Journal of Information Hiding and Multimedia Signal Processing ©2025 ISSN 2073-4212
Ubiquitous International Volume **16**, Number **3**, September **2025**

GaGL: Student behavior recognition based on Graph-embedded aggregation of Global and Local features

Thanh Nguyen Van

Department of Economic Information System
Academy of Finance
58 Le Van Hien, Hanoi, Vietnam
nguyenvanthanh@hvtc.edu.vn

Son Nguyen Thanh*

Department of Economic Information System
Academy of Finance
58 Le Van Hien, Hanoi, Vietnam
sonnt@hvtc.edu.vn

*Corresponding author: Son Nguyen Thanh

Received February 21, 2025, revised June 10, 2025, accepted June 13, 2025.

Abstract. *The recognition of classroom behavior for classroom management poses a significant challenge in the fields of image recognition and computer vision. While advanced technologies such as global and local image feature extraction, combined with text processing techniques, can improve the accuracy of behavior recognition, a systematic approach remains essential to effectively address this problem. In this paper, we introduce a novel network model, the Graph-embedded Aggregation of Global and Local features (GaGL) network, designed specifically for classroom behavior recognition. Our approach integrates both global and local feature information to construct an embedded graph representation, while also incorporating textual annotations within images to further refine recognition quality. Experimental results demonstrate the superiority of our method, achieving state-of-the-art performance on the MSCOCO dataset and D-Edu—a dataset collected from classroom camera footage. Additionally, we conduct comprehensive qualitative experiments and in-depth evaluations to analyze the contribution of each feature module to the proposed model, validating the effectiveness of our design.*
Keywords: Classroom behavior recognition, graph learning, image processing

1. **Introduction.** Developing methods to recognize student behavior [1] in classroom settings through camera-based monitoring has become increasingly critical in the field of computer vision. This capability enables educators and administrators to monitor learning conditions and enhance educational quality. Recent advancements in deep learning models [2, 3, 4] that integrate visual and textual data have significantly improved image recognition performance. However, despite notable progress, multimodal recognition tasks combining visual and textual information remain highly challenging due to data complexity and the inherent difficulty of accurately identifying nuanced student behaviors.

To address the precise integration of student behavior visual data with supplementary information such as image captions or annotations, researchers have explored mapping

global features [5] alongside local features [6] to improve recognition accuracy. Semantic strategies [7] have also been developed to establish relationships between objects in images through textual analysis. While some studies propose robust feature encoding techniques to align visual and textual modalities, the complexity of their interaction often limits recognition precision.

In this paper, we propose a novel convolutional network named the Graph-embedded Aggregation of Global and Local features (GaGL) for classroom student behavior recognition. Specifically, we combine global features with textual context to reconstruct enriched global representations. Similarly, local features are fused with localized textual information to generate refined local structures. An embedded graph is then constructed to aggregate these two feature types, capturing their interdependencies while minimizing redundant edges—effectively reducing noise from insignificant connections.

Our key contributions are summarized as follows:

- We propose a unified multimodal graph embedding framework (GaGL) that, for the first time, integrates global scene context, local object features, and semantic text annotations into a single, heterogeneous graph structure. This holistic approach enables nuanced reasoning about complex classroom behaviors.
- We introduce a novel Vectorized Similarity Learning mechanism that replaces conventional scalar metrics with a learnable, vector-based function. This allows the model to capture fine-grained cross-modal associations between visual and textual cues with much higher fidelity.
- We design an Efficient and Directed Graph Propagation method where graph edges are dynamically pruned based on a learnable criterion. This significantly reduces computational redundancy and noise from irrelevant connections, leading to a more scalable and accurate model.
- We introduce D-Edu, a new, large-scale dataset for educational behavior analysis, and demonstrate through extensive experiments that our GaGL model achieves state-of-the-art performance.

2. **Related Works.**

2.1. **Image-Text Feature Fusion.**

Feature Encoding Methods [8, 9] focus on extracting image features and text features separately before combining them. For image feature extraction [10], various techniques exist, including region-based feature extraction [11], segmentation-based feature extraction [12], and frequency-based feature extraction [13]. Basly et al. [14] employed attention-based convolutional networks with self-attention mechanisms to integrate global image features with local features. They also established correlations between region features based on the representational capabilities of both visual and textual data. While researchers have attempted to enhance critical feature acquisition by building such correlations, existing methods remain rudimentary and lack unification in encoding key features. These approaches are particularly limited in addressing the complexities of classroom behavior recognition, where harmonizing multimodal representations and converging on discriminative features remain significant challenges.

2.2. **Graph-Based Feature Embedding.**

Existing studies [15, 16, 17] focus on developing convolutional graph networks to link multimodal data such as images, text, and video. Mubarak et al. [18] proposed a Graph Convolutional Network (GCN) for computer vision tasks, while Li et al. [19] designed a graph-based convolutional framework to align image and text modalities. Feng et al. [20] further introduced a method to model semantic relationships between images and text, constructing image regions and graph-based representations to enrich contextual knowledge for image recognition. While these

efforts aim to enhance recognition quality by jointly modeling visual and textual data via graphs, most existing frameworks remain computationally heavy, incurring high storage and operational costs in real-world deployment.

3. **Proposed Method.** The core principle of our GaGL framework is the effective fusion of two complementary types of visual information: global features and local features. Before detailing the model architecture, we define these concepts in the context of our work.

Global Features refer to a single, compact feature vector that summarizes the entire input image. This vector captures the holistic, high-level context of the scene, such as the overall classroom layout and the general positioning of students. In our pipeline, we utilize a Convolutional Neural Network (CNN) like ResNet101 to process the whole image and generate this single global representation.

Local Features, in contrast, are a collection of feature vectors where each vector corresponds to a specific, semantically meaningful region or object within the image. These features provide fine-grained details necessary to identify key items or actions. They are extracted by first using models like Faster R-CNN or Mask R-CNN to detect objects of interest (e.g., phones, faces, papers) or segment pertinent areas (e.g., a hand holding a phone), and then deriving a distinct feature vector for each region.

The fundamental difference lies in their scope and the information they carry: global features provide the overall context, while local features provide specific evidence. Our proposed method is designed to leverage the synergy between them, as robust behavior recognition requires grounding local, detailed observations within the broader context of the scene.

3.1. **Problem Definition.** The integration of multimodal features, encompassing both visual and textual data, plays a pivotal role in achieving robust image recognition within complex domains such as classroom behavior analysis. Existing methodologies typically extract visual and textual representations in isolation—relying, for instance, on region-based image features and tokenized word embeddings—and subsequently merge these features through simplistic alignment strategies. However, such strategies often fail to fully capture the nuanced interactions between visual regions and textual tokens, leading to suboptimal performance in tasks requiring fine-grained semantic comprehension.

Moreover, conventional methods typically compute inter-modal similarity using scalar metrics (e.g., cosine distance), which provide limited expressiveness in modeling intricate relationships. Graph-based approaches for multimodal reasoning offer a promising alternative, yet many existing implementations suffer from two core drawbacks: (1) computational inefficiency arising from the inclusion of redundant edges, and (2) the use of undirected message propagation, which may introduce excessive noise and degrade scalability in real-world scenarios.

In response to these limitations, this work introduces a novel framework for classroom behavior recognition that addresses the shortcomings of current practices in three major components. First, Unified Feature Encoding combines region-level visual features $\{v_1, \ldots, v_K\}$ and word-level textual embeddings $\{t_1, \ldots, t_L\}$ by means of adaptive attention mechanisms, thereby generating enriched global $(\bar{v}, \bar{t})$ and local (a_j^v) representations that preserve both granular and holistic semantics. Second, Vectorized Similarity Learning replaces conventional scalar similarity metrics with learnable vector-based functions $s(x, y; W)$. This substitution enhances the model's capacity to capture fine-grained cross-modal associations, enabling more precise alignment of semantic cues between visual regions and textual tokens. Third, Efficient Graph Propagation employs a directed

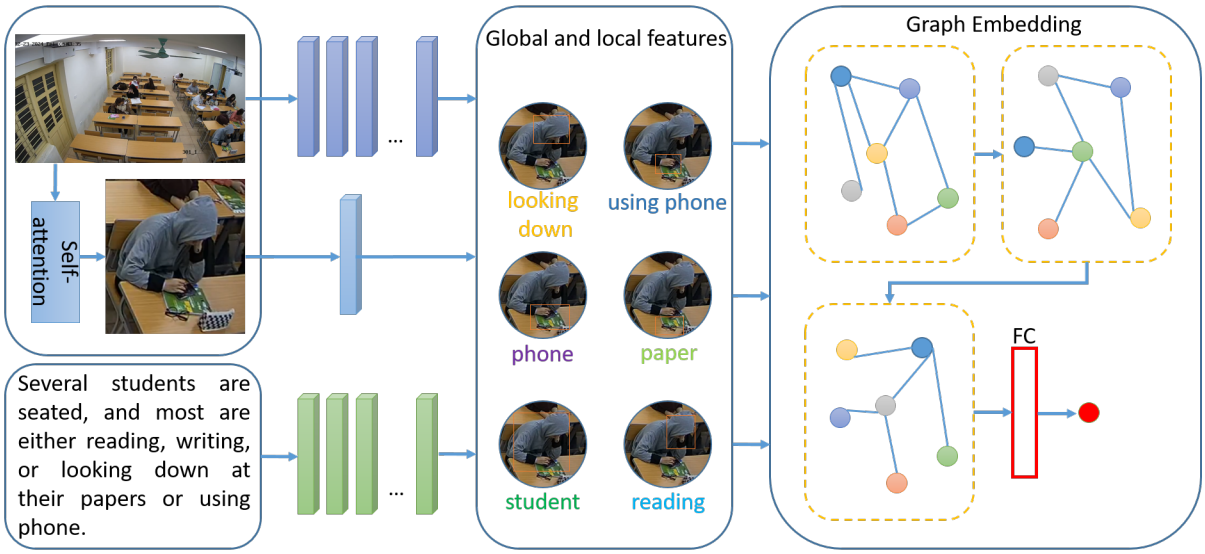

FIGURE 1. The framework of the proposed GaGL

similarity graph $N = \{s_1^l, \ldots, s_L^l, s_g\}$ whose edges are dynamically pruned according to a learnable criterion, thereby reducing redundancy while retaining critical node interactions.

By integrating these three components—unified encoding, vectorized similarity, and efficient graph-based propagation—the proposed framework overcomes the scalability issues and oversimplified similarity models characteristic of existing methods. Consequently, it achieves more precise behavior recognition in classroom environments without incurring prohibitive computational overhead.

3.2. **Model Architecture.** In an effort to advance classroom behavior recognition and enhance overall teaching quality, we propose a comprehensive, four-stage pipeline integrating multimodal data from both visual and textual sources, show in Figure 1. The first stage, the Input Layer, ingests video frames or still images captured by classroom cameras, depicting students in various postures and states of engagement. These images may show students looking down, holding a phone, reading papers, or sitting attentively. To enrich the contextual cues provided by the visual data, we incorporate supplementary textual annotations or labels (e.g., "using phone," "reading"), which can either be manually crafted or derived from existing knowledge bases that map specific behaviors to descriptive keywords.

In the second stage, Feature Extraction, we separately handle Visual and Textual features. On the visual side, a convolutional neural network (CNN) such as ResNet101 is employed to obtain global features, capturing holistic image characteristics (e.g., the classroom layout, student positioning, and high-level scene context). Subsequently, local features are derived through object detection using Faster R-CNN, which precisely identifies items of interest (e.g., phones, papers, and faces), and through semantic segmentation with Mask R-CNN, which isolates pertinent regions of the image (e.g., a hand holding a phone, eyes looking downward). To further refine the relevance of specific visual elements, we adopt a self-attention mechanism, inspired by Transformer-based architectures [21], allowing the system to focus on critical factors such as a phone in a student's hand. Meanwhile, on the textual side, Textual Features are extracted via a pretrained language model (BERT), which transforms captions, keywords, and descriptive labels into dense semantic embeddings. As a result, phrases like "using phone" are converted into high-dimensional vectors that preserve rich contextual information about the behavior in question.

The third stage, Multimodal Graph Embedding, fuses these visual and textual features by constructing a graph whose nodes represent global image embeddings $\bar{v}$, localized image embeddings $(v_1, \ldots, v_K)$, token-level text embeddings $(t_1, \ldots, t_L)$, and a global text embedding $\bar{t}$. Connections (edges) among these nodes are formed by assessing the semantic alignment between visual cues and textual descriptions. Crucially, instead of using a fixed similarity score, the weight and existence of these edges are determined by our learnable vectorized similarity module. Furthermore, during propagation, we apply our dynamic pruning mechanism to filter out weak or irrelevant connections, ensuring that the subsequent Graph Attention Network (GAT) operates on a refined, low-noise graph. This directed and pruned structure ensures that both global and localized data streams are maintained in a unified framework, enabling a richer contextual interplay between images and text. For instance, nodes corresponding to a detected phone in the visual domain may be strongly linked to the textual node representing the concept "using phone." Additionally, directed internal edges connect interdependent visual elements (e.g., "hand holding phone" $\rightarrow$ "eyes looking down") to reduce noise and highlight co-occurring features. This graph-based representation ensures that both global and localized data streams are maintained in a unified framework, enabling a richer contextual interplay between images and text.

Finally, in the Behavior Classification stage, a Graph Attention Network (GAT) aggregates information from the multimodal graph by adaptively weighing the significance of each node and edge. After the GAT refines and propagates these context-aware embeddings, a Softmax classifier produces the final behavior predictions. Common classroom-related behavior classes include using phone (characterized by phone detection plus a downward gaze), reading (presence of documents with focused eye direction), attention (upright posture and eye line aimed at the instructor or board), and taking notes (pen or pencil in hand alongside an open notebook or paper). By effectively synthesizing visual cues and textual annotations within a graph-based learning paradigm, this approach delivers robust, fine-grained recognition of student behaviors, ultimately contributing to improved monitoring, feedback, and pedagogical strategies in modern classroom environments.

3.3. **Loss function.** Below is an example of how the loss function can be formulated for the proposed behavior recognition pipeline. The overall objective typically consists of multiple components: (1) Detection and Segmentation Losses to guide object localization and region segmentation, (2) Multimodal Alignment Loss to ensure consistency between visual and textual features, and (3) Classification Loss for the final behavior prediction.

For the object detection task (e.g., identifying phones, papers, faces), the loss function usually follows the standard Faster R-CNN framework, which combines:

- Classification Loss, $\mathcal{L}_{\text{cls}}$: A cross-entropy loss that classifies each detected region as one of the object categories (e.g., "phone," "paper," "face") or as background
- Bounding Box Regression Loss, $\mathcal{L}_{\text{reg}}$: A smooth L1 (or L2) loss that refines the bounding box coordinates of detected objects.

The total detection loss is:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}} \tag{1}$$

For semantic or instance segmentation (e.g., segmenting a hand holding a phone or the region around the eyes), Mask R-CNN introduces an additional mask prediction branch. The segmentation loss, $\mathcal{L}_{\text{mask}}$, is typically a pixel-wise binary cross-entropy loss evaluated over each predicted mask compared to the ground-truth mask.

Once both visual and textual features are extracted, we establish correspondences between region-based visual embeddings $(v_1, \ldots, v_K)$ and textual embeddings $(t_1, \ldots, t_L)$. To strengthen the alignment between matching pairs (e.g., "phone" - visual region containing a phone) and push away non-matching pairs (e.g., "phone" - a region containing a book), a similarity-based loss or a contrastive loss can be used. One common strategy is a triplet loss:

$$\mathcal{L}_{\text{triplet}} = \max(0, \alpha + d(v_i, t^+) - d(v_i, t^-)), \tag{2}$$

where t^+ is a textual embedding that is relevant to the visual feature v_i, t^- is an irrelevant textual embedding, $d(\cdot, \cdot)$ is a distance metric (e.g., cosine distance or Euclidean distance), and α is the margin. This encourages correct matches to lie closer in the embedding space than incorrect ones.

Alternatively, a cosine similarity loss or cross-entropy loss over matching pairs can be adopted, depending on the specific design. The key objective is to ensure that each visual feature and its corresponding textual descriptor reinforce one another, improving the quality of graph-based node connections in the subsequent steps.

After constructing the multimodal graph and passing node representations through the Graph Attention Network (GAT), the final output layer classifies each sample into one of the predefined behavior categories (e.g., using phone, reading, attention, taking notes). The classification loss is typically a cross-entropy loss defined as:

$$\mathcal{L}_{\text{class}} = -\sum_{c=1}^{C} y_c \log(p_c), \tag{3}$$

where C is the number of classes (behavior categories), y_c is the binary indicator (0 or 1) for whether class c is the correct classification, and p_c is the predicted probability for class c.

These different loss terms are combined into a single multi-task objective, often as a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{triplet}} + \mathcal{L}_{\text{class}}, \tag{4}$$

The system is trained end to end (or in stages, depending on computational constraints), iteratively refining the network parameters to accurately detect objects, align visual and textual features, and classify student behaviors with high precision.

By coupling losses at the object, segmentation, multimodal alignment, and classification levels, this holistic loss formulation promotes robust feature learning that captures both fine-grained details (e.g., hands holding a phone) and high-level contextual cues (e.g., student posture, textual annotations). Consequently, the pipeline can more reliably differentiate nuanced classroom behaviors, ultimately empowering educators with richer insights for instructional improvement.

3.4. **Implementation Details.** To ensure the reproducibility and clarity of our proposed GaGL framework, this section outlines detailed architectural configurations and hyperparameters. Training-specific hyperparameters such as the optimizer and learning rate are elaborated in Section 4.2; here, we focus primarily on core model specifications and preprocessing steps.

Data Preprocessing: All input images from both datasets are resized uniformly to a resolution of 224×224 pixels. Before feature extraction, images undergo standard normalization procedures consistent with typical practices used in convolutional neural network training.

Feature Extractor Modules: Global visual features are extracted using a ResNet-101 model pre-trained on the ImageNet dataset, resulting in a 2048-dimensional global feature vector per image. For local visual features, a Faster R-CNN detector generates exactly 18 region proposals per image, each encoded into a 2048-dimensional feature vector. Textual features are derived from a BERT-base-uncased model. Specifically, the global text embedding t corresponds to the 768-dimensional representation of the [CLS] token, while local text embeddings $\{t_1, \ldots, t_L\}$ correspond to the 768-dimensional hidden states from the final layer for each word token.

Graph Construction and Propagation: The vectorized similarity between visual and textual features is computed using a two-layer Multi-Layer Perceptron (MLP). This similarity function $s(x, y; W)$ takes as input the concatenation of visual (2048-dimensional) and textual (768-dimensional) vectors, passing through a hidden layer of 512 neurons, and outputs a similarity vector of 256 dimensions. For graph pruning, we employ a top-k strategy, keeping only the five strongest connections per node based on similarity scores, thus constructing a refined graph before passing it to the Graph Attention Network (GAT). The GAT module comprises two layers, each featuring four parallel attention heads, and maintains a feature dimension of 512 throughout the network.

Classifier and Loss Function: The final behavioral prediction leverages a two-layer MLP classifier. The input to this classifier is the aggregated 512-dimensional graph feature vector, processed through a hidden layer of 256 neurons, and culminating in an output layer with C neurons (where C denotes the number of classes), followed by a softmax activation. The overall training loss is a weighted combination of multiple task-specific losses, defined as follows:

$$L_{total} = \lambda_{det} L_{det} + \lambda_{mask} L_{mask} + \lambda_{triplet} L_{triplet} + \lambda_{class} L_{class}. \tag{5}$$

In our experiments, we empirically set the loss coefficients to $\lambda_{det} = 1.0$, $\lambda_{mask} = 1.0$, $\lambda_{triplet} = 0.5$, and $\lambda_{class} = 1.0$.

4. **Experiments.**

4.1. **Dataset.** We evaluated our model on two datasets: MSCOCO [22] and D-Edu. The MSCOCO dataset includes 123,287 images, each annotated with 5 captions. It is divided into 113,287 images for training, 5,000 for validation, and 5,000 for testing. The captions in MSCOCO are descriptive of general scenes. For instance, a single image might have five captions such as: (1) 'A person is using a laptop on a wooden table,' (2) 'The kitchen is sunlit and a man is working on a computer,' (3) 'A person sits at a kitchen island with a silver laptop,' (4) 'A man types on a laptop in a modern kitchen,' and (5) 'The view of a person using a computer in a residential kitchen. We report performance by averaging over 5 folds of 1,000 test images and the entire 5,000 test images. On the other hand, the D-Edu dataset, which focuses on educational classroom imagery, contains 22,476 images each with 9 manually annotated captions. This dataset is split into 20,476 images for training, with 1,000 images each for validation and testing. D-Edu includes 9 specific action labels for students, such as using a phone, turning sideways or vertically, raising hand towards the board, talking, teasing, fighting, writing, and looking at the board, with each label averaging around 2,000 images, though numbers can slightly vary.

A key feature of the D-Edu dataset is the richness of its annotations. The nine captions for each image were intentionally designed to be multi-faceted, describing not just the core action but also the context, objects, and posture. This provides a diverse semantic signal for multimodal learning. For an image with the action label 'using a phone', the nine captions might include:

- Action-focused: 'A student is looking down and tapping on a bright screen.'
- Object-focused: 'The girl in the blue shirt is holding a smartphone under her desk.'
- Gaze-focused: 'Her gaze is directed at a personal electronic device, not the teacher.'
- Posture-focused: 'The student's body is hunched forward, a posture typical of phone use.'
- Interpretive: 'The student appears disengaged from the class activity and is using her phone.'
- Contextual: 'Although a textbook is open on the desk, her attention is on the mobile phone.'
- Formal/Neutral: 'An individual in a classroom setting is interacting with a handheld device.'
- Explicit Label: 'This image shows a student who is using a phone during class time.'
- Location-specific: 'A phone is partially visible in the student's lap below the table.'

This comprehensive annotation strategy is crucial for training a model like GaGL to understand the subtle cues of classroom behavior.

4.2. **Experiment setup.** Our experimental setup involves processing images with the Faster-RCNN using a ResNet-101 backbone from [24] to generate 18 region proposals, each with a 2048-dimensional feature vector, alongside semantic segmentation via Mask R-CNN. Textual data is handled by setting word embeddings to 224 dimensions and hidden states to 512 per sentence. The similarity vector m is configured with a dimension of 256, employing a smoothing temperature $\lambda = 6$, N = 3 reasoning steps, and a margin $\gamma = 0.1$. We utilize the AdamW optimizer [23] for training the GaGL network, with a batch size of 128. On the MSCOCO dataset, the learning rate starts at 0.00001 for the first 10 epochs, dropping to 0.000001 for the next 10. For the D-Edu dataset, the SGR (SAF) module begins training with a learning rate of 0.00001 for 30 (20) epochs, followed by a decay by a factor of 0.01 for an additional 10 epochs. Textual features are derived using a pretrained BERT model, converting text into dense semantic embeddings. We implement early stopping at epoch 58 to avoid overfitting, choosing the model snapshot with the best validation performance for final evaluation.

Our experiments were conducted to address two key research questions (RQs):

- **RQ1:** How well does the GaGL model perform compared to state-of-the-art models?
- **RQ2:** How does the GaGL model predict in the task of recognizing student behavior in a classroom setting?

4.3. **Performance Compare (RQ1).** The results presented in the table 1 provide a comparative evaluation of various object detection and segmentation methods across two datasets: MSCOCO and D-Edu. The evaluation metrics include Precision, Recall, and F1 measure, which are essential for assessing model performance in detecting classroom-related behaviors. To comprehensively evaluate the performance of our GaGL model and other baseline methods, we adopt three standard metrics: Precision, Recall, and the F1-score. The selection of these metrics is deliberately aligned with the practical requirements and ethical considerations of classroom behavior recognition.

- Precision is critical for ensuring the reliability of the system from an educator's perspective. It measures the accuracy of positive predictions, and a high precision rate minimizes the risk of falsely identifying a student as engaging in negative behavior (e.g., "using phone"), which is crucial for maintaining trust and avoiding unfair judgment.
- Recall is vital for ensuring the system's thoroughness. It measures the model's ability to detect all actual instances of a given behavior. High recall is necessary

TABLE 1. A comparative analysis against contemporary methods over the MS COCO and D-Edu dataset

	MSCOCO			D-Edu		
Method	**Precision**	**Recall**	**F1 measure**	**Precision**	**Recall**	**F1 measure**
Faster-RCNN [25]	-	0.93	-	0.9123	0.9276	0.9199
Mask R-CNN [26]	0.91	-	-	0.8951	0.9254	0.9095
SAM [27]	0.8585	0.9094	0.8834	0.8343	0.8877	0.8602
Yolov11 [28]	0.933	-	-	0.9275	0.9367	0.9321
GaGL (ours)	**0.9453**	**0.9564**	**0.95**	**0.9352**	**0.9583**	**0.9466**

for administrators who need a complete picture of classroom activities to effectively "monitor learning conditions and enhance educational quality".

- The F1-score, as the harmonic mean of Precision and Recall, offers a balanced assessment, which is especially important for datasets like D-Edu where certain behaviors may be less frequent than others. It ensures that the model's performance is robust and not skewed towards either precision or recall, making it a reliable indicator of overall effectiveness in real-world classroom scenarios.

The evaluation of traditional object detection models, including Faster R-CNN and Mask R-CNN, reveals that while these methods achieve high recall, their reported precision values are sometimes incomplete. Specifically, Faster R-CNN attains a recall of 0.93 on the MSCOCO dataset; however, its corresponding precision value is not documented. Conversely, Mask R-CNN reports a precision of 0.91 but lacks an explicit recall measurement, limiting a comprehensive assessment of its overall performance.

The Segment Anything Model (SAM) provides both precision and recall metrics; however, its performance remains suboptimal compared to other methods. Although the F1-score for MSCOCO is not reported, its performance on the D-Edu dataset (F1 = 0.8602) suggests challenges in accurately capturing classroom-related behaviors.

YOLOv11 demonstrates strong precision, achieving a value of 0.933 on MSCOCO, though its recall remains unreported. On D-Edu, it maintains robust detection capabilities with an F1-score of 0.9321, reinforcing its suitability for real-time object detection tasks.

The proposed GaGL model outperforms all other approaches across both datasets, achieving the highest recorded precision, recall, and F1-score. On MSCOCO, GaGL attains a precision of 0.9453, a recall of 0.9564, and an F1-score of 0.95. Similarly, on the more specialized D-Edu dataset, it maintains superior accuracy with a precision of 0.9352, a recall of 0.9583, and an F1-score of 0.9466. These high scores are not just numerically superior; they signify a model that is both reliable (high precision, minimizing false accusations) and comprehensive (high recall, capturing most behaviors), making it a practically viable tool for educators. This balanced excellence, reflected in the top-tier F1-score, confirms that our multimodal graph embedding strategy effectively addresses the nuanced challenges of classroom behavior recognition. These results indicate that the multimodal graph embedding strategy effectively integrates both global and local image-text interactions, leading to enhanced classification accuracy for classroom behavior recognition.

The superior performance of GaGL can be attributed to its four-stage pipeline, which effectively utilizes multimodal data fusion. By integrating CNN-based feature extraction, semantic segmentation, and graph-based attention mechanisms, the model achieves a comprehensive contextual understanding. The representation of student behaviors through graph embeddings further enhances classification accuracy, surpassing conventional methods that rely primarily on object detection.

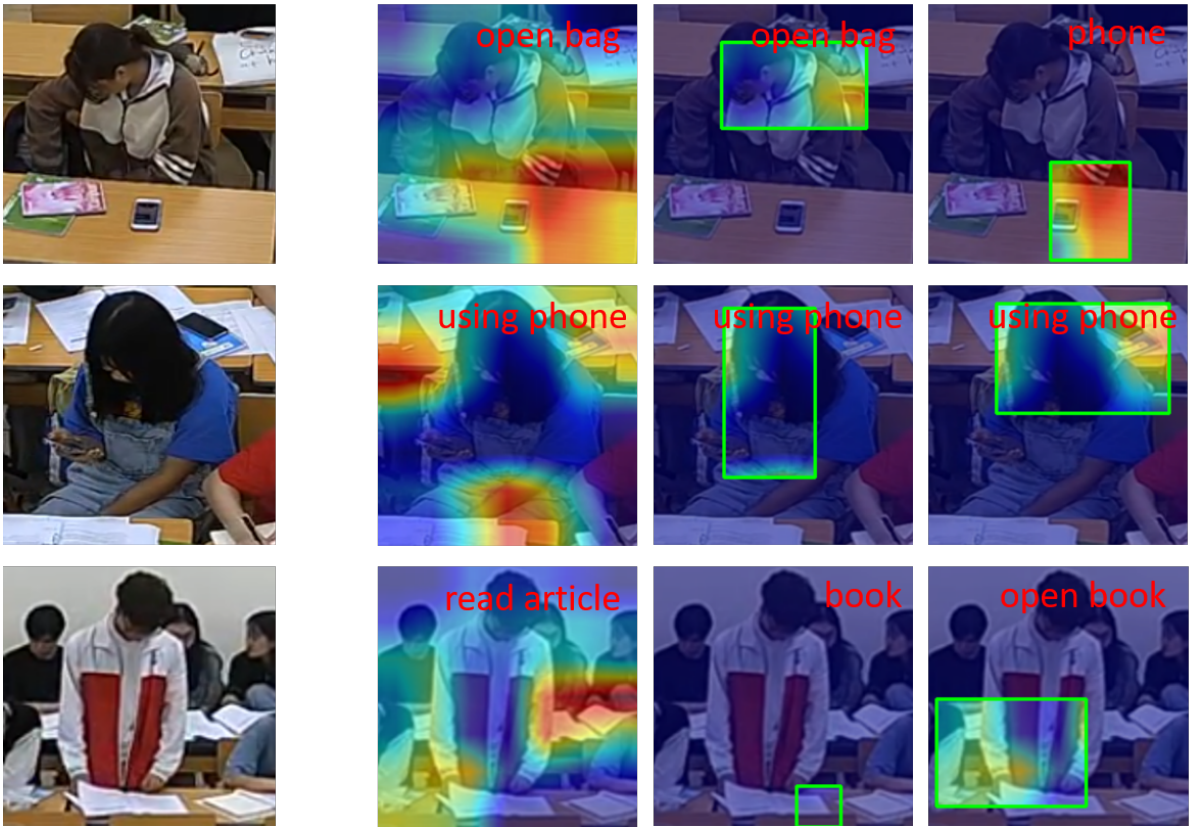

FIGURE 2. The detection and semantic results of GaGL model.

4.4. **Qualitative Study (RQ2).** Figure 2 illustrates qualitative examples of the GaGL model's performance in detecting and interpreting student behaviors within the classroom setting. Each column corresponds to a predicted behavior (e.g., "open bag," "using phone," "read article," "open book"), while the accompanying heatmaps and bounding boxes highlight salient regions that guided the classification. Notably, the model accurately localizes objects of interest such as phones and books, as well as relevant body parts (e.g., hands, faces) associated with these actions. For instance, in the top row, the bounding box around the phone and the high-intensity regions in the heatmap both indicate the model's focus on the student's hand and device, thereby reinforcing the "using phone" prediction. Similarly, in the lower rows, the highlighted areas around the student's torso and the detected books/papers confirm reading-related behaviors. These examples underscore the model's ability to capture nuanced interactions—such as hand-object contact or gaze direction—and fuse them with contextual cues (e.g., posture, surrounding objects) through its multimodal graph representation. Consequently, GaGL demonstrates robust and fine-grained behavior recognition, facilitating more accurate and interpretable monitoring of student engagement in classroom environments.

To assess the practical applicability of the GaGL framework, we analyzed its computational complexity by measuring the inference time. The experiments were conducted on a server equipped with an NVIDIA Tesla V100 GPU. The total processing time for a single image was approximately 156 ms, which translates to a throughput of roughly 6.4 frames per second (FPS). The breakdown of the average inference time across the main components of our pipeline is presented in Table 2.

The results indicate that the primary computational bottleneck lies in the visual feature extraction stage, which was designed to maximize accuracy using deep and complex

TABLE 2. Inference Time Analysis of the GaGL Pipeline

Module	**Description**	**Average Time (ms)**
Visual Feature Extraction	ResNet-101 + Faster/Mask R-CNN for global/local features	120
Textual Feature Extraction	BERT encoding for descriptive labels	10
Graph Construction & Propagation	Vectorized similarity, pruning, and GAT message passing	25
Behavior Classification	Final Softmax classifier	<1
Total Inference Time		**∼156 ms**

models. While the current performance is insufficient for live, real-time video processing, the GaGL framework is highly effective for offline analysis tasks, where high precision and detailed behavioral understanding are more critical than processing speed. This allows for in-depth post-session analysis of classroom recordings by educators.

For future real-time applications, several optimization strategies could be explored. These include replacing the current backbone with a more lightweight architecture (e.g., MobileNet, EfficientDet), applying model compression techniques like quantization and knowledge distillation, and implementing a temporal frame sampling strategy to reduce the processing load.

5. **Conclusions.** In this paper, we have presented the Graph-embedded Aggregation of Global and Local features (GaGL) network, a novel approach tailored for classroom behavior recognition. By systematically integrating global scene context, localized object and action cues, and textual annotations into a unified graph representation, the proposed model effectively captures both high-level and fine-grained behaviors. Experimental evaluations on MSCOCO and the newly introduced D-Edu dataset demonstrate the superiority of our method over existing techniques, validating its robustness and accuracy in real-world classroom scenarios. Furthermore, ablation studies confirm the complementary nature of the global, local, and textual feature modules, underscoring the benefits of fusing diverse information sources. Through these findings, GaGL not only addresses the complexities of classroom behavior recognition but also offers a scalable framework for broader applications in image recognition and computer vision. Future work will explore the integration of additional modalities and real-time processing capabilities, thereby further enhancing the potential impact of this approach on classroom management and educational research.

REFERENCES

[1] Shinoda, Hirofumi, Tsuyoshi Yamamoto, and Kyoko Imai-Matsumura. "Teachers' visual processing of children's off-task behaviors in class: A comparison between teachers and student teachers." PLoS One 16.11 (2021): e0259410.

[2] Chen, Wei, et al. "Deep learning for instance retrieval: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.6 (2022): 7270-7292.

[3] Maharana, Kiran, Surajit Mondal, and Bhushankumar Nemade. "A review: Data pre-processing and data augmentation techniques." Global Transitions Proceedings 3.1 (2022): 91-99.

[4] Ting-Ting Wu and Tien-Wen Sung, "Analysis of the Effects of a Game-Based Review System Integrated with the Hierarchy of Learning on Learning Outcomes in an Elementary Social Science Course," Interactive Learning Environments, Vol. 31, No. 6, August, 2023.

[5] Sun, Bo, et al. "Student Class Behavior Dataset: a video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes." Neural Computing and Applications 33 (2021): 8335-8354.

[6] Li, Weisheng, and Lin Huang. "YOLOSA: Object detection based on 2D local feature superimposed self-attention." Pattern Recognition Letters 168 (2023): 86-92.

[7] Liu, Yang, et al. "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition." IEEE Transactions on Image Processing 30 (2021): 5573-5588.

[8] Meng, Linhao, et al. "Class-constrained t-sne: Combining data features and class probabilities." IEEE Transactions on Visualization and Computer Graphics (2023).
[9] Hamza, Ameer, et al. "Multimodal religiously hateful social media memes classification based on textual and image data." ACM Transactions on Asian and Low-Resource Language Information Processing 23.8 (2024): 1-19.
[10] Abdallah, Taoufik Ben, Islam Elleuch, and Radhouane Guermazi. "Student behavior recognition in classroom using deep transfer learning with VGG-16." Procedia Computer Science 192 (2021): 951-960.
[11] Pabba, Chakradhar, Vishal Bhardwaj, and Praveen Kumar. "A visual intelligent system for students' behavior classification using body pose and facial features in a smart classroom." Multimedia Tools and Applications 83.12 (2024): 36975-37005.
[12] Li, Liangwei, et al. "Analysis of Classroom Teaching Interaction Patterns Based on Behavior Recognition." 2024 Twelfth International Conference on Advanced Cloud and Big Data (CBD). IEEE, 2024.
[13] Arpasat, Poohridate, et al. "Applying process mining to analyze the behavior of learners in online courses." International Journal of Information and Education Technology 11.10 (2021): 436-443.
[14] Basly, Hend, Mohamed Amine Zayene, and Fatma Ezahra Sayadi. "Spatiotemporal Self-Attention Mechanism Driven by 3D Pose to Guide RGB Cues for Daily Living Human Activity Recognition." Journal of Intelligent & Robotic Systems 109.1 (2023): 2.
[15] Pang, Shiyan, et al. "Graph convolutional network for automatic detection of teachers' nonverbal behavior." Computers and Education: Artificial Intelligence 5 (2023): 100174.
[16] Duhme, Michael, Raphael Memmesheimer, and Dietrich Paulus. "Fusion-gcn: Multimodal action recognition using graph convolutional networks." DAGM German conference on pattern recognition. Cham: Springer International Publishing, 2021.
[17] Hu, Kun, et al. "Graph fusion network-based multimodal learning for freezing of gait detection." IEEE Transactions on Neural Networks and Learning Systems 34.3 (2021): 1588-1600.
[18] Mubarak, Ahmed A., et al. "Modeling students' performance using graph convolutional networks." Complex & Intelligent Systems 8.3 (2022): 2183-2201.
[19] Li, Yujie, et al. "Fuzzy Multimodal Graph Reasoning for Human-Centric Instructional Video Grounding." IEEE Transactions on Fuzzy Systems (2024).
[20] Feng, Duoduo, Xiangteng He, and Yuxin Peng. "MKVSE: Multimodal knowledge enhanced visual-semantic embedding for image-text retrieval." ACM Transactions on Multimedia Computing, Communications and Applications 19.5 (2023): 1-21.
[21] Bani-Almarjeh, Mohammad, and Mohamad-Bassam Kurdy. "Arabic abstractive text summarization using RNN-based and transformer-based architectures." Information Processing & Management 60.2 (2023): 103227.
[22] Tong, Kang, and Yiquan Wu. "Rethinking PASCAL-VOC and MS-COCO dataset for small object detection." Journal of Visual Communication and Image Representation 93 (2023): 103830.
[23] Llugsi, Ricardo, et al. "Comparison between Adam, AdaMax and Adam W optimizers to implement a Weather Forecast based on Neural Networks for the Andean city of Quito." 2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM). IEEE, 2021.
[24] Sahu, Sonal, Satya Prakash Sahu, and Deepak Kumar Dewangan. "Pedestrian detection using ResNet-101 based Mask R-CNN." AIP Conference Proceedings. Vol. 2705. No. 1. AIP Publishing, 2023.
[25] Liu, Yuhang, et al. "Spark plug defects detection based on improved Faster-RCNN algorithm." Journal of X-Ray Science and Technology 30.4 (2022): 709-724.
[26] Yayla, Ridvan, Emir Albayrak, and Ugur Yuzgec. "Vehicle detection from unmanned aerial images with deep mask R-CNN." Computer Science Journal of Moldova 89.2 (2022): 148-169.
[27] Chen, Tianle, et al. "Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation." arXiv preprint arXiv:2305.05803 (2023).
[28] Sapkota, Ranjan, et al. "Comprehensive performance evaluation of yolo11, yolov10, yolov9 and yolov8 on detecting and counting fruitlet in complex orchard environments." arXiv preprint arXiv:2407.12040 (2024).