



Automatic Classroom Learning Behavior Recognition Based on a Spatiotemporal Attention Graph Convolutional Network for Music Teaching Decision Support

Yang Cao 

Xi'an University, Xi'an 710065, China

Corresponding Author Email: yanzhao@xawl.edu.cn

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

<https://doi.org/10.18280/ts.430428>

ABSTRACT

Received: 29 March 2026

Revised: 18 June 2026

Accepted: 24 August 2026

Available online: 31 August 2026

Keywords:

classroom behavior recognition, spatiotemporal graph convolutional network, spatial attention, temporal attention, skeleton-based action recognition, observable classroom activity profile, music education, instructional decision support

Automatic recognition of student behavior provides a systematic means of characterizing observable classroom activity and may offer useful information for instructional decision-making. However, many existing classroom behavior recognition methods rely primarily on appearance features extracted from individual images, limiting their ability to represent the temporal evolution of student actions. This study develops a spatiotemporal attention graph convolutional network (STAGCN) for skeleton-based recognition of continuous student behaviors and investigates its potential application to music teaching decision support. The framework is developed in the context of the publicly available CStudentAct dataset, which contains temporally connected instances of four classroom behaviors: raising hand, using phone, sleeping, and standing. Student-centered sequences are constructed from annotated action instances, and two-dimensional human poses are extracted to form spatiotemporal skeleton graphs. STAGCN combines adaptive graph convolution (AGC) with spatial and temporal attention (TA) to model behavior-dependent joint relationships, emphasize informative skeletal regions, and represent temporal dependencies across action sequences. Its architectural characteristics are examined in relation to representative skeleton-based action-recognition approaches through methodological comparison, class-level behavioral analysis, architectural decomposition, and mathematical analysis of the spatial and TA mechanisms. This analysis clarifies the complementary representational roles of adaptive topology learning, spatial attention (SA), and TA without treating results obtained under different datasets or experimental protocols as directly comparable numerical evidence. Beyond behavior recognition, the four classroom behavior categories are organized into an observable classroom activity profile comprising participatory activity (PA), movement-related activity (MA), and off-task activity (OT). Their temporal distributions are incorporated into a behavior-to-strategy mapping framework that links observable classroom conditions with candidate music teaching responses while retaining teacher judgement in the final instructional decision. The framework does not treat visual behavior as a direct measure of motivation, comprehension, cognitive engagement, or musical achievement, nor does it assume that the proposed teaching responses constitute empirically validated pedagogical interventions. Instead, the study establishes a technically grounded connection between continuous classroom behavior recognition and data-informed instructional decision support, providing a foundation for future validation using dedicated music-classroom data and controlled teaching interventions.

1. INTRODUCTION

The increasing use of computer vision and artificial intelligence in educational environments has created new possibilities for observing how students participate in classroom activities. Learning behaviors such as listening, reading, writing, raising a hand, interacting with peers, and engaging in off-task activities provide observable information about students' participation during instruction. Traditionally, these behaviors have been examined through classroom observation, questionnaires, video coding, or teacher judgement. Although these approaches remain useful for pedagogical evaluation, continuous observation of multiple students is labor-intensive and may be influenced by

differences among observers. Automated visual analysis provides an alternative means of extracting behavioral information from classroom recordings and has consequently received increasing attention in intelligent education and learning analytics [1-3].

Recent advances in deep learning have substantially improved the detection and recognition of student behaviors in complex classroom scenes. Convolutional neural networks and object-detection architectures have been applied to classroom images and videos, while multi-scale feature representations, attention mechanisms, pose information, and relational modeling have been introduced to address variations in student scale, partial occlusion, crowded seating, and complex backgrounds [2-5]. The development of classroom-

specific datasets has further supported this line of research. The Student Classroom Behavior (SCB) dataset series, for example, provides densely annotated classroom images covering a range of common student behaviors [6, 7]. These datasets and related classroom-video studies provide a practical basis for evaluating visual recognition methods under conditions that are considerably closer to authentic educational environments than conventional laboratory action-recognition settings [4-7].

Despite this progress, a large proportion of classroom behavior recognition methods formulate the task primarily as image detection or frame-level classification. Such approaches are effective for visually distinctive states but provide a limited representation of behaviors whose characteristics evolve over time. Raising a hand, for example, involves a coordinated change in the positions of the shoulder, elbow, and wrist, while standing contains a transition between substantially different body configurations. Mobile-phone use and sleeping may instead involve more localized movements or sustained postural patterns. Classroom behavior is therefore not merely a collection of static visual states but a dynamic process in which body configurations evolve across consecutive frames.

Skeleton-based action recognition provides a structured representation for modeling this temporal process. Human pose estimation converts RGB observations into anatomical keypoints, reducing dependence on background appearance, clothing, illumination, and other scene-specific visual information. A sequence of poses can naturally be represented as a graph in which body joints constitute nodes, anatomical connections form spatial edges, and corresponding joints across consecutive frames form temporal edges. Spatial-temporal graph convolutional networks (ST-GCNs) established an influential framework for learning spatial and temporal patterns directly from skeleton sequences [8]. Subsequent methods extended this formulation through adaptive graph structures, multi-scale spatial-temporal reasoning, and channel-wise topology refinement [9-11]. These developments demonstrate the capacity of graph-based models to capture complex relationships among human joints and their temporal evolution.

Directly applying conventional skeleton-based action recognition methods to classroom environments nevertheless presents several challenges. First, the contribution of individual body joints varies among classroom behaviors. Upper-limb configuration is particularly informative for raising-hand behavior, whereas mobile-phone use may depend on more localized relationships among the hands, arms, head, and upper torso. Standing and sleeping exhibit different postural characteristics and may involve broader or more sustained skeletal patterns. A fixed anatomical graph does not explicitly determine which body regions should receive greater emphasis for a particular behavioral sequence. Second, discriminative information is unevenly distributed across time. Transitional frames may contain information that differs substantially from periods in which a posture remains relatively stable. Third, classroom recordings commonly involve restricted motion ranges, partial occlusion, variations in student scale, and complex visual surroundings, increasing the difficulty of distinguishing behaviors using appearance information or fixed skeletal topology alone.

Attention mechanisms provide a means of addressing the unequal distribution of behavioral information across joints and time. Skeleton-based attention models have shown that spatial mechanisms can selectively emphasize discriminative

joints, while temporal mechanisms can assign different importance to frames or temporal segments [12-14]. When combined with adaptive graph learning (AG), attention allows the representation to move beyond uniform treatment of predefined anatomical connections and to emphasize behavior-dependent skeletal relationships and temporally informative motion patterns. Such a formulation is particularly relevant to classroom activity, where discriminative information may be concentrated in a limited number of joints or in relatively short portions of an action sequence.

A further limitation of automated classroom behavior recognition concerns the interpretation of model outputs. High classification accuracy alone does not explain how recognized behaviors can contribute to instructional decision-making. Observable behaviors should not be treated as direct measurements of motivation, comprehension, cognitive engagement, or learning achievement. A more defensible approach is to aggregate model predictions into transparent descriptions of observable classroom activity while maintaining a clear distinction between visual measurement and pedagogical interpretation.

This distinction is particularly important in music education. Music teaching frequently alternates among explanation, listening, demonstration, rhythmic exercises, movement-based activities, individual practice, and group performance. Artificial intelligence and digital technologies are increasingly being investigated in music education for personalized instruction, feedback, interactive learning, and technology-supported teaching and assessment [15, 16]. However, the instructional meaning of an observable classroom behavior depends strongly on the activity being conducted. Standing during a movement-based rhythmic exercise, for example, may be entirely consistent with the intended task, whereas persistent standing during a seated instructional segment may require a different interpretation. Similarly, an increase in observable off-task behavior may provide useful contextual information without revealing why the behavior occurred.

Automated behavioral analysis can therefore provide supplementary observational evidence for music teaching, but it should not replace professional pedagogical judgement. Recognized behavioral patterns may help teachers consider whether an instructional segment should be maintained, shortened, reorganized, or replaced by a more participatory activity (PA). The final interpretation must nevertheless incorporate information unavailable to the visual recognition model, including lesson objectives, musical material, student ability, classroom organization, and instructional stage.

Directly establishing the effectiveness of behavior-informed music teaching requires dedicated music-classroom data and controlled educational evaluation. Public continuous classroom datasets currently provide a more reliable basis for developing and testing the underlying behavior-recognition methodology than for claiming intervention effects in music education. It is therefore necessary to distinguish experimentally evaluated recognition performance from the subsequent instructional interpretation of recognition outputs. Maintaining this distinction avoids attributing educational effects to a computer-vision model without direct empirical evidence.

Against this background, this study develops a spatiotemporal attention graph convolutional network (STAGCN) for automatic recognition of continuous student behaviors in authentic classroom environments and

investigates how the resulting behavioral information can be organized for potential application to music teaching decision support. Student-centered action sequences are represented as spatiotemporal skeleton graphs. An adaptive graph-convolution mechanism is used to model behavior-dependent relationships among body joints, while spatial attention (SA) identifies informative skeletal relationships and temporal attention (TA) models dependencies among different portions of the action sequence. The resulting representation is used to classify four observable classroom behaviors: raising hand, using phone, sleeping, and standing.

Beyond individual behavior classification, the recognition outputs are aggregated into an observable classroom activity profile consisting of PA, MA, and OT. Their temporal variation is subsequently incorporated into a behavior-to-strategy mapping framework that connects persistent observable patterns with candidate music teaching responses. The framework is designed as a decision-support mechanism rather than an automated teaching system, with teacher judgement retained as the final source of instructional decision-making.

The main contributions of this study are summarized as follows. First, a skeleton-based spatiotemporal representation is established for continuous student behavior recognition in authentic classroom scenes, allowing dynamic behavioral information to be modeled beyond single-frame appearance features. Second, an STAGCN architecture combining adaptive graph convolution (AGC) with spatial and TA is developed to capture behavior-dependent joint relationships and temporally discriminative motion patterns. Third, the proposed architecture is examined in relation to established skeleton-based action-recognition approaches through reference, architectural, and mechanism-level analyses, clarifying the roles of adaptive topology learning and spatial-temporal attention in continuous classroom behavior representation. Finally, the four classroom behavior categories are organized into an observable activity profile (OAP) and incorporated into a transparent behavior-to-strategy framework for music teaching. This design maintains a clear separation between computational behavior recognition and prospective instructional application, providing a basis for future validation using dedicated music-classroom data and controlled teaching interventions.

2. RELATED WORK

2.1 Vision-based classroom behavior recognition

Computer vision has increasingly been used to identify observable student behaviors from classroom images and videos. Recent approaches predominantly employ deep convolutional networks and object-detection architectures to localize students and classify their behaviors in complex classroom scenes. Variants of the YOLO family have been applied to behaviors such as reading, writing, raising hands, standing, and mobile-phone use, while attention mechanisms, multi-scale feature fusion, pose information, and relational modeling have been introduced to address differences in student scale, partial occlusion, and crowded seating arrangements [2-5].

The development of classroom-specific datasets has accelerated this line of research. The SCB dataset series was constructed from classroom scenes and provides annotations

for common student behaviors [6]. Subsequent versions expanded the available images, annotations, and behavior categories, providing benchmarks for evaluating detection models under variations in posture, scale, occlusion, and classroom position [6, 7]. Classroom-video studies have further demonstrated the feasibility of recognizing student activities under realistic educational conditions, where multiple students, background interference, and interactions among individuals increase the difficulty of visual analysis [5].

Although image-based detectors can identify visually distinctive behaviors effectively, they generally emphasize information available within individual frames. This limits their ability to represent how body configurations evolve throughout an action. Raising a hand, for example, contains a transition from a resting arm configuration to an elevated posture, while standing involves a broader transition in body configuration. Mobile-phone use and sleeping may contain smaller or more sustained postural changes. Such temporal characteristics cannot always be represented adequately by isolated images.

Video-based recognition can incorporate temporal information, but directly processing RGB sequences also introduces substantial appearance information related to classroom background, furniture, clothing, illumination, and neighboring students [5]. Much of this information is not directly related to the skeletal movement being recognized. Pose-based representations provide an alternative by describing the student primarily through body structure and motion, thereby motivating the use of skeleton-based modeling for continuous classroom behavior recognition.

2.2 Skeleton-based human action recognition

Skeleton-based action recognition represents human motion as sequences of anatomical keypoints rather than complete RGB appearances. Skeletal representations provide a compact description of body configuration and movement and are relatively insensitive to variations in background and clothing. Earlier approaches represented joint coordinates using sequential or convolutional structures, whereas graph neural networks provided a more natural formulation because the human skeleton possesses an explicit relational topology [17].

ST-GCN established an influential graph-based framework by treating body joints as graph nodes, anatomical connections as spatial edges, and corresponding joints across consecutive frames as temporal edges [8]. Graph convolution over this spatiotemporal structure allows body configuration and temporal movement to be learned jointly and has provided the foundation for many subsequent skeleton-based action-recognition architectures.

A limitation of the original formulation is that the graph topology is largely determined by predefined anatomical connectivity. 2s-AGCN addressed this issue by introducing AG, allowing additional joint relationships to be learned from data rather than relying exclusively on the physical skeleton [9]. MS-G3D extended graph reasoning across multiple spatial and temporal scales, enabling information to propagate across different graph distances and temporal ranges [10]. CTR-GCN further introduced channel-wise topology refinement, allowing skeletal relationships to vary according to feature channels and action-dependent information [11].

Attention-based and Transformer-related approaches have further expanded skeleton action recognition by selectively modeling discriminative joints, frames, temporal segments,

and longer-range dependencies [12-14, 17]. These developments indicate that action recognition depends not only on the physical connectivity of body joints but also on identifying which skeletal relationships and temporal patterns are most informative for a particular action.

Classroom behavior recognition nevertheless differs from many conventional skeleton-action benchmarks. Student activities frequently occur in seated or spatially restricted conditions, and some behavior classes contain relatively localized or sustained movements. Raising hand is strongly associated with upper-limb configuration, whereas mobile-phone use may depend on coordinated relationships among the hands, arms, head, and upper torso. Sleeping and standing involve different degrees of postural change. These characteristics motivate a model capable of adapting skeletal relationships while selectively emphasizing informative joints and temporal segments.

2.3 Spatiotemporal attention for behavior recognition

Attention mechanisms have been incorporated into skeleton-based action recognition to selectively emphasize informative spatial and temporal features. Song et al. [12] introduced an end-to-end spatiotemporal attention model in which SA identifies discriminative joints and TA assigns different importance to frames. Subsequent attention-aware skeleton models further explored the selective modeling of spatial and temporal information [13], while segment-based attention approaches investigated dependencies across informative portions of skeletal action sequences [14].

For classroom behavior recognition, SA has a direct functional interpretation because different behaviors depend on different skeletal regions. Upper-limb joints are particularly relevant to raising-hand behavior, while mobile-phone use may involve more localized relationships among the hands, arms, head, and upper torso. Sleeping may be associated with sustained head and upper-body configuration, whereas standing produces a broader postural change. Treating all skeletal joints uniformly may therefore introduce features that contribute little to a particular behavioral decision.

TA addresses a related problem. Given a skeleton sequence $X = \{X_1, X_2, \dots, X_T\}$, the contribution of each frame to behavior recognition need not be identical. A temporal weighting mechanism can be expressed generally as:

$$F_T = \sum_{t=1}^T \alpha_t X_t$$

where, α_t represents the learned importance of frame t and satisfies

$$\sum_{t=1}^T \alpha_t = 1$$

Frames containing characteristic motion transitions can consequently receive different weights from repetitive or relatively stationary portions of a sequence. This principle has been explored in skeleton-based spatiotemporal attention models that learn discriminative temporal information jointly with spatial skeletal features [12-14].

Attention alone, however, does not address all limitations associated with a fixed anatomical graph. The physical

skeleton specifies direct anatomical connections, but action-discriminative relationships may also occur between joints that are not immediately adjacent. Adaptive graph methods therefore learn additional relationships from data, allowing the effective topology to vary with action characteristics [9, 11]. Combining adaptive topology learning with spatial and TA provides a mechanism for jointly modeling anatomical structure, behavior-dependent joint relations, and temporally uneven motion information.

Based on this reasoning, the present study integrates AGC with spatial and TA within a unified STAGCN architecture. AG supplements the physical skeleton topology with sequence-dependent joint relationships, SA models joint-to-joint relevance within individual frames, and TA models dependencies among different time steps. These three components address complementary aspects of continuous classroom behavior recognition.

2.4 Intelligent behavior analysis in music education

Artificial intelligence and digital technologies have increasingly been investigated in music teaching and learning. Recent reviews have identified applications involving personalized learning, automated feedback, interactive instructional environments, technology-supported assessment, and other forms of AI-assisted music education [15, 16]. These developments indicate growing interest in using computational methods to provide additional information for music teaching and learning.

Music education also differs from continuously lecture-oriented instruction because lessons may alternate among explanation, listening, demonstration, rhythmic exercises, movement-based activities, individual practice, and group performance. Student participation is therefore expressed through multiple forms of observable behavior and physical activity. Pose- and skeleton-based action-recognition research demonstrates that structured joint representations can capture meaningful information from human movement [8, 12, 17], suggesting a technical basis for analyzing observable motor behavior in instructional settings.

An important distinction must nevertheless be maintained between recognizing observable behavior and inferring educational outcomes. A visual action does not by itself establish motivation, comprehension, cognitive engagement, or musical achievement. Automated behavior recognition should therefore be used to quantify observable patterns rather than assign unmeasured psychological states to individual students.

This distinction is particularly important when transferring general classroom behavior recognition to music education. The behaviors available in the dataset used in this study—raising hand, using phone, sleeping, and standing—can occur in music classrooms, but they represent only a limited subset of music-related activity. Music-specific behaviors such as score following, instrumental execution, singing, rhythmic synchronization, conducting gestures, and ensemble coordination are not represented by these categories. Consequently, recognition performance on general classroom behavior cannot by itself demonstrate effectiveness in music teaching.

Within these constraints, automatically recognized behaviors can still provide structured observational information for instructional decision support. Persistent changes in observable participation, physical activity, or off-

task behavior can be organized temporally and interpreted in relation to the current lesson stage. The purpose of such analysis is not to automate pedagogical judgement but to provide teachers with additional behavioral information that can be considered together with lesson objectives, musical material, student ability, and classroom context.

2.5 Research gap

The literature reviewed above reveals several limitations that motivate the present study. First, much existing classroom behavior research remains centered on image-based detection or frame-level recognition. Although these approaches are effective for localizing students and identifying visually distinctive behaviors, they do not fully represent the temporal evolution of continuous student actions.

Second, conventional skeleton-based graph networks were largely developed for general human action recognition. Classroom activities present a different recognition environment characterized by restricted motion, crowded scenes, partial occlusion, and behavior classes that may depend on localized or sustained skeletal patterns.

Third, the relative importance of body joints and temporal segments varies among classroom behaviors. Fixed anatomical topology and uniform feature weighting may therefore be insufficient for capturing behavior-dependent joint relationships and temporally discriminative information. AG and spatial-temporal attention provide complementary mechanisms for addressing these limitations.

Fourth, recognition performance and instructional interpretation are often treated as separate problems. A technically accurate behavior classifier does not automatically establish the educational meaning of its predictions. There remains a need for transparent mechanisms that transform observable behavior predictions into structured classroom-level information without treating visual actions as direct measurements of internal learning states.

To address these issues, this study develops an STAGCN that combines AGC with spatial and TA for continuous classroom behavior recognition. The model is formulated using the authentic continuous classroom activity setting provided by CStudentAct [18], while the recognized behavior categories are subsequently organized into an observable classroom activity profile consisting of PA, MA, and OT. This profile is incorporated into a behavior-to-strategy mapping framework for potential application to music teaching. The design deliberately separates computational behavior recognition from prospective instructional application, allowing the technical recognition model and the pedagogical decision-support framework to be assessed at appropriate levels of evidence.

3. DATASET AND DATA PREPROCESSING

3.1 Classroom behavior dataset

The proposed recognition framework is developed in the context of the CStudentAct dataset [18, 19], a continuous student activity dataset collected in authentic classroom environments. Unlike image-based classroom behavior datasets, CStudentAct provides temporally connected action instances, making it suitable for constructing continuous student-centered sequences for skeleton-based action

recognition.

The dataset contains four observable classroom behaviors: raising hand, using phone, sleeping, and standing (Table 1). It comprises 11,357 annotated frames, 142 action instances, and 32,777 bounding-box annotations [18]. Each action instance links the same behavioral event across consecutive frames, providing the temporal information required for pose-sequence construction and subsequent STAGCN modeling.

The original behavior labels are retained without semantic relabeling. Bounding-box trajectories associated with each action instance are used to extract student-centered image sequences, from which two-dimensional skeletal keypoints are subsequently estimated.

Table 1. Statistics of the CStudentAct dataset

Behavior	Annotated Frames	Action Instances	Bounding Boxes
Raising hand	296	39	3,991
Using phone	3,125	26	11,998
Sleeping	3,313	31	10,162
Standing	4,623	46	6,626
Total	11,357	142	32,777

3.2 Action-instance localization and trajectory construction

The CStudentAct annotations provide frame-level bounding boxes and action-instance identifiers for temporally connected student behaviors [18]. Therefore, an additional student detector or multi-object tracker is not required for constructing the behavioral sequences used in this study. Instead, bounding boxes sharing the same action-instance identifier across consecutive frames are directly associated to form a student-centered trajectory.

For action instance i , the annotated bounding box at frame t is represented as

$$B_t^i = (x_t^i, y_t^i, w_t^i, h_t^i)$$

where, x_t^i and y_t^i denote the horizontal and vertical coordinates of the upper-left corner of the bounding box, respectively, while w_t^i and h_t^i denote its width and height.

$$T_i = \{B_1^i, B_2^i, \dots, B_{T_i}^i\}$$

where, T_i denotes the number of annotated frames belonging to action instance i .

Each annotated bounding box is used to define a student-centered region for pose estimation. The resulting regions preserve the temporal correspondence supplied by the original dataset annotations and avoid introducing additional identity-association errors from an independent tracking algorithm.

3.3 Human pose estimation and skeleton representation

Two-dimensional human poses are extracted from each student-centered sequence using RTMPose-m implemented within the MMPose framework [20]. The COCO body-keypoint configuration is adopted, providing 17 anatomical keypoints for each student. Because CStudentAct already provides frame-level bounding-box annotations, these annotated regions are supplied directly to the top-down pose estimator without introducing an additional person-detection

stage.

For frame t , the estimated pose of action instance i is represented as

$$P_t^i = \{p_{t,1}^i, p_{t,2}^i, \dots, p_{t,17}^i\}$$

where, each keypoint j is defined as

$$p_{t,j}^i = (x_{t,j}^i, y_{t,j}^i, c_{t,j}^i), j = 1, \dots, 17$$

and $x_{t,j}^i$ and $y_{t,j}^i$ denote the two-dimensional image coordinates of joint j , while $c_{t,j}^i$ is the corresponding pose-estimation confidence.

To reduce variations caused by student position, camera distance, and bounding-box scale, the joint coordinates are normalized relative to the annotated bounding box. The normalized coordinates are calculated as

$$\tilde{x}_{t,j}^i = \frac{x_{t,j}^i - x_t^i}{w_t^i}$$

and

$$\tilde{y}_{t,j}^i = \frac{y_{t,j}^i - y_t^i}{h_t^i}$$

The normalized joint feature is therefore represented as

$$\tilde{p}_{t,j}^i = (\tilde{x}_{t,j}^i, \tilde{y}_{t,j}^i, c_{t,j}^i)$$

Pose-confidence values are retained as an input channel so that uncertainty in individual keypoints remains available to the recognition model. Pose-estimation failures and missing keypoints are handled during temporal sequence preprocessing according to the implementation procedure described in Section 3.4.

3.4 Sequence construction and data partitioning

The normalized pose estimates are organized into temporally ordered skeleton sequences according to the action-instance identifiers provided by CStudentAct. Each frame contains 17 body joints, and each joint is represented by three channels: normalized horizontal coordinate, normalized vertical coordinate, and pose-estimation confidence.

Before temporal normalization, action instance i is represented as

$$X_i \in \mathbb{R}^{3 \times T_i \times 17}$$

where, T_i denotes the original number of annotated frames belonging to that action instance.

Because action-instance durations vary, all sequences are converted to a common temporal length T before being supplied to STAGCN. The resulting model input is represented as

$$\hat{X}_i \in \mathbb{R}^{3 \times T \times 17}$$

The temporal length T is treated as an implementation parameter determined from the distribution of action-instance durations after complete CStudentAct preprocessing. Action

instances longer than T are temporally sampled across the complete action interval, whereas shorter sequences can be normalized through padding or temporal interpolation. The same temporal-normalization rule should be applied consistently to all action instances before they are supplied to STAGCN.

For model implementation, skeleton-sequence augmentation can be restricted to the training subset. Suitable transformations include small coordinate perturbations, moderate spatial scaling, and temporal sampling, provided that the semantic identity of the original classroom behavior is preserved. Validation and test sequences should remain unaugmented so that quantitative evaluation, when conducted, reflects the original data distribution.

Dataset partitioning is defined at the action-instance level to prevent information leakage between subsets. Frames belonging to the same action instance should not be distributed across different subsets. Where source-video information permits multiple action instances from the same continuous recording context to be identified, these instances should be grouped before partitioning to further reduce contextual leakage.

For quantitative implementation, the processed action instances should be divided into training, validation, and independently held-out test subsets after complete preprocessing. A fixed partition should be retained across STAGCN and any comparative or component-level evaluations so that observed differences are not introduced by inconsistent data assignment.

4. PROPOSED METHOD

4.1 Overall architecture

The proposed STAGCN is designed to capture structural and temporal characteristics of continuous student behaviors from two-dimensional skeleton sequences. Following the preprocessing procedure described in Section 3, each fixed-length action instance is represented

$$\hat{X}_i \in \mathbb{R}^{3 \times T \times 17} \quad (1)$$

where, the three input channels correspond to the normalized horizontal coordinate, normalized vertical coordinate, and pose-estimation confidence; T denotes the fixed sequence length; and 17 corresponds to the body joints extracted using the COCO configuration of RTMPose-m.

For clarity, the batch dimension is omitted from the following formulation. The input skeleton sequence is first represented as a spatiotemporal graph G_i . AGC is then applied to learn both anatomically defined and data-dependent relationships among body joints, producing the graph feature representation F_G . A SA module subsequently emphasizes behavior-relevant joints and joint relationships, yielding F_S . TA is then applied along the frame dimension to obtain F_T . Finally, the graph and attention features are fused into the spatiotemporal representation F_{ST} , which is subsequently used for behavior classification.

The complete feature-processing pipeline is summarized as

$$\hat{X}_i \rightarrow G_i \rightarrow F_G \rightarrow F_S \rightarrow F_T \rightarrow F_{ST} \rightarrow \hat{y}_i \quad (2)$$

where, G_i denotes the spatiotemporal skeleton graph of action

instance i ; F_G is the adaptive graph-convolutional representation; F_S and F_T denote the spatially and temporally refined features, respectively; F_{ST} is the fused spatiotemporal representation; and $\hat{y}_i$ is the predicted probability distribution

over the four classroom behavior categories.

The overall architecture of STAGCN and the complete processing flow from skeleton-sequence input to four-class behavior prediction are illustrated in Figure 1.

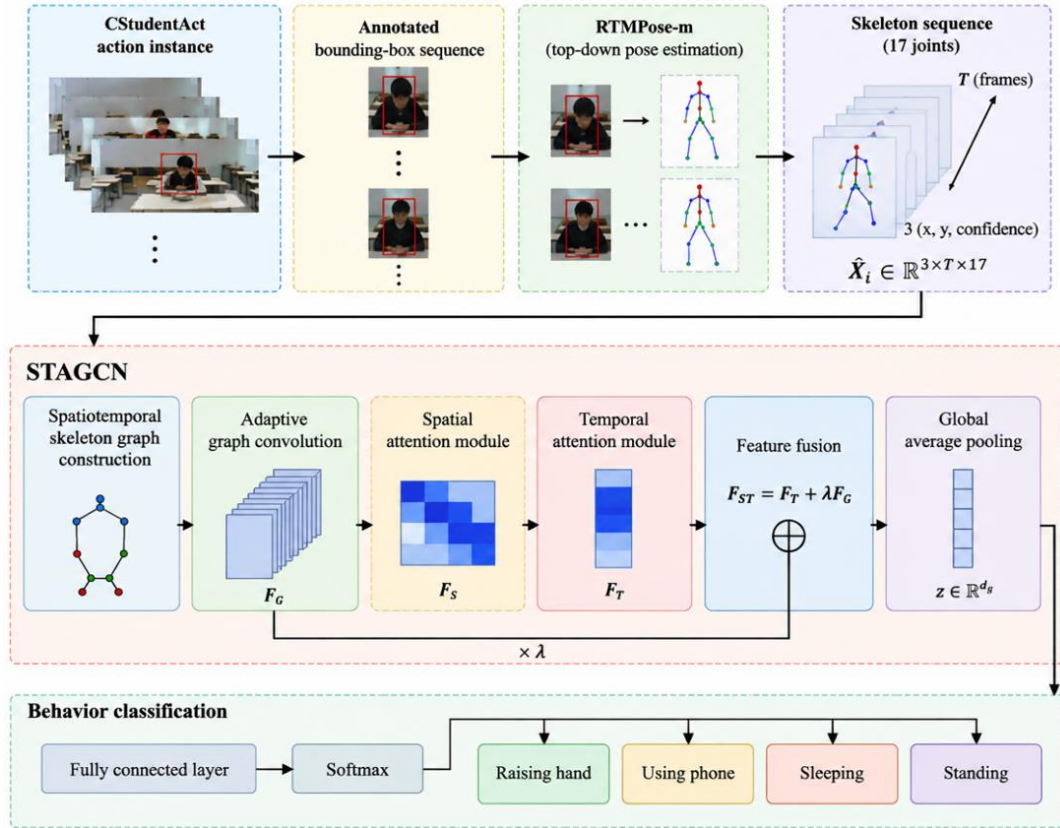


Figure 1. Overall architecture of the proposed STAGCN

4.2 Spatiotemporal skeleton graph construction

Each fixed-length skeleton sequence is represented as a spatiotemporal graph

$$G_i = (\mathcal{V}_i, \mathcal{E}_i) \quad (3)$$

where, $\mathcal{V}_i$ denotes the set of skeletal nodes and $\mathcal{E}_i$ denotes the corresponding edge set. For a sequence containing T frames and $V = 17$ body joints, each node corresponds to a particular joint at a particular time step:

$$\mathcal{V}_i = \{v_{t,j} \mid t = 1, \dots, T; j = 1, \dots, 17\} \quad (4)$$

The feature vector associated with node $v_{t,j}$ is

$$h_{t,j} = [\tilde{x}_{t,j} \quad \tilde{y}_{t,j} \quad c_{t,j}]^T \in \mathbb{R}^3 \quad (5)$$

where, $\tilde{x}_{t,j}$ and $\tilde{y}_{t,j}$ are the bounding-box-normalized joint coordinates defined in Section 3.3, and $c_{t,j}$ is the corresponding RTMPose confidence score.

The complete edge set consists of spatial and temporal connections:

$$\mathcal{E}_i = \mathcal{E}_{S,i} \cup \mathcal{E}_{T,i} \quad (6)$$

Spatial edges connect anatomically related joints within the same frame. Let $\mathcal{B}$ denote the set of anatomical joint pairs

corresponding to the 17-joint COCO body-keypoint configuration adopted in Section 3.3 [20]. The spatial edge set is written as

$$\mathcal{E}_{S,i} = \{(v_{t,j}, v_{t,k}) \mid (j, k) \in \mathcal{B}, t = 1, \dots, T\} \quad (7)$$

Temporal edges connect the same anatomical joint in consecutive frames:

$$\mathcal{E}_{T,i} = \{(v_{t,j}, v_{t+1,j}) \mid t = 1, \dots, T-1, j = 1, \dots, 17\} \quad (8)$$

Following the general spatiotemporal skeleton-graph formulation established in ST-GCN [8], the resulting graph contains both the anatomical structure of the human body and the temporal evolution of each joint. The fixed anatomical topology provides a stable structural prior, while additional behavior-dependent relationships are learned through the adaptive graph component introduced in Section 4.3. The construction of the spatial and temporal connections within the resulting skeleton graph is illustrated in Figure 2.

4.3 AGC

Predefined anatomical topology does not necessarily capture all joint relationships that are discriminative for different actions. Adaptive graph-learning methods have therefore been developed to supplement fixed skeletal connectivity with data-dependent relationships [9, 11]. Following this general principle, the present model introduces

an adaptive graph component to learn additional joint-to-joint

dependencies from the skeleton features.

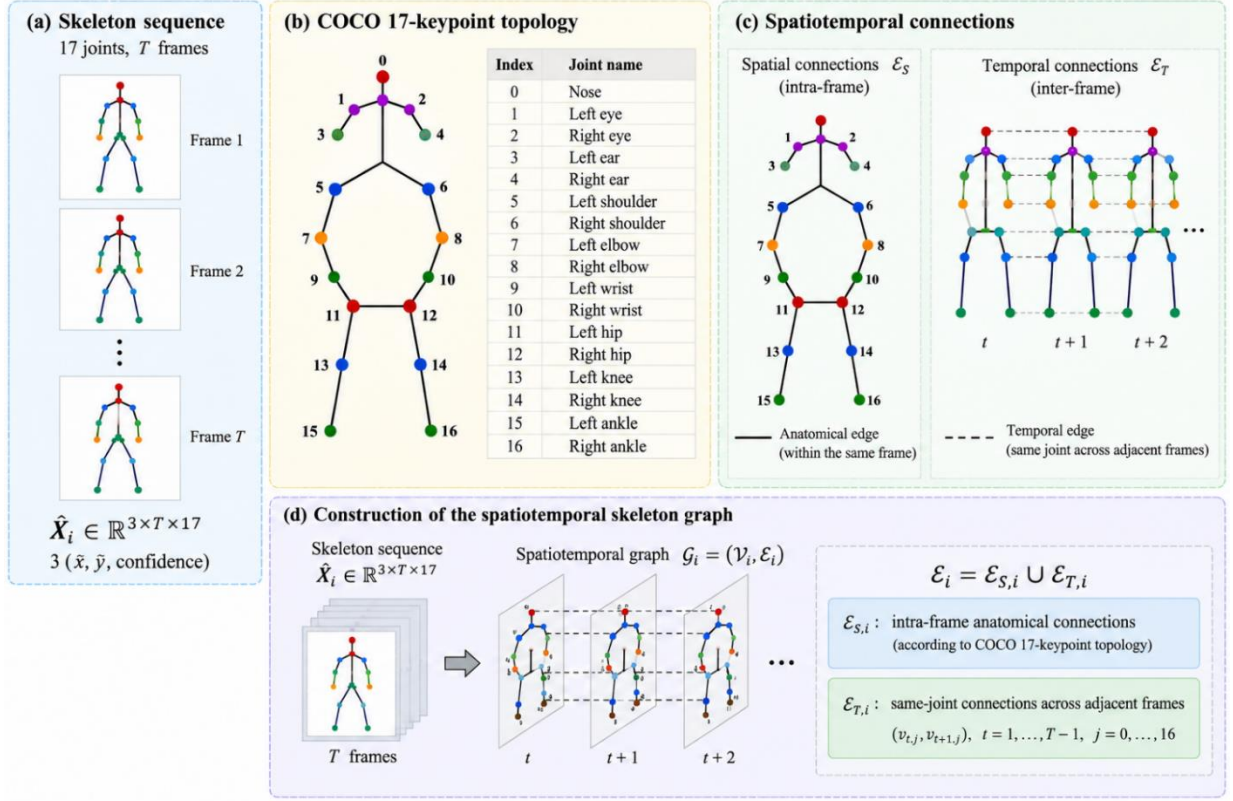


Figure 2. Construction of the spatiotemporal skeleton graph

Let the input feature tensor of graph-convolution layer l be

$$H^{(l)} \in \mathbb{R}^{T \times V \times d_l}, V = 17 \quad (9)$$

where, T is the sequence length, V is the number of body joints, and d_l denotes the feature dimension of layer (l) .

To construct a sequence-level adaptive spatial topology, the features are first aggregated along the temporal dimension:

$$\bar{H}^{(l)} = \frac{1}{T} \sum_{t=1}^T H_t^{(l)} \quad (10)$$

where, $\bar{H}^{(l)} \in \mathbb{R}^{V \times d_l}$ denotes the sequence-level joint representation obtained by temporal averaging.

Query and key representations are then obtained through learnable linear transformations:

$$Q^{(l)} = \bar{H}^{(l)} W_Q^{(l)}, K^{(l)} = \bar{H}^{(l)} W_K^{(l)} \quad (11)$$

where, $W_Q^{(l)}, W_K^{(l)} \in \mathbb{R}^{d_l \times d_a}$ are learnable projection matrices, $Q^{(l)}, K^{(l)} \in \mathbb{R}^{V \times d_a}$, and d_a denotes the adaptive embedding dimension.

The adaptive joint-relation matrix is calculated as

$$A_{\text{adp}}^{(l)} = \text{Softmax}_{\text{row}} \left(\frac{Q^{(l)} K^{(l)T}}{\sqrt{d_a}} \right) \quad (12)$$

where, $A_{\text{adp}}^{(l)} \in \mathbb{R}^{V \times V}$.

The row-wise softmax normalizes the learned outgoing relation weights of each joint.

Let $A_{\text{phy}} \in \mathbb{R}^{V \times V}$ denote the predefined physical adjacency matrix derived from the COCO skeleton topology. After including self-connections through the identity matrix (I), the effective spatial adjacency matrix is defined as

$$A_*^{(l)} = A_{\text{phy}} + I + \gamma_l A_{\text{adp}}^{(l)} \quad (13)$$

where, γ_l is a learnable scalar controlling the contribution of the adaptive topology.

The corresponding degree matrix is

$$D_{*,jj}^{(l)} = \sum_{k=1}^V A_{*,jk}^{(l)} \quad (14)$$

For each frame t , spatial graph convolution is performed as

$$F_{G,t}^{(l)} = \sigma \left[(D_*^{(l)})^{-\frac{1}{2}} A_*^{(l)} (D_*^{(l)})^{-\frac{1}{2}} H_t^{(l)} W_G^{(l)} \right] \quad (15)$$

where, $W_G^{(l)}$ denotes the trainable graph-convolution matrix and $\sigma(\cdot)$ denotes a nonlinear activation function.

Applying Eq. (15) to all (T) frames yields the graph feature tensor

$$F_G^{(l)} \in \mathbb{R}^{T \times V \times d_{l+1}} \quad (16)$$

This formulation preserves the physical human-body topology while allowing the network to learn sequence-dependent relationships between joints that are not necessarily connected anatomically.

4.4 SA module

Skeleton-based attention studies have shown that different joints can contribute unequally to action recognition and that selectively weighting spatial information can improve the representation of discriminative body regions [12-14]. Accordingly, the SA module is introduced after AGC to model joint-to-joint dependencies within each frame and selectively emphasize behavior-relevant skeletal regions.

For frame t , the graph feature matrix is

$$F_{G,t} \in \mathbb{R}^{V \times d_g}, V = 17 \quad (17)$$

where, d_g denotes the graph-feature dimension.

The spatial query, key, and value representations are obtained as

$$Q_{S,t} = F_{G,t} W_Q^S, K_{S,t} = F_{G,t} W_K^S, V_{S,t} = F_{G,t} W_V^S \quad (18)$$

where, $W_Q^S, W_K^S \in \mathbb{R}^{d_g \times d_s}$ and $W_V^S \in \mathbb{R}^{d_g \times d_g}$ are learnable projection matrices, $Q_{S,t}, K_{S,t} \in \mathbb{R}^{V \times d_s}$, $V_{S,t} \in \mathbb{R}^{V \times d_g}$, and d_s denotes the SA embedding dimension.

The frame-specific SA matrix is calculated as

$$A_{S,t} = \text{Softmax}_{\text{row}} \left(\frac{Q_{S,t} K_{S,t}^T}{\sqrt{d_s}} \right) \quad (19)$$

where, $A_{S,t} \in \mathbb{R}^{V \times V}$.

Thus, each element of $(A_{S,t})$ represents the learned relevance between two skeletal joints within the same frame.

The spatially refined feature is obtained through

$$F_{S,t} = A_{S,t} V_{S,t} + F_{G,t} \quad (20)$$

where, the residual connection preserves the original graph representation while allowing the attention mechanism to selectively strengthen informative joint relationships.

Applying Eqs. (18)-(20) to all T frames produces

$$F_S \in \mathbb{R}^{T \times V \times d_g} \quad (21)$$

This frame-wise formulation ensures that SA operates explicitly along the joint dimension. For example, the model may assign greater importance to upper-limb joints when recognizing hand raising, whereas mobile-phone use and sleeping may involve different distributions of attention across the hands, head, shoulders, and upper torso. The frame-wise computation of the SA module is summarized in Figure 3.

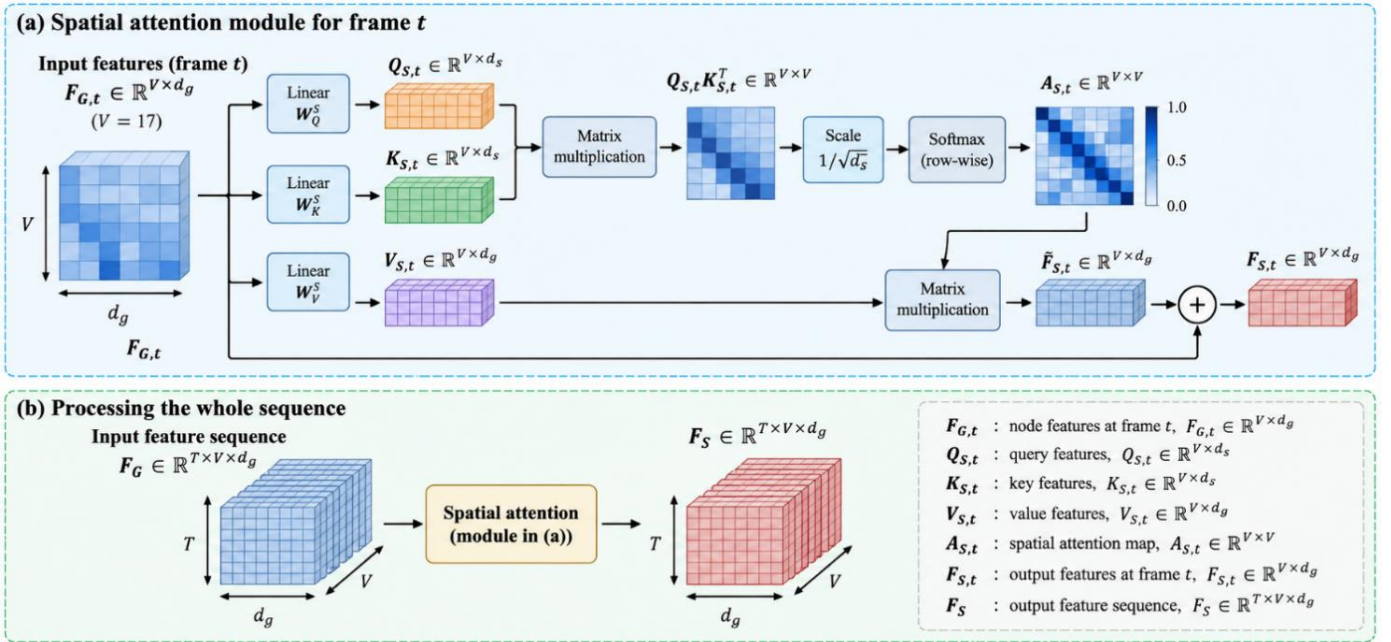


Figure 3. Structure of the SA module

4.5 TA module

TA mechanisms in skeleton-based action recognition are designed to account for the unequal contribution of different frames or temporal segments to an action sequence [12-14]. This consideration is particularly relevant to classroom behaviors, where characteristic transitions may be surrounded by relatively stable postures. The TA module is therefore introduced after spatial refinement to model relationships among different time steps.

Given the spatially refined feature tensor $F_S \in \mathbb{R}^{T \times V \times d_g}$, a frame-level descriptor is first obtained by averaging over the $V = 17$ body joints:

$$u_t = \frac{1}{V} \sum_{j=1}^V F_S(t, j, :), t = 1, \dots, T \quad (22)$$

where, $u_t \in \mathbb{R}^{d_g}$ denotes the joint-averaged descriptor of frame t .

Stacking all frame descriptors gives

$$U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix} \in \mathbb{R}^{T \times d_g} \quad (23)$$

Temporal query and key representations are then obtained

as

$$Q_T = UW_Q^T, K_T = UW_K^T \quad (24)$$

where, $W_Q^T, W_K^T \in \mathbb{R}^{d_g \times d_t}$ are learnable temporal projection matrices, $Q_T, K_T \in \mathbb{R}^{T \times d_t}$, and d_t denotes the TA embedding dimension.

The TA matrix is calculated as

$$A_T = \text{Softmax}_{\text{row}} \left(\frac{Q_T K_T^T}{\sqrt{d_t}} \right) \quad (25)$$

where, $A_T \in \mathbb{R}^{T \times T}$.

Each element $A_T(t, \tau)$ therefore represents the learned relevance of frame τ to frame t .

The TA weights are applied to the complete spatially refined skeleton features rather than only to the pooled frame descriptors. The temporally refined feature at frame (t) is defined as

$$F_T(t, j, :) = \sum_{\tau=1}^T A_T(t, \tau) F_S(\tau, j, :) + F_S(t, j, :) \quad (26)$$

$$t = 1, \dots, T, j = 1, \dots, V$$

The complete temporal feature tensor therefore satisfies

$$F_T \in \mathbb{R}^{T \times V \times d_g} \quad (27)$$

This formulation allows TA to model relationships among frames while preserving joint-specific spatial information. It enables the model to assign different levels of influence to characteristic behavioral transitions and relatively stationary portions of a sequence. For example, the transition from a resting arm position to a raised-hand posture may contain discriminative information that differs from subsequent frames in which the arm remains stationary.

The complete TA process, from joint-averaged frame descriptors to temporally refined skeleton features, is illustrated in Figure 4.

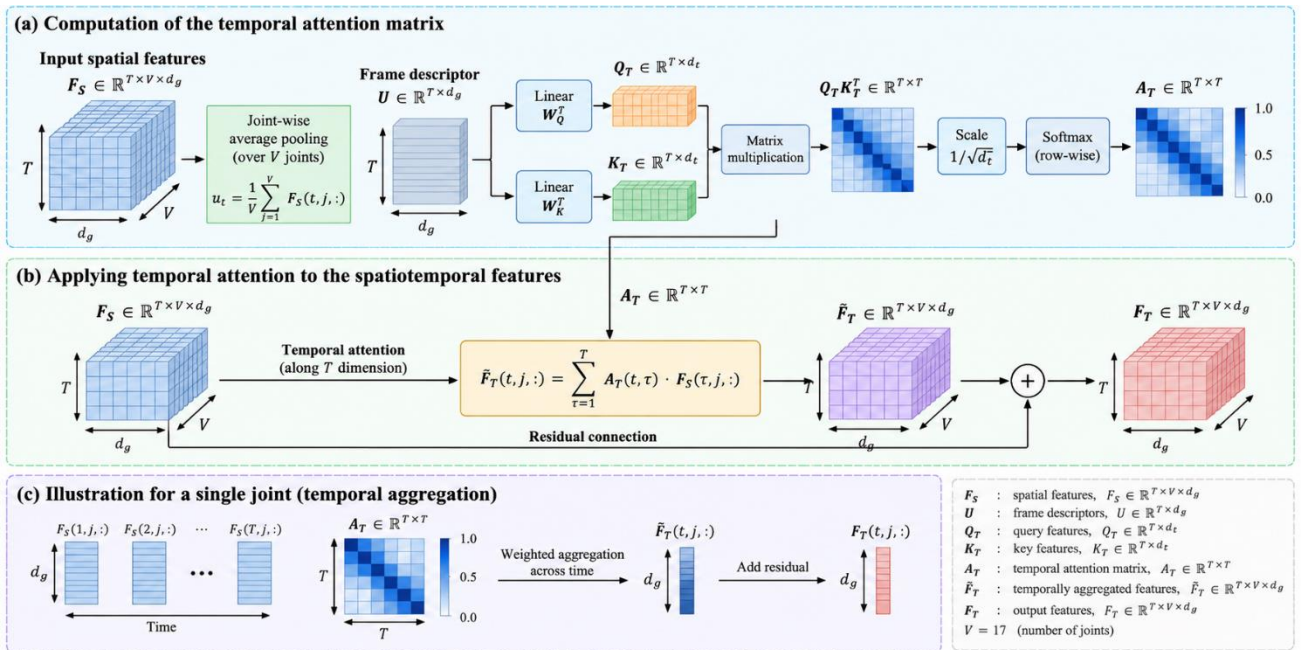


Figure 4. Structure of the TA module

4.6 Feature fusion and behavior classification

The TA module produces the refined feature tensor $F_T \in \mathbb{R}^{T \times V \times d_g}$, which has the same dimensionality as the graph-convolutional representation F_G . To preserve the structural information learned by AGC while incorporating the spatially and temporally refined features, a residual feature-fusion operation is employed:

$$F_{ST} = F_T + \lambda F_G \quad (28)$$

where, $F_{ST} \in \mathbb{R}^{T \times V \times d_g}$ and λ is a learnable scalar controlling the contribution of the original graph representation. Because F_T and F_G have identical dimensions, the fusion can be performed directly without an additional feature projection.

The fused representation is subsequently aggregated over the temporal and joint dimensions using global average pooling:

$$z = \frac{1}{TV} \sum_{t=1}^T \sum_{j=1}^V F_{ST}(t, j, :) \quad (29)$$

where, $z \in \mathbb{R}^{d_g}$.

The pooled feature vector is passed through a fully connected classification layer. The class logits are calculated as

$$o = W_c z + b_c \quad (30)$$

where, $W_c \in \mathbb{R}^{C_b \times d_g}$, $b_c \in \mathbb{R}^{C_b}$, and $C_b = 4$ denotes the number of classroom behavior categories.

$$\hat{y}_{i,c} = \frac{\exp(o_{i,c})}{\sum_{k=1}^{C_b} \exp(o_{i,k})}, c = 1, \dots, C_b \quad (31)$$

The predicted behavior category for action instance i is

therefore

$$\hat{c}_i = \arg \max_{c \in \{1, \dots, C_b\}} \hat{y}_{i,c} \quad (32)$$

The four output categories correspond to raising hand, using phone, sleeping, and standing. Through this procedure, the final prediction incorporates anatomical graph structure, adaptive joint relationships, spatially selective skeletal information, and temporally discriminative motion patterns.

4.7 Objective function

The four behavior categories in CStudentAct contain unequal numbers of action instances. To reduce the influence of class imbalance during model training, a class-weighted cross-entropy loss is used. For a mini-batch containing N_b action instances, the classification loss is defined as

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{c=1}^{C_b} w_c y_{i,c} \log(\hat{y}_{i,c} + \varepsilon) \quad (33)$$

where, $C_b = 4$ is the number of behavior categories, $y_{i,c} \in \{0,1\}$ is the one-hot ground-truth label, $\hat{y}_{i,c}$ is the predicted probability defined in Eq. (31), w_c is the weight assigned to class c , and ε is a small numerical constant used to avoid evaluating the logarithm at zero.

The class weights are calculated exclusively from the training subset. Let N_c denote the number of training action instances belonging to class c , and let

$$N_{\text{tr}} = \sum_{c=1}^{C_b} N_c \quad (34)$$

denote the total number of training instances. The inverse-frequency class weight is defined as

$$w_c = \frac{N_{\text{tr}}}{C_b N_c} \quad (35)$$

This formulation assigns larger weights to less frequent behavior categories while keeping the average contribution of the class weights on a comparable scale. Validation and test samples are not used to calculate w_c , thereby avoiding information leakage from the evaluation subsets.

The complete model is trained end-to-end by minimizing

$$\mathcal{L} = \mathcal{L}_{\text{cls}} \quad (36)$$

so that AG, SA, TA, feature fusion, and behavior classification are jointly optimized.

5. IMPLEMENTATION AND EVALUATION FRAMEWORK

5.1 Implementation details

The implementation specification of STAGCN follows the preprocessing and model formulations described in Sections 3 and 4. The model input consists of temporally normalized two-dimensional skeleton sequences containing 17 COCO body

joints extracted using RTMPose-m [20]. Each joint is represented by three channels: normalized horizontal coordinate, normalized vertical coordinate, and pose-estimation confidence.

STAGCN is formulated for implementation in PyTorch with GPU acceleration. The reference training configuration uses AdamW optimization with an initial learning rate of 1×10^{-3} , a batch size of 16, a weight decay of 5×10^{-4} , cosine-annealing learning-rate scheduling, and a maximum of 100 training epochs. These settings provide a reproducible implementation specification for the proposed architecture and its component configurations.

Dataset partitioning is defined at the action-instance level so that frames belonging to the same temporally connected behavior instance are not distributed across different subsets. Hyperparameter selection should be performed using only the training and validation subsets, while an independently held-out test subset should be reserved for quantitative evaluation. This separation prevents test information from influencing model configuration or parameter selection.

The principal implementation settings are summarized in Table 2. Parameters that depend on complete preprocessing of the original CStudentAct action-instance annotations, particularly the final temporal sequence length and exact subset composition, should be determined from the processed dataset rather than inferred from aggregate dataset statistics [18].

Table 2. Reference implementation configuration of the proposed STAGCN

Parameter	Setting
Deep learning framework	PyTorch
Dataset	CStudentAct
Input representation	2D skeleton sequence
Pose estimation model	RTMPose-m (COCO)
Number of joints	17
Input channels	$3(\hat{x}, \hat{y}, c)$
Number of behavior classes	4
Dataset partition	Action-instance level
Sequence length T	Determined from processed action-instance durations
Batch size	16
Optimizer	AdamW
Initial learning rate	1×10^{-3}
Weight decay	5×10^{-4}
Learning-rate schedule	Cosine annealing
Maximum epochs	100

5.2 Reference recognition methods

Four representative skeleton-based action-recognition architectures are selected as methodological references for positioning the proposed STAGCN. These methods cover fixed skeletal topology, AG, multi-scale spatiotemporal modeling, and topology refinement.

ST-GCN provides the fundamental graph-convolution reference. It represents body joints as graph nodes, anatomical relationships as spatial edges, and corresponding joints across consecutive frames as temporal connections, establishing the basic spatiotemporal skeleton-graph formulation [8].

2s-AGCN extends fixed skeletal topology through AG and combines joint and bone information, allowing action-dependent relationships among body joints to be learned from data [9]. Its adaptive topology mechanism provides an important methodological reference for the adaptive graph

component of STAGCN.

MS-G3D models spatial and temporal dependencies across multiple graph scales, allowing information to propagate over different spatial distances and temporal ranges [10]. It therefore provides a reference for multi-scale spatiotemporal skeleton modeling.

CTR-GCN introduces channel-wise topology refinement, allowing skeletal relationships to vary according to feature channels and action-dependent information [11]. It represents a stronger adaptive graph formulation against which the topology-learning characteristics of STAGCN can be conceptually compared.

In addition, the published CStudentAct recognition framework provides a classroom-specific reference because it addresses continuous student activity recognition directly from classroom video [18, 19]. Unlike STAGCN, which represents annotated action instances as temporally ordered skeleton sequences, the published framework uses appearance-based detection and temporal association. The two approaches therefore address related classroom-recognition problems through different representations and modeling assumptions.

These architectures were originally developed under different datasets and experimental protocols. They are therefore used in the present study as methodological references rather than as sources of directly comparable numerical performance. Their architectural characteristics are compared with STAGCN in Section 6.1.

5.3 Quantitative evaluation criteria

For quantitative evaluation of the proposed recognition architecture, accuracy, precision, recall, and F1-score provide the principal classification criteria. Because the four CStudentAct behavior categories contain unequal numbers of action instances [18], macro-averaged metrics are particularly appropriate in addition to overall accuracy, as they assign equal importance to each behavior category.

For behavior category c , precision is defined as

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}$$

where, TP_c and FP_c denote the numbers of true-positive and false-positive predictions for category c , respectively.

Recall is defined as

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$$

where, FN_c denotes the number of false-negative predictions.

The class-wise F1-score is defined as

$$F1_c = \frac{2 \text{ Precision}_c \text{ Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

For the four-class classification task, overall accuracy is defined as

$$\text{Accuracy} = \frac{\sum_{c=1}^{C_b} TP_c}{N_{\text{test}}}, C_b = 4$$

where, N_{test} denotes the total number of evaluated action

instances.

Macro-F1 assigns equal importance to the four behavior categories and is defined as

$$\text{Macro-F1} = \frac{1}{C_b} \sum_{c=1}^{C_b} F1_c$$

Macro-precision and macro-recall are defined analogously. These metrics provide a standardized quantitative evaluation protocol for future controlled implementation of STAGCN and facilitate comparison with skeleton-based action-recognition methods evaluated under the same dataset partition and experimental conditions [8-11, 17].

5.4 Architectural decomposition

To clarify the functional organization of STAGCN, the architecture is decomposed progressively from a physical skeleton-graph baseline. AG, SA, and TA are introduced sequentially to distinguish the representational role of each component. The four configurations are summarized in Table 3.

Table 3. Architectural configurations of STAGCN

Configuration	AG	SA	TA
Baseline	—	—	—
Baseline + AG	✓	—	—
Baseline + AG + SA	✓	✓	—
Full STAGCN	✓	✓	✓

The baseline retains the predefined physical skeleton topology following the general ST-GCN formulation [8]. Baseline + AG supplements this topology with data-dependent joint relationships, following the motivation of adaptive graph-learning approaches [9, 11]. Baseline + AG + SA additionally introduces frame-wise SA to model unequal contributions among skeletal joints [12-14]. The full STAGCN further incorporates TA to model relationships among different portions of the action sequence [12-14, 17].

This decomposition provides the basis for the architectural component analysis in Section 6.3. It identifies how the representational capability of the model changes as AG, SA, and TA are introduced, without treating architectural inclusion alone as evidence of quantitative performance improvement.

6. MODEL ANALYSIS AND REFERENCE EVALUATION

6.1 Reference evaluation and methodological comparison

The proposed STAGCN is examined in relation to both the published CStudentAct recognition framework and representative skeleton-based action-recognition architectures [8-11]. Because these methods were developed under different input representations, task formulations, and evaluation protocols, numerical results reported in the original studies are not treated as directly comparable performance scores. Instead, the comparison focuses on the methodological capabilities that are relevant to continuous classroom behavior recognition.

CStudentAct was developed specifically for continuous student activity recognition and contains four annotated

classroom behaviors—raising hand, using phone, sleeping, and standing—with 11,357 annotated frames, 142 action instances, and 32,777 bounding boxes [18, 19]. The published STrack4Re framework formulates continuous student activity recognition as a detection-and-tracking problem [19]. YOLOv5 is used to identify activity regions, while OC-SORT associates detected activity instances across successive frames [19]. This formulation directly addresses the temporal localization of activities in continuous classroom video.

The present STAGCN follows a different modeling strategy. Rather than performing activity detection and tracking directly from RGB appearance, annotated action

instances are converted into temporally ordered 17-joint skeleton sequences using RTMPose-m [20]. AGC models both physical and data-dependent relationships among body joints, while spatial and TA separately model behavior-relevant skeletal relationships and temporal dependencies.

Table 4 summarizes the methodological relationship among the published CStudentAct framework, representative skeleton-based action-recognition architectures, and the proposed STAGCN. The comparison is intended to clarify differences in representation and modeling capability rather than to imply numerical superiority across incompatible experimental protocols.

Table 4. Methodological comparison of reference approaches relevant to continuous classroom behavior recognition

Method	Primary Representation	Adaptive Topology	SA	Temporal Modeling	Continuous Classroom Orientation
STrack4Re [19]	RGB activity regions	No	No	OC-SORT tracking	Yes
ST-GCN [8]	Skeleton	No	No	Spatiotemporal graph convolution	No
2s-AGCN [9]	Joint and bone skeleton streams	Yes	No	AGC	No
MS-G3D [10]	Skeleton	Multi-scale graph modeling	No	Multi-scale spatiotemporal graph convolution	No
CTR-GCN [11]	Skeleton	Yes	Channel-wise topology refinement	Graph-based temporal modeling	No
STAGCN (proposed)	2D skeleton sequence	Yes	Yes	Explicit TA	Adapted to CStudentAct

The comparison highlights two complementary directions in continuous classroom behavior analysis. STrack4Re preserves RGB appearance and explicitly solves the localization and temporal association problem, whereas skeleton-based methods abstract the observed student into a structured representation of body joints and motion. STAGCN adopts the latter representation while introducing adaptive joint relations and separate spatial and TA mechanisms. Consequently, its contribution lies in the architecture used to represent continuous skeletal behavior rather than in claiming direct numerical superiority over results obtained under different published protocols.

6.2 Class-level behavioral characteristics and recognition considerations

The four CStudentAct categories differ substantially in both their sample distributions and their underlying skeletal characteristics. As summarized in Table 5, standing contains the largest number of annotated frames and action instances, whereas raising hand contains relatively few frames despite accounting for 39 action instances. Using phone and sleeping contain fewer action instances but substantially more annotated frames. These differences indicate that class frequency should be considered at the action-instance level rather than inferred only from the number of annotated frames [18].

From a skeletal perspective, raising hand is characterized primarily by changes in upper-limb configuration. The shoulder, elbow, and wrist joints provide a direct representation of the transition from a resting arm position to an elevated posture. This behavior is therefore expected to contain comparatively localized but temporally distinctive skeletal motion. Standing involves a broader postural transition and can affect the relative positions of the upper

body, hips, knees, and ankles. The resulting change in whole-body configuration provides a different type of structural cue from the predominantly upper-limb motion associated with raising hand. Such differences in the spatial and temporal contributions of body joints are consistent with the motivation underlying skeleton-based graph convolution and attention mechanisms [8, 12-14].

Using phone and sleeping present a different recognition setting. Mobile-phone use may involve comparatively small changes in the hands, wrists, elbows, head, and upper torso, while the visual object itself is not represented in the skeleton input. Consequently, the skeleton representation captures posture associated with phone use but not the appearance of the mobile device. Sleeping may similarly be characterized by relatively sustained head and upper-body configurations rather than large body movements. These characteristics provide a rationale for jointly modeling spatial skeletal relations and temporal information rather than relying only on instantaneous pose [12-14, 17].

The class-level characteristics also motivate the use of class-weighted cross-entropy defined in Section 4.7. Because the number of action instances varies from 26 for using phone to 46 for standing [18], weighting is calculated from the training subset rather than from frame counts. This prevents long-duration action instances from being treated as independent samples merely because they contain more annotated frames.

These observations describe structural characteristics of the four recognition categories rather than measured class-wise recognition performance. Quantitative claims concerning the relative difficulty of individual behaviors would require predictions from a held-out test subset and are therefore not inferred from class frequency or qualitative skeletal characteristics alone.

Table 5. Class-level composition and skeletal characteristics of the CStudentAct behaviors

Behavior	Annotated Frames	Action Instances	Main Skeletal Characteristics	Temporal Characteristic
Raising hand	296	39	Predominantly shoulder-elbow-wrist configuration	Distinct upper-limb transition
Using phone	3,125	26	Hands, wrists, elbows, head, and upper torso	Localized or sustained posture
Sleeping	3,313	31	Head and upper-body configuration	Predominantly sustained posture
Standing	4,623	46	Broad upper- and lower-body configuration	Whole-body postural transition

The class-level dataset statistics reported in Table 5 are derived from the published CStudentAct annotations [18]. Table 5 also illustrates why frame count alone is not an appropriate measure of class frequency for the present recognition task. Raising hand contains only 296 annotated frames but 39 action instances, whereas using phone contains 3,125 frames but only 26 action instances. Because STAGCN classifies temporally connected action instances rather than independent frames, the training loss and dataset partitioning are defined at the action-instance level.

6.3 Architectural component analysis

The proposed STAGCN is constructed progressively from a physical skeleton graph by introducing AG, SA, and TA. These components address different limitations of fixed-

topology skeleton modeling and therefore play complementary roles within the complete architecture. Table 6 summarizes the functional contribution of each configuration. The analysis is based on the mathematical formulation developed in Section 4 and on established findings from skeleton-based graph convolution and attention research [8-14, 17].

The baseline configuration retains only the predefined physical skeleton topology. This formulation follows the general principle established by ST-GCN, in which anatomically connected body joints define the spatial structure and corresponding joints across successive frames define temporal connections [8]. Such a representation preserves the natural topology of the human body but does not explicitly account for action-dependent relationships between anatomically nonadjacent joints.

Table 6. Architectural component analysis of the proposed STAGCN

Configuration	AG	SA	TA	Representation Capability	Principal Role
Baseline	—	—	—	Fixed physical skeleton topology	Models predefined anatomical and temporal connections
Baseline + AG	✓	—	—	Physical + adaptive joint topology	Introduces action-dependent relationships among joints
Baseline + AG + SA	✓	✓	—	Adaptive topology + frame-wise spatial refinement	Emphasizes behavior-relevant joint relationships
Full STAGCN	✓	✓	✓	Adaptive spatial topology + spatial and temporal refinement	Jointly models behavior-dependent skeletal and temporal dependencies

Introducing AG extends the physical topology with data-dependent joint relationships. Adaptive graph approaches such as 2s-AGCN and CTR-GCN have demonstrated the importance of allowing skeletal connectivity to vary according to action features rather than relying exclusively on predefined anatomical edges [9, 11]. In STAGCN, this capability is represented by the adaptive relation matrix $A_{\text{adp}}^{(l)}$, which is combined with the physical adjacency matrix through Eq. (13). The resulting topology therefore preserves the anatomical prior while permitting additional sequence-dependent relationships among joints.

SA is subsequently introduced to refine the graph representation at the frame level. Previous skeleton-based attention studies have shown that different joints can contribute unequally to action recognition and that selective spatial weighting can emphasize action-relevant body regions [12-14]. The SA matrix $A_{s,t}$ defined in Eq. (19) models joint-to-joint relevance within each frame. This mechanism is particularly relevant to the CStudentAct categories because raising hand, using phone, sleeping, and standing involve different distributions of skeletal information across the upper and lower body [12-14, 18].

TA provides the final refinement stage by modeling relationships among different portions of the skeleton sequence. Temporal weighting has been widely used in

skeleton-based action recognition to distinguish informative transitions from repetitive or relatively stationary frames [12-14, 17]. In the proposed architecture, the TA matrix A_T defined in Eq. (25) is estimated from joint-averaged frame descriptors and subsequently applied to the complete spatially refined skeleton representation. This design allows temporal dependencies to be modeled without discarding joint-specific spatial information.

The full STAGCN therefore combines three complementary mechanisms: AG modifies the effective skeletal topology, SA redistributes representation capacity across joints within individual frames, and TA redistributes information across the temporal dimension. These components operate at different levels of the representation rather than performing interchangeable functions. The progressive configurations in Table 6 provide a structured architectural decomposition of the proposed model. Quantitative attribution of recognition improvements to individual components would additionally require controlled ablation experiments under an identical training and evaluation protocol and is therefore not inferred from the architectural analysis alone.

The decomposition in Table 6 is consistent with the evolution of skeleton-based action-recognition architectures reported in the literature. ST-GCN established fixed spatiotemporal skeletal graph modeling [8], while subsequent approaches introduced adaptive topology learning [9, 11],

multi-scale spatiotemporal reasoning [10], and selective spatial and TA [12-14]. Recent reviews further indicate that adaptive graph structures and attention-based temporal modeling have become important directions for representing action-dependent skeletal relationships and long-range temporal information [17]. STAGCN integrates these principles within a single architecture tailored to the four continuous classroom behaviors represented in CStudentAct [18].

6.4 Potential inter-class ambiguity analysis

The four CStudentAct behavior categories exhibit different degrees of similarity in their skeletal representations. Because STAGCN operates on two-dimensional joint coordinates and pose-estimation confidence rather than complete RGB appearance, potential recognition ambiguity is determined primarily by similarities in body configuration, motion pattern, and temporal evolution. This differs from RGB-based classroom behavior recognition, in which clothing, surrounding objects, and local appearance can also contribute directly to classification [2-5].

Raising hand is characterized by a relatively distinctive change in upper-limb configuration. The transition from a resting arm position to an elevated posture changes the relative arrangement of the shoulder, elbow, and wrist joints and therefore provides both spatial and temporal skeletal information. Skeleton-based graph convolution is well suited to representing coordinated changes among anatomically connected joints [8-11], while spatial and TA can further distinguish informative joints and temporal segments [12-14]. Consequently, the principal representation challenge for raising hand is expected to arise when the arm is only partially visible, the elevation is small, or pose estimation fails to recover reliable upper-limb keypoints.

Standing also produces a comparatively broad skeletal change because the transition between seated and standing postures affects both upper- and lower-body configuration. However, standing is inherently context dependent. A student may remain upright for an extended interval after the initial transition, causing later frames to contain substantially less motion than the transition itself. Temporal modeling is therefore important for preserving information about how the standing posture develops across the action sequence rather than representing the behavior only through isolated poses [8, 12-14, 17].

Greater representational ambiguity may occur between using phone and sleeping because both behaviors can contain relatively limited whole-body movement. Mobile-phone use may be expressed primarily through the configuration of the hands, wrists, elbows, head, and upper torso, whereas sleeping may involve sustained head lowering or changes in upper-body posture. In addition, the mobile device itself is absent from the skeleton representation. Skeleton-based input therefore cannot directly exploit object appearance, even though such information may be available to RGB-based recognition methods [2-5]. Distinguishing these behaviors consequently depends on subtle differences in joint configuration and their temporal persistence.

Pose-estimation uncertainty introduces an additional source of potential ambiguity. RTMPose provides structured two-dimensional keypoints, but classroom scenes can contain partial occlusion, crowded seating, restricted visibility, and variations in camera perspective [20]. Errors in the estimated

positions or confidence values of the hands, wrists, head, or lower-body joints may reduce the distinctiveness of the corresponding skeletal patterns. Retaining pose-estimation confidence as the third input channel allows STAGCN to preserve information about keypoint uncertainty rather than treating all estimated coordinates as equally reliable.

The adaptive graph, SA, and TA components address different aspects of these potential ambiguities. Adaptive topology permits behavior-dependent relationships among joints beyond fixed anatomical connections [9, 11]; SA allows different skeletal regions to receive different levels of representation according to the observed sequence [12-14]; and TA models the unequal contribution of different frames and transitions [12-14, 17]. These mechanisms provide a structural basis for distinguishing classroom behaviors with heterogeneous motion characteristics, although measured confusion rates between individual categories would require predictions obtained from an independently evaluated test subset.

6.5 SA mechanism analysis

The SA module is designed to represent the unequal contribution of skeletal relationships within individual frames. Previous skeleton-based attention studies have demonstrated that action recognition can be improved by selectively emphasizing informative joints or spatial regions rather than treating all skeletal features uniformly [12-14]. In STAGCN, this principle is implemented through the frame-specific attention matrix $A_{s,t} \in \mathbb{R}^{V \times V}$, defined in Eq. (19), where $V = 17$ corresponds to the COCO body joints used in the present skeleton representation [20].

Because $A_{s,t}$ is normalized row-wise, each row represents the distribution of attention from one query joint to the complete set of skeletal joints. The attention received by joint j at frame t can therefore be summarized as

$$a_{t,j} = \frac{1}{V} \sum_{k=1}^V A_{s,t}(k,j), V = 17$$

A sequence-level joint-importance descriptor can subsequently be obtained by averaging over the T frames:

$$\bar{a}_j = \frac{1}{T} \sum_{t=1}^T a_{t,j}, j = 1, \dots, V$$

Since each row of $A_{s,t}$ is normalized by the softmax operation, the resulting sequence-level scores satisfy

$$\sum_{j=1}^V \bar{a}_j = 1$$

The vector $\bar{\mathbf{a}} = (\bar{a}_1, \bar{a}_2, \dots, \bar{a}_V)$, therefore provides a normalized summary of how SA is distributed across the 17 body joints for an action sequence. This formulation is consistent with previous attention-based skeleton-recognition approaches in which discriminative skeletal regions are weighted according to their contribution to action representation [12-14].

The expected functional role of SA differs among the four CStudentAct categories. Raising hand contains prominent

upper-limb configuration changes, making relationships among the shoulder, elbow, and wrist joints particularly relevant to its skeletal representation. Using phone may depend more strongly on localized configurations involving the hands, wrists, elbows, head, and upper torso, while sleeping may be associated with sustained relationships between the head, shoulders, and upper-body joints. Standing produces a broader postural configuration involving both upper- and lower-body joints. These biomechanical differences provide the motivation for learning behavior-dependent spatial relationships rather than assigning identical importance to all joints [8, 12-14, 18].

The spatial-attention formulation should nevertheless be distinguished from an empirical claim about which joints the trained model actually emphasizes. Without learned attention matrices obtained from a trained STAGCN, specific joint-importance distributions cannot be reported as observed results. The analysis above therefore characterizes the mathematical behavior and intended representational role of the SA mechanism rather than presenting unobserved attention patterns as experimental evidence.

6.6 TA mechanism analysis

The TA module is designed to represent the unequal contribution of different frames and temporal segments within an action sequence. Previous skeleton-based attention studies have shown that discriminative information is not necessarily distributed uniformly over time and that temporal weighting can emphasize informative transitions or action segments [12-14, 17]. In STAGCN, temporal dependencies are represented by the attention matrix $A_T \in \mathbb{R}^{T \times T}$ defined in Eq. (25), where each element $A_T(t, \tau)$ represents the learned relevance of frame (τ) to query frame t .

Because A_T is normalized row-wise, the attention received by frame τ across all query frames can be summarized as

$$\alpha_\tau = \frac{1}{T} \sum_{t=1}^T A_T(t, \tau), \tau = 1, \dots, T$$

The resulting sequence-level temporal-attention profile is represented as

$$\alpha = (\alpha_1, \alpha_2, \dots, \alpha_T)$$

Since every row of A_T is normalized by the softmax operation,

$$\sum_{\tau=1}^T A_T(t, \tau) = 1$$

Thus, α provides a normalized description of how TA is distributed across an action sequence. Unlike simple temporal averaging, this formulation permits different portions of the sequence to contribute unequally to the temporally refined representation. TA has been shown to provide useful mechanisms for representing discriminative temporal information in skeleton-based action recognition [12-14, 17].

The functional relevance of this mechanism differs among the four CStudentAct behaviors [18]. Raising hand contains a characteristic transition from a resting upper-limb configuration to an elevated arm posture, while standing may

contain a broader transition from a seated to an upright body configuration. In such sequences, transitional portions may contain skeletal information that differs substantially from subsequent relatively stable postures. By contrast, using phone and sleeping may contain smaller or more sustained postural changes, making relationships among temporally separated but structurally similar frames potentially relevant to their representation.

An important feature of the proposed temporal module is that A_T is estimated from joint-averaged frame descriptors but is subsequently applied to the complete spatially refined skeleton tensor, as defined in Eq. (26). Consequently, temporal relationships are estimated at the frame level without discarding the joint-specific information contained in F_S . This distinguishes the TA operation from a simple weighted pooling procedure and preserves the $V \times d_g$ skeletal representation throughout temporal refinement.

As with the spatial-attention analysis, the mathematical formulation should be distinguished from an empirical claim about the attention profile learned by a trained model. Specific high-attention frames or behavior-dependent temporal distributions cannot be reported without learned A_T matrices obtained from an evaluated STAGCN. The present analysis therefore characterizes the mathematical properties and intended representational role of TA rather than treating unobserved attention distributions as experimental findings.

6.7 Classroom-level temporal aggregation

Although STAGCN operates at the level of individual action instances, its categorical outputs can be aggregated over predefined temporal intervals to construct a classroom-level description of observable activity. This aggregation provides the interface between skeleton-based behavior recognition and the decision-support framework introduced in Section 7. The four behavior categories are retained directly from CStudentAct—raising hand (RH), using phone (UP), sleeping (SL), and standing (ST)—without introducing additional semantic behavior labels [18].

Let $n_{c,t}$ denote the number of recognized action instances assigned to behavior category c within temporal interval t . The proportion of category c in that interval is defined as

$$r_{c,t} = \frac{n_{c,t}}{\sum_{k=1}^{C_b} n_{k,t}}, C_b = 4$$

where, $c \in \{\text{RH}, \text{UP}, \text{SL}, \text{ST}\}$.

For any temporal interval containing at least one recognized action instance, the four proportions satisfy

$$\sum_{c=1}^{C_b} r_{c,t} = 1$$

The vector $\mathbf{r}_t = (r_{\text{RH},t}, r_{\text{UP},t}, r_{\text{SL},t}, r_{\text{ST},t})$ therefore provides a normalized description of the distribution of recognized classroom behaviors within interval t . Retaining the individual category proportions at this stage preserves the direct correspondence between the recognition outputs and the original CStudentAct labels [18].

The interval length is a decision-support parameter rather than a property of the STAGCN classifier itself. Shorter intervals provide finer temporal resolution but may contain

relatively few recognized action instances, whereas longer intervals provide more stable aggregate proportions at the cost of reduced temporal resolution. The appropriate interval duration should therefore be selected according to the observation period and instructional context when the framework is applied to continuous classroom recordings.

This aggregation should not be interpreted as a measurement of latent educational states. In particular, $r_{c,t}$ describes only the relative occurrence of recognized observable behaviors within an interval and does not directly quantify motivation, comprehension, cognitive engagement, or academic achievement. This distinction is consistent with the broader requirement that AI-assisted educational analysis should preserve a clear boundary between observable computational outputs and subsequent pedagogical interpretation [1, 15, 16].

The normalized behavior proportions provide the direct inputs to the observable classroom activity model developed in Section 7. Raising-hand and standing proportions are retained as indicators of participatory and MA, respectively, while using-phone and sleeping proportions are combined operationally within the OT dimension. This transformation is defined explicitly in Section 7.1 and is used only as a transparent organizational layer for decision support rather than as a validated psychological or educational measurement model.

7. BEHAVIOR-DRIVEN DECISION SUPPORT FOR MUSIC TEACHING

7.1 Observable classroom activity modeling

The output of STAGCN consists of temporally organized predictions for four directly observable classroom behaviors: raising hand, using phone, sleeping, and standing [18]. Rather than converting these behaviors into psychological or cognitive states that are not directly supported by the dataset, the recognized classes are organized into a compact observable classroom activity representation.

Three behavioral dimensions are defined: PA, MA, and OT. Raising hand is used as an observable indicator of overt classroom participation, while standing represents a change in physical activity state. Using a mobile phone and sleeping are grouped as observable off-task behaviors under the labeling scheme used in the present analysis.

Using the behavior proportions defined in Section 6.7, the three components are calculated as

$$\begin{aligned} PA_t &= r_{RH,t} \\ MA_t &= r_{ST,t} \\ OT_t &= r_{UP,t} + r_{SL,t} \end{aligned}$$

where, $r_{RH,t}$, $r_{ST,t}$, $r_{UP,t}$, and $r_{SL,t}$ denote the proportions of raising-hand, standing, using-phone, and sleeping behaviors within temporal interval t , respectively.

This representation maintains a direct correspondence between the STAGCN outputs and the original CStudentAct behavior categories. It therefore provides a time-varying description of observable classroom activity without treating visual behavior as direct evidence of motivation, comprehension, cognitive engagement, or learning achievement.

7.2 Observable classroom activity profile

Rather than combining heterogeneous classroom behaviors into a single engagement score, the three observable behavioral dimensions are retained as a multidimensional classroom activity profile. For temporal interval (t), the OAP is defined as $OAP_t = [PA_t, MA_t, OT_t]$.

This representation avoids assigning arbitrary numerical weights to behaviors that do not necessarily lie on a common educational scale. In particular, raising hand, standing, using phone, and sleeping cannot be assumed to represent different numerical levels of a single latent construct without independent educational validation.

For an analyzed instructional period containing L temporal intervals, the mean values of the three dimensions are calculated as

$$\begin{aligned} \overline{PA} &= \frac{1}{L} \sum_{t=1}^L PA_t \\ \overline{MA} &= \frac{1}{L} \sum_{t=1}^L MA_t \\ \overline{OT} &= \frac{1}{L} \sum_{t=1}^L OT_t \end{aligned}$$

The temporal profiles are represented as

$$\begin{aligned} &\{PA_1, PA_2, \dots, PA_L\} \\ &\{MA_1, MA_2, \dots, MA_L\} \\ &\{OT_1, OT_2, \dots, OT_L\} \end{aligned}$$

Compared with a single lesson-level average, these temporal profiles preserve information about when observable participation, physical activity, and off-task behavior change during the analyzed period. They therefore provide the behavioral input to the music teaching strategy mapping described in Section 7.3.

7.3 Behavior-to-strategy mapping for music teaching

Music lessons frequently involve listening, demonstration, rhythmic activity, movement, active music making, individual practice, and group performance, with bodily participation forming an important component of many music-learning activities [21, 22]. Consequently, the instructional interpretation of an observable behavioral pattern depends on the activity being conducted. The OAP is therefore used as contextual evidence for possible teaching adjustments rather than as an automatic instruction command.

Table 7 establishes a transparent mapping between persistent observable behavioral patterns and candidate music teaching responses. The candidate responses are formulated as context-dependent instructional options rather than empirically established causal interventions. Their design draws on the broader emphasis on active, participatory, and embodied learning in music education.

The mapping does not assume that a single behavior has a universal educational meaning. In particular, standing may represent appropriate participation during a movement-based music activity but may have a different interpretation during an instructional stage in which students are expected to remain seated. Behavioral patterns are therefore interpreted according

to their duration and the current instructional context.

For candidate strategy q , a general profile-based suitability score can be represented as

$$R_q(t) = \beta_{q,1}PA_t + \beta_{q,2}MA_t + \beta_{q,3}OT_t$$

where, $\beta_{q,1}$, $\beta_{q,2}$, and $\beta_{q,3}$ represent the associations between candidate strategy (q) and the three observable activity dimensions. This profile-based score provides a general mathematical representation of strategy suitability; it does not encode all contextual conditions listed in Table 7. In particular, distinctions between individual off-task behaviors, temporal persistence, and the current instructional activity remain explicit contextual inputs to the decision-support process.

The strategy associated with the highest suitability score is

represented as $q_t^* = \arg \max_q R_q(t)$.

The coefficients $\beta_{q,m}$ define the mathematical structure of the decision-support framework but are not treated as empirically validated pedagogical effects in the present study. Their numerical calibration requires independent validation using dedicated music-classroom data.

Accordingly, q_t^* represents a candidate response generated from the OAP rather than an automatically executed teaching decision. Consistent with broader discussions of responsible AI-assisted music education, automated analysis is treated as instructional support rather than a replacement for the teacher's pedagogical role [15, 16]. The teacher retains responsibility for determining whether the suggested response is appropriate to the instructional context.

Table 7. Behavior-to-strategy mapping for music teaching

Observable Behavioral Pattern	Possible Classroom Interpretation	Candidate Music Teaching Response
Sustained low raising-hand activity	Limited overt voluntary participation	Introduce guided questioning, call-and-response, or a short rhythm-response task
Increase in standing during an intended movement-based activity	Increased observable physical participation	Continue or extend the movement-based rhythmic or performance activity when consistent with the lesson objective
Persistent standing outside the intended activity stage	Possible disruption or unplanned activity transition	Re-establish task structure or clarify the next activity
Persistent increase in mobile-phone use	Persistent observable OT	Shorten the current segment, introduce direct participation, or shift to a more interactive task
Persistent increase in sleeping behavior	Sustained observable inactivity during the current activity	Change the activity format, introduce a short demonstration, or provide a more active practice task
Concurrent increase in phone use and sleeping	Broad increase in observable OT	Consider an activity transition, guided participation, or differentiated task support
Stable raising-hand activity with low OT	Observable participation remains stable	Maintain the current instructional sequence

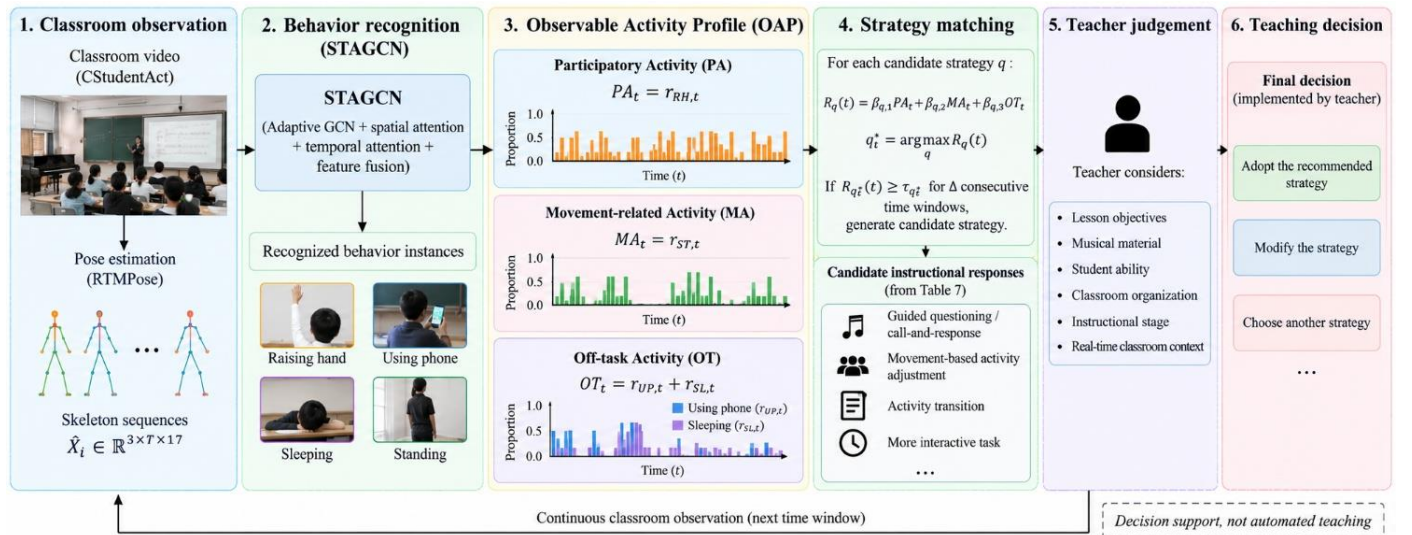


Figure 5. Behavior-driven music teaching decision-support framework

7.4 Adaptive music teaching decision framework

The behavior-recognition and instructional-interpretation components are integrated into a six-stage decision-support framework:

classroom observation → behavior recognition → OAP → strategy matching → teacher judgement → teaching decision.

At the observation stage, classroom video is processed to obtain student-centered skeletal sequences. STAGCN recognizes raising-hand, mobile-phone-use, sleeping, and

standing behaviors and produces temporally organized predictions. These predictions are aggregated according to Section 6.7 and subsequently transformed into the three-dimensional OAP defined in Sections 7.1 and 7.2.

The strategy-matching stage compares the current activity profile with the behavior-to-strategy relationships defined in Table 7. Rather than automatically modifying instruction, the framework presents the detected behavioral pattern and a candidate instructional response. Contextual information unavailable to the recognition model—including lesson

objectives, musical material, student ability, classroom organization, and instructional stage—remains part of the teacher's final decision.

The complete six-stage decision-support process and the relationship between behavior recognition, observable activity profiling, strategy matching, teacher judgement, and the final teaching decision are summarized in Figure 5.

To reduce responses to short-term fluctuations, a candidate strategy is considered only when the corresponding behavioral condition persists over a predefined temporal window. Let $R_q(t)$ denote the suitability score of strategy q . A candidate recommendation may be activated when $R_q(t) \geq \tau_q$ for a sustained interval, where τ_q denotes a strategy-specific decision threshold.

Because the present dataset does not contain direct measures of music learning outcomes or teacher intervention effects, neither the coefficients $\beta_{q,m}$ nor the thresholds τ_q are treated as validated pedagogical parameters. Their empirical calibration requires future studies conducted in authentic music-classroom settings.

The resulting framework maintains a clear separation between computational behavior recognition and prospective instructional application. STAGCN provides a structured mechanism for representing observable classroom behavior, whereas the strategy layer translates these representations into information that may support, but does not replace, professional teaching judgement [15, 16].

The present study therefore does not claim that the candidate strategies improve musical achievement, performance quality, or long-term learning outcomes. Establishing such effects would require controlled intervention studies in authentic music classrooms.

8. DISCUSSION

8.1 Spatiotemporal modeling of classroom learning behaviors

Classroom behavior recognition differs from conventional human action recognition in both motion scale and observational conditions. Many general action-recognition datasets contain large and visually distinctive movements, whereas classroom behaviors are frequently performed within restricted spaces and may involve relatively subtle changes in posture. The four behaviors considered in the present study also exhibit different motion characteristics. Raising hand is primarily characterized by coordinated upper-limb movement, mobile-phone use often involves localized movements of the hands, arms, and head, sleeping may be reflected by sustained changes in head and upper-body posture, and standing involves a broader change in body configuration.

The proposed STAGCN represents these behaviors as temporally ordered skeletal graphs rather than isolated image appearances. The 17-joint poses extracted using RTMPose-m provide the skeletal input to the model, while the resulting skeleton-based representation reduces direct dependence on classroom background, clothing, illumination, and other appearance variations [8, 17]. Spatial connections preserve the physical structure of the human skeleton, whereas temporal connections represent the evolution of individual joints across consecutive frames.

A fixed anatomical graph alone may not fully describe the joint relationships that are informative for different classroom

behaviors [9, 11]. For this reason, the proposed adaptive graph component supplements the predefined COCO skeleton topology with sequence-dependent joint relations. This mechanism allows joints that are not directly connected anatomically to exchange information when their coordinated patterns are relevant to behavior recognition. The learned adaptive topology therefore complements rather than replaces the physical skeletal structure.

The proposed representation is particularly relevant to the heterogeneous motion patterns present in CStudentAct [18]. Raising hand and standing contain relatively apparent motion transitions, whereas mobile-phone use and sleeping may involve more localized or sustained postural patterns. A representation that jointly models anatomical structure and temporal evolution may therefore provide more discriminative information than a purely frame-based description.

The methodological comparison in Section 6.1 places these design choices in relation to established skeleton-based action-recognition architectures. ST-GCN provides the reference formulation based on fixed spatiotemporal skeletal topology [8], whereas 2s-AGCN and CTR-GCN demonstrate different forms of adaptive topology learning [9, 11]. MS-G3D further illustrates the use of multi-scale spatiotemporal graph reasoning [10]. Against this background, STAGCN combines adaptive joint-relation learning with explicit spatial and TA within a unified architecture. The resulting contribution should therefore be interpreted in terms of representational design and methodological integration rather than as a claim of numerical superiority over methods evaluated under different datasets or experimental protocols.

8.2 Role of spatial and TA

Spatial and TA provide complementary mechanisms for emphasizing discriminative skeletal regions and informative portions of an action sequence, as demonstrated in previous skeleton-based attention models [12-14]. SA determines how information is distributed among skeletal joints within individual frames, whereas TA models the relative importance of different portions of an action sequence. Their separation is particularly relevant to classroom behaviors because discriminative information may be localized both anatomically and temporally.

The SA module produces a joint-to-joint relation matrix for each frame. This allows the contribution of individual skeletal regions to vary according to the observed behavior rather than assigning equal importance to all 17 joints. Raising hand, for example, is expected to depend strongly on coordinated upper-limb configuration, whereas mobile-phone use may involve more localized relationships among the hands, arms, head, and upper torso. Sleeping and standing exhibit different postural characteristics and may therefore produce different SA distributions. These examples describe plausible biomechanical patterns rather than empirically observed attention distributions. Section 6.5 therefore analyzes the mathematical properties and intended representational role of the SA mechanism without attributing unobserved joint-importance patterns to a trained model.

TA addresses a different problem. Classroom action sequences may contain both informative transitions and relatively stationary periods. In a raising-hand sequence, the transition from a resting arm position to an elevated posture may carry information that differs from subsequent frames in which the arm remains raised. Similarly, the transition

between seated and standing postures may contain temporal information that is not fully represented by either posture in isolation. The TA matrix therefore allows relationships among different frames to be learned explicitly.

The mechanism analyses in Sections 6.5 and 6.6 provide a mathematical interpretation of how spatial and TA operate within STAGCN. The spatial analysis derives a normalized sequence-level joint-importance representation from the frame-specific attention matrices, while the temporal analysis derives a normalized frame-importance profile from the TA matrix. These formulations clarify how the architecture can redistribute representation across skeletal joints and temporal positions, but they do not substitute for empirical visualization of attention learned by a trained model [12-14, 17].

The architectural decomposition in Section 6.3 further clarifies the distinct roles of AG, SA, and TA. AG modifies the effective relationships among skeletal joints [9, 11], SA models unequal contributions among joint relationships [12-14], and TA represents unequal contributions among different portions of an action sequence [12-14, 17]. Their integration within STAGCN is therefore motivated by complementary representational functions. Quantifying the independent performance contribution of each component would require controlled ablation experiments under a common implementation and evaluation protocol.

8.3 Implications for music teaching

The educational relevance of the proposed framework lies in converting automatically recognized classroom behaviors into structured observational information that can be interpreted in relation to instructional activity. Music lessons frequently alternate among explanation, listening, demonstration, rhythmic exercises, movement-based activities, individual practice, and group performance [15, 16, 21, 22]. Consequently, the instructional meaning of an observable behavior depends strongly on when it occurs and on the activity being conducted at that time.

The observable activity profile introduced in Section 7 provides a conservative means of organizing the recognition outputs. PA, MA, and OT are derived directly from the four behaviors recognized by STAGCN rather than from inferred psychological states. Their temporal variation can therefore describe changes in observable classroom activity without treating visual behavior as direct evidence of motivation, comprehension, cognitive engagement, or musical achievement.

Within this framework, raising-hand activity provides an observable indication of overt voluntary participation. A sustained reduction in raising-hand activity may therefore provide a reason to consider more participatory instructional formats, such as guided questioning, call-and-response, or short rhythm-response tasks. This interpretation does not imply that students who do not raise their hands are disengaged; rather, the detected pattern provides one source of observational evidence that may be considered together with the teacher's knowledge of the lesson context.

Bodily movement can form an integral part of music learning and performance activities [21, 22]. Standing therefore requires particularly context-dependent interpretation in music education. During movement-based rhythm exercises, singing activities, conducting practice, or performance tasks, standing may be entirely consistent with the intended instructional design. The same behavior

occurring persistently during a seated instructional segment may have a different meaning. MA is therefore not assigned a fixed positive or negative interpretation in the proposed framework. Its instructional relevance is determined in relation to the current lesson stage.

Mobile-phone use and sleeping are combined within the OT dimension as an operational grouping adopted for the present decision-support framework. This grouping describes observable behavioral patterns and does not establish the underlying reason for either behavior. A persistent increase in these behaviors may provide a reason to consider whether the current instructional segment should be shortened, reorganized, or replaced by a more active form of participation. Nevertheless, the visual model cannot determine why such behaviors occur, and the corresponding instructional response remains subject to teacher judgement.

The temporal nature of the OAP is particularly relevant to music teaching. Rather than reducing an entire lesson to a single engagement score, the framework retains changes in PA, MA, and OT across successive intervals. These temporal profiles can be compared with changes in instructional activity, allowing teachers to examine whether observable classroom behavior changes around demonstrations, rhythm exercises, activity transitions, or other teaching stages.

The scope of these implications should remain clearly defined. CStudentAct was developed for general classroom activity recognition rather than music education and contains only raising hand, using phone, sleeping, and standing. It does not represent music-specific behaviors such as score following, instrumental execution, singing participation, rhythmic synchronization, conducting gestures, or ensemble coordination. The present framework therefore demonstrates how recognized classroom behaviors can be transformed into structured decision-support information, but it does not establish that STAGCN can recognize the complete behavioral repertoire of a music classroom.

For the same reason, the behavior-to-strategy relationships defined in Section 7 should be interpreted as a transparent decision-support framework rather than experimentally validated pedagogical effects. This positioning is consistent with broader discussions of AI-assisted music education in which technological systems are treated as tools supporting teaching and learning rather than substitutes for pedagogical judgement [15, 16]. Demonstrating that behavior-informed strategy adjustment improves music learning would require dedicated music-classroom data and controlled intervention studies. The present work provides the technical and analytical foundation for such subsequent validation.

8.4 Limitations

Several limitations should be considered when interpreting the present study. First, the recognition framework is developed using a single public classroom behavior dataset. Although CStudentAct provides temporally connected action instances from authentic classroom scenes [18], its behavior categories, recording conditions, class distribution, and sequence characteristics were determined by the original data-collection protocol. The present study therefore cannot control all characteristics of the source data, and evaluation on additional continuous classroom datasets would be necessary to establish broader generalization.

Second, CStudentAct is a general classroom activity dataset rather than a dedicated music-classroom dataset. The present

recognition setting is limited to raising hand, using phone, sleeping, and standing. Performance on these four categories cannot be interpreted as evidence that STAGCN can recognize music-specific behaviors such as score following, instrumental practice, singing, rhythmic movement, conducting, or ensemble coordination.

Third, the recognition pipeline depends on two-dimensional pose estimation. Although RTMPose-m [20] provides a structured two-dimensional representation of body configuration, pose estimates can be affected by partial occlusion, camera perspective, image resolution, crowded seating arrangements, and incomplete visibility of individual students. Errors introduced during pose estimation may propagate into the subsequent graph representation and behavior classifier.

Fourth, skeleton representations intentionally discard much of the appearance and object information contained in RGB images [8, 17]. This can improve robustness to irrelevant variations in background and clothing but may also remove contextual information useful for particular behaviors. Mobile-phone use, for example, may sometimes be distinguished more reliably when the visual appearance of the phone is available in addition to hand, arm, and head posture. Multimodal integration of skeletal, RGB, object, and audio information may therefore provide a more complete representation.

Fifth, the observable classroom activity profile and behavior-to-strategy mapping proposed in Section 7 are descriptive decision-support components rather than validated measures of learning achievement. The associations between behavioral patterns and candidate teaching responses, together with the coefficients $\beta_{q,m}$ and decision thresholds τ_q , require independent empirical calibration before they can be used for consequential educational decisions. The framework should not be interpreted as an automated assessment of student motivation, comprehension, engagement, or musical ability.

Finally, classroom organization, student age, interaction patterns, camera configuration, and behavioral conventions may vary substantially across institutions and educational settings. Cross-dataset, cross-institutional, and cross-context evaluation is therefore required to determine the external validity of the proposed recognition framework.

8.5 Future work

Future research should first extend the present four-category recognition setting by constructing a dedicated music-classroom behavior dataset containing synchronized classroom video, skeletal trajectories, instructional-stage labels, and carefully defined music-specific behaviors. Such a dataset should provide substantially richer behavioral coverage than CStudentAct and may include score following, rhythmic imitation, singing participation, instrumental practice, movement-based musical activity, ensemble interaction, conducting-related gestures, and teacher-student musical exchange. Annotation should distinguish directly observable behavior from inferred cognitive or psychological states.

A second direction concerns multimodal modeling. Music learning inherently involves visual, auditory, and contextual information. Skeleton features can represent body configuration and movement, RGB or object features can capture interactions with instruments, scores, or mobile devices, and audio features can characterize musical events

and rhythmic structure. Integrating these modalities may allow music-specific behaviors such as rhythmic synchronization, instrumental execution, and coordinated ensemble activity to be modeled more effectively than with two-dimensional skeleton information alone.

Future studies should also evaluate the instructional component through controlled classroom interventions. A possible design would compare conventional teaching with behavior-informed instructional decision support across multiple lessons. The coefficients used in the strategy-mapping model and the corresponding decision thresholds could then be calibrated using authentic classroom observations rather than predefined assumptions. Outcomes may include observable participation, task completion, music-performance measures, teacher evaluation, and other independently defined learning indicators.

Model generalization should additionally be examined across different classrooms, student age groups, camera configurations, lesson types, institutions, and educational contexts. Cross-dataset evaluation would help determine whether the learned skeletal representations capture transferable behavioral patterns or depend strongly on the characteristics of CStudentAct.

Finally, future systems should investigate how automated behavioral information can be presented to teachers without increasing instructional burden. Rather than automatically changing teaching activities, an effective system should provide concise and interpretable behavioral summaries that allow teachers to combine model outputs with pedagogical knowledge and situational judgement.

9. CONCLUSIONS

This study developed a STAGCN for automatic recognition of continuous student behaviors in classroom environments and established a behavior-driven decision-support framework for the potential application of recognition outputs to music teaching. The proposed method represents student behavior using temporally ordered 17-joint skeleton sequences extracted with RTMPose-m. AGC supplements the predefined physical skeleton topology with data-dependent joint relationships, while spatial and TA mechanisms model behavior-relevant skeletal regions and informative temporal dependencies. The resulting features are integrated to classify four observable classroom behaviors: raising hand, using phone, sleeping, and standing.

The proposed architecture is developed using the four-category CStudentAct setting as its classroom-recognition context [18]. Its methodological characteristics are examined in relation to representative skeleton-based action-recognition architectures, including ST-GCN, 2s-AGCN, MS-G3D, and CTR-GCN [8-11]. The analysis shows that these approaches provide complementary methodological references involving fixed skeletal topology, AG, multi-scale spatiotemporal reasoning, and topology refinement. STAGCN integrates adaptive joint-relation learning with explicit spatial and TA, providing a unified representation for modeling heterogeneous skeletal and temporal characteristics of continuous classroom behaviors. The architectural and mechanism analyses further clarify the distinct representational roles of AG, SA, and TA without treating methodological differences as evidence of unmeasured numerical performance superiority.

Beyond individual action recognition, the four recognition

categories are organized into an observable classroom activity profile consisting of participatory activity, MA, and OT. The profile preserves temporal changes in observable classroom behavior rather than reducing an instructional period to a single engagement score. On this basis, a behavior-to-strategy mapping framework is established to associate persistent behavioral patterns with candidate music teaching responses. The framework retains teacher judgement as the final source of instructional decision-making and does not interpret visual behavior as a direct measure of motivation, comprehension, cognitive engagement, or musical ability.

The educational scope of the study remains deliberately limited. CStudentAct represents general classroom activity rather than music-specific learning behavior, and the present recognition framework does not represent score following, instrumental performance, singing, rhythmic synchronization, ensemble coordination, or other activities characteristic of music education. Moreover, the coefficients and decision thresholds used in the strategy-mapping framework have not been empirically calibrated through music-classroom interventions. Consequently, the present study does not establish that behavior-informed strategy adjustment improves musical achievement or learning outcomes.

Future research should extend the recognition framework using dedicated music-classroom datasets with richer behavioral annotations and synchronized visual, skeletal, and auditory information. Controlled classroom studies are also required to calibrate the proposed strategy-mapping mechanism and determine whether behavior-informed decision support produces measurable changes in participation, musical performance, or other independently defined learning outcomes. Such extensions would provide the empirical basis required to move from general classroom behavior recognition toward validated data-informed support for music teaching.

ACKNOWLEDGMENT

This paper was supported by Research Project of Shaanxi Provincial Department of Education (Grant No.: 22JK0168) and National College Students' Innovation and Entrepreneurship Training Program Project (Grant No.: 202411080008).

REFERENCES

- [1] Alfredo, R., Echeverria, V., Jin, Y., et al. (2024). Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6: 100215. <https://doi.org/10.1016/j.caeai.2024.100215>
- [2] Chen, H., Zhou, G., Jiang, H. (2023). Student behavior detection in the classroom based on improved YOLOv8. *Sensors*, 23(20): 8385. <https://doi.org/10.3390/s23208385>
- [3] Jia, Q., He, J. (2024). Student behavior recognition in classroom based on deep learning. *Applied Sciences*, 14(17): 7981. <https://doi.org/10.3390/app14177981>
- [4] Tang, L., Xie, T., Yang, Y., Wang, H. (2022). Classroom behavior detection based on improved YOLOv5 algorithm combining multi-scale feature fusion and attention mechanism. *Applied Sciences*, 12(13): 6790. <https://doi.org/10.3390/app12136790>
- [5] Li, Y., Qi, X., Saudagar, A.K.J., Badshah, A.M., Muhammad, K., Liu, S. (2023). Student behavior recognition for interaction detection in the classroom environment. *Image and Vision Computing*, 136: 104726. <https://doi.org/10.1016/j.imavis.2023.104726>
- [6] Yang, F. (2023). SCB-dataset: A dataset for detecting student classroom behavior. *arXiv preprint*, arXiv:2304.02488.
- [7] Yang, F., Wang, T. (2023). SCB-Dataset3: A benchmark for detecting student classroom behavior. *arXiv preprint*, arXiv:2310.02522.
- [8] Yan, S., Xiong, Y., Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1): 7444-7452. <https://doi.org/10.1609/aaai.v32i1.12328>
- [9] Shi, L., Zhang, Y., Cheng, J., Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 12018-12027. <https://doi.org/10.1109/CVPR.2019.01230>
- [10] Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W. (2020). Disentangling and unifying graph convolutions for skeleton-based action recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 140-149. <https://doi.org/10.1109/CVPR42600.2020.00022>
- [11] Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W. (2021). Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 13339-13348. <https://doi.org/10.1109/ICCV48922.2021.01311>
- [12] Song, S., Lan, C., Xing, J., Zeng, W., Liu, J. (2017). An end-to-end spatio-temporal attention model for human action recognition from skeleton data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1): 4263-4270. <https://doi.org/10.1609/aaai.v31i1.11212>
- [13] Cui, R., Zhu, A., Wu, J., Hua, G. (2020). Skeleton-based attention-aware spatial-temporal model for action detection and recognition. *IET Computer Vision*, 14(5): 177-184. <https://doi.org/10.1049/iet-cvi.2019.0751>
- [14] Qiu, H., Hou, B., Ren, B., Zhang, X. (2023). Spatio-temporal segments attention for skeleton-based action recognition. *Neurocomputing*, 518: 30-38. <https://doi.org/10.1016/j.neucom.2022.10.084>
- [15] Zhang, Y., Beh, W.F., Zhang, C., Pi, S. (2024). Transforming music education through artificial intelligence: A systematic literature review on enhancing music teaching and learning. *International Journal of Interactive Mobile Technologies*, 18(18): 76-93. <https://doi.org/10.3991/ijim.v18i18.50545>
- [16] Merchán Sánchez-Jara, J.F., González Gutiérrez, S., Cruz Rodríguez, J., Syroyid Syroyid, B. (2024). Artificial intelligence-assisted music education: A critical synthesis of challenges and opportunities. *Education Sciences*, 14(11): 1171. <https://doi.org/10.3390/educsci14111171>
- [17] Xin, W., Liu, R., Liu, Y., Chen, Y., Yu, W., Miao, Q. (2023). Transformer for skeleton-based action recognition: A review of recent advances. *Neurocomputing*, 537: 164-186.

- <https://doi.org/10.1016/j.neucom.2023.03.001>
- [18] SigM Lab. (2024). CStudentAct dataset: Continuous student activity recognition dataset. Hanoi University of Science and Technology, Hanoi, Vietnam. <https://sigmlab.com/datasets/CStudentAct/>.
- [19] Nguyen, P.D., Le, N.T., Bui, K.H., Nguyen, H.Q., Nguyen, H.Q., Le, T.L. (2025). A method for continuous student activity recognition from classroom videos. *International Journal of Machine Learning and Cybernetics*, 16(10): 7913-7937. <https://doi.org/10.1007/s13042-025-02695-w>
- [20] Jiang, T., Lu, P., Zhang, L., et al. (2023). RTMPose: Real-time multi-person pose estimation based on MMPose. arXiv preprint, arXiv:2303.07399.
- [21] Leman, M. (2007). *Embodied Music Cognition and Mediation Technology*. MIT Press. <https://doi.org/10.7551/mitpress/7476.001.0001>
- [22] Bremmer, M., Nijjs, L. (2020). The role of the body in instrumental and vocal music pedagogy: A dynamical systems theory perspective on the music teacher's bodily engagement in teaching and learning. *Frontiers in Education*, 5: 79. <https://doi.org/10.3389/feduc.2020.00079>