IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 943-954
Journal homepage: http://iieta.org/journals/ts

Non-Contact Assessment of Social Anxiety in University Students via Video-Based Crowd Behavior Analysis

Lu Wang

Faculty of Humanities and Arts, Xi'an Fanyi University, Xi'an 710105, China

Corresponding Author Email: wl@xafy.edu.cn

https://doi.org/10.18280/ts.430231

Received: 30 October 2025
Revised: 7 February 2026
Accepted: 28 February 2026
Available online: 30 April 2026

Keywords:
crowd behavior analysis, social anxiety assessment, non-contact detection, adaptive graph convolution, psychology-driven attention mechanism, image processing

ABSTRACT

Traditional assessment of social anxiety among university students is inherently subjective and lacks continuous real-time monitoring. Consequently, the development of non-contact assessment approaches is critical in mental health monitoring. Video-based image analysis has emerged as an effective paradigm for modeling crowd behavior; however, existing methods have several limitations. Therefore, a novel framework integrating computer vision, graph neural networks, and psychologically grounded behavioral indicators was proposed for the non-contact quantitative evaluation of social anxiety in university students. An adaptive spatiotemporal graph convolutional network was designed, in which the adjacency matrix was dynamically learned through a weighted fusion of spatial proximity and feature similarity, enabling precise modeling of temporal–spatial interaction patterns within groups. Furthermore, a psychology-driven attention mechanism was introduced, whereby behavioral indicators associated with social anxiety were quantified and incorporated as attention-guiding signals, thereby enhancing the representation of salient behavioral features and improving model interpretability. In addition, a dedicated dataset of university student social behavior was constructed, comprising multi-scenario video data annotated with standardized social anxiety scale labels, thus addressing the lack of publicly available domain-specific resources. A multi-task learning framework was further developed to jointly optimize social anxiety score regression and behavioral indicator prediction, with a graph structure consistency loss imposed to regularize the learning process. The proposed approach provides a novel paradigm for dynamic modeling of crowd behavior and establishes an efficient, non-invasive pathway for mental health monitoring, demonstrating significant potential in the interdisciplinary domain of image processing and psychological health.

1. INTRODUCTION

Social anxiety represents a prevalent mental health concern among university student populations [1-3]. Conventional assessment approaches have predominantly relied on contact-based psychometric scales, which are characterized by strong subjectivity and limited capacity for continuous, real-time, and non-invasive monitoring [4, 5]. As a result, the development of non-contact assessment techniques has become an urgent priority in the domain of mental health evaluation. Video-based image analysis has been widely recognized as an effective approach for modeling crowd behavior [6, 7]. However, existing methods remain insufficient in capturing the dynamic nature of group social interactions and in establishing robust mappings between observable behavioral patterns and underlying psychological states. Consequently, their direct applicability to the quantitative assessment of social anxiety in university students remains constrained. To overcome these limitations, a novel technical framework is introduced, in which dynamic modeling of group interactions is advanced within the field of image processing, while high-precision non-contact assessment of social anxiety is simultaneously achieved. This approach provides a non-invasive technological foundation for mental health monitoring. Positioned at the intersection of image processing and psychological health, the proposed methodology aligns with the interdisciplinary innovation paradigm emphasized by leading international journals and demonstrates substantial academic significance and application potential.

Despite recent advances, existing video-based non-contact approaches for social anxiety assessment and crowd behavior analysis remain constrained by four critical limitations, which collectively hinder both evaluation accuracy and practical applicability. First, substantial deficiencies persist in modeling group interactions [8, 9]. Current graph convolution-based methods typically rely on predefined and static graph structures [10, 11], in which inter-individual relationships are constructed using simplified criteria such as spatial proximity [12]. Such approaches fail to dynamically adapt to the spatiotemporal evolution of social relationships within groups and are therefore unable to accurately capture variations in interpersonal affinity. As a result, the extracted group interaction features remain incomplete and insufficiently expressive. Second, the integration of psychological priors

remains inadequate. Most existing approaches are predominantly data-driven [13, 14] and do not explicitly incorporate behaviorally grounded indicators associated with social anxiety into image processing models. Consequently, insufficient emphasis is placed on behavior patterns that are strongly correlated with anxiety states, leading to limited task specificity and reduced interpretability. This limitation restricts their applicability in clinical and real-world settings, where interpretability is essential. Third, a lack of domain-specific datasets constitutes a significant bottleneck. Currently, dedicated datasets for group behavior analysis targeting social anxiety assessment in university students are largely unavailable [15, 16]. Existing datasets primarily focus on general crowd behavior analysis and do not establish precise correspondences between video-derived behavioral features and standardized social anxiety scale scores. This gap prevents the provision of reliable data support for model training and validation. Finally, model generalization remains limited. Most existing studies adopt single-task learning frameworks [17, 18], focusing primarily on social anxiety score regression [19]. Such approaches do not enable the joint optimization of behavioral feature extraction and anxiety assessment, resulting in suboptimal performance under complex campus scenarios involving variations in illumination, camera viewpoints, and crowd density. Consequently, the generalization capability of these models remains insufficient and requires further enhancement.

The objective of this study is to develop an integrated framework combining computer vision, graph neural networks, and psychologically grounded behavioral indicators to enable non-contact quantitative assessment of social anxiety in university students, while simultaneously improving evaluation accuracy and model interpretability. The principal innovations are summarized below. An adaptive spatiotemporal graph convolutional network is proposed, in which the adjacency matrix is dynamically learned through the fusion of spatial proximity and feature similarity, thereby overcoming the limitations associated with fixed graph structures. A psychology-driven attention mechanism is designed, wherein behavioral indicators associated with social anxiety are quantitatively encoded as guidance signals, enhancing both task specificity and interpretability. Furthermore, a dedicated dataset of group behavior for social anxiety assessment in university students is constructed, addressing the absence of precise correspondences between observable behaviors and psychological states. In addition, a multi-task learning framework is introduced to jointly optimize behavioral indicator prediction and social anxiety regression, with a graph structure consistency loss incorporated to improve model generalization. Detailed technical formulations of these innovations are presented in the subsequent section.

The remainder of this study is organized below. Section 2 provides a comprehensive description of the proposed framework, with emphasis placed on the design of key modules and the dataset construction methodology. Section 3 presents experimental evaluations to validate the effectiveness of the approach. Section 4 analyzes the experimental results, discusses existing limitations, and outlines potential directions for future research. Section 5 concludes with a summary of the core contributions and highlights the academic significance and application prospects of the proposed framework.

2. METHODOLOGY

2.1 Overall framework

A non-contact assessment framework for social anxiety in university students is proposed, in which video-based image data are utilized as the input. A comprehensive technical pipeline is constructed, encompassing data processing, feature extraction, interaction modeling, attention enhancement, and multi-task evaluation, thereby enabling an accurate mapping from observed group behavior to quantified social anxiety levels. The proposed framework consists of six sequential modules: data acquisition and preprocessing, individual detection and tracking, multi-dimensional behavioral feature extraction, adaptive spatiotemporal graph convolutional network modeling, a psychology-driven attention mechanism, and multi-task social anxiety assessment. Among these, the adaptive spatiotemporal graph convolutional network, the psychology-driven attention mechanism, and the multi-task learning framework constitute the core innovative components, which play a dominant role in enhancing the accuracy of group interaction modeling and social anxiety evaluation. In contrast, modules such as individual detection and tracking, as well as multi-dimensional behavioral feature extraction, are implemented using established techniques and serve as foundational support for the proposed framework, without introducing additional methodological modifications.

Figure 1 illustrates the overall framework of the proposed non-contact social anxiety assessment method for university students. A tightly coupled logical pipeline is established across all modules. The data acquisition and preprocessing module is responsible for constructing a dedicated dataset and performing data cleaning and standardization, thereby providing high-quality inputs for subsequent processing. The individual detection and tracking module extracts trajectory and posture information of individuals from video sequences, while the multi-dimensional behavioral feature extraction module generates individual feature vectors based on these inputs. Together, these modules establish a robust data foundation for group interaction modeling. The adaptive spatiotemporal graph convolutional network is employed to dynamically model group interaction relationships and to extract high-order interaction features. Subsequently, the psychology-driven attention mechanism is applied to selectively refine and enhance feature representations, with emphasis placed on key behavioral patterns associated with social anxiety. The multi-task learning framework further enables the joint optimization of behavioral indicator prediction and social anxiety score regression, ultimately producing quantitative assessments of individual social anxiety levels. Through the integrated design of these innovative modules, a unified framework is achieved that simultaneously ensures technical feasibility and psychological relevance, thereby significantly improving both assessment accuracy and model interpretability.

2.2 Design of the adaptive spatiotemporal graph convolutional network

To address the limitations of existing graph convolutional methods, in which fixed graph structures fail to adapt to the spatiotemporal dynamics of group social relationships [20, 21], an adaptive spatiotemporal graph convolutional network is proposed. The core innovation lies in the dynamic learning

of graph structures that simultaneously account for spatial proximity and social affinity. In addition, spatiotemporal convolution is employed to collaboratively extract high-order interaction features. The adaptive spatiotemporal graph convolutional network is primarily composed of a dynamic graph structure learning mechanism and a spatiotemporal convolution module. Through the integrated operation of these components, precise modeling of group interaction behaviors is achieved, thereby providing high-quality feature representations for subsequent social anxiety assessment. The network architecture is illustrated in Figure 2.

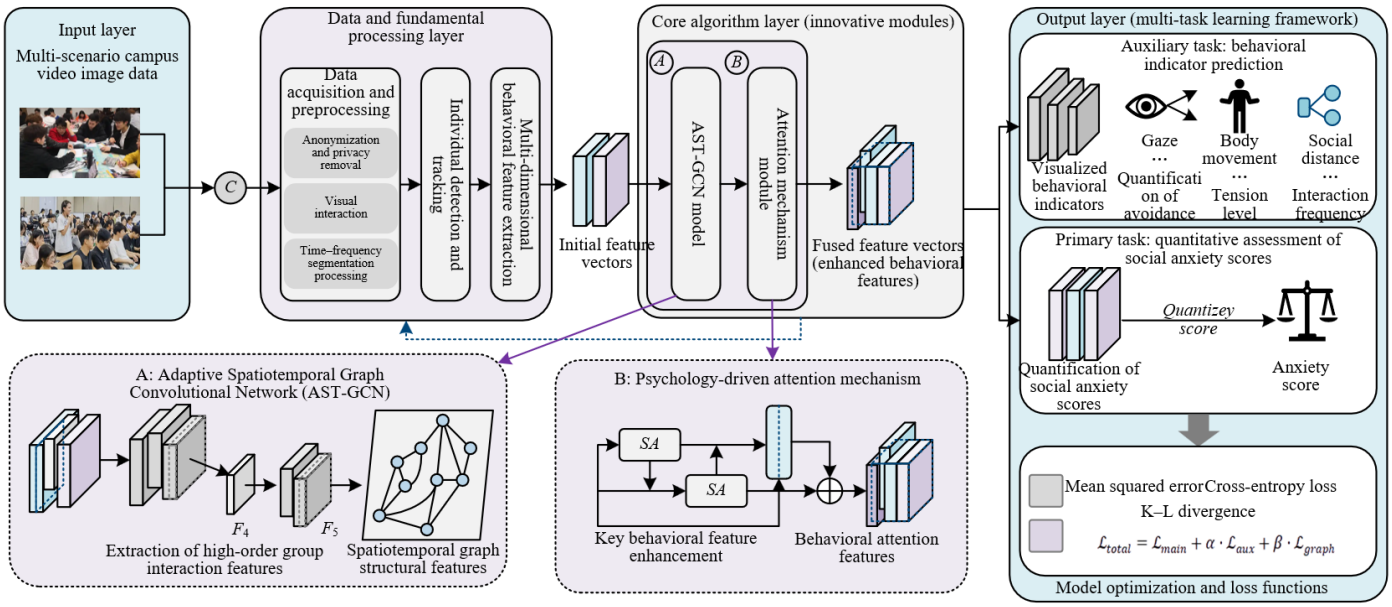

Figure 1. Overall framework of the non-contact social anxiety assessment method for university students

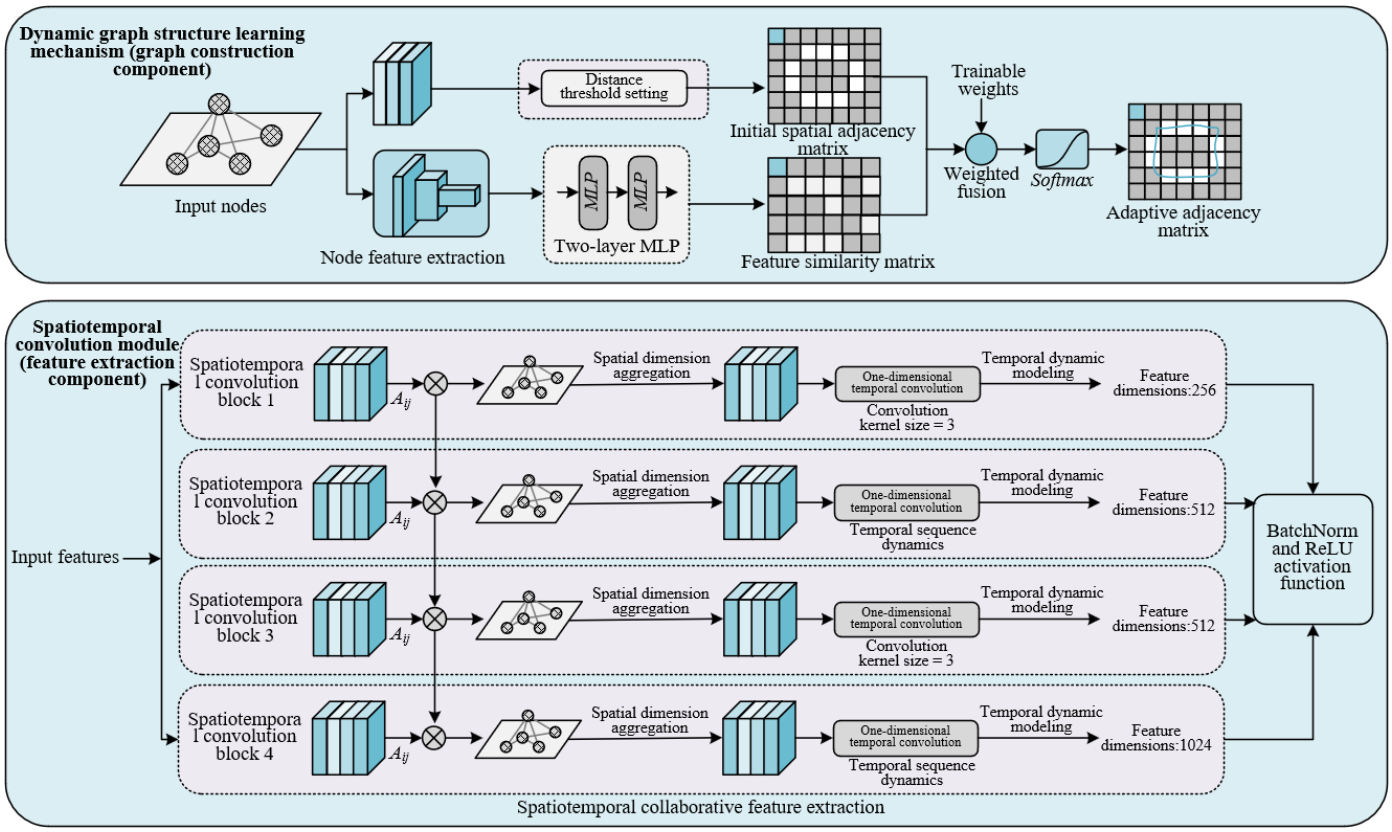

Figure 2. Schematic diagram of the adaptive spatiotemporal graph convolutional network

The dynamic graph structure learning mechanism constitutes the core innovation of the adaptive spatiotemporal graph convolutional network, in which an adaptive adjacency matrix is constructed through the dual integration of spatial proximity and feature similarity. Initially, an initial adjacency matrix A_{dist} is established based on the spatial distances between individuals. A predefined distance threshold is introduced, such that when the Euclidean distance between the centers of two individuals is below this threshold, the corresponding matrix element is assigned a value of 1;

otherwise, it is set to 0. In this manner, the spatial rationality of the graph structure is ensured. Subsequently, a trainable similarity matrix $M \in R^{N \times N}$ is incorporated, where each element M_{ij} is computed using a multilayer perceptron. The input to the multilayer perceptron is formed by concatenating the feature vectors of nodes i and j. A two-layer fully connected architecture is adopted, with the hidden layer dimension set to 128 and the rectified linear unit function employed as the activation function. The output is normalized to produce similarity scores. The final adaptive adjacency matrix A_{ij} is obtained through a weighted fusion strategy, defined as $A_{ij} = softmax(\alpha A_{dist} + (1 - \alpha)M_{ij})$, where α denotes a trainable weighting parameter that balances the contributions of spatial proximity and feature similarity. The softmax normalization ensures that the sum of each row of the adjacency matrix equals 1. Through this formulation, dynamic adaptation to the spatiotemporal evolution of group social relationships is achieved, enabling precise characterization of interpersonal affinity variations among individuals.

The spatiotemporal convolution module is constructed based on the adaptive adjacency matrix to enable collaborative extraction of group interaction features across both spatial and temporal domains, thereby further enhancing feature representation capacity. For spatial graph convolution, an improved graph convolution kernel is adopted, in which node features are aggregated using the adaptive adjacency matrix Aij. The feature update is defined as:

$$X' = \hat{A}\sigma(\hat{A}XW_0 + b_0)W_1 + b_1 \quad (1)$$

where, $\hat{A} = A + I$ denotes the adjacency matrix with added self-loops, W_0 and W_1 represent the convolutional kernel weights, b_0 and b_1 are bias terms, and σ denotes the rectified linear unit activation function. The input dimension is consistent with the node feature dimension, and the output dimension is set to 256, enabling effective extraction of spatial features associated with group interactions. Following spatial convolution, a one-dimensional temporal convolution layer is applied in sequence. The convolution kernel size is set to 3, with a stride of 1, and padding is implemented using the "same" mode. This design facilitates the capture of temporal dynamics in both individual behaviors and group interactions, thereby addressing the limitation of static graph convolution in modeling temporal dependencies. The network adopts a stacked architecture composed of four spatiotemporal convolution blocks, with feature dimensions progressively set to 256, 512, 512, and 1024. After each convolution block, a batch normalization layer and a rectified linear unit activation function are incorporated to suppress overfitting and accelerate convergence during training. Through this hierarchical structure, high-order spatiotemporal features of group interactions are progressively extracted, providing a robust feature foundation for subsequent attention enhancement and social anxiety assessment.

2.3 Psychology-driven attention mechanism

To address the limitations of existing approaches, in which psychological priors are insufficiently integrated and model specificity and interpretability remain limited, a psychology-driven attention mechanism is designed. The core innovation lies in the quantification of representative behavioral indicators associated with social anxiety into computable attention-guiding signals. Through dedicated network modeling and feature reweighting, deep integration between psychological priors and image-derived features is achieved, enabling the model to selectively focus on behavioral patterns that are strongly correlated with anxiety states. The proposed mechanism is composed of three components: attention-guiding signal quantification, behavioral attention network design, and an attention weighting mechanism. Through the coordinated operation of these components, both the accuracy of social anxiety assessment and the interpretability of the model are significantly enhanced. The detailed architecture is illustrated in Figure 3.

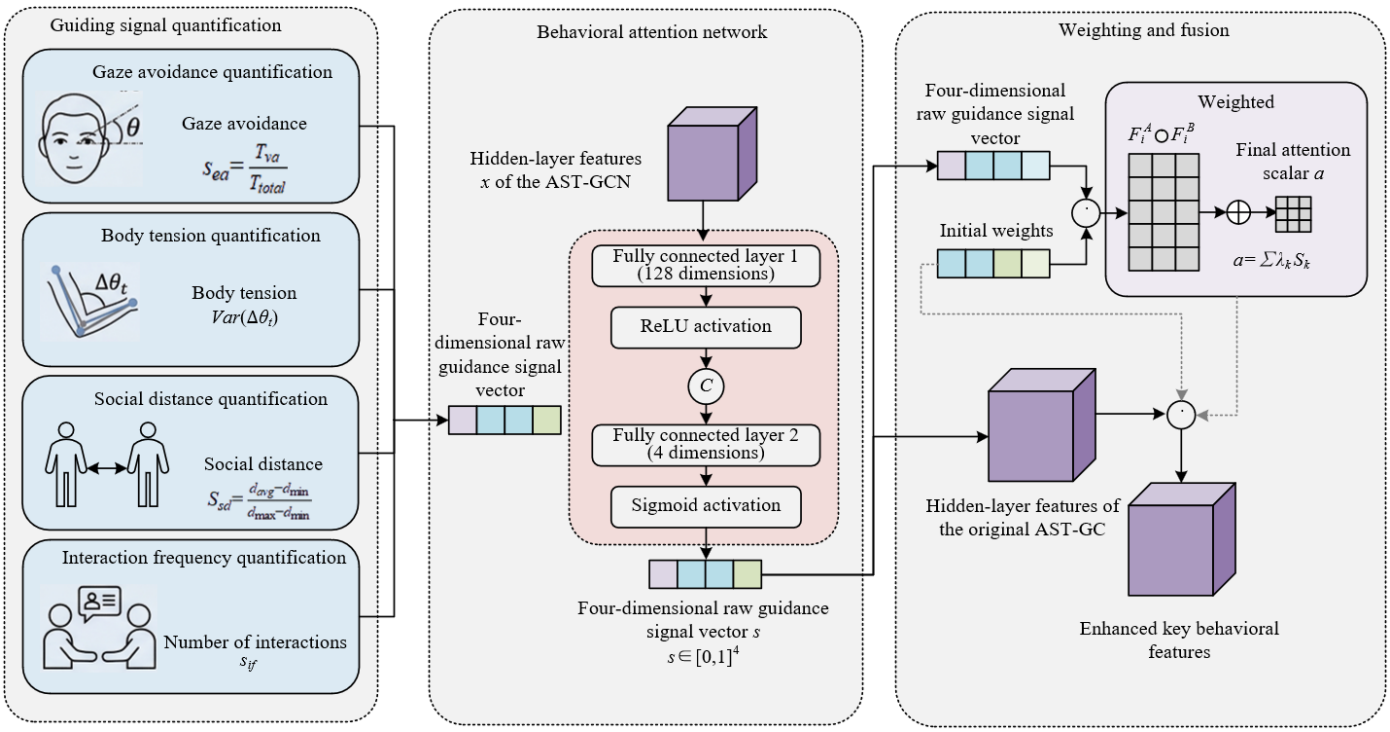

Figure 3. Architecture of the psychology-driven attention mechanism module

The quantification of attention-guiding signals constitutes the foundation of the proposed mechanism. The core objective is to transform four categories of behavioral indicators, which are strongly associated with social anxiety in clinical psychology, into model-compatible quantitative signals, thereby ensuring both scientific validity and effectiveness. Gaze avoidance is quantified as the proportion of time during which the angle between an individual's gaze direction and the normal vector of surrounding faces exceeds a predefined threshold. This threshold is determined based on established psychological experimental standards. The quantification is defined as:

$$s_{ea}=\frac{T_{va}}{T_{total}} \tag{2}$$

where, T_{ea} denotes the duration during which the gaze angle exceeds the threshold, and T_{total} represents the total observation time. Body tension is quantified based on pose keypoints. Core joints, including the elbow and shoulder joints, are selected, and the variance of joint angle change rates is computed. The formulation is given by $s_{mt}=Var(\Delta\theta_t)$, where $\Delta\theta_t$ denotes the change in joint angles between consecutive frames. Social distance is measured as the average Euclidean distance between an individual and its nearest neighbors, followed by Min–Max normalization. The formulation is expressed as:

$$S_{sd}=\frac{d_{avg}-d_{\min}}{d_{\max}-d_{\min}} \tag{3}$$

Interaction frequency is quantified by counting the occurrences of face-to-face orientation and proximity events, and is subsequently normalized to obtain s_{if}. These four quantified signals collectively form the core basis for attention guidance within the proposed framework.

The behavioral attention network and weighting mechanism enable precise selection and enhancement of features guided by attention signals, constituting the central innovation of this module. A lightweight multilayer perceptron architecture is adopted for the behavioral attention network, with the input defined as the individual hidden-layer features produced by the adaptive spatiotemporal graph convolutional network. The network consists of two fully connected layers. The first layer has an input dimension equal to that of the hidden-layer features and an output dimension of 64. The second layer produces an output dimension of 4, corresponding to the predicted intensity values of the four behavioral indicators. A sigmoid activation function is employed to constrain the outputs within the interval [0, 1]. The network output is formulated as $s=sigmoid(W_2\cdot ReLU(W_1x+b_1)+b_2)$, where W_1 and W_2 denote the weight matrices, b_1 and b_2 represent bias terms, and x corresponds to the hidden-layer features generated by the adaptive spatiotemporal graph convolutional network. To constrain network training, an auxiliary loss is introduced. The mean squared error loss is employed to measure the discrepancy between the predicted behavioral indicators and the corresponding ground-truth annotations, defined as:

$$L_{aux}=\frac{1}{4N}\sum_{i=1}^{N}\sum_{k=1}^{4}(s_{i,k}-\hat{s}_{i,k})^2 \tag{4}$$

where, $s_{i,k}$ denotes the predicted value, $\hat{s}_{i,k}$ represents the ground-truth value, and N is the number of samples. The attention scalar is computed as a weighted sum of the four predicted behavioral indicators, expressed as:

$$\omega=\sum_{k=1}^{4}\lambda_k s_{i,k} \tag{5}$$

where, λ_k denotes the initial weights assigned according to psychological importance, which can be further refined during model training. Finally, feature reweighting is performed through element-wise multiplication, given by $x'=\omega\cdot x$. In this manner, key behavioral features associated with social anxiety are emphasized, while irrelevant features are suppressed. Furthermore, the model decision process can be interpreted through visualization of attention weights, thereby achieving a deep integration of psychological priors and image processing techniques.

2.4 Multi-task learning and social anxiety assessment framework

To address the limitations of single-task learning, including restricted generalization capability and insufficient evaluation accuracy, a multi-task learning and social anxiety assessment framework is designed. The core innovation lies in the joint optimization of the primary and auxiliary tasks, combined with a graph structure consistency constraint, thereby enabling accurate regression of social anxiety scores and enhanced model generalization. The framework takes attention-weighted group interaction features as input. Through multi-task collaborative training, the discriminative capacity of feature representations is strengthened, while consistency between the adaptively learned graph structure and real social relationships is enforced. In this manner, a social anxiety assessment system that balances accuracy and generalization is established, providing a robust foundation for non-contact evaluation.

Figure 4 illustrates the proposed multi-task learning and social anxiety assessment framework, along with the associated loss computation. A primary–auxiliary task design is adopted, in which task boundaries are explicitly defined and complementary optimization is achieved. The primary task focuses on social anxiety score regression. The attention-weighted individual hidden feature sequences are first aggregated using global average pooling, with the pooling window size set to 5, enabling temporal feature aggregation into fixed-dimensional individual feature vectors. The aggregated features are then fed into fully connected layers, where the first layer has an output dimension of 128 and the second layer produces a single scalar corresponding to the Liebowitz social anxiety scale score. The mean squared error loss is employed to measure the discrepancy between the predicted scores and the ground-truth scale values, defined as:

$$L_{main}=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \tag{6}$$

where, y_i denotes the ground-truth anxiety score, $\hat{y}_i$ represents the predicted score, and N is the number of samples. The auxiliary task involves the prediction of four categories of behavioral indicators. Dedicated prediction heads are designed

according to the characteristics of each indicator. Gaze avoidance and interaction frequency are modeled using binary classification heads, which output the probability of occurrence, whereas body tension and social distance are modeled using regression heads, which output continuous quantitative values. For classification heads, the cross-entropy loss is adopted, defined as:

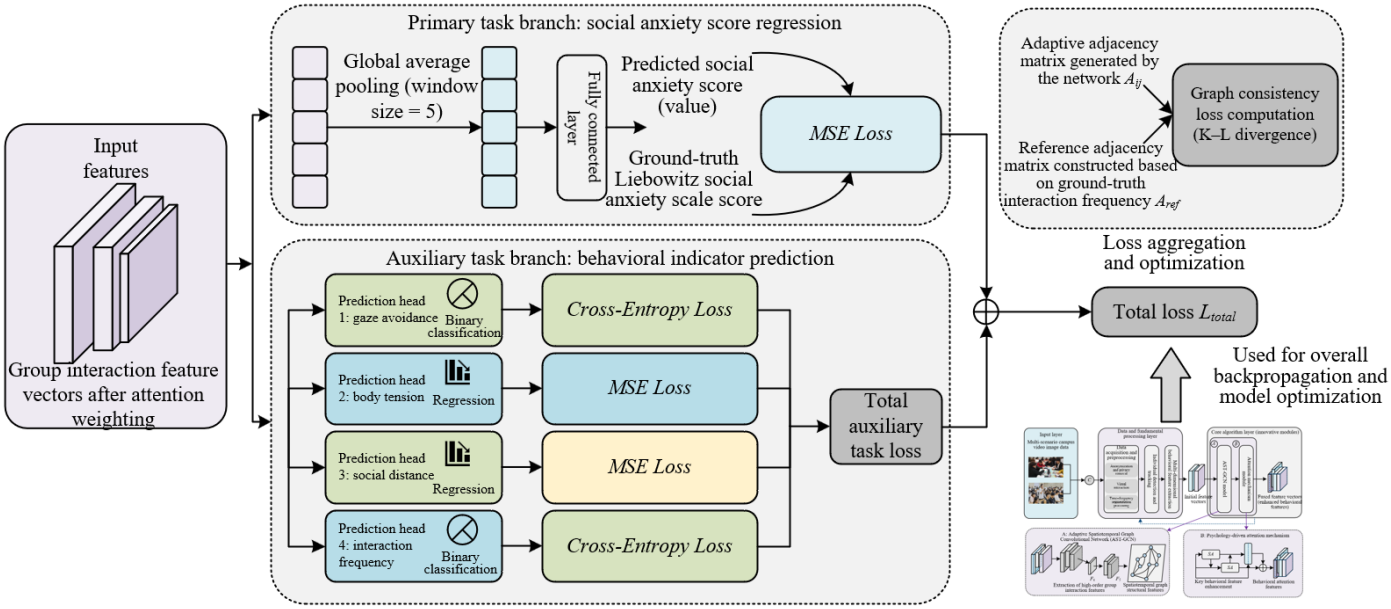

Figure 4. Multi-task learning and social anxiety assessment framework with loss computation

$$L_{ce}=-\frac{1}{N}\sum_{i=1}^{N}\hat{s}_{i,k}\log(s_{i,k}) \tag{7}$$

For regression heads, the mean squared error loss is employed. The total auxiliary loss is defined as the sum of the losses corresponding to all behavioral indicators. Through the auxiliary tasks, the model is guided to learn behavior patterns associated with social anxiety, thereby indirectly improving the regression accuracy of the primary task.

The design of the graph structure consistency loss and the overall loss function constitutes a key innovation of the framework, enabling coordinated optimization across tasks and graph modeling. A reference graph is constructed based on manually annotated interaction frequencies, where higher interaction frequency corresponds to greater edge weights between nodes, resulting in a reference adjacency matrix A_{ref} that reflects real social relationships. The graph structure consistency loss is defined using the Kullback–Leibler divergence to measure the discrepancy between the adaptively learned adjacency matrix A_{ij} and the reference adjacency matrix A_{ref}, formulated as:

$$L_{graph}=KL(A_{ij}|A_{ref})=\sum_{i,j}A_{ij}\log\frac{A_{ij}}{A_{ref}} \tag{8}$$

This formulation enforces consistency between the self-adaptive graph structure and real social interactions, thereby improving the accuracy of group interaction modeling. The total loss function is defined as a weighted sum of the individual loss components, expressed as:

$$L_{total}=L_{main}+\alpha\cdot L_{aux}+\beta\cdot L_{graph} \tag{9}$$

where, α and β are weighting coefficients determined through validation-based tuning, with search ranges set to [0.1, 1.0]. The optimal values are selected as 0.5 and 0.3, respectively. Through this formulation, joint optimization of the primary task, auxiliary tasks, and graph structure modeling is achieved. As a result, the accuracy of social anxiety score regression is ensured, while the learning of behavior features and the rationality of the graph structure are simultaneously enhanced, leading to significant improvements in model generalization capability and evaluation reliability.

2.5 Construction of the university student social anxiety group behavior dataset

Existing studies lack a dedicated group behavior dataset specifically designed for social anxiety assessment in university students, resulting in the absence of precise correspondence between video-derived behavioral features and social anxiety scale scores. This limitation constrains the reliability of model training and validation. To address this gap, a specialized dataset for group behavior associated with social anxiety in university students is constructed. The core innovation lies in focusing on typical university social scenarios, enabling multi-dimensional scene coverage, precise label alignment, and standardized preprocessing. This dataset fills the existing gap in aligning behavioral observations with psychological states, providing reliable data support for the proposed framework and serving as a benchmark resource for related research.

Dataset collection was conducted according to the principles of scenario diversity, parameter standardization, and labeling accuracy, ensuring both representativeness and practical applicability. Four typical campus social scenarios were selected, including group discussions, cafeteria dining, student organization activities, and classroom interactions. These scenarios encompassed variations in lighting conditions, camera viewpoints, and group sizes, thereby comprehensively simulating daily social environments encountered by university students. Data acquisition was

performed using high-definition cameras, with a resolution of 1920 × 1080 and a frame rate of 30 frames per second. Video recordings were conducted from either top-down or horizontal perspectives. Each participant was recorded for a duration of no less than 5 minutes. The dataset consisted of 200 university students, including 98 males and 102 females, aged between 18 and 24 years, and representing diverse academic disciplines such as humanities, sciences, engineering, and medicine, thereby ensuring sample diversity. A dual-annotation strategy was adopted for label acquisition. The Liebowitz social anxiety scale scores completed by participants were collected as the primary supervision labels. In parallel, video segments were annotated by two psychology professionals to generate auxiliary supervision labels for four categories of behavioral indicators: gaze avoidance, body tension, social distance, and interaction frequency. Annotation consistency was evaluated using the Kappa coefficient, thereby ensuring the accuracy of the labels.

Data preprocessing was conducted using a standardized pipeline that simultaneously addresses privacy preservation, data compatibility, and training stability, representing a critical component of the dataset construction process. Privacy protection was achieved through Gaussian blurring, with the kernel size set to 5×5, applied to all facial regions in the video to ensure anonymization. This approach effectively safeguarded participant privacy while preserving essential behavioral features required for analysis. Video segmentation was performed by dividing the recordings into fixed-length clips in chronological order, with each segment set to a duration of 10 seconds. This strategy prevented both the loss of temporal information and redundancy. A total of 12,000 video segments were obtained and subsequently divided into training, validation, and test sets according to a ratio of 7:2:1, ensuring a balanced and representative dataset split. Feature normalization was implemented using Z-score standardization to process the extracted multi-dimensional behavioral features, thereby eliminating the influence of differing scales. The normalization is defined as $x_{norm} = (x - \mu)/\sigma$, where x denotes the original feature value, μ represents the mean of the feature, and σ denotes the standard deviation. Through this process, all features were transformed to a unified scale, enhancing training stability and accelerating convergence. The standardized dataset can be directly integrated with the proposed adaptive spatiotemporal graph convolutional network, attention mechanism, and multi-task learning framework, thereby providing a robust foundation for effective model training and reliable performance evaluation.

3. EXPERIMENTS AND RESULTS ANALYSIS

3.1 Experimental setup

The experiments were conducted based on the constructed dataset of group behavior for social anxiety assessment in university students. Standard experimental protocols in the field of image processing were strictly followed to ensure reproducibility and fairness. The dataset was partitioned into training, validation, and test sets according to a ratio of 7:2:1. Specifically, the training set contained 8,400 video segments, the validation set contained 2,400 segments, and the test set contained 1,200 segments, collectively covering all campus social scenarios and sample types.

The hardware environment consisted of an NVIDIA RTX 3090 graphics processing unit with 24 GB of memory, an Intel Core i9-12900K central processing unit, and 64 GB of system memory. The software environment was implemented using the PyTorch 1.12.0 deep learning framework with Python 3.8, and is supported by tools such as OpenCV and Scikit-learn for data processing and model evaluation. Key model parameters were configured below. The adaptive spatiotemporal graph convolutional network adopted a stacked architecture of four spatiotemporal convolution blocks, with the spatial graph convolution input dimension set to 128 and output dimensions sequentially set to 256, 512, 512, and 1024. In the dynamic graph learning module, the hidden layer dimension of the multilayer perceptron was set to 128, and the initial value of the trainable weighting parameter α was initialized to 0.5. In the psychology-driven attention mechanism, the hidden layer dimension of the multilayer perceptron was set to 64, with an output dimension of 4. The optimal weights for the multi-task loss were determined as $\alpha = 0.5$ and $\beta = 0.3$. The training batch size was set to 32, and the number of training epochs was set to 100. The Adam optimizer was employed, with an initial learning rate of 1e-4, which is further adjusted using a cosine annealing schedule.

3.2 Comparative experimental results and analysis

The comparative experiments were designed to evaluate the superiority of the proposed full model over existing baseline methods. Both quantitative and qualitative analyses were conducted to highlight the core contributions of the proposed innovative modules. The experimental results are presented in Table 1.

Table 1. Performance comparison between baseline methods and the proposed model

Method	Mean Absolute Error (MAE)	Root Mean Squared Error (MSE)	Coefficient of Determination R^2	Improvement over Method 3 (MAE ↓)	Improvement over Method 4 (MAE ↓)
Method 1: Handcrafted features + support vector regression	5.82	7.36	0.612	-	-
Method 2: Individual temporal model based on long short-term memory	4.95	6.28	0.703	18.7%	-
Method 3: Fixed-graph graph convolutional network	4.21	5.57	0.765	-	-
Method 4: Adaptive spatiotemporal graph convolutional network without the attention mechanism	3.15	4.02	0.848	25.2%	-
Proposed full model	2.38	3.05	0.912	43.5%	24.4%

The quantitative analysis demonstrates that the proposed full model significantly outperforms all baseline methods across all three evaluation metrics, highlighting the synergistic effect of the proposed innovative modules. Compared with traditional handcrafted feature-based methods, the mean absolute error is reduced by 3.44, the root mean squared error is reduced by 4.31, and the coefficient of determination (R^2) is improved by 0.300. These results indicate that deep learning approaches, when combined with the proposed innovations, are capable of extracting more discriminative group behavioral features. In comparison with the long short-term memory-based individual temporal model, the mean absolute error is reduced by 2.57 and the R^2 is increased by 0.209, suggesting that group interaction modeling is more effective than individual temporal modeling in capturing social anxiety states. Relative to the fixed-graph graph convolutional network, the mean absolute error is reduced by 1.83, corresponding to an improvement of 43.5%, thereby validating the effectiveness of the adaptive spatiotemporal graph convolutional network in dynamically modeling group social relationships. Furthermore, when compared with the adaptive spatiotemporal graph convolutional network without the attention mechanism, the mean absolute error is reduced by 0.77, representing an improvement of 24.4%. This result confirms that the psychology-driven attention mechanism effectively emphasizes key behavioral features and enhances evaluation accuracy. The proposed full model achieves an R^2 value of 0.912, indicating that the predicted social anxiety scores closely align with the true distribution of the ground-truth values, thereby satisfying the practical requirements of non-contact assessment.

3.3 Ablation study

The ablation study was conducted by systematically removing each core innovative module to analyze its contribution to overall model performance, thereby validating the necessity of each component. The experimental results are presented in Table 2.

Table 2. Results of the ablation study

Method	Mean Absolute Error (MAE)	Root Mean Squared Error (MSE)	Coefficient of Determination (R^2)	Performance Degradation Compared to the Full Model (MAE ↑)
Proposed full model	2.38	3.05	0.912	-
Ablation 1: without the adaptive graph module (fixed graph)	3.67	4.72	0.805	54.2%
Ablation 2: without the psychology-driven attention mechanism	3.21	4.18	0.853	34.9%
Ablation 3: without multi-task learning (primary task only)	3.02	3.94	0.867	26.9%
Ablation 4: without the graph structure consistency loss	2.85	3.61	0.889	19.7%

The results of the ablation study demonstrate that all proposed innovative modules contribute significantly to overall model performance, and their removal leads to varying degrees of performance degradation. In Ablation 1, where the adaptive graph module is replaced with a fixed graph structure, the mean absolute error increases by 1.29, corresponding to the largest performance decline of 54.2%, and the R^2 decreases to 0.805. This finding indicates that the dynamic graph structure learning mechanism plays a critical role in improving the accuracy of group interaction modeling. Fixed graph structures are unable to adapt to the spatiotemporal dynamics of social relationships and fail to capture variations in interpersonal affinity. In Ablation 2, the removal of the psychology-driven attention mechanism results in a mean absolute error increase of 0.83 and a performance degradation of 34.9%, demonstrating that the attention mechanism effectively enhances behavioral features associated with social anxiety and improves model specificity. In Ablation 3, when multi-task learning is excluded, the mean absolute error increases by 0.64, corresponding to a performance decline of 26.9%. This result confirms that the collaborative optimization of primary and auxiliary tasks enhances model generalization capability, and that auxiliary tasks facilitate the learning of behavior features, thereby indirectly improving the accuracy of social anxiety score regression. In Ablation 4, the removal of the graph structure consistency loss leads to a mean absolute error increase of 0.47 and a performance decline of 19.7%, indicating that this loss function effectively constrains the alignment between the adaptive graph structure and real social relationships, thereby further improving the accuracy of group interaction modeling. The results of all ablation experiments are mutually consistent, demonstrating that the four proposed innovative modules operate in a complementary and synergistic manner, jointly contributing to a high-accuracy and highly interpretable social anxiety assessment model.

3.4 Robustness evaluation

To validate the effectiveness of the individual detection, multi-object tracking, and pose and facial feature extraction modules in supporting group behavior modeling, a visual verification experiment was conducted. As illustrated in Figure 5, the processing results of four university students in an outdoor campus social scenario are clearly presented. The light-blue detection bounding boxes are observed to accurately align with human body contours, while unique identity labels are assigned without confusion or omission. These results demonstrate that the individual detection and tracking module exhibits strong robustness in complex outdoor environments, enabling stable output of continuous spatial positions and identity information. This capability provides a reliable data foundation for the subsequent dynamic modeling of group interaction relationships. Furthermore, human pose skeletons are reconstructed using colored keypoints and white connecting lines, with no observable misalignment or missed detections. Facial keypoints accurately cover critical regions such as the eyes, eyebrows, and nose, which are highly relevant to social anxiety analysis. This ensures the effective

extraction of fine-grained behavioral features, thereby providing precise visual evidence for the quantitative computation of psychological indicators, including gaze avoidance, body tension, and social distance.

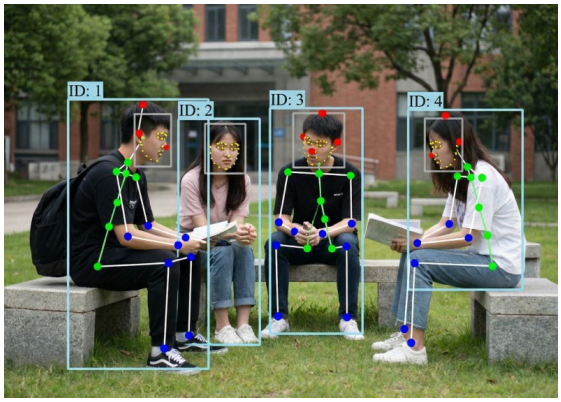

Figure 5. Human pose and facial key point extraction

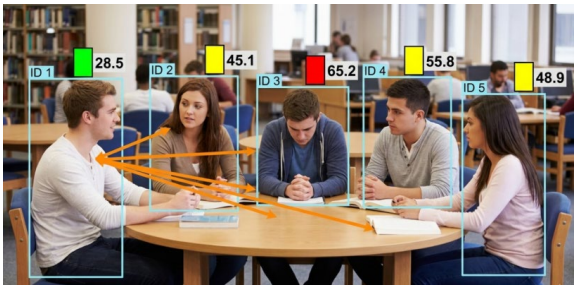

Figure 6. Mapping of social anxiety level assessment results

To validate the accuracy of group interaction modeling achieved by the adaptive spatiotemporal graph convolutional network and the practical effectiveness of the multi-task social anxiety assessment framework, a visual verification experiment was conducted. As illustrated in Figure 6, the complete assessment results in a campus library group discussion scenario are clearly presented. Light-blue detection bounding boxes, together with unique identity labels, enable precise localization and continuous tracking of individuals. The dynamically generated social relationship edges, represented in orange with varying thickness, intuitively reflect the adaptively learned interpersonal affinity within the group. Stronger interaction relationships are indicated by thicker edges, whereas sparse connections are observed among socially avoidant individuals. These observations confirm that the dynamic graph structure learning mechanism accurately captures group interactions, overcoming the limitations of traditional fixed graph structures. Social anxiety levels are visually mapped through color-coded markers and corresponding quantitative Liebowitz social anxiety scale scores. For example, Identity 3 is associated with a red marker and a high score of 65.2, corresponding to typical high-anxiety behavioral patterns such as downward gaze, body stiffness, and sparse social connections. In contrast, Identity 1 is associated with a green marker and a low score of 28.5, reflecting a relaxed posture and frequent interactions characteristic of low anxiety. Other individuals, represented by yellow markers and moderate scores, exhibit behavioral patterns consistent with intermediate anxiety levels.

To further evaluate robustness, experiments were conducted under common sources of disturbance in image processing, simulating real-world campus environments. Three typical types of interference were considered: variations in illumination, changes in camera viewpoint, and fluctuations in group size. Different levels of disturbance intensity were introduced for each factor, and the corresponding experimental results are summarized in Table 3.

Table 3. Robustness evaluation of the proposed model under different disturbance conditions

Disturbance Condition	Disturbance Intensity	Mean Absolute Error (MAE)	Root Mean Squared Error (MSE)	Coefficient of Determination (R^2)	Performance Degradation Compared to Normal Condition (MAE ↑)
Normal condition	No disturbance	2.38	3.05	0.912	-
Illumination variation	Low illumination (50 lux)	2.65	3.37	0.894	11.3%
Illumination variation	High illumination (1500 lux)	2.71	3.45	0.889	13.9%
Viewpoint variation	Top-down view (45°)	2.52	3.21	0.903	5.9%
Viewpoint variation	Side view (30°)	2.83	3.58	0.882	18.9%
Group size variation	Small group (2–3 individuals)	2.41	3.11	0.909	1.3%
Group size variation	Large group (10–12 individuals)	2.97	3.72	0.875	24.8%

The robustness evaluation demonstrates that the proposed model maintains high assessment accuracy under various disturbance conditions, indicating strong stability. Under illumination variations, the mean absolute error increases by 0.27 and 0.33 under low-light and high-light conditions, respectively, with performance degradation remaining below 14% in both cases. This result indicates that the model exhibits strong adaptability to illumination changes, which can be attributed to the complementary nature of multi-dimensional behavioral features and the application of feature normalization. Under viewpoint variations, a performance degradation of only 5.9% is observed at a top-down angle of 45°, while a larger degradation of 18.9% occurs at a side-view angle of 30°. Despite this decrease, the R^2 value remains above 0.88, indicating that the model is capable of adapting to diverse camera viewpoints commonly encountered in campus environments. Under group size variations, model performance remains stable in small-group scenarios, whereas in large-group scenarios, the mean absolute error increases by 0.59, corresponding to a performance degradation of 24.8%. Nevertheless, the performance remains superior to that of most baseline methods. This degradation is primarily attributed to

the increased complexity of individual detection and interaction modeling in larger groups. However, the adaptive graph structure and psychology-driven attention mechanism effectively focus on key individuals and behaviors, thereby mitigating the impact of such disturbances. Overall, the proposed model demonstrates strong robustness under common disturbances in complex campus environments, maintaining high evaluation accuracy and satisfying the practical requirements of non-contact social anxiety monitoring.

3.5 Performance comparison across different campus scenarios

To evaluate the adaptability of the proposed full model to different social environments, performance was assessed across the four typical campus social scenarios included in the dataset. The results are illustrated in Figure 7, where the mean absolute error is compared across scenarios using a bar chart.

As shown in Figure 7, high evaluation accuracy is consistently maintained across all four campus scenarios. The group discussion scenario exhibits the best performance, owing to clearly defined interaction patterns. In the cafeteria dining scenario, where crowd density is higher, a slight decrease in performance is observed; however, the overall accuracy remains at a high level. In the classroom interaction scenario, the influence of viewpoint constraints is minimal, and the R^2 remains above 0.89. The performance variation across different scenarios is observed to be less than 5%, indicating that the adaptive spatiotemporal graph convolutional network and the psychology-driven attention mechanism effectively adapt to diverse group behavior patterns. These results demonstrate strong scenario generalization capability of the proposed model.

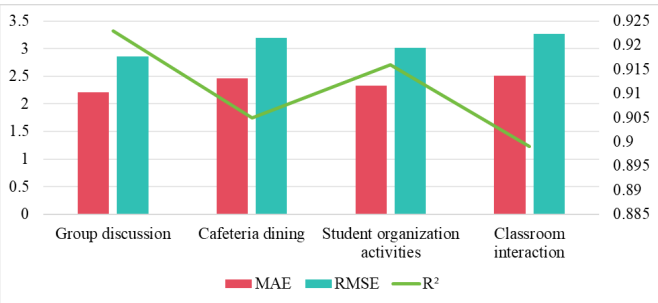

Figure 7. Model performance comparison across different campus scenarios

3.6 Sensitivity analysis of key parameters

A sensitivity analysis was conducted on the core trainable parameter α in the adaptive spatiotemporal graph convolutional network. Different values of α were evaluated to examine their impact on model performance, particularly in terms of the mean absolute error. The corresponding results are presented in Figure 8, where the relationship between the parameter and performance is illustrated using a line chart.

As illustrated in Figure 8, model performance exhibits a non-monotonic trend with respect to the parameter α, increasing initially and subsequently decreasing. Optimal performance is achieved at $\alpha = 0.5$, where a balanced integration between spatial proximity and feature similarity is attained. When α is set to a smaller value, the influence of spatial structural constraints is weakened, whereas excessively large values of α reduce the model's ability to represent social affinity through feature similarity. Within the interval [0.3, 0.7], performance variation remains below 8%, indicating that the proposed model demonstrates strong robustness to this key parameter. Furthermore, the relatively wide range of stable performance suggests sufficient flexibility for parameter tuning, facilitating practical deployment in real-world applications.

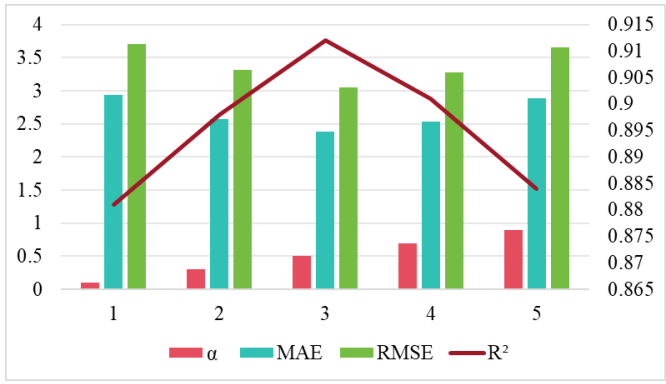

Figure 8. Sensitivity analysis of the fusion weight parameter α

4. DISCUSSION

A comprehensive analysis of the experimental results indicates that the high accuracy achieved in the non-contact assessment of social anxiety among university students can be primarily attributed to the synergistic integration of the four proposed innovative modules. These modules are found to align closely with the core requirements of the image processing domain, particularly in terms of dynamic modeling of group behavior and the selective extraction of critical features. The adaptive spatiotemporal graph convolutional network enables the dynamic learning of an adjacency matrix that integrates spatial proximity and feature similarity, thereby overcoming the limitations of traditional fixed graph structures in capturing the spatiotemporal evolution of social relationships. Through this mechanism, group interaction modeling is advanced from static representation to dynamic adaptation, allowing for the accurate characterization of interpersonal affinity variations and providing more discriminative behavioral features for social anxiety assessment. The psychology-driven attention mechanism facilitates deep integration between psychological priors and image processing techniques. By quantifying behavior indicators associated with social anxiety and selectively enhancing key features, the limitations of purely data-driven models in terms of specificity and interpretability are effectively addressed. Furthermore, the model decision process can be validated through attention weight visualization, aligning with emerging research trends at the intersection of image processing and mental health. The multi-task learning framework, combined with the graph structure consistency constraint, enables the coordinated optimization of primary and auxiliary tasks. This design not only enhances the learning of behavior-related features but also improves model generalization capability. Compared with existing approaches, significant improvements are achieved across the mean absolute error, root mean squared error, and R^2 metrics.

In addition, superior interpretability and robustness are demonstrated. Collectively, these findings establish the academic contribution of the proposed approach within the interdisciplinary domain of group behavior analysis and non-contact psychological assessment. A novel technical paradigm is thereby provided for the application of image processing technologies in mental health monitoring.

Although favorable performance has been achieved, several limitations remain, and a critical analysis of these issues provides valuable directions for future research. In scenarios where faces are not visible, the absence of facial features leads to inaccurate quantification of attention-guiding signals, particularly for indicators such as gaze avoidance. This limitation reduces the effectiveness of attention weighting and consequently degrades assessment accuracy, representing a common challenge in non-contact video-based evaluation methods. In complex and crowded campus environments, frequent occlusions of individuals can constrain the accuracy of individual detection and tracking, resulting in deviations in trajectory and pose feature extraction. These inaccuracies may further propagate to group interaction modeling and ultimately affect the reliability of social anxiety assessment. In addition, the current dataset is limited to university students from a single geographic region, with a relatively restricted sample size and diversity. This limitation may constrain the generalization capability of the model, making it less adaptable to student populations from different regions and types of institutions.

To address these limitations, several future research directions are proposed, informed by emerging trends in the field of image processing. First, the integration of Transformer-based architectures is suggested to leverage their strong capability in modeling temporal dependencies, improving the extraction of temporal features in group behavior and overcoming the limitations of spatiotemporal convolution in long-sequence modeling, thereby improving the accuracy of group interaction modeling. Second, the incorporation of multimodal data is recommended, including audio signals (e.g., speech prosody) and physiological signals (e.g., heart rate variability), which can be fused with video-based behavioral features to overcome the constraints of a single modality and further enhance the accuracy and robustness of social anxiety assessment. Furthermore, improvements in individual detection and tracking algorithms under complex conditions are required. By incorporating attention mechanisms and contextual information, the impact of occlusion and crowd density can be mitigated, thereby improving the reliability of feature extraction. Finally, expansion of the dataset is essential, including the inclusion of participants from diverse regions and types of higher education institutions, as well as broader coverage of age groups, academic disciplines, and social scenarios. Simultaneously, refinement of the annotation system is necessary to further enhance model generalization capability and facilitate the practical deployment of non-contact social anxiety assessment technologies.

5. CONCLUSION

A novel framework for non-contact assessment of social anxiety in university students was developed to address practical demands in this domain, while targeting key challenges at the intersection of image processing and mental health research. The proposed approach integrates computer vision, graph neural networks, and psychologically grounded behavioral indicators. Four core components are introduced, including the adaptive spatiotemporal graph convolutional network, the psychology-driven attention mechanism, a dedicated group behavior dataset for social anxiety assessment in university students, and a multi-task learning framework. Experimental results demonstrate that superior performance is achieved in the social anxiety score regression task. Compared with existing baseline methods, significant improvements are observed across multiple evaluation metrics, including mean absolute error, root mean squared error, and coefficient of determination. These findings validate both the effectiveness of each proposed module and their synergistic interaction, enabling high-accuracy and highly interpretable non-contact assessment of social anxiety. From a methodological perspective, a novel paradigm for dynamic modeling of group behavior is established within the field of image processing, overcoming the limitations of traditional fixed graph structures and enabling deep integration of psychological priors with image processing techniques. From an application perspective, an efficient and non-invasive technical pathway is provided for non-contact mental health monitoring. The proposed framework aligns with the interdisciplinary innovation and practical applicability emphasized by leading Science Citation Index (SCI) journals in image processing and offers a valuable technical reference and data foundation for future research in group behavior analysis and psychological assessment.

FUNDINGS

This paper was suppored by the 14th Five-Year Plan Educational Science Project of Shaanxi Province (2025) (Grant No.: SGH25Y3228) and the Regular Research Project on Sports Science of Shaanxi Province (Grant No.: 20251591).

REFERENCES

[1] Ham, L.S., Zamboanga, B. L., Olthuis, J.V., Casner, H.G., Bui, N. (2010). No fear, just relax and play: Social anxiety, alcohol expectancies, and drinking games among college students. Journal of American College Health, 58(5): 473-479. https://doi.org/10.1080/07448480903540531

[2] Ou, Y., Guo, K., Cheng, Y. (2025). Physical exercise and social anxiety in college students: Cell phone addiction tendency and self-control as mediators. Social Behavior and Personality: An International Journal, 53(10): 1-9. https://doi.org/10.2224/sbp.14547

[3] Fang, H., Xu, X., Yang, S. (2025). The role of loneliness and self-concept clarity in the relationship between problematic mobile social network usage and social anxiety among college students. Frontiers in Psychology, 16: 1600474. https://doi.org/10.3389/fpsyg.2025.1600474

[4] Dechant, M.J., Frommel, J., Mandryk, R.L. (2021). The development of explicit and implicit game-based digital behavioral markers for the assessment of social anxiety. Frontiers in Psychology, 12: 760850. https://doi.org/10.3389/fpsyg.2021.760850

[5] Alvi, T., Kouros, C.D., Lee, J., Fulford, D., Tabak, B.A.

(2020). Social anxiety is negatively associated with theory of mind and empathic accuracy. Journal of Abnormal Psychology, 129(1): 108. https://psycnet.apa.org/doi/10.1037/abn0000493

[6] Yokoyama, K., Yamamoto, G., Liu, C., Kishimoto, K., Kuroda, T. (2023). Operating room surveillance video analysis for group activity recognition. Advanced Biomedical Engineering, 12: 171-181. https://doi.org/10.14326/abe.12.171

[7] Hsieh, C.C., Lai, W.R., Pao, T.L. (2010). An object behavior analysis system based on decoded motion vectors and Boolean operations. International Journal of Innovative Computing Information and Control, 6(10): 4565-4578.

[8] Hoffman, M., Block, P., Elmer, T., Stadtfeld, C. (2020). A model for the dynamics of face-to-face interactions in social groups. Network Science, 8(S1): S4-S25. https://doi.org/10.1017/nws.2020.3

[9] Löwe, M., Schubert, K., Vermet, F. (2020). Multi-group binary choice with social interaction and a random communication structure—A random graph approach. Physica A: Statistical Mechanics and its Applications, 556: 124735. https://doi.org/10.1016/j.physa.2020.124735

[10] Muduli, K., Ghosh, I., Ukkusuri, S.V. (2026). A graph-based spatio-temporal framework for predicting safety-critical pedestrian–vehicle interactions at unsignalized crosswalks. Accident Analysis & Prevention, 228: 108409. https://doi.org/10.1016/j.aap.2026.108409

[11] Choi, J., Kwon, J., Kim, Y., Kim, Y. (2025). Hypergraph temporal multi-behavior recommendation. Engineering Applications of Artificial Intelligence, 145: 110112. https://doi.org/10.1016/j.engappai.2025.110112

[12] Liu, C., Fu, R., Li, Y., Gao, Y., Shi, L., Li, W. (2021). A self-attention augmented graph convolutional clustering networks for skeleton-based video anomaly behavior detection. Applied Sciences, 12(1): 4. https://doi.org/10.3390/app12010004

[13] Liu, Y., Zhang, Y.W., Wang, Y. (2024). Application of deep learning-based image processing in emotion recognition and psychological therapy. Traitement Du Signal, 41(6): 2923-2933. https://doi.org/10.18280/ts.410612

[14] Guo, X.D. (2020). Cognitive psychological analysis based on multilayer semantics of web video and feature extraction of psychological images. Multimedia Tools and Applications, 79(13): 9207-9223. https://doi.org/10.1007/s11042-019-7245-9

[15] Liu, M., Shi, B. (2023). The effect of physical exercise on the anxiety of college students in the post-pandemic era: The mediating role of social support and proactive personality. Frontiers in Psychology, 14: 1128748. https://doi.org/10.3389/fpsyg.2023.1128748

[16] Sava, I.N. (2023). Social value of pathology: Adapting primary health care to reduce stress and social anxiety in college students exposed to social distancing. Frontiers in Psychology, 14: 1143221. https://doi.org/10.3389/fpsyg.2023.1143221

[17] Ulyev, A.D., Rozaliev, V.L., Zaboleeva-Zotova, A.V., Orlova, Y.A. (2021). An intelligent video surveillance system for human behavior. Scientific and Technical Information Processing, 48(5): 388-397. https://doi.org/10.3103/S0147688221050117

[18] Zhou, Y., Liu, L., Sun, X. (2022). The effects of perception of video image and online word of mouth on tourists' travel intentions: Based on the behaviors of short video platform users. Frontiers in Psychology: 13, 984240. https://doi.org/10.3389/fpsyg.2022.984240

[19] Dönmezdil, S., Toprak, S.F. (2025). Hoarseness, quality of life, and social anxiety: A case–control study. Behavioral Sciences, 15(9): 1160. https://doi.org/10.3390/bs15091160

[20] Chen, L., Shi, P., Li, G., Qi, T. (2022). Traffic flow prediction using multi-view graph convolution and masked attention mechanism. Computer Communications, 194: 446-457. https://doi.org/10.1016/j.comcom.2022.08.008

[21] Li, M., Ma, Z., Wang, Y.G., Zhuang, X. (2020). Fast Haar transforms for graph neural networks. Neural Networks, 128: 188-198. https://doi.org/10.1016/j.neunet.2020.04.028