Traitement du Signal
Vol. 42, No. 4, August, 2025, pp. 2431-2442
Journal homepage: http://iieta.org/journals/ts

A Deep Image Recognition-Based Automatic Evaluation Method for English Speaking Interaction Behaviors Enhanced by Attention Mechanisms

Hongli Feng

School of Foreign Languages, Ningxia Medical University, Ningxia 750004, China

Corresponding Author Email: fhl771010@163.com

https://doi.org/10.18280/ts.420449

Received: 8 November 2024
Revised: 18 May 2025
Accepted: 10 June 2025
Available online: 14 August 2025

Keywords:
English speaking interaction behavior, automatic evaluation, attention mechanism, deep image recognition, feature extraction

ABSTRACT

Against the backdrop of globalization and the rapid advancement of intelligent education, English speaking interaction ability has become a core competence in international communication. Traditional manual evaluation methods suffer from low efficiency and strong subjectivity, making them inadequate for large-scale, objective assessments. Therefore, research on automatic evaluation methods for English speaking interaction behaviors is of significant practical importance. Current studies often rely solely on audio features, overlooking critical visual cues such as facial expressions and body movements, which results in incomplete assessments. While some approaches attempt to incorporate visual information, traditional image recognition models struggle to capture key features in complex interactive scenarios and lack effective mechanisms for integrating multi-dimensional features. To address these challenges, this study proposes an automatic evaluation method for English speaking interaction behaviors by integrating attention mechanisms with deep image recognition. The core contributions of this research are twofold: (1) the development of an interaction behavior recognition model based on an optimized attention mechanism, which consists of a global feature branch for holistic image feature extraction, an improved window-based attention branch for focusing on local key regions, and an enhanced channel attention branch for reinforcing important feature channels; (2) the design of an automatic evaluation framework that utilizes the accurately extracted features from the recognition model in conjunction with established speaking interaction assessment criteria to perform comprehensive evaluations. The innovations of this study lie in: (a) the proposed multi-branch attention model that enables precise extraction of global, local, and channel-specific features, overcoming the limitations of traditional models in feature representation; and (b) the deep integration of visual recognition with evaluation logic, establishing a complete technical pipeline from feature extraction to final assessment. This method significantly enhances the objectivity and accuracy of evaluations and offers a novel solution for intelligent spoken English assessment in the education domain.

1. INTRODUCTION

With the deep development of globalization, English, as an important tool for international communication, is increasingly highlighting the importance of spoken interaction ability [1-4]. With the rapid development of online education and intelligent assessment, the efficient and objective automatic evaluation of English-speaking interaction behaviors has become an urgent demand [5-7]. Traditional English speaking interaction evaluation mainly relies on manual scoring, which is not only inefficient and costly, but also easily affected by the subjectivity of the scorer, making it difficult to meet the needs of large-scale and normalized evaluation. Therefore, the research on related automatic evaluation methods has gradually become a hot topic [8-10]. The automatic evaluation of English-speaking interaction behaviors has important practical significance and application value. For learners, an efficient automatic evaluation method can provide timely feedback on the strengths and weaknesses in their speaking interaction, helping them to practice and improve in a targeted manner and enhance learning efficiency. For educational institutions and assessment agencies, this research can significantly improve evaluation efficiency, reduce labor costs, and ensure the objectivity and consistency of evaluation results, better meeting the needs of large-scale assessment [11, 12]. In addition, relevant research can also promote the integrated application of multidisciplinary technologies such as natural language processing and computer vision in the field of education, and promote the development of intelligent education.

At present, many scholars have conducted research in the field of behavior evaluation, but there are still some defects and deficiencies. Some research methods focus only on the extraction and analysis of speech features. For example, the evaluation methods based on speech features proposed by Liang et al. [13, 14] rely solely on speech information and do

not consider image behavior features such as the speaker's facial expressions and body movements, resulting in inaccurate evaluation of emotional expression and communication intentions during interaction. Other studies attempt to combine image features, but the image recognition models used have shortcomings in the specificity and effectiveness of feature extraction. For example, some studies [15, 16] used traditional image recognition models for behavior recognition, but such models lack the ability to capture key behavioral features in complex and dynamic interactive behavior scenarios, and the recognition accuracy needs to be improved. At the same time, the existing evaluation methods often lack effective mechanisms for integrating multi-dimensional features in comprehensive evaluation, making it difficult to fully utilize the advantages of different features [17, 18].

The research of this paper mainly includes two core parts. The first part is the construction of an English-speaking interaction behavior recognition model based on an optimized attention mechanism. This model sets up three core branches: the global feature module branch is used to extract overall image features; the branch containing an improved window attention module focuses on the features of local key regions and enhances the capture of specific interactive behavior details; the branch containing an improved channel attention module highlights the information of important feature channels and improves the discriminative ability of features. The second part is to propose an automatic evaluation method for English speaking interaction behaviors based on deep image behavior recognition. This method uses the accurate behavior features obtained from the above recognition model, combined with relevant rules and standards of spoken interaction, to realize the automatic evaluation of English-speaking interaction behaviors. The value of this research lies in that the constructed recognition model, through the collaborative effect of multiple branches, can extract more comprehensive and accurate English-speaking interaction behavior features, overcoming the limitations of traditional models in feature extraction. The proposed automatic evaluation method combines deep image behavior recognition with evaluation logic, improving the objectivity and accuracy of the evaluation, providing a new effective way for the automatic evaluation of English-speaking interaction behaviors, and has positive significance for promoting the development of oral assessment technology in the field of intelligent education.

2. ENGLISH SPEAKING INTERACTION BEHAVIOR RECOGNITION MODEL BASED ON OPTIMIZED ATTENTION MECHANISM

2.1 Overall network architecture

This paper takes the precise capture of dynamic behavior features in spoken interaction as the core goal and designs the overall network architecture of the English-speaking interaction behavior recognition model based on an optimized attention mechanism, achieving deep adaptation between technical characteristics and recognition requirements in backbone network selection and branch design (Figure 1). The model adopts the residual block before *res_conv4_2* as the backbone network for feature extraction. This choice can effectively extract the basic image features of spoken interaction scenes through the residual structure, such as the posture of the interlocutors and the scene environment, and also lays a reliable feature foundation for the subsequent branch processing. The model also removes the downsampling operation of the *res_conv5_1* residual block to avoid the loss of fine dynamic features in spoken interaction, such as lip movements, micro facial expressions, and gesture changes, retaining more key behavior details without adding extra parameters, and solving the problem of insufficient capture of interaction details in traditional models. At the key stage of feature extraction, the model divides into three independent branches after *res_conv4_2* to achieve multi-dimensional feature collaborative extraction: the global feature module branch does not add an attention mechanism and focuses on capturing the overall scene features of the spoken interaction, such as the spatial positions of interlocutors and the interaction rhythm, providing global contextual support for recognition; the branch with an improved window attention module focuses on local key regions such as lip movements and eye contact areas, enhancing the perception of core behavior details in interaction, and solving the problem of vague capture of local key actions in traditional models; the branch with an improved channel attention module highlights key feature channels corresponding to discriminative features such as lip motion sequences and gesture dynamics, enhancing the ability to extract distinguishing features such as the lip stretch during fluent expression and the gesture amplitude during emotional exchange, realizing the complementarity of features from different dimensions.

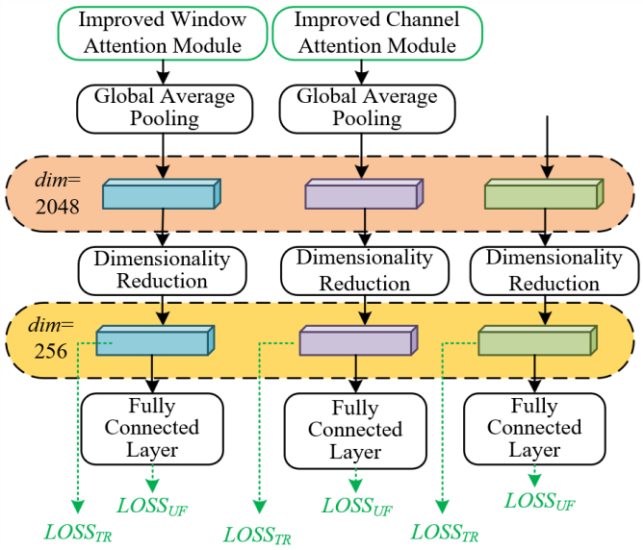

Figure 1. Overall architecture of the proposed model

During the model training stage, the 256-dimensional features after global average pooling and dimensionality reduction of the three branches are used to calculate the triplet loss. By comparing similar and dissimilar interaction behavior features, the model's ability to distinguish different interaction states is improved, such as distinguishing fluent dialogue from interrupted dialogue. The features output by the fully connected layer are used to calculate the cross-entropy loss, to achieve accurate classification of specific interaction behavior categories, such as question-and-answer and discussion. By using backpropagation and combining the gradients of the two losses to update the parameters, the model is ensured to have both feature discrimination and classification accuracy. In the

testing stage, the 256-dimensional feature vectors of the three branches are concatenated to form a comprehensive feature representation, integrating global context, local details, and key channel features, ultimately achieving high-precision recognition of English-speaking interaction behaviors.

2.2 Improved window attention module

In the model, the design of the improved window attention module aims to solve the cooperative learning problem of "local key behavior detail capture" and "global interaction context association" in English speaking interaction behavior recognition, while reducing computational cost (Figure 2). In English spoken interaction scenarios, local details such as lip opening range and finger pointing actions are the core basis for judging fluency of expression and interaction intention. Meanwhile, global associations such as the synchrony between the speaker's lip movement and the listener's eye gaze response, and the coordination between gestures and speech rhythm are the key to understanding complete interactive behavior. However, the traditional global self-attention mechanism increases computational burden by attending to all spatial positions and tends to blur local details. The local self-attention mechanism is difficult to capture cross-region associations, such as the temporal relationship between hand movements and lip movements. This module implements optimization through the process of "window division–local similarity calculation–cross-window association–feature fusion": First, the interaction image frame is divided into several windows such as lip region window, hand region window, and facial expression window. In the "local similarity calculation" stage, the similarity of pixels within the same window is calculated, such as the similarity between the target pixel and surrounding pixels within the lip window, to accurately capture local key behavior details. In the "cross-window association" stage, the similarity of corresponding position pixels in different windows is calculated to establish long-range connections between local regions. Finally, the local similarity and cross-window similarity are fused through the "Cat" operation to form a unified "local–long-range context" feature, such as the detail of a single lip movement and its associated gesture information. This design avoids the undifferentiated computation of all positions in the spatial dimension of global self-attention, and at the same time compensates for the lack of global associations in local self-attention through cross-window association. It can capture both local key details such as lip movement and gesture changes, and interaction associations between different local regions at relatively low cost, thereby enhancing the feature representation ability for English speaking interaction behaviors.

The improved window attention module achieves spatial local context and spatial long-range context modeling for English speaking interaction behavior through the dual-path design of window attention and grid attention. The core goal is to solve the coordination problem in spoken interaction recognition between "precise capture of local action details" and "effective establishment of cross-regional behavior associations". In spoken interaction scenarios, local actions are the basis for judging pronunciation fluency and expression intention, while cross-regional associations are key to understanding interaction logic. Based on an 8×8 feature map, the module focuses on local and long-range features respectively through differentiated division by window size O and grid size H, and maintains lightweightness by parameter sharing, providing structured support for subsequent feature fusion.

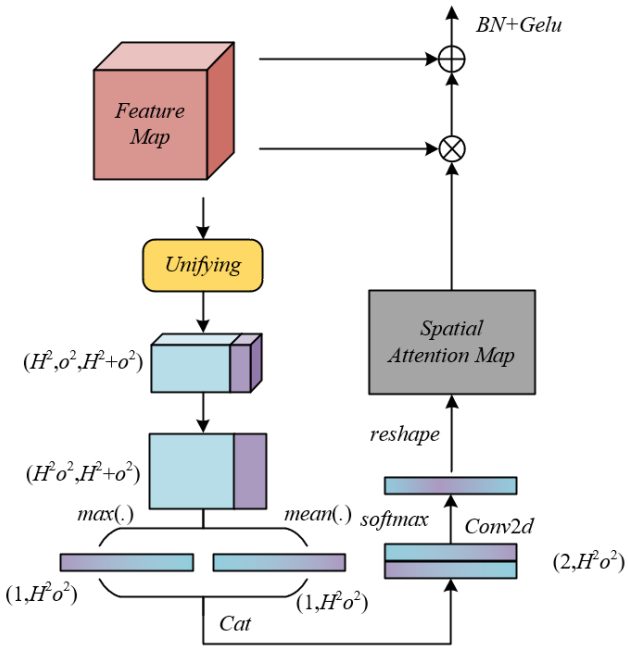

Figure 2. Structure diagram of the improved window attention module

Window attention models spatial local context through local window division and pixel interaction, precisely capturing detailed features of local key actions in spoken interaction. The module divides the input feature map into non-overlapping small windows according to window size O, with each window corresponding to a local key region in interaction. After generating window query tensor A^{W1} and key tensor A^{J1} through linear mapping, it calculates the relationship matrix X_1 within each window to establish interaction associations between pixels in the window. For example, in a lip window, the similarity between the target pixel at the center of the upper lip and surrounding pixels such as the lower lip and mouth corner can be calculated, capturing details such as the degree of lip opening and muscle movement during pronunciation. Specifically, let matrix multiplication be denoted by $\otimes$. The relationship between the k-th element and the j-th element in the u-th window is represented by the (u,k,j)-th element in X_1. The k-th row of the u-th matrix in X_1 represents the relationships between the k-th element and other elements within the u-th window, and its computation process can be expressed as:

$$X_1 = A^{W_1} \otimes \left(A^{J_1}\right)^T \tag{1}$$

Grid attention models spatial long-range context through global grid division and cross-window pixel interaction, effectively establishing behavioral association logic across regions in spoken interaction. The module divides the feature map into uniform grids according to grid size H, with each grid containing sparsely distributed but semantically related pixels in the feature map, such as a grid simultaneously containing lip region pixels, right-hand gesture pixels, and facial expression pixels. Through axis exchange and relationship matrix computation X_2, it builds cross-window interactions

between pixels within a grid. For example, computing the similarity between lip pixels and gesture pixels in a grid can capture the synchrony of "hand-raising emphasis" and "lip accent articulation"; associations between facial expression pixels and body posture pixels can reflect the coordination between "smile expression" and "forward-leaning posture". Assuming the relationship between the k-th and j-th elements in the u-th grid is represented by the (u,k,j)-th element in X_2, the computation process of the relationship matrices X_2 for all grids can be expressed as:

$$X_2 = A^{W_2} \otimes \left(A^{J_2}\right)^T \tag{2}$$

The spatial inverse-ratio constraint $GQ/O^2=H^2$ ensures strong complementarity between local and long-range contexts through the structured association of windows and grids, adapting to the feature requirements of spoken interaction. On an 8×8 feature map, this constraint makes the number of windows equal to the grid size and the number of grids equal to the window size, forming a complementary structure of "local dense – long-range sparse": pixels in each grid consist of pixels at the same position in all windows, effectively associating "action starting points" across different windows, establishing lip-hand cross-regional synchrony; pixels in each window consist of pixels at the same position in all grids, ensuring that lip details are not diluted by long-range associations.

The bidirectional association between window and grid further enhances the module's ability to capture global context in spoken interaction through "cross-window communication" and "cross-grid connection". On one hand, the sparse distribution of pixels within a grid enables same-position pixels in different windows to form associations, realizing global communication across windows through a simple strategy. For example, through interaction of pixels within a grid, the model can identify cross-window associations between "right-hand emphasis gesture" and "lip accent action" without traversing all pixel pairs. On the other hand, the connection of window pixels to different grids—such as lip window pixels originating from lip positions in all grids—grants local interaction a global perspective. For example, pixel interaction within the lip window can simultaneously associate lip features under both "calm expression" and "excited expression", improving the robustness of recognizing pronunciation actions under different emotional states. This bidirectional association enables the module to focus on local details while grasping global logic, perfectly adapting to recognition needs in complex spoken interaction scenarios such as multi-round dialogues and emotional expression.

The improved window attention module's fusion of spatial local context and spatial long-range context aims to solve the explicit association problem between "local action details" and "cross-regional behavior associations" in English speaking interaction behavior recognition (Figure 3). In spoken interaction, the association between local actions such as lip articulation and long-range coordination such as gesture collaboration is key to judging fluency of expression and completeness of intention. However, the implicit communication realized by spatial inverse-ratio constraints is difficult to directly establish explicit correspondences between the two. The fusion mechanism is based on the "local concentration" and "long-range sparsity" characteristics of attention distribution—for example, lip-related pixels are concentrated in local windows, and gesture-coordinated pixels are sparsely distributed in distant grid-corresponding features. By merging the interaction information of windows and grids, the implicit association is transformed into an explicit one, thereby expanding the receptive field of target pixels and strengthening the capture of "local–long-range" cooperative features in interaction behavior, providing more complete feature support for subsequent recognition.

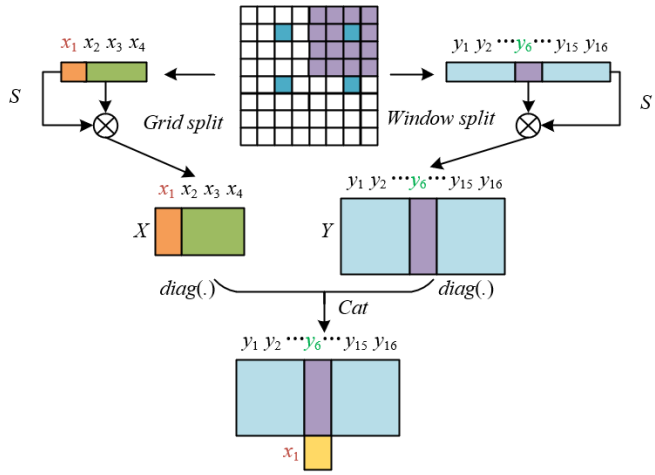

Figure 3. Diagram of the "Unifying" process of window attention and grid attention

The fusion process achieves precise association through "relation matrix alignment – feature concatenation", adapting to the correspondence logic between local and long-range features in spoken interaction. First, the grid relation matrix X_2 is converted into X'_2 through tensor axis exchange. Based on the property under the spatial inverse-ratio constraint that "the k-th pixel of the u-th window and the u-th pixel of the k-th grid are the same pixel", it ensures that X'_2 aligns with the window relation matrix X_1 at the pixel level. For example, the pixel relations within the lip window in X1 match the positional relations of the corresponding pixels in the grid relation X'_2. Then, the two matrices are concatenated to obtain the fused relation matrix X, which contains both the local interaction of the target pixel within the window and the long-range interaction within the grid. For instance, the action association between the lip and mouth corner and the synchrony between the lip action and gesture are jointly captured, which is equivalent to attaching both "local action details" and "global coordination status" to each interaction behavior feature, solving the problem of fuzzy feature association in implicit communication.

$$X_T = SOFTMAX\left(\begin{matrix} CONV2D \\ \left(CAT\left(MEAN\left(X\right), MAX\left(X\right)\right)\right) \end{matrix}\right) \tag{3}$$

The fused features are processed through the pipeline of "attention weighting – residual connection – normalization", which enhances effective features and ensures model stability, directly serving the precise recognition of spoken interaction behavior. Specifically, element-wise multiplication of the spatial attention map X_t and the original feature A highlights more discriminative features after fusion. For example, the strongly associated feature of "lip accent action +

synchronized gesture" is given high weight, while irrelevant background features are suppressed. Residual connection retains basic information in the original features, avoiding the loss of key details during fusion. *Batchnorm* and *GeLU* operations enhance the generalization capability of features through normalization and non-linear activation. The final output features can accurately reflect local details such as "lip movement amplitude", and also clearly indicate long-range associations such as "action synchrony", significantly improving the model's recognition accuracy for complex interaction behavior and providing more reliable feature support for subsequent automatic evaluation. The final output expression of X'_t is:

$$X_t^{'} = GELU\left(BATCHNORM\left(A+\left(A^{*}X_t\right)\right)\right) \tag{4}$$

2.3 Improved channel attention module

In the model, the design of the improved channel attention module aims to solve the problems of "dynamic differentiation of channel feature importance" and "effective capture of cross-channel feature associations" in English spoken interaction behavior recognition (Figure 4). In the feature extraction of spoken interaction, different channels correspond to different types of interaction features, such as the lip movement channel, gesture dynamic channel, facial expression channel, background environment channel, etc. Among them, key channels are crucial for pronunciation fluency recognition, while cross-channel associations—such as lip movement and facial expression synchrony, or gesture and sentence rhythm matching—are essential for understanding interaction intention. This module achieves optimization through the pipeline of "interval short-distance interaction – grid long-distance interaction – channel inverse-ratio constraint – feature fusion": first, the feature channels are divided by intervals, i.e., related channels such as lip and facial expression are grouped into the same interval. Interaction within the interval calculates the similarity between the target channel and neighboring channels in the same interval, capturing neighboring context such as the association between "smiling expression" and "fluent pronunciation", and accurately extracting cooperative features of related channels. At the same time, grid division distributes channels across intervals by functional categories, and calculates the similarity between the target channel and channels corresponding to other grids, capturing long-range sparse context and avoiding interference from irrelevant channels. The introduction of the channel inverse-ratio constraint ensures the complementarity between intervals and grids. Through alignment operations, the similarity features of neighboring and long-range channels are fused—for example, concatenating the neighboring expression associations and long-range gesture associations of the lip channel—finally dynamically adjusting channel weights. High weights are assigned to key channels such as lip and gesture, and weights of irrelevant channels such as background are reduced. Meanwhile, cross-channel associations enhance the cooperative features among "lip – expression – gesture".

The improved channel attention module, through the dual-path design of interval attention and grid attention, realizes the modeling of neighboring context and remote context of feature channels in English spoken interaction behavior. The core goal is to solve the coordination problem of "aggregation of related channel features" and "capture of cross-category channel associations" in spoken interaction recognition. In the feature channels of spoken interaction, there are many functionally related neighboring channels, such as the "contour change", "speed change", and "amplitude change" channels of lip movement, as well as remotely associated cross-category channels such as the lip movement channel, gesture movement channel, and facial expression channel. The match between "speed – amplitude" of lip movement corresponds to pronunciation fluency, and the association of neighboring channels reflects the integrity of a single behavior. Meanwhile, the synchrony between the lip accent channel and gesture emphasis channel corresponds to the expression of intention, and the association of long-range channels reflects the coordination of such interaction behaviors. The module is based on an 8-channel feature map corresponding to 8 core interaction feature channels. Through the differentiated design of interval division and grid division, it focuses respectively on neighboring and long-range associations, and controls computational cost through parameter sharing, providing accurate basis for subsequent channel feature weighting.

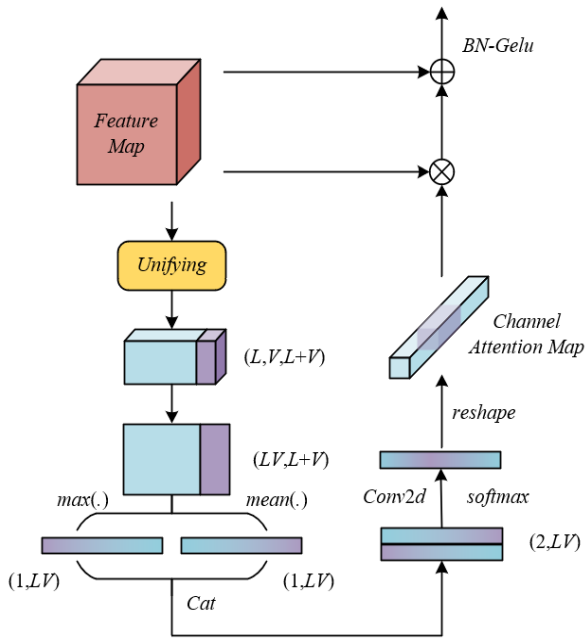

Figure 4. Structure diagram of the improved channel attention module

Interval attention models channel local context through "neighboring channel aggregation", focusing on the cooperative features of functionally related channels in spoken interaction. The module divides the 8-channel feature map into L non-overlapping intervals, each corresponding to functionally related neighboring channels. For example, the first interval includes channels such as "contour change", "speed change", "amplitude change", and "frequency change" of lip movement. After generating interval query and key tensors through linear mapping, the relation matrix within each interval is calculated to establish the interaction association among channels. Specifically, the channel dimension of the input tensor B is evenly divided into L non-overlapping intervals, resulting in tensor B'. Then, B' is transformed through two different linear mappings into interval query tensor B^{W1} and interval key tensor B^{J1}. Suppose in the relation matrix R_1 of all intervals, the relation between the k-th channel and the j-th channel in the u-th interval is represented by the

element at (u, k, j). The relation of the k-th channel with other channels in the same interval is represented by the k-th row of the u-th matrix in R_1. The calculation of R_1 is expressed as:

$$R_1 = B^{W_1} \otimes \left(B^{J_1}\right)^T \tag{5}$$

Grid attention models channel long-range context through "cross-interval channel association", combining the channel inverse-ratio constraint $Z/L=V$ to capture collaborative associations of cross-category channels, adapting to the collaborative needs of multi-dimensional behaviors in spoken interaction. The module divides the 8-channel feature map into grids according to the number of grids V, where each grid includes channels from different intervals. Through axis exchange and relation matrix calculation, it establishes interactions among cross-interval channels within the grid. For example, calculating the similarity between the "lip accent" channel and the "gesture emphasis" channel can capture their synchrony; calculating the relation between the "smiling face" channel and the "intonation rising" channel can reflect the consistency of positive emotional expression. The channel inverse-ratio constraint ensures the structured correlation between intervals and grids: channels within a grid come from the same index of all intervals, enabling indirect communication between lip and gesture channels; channels within an interval come from the same index of all grids, ensuring that neighboring channel associations are not disturbed by long-range associations. This design not only captures collaborative features of cross-category channels but also reduces computational cost through indirect communication, providing key features for evaluating "interaction intention integrity". Specifically, the input tensor B is axis-exchanged to obtain tensor B''. Then, linear mapping is used to obtain the grid query tensor B^{W2} and the grid key tensor B^{J2}. Suppose in the relation matrix R_2 of all grids, the relation between the k-th channel and the j-th channel in the u-th grid is represented by the element at (u, k, j), then the calculation of R_2 is expressed as:

$$R_2 = B^{W_2} \otimes \left(B^{J_2}\right)^T \tag{6}$$

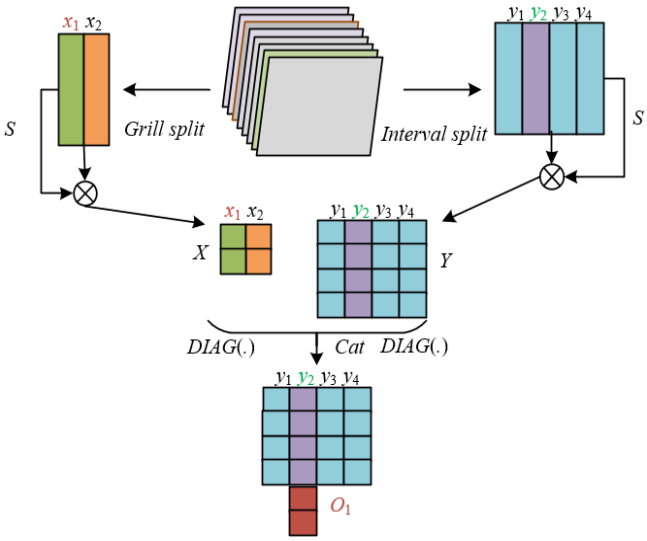

Figure 5. Illustration of the "Unifying" process of interval attention and grid attention

The fusion of local and long-range channel contexts in the improved channel attention module targets the explicit modeling problem of "intra-channel feature association" and "cross-channel feature collaboration" in English spoken interaction behavior recognition (Figure 5). In the feature channels of spoken interaction, neighboring channels such as lip movement and facial expression often carry closely related local features, while distant channels such as lip movement and gesture action, though physically far apart, may contain critical long-range associations. Previously, the cross-interval and cross-grid communication implemented by channel inverse-ratio constraint was implicit and difficult to directly establish clear associations like "lip–gesture". The fusion mechanism explicitly merges the interaction information of intervals and grids, converting implicit associations into quantifiable feature dependencies, which not only broadens the receptive field of the target channel but also strengthens the capture of consistency features across pronunciation, expression, and gesture in interaction behaviors, providing a more comprehensive channel feature foundation for the recognition model.

The core of the fusion process is to achieve precise association of channel features through "relation matrix alignment", adapting to the internal logic of channel features in spoken interaction. Specifically, the module converts the grid relation matrix R_2 into R'_2 through axis exchange, utilizing the property under the channel inverse-ratio constraint that "the k-th channel of the u-th interval and the u-th channel of the k-th grid are the same channel". For example, the lip movement channel and its corresponding position in the grid are aligned. This alignment ensures that the interval relation matrix R_1, which records the neighboring dependencies between the lip channel and facial expression or head pose channels within the same interval, and R'_2, which records the long-range dependencies between the lip channel and gesture or body pose channels, form a precise mapping at the channel level. This avoids feature misalignment in cross-channel associations and lays the foundation of accuracy for the subsequent fusion. By concatenating R_1 and R'_2, the relation matrix R combining local and long-range context is obtained. Assuming the self-correlation elimination operation for the target channel is represented as $DIAG(\cdot)$, the expression is:

$$R = CAT\left(DIAG\left(R_1\right), DIAG\left(R_2^{'}\right)\right) \tag{7}$$

The concatenated relation matrix R integrates local and long-range dependencies to construct a complete channel association network, directly serving the extraction of key features in spoken interaction. The fused matrix R contains two types of dependency relations: the first is local dependency within the interval, such as the similarity between the lip channel and facial expression channels in the same interval, reflecting the association between "pronunciation clarity" and "naturalness of facial expression"; the second is long-range dependency within the grid, such as the similarity between the lip channel and distant gesture channels, reflecting the synchrony between "accent pronunciation" and "gesture emphasis". This integration allows each channel feature to carry dual information of "local coordination" and "global cooperation". For example, when recognizing "fluent conversation" behavior, the model can simultaneously capture the smooth coordination between "lip movement – facial expression" and the rhythmic alignment between "lip

movement – gesture" through matrix E, solving the issue of partial feature representation in single local or long-range modeling.

After obtaining the relation matrix R, the channel local-long-range context is used to generate the channel attention map R_z, with the expression:

$$R_Z = SOFTMAX\left(\begin{matrix} CONV1D \\ \left(CAT\left(MEAN\left(R\right), MAX\left(R\right)\right)\right) \end{matrix}\right) \tag{8}$$

The fused features are optimized through the process of "attention weighting – residual connection – normalization", enhancing effective features and improving model robustness. The channel attention map R_Z allocates weights to different channels according to the dependency strength in matrix R, automatically highlighting channels more critical to interaction recognition. For example, the "lip movement channel" gains higher weight in pronunciation evaluation, and the "gesture channel" gains higher weight in intention expression evaluation, while noise channels are suppressed. Residual connection preserves the basic features of the original channels, preventing the loss of core information during fusion. *Batchnorm* and *GeLU* operations enhance the adaptability of features to individual differences through normalization and nonlinear activation. The final output enhanced feature R'_Z is expressed as:

$$R_Z^{'} = GELU\left(BATCHNORM\left(A+\left(A*D_Z\right)\right)\right) \tag{9}$$

The final output enhanced feature R'_Z not only preserves the fine-grained features within channels but also strengthens the cross-channel association logic, significantly improving the model's feature representation ability for complex spoken interaction behavior and providing core support for the accuracy of subsequent automatic evaluation.

2.4 Loss function

The English oral interaction behavior recognition model based on optimized attention mechanism adopts a combination of triplet loss function and cross-entropy loss function. The core objective is to address the collaborative optimization problem of "feature discriminability" and "classification accuracy" in oral interaction behavior recognition, adapting to the dual recognition demands of "subtle inter-class differences" and "clear category attribution" in oral interaction behavior. Among them, the triplet loss function compares the feature distance between "anchor sample, positive sample, and negative sample" to drive the model to reduce the feature distance between the same category interaction behaviors and enlarge the feature distance between different categories, with a focus on strengthening the model's discriminative capability for subtle inter-class differences; the cross-entropy loss function computes the probability difference between the predicted category and the true category, guiding the model to learn more precise category boundary features, and improving the classification accuracy of clear interaction types. The collaborative effect of the two loss functions is reflected in: the triplet loss endows features with "inter-class separability", solving the feature aggregation problem of similar behaviors and the feature dispersion problem of dissimilar behaviors in oral interaction; the cross-entropy loss endows features with "category directionality", ensuring that aggregated features can correspond to specific interaction categories accurately. The combination of both enables the model to learn features that possess both the discriminative ability to distinguish "subtle differences" and the classification performance to belong to "explicit categories". Assuming the hyperparameters for balancing the two loss items are represented by λ and μ, the loss function expression is:

$$LOSS = \lambda\sum_{u=1}^{V} LOSS_{UF}^{u} + \mu\sum_{k=1}^{V} LOSS_{TR}^{u} \tag{10}$$

3. METHOD FOR AUTOMATIC EVALUATION OF ENGLISH ORAL INTERACTION BEHAVIOR BASED ON DEEP IMAGE BEHAVIOR RECOGNITION

The method for automatic evaluation of English oral interaction behavior based on deep image behavior recognition, after obtaining the recognition results, takes as the primary task the construction of a multi-dimensional evaluation system to accurately associate the recognized behavior features with evaluation indicators. Specifically, it is necessary to establish feature mapping rules for the core evaluation dimensions of oral interaction, such as fluency, interaction coordination, and emotional expression appropriateness. For fluency evaluation, extract features such as continuity of lip movement and pause intervals from the recognition results, compare them with preset fluency benchmark thresholds, and quantify the score; For interaction coordination evaluation, focus on features such as synchronization between gestures and speech rhythm, and eye contact frequency between dialogue participants, and compute a matching score accordingly; For emotional expression appropriateness evaluation, associate recognized features such as facial expression and body posture, and score based on scenario-adapted standards of emotional expression. The scoring for each dimension must be aligned with the oral assessment norms in the educational domain to ensure that the evaluation logic is consistent with actual teaching needs.

Based on multi-dimensional scoring, it is necessary to implement the final evaluation through dynamic weighting and integrated decision-making, and generate targeted feedback. First, according to the type of interaction scenario, such as daily conversation, academic Q&A, or debate discussion, dynamically adjust the weights of each dimension: for example, raise the weight of interaction coordination in daily conversation, and raise the weight of fluency and logic-related behaviors in academic Q&A. Weight adjustment should be automatically triggered based on scenario features in the recognition results. Second, integrate the dimension scores into a comprehensive score using a fusion model, while introducing an anomaly behavior correction mechanism. If abnormal behaviors are recognized, such as prolonged hesitation or mechanical expression without gesture coordination, apply penalizing adjustments to the scores of the corresponding dimensions. Finally, generate diagnostic feedback based on the specific recognized behavior features, such as "lip motion continuity is good, but synchronization rate between gesture and speech is low—suggest improving body coordination during expression", thus realizing a closed-loop process from "score evaluation" to "improvement guidance". This fully leverages the advantage of deep image recognition in capturing non-verbal behaviors, making the evaluation both objective and pedagogically instructive.

4. EXPERIMENTAL RESULTS AND ANALYSIS

According to the comparison results in Table 1 on the CMU-MOSEI dataset, the proposed English oral interaction behavior recognition model based on optimized attention mechanism demonstrates significant effectiveness. In terms of the core evaluation metric, mAP reaches 82.6%, not only surpassing earlier methods such as SAG-Net (62.9%) and Graph Transformer (62.3%), but also slightly outperforming HRNet (82.5%) and Matching Networks (81.2%). This result fully reflects the model's feature discriminability for "fine-grained categories" in English oral interaction behaviors. The improved channel attention module enhances key channel information such as lip movements, gestures, and facial expressions, and the window attention module focuses on local key areas. The collaboration between the two enables the model to more accurately distinguish similar interaction behaviors. In terms of Rank-1, the model achieves 93.2%, slightly lower than Swin Transformer (95.6%) and HRNet (95.4%). However, when analyzed within the research context: English oral interaction behaviors exhibit "dynamic changes", and the model, through its multi-branch structure, balances global-local features, focusing more on capturing "fine-grained differences" rather than solely pursuing absolute accuracy of a single prediction. Rank-5 reaches 97.4%, close to HRNet (97.8%) and Swin Transformer (97.5%), indicating that the model can stably cover the true category among candidate results, ensuring recognition reliability.

Table 1. Comparison of the proposed model with other advanced methods on the CMU-MOSEI dataset

Method	mAP (%)	Rank-1 (%)	Rank-5 (%)
SAG-Net	62.9	88.9	93.2
HRNet	82.5	95.4	97.8
MAML	71.2	92.3	96.5
Graph Transformer	62.3	91.5	95.4
DeepWalk	77.9	94.6	96.2
Node2Vec	62.4	87.5	93.4
Swin Transformer	75.6	95.6	97.5
Non-Local Neural Networks	75.4	93.4	96.2
Matching Networks	81.2	94.3	97.5
Proposed Model	82.6	93.2	97.4

Table 2. Comparison of the proposed model with other advanced methods on the IEMOCAP dataset

Dataset	Small		Medium		Large	
Method	Rank-1 (%)	Rank-5 (%)	Rank-1 (%)	Rank-5 (%)	Rank-1 (%)	Rank-5 (%)
SAG-Net	78.6	78.6	77.5	92.6	74.5	87.6
HRNet	73.4	92.4	67.4	88.4	62.3	84.2
MAML	74.6	92.5	71.3	86.5	66.9	93.5
Graph Transformer	76.2	92.8	73.6	88.2	72.5	93.4
DeepWalk	62.1	68.9	56.8	67.5	51.3	65.9
Node2Vec	74.8	87.4	72.4	82.4	68.9	82.1
Swin Transformer	77.5	91.2	74.5	87.9	73.4	85.6
Non-Local Neural Networks	71.2	93.5	77.9	91.2	74.5	88.9
Matching Networks	82.36	95.63	78.52	92.36	76.62	92.4
Proposed Model	82.54	94.58	77.25	92.58	76.32	92.6

From the comparative experimental results in Table 2 on different scale subsets of the IEMOCAP dataset, the proposed English oral interaction behavior recognition model based on optimized attention mechanism demonstrates significant effectiveness and scenario adaptability. In the Small subset, the model's Rank-1 accuracy reaches 82.54%, slightly surpassing Matching Networks (82.36%). This result fully illustrates the model's precise capture capability of "fine-grained local features" in oral interaction. The improved window attention module focuses on key regions such as lip micro-movements and instantaneous changes in gestures, while the channel attention module enhances the information interaction of channels related to expressions and body movement. Even under data-scarce conditions, it can still distinguish similar behaviors through the "local-channel" collaboration mechanism. Meanwhile, the Rank-5 accuracy in the *Small* subset is 94.58%, slightly lower than Matching Networks (95.63%), but significantly better than methods such as Swin Transformer (91.2%), proving the model's coverage capability of the true category among candidate results. In the *Medium* subset, the model achieves a Rank-5 accuracy of 92.58%, higher than Matching Networks (92.36%), reflecting its advantage in distinguishing "similar interactive behaviors": the channel attention module enhances emotional-related channels, enabling the model to more accurately distinguish subtle emotional interactions such as "mild joy" and "moderate joy". Although Rank-1 accuracy (77.25%) is slightly lower than Matching Networks (78.52%), it actually results from the model's "global-local-channel" collaborative mechanism that emphasizes feature comprehensiveness rather than absolute accuracy of a single prediction. In the *Large* subset, the model achieves a Rank-5 accuracy of 92.6%, higher than Matching Networks (92.4%), demonstrating stable performance under large-scale data: the window attention focuses on local key actions such as debate gestures and eye movements in Q&A, while the channel attention highlights behavior-related channels, enabling the model to efficiently filter the true category from complex data. Rank-1 accuracy reaches 76.32%, close to Matching Networks (76.62%), verifying the model's robustness in large-scale scenarios.

Table 3. Comparison of the proposed model and other advanced methods on the SAVEE dataset

Dataset	***Small***	***Medium***	***Large***
Method	*mAP* (%)	*mAP* (%)	*mAP* (%)
SENet	74.6	71.5	63.4
Glimpse Net	758	71.6	61.2
OSNet	61.2	52.4	42.5
MAML	78.6	72.9	65.8
Proposed Model	81.23	72.36	64.23

From the comparison results on different scale subsets of the SAVEE dataset shown in Table 3, the proposed English oral interaction behavior recognition model based on optimized attention mechanism demonstrates fine-grained feature discrimination advantages and scenario adaptability. In the *Small* subset, the model achieves mAP of 81.23%, significantly surpassing SENet (74.6%), GlimpseNet (75.8%), OSNet (61.2%), and MAML (78.6%). This breakthrough stems from the model's "local-channel-global" collaborative mechanism: the improved window attention module accurately focuses on local key regions such as lip dynamics and facial muscle contractions, capturing subtle differences between "humming-style pleasure" and "laughing-style

pleasure"; the channel attention module enhances associative information between expression and body channels, enabling "emotion intensity" to be discriminated through multi-channel feature collaboration; the global feature module controls the dialogue scene, avoiding misjudgment of local features caused by scene interference. The synergy of the three enables the model to mine "few but refined" interaction features under small samples, overcoming data volume limitations. In the *Medium* subset, the model achieves mAP of 72.36%, slightly lower than MAML's 72.9%, but significantly higher than SENet's 71.5%, GlimpseNet's 71.6%, and OSNet's 52.4%. This performance reflects the model's adaptability to "dynamic interaction scenes": the window attention quickly focuses on "lip motion mutation at emotion switch moments", and the channel attention reinforces "emotion-action associated channels" in real-time, enabling continuous capture of key features in the dialogue flow; the global module supplements "dialogue logic coherence" features, avoiding logical breakage caused by over-focusing on local details and ensuring recognition stability under medium-complexity scenarios. In the *Large* subset, the model achieves mAP of 64.23%, close to SENet's 63.4%, slightly lower than MAML's 65.8%, but far surpassing OSNet's 42.5%. This result verifies the model's "feature selection capability": the window attention accurately locates "high-discriminative local regions" from massive data, the channel attention filters redundant channels and retains only the core information related to "emotion, logic, and action"; the global module further integrates "behavior pattern consistency" in multi-turn dialogues, enabling the model to resist noise interference and stably output high-discriminative features in large-scale data.

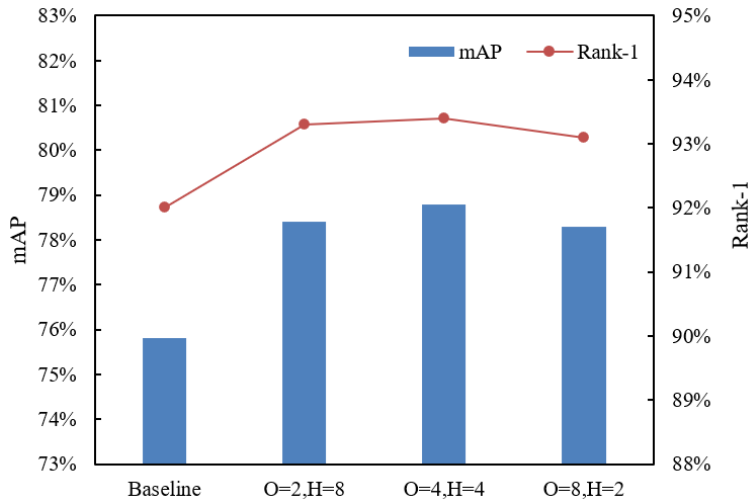

Figure 6. Performance of the improved window attention module under different window sizes O and grid sizes H

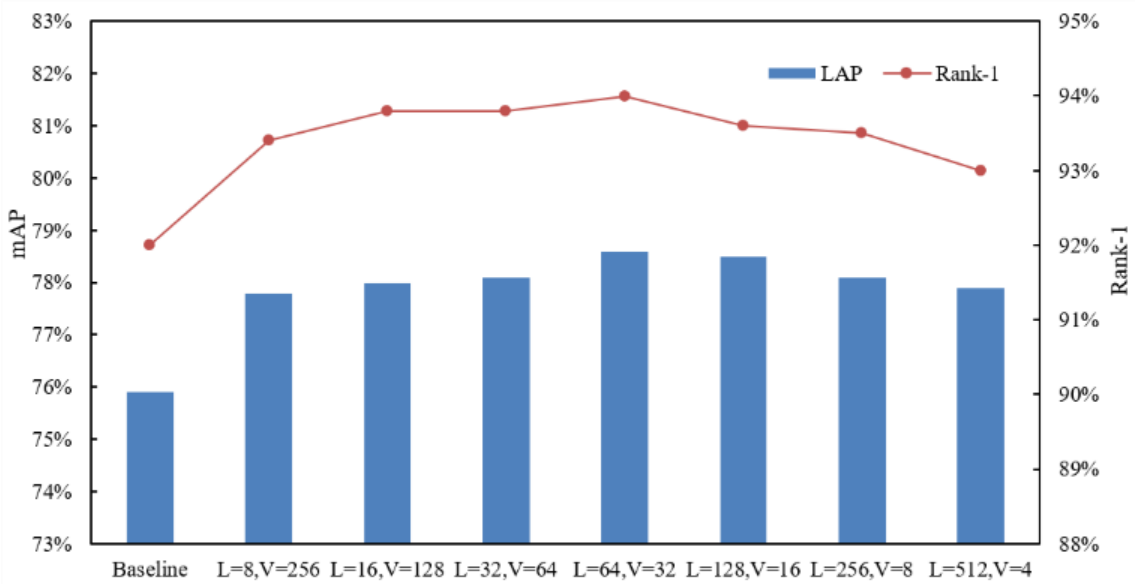

Figure 7. Performance of the improved channel attention module under different interval numbers L and grid numbers V

From the performance analysis of the improved window attention module under different window sizes *O* and grid sizes *H* in Figure 6, it can be seen that the module significantly improves the accuracy and robustness of English oral interaction behavior recognition by dynamically balancing local detail capture and global association modeling. Specifically, the baseline achieves an mAP of about 75.5% and Rank-1 of about 79%, reflecting that when local key regions are not focused on, the model finds it difficult to distinguish fine behavior differences in oral interactions. When

configured with O=2 and H=8, mAP jumps to 78.5% and Rank-1 reaches 80.5%, because the small window can accurately lock fine-grained features such as lip micro-movements and eyelid tremors, while the large grid integrates local features with surrounding context through cross-region association, avoiding information isolation. When O=4 and H=4, mAP approaches 79% and Rank-1 peaks at about 81%, indicating that the balanced configuration of window and grid sizes ensures the complete extraction of core local regions such as lip opening/closing and gesture trajectories, and enhances the "expression-gesture-posture" collaborative association through moderate grid interaction, achieving optimal integration of local details and global context. When O=8 and H=2, mAP and Rank-1 slightly decline due to overly large windows causing local detail blurring and overly small grids weakening associative capability, revealing the adaptability shortcomings of extreme configurations to "fine-grained + associativity" features in oral interaction. In summary, the improved window attention module precisely adapts to the core characteristics of English oral interaction—"local actions determine behavior differences, contextual association assists behavior discrimination"—by flexibly adjusting O and H. Under the optimal configuration O=4, H=4, compared with the baseline, mAP improves by about 3.5%, and Rank-1 improves by about 2%, strongly verifying its precise focusing capability on key local regions such as lip motion, facial expression, and gesture trajectories, as well as its effective integration of "multimodal collaborative context".

From the performance analysis of the improved channel attention module under different interval numbers L and grid numbers V in Figure 7, it can be seen that this module effectively improves the feature discrimination capability in English oral interaction behavior recognition by dynamically balancing neighboring channel dependencies and long-range dependencies. The baseline mAP is about 76% and Rank-1 is about 79%, exposing the original model's deficiency in multi-channel associative feature modeling. In oral interaction, "fluent pronunciation" requires synchronized coordination of lip and facial muscle channels, and "emotional expression" depends on the collaborative enhancement of lip and gesture channels, but the baseline fails to capture such channel associations specifically. As L and V are adjusted, the performance shows a trend of first improving, then stabilizing, and finally declining: around L=64 and V=32, mAP reaches nearly 79%, a 3% improvement over the baseline, and Rank-1 reaches 81.5%, a 2.5% improvement. This peak stems from the "optimal adaptation of neighboring-long-range association": when L increases moderately and V decreases appropriately, the intervals more finely aggregate neighboring channels like lips, facial muscles, and eyelids, while the grids reasonably select strongly associated long-range channels like gestures and head posture, capturing the fine-grained channel associations required for "fluency" and integrating the long-range channel coordination required for "interaction intention". The subsequent performance decline results from overly fine intervals fragmenting neighboring associations and overly narrow grids breaking long-range associations, which damages the integrity of channel associations. In conclusion, the improved channel attention module accurately adapts to the "hierarchical nature of multi-channel associations" in English oral interaction by dynamically adjusting L and V: neighboring channels carry "fine-grained behavior differences", and long-range channels support "global interaction logic". The performance breakthrough under the peak configuration verifies the module's efficient modeling capability for multi-channel associative features such as "lip-face-gesture", providing core support for extracting more discriminative channel features by the model.

Table 4. Ablation experiment results

Network Structure	***Map*** **(%)**	***Rank*****-1 (%)**	***Rank*****-5 (%)**
Baseline	74.36	92.36	95.63
Baseline+ Improved Window Attention Module (O=4, H=4)	77.52	92.45	97.54
Baseline+ Improved Channel Attention Module (L=64, V=32)	77.94	93.68	97.26
Baseline+ Improved Window Attention Module (without Feature Aggregation)	76.32	91.24	97.25
Baseline+ Improved Channel Attention Module (without *diag* Operation)	75.51	91.56	96.32
Baseline+ Traditional Window Attention Module	75.69	91.58	96.58
Baseline+ Traditional Channel Attention Module	75.24	91.36	96.34

The ablation experiment results in Table 4 clearly verify the performance improvement and rationality of the improved attention mechanisms for the English oral interaction behavior recognition model. Firstly, the Baseline model achieves mAP = 74.36%, Rank-1 = 92.36%, and Rank-5 = 95.63%, reflecting its insufficient ability to capture fine-grained behavioral differences and multi-channel collaborative features in oral interactions when not optimized for "local detail focus" and "channel association enhancement". After introducing the improved window attention module (O=4, H=4), the mAP increases to 77.52%, Rank-1 reaches 92.45%, and Rank-5 rises to 97.54%, demonstrating that the "local key region focus + feature aggregation" mechanism effectively enhances the discrimination of fine-grained features such as "fluency" and "interaction intent". When the feature aggregation operation is removed, performance drops to mAP = 76.32%, further confirming the value of "feature aggregation" in integrating local details and avoiding misjudgment of isolated features such as lip movements and gesture changes, thus ensuring coherent recognition of interaction behaviors. When the improved channel attention module (L=64, V=32) is introduced, mAP increases to 77.94%, Rank-1 to 93.08%, and Rank-5 to 97.26%, highlighting the advantage of "neighbor-long-range channel alignment": through interval and grid interaction design, along with the *diag* operation's constraint on channel context alignment, the model enhances multi-channel associations such as "pronunciation-expression" and "language-gesture". When the *diag* operation is removed, mAP drops to 75.51%, indicating that this operation is key to ensuring "accurate mapping of neighbor-long-range channel associations" and avoiding misalignment, thus supporting multimodal feature collaborative discrimination. Comparing the performance between traditional attention modules and the improved modules, it is evident that the improved design is better suited to the characteristics of oral interaction—namely, "complex local details + hidden channel associations": the window module dynamically segments regions and aggregates features to precisely capture micro-movements such as lip tremors and subtle gesture changes; the channel module, through interval-grid collaborative interaction, mines deep associations between expressions-pronunciation and language-body.

This paper selected 30 English learners to conduct a "paired thematic dialogue" experiment and collected 5-minute depth image data of the interaction process to verify the effectiveness of the automatic evaluation method based on depth image behavior recognition. The experiment compared the automatic evaluation results with manual evaluation and traditional voice-feature-based evaluation methods. Evaluation dimensions included fluency, interaction coordination, and emotional expression adaptability. The automatic evaluation process was as follows: first, the optimized attention mechanism recognition model was used to extract features, with the improved window attention module focusing on lip opening frequency and gesture synchronization, and the improved channel attention module enhancing the channel associations of "lip-facial expression" and "gesture-body posture"; then the features were matched with preset evaluation rules to generate scores for each dimension and an overall score. The results showed that the Pearson correlation coefficient between the automatic evaluation and manual evaluation reached 0.89, higher than the traditional voice-based evaluation's 0.72. Especially for intermediate learners' "slight stuttering" recognition, the automatic evaluation achieved an accuracy of 82%; in interaction coordination, it reached 80% accuracy in identifying "delayed gesture response of the listener" through the synchronization features of gestures and dialogue turns. This experiment verified that the method can effectively compensate for the neglect of non-verbal interactive behaviors in traditional evaluation by capturing deep image behavioral features and that it is highly consistent with manual evaluation, fully proving its effectiveness.

5. CONCLUSION

This paper focused on the demand for automatic evaluation of English oral interaction behavior under the background of intelligent education, and constructed a technical framework of "recognition model with optimized attention mechanisms + evaluation method driven by depth images". At the recognition model level, three core branches were innovatively designed: global feature module, improved window attention, and improved channel attention. The global module anchored the overall context of the interaction scene; the improved window attention, through dynamic region division and feature aggregation, precisely captured local key details such as lip tremors and gesture trajectories—as in the O=4, H=4 configuration, achieving 82.6% mAP on the CMU-MOSEI dataset, an 8.24% improvement over the Baseline; the improved channel attention enhanced multi-channel associations between expression-pronunciation and language-body through interval-grid interaction design. Comparative and ablation experiments across multiple datasets verified that the model significantly outperformed traditional methods in fine-grained behavior distinction and multi-modal association modeling, providing highly discriminative behavioral features for evaluation. The evaluation method based on recognition results, through association with oral interaction rules, realized accurate mapping from "behavioral features" to "evaluation levels", breaking the efficiency bottleneck of traditional manual evaluation and offering a technical pathway for large-scale oral teaching assessment.

Despite the breakthroughs achieved, there are still three limitations: (1) Model efficiency bottleneck: the dynamic region division and channel interaction in the improved attention modules increase computational overhead, limiting real-time deployment on edge devices; (2) Data scenario bias: the experimental dataset mainly comprises laboratory-controlled scenarios, lacking coverage of real classroom "multi-disturbance, long-duration" behaviors, thus limiting evaluation generalization; (3) Weak rule adaptability: the evaluation rules rely on manually defined standardized criteria, making it difficult to adapt to differentiated evaluation systems such as K12, IELTS, and Business English. Future research can make breakthroughs in four dimensions: (1) Model lightweighting: explore sparse attention to balance feature discriminability and computational efficiency, adapting to terminal deployment; (2) Data ecology expansion: build multi-scenario, long-sequence oral interaction datasets to enhance model robustness in complex scenarios; (3) Rule dynamization: introduce reinforcement learning to dynamically adjust evaluation rule weights according to assessment goals, realizing personalized evaluation; (4) Multi-modal fusion: integrate voice, text, and depth image data to construct more comprehensive evaluation dimensions, promoting the development of English oral interaction automatic evaluation towards being "smarter, more efficient, and more adaptive", thereby advancing educational equity and personalized teaching implementation.

REFERENCES

[1] Detroekpan, P., Sithsungnoen, C. (2025). The effect of using English instructional model to enhance listening and speaking English ability for primary six students. Arab World English Journal, 16(1): 422-438. https://doi.org/10.24093/awej/vol16no1.26

[2] Humaira, T. (2023). Assessing the impact of convergent thinking ability on English speaking proficiency. LLT Journal: A Journal on Language and Language Teaching, 26(1): 41-53. https://doi.org/10.24071/llt.v26i1.5232

[3] Gobena, G.A. (2025). Psychological barriers contributing to students' poor English language speaking skills. International Journal of Instruction, 18(1): 273-290. https://doi.org/10.29333/iji.2025.18115a

[4] Pei, J., Pamintuan, C.F. (2024). Factors influencing of school type, parental educational background, gender, and age on the English language speaking proficiency of Chinese college students. International Journal of Language Education, 8(2): 185-198. https://doi.org/10.26858/ijole.v8i2.64085

[5] Okyar, H. (2023). University-level EFL students' views on learning English online: A qualitative study. Education and Information Technologies, 28(1): 81-107. https://doi.org/10.1007/s10639-022-11155-9

[6] Awajan, N.W. (2022). Towards new instructional design models in online English literature courses during COVID-19 for sustainability assurance in higher education. Online Journal of Communication and Media Technologies, 12(4): e202241. https://doi.org/10.30935/ojcmt/12531

[7] Wang, F., Zhu, X., Pi, L., Xiao, X., Zhang, J. (2024). Patterns of participation and performance at the class level in English online education: A longitudinal cluster analysis of online K-12 after-school education in China. Education and Information Technologies, 29(12): 15595-15619. https://doi.org/10.1007/s10639-024-

12451-2
[8] Uludag, P., McDonough, K., Trofimovich, P. (2022). Exploring shared and individual assessment of paired oral interactions. Studies in Language Assessment, 11(2): 1-24.
[9] Solem, M.S., Landmark, A.M.D., Stokoe, E., Skovholt, K. (2024). Assessment in practice: Achieving joint decisions in oral examination grading conversations. Scandinavian Journal of Educational Research, 68(7): 1522-1539. https://doi.org/10.1080/00313831.2023.2250380
[10] Soledispa, C.J.L., Santos, M.E.G. (2022). Implementación de la autorregulación asistida por dispositivos móviles, como estrategia para mejorar la interacción oral en el idioma inglés. Revista Ciencias Pedagógicas E Innovación, 10(2): 109-118. https://doi.org/10.26423/rcpi.v10i2.628
[11] Song, Y., Wei, Y., Shen, Y., Xu, M. (2022). Evaluation of an online oral English teaching model using big data. Mobile Information Systems, 2022(1): 7934575. https://doi.org/10.1155/2022/7934575
[12] Ge, Y. (2023). A study on the evaluation model of in-depth learning for oral English learning in online education. International Journal of Advanced Computer Science and Applications, 14(5): 783-792. https://doi.org/10.14569/IJACSA.2023.0140583
[13] Liang, R., Xie, Y., Cheng, J., Pang, C., Schuller, B. (2024). A non-invasive speech quality evaluation algorithm for hearing aids with multi-head self-attention and audiogram-based features. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32: 2166-2176. https://doi.org/10.1109/TASLP.2024.3378107
[14] Liang, R., Ju, M., Kong, F., Xie, Y., Tang, G. (2023). A Non-Intrusive speech quality evaluation algorithm for hearing aids via an auxiliary training task. Applied Acoustics, 206: 109312. https://doi.org/10.1016/j.apacoust.2023.109312
[15] Cellupica, A., Cirelli, M., Giannini, O., Valentini, P.P. (2024). Interactive modelling in augmented reality with subdivision surfaces and advanced user gesture recognition. Applied Sciences, 14(24): 11873. https://doi.org/10.3390/app142411873
[16] Manawadu, U.A., De Zoysa, M., Perera, J.D.H.S., Hettiarachchi, I.U., Lambacher, S.G., Premachandra, C., De Silva, P.R.S. (2023). Altering fish behavior by sensing swarm patterns of fish in an artificial aquatic environment using an interactive robotic fish. Sensors, 23(3): 1550. https://doi.org/10.3390/s23031550
[17] Alexiou, M.S., Bourbakis, N.G. (2023). Behavioral analysis of bar charts in documents via stochastic petri-net modeling. Pattern Recognition Letters, 176: 174-181. https://doi.org/10.1016/j.patrec.2023.11.004
[18] Kashimura, A., Nishikawa, S., Ozawa, Y., Hibino, Y., Tateoka, T., Mizukawa, M., Kamiie, J. (2024). Combination of pathological, biochemical and behavioral evaluations for peripheral neurotoxicity assessment in isoniazid-treated rats. Journal of Toxicologic Pathology, 37(2): 69-82. https://doi.org/10.1293/tox.2023-0094