for audio and visual affective computing: A scoping review. *IEEE Transactions on Affective Computing*, 16(2): 518-536. <https://doi.org/10.1109/taffc.2024.3505269>
- [19] Deramgozin, M.M., Jovanovic, S., Arevalillo-Herráez, M., Ramzan, N., Rabah, H. (2023). Attention-enabled lightweight neural network architecture for detection of action unit activation. *IEEE Access*, 11: 117954-117970. <https://doi.org/10.1109/access.2023.3325034>
- [20] Fang, B., Li, X., Han, G., He, J. (2023). Facial expression recognition in educational research from the perspective of machine learning: A systematic review. *IEEE Access*, 11: 112060-112074. <https://doi.org/10.1109/access.2023.3322454>