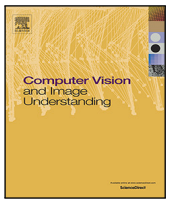
Computer Vision and Image Understanding 257 (2025) 104376

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Classroom teacher behavior analysis: The TBU dataset and performance evaluation

Ting Cai [a,b], Yu Xiong [b,*], Chengyang He [b], Chao Wu [b], Linqin Cai [b]

[a] The School of Communication and Information Engineering of Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
[b] The Artificial Intelligence and Intelligent Education Research Center of Chongqing University of Posts and Telecommunications, Chongqing, 400065, China

ARTICLE INFO

Communicated by Junsong Yuan

Dataset link: https://github.com/cai-KU/TBU

Keywords:
Teacher behavior
Classroom space
Behavior recognition
Temporal action localization
Behavior description

ABSTRACT

Classroom videos are objective records of teaching behaviors, which provide evidence for teachers' teaching reflection and evaluation. The intelligent identification, tracking and description of teacher teaching behavior based on classroom videos have become a research hotspot in the field of intelligent education to understand the teaching process of teachers. Although the recent attempts propose several promising directions for the analysis of teaching behavior, the existing public datasets are still insufficient to meet the need for these potential solutions due to lack of varied classroom environment, fine-grained teaching scene behavior data. To address this, we analyzed the influencing factors of teacher behavior and related video datasets, and constructed a diverse, scenario-specific, and multi-task dataset named TBU for Teacher Behavior Understanding. The TBU contains 37,026 high-quality teaching behavior clips, 9422 annotated teaching behavior clips with precise time boundaries, and 6098 teacher teaching behavior description clips annotated with multi-level atomic action labels of fine-grained behavior, spatial location, and interactive objects in four education stages. We performed a comprehensive statistical analysis of TBU and summarized the behavioral characteristics of teachers at different educational stages. Additionally, we systematically investigated representative methods for three video understanding tasks on TBU: behavior recognition, behavior detection, and behavior description, providing a benchmark for the research towards a more comprehensive understanding of teaching video data. Considering the specificity of classroom scenarios and the needs of teaching behavior analysis, we put forward new requirements for the existing baseline methods. We believe that TBU can facilitate in-depth research on classroom teacher teaching video analysis. TBU is available at: https://github.com/cai-KU/TBU.

1. Introduction

The analysis of classroom videos is a comprehensive task that aims to identify, detect, and describe the behavior of teaching subjects (teachers or students) to discover teaching patterns and enhance teaching quality (Rich and Hannafin, 2009) . Essentially, teacher behavior in the classroom reflects the nature of the teacher's relationships with students, as well as their pedagogical strategies (Osterman, 2023), which directly affects the teaching effect (Chen, 2020). Furthermore, teacher behavior can fruitfully reflect their cognitive and emotional status (Tammets et al., 2022), which has received attention from educational scholars. In recent years, because of its great research and application value, increasing efforts have been devoted to exploring different aspects of teacher behavior analysis, involving behavior identification (Gang et al., 2021; Li et al., 2022), classroom spatial location (Li et al., 2023; Karumbaiah et al., 2023), sequential actions (Karumbaiah et al., 2023; Sharpe et al., 1997; Martinez-Maldonado et al., 2022), behavior description (Martinez-Maldonado et al., 2022; Berliner, 2004). These studies indicate that teacher behavior can reflect a teacher's teaching skills (Gaudin and Chaliés, 2015) and cognitive status (Hollingsworth and Clarke, 2017), and has a positive impact on students' perceived of support and participation (Warren, 2019; Dahleez et al., 2021).

Traditional methods for analyzing teacher behavior primarily rely on video annotation tools to identify behavior patterns in individual or entire libraries of classroom teaching videos. Rodgers et al. (2019) utilized technology-supported observation and coaching tools to record teachers' classroom teaching actions and generate descriptive language about their teaching processes, aiding in reflecting on their teaching practices. However, these methods mainly rely on manual observation,

* Corresponding author.
E-mail addresses: caiting@cqupt.edu.cn (T. Cai), xiongyu@cqupt.edu.cn (Y. Xiong), s220331036@stu.cqupt.edu.cn (C. He), wuchao@cqupt.edu.cn (C. Wu), cailq@cqupt.edu.cn (L. Cai).

https://doi.org/10.1016/j.cviu.2025.104376
Received 5 November 2024; Received in revised form 16 February 2025; Accepted 23 April 2025
Available online 28 April 2025

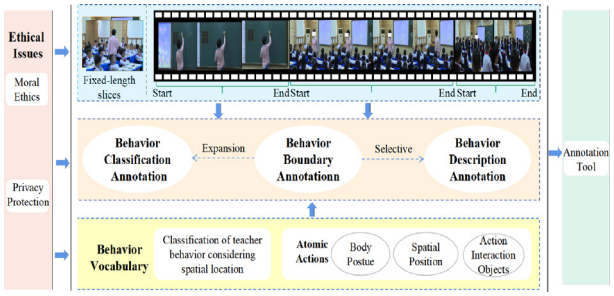


Fig. 1. A multi-task video dataset framework for classroom teacher teaching behavior.

which is time-consuming and labor-intensive. Additionally, these methods are often limited to small-scale teaching videos, which reduces the accuracy and reliability of analysis outcomes. With the integration of information technology in education, many researchers have collected datasets for automatic analysis of teacher behavior (Li et al., 2022; Sharma et al., 2021; Jia et al., 2023; Zhao et al., 2024; Wu et al., 2024). Although these datasets are helpful for exploring teacher behavior, they still have the following problems: (1) Insufficient consideration of teacher behavior characteristics. Teacher behavior is driven by teaching scenarios and students' acceptance of classroom learning, which are manifested through changes in spatial movement, interactive object selection, and fine-grained actions (Sharpe et al., 1997; Martinez-Maldonado et al., 2022). The existing datasets do not fully consider the behavioral characteristics of teachers, ignoring spatial location and interaction objects. (2) Limited support for behavior understanding tasks. Currently, most teacher behavior datasets focus on behavioral classification, overlooking the need for behavior descriptions, which are essential for capturing the progression of teacher behavior. This is because the efficient behavior description is the first step to interpret teaching behavior, including the description of specific behaviors, interaction objects (Martinez-Maldonado et al., 2022; Berliner, 2004). (3) Uniform classroom environment and fixed shooting mode. The single classroom environment and unchanged shooting mode often come with fixed classroom settings and student compositions, which restrict the diverse expression of teacher behavior and constrain the generalization ability of algorithm models. Therefore, it is expected to establish a large-scale, diverse, multi-task and scenario-specific video dataset to drive the comprehension of teacher behavior.

In our previous work (Cai et al., 2024), we introduced the Teacher Behavior Understanding (TBU) dataset along with initial findings. Building on this foundation, this article presents a comprehensive multi-task video dataset framework for classroom teacher teaching behavior (see Fig. 1). This study extends prior work in the following ways: (1) A comprehensive analysis of factors influencing teacher behavior, recent advancements in video-based teacher behavior understanding, and relevant datasets. (2) A multi-task annotation framework for teaching behavior is proposed, supported by a custom-built collaborative annotation tool. (3) The dataset formation process is expanded with additional dimensions of statistical analysis; and (4) Relevant experiments and comparative analyses are conducted to assess the dataset's practical value. The main contributions of this article are summarized as follows:

1. This article investigates factors that influence teacher behavior in the classroom and proposes a multi-task framework for classroom teacher teaching behavior, providing a reference for a comprehensive understanding of teacher behavior.
2. We construct the TBU, a novel large-scale dataset of teacher behavior videos across multiple educational stages. The dataset considers the teacher's spatial location and supports video comprehension tasks such as behavior recognition, behavior detection, and behavior description, offering rich, high-quality data samples for analyzing teacher behavior.
3. The TBU dataset is evaluated using 6 behavior recognition models, 2 behavior detection models, and 3 behavior description models. The experiments highlight the practicality and challenges associated with the TBU dataset.

2. Relate work

In this section, the understanding of teacher behavior based on video, and video datasets of classroom teaching behavior are analyzed in detail.

2.1. The understanding of teacher teaching behavior based on video

Analyzing classroom teaching videos helps teachers understand their own behaviors, which can ultimately enhance their teaching practice (Hamel and Viau-Guay, 2019). Advancements in data-driven AI for education has promoted the automatic analysis of teacher behavior in large-scale classroom teaching videos (Ahmad et al., 2024).

(1) Categories of Teacher Behavior: Currently, the analysis of classroom teacher behavior is carried out from three aspects. The first is the identification of overall information behaviors, such as bowing, standing, walking, looking, board and wiping the blackboard (Gang et al., 2021; Sharma et al., 2021; Shi et al., 2016; Wu et al., 2020; Li et al., 2022). The second is the recognition of local behaviors, which is gestures, head posture, and eye gaze (Fernandez-Nieto et al., 2022; Chen et al., 2022). The third is the reasoning of teaching significance, which is schematic, guiding, evaluative (Pang et al., 2022; Liu et al., 2024). From the analysis results, the typical progression is from coarse to fine. Specifically, it is from single behavior to multi-modal behavior, from behavior classification to fine-grained behavioral descriptions, and from behavioral conceptual reasoning with educational characteristics to scene reasoning. According to the "spatial pedagogy" (Lim et al., 2012) and integrative view (Korthagen, 2010), the teacher's pedagogical state is closely related to their teaching behaviors in different classroom locations. Therefore, a teacher behavior dataset that adheres to the characteristics of classroom space is a foundation for studying teachers' teaching status.

(2) Identification of Classroom Teacher Behavior: Methods for automatically identifying teacher behavior can be divided into two types: body skeleton-based and spatiotemporal context-based approaches. The former mainly extracts the skeletal feature information of the entire body or multiple parts of the body in the image to achieve the purpose of behavioral classification (Zhao et al., 2024; Wu et al., 2024; Shi et al., 2016; Pang et al., 2022; Liu et al., 2024). Although the recognition methods based on body skeleton have shown good performance in single-person behavior recognition, but they are vulnerable to lighting interference and object occlusion, and cannot effectively identify teacher–student interaction behaviors in complex backgrounds. Spatiotemporal context-based recognition methods address these issues to some extent by leveraging strong spatiotemporal modeling capabilities. Gang et al. (2021) and Li et al. (2022) designed spatial–temporal information fusion network based on 3DCNN to improve the accuracy of teacher behavior recognition. Currently, notable video behavior recognition models include 3D CNNs for robust spatiotemporal modeling (e.g., C3D (Tran et al., 2015) and I3D (Carreira and Zisserman, 2017)), Temporal Segmentation Networks (TSN) (Wang et al., 2016), and two-stage networks with different speeds (SlowFast (Feichtenhofer et al., 2019)), self-attention mechanism model (Timesformer (Bertasius et al., 2021)), and large pre-trained visual language models like "Bldirectional cross-modal Knowledge Exploration(BIKE)" (Wu et al., 2023).

(3) Detection of Classroom Teacher Behavior: Different from behavior recognition, behavior detection not only identifies the behavior category, but also detects the start time and end time of the behavior. Current method for time-series teacher behavior detection is mainly carried out based on the S–T (Student–Teacher). The S–T analysis

method identifies the teacher behavior categories in fixed-length clips based on S–T behavior coding, and then derives the sequence of teacher behaviors in the classroom videos (Zhan et al., 2021). The popularization and application of information technology has facilitated the gradual development of the S–T analysis method from manual observation and recording to automatic identification. Due to its limitation to fixed-length video clips, the S–T analysis method cannot accurately capture the true duration of behaviors. Temporal behavior detection in AI can identify behaviors in videos of any length and primarily includes two approaches: single-stage and two-stage methods. The single-stage method directly predicts the presence and location of actions within the video, as seen in Zhang et al. (2022). In the two-stage method, candidate frames are first selected from the original video and then processed by a network to detect behavioral boundaries, as demonstrated in Zhao et al. (2017).

(4) Description of Classroom Teacher Behavior: Teacher behavior descriptions capture the fine-grained details of teachers' actions during teaching activities or events, providing a foundation for exploring the relationship between behavior and classroom outcomes. Teacher behavior in the classroom can reflect the teacher's professionalism, and effective descriptions of teachers' teaching behavior characteristics have been found to be of heuristic value (Berliner, 2004; Van der Lans et al., 2018). The descriptions of teacher behavior based on classroom observation are manual records of behaviors and their corresponding teaching effects. Manual recording methods provide more accurate analysis results within a shorter time, but it is limited to small-scale short-duration teaching videos due to heavy manual recording workload and difficulty in tracking. Precise teaching behavior descriptions supported by information technology effectively demonstrate teaching situations in recorded video clips, facilitating teaching reflection activities (Almodaires, 2009). The existing AI-based analysis of teacher behaviors focuses on statistics and clustering methods, which fail to effectively understand the mechanism behind teaching behaviors and do not support teachers' professional development. However, teacher-centered behavior descriptions are essential for tracking the evolution of these behaviors (Berliner, 2004). It can be seen that the description of teacher behavior is a detailed record of behavior to explore the rules behind teacher behavior.

2.2. Video datasets of classroom teaching behavior

Researchers have produced some classroom teaching behavior video datasets based on actual needs, including TAD-08 (Gang et al., 2021), EduNet (Sharma et al., 2021), TAR (Jia et al., 2023), CTA (Wu et al., 2024), PTPD (Zhao et al., 2024), CCTV (Rafique et al., 2022), BNU-LCSAD (Sun et al., 2019), Student Class Behavior Dataset (Sun et al., 2021), IAVID-I (Nida et al., 2019), etc. (see Table 1).

As shown in Table 1, teaching behavior video datasets are evolving towards greater diversity. Specifically, they have expanded from behavior recognition and detection to behavior description tasks, cover both simulated and real classrooms, and span multiple educational stages. This shift reflects the growing emphasis on diverse classroom settings and educational stages in evaluating teaching effectiveness.

(1) Behavior Classification Dataset: The dataset for teacher behavior primarily consists of image-based datasets and video-based datasets. Image-based datasets usually provide annotations for a single image. However, such static individual images are insufficient to support in-depth analysis and reasoning of the behavior. The video-based datasets offer annotations for temporal clips, whose information is dynamic and can infer the evolution of behavior at diverse stages. Then, a number of representative video-based teacher behavior datasets are proposed. TAD-08 (Gang et al., 2021) dataset is a classification annotation of eights behaviors in 2048 behavior sequences from a middle school. It provides a uniform classroom environment with no camera motion. EduNet (Sharma et al., 2021) collects 7851 teaching video clips from primary and secondary schools (excluding university stage). It contains nine categories of teacher behaviors, but some classifications have less relevance to behavior reasoning (e.g., Hold_Mobile_Phone). TAR (Jia et al., 2023) collates 13,288 valid video samples from 25 teachers in eight classrooms of vocational and technical schools, and contains 5 categories of normative teaching behaviors and 3 categories of misbehaviors. The CTA dataset (Wu et al., 2024) includes 12 types of behaviors in university classrooms, such as writing on the blackboard, wiping the blackboard, and gesturing with hands. Although PTPD (Zhao et al., 2024) innovatively proposes understanding teacher–student interactions from different perspectives of teachers, it only supports the classification of behaviors. Due to limited consideration of classroom setting diversity, educational stages, and the comprehensiveness of video comprehension tasks, datasets like TAD-08, EduNet, TAR, CTA and PTPD cannot be well generalized to other classroom scenarios.

(2) Temporal Behavior Detection Dataset and Video Behavior Description Dataset: Temporal behavior detection datasets are typically labeled with behavior categories, behavior start and end times for unedited videos, with each video segment including at least one behavior. However, only a few datasets support temporal action localization in teacher behavior due to the intensive manual annotation required. Gang et al. (2021) and Jia et al. (2023) generated behavior classification segments by marking behavior boundaries; however, they do not support behavior detection tasks due to the extensive workload involved.

In video behavior description datasets, each video clip sample is associated with one or more description sentences. Objective recordings of teaching behaviors support reflective teaching practices. Rodgers et al. (2019) described the process of teachers' teaching activities in the classroom by recording the frequency of teachers' teaching behaviors; however, this approach is primarily a clustering of behaviors. Most descriptions of teaching activities are based on classroom observation that record teacher behavior (such as body postures, interactions, sound data, etc.) and student feedback (Hollingsworth and Clarke, 2017; Allas et al., 2020; Almodaires, 2009). The Measures of Effective Teaching (MET) dataset, a large-scale and impactful educational research resource funded by the Bill & Melinda Gates Foundation (Cantrell and Kane, 2013), records scores in two classroom-focused domains: classroom environment and instruction. The 'classroom environment' domain includes managing classroom procedures, student behavior, and organizing physical space, while the 'instruction' domain encompasses communicating with students, lesson engagement, and responsive teaching. The MET can be regarded as a description and evaluation of teachers' classroom behavior based on manual observation. Advances in information technology have enabled the automated analysis of teacher behavior descriptions. To our knowledge, no existing video dataset is specifically dedicated to detailed teacher behavior descriptions.

Table 1 shows the comparison of TBU with other existing datasets. Compared with previous datasets, TBU is sourced from real classroom environments across four educational stages, encompassing a variety of classroom settings. Additionally, TBU supports multiple video understanding tasks, including behavior recognition, temporal action localization, and video behavior description. Moreover, the dataset incorporates spatial location information for behavior categorization and provides fine-grained annotations of teaching behaviors along with multi-level atomic actions. These features enable TBU to accurately model teachers' specific behavior patterns across different instructional contexts, offering richer semantic information for understanding teaching behaviors. In the next section, we will further explore the influencing factors of teacher behavior, providing theoretical support and rationale for the design of TBU.

3. The influencing factors of teacher behavior

Video-based teaching reflection is a learning process closely associated with classroom practice and personal nature (Hollingsworth and Clarke, 2017). Teachers' teaching behaviors and decisions are affected

Table 1
Some video datasets related to classroom teaching behavior.

Database	Action subject	Behavior type	Number of samples	Supported task type
TAD-08 (Gang et al., 2021)	College teachers	8 types: Bowing to students; Pointing to the blackboard; Writing on blackboard; Cleaning the blackboard; Operating the interactive whiteboard; Inviting students to answer questions; Walking around classroom; Operating the realia	2048	Behavior recognition; Temporal action localization
EduNet (Sharma et al., 2021)	Primary and secondary school teachers	9 types: Explaining_the_subject; Hitting; Holding_Book; Holding_Mobile_Phone; Holding_Sticks; Sitting_on_Chair; Slapping; Walking_in_Classroom; Writing_On_Board	7851	Behavior recognition
TAR (Jia et al., 2023)	College teachers	8 types: Upright; Lean forward; Turn around; Small intrathoracic movements; Large extrathoracic movements; Leaning the platform; Hand in pocket; Hands at the back	13 288	Behavior recognition; Temporal action localization
CTA (Wu et al., 2024)	University teachers	12 types: Writing on the blackboard; Wipe the blackboard; Point to the blackboard; View/Operate the equipment; Gesturing with the hands; Side teaching; Move teaching; Seated teaching; Holding the book; Holding the phone ; Hand supporting while teaching	13 000	Behavior recognition
CCTV (Rafique et al., 2022)	College teachers	10 types: Standing; Writing; Pointing; Talking; Cleaning; Delivering Presentation; Presentation, Writing; Delivering; Presentation, Pointing on Board; Presentation, Talking; Presentation; Cleaning; Writing, Talking	1050	Behavior recognition
PTPD (Zhao et al., 2024)	Primary students and teachers; Middle students and teachers; University students and teachers	19 types: Point to multimedia-Look up; Point to multimedia-writing; Writing on the blackboard-Look up; Writing on the blackboard-Writing, Demo-Imitate; Demo-Look up; Visits-Writing; Visits-Read; Visits-Turned; Questions-Hands; Point to multimedia-Read; Point to multimedia-Turned; Writing on the blackboard-Read; Writing on the blackboard-Turned; Demo-writing; Demo-Turned; Visits-Look up; Questions-Look up; Questions-Read	13 500	Behavior recognition
BNU-LCSAD (Sun et al., 2019)	College students	8 types: Taking notes; Listening carefully; Using computers; Looking around; Playing mobile phones; Sleeping or snoring; Eating or drinking; Reading, discussing, and yawning	1670	Behavior recognition
Student Class Behavior Dataset (Sun et al., 2021)	College students	10 types: Listening carefully; Taking notes; Using mobile phones; Yawning, eating or drinking; Reading; Discussing; Looking around; Using computers; Sleeping or snoring; Raising hands	16 457	Behavior recognition; Temporal action localization; Video behavior description
IAVID-I (Nida et al., 2019)	Actors	9 types: Interacting or idle; Pointing towards the board; Pointing towards the screen; Using a mobile phone; Using a laptop; Reading notes; Sitting; Walking; Writing on the board	2898	Behavior recognition
TBU (Cai et al., 2024)	Primary teachers; Junior teachers; High teachers; University teachers	13 types: Board writing; Erasing the blackboard; Operating multimedia; Multimedia teaching; Teacher bowing; Displaying teaching aids; Lecture on the podium; Classroom inspection around on the podium; Interacting with students on the podium; Lecture underneath the podium; Classroom inspection underneath the podium; Interact with students outside the podium; Pointing to the blackboard	52 546	Behavior recognition; Temporal action localization; Video behavior description

by classroom space situations and teachers' cultural factors (Allas et al., 2020).

Classroom space significantly influences teachers' behaviors and decisions, acting as the main environment for teaching practices. The term "spatial pedagogy" refers to the patterns of orientation and movement in learning spaces when teachers interact with resources and students (Lim et al., 2012). The "spatial pedagogy" holds that the expression of teacher behavior is closely related to spatial location and interactive objects. Based on the spatial pedagogy, researchers have conducted an in-depth analysis of the relationship between teachers' classroom spatial location and teaching effectiveness through case studies or special sensor equipment. Morell (2018) analyzed an example of a lecturer's five-and-a-half-minute classroom activities and found that teachers need to use classroom space effectively to achieve specific pedagogical. Fernandez-Nieto et al. (2022) constructed a "Classroom Dandelions" system to enhance the understanding of the impact of classroom space on teacher teaching. They found that the teacher's classroom position, trajectory, and body direction contribute to teaching reflection. Li et al. (2023) found that teachers' teaching behaviors in different classroom spatial locations are positively related to the support they provide to students. After understanding the relationship between teacher teaching and classroom spatial location, scholars further explored the role of teacher–student interaction in different spatial locations on student learning effects. Karumbaiah et al. (2023) collected fine-grained data on teacher location and interactions with students in the physics classroom, and revealed that the close interaction between teachers and students has a positive impact on student behavior. Wang et al. (2024) investigated the relationship between classroom interaction and instructional proxemics, and showed that there are significant differences in the spatial location (e.g., as standing on the podium, student area, etc.) and the interaction object (multimedia PowerPoint, teaching aids, etc.) of teachers for different types of classroom interaction.

Teacher behavior is also affected by personal experience, knowledge, attitude and other factors (Clark, 1986). In many cases, teachers may be not fully aware of the impact of their behavior and its underlying causes (MacKay et al., 1980). In other words, teachers' explicit teaching behavior is an implicit representation of teachers' personal status. Korthagen (2010) proposed a three-level learning model including holistic learning, element relationship network and theory, to analyze the friction between teacher behavior in practice and teachers' practices in theory. This model reconciled the situated learning perspective with cognitive theory, indicating that empirical data on teaching behavior and brain research should be fully utilized in teaching practice to clarify

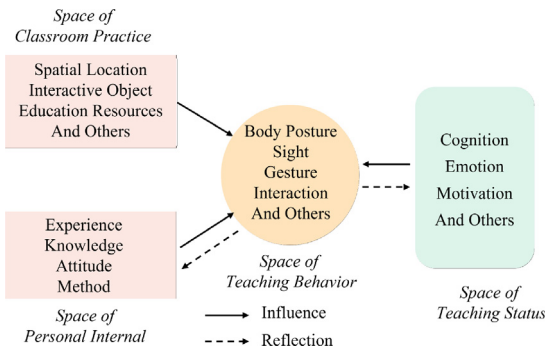


Fig. 2. The relationship between teacher behavior and teaching reflection.

the relationship between theory and cognition in teaching reflective learning. It is evident that the unconscious teaching behaviors can be developed into conscious cognitive patterns in the teaching reflection supported by behavioral data. Therefore, teachers' variable spatial locations and interactive object behaviors in classroom mirror their different mental states.

In general, classroom space is an explicit factor that affects teachers behavior, and teachers' personal factors are an implicit factor. The teacher's teaching status represents an internal aspect of their behavior and reflects a conceptual transformation of their personal space. Classroom practice and internal personal factors together manifest as external expressions of teachers' instructional states (see Fig. 2). Teacher behaviors are systematically recorded in classroom teaching video. Analysis of teaching behaviors based on classroom videos is a powerful way to support teachers' professional development.

Drawing from the preceding analysis, we believe that a video dataset encompassing diverse educational stages and multiple analytical tasks is crucial for comprehensive teacher behavior analysis. To effectively analyze and apply real classroom teacher behaviors, the following points need to be considered in the production of this dataset:

1. The classification of teacher behavior should consider the teacher's spatial location in the classroom, as it directly influences interaction dynamics and teaching efficacy.
2. The video data should encompass multiple educational stages and classroom environments to ensure the dataset's generalizability.
3. The dataset should support various video understanding tasks, including behavior recognition, behavior detection, and behavior description.
4. The teacher behavior description dataset should record multi-level relationships, such as fine-grained behaviors, spatial locations, and interactive objects.
5. High-quality data annotation tools should be available to meet actual needs.

4. The TBU dataset

Our intention is to create a large-scale, high-quality dataset for teacher behavior that supports behavior classification, detection, and description tasks. Informed by detailed analysis and research, we proposed a multi-task framework for teacher behavior in the classroom (see Fig. 1). Our previous work (Cai et al., 2024) briefly introduced the dataset formation process. In this study, we further refine the annotation process and clarified the relationship between each annotation stage. As shown in Fig. 3, this carefully designed annotation framework aims to enhance the consistency and accuracy of fine-grained annotations. This section further expands the framework from multiple perspectives.

4.1. Action vocabulary generation

As discussed in Section 3, teaching behavior is influenced by classroom space and personal factors, creating a reflection of the classroom environment and individual aspects of the teacher. In classroom teaching practice, spatial positioning plays a crucial role in behavior classification, as elaborated in Section 2.1. Following prior research (Gang et al., 2021; Sharma et al., 2021; Jia et al., 2023; Shi et al., 2016; Wu et al., 2020; Li et al., 2022), we identified 13 distinct types of teacher behaviors (see Table 2). Unlike other video datasets, our classifications account for behaviors "on the podium" and "underneath the podium (among students)," enhancing relevance for diverse classroom dynamics.

Atomic action theory suggests that human behavior is composed of multiple atomic actions (Gu et al., 2018), including: (1) body posture and displacement; (2) interaction behaviors with objects; and (3) interaction behaviors with other people. Similarly, teacher behavior is also composed of various atomic actions, such as body posture, spatial position, and action interaction objects, which is associated with the spatial pedagogy. At the level of behavioral object, many factors are involved: students, teaching progress, teaching resources, and teaching environment. At behavioral expression level, it is embodied in physical contact, posture, proximity, body, physical, and eye movement. Unlike atomic actions in other behavioral scenes, atomic actions in teaching have distinct contextual characteristics and educational significance.

Guided by the concept of spatial pedagogy (Lim et al., 2012) and insights from MET-based teaching effectiveness analysis reports, atomic actions of teacher behavior in the classroom have been designed (see Table 3). The behavioral hierarchy of body posture includes fine-grained body posture, eye interaction objects and body orientation. It is essential to establish the main categories of teacher behavior before determining specific atomic actions, as this prevents arbitrary annotation and ensures systematic analysis.

4.2. Data collection

(1) Source Data: We obtained the classroom teacher videos from the National Public Service Platform for Educational Resources (https://www.eduyun.cn), which provides open source, high-quality teacher teaching videos of real classrooms at different educational stages and regions. These teaching videos are shot in various ways, including fixed shots, following shots, and close-up shots. To clearly identify teachers' verbal and gestural behaviors, all videos are required to have a minimum resolution of 720p and a frame rate of 25 fps.

(2) Data Sample Formation Method: The tasks for understanding teacher behavior include classification, detection, and description. For video behavior classification samples, the commonly used fixed-length cropping method is adopted, as seen in Gang et al. (2021), Sharma et al. (2021) and Rafique et al. (2022). The following describes methods for creating samples for behavior detection and description.

We use two methods to form behavior detection samples: dense frame-level annotation and video clip segment annotation, depending on the scene's characteristics. In dense frame-level annotation, video actions are characterized by rigorous temporal ordering and instructional sequence guidance, such as FineAction (Liu et al., 2021). Video clip segment annotation is more concerned with whether the action occurs rather than strict temporal sequencing. It emphasizes action localization and camera transitions, as seen in datasets like ActivityNet, THUMOS'14, etc. Teacher behavior is influenced by classroom environment, teaching progress, and the teacher's understanding of instructional objectives. Unlike some other instructional video, classroom teaching lacks a strict temporal sequence. Therefore, the duration of actions and the reasoning of contextual information are more concerned in analyzing actual teacher behavior. Accordingly, this article provides temporal behavior detection samples based on video clip annotation.

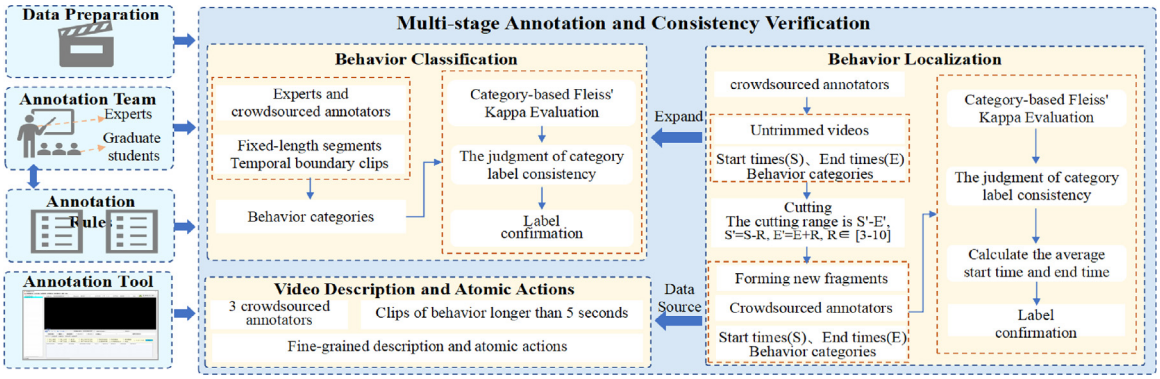


Fig. 3. Detailed annotation framework for TBU.

Table 2
Teacher behavior categories and category descriptions.

Behavior category	Description
Board Writing	Teachers write content on the blackboard or whiteboard to show and record relevant knowledge points, concepts, formulas, charts, etc. to students.
Erasing the Blackboard	Teachers use tool to erase content written on blackboard or multimedia whiteboard.
Operating Multimedia	It is mainly for teachers to open, close, switch different multimedia devices, and adjust multimedia related parameters.
Multimedia Teaching	Teachers teach students based on course content (such as pictures, text, videos, etc.) displayed and explained on multimedia devices (mainly projectors).
Teacher Bowing	The teacher bends forward and bows slightly to the students. It usually occurs before the formal class and at the end of the class.
Displaying Teaching Aids	Teachers use physical objects, models, charts, pictures or other teaching aids to show and explain course content to students during the teaching process.
Lecture On the Podium	It refers to teaching activities in which teachers stand on the podium and impart knowledge and skills to students through oral explanation as the main method.
Classroom Inspection Around On the Podium	It refers to the teaching activities in which teachers inspect students' learning status and behavioral performance on the podium.
Interacting with Students On the Podium	Teachers interact with students on the podium.
Lecture Underneath the Podium	It refers to teaching activities in which teachers leave the podium and impart knowledge and skills to students through oral explanation as the main method.
Classroom Inspection Underneath the Podium	It refers to the teaching activities in which teachers move around the classroom and inspect students' learning status and behavioral performance.
Interact with Students Outside the Podium	Teachers interact with students in the middle of the classroom (or among students).
Pointing To the Blackboard	The behavior of teachers pointing to the blackboard or whiteboard with their hands, fingers or teaching aids in class.

Table 3
The atomic actions of teacher behavior.

Behavioral hierarchy		Atomic actions labels
Body Posture	Fine-Grained Body Posture	Sitting; Standing; Bending Down; Standing to Bending; Standing to Sitting; Walking on the podium; Walking from above podium to below podium; Walking from below podium to above podium; Walking around under the podium; Walking behind the class.
	Eye Interaction Objects	Scan students; Staring at a student; Students in classroom (not scanning, no target students); Group or team; Blackboard; Multimedia PPT; Teaching materials and aids; Podium items.
	Body Orientation	Facing the students; Sideways; Facing the blackboard; Facing the multimedia PPT; Others.
Spatial Position		Behind the podium; Beside the podium; In front of the blackboard; In front of the multimedia PPT; Next to the lecture table; Among the students; At the back of the classroom; Side of the classroom; Next to the students.
Action Interaction Objects		Students; Podium; Blackboard; Chalk; Teaching aids; Multimedia PPT; Books or examination paper.

Behavior description sample annotation methods are divided into dense annotation and single activity annotation. The former provides descriptive information for multiple stages within a long video, such as Youcook II (Zhou et al., 2018). The latter uses multiple descriptive sentences for the same video clip, as found in dataset like TACoS (Rohrbach et al., 2012), MSR-VTT (Zhou et al., 2019), and MSVD (Chen and Dolan, 2011). Typically, a behavior description sentence includes video objects, actions categories, and scene descriptions. For annotations focusing on 'teacher behavior', it is essential to emphasize both the behavioral process and level of granularity within the scene context.

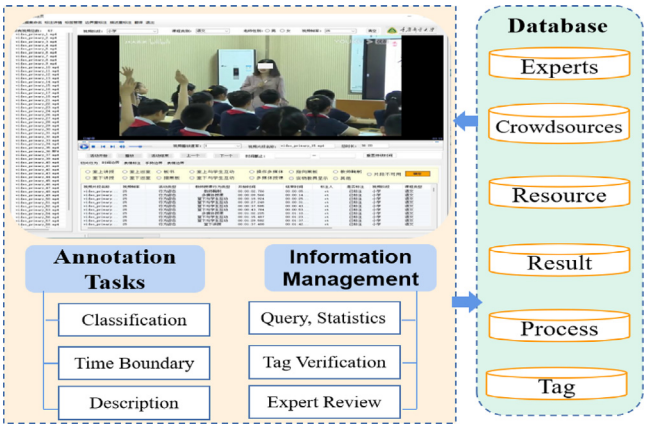


Fig. 4. Video annotation tool.

At the same time, the diversity of description sentences significantly impacts the performance of the behavior description model. Therefore, this article proposes a high-quality corpus of teacher behavior descriptions for single sentence generation (the latter). Additionally, to gain a deeper understanding of classroom teachers' teaching behaviors, fine-grained atomic actions representing teacher behavior combined with actual classroom teaching scenarios are provided.

(3) Data Sample: Data sample consists of fixed-length slices and raw videos, where fixed-length slices (3, 5, and 10 s) are used for behavior category annotation, and full classroom videos (30–45 min) are used for temporal boundary annotation.

4.3. Annotation tool

To create a high-quality dataset, we developed a multi-task collaborative annotation tool for fine-grained annotation of classroom teacher behaviors (see Fig. 4). Multiple data tables (experts, crowdsources, data resources, annotation results, annotation processes, and tags) were established to facilitate the maintenance of annotation data and annotation processes. The tool has two main functions: constructing annotation tasks and managing annotation information. The former supports tasks including classification, video time boundaries, and descriptions. The latter embeds functional modules for querying, statistics, crowd annotation, verification and expert review. This tool is adaptable for annotating images and videos across various scenes, covering categories, descriptions, and behavior time boundaries.

4.4. Data annotation

The data samples of TBU mainly consist of fixed-length slices and raw videos, where fixed-length slices (3 s, 5 s, 10 s) are used for behavior category annotation, and raw videos (30–45 min) are used for temporal boundary annotation.

TBU was formed through a combination of crowdsourced and expert annotations. Experts trained crowdsourcing annotators, emphasizing accurate labeling of visible body segments, and explained the meaning of behavior categories and atomic actions. Before the formal annotation, crowdsourcing annotators completed a two-day practice and feedback session to ensure comprehension of teacher behavior, classroom space, and interaction objects. TBU includes three sub-datasets: Behavior Recognition (TBU_R), Behavior Localization (TBU_L), and Behavior Description (TBU_D), as shown in Table 4. And, TBU_R supports behavior classification, TBU_L supports behavior detection, and TBU_D supports behavior detection. TBU is available at: https://github.com/caiKU/TBU.

(1) Behavior Recognition Annotation: The TBU_R consists of randomly fixed-length segments and temporal boundary clips. The labels for the fixed-length segments were confirmed by seven crowdsourced annotators and two experts. After completing the annotations, Fleiss' Kappa was used to verify the consistency of the category labels. Once consistency reached a high value, the data range was confirmed. For category labels with ≥5/7 consistency, the segment label was directly confirmed. Categories with 4/7 or 3/7 agreement were reviewed by experts, and categories with fewer than 3/7 agreement were discarded. In the end, 28,636 usable fixed-length instances and 8390 valid time boundary clips were obtained.

Table 4
The composition of the TBU dataset.

TBU	Total samples	Labels
TBU_R	37,026	Behavior category, Educational stage, Teacher gender, Course name
TBU_L	9,422	Behavior Start time, Behavior End time, Behavior category, Educational stage, Teacher gender, Course name
TBU_D	6,098	Behavior category, Behavior start time, Behavior end time, Fine-grained behavior, Spatial location, Interaction object, Description sentences (no fewer than three), Educational stage, Teacher gender, Course name

(2) Behavior Localization Annotation: Four annotators completed the annotation of start and end times and behavior categories on untrimmed videos. First, a single annotator marked the start time (S), end time (E), and behavior category for raw videos, and the annotation tool automatically calculates the cutting range [S', E'] = [S - R, E + R] (where $R \in [3, 10]$). Subsequently, three annotators independently annotate the new segments. After all annotations are complete, Fleiss' Kappa is used to evaluate the consistency of the category labels to ensure overall annotation quality. If 3/4 or more annotators select the same category, the category is confirmed. Otherwise, the segment is discarded. Next, four annotation statistics were calculated: average start time, end time, frame duration, and labels with ≥ 75 consistency, which determine the final behavior boundary. Finally, 9422 time boundary behavior detection segments were formed for TBU_L (including 8390 single behavior boundary segments and 1032 segments with 2 to 12 behaviors). These 180 datasets serve as a temporal behavior detection resource.

(3) Behavior Descriptive Annotation: Based on behavior localization annotation, behavioral clips longer than 5 s were selected for descriptive annotations. Teacher behavior was described by crowdsourced annotators from dimensions such as body posture, positional relationships, and interaction objects. The descriptive sentences do not involve the annotator's personal subjective emotions, the teacher's emotions (e.g., seriousness, happiness), and the classroom effects (e.g.,positive feedback from students). At the same time, atomic actions annotation was performed for each description clips. According to Table 3, the atomic actions dictionary contains five atomic label types across three behavioral levels. Each description includes at least five atomic action labels, with some behavioral levels containing multiple atomic labels in various contexts. Finally, a total of 6098 descriptive behavior clips were compiled.

4.5. Data statistics

(1) Comparison with other datasets: As shown in Table 5, a brief comparison of the overall statistics of existing datasets for recognition and detection is presented. Notably, TBU_R comprises random fixed-length slices and time-boundary clips, increasing the learning demands on action recognition models across diverse scenes.

(2) Instance statistics of TBU_R: Fig. 5 shows the statistics of instances across four educational stages in TBU_R. Despite the same number of original videos, there are notable variations in behavior categories

Table 5
Comparison of video datasets for action classification and localization.

Main task	Database	Category	Instance	Anno type	Avg duration	Action type
Action Classification	UCF101-24 (Soomro et al., 2012)	24	4458	Segments	5.1s	Daily events
	Kinetics 400 (Carreira and Zisserman, 2017)	400	39 596	Segments	10s	Daily events
	HMDB51 (Kuehne et al., 2011)	51	6766	Segments	3.2s	Daily events
	TBU_R	13	37 026	Clip; Segments	10s	Education
Temporal Action Localization	ActivityNet (Caba Heilbron et al., 2015)	200	30 791	Segments	49.2s	Daily events
	THUMOS14 (Idrees et al., 2017)	20	6363	Segments	24s	Sports
	TBU_L	13	9422	Segments	13s	Education

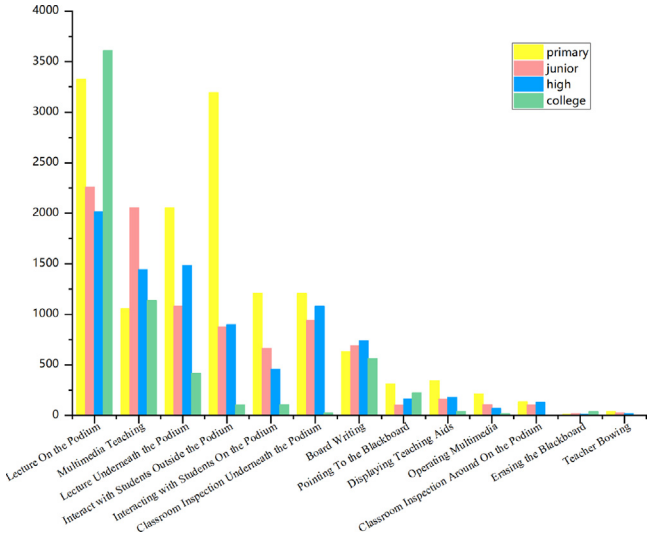


Fig. 5. The different action instances in different educational stages.

across the four educational stages. Specifically, the primary school teachers frequently utilize interaction and spatial movement to boost teaching effectiveness. In contrast, the behavioral differences between junior high school and high school teachers are similar, reflecting a shared focus on systematic knowledge transfer. However, a detailed comparison reveals that junior high school teachers utilize spatial movement more effectively to engage students compared to high school teachers. College teachers, on the other hand, rely more on multimedia, aligning with the specialized and theoretical nature of university content. At the same time, the number of behavior instances exhibits a clear long-tail distribution, such as the head category "Lecture on the podium" accounts for 29.68% of the sample, while the tail category "Teacher Bowing" accounts for only 0.23%. This distribution pattern aligns with the natural characteristics of real classroom environments, where certain teacher behaviors occur more frequently than others due to the inherent structure and dynamics of classroom activities.

(3) Duration statistics of TBU_L: Fig. 6 displays the distribution of behavior duration in TBU_L. As shown on the left side of Fig. 6, the behavioral changes in the four educational stages occurred within approximately 20 s, with primary school teachers typically exhibiting changes within 5–10 s. The right side of Fig. 6 reveals that the duration of different behaviors varies significantly, adding to the challenge of accurately locating behavior boundaries.

(4) The statistics of TBU_D: Table 6 gives a brief statistic of TBU_D and other scene description datasets. To the best of our knowledge, TBU_D is the first fine-grained descriptive dataset for teacher behavior, providing detailed records of behavior changes in three dimensions: body posture, spatial position, and action interaction objects. This ensures that descriptions within the same video are unique, complex, and diverse.

(5) Atomic action statistics of TBU_D: Based on the data annotation guidelines, it can be seen that each teacher behavior contains at least five atomic labels, and multiple atomic actions at the same hierarchy represent more detailed sequences of teaching actions. Fig. 7 displays the percentage of atomic action labels in interaction objects of different behavior categories. The interaction objects of erasing the blackboard and teacher bowing are relatively fixed, while the interaction objects of the other behaviors are variable. It is noteworthy that a single behavior instance may involve two to three eye interaction objects and up to two action interaction objects, which are critical for teacher behavior reasoning. Combined with rich behavioral description sentences, TBU_D can more accurately track the behavioral changes of teachers during the teaching process, providing robust support for understanding teacher behavior.

4.6. Data attributes

Challenges: (1) Multiple video angles. Various shots, viewpoints, and pixel values introduce uncertainty that can affect the learning of temporal features. (2) Long-tail distribution. The long-tailed distributions of both the number and duration of instances lead to greater variability in duration, behavioral amplitude, and internal coherence of similar behaviors. (3) Overlap of behavioral boundaries. Due to teachers' free movement and rapid camera switching, overlapping behavior boundaries pose a challenge for temporal action localization. (4) Rich and complex description sentences. Fine-grained behavior descriptions focus on specific aspects, such as body posture, spatial position, interaction objects, impose higher demands on behavior description models. (5) Multi-level atomic action labels. Each behavior segment contains at least five atomic behaviors that reflect the environment, fine-grained postures, interacting objects, etc. Combining atomic actions to infer teacher behavior patterns puts higher requirements on understanding teacher behavior models.

Large Scale and High Quality: Our annotation team consists of experts and graduate students in the field of education. In order to meet the needs of classroom teaching evaluation, we developed a special annotation tool to annotate video instances from three tasks: behavior classification, temporal action localization, and behavior description. We used Fleiss Kappa to verify annotation consistency, achieving values of 0.892 for fixed-length segments and 0.947 for temporal boundary clips. After several checks and calibrations over 11 months, TBU was created. To the best of our knowledge, TBU is the largest high-quality dataset in the field of teacher behavior understanding.

Diversity and Practicality: (1) Dataset selection. TBU is composed of real classroom teaching videos in 4 educational stages, with diverse shooting methods and varying classroom spaces, enhancing model generalization and robustness. (2) Data annotation. First, spatial positions are added to address the lack of spatial information in the dataset. Second, multi-level atomic action labels are designed for classroom scenarios. The teaching strategies of teachers can be revealed through the changes and combinations of these atomic actions, which is crucial for tracking teacher behavior.

4.7. Ethical issues

The TBU dataset adheres to data privacy and ethical issues related to data usage. The source video data consist of open-source classroom

Table 6
Comparative overview of relevant datasets for Behavior Description.

Main task	Database	Category	Segments	Sent-Len	Sent	Sent-Clip	Behavior type
Behavior description	TACoS (Rohrbach et al., 2012)	26	3.5k	21	11.8K	_	cooking
	Youcook II (Zhou et al., 2018)	89	15.4K	_	15.4K	_	Cooking
	MSR-VTT (Zhou et al., 2019)	257	10k	9.28	200K	20	Daily events
	MSVD (Chen and Dolan, 2011)	_	1.97K	7.1	78.8K	40	Daily events
	Ours	13	6.098K	18.2	18.294K	3	Education

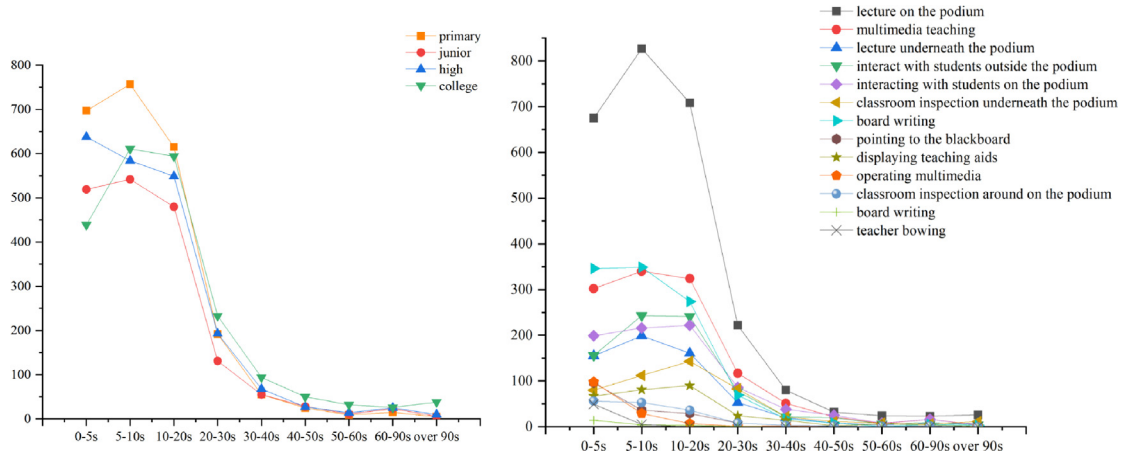


Fig. 6. Left: The Behavior duration at each educational stage in TBU_L. Right: The duration of different behaviors in TBU_L.

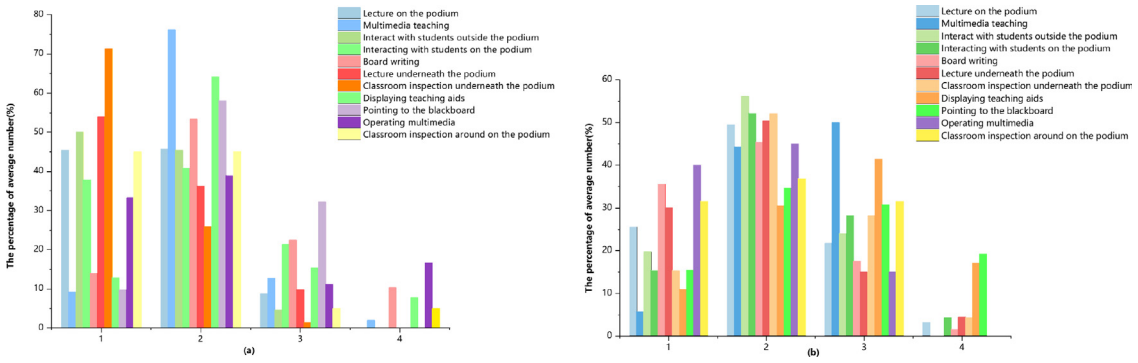


Fig. 7. Left: The semantic percentage of gesture interaction objects in TBU_D. Right: The semantic percentage of sight interaction objects in TBU_D.

teaching videos, which are collected from public websites. During the annotation process, annotators received data security training and strictly followed relevant guidelines. The researchers also received ethical training to ensure compliance with all ethical standards. The use of the data and the publication of experimental results do not involve any personal privacy of teachers and students.

5. Experiments

The practicality of TBU was evaluated from three aspects: behavior recognition, behavior detection, and behavior description. Our experiments have been extended based on the previous work (Cai et al., 2024).

5.1. Behavior recognition

Evaluation Metrics: The evaluation indicators for behavior recognition include the accuracy of a single category (Acc/top1), the accuracy averaged across all samples (Acc/mean1), and the F1-Score.

Benchmark Models: C3D (Tran et al., 2015), I3D (Carreira and Zisserman, 2017), TSN (Wang et al., 2016), Slowfast (Feichtenhofer et al., 2019), Timesformer (Bertasius et al., 2021) and BIKE (Wu et al., 2023) are selected as behavior recognition baseline models to evaluate the usability of TBU_R. 7/10 of clips are utilized for training and the others for validation. *Results and Analysis:* Due to the complex background of the TBU dataset and significant variations in the same behavior among different teachers, there are challenges such as small inter-class and large intra-class differences, alongside class imbalance. Therefore, the classification results were compared using RGB, Flow, and RGB+Flow input features. As shown in Table 7, "Pre_train" represents the pre-training operation on the corresponding dataset. From the performance of benchmarks, we can observe: (1) Using the single RGB feature, BIKE has the largest F1-Score value of 74.7%. This denotes that the spatial information in the TBU_R is crucial and cannot be ignored. (2) With a single Flow feature, TSN achieves the highest F1-Score value of 76.3%. This demonstrates that the segmented temporal information in the TBU_R dataset is highly effective for capturing time-related aspects. (3) On the two-stream input, the performance on all metrics improves with two-stream input, though the enhancement remains limited. This may be due to minimal amplitude variations or feature expression overlap among behaviors, along with semantic overlap within behavior categories. (4) On the evaluation indicators, the Acc/top1 is much higher than the Acc/mean1 and F1-Score, which indicates that category imbalance within TBU_R.

Table 7
The performance of TBU_R on different action recognition baseline models.

Model		Acc/top1	Acc/mean1	F1_score	Pre_train
C3D (Tran et al., 2015)	RGB	0.8132	0.6683	0.7104	Spores1 m
TSN (Wang et al., 2016)	RGB	0.8406	0.7093	0.7416	Imagenet
	Flow	0.851	0.721	0.763	
	RGB+Flow	0.872	0.756	0.784	
I3D (Carreira and Zisserman, 2017)	RGB	0.8171	0.6694	0.6832	Imagenet
	Flow	0.835	0.695	0.706	
	RGB+Flow	0.862	0.712	0.733	
Slowfast (Feichtenhofer et al., 2019)	RGB	0.8209	0.6687	0.6815	Kinetics400
	Flow	0.803	0.657	0.669	
	RGB+Flow	0.844	0.726	0.752	
Timesformer (Bertasius et al., 2021)	RGB	0.8288	0.7003	0.7333	Imagenet
BIKE (Wu et al., 2023)	RGB	0.8341	0.727	0.747	Kinetics400

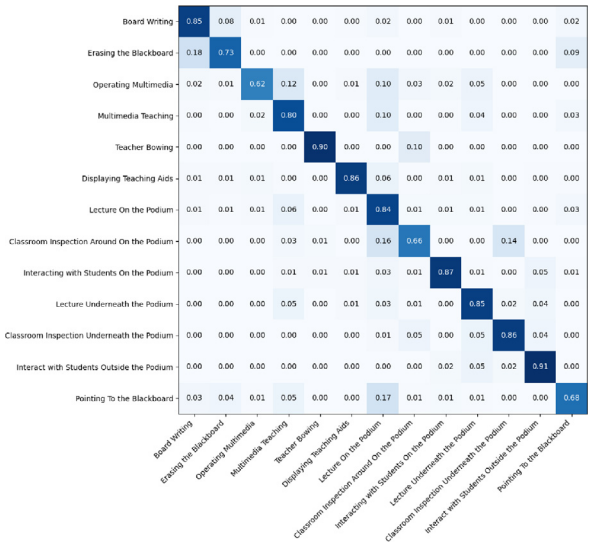


Fig. 8. Confusion matrix of BIKE on TBU_R.

As shown in Fig. 8, there are differences in recognition rates across various teacher behaviors in TBU_R dataset. On the one hand, for categories with fewer samples (e.g., teacher bowing), the recognition accuracy is also comparable to that of categories with more samples (e.g., speech on the podium). This may be due to the sufficient diversity in the dataset, which allows the model to effectively learn the behaviors of less-represented categories. On the other hand, the recognition rates are lower for "Operating Multimedia", "Classroom Inspection Around On the Podium" and "Pointing To the Blackboard". "Operating Multimedia" is often confused with "Multimedia Teaching", "Classroom Inspection Around On the Podium" and "Pointing To the Blackboard" are frequently mistaken for "Lecture On the Podium". The low recognition rate may result from subtle visual differences between these behaviors and similar misclassified actions. However, notable differences exist in the atomic actions of "Operating Multimedia", "Classroom Inspection Around On the Podium" and "Pointing To the Blackboard" compared to similar behaviors. Therefore, for behavior classification models on the TBU_R, it is necessary to account for semantic variations in fine-grained time series.

Fig. 9 shows example videos from six action categories. Affected by the teaching situation and shooting methods, the spatial context is complex and varied, and teacher's body orientations are dynamic, providing a richer and more diverse dataset for understanding teacher behavior. Therefore, considering the characteristics of TBU_R, although recognition accuracy varies among different behaviors, these variations remain within an acceptable range for this study.

5.2. Behavior detection

TBU_D is a teacher-centered behavior detection dataset that records teaching behaviors in various classroom environments. These videos have complex environments, multiple individuals, as well as scene transitions and camera switches, which pose new challenges for locating teaching behaviors. To evaluate dataset quality, initial experiments were conducted using annotations from a single annotator, followed by experiments with averaged annotations from four annotators.

Evaluation Metrics: Mean average precision (mAP) is employed to measure the performance of behavior detection.

Benchmark Models: Actionformer(AF) (Zhang et al., 2022),SSN (Zhao et al., 2017) and TriDet (Shi et al., 2023) are used as baseline models for temporal behavior detection in TBU_L. ActionFormer utilizes a Transformer network to integrate multi-scale feature representation and local self-attention, enabling simultaneous detection of action classification and action boundaries. SSN model the temporal structure of each video instance by constructing a structured temporal pyramid, effectively distinguishing action boundaries from the background, which facilitates accurate recognition and precise localization. TriDet processes information across different temporal granularities by estimating the relative probability distribution of action boundaries and performing scalable feature aggregation, thus achieving efficient temporal action detection. 5/10 of clips are utilized for training and the others for validation. We compared the mAP@IoU of the baseline models with different inputs, such as RGB, Flow, and RGB+Flow, with IoU thresholds set to 0.3, 0.4, 0.5, 0.6, and 0.7.

Results and Analysis: As shown in Table 8, the evaluation of TBU_L can be approached from two aspects. First is the performance comparison of the models. (1) For SSN, the performance of RGB is better than that of Flow where the IoU is set to less than or equal to 0.6. This is likely because SSN, as a TSN architecture, effectively extracts motion information from video segments. (2) For ActionFormer, the performance of Flow is superior to RGB across all IoU thresholds. The reason probably is that ActionFormer is more adept at modeling long-distance temporal dependencies. (3) For TriDet, its performance is consistently the best across all input modalities. This is primarily due to TriDet's use of relative probability prediction and multi-scale feature aggregation, which allows it to better handle blurred boundaries and complex spatiotemporal relationships. (4) SSN vs. AF, in the case of a low IoU threshold (0.3, 0.4, 0.5), ActionFormer performs better than SSN with Flow input, but worse with both RGB and RGB+Flow inputs. The comparison results highlight the critical role of spatially relevant motion information and temporal motion information in TBU_L. (5) TriDet vs. SSN, TriDet outperforms SSN across most input modalities, particularly with Flow input. The key advantage of TriDet lies in its relative probability prediction mechanism, implemented through the Trident-head, which enables more accurate estimation of action boundaries. In contrast, SSN's proposal generation method distinguishes between action and background boundaries, which leads to less precise boundary localization compared to TriDet's approach. The second aspect is the method of data formation. Table 8 shows that data annotated by a single annotator performs significantly worse than data averaged from multiple annotators, indicating that using multiple annotators enhances the reliability and utility of TBU_L. Due to the lower performance of single annotation data, no single annotator data experiment was conducted on the TriDet model.

In the TriDet model, a comparative experiment on behavior detection based on spatial location was carried out. The original spatially labeled "Interacting with Students Outside the Podium" and "Interacting with Students On the Podium" were merged into a non-spatially labeled category "Interacting with Students". As shown in Fig. 10,

Table 8
The mAP performance of the detection dataset on SSN, Actionformer and TriDet.

Method	Modality	Single annotator data					Average data from four annotators				
		mAP@0.3	mAP@0.4	mAP@0.5	mAP@0.6	mAP@0.7	mAP@0.3	mAP@0.4	mAP@0.5	mAP@0.6	mAP@0.7
SSN (Zhao et al., 2017)	RGB	8.39	7.27	6.85	4.59	3.01	65.54	60.87	55.34	48.29	34.24
	Flow	6.38	5.89	5.01	4.21	3.69	67.46	59.68	53.21	39.18	30.51
	RGB+Flow	10.29	8.63	8.23	5.68	4.23	70.61	63.57	57.54	45.37	33.98
ActionFormer (Zhang et al., 2022)	RGB	5.81	5.45	4.75	4.06	3.12	63.51	58.43	49.72	37.82	30.89
	Flow	6.23	6.59	5.36	5.28	4.36	65.28	59.12	51.34	39.67	31.07
	RGB+Flow	7.56	7.25	6.69	6.58	5.2	68.97	61.59	53.7	42.58	33.51
TriDet (Shi et al., 2023)	RGB	–	–	–	–	–	65.49	61.13	52.27	42.57	34.58
	Flow	–	–	–	–	–	68.25	62.39	54.36	43.83	35.26
	RGB+Flow	–	–	–	–	–	70.83	64.25	58.67	46.17	36.56

Fig. 9. Example videos of six behaviors categories in TBU_R.

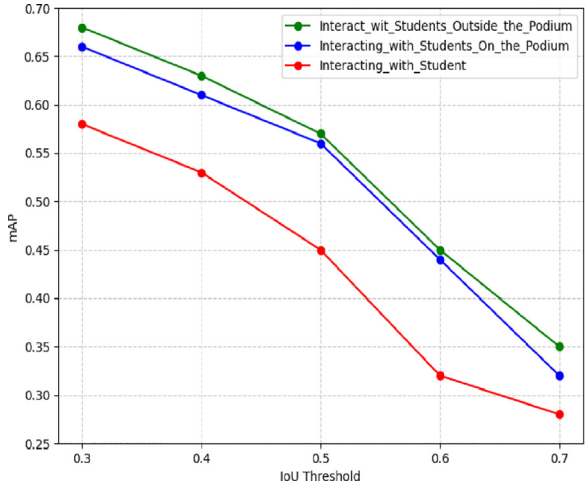


Fig. 10. The performance of behavior localization after merging spatial locations on TriDet.

the detection performance comparison across three behavioral categories revealed that: (1) The localization accuracy of the coarse-grained category without spatial consideration was significantly lower than the two fine-grained subcategories; (2) In fine-grained classification, the detection performance of "Interacting with Students Outside the Podium" surpassed that of "Interacting with Students On the Podium". The experimental results indicate that when teachers interact with students below the podium, their body movements exhibit greater amplitude (e.g., bending to instruct or moving around), resulting in more distinctive behavioral representations. Conversely, podium-based interactions demonstrate relatively constrained postures, leading to reduced feature distinctiveness. This conclusion validates the critical role of spatial positioning information in teacher behavior classification — distinct spatial zones correspond to unique motion patterns, and fine-grained spatial partitioning can provide a more discriminative feature basis for classroom behavior analysis.

The analysis reveals that prominent motion expression and dependency on the relative positioning of classroom objects are distinctive features of TBU_L, distinguishing it from datasets with strictly temporal action sequences. Specifically, the interpretation of teacher behaviors relies more on the spatial positioning of objects rather than on strict temporal action sequence. Additionally, the use of averaged boundary annotations from multiple individuals confirms the validity of TBU_L.

5.3. Behavior description

Evaluation Metrics: We employ standard evaluation metrics including BLEU-4, METEOR, ROUGE-L, and CIDEr for behavior description generation.

Benchmark Models: We utilize RecNet (Wang et al., 2018), SGN (Ryu et al., 2021) and CoCap (Shen et al., 2023) as baseline models for behavior description. RecNet network is a common video subtitle generation model that utilized temporal attention mechanisms and reconstructions to improve the performance of video subtitle generation. SGN network achieves a comprehensive understanding of the subtitle context by encoding the video as a semantic group consisting of partially decoded subtitle phrases and related frames. CoCap learns spatial, motion, and contextual features of the entire video from a video compression perspective. For training, 7/10 of clips are utilized, with the remaining clips set aside for validation. The MSVD behavior description dataset, consisting of 1970 video clips averaging 10 s each, is used for comparison.

Results and Analysis: As depicted in Table 9, CoCap achieves the best results, possibly due to its consideration of contextual and action features throughout the entire video. This is consistent with the attributes of TBU_D, which describes posture changes, spatial motion, and interactive objects within behavioral sequences. Across all evaluation metrics, TBU_D performs lower than MSVD, which is due to the fact

Table 9
The performance of the description dataset on RecNet, SGN, and CoCap.

Method	MSVD dataset				TBU_D dataset			
	BLEU-4	METEOR	ROUGE-L	CIDEr(%)	BLEU-4	METEOR	ROUGE-L	CIDEr(%)
RecNet (Wang et al., 2018)	51.1	34.0	69.4	79.7	21.5	18.1	36.6	32.2
SGN (Ryu et al., 2021)	51.7	35.5	72.2	93.4	23.4	20.5	40.9	35.4
CoCap (Shen et al., 2023)	59.8	40.6	77.8	111.5	25.3	23.8	46.3	38.7

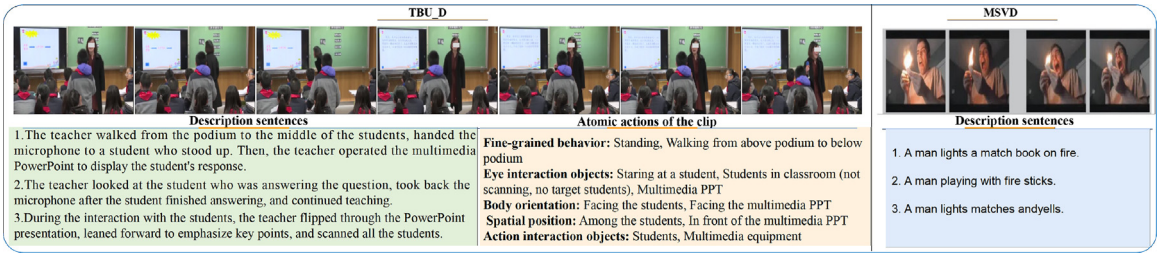


Fig. 11. An example of a video clip from TBU_D and MSVD.

that the description sentences of TBU_D are more detailed than those of MSVD.

Fig. 11 shows an example of the TBU_D and MSVD. It can be observed that TBU_D features a more complex background with more interactive objects. TBU_D goes beyond describing single behaviors, instead capturing atomic action combinations of teacher location, fine-grained postures, and interactive objects over time. In contrast, MSVD is visually simpler, with more straightforward description sentences. Therefore, behavior description models on TBU_D must closely track context, spatial changes, and fine-grained features. Additionally, the merging and inference of atomic actions in TBU_D enhance the understanding of teacher behavior.

6. Conclusion

Reflective practice based on teaching videos has brought significant benefits to teachers' professional development (Hamel and Viau-Guay, 2019). The integration of AI and education meets the need for process-oriented, fine-grained, and multi-dimensional analysis of teacher behavior. However, due to external classroom factors and internal teacher factors, classroom behaviors vary in category, timing, and performance granularity. Therefore, constructing a dataset that meets these needs using existing teaching video data is the first step in applying AI to teacher behavior analysis.

This article provides detailed evidence that a multi-task teacher behavior dataset that considers spatial location can effectively facilitate research and applications of AI in teaching. With well-designed behavior classification and atomic actions, the process of teacher's teaching behavior can be objectively recorded. Then, a multi-task framework for annotating classroom teaching videos is proposed, aiming to provide a comprehensive understanding of teacher behavior. Based on this, a multi-task comprehensive video dataset, termed TBU, is introduced for understanding classroom teachers' teaching behaviors. TBU comprises three distinct subsets: a behavior recognition dataset (TBU_R), a behavior localization dataset (TBU_L), and a behavior description dataset (TBU_D). Compared with other datasets in similar scenarios, TBU has the following key features: (1) Consideration of teachers' spatial locations in the classroom. (2) It provides precise time boundaries for teacher behavior in complex scenarios. (3) It includes fine-grained behavior descriptions and multi-level atomic action phrases. (4) It offers highly diverse biased datasets covering four educational stages, multiple types of courses, different genders, various regions and diverse shooting methods.

Additionally, differences in behavior categories, behavior duration, and atomic actions were analyzed within TBU. Primary school teachers exhibit a higher number of spatial movements, more frequent interactions with students, rapidly changing behavior categories over time, and more diverse atomic actions. In contrast, college teachers more commonly engage in lecturing at the podium and multimedia teaching, which tend to last longer. Furthermore, several commonly used models for behavior recognition, behavior detection, and behavior description were empirically evaluated on TBU_R, TBU_L, and TBU_D, respectively. The experiments demonstrate the practicability and effectiveness of TBU. Specifically, in the behavior recognition experiment, it was found that only the features of the RGB modality achieved higher classification results. This may be due to the influence of complex classroom scenarios on the interpretation of teacher behavior. Meanwhile, the recognition experiments on TBU_R also reveal similarities within behavioral categories and differences between behavioral categories. In the behavior detection experiment, it was found that the Temporal Segmentation Network and the Boundary Probability Prediction Network were more conducive to TBU_L. This is due to the fact that the teacher's behavior is highly dynamic, often accompanied by behavioral concurrency or time overlap. It further illustrates that the pattern of teacher behavior is not strictly preset, but is affected by student feedback and the distribution of teaching resources in real time, showing flexible adaptive characteristics. In the video captioning experiment, it was shown that the current description models performed poorly due to the fine granularity and obvious procedural nature of TBU_D behavioral statements.

6.1. Future work

For future research, we believe that research can be carried out from two aspects. First, behavior understanding based on the dataset. We believe behavior understanding based on the spatial location, multi-shot temporal action localization, fine-grained behavior description, and teacher tracking in complex classroom environments can improve the performance of TBU. A feasible approach is to separate the background from the motion foreground or use a visual language large model pre-training method to improve teacher behavior classification and detection. For behavior description, a viable alternative is to embed domain expert knowledge into the description model to enhance the procedural and fine-grained aspects of the generated descriptive sentences. In addition to behavioral classification, detection, and description, there are also some other potentials applications of TBU. For example, each video instance of TBU_D contains multi-level, fine-grained atomic action labels, a promising approach is to generate multiple behavioral description statements for specific atomic actions through a joint model of prompt learning and generative models to enhance the interpretability of teacher behavior comprehension. This

is because description of teaching activities or teaching practices has an indirect impact on reflective practice, which can help to better understand and contextualize the teaching process and provide effective explanations for reflective activities (Almodaires, 2009).

Second, alleviate the limitations of the dataset. To mitigate the long-tail distribution in the TBU dataset, we suggest introducing a dynamic sampling framework and an automatic category discovery method based on self-supervision to improve the recognition of low-frequency behaviors. Concurrently, we plan to introduce multiple teacher–student interaction dimensions, such as emotional engagement, dialogic interaction, and collaborative learning, to expand the annotation schema and more comprehensively capture the complexity of classroom interactions. Despite certain limitations, the TBU dataset offers significant advantages in spatial, fine-grained, and multi-tasking capabilities, enabling research on teacher–student interaction patterns across educational stages, the impact of classroom spatial location on interaction effects, and location-driven cross-stage personalized teaching.

The formation and research of TBU have been conducted in compliance with relevant data privacy and ethical guidelines. Future research will be conducted in the same manner. We hope that TBU will serve as basic data for understanding teacher behavior in actual classroom teaching scenarios to advance AI applications in smart education.

CRediT authorship contribution statement

Ting Cai: Writing – original draft, Validation, Software, Data curation. **Yu Xiong:** Writing – review & editing, Project administration, Methodology. **Chengyang He:** Validation, Data curation. **Chao Wu:** Formal analysis, Data curation. **Linqin Cai:** Validation, Software.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by National Natural Science Foundation of China “Research on Intelligent Recognition and Explainable Evaluation on Teachers’ Classroom Teaching Engagement” (No. 62377007), Chongqing Key Research Project for Higher Education Teaching Reform “Research and Exploration on Intelligent Evaluation of Students’ Comprehensive Quality under the Background of Digital Education Transformation” (No. 232073), and Chongqing Municipal Education Commission Science and Technology Research Youth Project “Research on video representation learning based on context perception and its application in understanding teachers’ teaching behavior” (No. KJQN202400634).

Data availability

https://github.com/cai-KU/TBU.

References

Ahmad, K., Iqbal, W., El-Hassan, A., Qadir, J., Benhaddou, D., Ayyash, M., Al-Fuqaha, A., 2024. Data-driven artificial intelligence in education: A comprehensive review. IEEE Trans. Learn. Technol. 17, 12–31.

Allas, R., Leijen, Ä., Toom, A., 2020. Guided reflection procedure as a method to facilitate student teachers’ perception of their teaching to support the construction of practical knowledge. Teach. Teach. 26, 166–192.

Almodaires, A.A., 2009. Technology-supported reflection : towards bridging the gap between theory and practice in teacher education.

Berliner, D.C., 2004. Describing the behavior and documenting the accomplishments of expert teachers. Bull. Sci. Technol. Soc. 24 (3), 200–212.

Bertasius, G., Wang, H., Torresani, L., 2021. Is space-time attention all you need for video understanding? In: ICML, vol. 2, p. 4.

Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J., 2015. Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition. pp. 961–970.

Cai, T., Xiong, Y., He, C., Wu, C., Zhou, S., 2024. TBU: A large-scale multi-mask video dataset for teacher behavior understanding. In: 2024 IEEE International Conference on Multimedia and Expo. ICME, IEEE, pp. 1–6.

Cantrell, S., Kane, T., 2013. Ensuring fair and reliable measures of effective teaching.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 4724–4733.

Chen, G., 2020. A visual learning analytics (VLA) approach to video-based teacher professional development: Impact on teachers’ beliefs, self-efficacy, and classroom talk practice. Comput. Educ. 144, 103670.

Chen, W., Chen, L., Li, J., Jiang, L.L., Xue, Z., 2022. Nonverbal behavior analysis for instructors with deep semantic segmentation and pose classification. SSRN Electron. J.

Chen, D., Dolan, W.B., 2011. Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 190–200.

Clark, C.M., 1986. Ten years of conceptual development in research on teacher thinking. Adv. Res. Teach. Think. 7–20.

Dahleez, K.A., El-Saleh, A.A., Al Alawi, A.M., Abdelmuniem Abdelfattah, F., 2021. Higher education student engagement in times of pandemic: the role of e-learning system usability and teacher behavior. Int. J. Educ. Manag. 35 (6), 1312–1329.

Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6202–6211.

Fernandez-Nieto, G., An, P., Zhao, J., Shum, B., et al., 2022. Classroom dandelions: Visualising participant position, trajectory and body orientation augments teachers’ sensemaking. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–17.

Gang, Z., Wenjuan, Z., Biling, H., Jie, C., Hui, H., Qing, X., 2021. A simple teacher behavior recognition method for massive teaching videos based on teacher set. Appl. Intell. 51, 8828–8849.

Gaudin, C., Chaliés, S., 2015. Video viewing in teacher education and professional development: A literature review. Educ. Res. Rev. 16, 41–67.

Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., et al., 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6047–6056.

Hamel, C., Viau-Guay, A., 2019. Using video to support teachers’ reflective practice: A literature review. Cogent Educ. 6 (1), 1673689.

Hollingsworth, H., Clarke, D., 2017. Video as a tool for focusing teacher self-reflection: supporting and provoking teacher learning. J. Math. Teach. Educ. 20, 457–475.

Idrees, H., Zamir, A.R., Jiang, Y.G., Gorban, A., Laptev, I., Sukthankar, R., Shah, M., 2017. The thumos challenge on action recognition for videos “in the wild”. Comput. Vis. Image Underst. 155, 1–23.

Jia, J., Song, J., Hu, Q., Tang, S., Xu, S., 2023. TAR: A dataset of teacher-teaching action recognition. In: 2023 8th International Conference on Image, Vision and Computing. ICIVC, IEEE, pp. 676–681.

Karumbaiah, S., Borchers, C., Shou, T., Falhs, et al., 2023. A spatiotemporal analysis of teacher practices in supporting student learning and engagement in an AI-enabled classroom. In: International Conference on Artificial Intelligence in Education. Springer, pp. 450–462.

Korthagen, F.A., 2010. Situated learning theory and the pedagogy of teacher education: Towards an integrative view of teacher behavior and teacher learning. Teach. Teach. Educ. 26 (1), 98–106.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., 2011. HMDB: a large video database for human motion recognition. In: 2011 International Conference on Computer Vision. IEEE, pp. 2556–2563.

Van der Lans, R.M., Van de Grift, W.J., van Veen, K., 2018. Developing an instrument for teacher feedback: using the rasch model to explore teachers’ development of effective teaching strategies and behaviors. J. Exp. Educ. 86 (2), 247–264.

Li, Y., Deng, Z., Liu, M., He, S., Wang, Y., Jiang, W., 2022. A method for analyzing teacher behavior in classroom based on the long-and short-term features of pose sequences. In: 2022 9th International Conference on Digital Home. ICDH, IEEE, pp. 233–238.

Li, X., Yan, L., Zhao, L., Martinez-Maldonado, R., Gasevic, D., 2023. CVPE: A computer vision approach for scalable and privacy-preserving socio-spatial, multimodal learning analytics. In: LAK23: 13th International Learning Analytics and Knowledge Conference. pp. 175–185.

Lim, F.V., O’Halloran, K.L., Podlasov, A., 2012. Spatial pedagogy: mapping meanings in the use of classroom space. Camb. J. Educ. 42, 235–251.

Liu, Y., Wang, L., Ma, X., Wang, Y., Qiao, Y., 2021. FineAction: A fine-grained video dataset for temporal action localization. IEEE Trans. Image Process. 31, 6937–6950.

Liu, Q., Zheng, X., Liu, Y., Wu, L., et al., 2024. Exploration of the characteristics of teachers’ multimodal behaviours in problem-oriented teaching activities with different response levels. Br. J. Educ. Technol. 55 (1), 181–207.

MacKay, A., Peterson, P.L., Walberg, H.J., 1980. Research on teaching: Concepts, findings, and implications. Can. J. Educ. 5, 114.

Martinez-Maldonado, R., Echeverria, V., Mangaroska, K., Shibani, et al., 2022. Moodoo the tracker: Spatial classroom analytics for characterising teachers' pedagogical approaches. Int. J. Artif. Intell. Educ. 1–27.

Morell, T., 2018. Multimodal competence and effective interactive lecturing. System 77, 70–79.

Nida, N., Yousaf, M.H., Irtaza, A., Velastin, S.A., 2019. Instructor activity recognition through deep spatiotemporal features and feedforward extreme learning machines. Math. Probl. Eng. 2019 (1), 2474865.

Osterman, K.F., 2023. Teacher practice and students' sense of belonging. In: Second International Research Handbook on Values Education and Student Wellbeing. Springer, pp. 971–993.

Pang, S., Zhang, A., Lai, S., Yang, Y., 2022. Automatic recognition of teachers' nonverbal behaviors based on graph convolution neural network. In: Proceedings of the 14th International Conference on Education Technology and Computers. pp. 429–435.

Rafique, M.A., Khaskheli, F., Hassan, M.T., Naseer, S., Jeon, M., 2022. Employing automatic content recognition for teaching methodology analysis in classroom videos. PLoS One 17 (2), e0263448.

Rich, P.J., Hannafin, M., 2009. Video annotation tools: Technologies to scaffold, structure, and transform teacher reflection. J. Teach. Educ. 60 (1), 52–67.

Rodgers, W.J., Kennedy, M.J., VanUitert, V.J., Myers, A., 2019. Delivering performance feedback to teachers using technology-based observation and coaching tools. Interv. Sch. Clin. 55, 103–112.

Rohrbach, M., Regneri, M., Andriluka, M., Amin, S., et al., 2012. Script data for attribute-based recognition of composite activities. In: 12th European Conference on Computer Vision. ECCV, Springer, pp. 144–157.

Ryu, H., Kang, S., Kang, H., Yoo, C.D., 2021. Semantic grouping network for video captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2514–2522.

Sharma, V., Gupta, M., Kumar, A., Mishra, D., 2021. EduNet: a new video dataset for understanding human activity in the classroom environment. Sensors 21 (17), 5699.

Sharpe, T., Lounsbery, M., Bahls, V., 1997. Description and effects of sequential behavior practice in teacher education. Res. Q. Exerc. Sport 68 (3), 222–232.

Shen, Y., Gu, X., Xu, K., Fan, H., Wen, L., Zhang, L., 2023. Accurate and fast compressed video captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15558–15567.

Shi, Y., Liu, J., Zhang, Y., Zhao, X., Yan, C., 2016. Teaching behavior evaluation model establishment and feature extraction. In: 2016 Chinese Control and Decision Conference. CCDC, IEEE, pp. 2828–2833.

Shi, D., Zhong, Y., Cao, Q., Ma, L., Lit, J., Tao, D., 2023. TriDet: Temporal action detection with relative boundary modeling. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 18857–18866.

Soomro, K., Zamir, A.R., Shah, M., 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. In: Computer Vision and Pattern Recognition. pp. 1–6.

Sun, B., Wu, Y., Zhao, K., He, J., Yu, L., Yan, H., Luo, A., 2021. Student Class Behavior Dataset: a video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. Neural Comput. Appl. 33, 8335–8354.

Sun, B., Zhao, K., Xiao, Y., He, J., Yu, L., Wu, Y., Yan, H., 2019. BNU-LCSAD: a video database for classroom student action recognition. In: Optoelectronic Imaging and Multimedia Technology VI. vol. 11187, SPIE, pp. 417–424.

Tammets, K., Khulbe, M., Sillat, L.H., Ley, T., 2022. A digital learning ecosystem to scaffold teachers' learning. IEEE Trans. Learn. Technol. 15, 620–633.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4489–4497.

Wang, M., Long, T., Chen, Z., Wu, X., Shi, Y., Xu, L., 2024. Investigating the interaction types and instructional proxemics in information technology enhanced exemplary lessons. Asia-Pac. Educ. Res. 33 (1), 129–141.

Wang, B., Ma, L., Zhang, W., Liu, W., 2018. Reconstruction network for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7622–7631.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L., 2016. Temporal segment networks: Towards good practices for deep action recognition. In: European Conference on Computer Vision. Springer, pp. 20–36.

Warren, L.L., 2019. Behaviors of teacher leaders in the classroom. Psychol. Behav. Sci. 12 (1), 104–108.

Wu, D., Chen, J., Deng, W., Wei, Y., et al., 2020. The recognition of teacher behavior based on multimodal information fusion. Math. Probl. Eng. 2020 (1), 8269683.

Wu, W., Wang, X., Luo, H., Wang, J., Yang, Y., Ouyang, W., 2023. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6620–6630.

Wu, D., Wang, J., Zou, W., Zou, S., Zhou, J., Gan, J., 2024. Classroom teacher action recognition based on spatio-temporal dual-branch feature fusion. Comput. Vis. Image Underst. 247, 104068.

Zhan, Z., Wu, Q., Lin, Z., Cai, J., 2021. Smart classroom environments affect teacher-student interaction: Evidence from a behavioural sequence analysis. Australas. J. Educ. Technol. 37 (2), 96–109.

Zhang, C.L., Wu, J., Li, Y., 2022. Actionformer: Localizing moments of actions with transformers. In: European Conference on Computer Vision. Springer, pp. 492–510.

Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D., 2017. Temporal action detection with structured segment networks. Int. J. Comput. Vis. 128, 74–95.

Zhao, J., Xu, M., Wang, X., 2024. A novel dataset based on indoor teacher-student interactive mode using AIoT. Internet Things 25, 101044.

Zhou, L., Kalantidis, Y., Chen, X., Corso, J.J., Rohrbach, M., 2019. Grounded video description. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6578–6587.

Zhou, L., Xu, C., Corso, J.J., 2018. Towards automatic learning of procedures from web instructional videos. In: AAAI Conference on Artificial Intelligence. pp. 7590–7598.