TS-Interact: A Multimodal Dataset for Characterizing Classroom Teacher- and Student-Led Instructional Behaviors

Weigang Lu
Department of Education
Ocean University of China
luweigang@ouc.edu.cn

Yaoxuan Peng
Department of Education
Ocean University of China
pengyx@stu.ouc.edu.cn

Saisai Ye
Department of Education
Ocean University of China
yesaisai@stu.ouc.edu.cn

Cunling Bian*
Department of Education
Ocean University of China
clbian@ouc.edu.cn

ABSTRACT

Classroom instructional behavior analysis seeks to model teacher- and student-led behaviors to support instructional quality assessment, teaching feedback, and AI-assisted classroom analytics. With the rapid progress of Large Language Models (LLMs) and multimodal learning, reliable analysis increasingly depends on well-structured multimodal datasets with pedagogically grounded behavior definitions. However, existing classroom datasets are often limited to single modalities or coarse, task-agnostic behavior categories, failing to capture fine-grained teacher–student instructional dynamics and limiting model interpretability and generalizability. To address these issues, we introduce TS-Interact (Teacher–Student Interaction Dataset), a multimodal dataset for characterizing classroom teacher- and student-led instructional behaviors. TS-Interact integrates synchronized video, audio, and text from real classroom scenarios, comprising approximately 1,000 instructional segments across multiple core subjects. Built upon an S–T interaction analysis framework, instructional behaviors are systematically decomposed into teacher-led and student-led categories, each further organized into four pedagogically motivated behavior types. All segments are manually annotated, with annotation reliability validated using the Kappa coefficient. Benchmark experiments with representative multimodal behavior recognition models demonstrate that TS-Interact achieves higher annotation consistency and improved recognition accuracy than commonly used generic classroom datasets. Overall, TS-Interact provides a high-quality multimodal benchmark with theoretically grounded behavior definitions, enabling precise and interpretable LLM-based classroom behavior analysis and supporting future research in intelligent teaching evaluation.

Keywords

Classroom Instructional Behavior Analysis, Multimodal Educational Dataset, Teacher–Student Interaction, Pedagogically Grounded Annotation, LLM-based Classroom Analytics

1. INTRODUCTION

Classroom teaching behavior analysis, which involves the quantitative modeling and qualitative characterization of teacher-initiated and student-initiated behaviors during instructional activities, serves as a cornerstone for advancing modern educational practices[11]. Its application value spans three core domains: in teaching evaluation, it provides objective, data-driven evidence to complement subjective observation scales; in teaching feedback, it enables targeted improvement suggestions for teachers by identi-fying interaction bottlenecks in classroom processes; in intelligent education systems, it underpins the development of adaptive teaching modules and automated classroom monitoring tools[12]. With the rapid advancement of large language models (LLMs) and multimodal learning technologies, the integration of cross-modal information (video, audio, text) has revolutionized the way class-room behaviors are analyzed—LLMs, in particular, empower the mining of implicit semantic relationships behind behavioral sig-nals, while multimodal fusion enhances the comprehensiveness of behavior recognition[10]. This technological evolution has put forward new demands for classroom analysis research: the need for high-quality multimodal datasets with pedagogically grounded behavior definitions, which can bridge the gap between technical model training and practical educational scenarios.

However, a critical review of existing classroom datasets reveals three prominent limitations that hinder the advancement of LLM-driven multimodal classroom behavior analysis. First, in terms of modality coverage, most datasets prioritize single-modal signals (e.g., learning management system logs or video-only recordings), ignoring the comple-

*Corresponding author

Weigang Lu, Yaoxuan Peng, Saisai Ye, and Cunling Bian. TS-Interact: A Multimodal Dataset for Characterizing Classroom Teacher- and Student-Led Instructional Behaviors. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 549–555. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21039742

mentary value of synchronized audio and text information in capturing complex interaction dynamics such as teacher-student dialogue and experimental operation sounds[7]. Second, behavior definition frameworks are often coarse-grained and task-irrelevant. Many datasets categorize classroom behaviors into overly general types (e.g., "interaction" and "non-interaction"), failing to distinguish subject-specific behavioral characteristics such as science experiment operations and art creation practices[13]. Third, there is a lack of solid pedagogical theoretical support for behavior classification systems. Existing datasets rarely integrate classic educational theories such as student-teacher (S-T) interaction analysis, resulting in behavior taxonomies that are disconnected from teaching practice and reducing the interpretability of downstream analysis models[9]. These limitations collectively restrict the performance of multimodal and LLM-based analysis models in real classroom scenarios, leading to a mismatch between technical solutions and educational application needs.

To address the aforementioned gaps, this study introduces TS-Interact, a multimodal teacher-student interaction dataset dedicated to classroom teaching behavior analysis. The core contributions of this study are threefold. First, we construct TS-Interact, a high-quality multimodal dataset comprising approximately 1,000 teaching segments across seven core middle school subjects, which integrates synchronized video, audio, and text data collected from authentic classroom environments. Second, based on the classic S-T interaction analysis framework, we design a dual-dominant teaching behavior classification system, which divides classroom behaviors into teacher-dominant and student-dominant categories, with each category further refined into four pedagogically moti-vated subtypes, ensuring both fine-grained modeling and semantic interpretability. Third, we conduct comprehensive experiments to verify the dataset's effectiveness: we quantify annotation consistency using Kappa coefficients to ensure data quality, and per-form benchmark tests with representative multimodal behavior recognition models to demonstrate that TS-Interact outperforms general-purpose datasets in behavior recognition accuracy.

2. RELATED WORK

Classroom Instructional Behavior Analysis. Classroom instructional behavior analysis is a central topic in educational technology, aiming to model and interpret the dynamic interactions between teachers and students to support instructional evaluation and intelligent educational systems. Existing studies largely rely on single-modal data sources—video, audio, or text—enabled by advances in computer vision, speech processing, and natural language processing. Video-based approaches leverage visual cues such as posture, gestures, and movement patterns to recognize classroom behaviors[6]. While automatic recognition models (e.g., CNNs and pose-based methods) have reduced manual annotation costs, most studies adopt coarse-grained behavior categories and often fail to distinguish teacher and student roles, limiting interpretability of interaction dynamics. Audio-based methods focus on speech features and dialogue structures to analyze verbal interaction and emotional states[4]. However, these approaches are highly sensitive to noise and typically rely on shallow behavior definitions (e.g., speaking vs. silence), overlooking pedagogically meaningful interaction types. Text-based studies exploit transcribed classroom discourse and learning logs to analyze teaching content and participation quality [3]. Although semantic modeling offers interpretability advantages, such approaches are constrained by limited data scale and the lack of integration with visual and auditory cues. Overall, prior work exhibits three common limitations: coarse behavioral granularity, ambiguous subject distinction between teachers and students, and insufficient multimodal integration. These issues restrict the applicability of existing models for fine-grained, theory-aligned instructional behavior analysis.

Classroom Instructional Behavior Datasets. Several datasets have been proposed to support classroom behavior analysis; however, notable limitations remain in modality coverage, annotation granularity, and theoretical grounding. Existing datasets either focus on engagement estimation or general instructional actions, with limited representation of authentic classroom interaction dynamics[5]. Most available datasets employ single or weakly integrated modalities[8], preventing effective modeling of cross-modal behavioral cues. In addition, behavior labels are typically broad (e.g., "engaged" or "active"), lacking alignment with pedagogical constructs and obscuring distinctions between different instructional intents. More critically, many datasets are not grounded in established educational theories, resulting in annotations that reflect surface-level observations rather than interpretable instructional behaviors[1]. These limitations highlight the need for a multimodal classroom dataset that integrates synchronized video, audio, and text, adopts fine-grained and role-aware behavior definitions, and is explicitly informed by instructional theory. Addressing these gaps forms the motivation for the TS-Interact dataset proposed in this work.

3. METHODOLOGY

To provide a clear and systematic understanding of how TS-Interact is designed, constructed, and validated, this section presents the methodological framework of the proposed dataset from theory to practice. As illustrated in Fig.1, our method follows a structured pipeline that starts from a pedagogically grounded instructional behavior modeling framework, proceeds through multimodal data collection, preprocessing, and rigorous annotation, and culminates in dataset validation via representative multimodal baseline models. This end-to-end design ensures that the dataset is not only empirically reliable but also theoretically interpretable, enabling fine-grained characterization of teacher–student instructional behaviors. In the following subsections, we first introduce the theoretical foundation underlying the instructional behavior taxonomy, then describe the dataset construction and annotation process in detail, and finally report validation experiments that demonstrate the effectiveness and adaptability of TS-Interact for multimodal classroom behavior analysis

3.1 Theoretical Foundation of Instructional Behavior Modeling

TS-Interact is built based on the Teacher-Student (S-T) interaction framework, which originates from the classic classroom interaction theory [2] and has been modernized in recent multimodal educational research. The framework

Table 1: Classroom Teaching Behavior Categories

Major Category	Subcategory	Category
Teacher-led Teaching Behavior	Knowledge Presentation	Explain and state; Summarize the lecture; Evaluation summary; Demonstrate and explain
	Guiding Questions	Raise a question; Assign tasks; Follow up with a counter-question
	Resource Display	Physical display; Experimental demonstration; Text display; Picture display; Data display; Blackboard writing display
	Process Management	Activity guidance; Communication guidance; Order management; Emotional motivation; Performance evaluation
Student-led Teaching Behavior	Independent Exploration	Independent thinking; Information query
	Oral Expression	Student read aloud; Speak and answer together; Student's direct answers
	Operational Practice	Start writing by hand; Hands-on operation
	Cooperation and Exchange	Group discussion; Report conversation; Work display

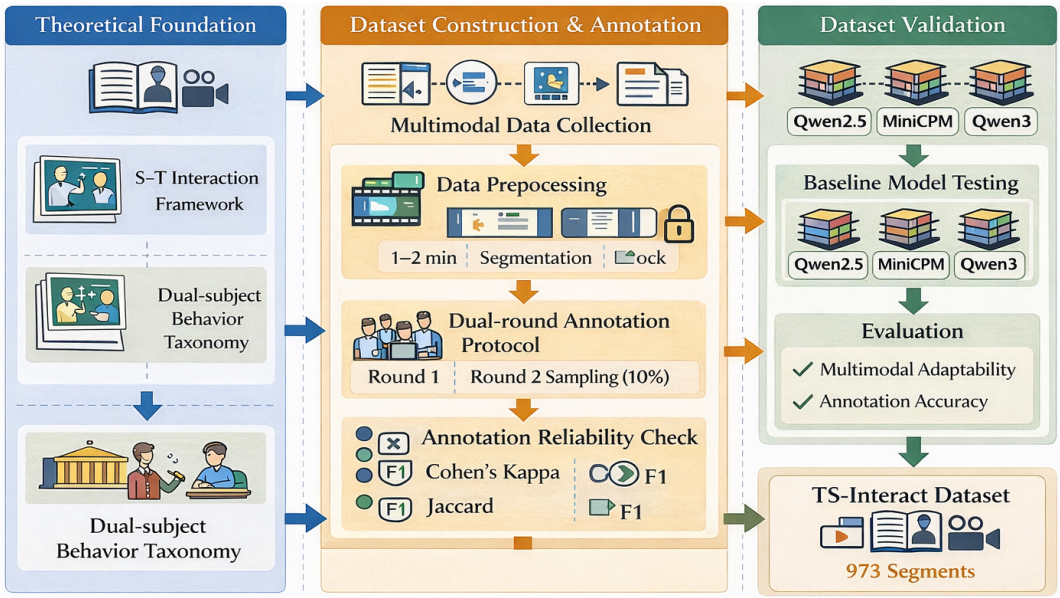

Figure 1: The workflow for constructing the dataset of T-S interaction

clearly defines teacher-led and student-led teaching behaviors, providing a rigorous theoretical foundation for multimodal observation and teaching semantic alignment. Both the behavior classification system and the three-layer annotation scheme in this paper are strictly designed based on this theory. The instructional behavior annotation system constitutes the theoretical foundation of the TS-Interact dataset, enabling structured and interpretable modeling of classroom teaching and learning processes. Existing classroom behavior annotation schemes often adopt generalized, subject-agnostic categories, which fail to capture the core logic of teacher–student interaction and limit their applicability to multimodal analysis. To address this issue, TS-Interact grounds its annotation design in an interaction-oriented pedagogical framework, ensuring both theoretical validity and annotation feasibility.

Student–Teacher Interaction Framework. TS-Interact adopts the Student–Teacher (S–T) interaction analysis framework as its theoretical basis, viewing classroom instruction as a dynamic process formed through continuous interaction between teachers and students. In this framework, teachers primarily guide and regulate instructional processes, while students actively participate in learning activities. Since instructional effectiveness emerges from this bidirectional interaction, instructional behaviors must be explicitly associated with their initiating subjects and pedagogical functions. This subject-aware design provides a clear theoretical foundation for aligning multimodal observations with instructional semantics.

Dual-subject Instructional Behavior Taxonomy. Based on the S–T interaction framework, TS-Interact employs a dual-subject instructional behavior taxonomy that distinguishes teacher-led and student-led behaviors, as shown in Table 1. Teacher-led behaviors describe instructional actions such as

content presentation, task guidance, resource support, and process management, while student-led behaviors capture learning activities including exploration, expression, operation, and collaboration. This hierarchical taxonomy ensures clear behavioral boundaries, reduces annotation ambiguity, and aligns each behavior category with observable multimodal signals from video, audio, and text. As a result, the proposed taxonomy supports reliable annotation, interpretable modeling, and effective multimodal behavior recognition.

3.2 Dataset Construction and Annotation

Multimodal Classroom Data Collection. To ensure the authenticity, representativeness, and disciplinary diversity of the dataset, all classroom videos used in this study were collected from the National Smart Education Platform for Primary and Secondary Schools and several provincial-level smart education platforms. All recorded lessons available on these platforms had been officially reviewed and approved by educational administrative authorities, ensuring high instructional quality and authoritative reliability. The collected videos represent real-world daily teaching scenarios rather than artificially designed experimental classrooms, with clear and stable audio–visual signals. A total of 41 complete classroom recordings were initially selected, covering seven academic subjects: physics, art, mathematics, chemistry, Chinese, history, and biology. This subject diversity enables the dataset to reflect varied instructional styles and interaction patterns across science, humanities, and arts disciplines, supporting broad applicability in multimodal classroom behavior analysis.

Data Preprocessing and Segmentation. To accurately capture meaningful instructional behavior units and meet the requirements of fine-grained behavior recognition, all collected classroom videos were segmented by trained personnel with an educational technology background. Using professional video editing tools, each full-length recording was divided into short clips with an average duration of 1–2 minutes. During segmentation, special care was taken to avoid splitting a coherent teacher–student interaction across multiple segments, ensuring that each clip contains a complete behavioral unit or a continuous interaction sequence. Segments with blurred visuals, severe background noise, missing instructional content, or disrupted audio–visual signals were excluded. For science experiment lessons, clips with unclear experimental procedures or missing key operational steps were further removed. All original videos had already undergone privacy protection procedures on the hosting platforms, and additional audio de-sensitization was applied to strictly comply with educational data ethics and privacy regulations. Classroom automated monitoring is now widely used in school education and has been generally accepted by teachers and students. Therefore, the potential negative impact of automated monitoring on teacher-student teaching interaction is negligible.

Dual-round Annotation Protocol. To enhance annotation reliability, a dual-round independent annotation protocol was adopted. The annotation team consisted of four annotators holding master's degrees in educational technology, all of whom had solid theoretical training and experience in multimodal data analysis. Prior to formal annotation, the annotators participated in a one-day centralized training session, which included theoretical instruction on the annotation taxonomy, detailed explanation of category definitions and decision rules, and hands-on practice with annotated and unannotated samples. In the first round, all annotators independently labeled the full dataset in isolated environments, with no communication permitted. One month later, to mitigate memory bias, a second-round annotation was conducted on a randomly sampled subset of 10% of the data (100 segments). The sampled segments preserved the original subject distribution and were randomly assigned to annotators such that no annotator re-labeled their own previous samples. All annotations in the second round were completed strictly according to the same guidelines, ensuring independence between rounds.

Annotation Reliability and Quality Control. To verify the consistency and reliability of the multi-label annotations, a two-dimensional evaluation framework was employed, addressing both label agreement and label set similarity. At the label level, Cohen's Kappa coefficient was used to assess whether individual behavior labels were consistently assigned. Following standard practice, values above 0.8 indicate almost perfect agreement, while values between 0.6 and 0.8 indicate high agreement. At the set level, the Jaccard coefficient and set-level F1 score were adopted to evaluate the overlap and balance of the complete label sets assigned to each segment, accommodating variations in the number of labels per sample. All annotation results were organized into a structured dataset, converted into binary label representations, and evaluated using Python-based scripts. The results demonstrate high consistency across all metrics, confirming the reliability of the annotation process.

Dataset Statistics and Content. As shown in Table 2, TS-Interact consists of 973 multimodal classroom segments, each with an average duration of 1–2 minutes, enabling fine-grained quantitative analysis of instructional behaviors at the interaction unit level. The dataset covers seven academic subjects, with physics (25.4%) and art (21.3%) forming the largest portions, followed by mathematics and chemistry (both 14.3%), ensuring balanced cross-disciplinary representation across science, humanities, and art domains. From a behavioral perspective, teacher-led behaviors dominate the dataset, with guiding question behaviors occurring most frequently (1,155 instances), followed by process management (539), knowledge presentation (538), and resource display (428). Student-led behaviors exhibit a complementary but less frequent distribution, where oral expression (690) constitutes the primary form of learner participation, while independent exploration (158), practical operation (187), and collaborative communication (16) reflect varying levels of learner autonomy and interaction intensity. Annotation quality is quantitatively supported by high inter-annotator agreement. Label-level consistency achieves an average Cohen's κ of 0.779 and a micro-average κ of 0.808, while set-level evaluation yields a Jaccard coefficient of 0.787 and an average set-level F1 score of 0.834. These quantitative indicators demonstrate that TS-Interact provides statistically reliable, behaviorally diverse, and well-balanced data support for multimodal instructional behavior analysis.

Fig.2 provides representative visual samples from TS-Interact,

Table 2: Core Dataset Statistics of TS-Interact

Dimension	Attribute	Description / Statistics
Data Scale	Total teaching segments	973 multimodal classroom segments
	Average segment duration	1–2 minutes per segment
	Data source	National and provincial smart education platforms
	Classroom setting	Authentic middle school classrooms (non-experimental)
	Privacy protection	Platform anonymization and audio de-sensitization
Subject Coverage	Subject distribution	Physics (247), Art (207), Mathematics (139), Chemistry (139), Chinese (91), History (87), Biology (63)
Behavior Distribution	Teacher-led behaviors	Knowledge presentation (538), guiding questions (1155), resource display (428), process management (539)
	Student-led behaviors	Independent exploration (158), oral expression (690), practical operation (187), collaborative communication (16)
Annotation Quality	Annotators	4 trained annotators with educational technology background
	Annotation protocol	Dual-round independent annotation with a one-month interval
	Label-level consistency	Average Cohen's Kappa Coefficient:0.779, Micro-average Cohen's Kappa Coefficient:0.808
	Set-level consistency	Average Jaccard Coefficient:0.787, Average Set-level F1 Score:0.834, Average Difference in Number of Labels:0.660

directly demonstrating the authenticity and high quality of the dataset. All samples are extracted from real, non-experimental middle school classrooms, preserving natural instructional rhythms, teacher–student interactions, and classroom layouts without staged or scripted behaviors. The visual examples show clear correspondence between annotated behavior labels and observable multimodal cues, such as teacher gestures, student postures, gaze directions, and interactions with instructional resources. This alignment indicates that the annotations are grounded in concrete visual and contextual evidence rather than abstract or inferred descriptions. Moreover, the diversity of scenes, subjects, and interaction patterns visible in the samples reflects the broad subject coverage and behavioral variability reported in Table 2. Together, these image-based samples substantiate the quantitative statistics by providing intuitive evidence of data realism, annotation reliability, and semantic clarity, reinforcing TS-Interact as a high-quality benchmark for multimodal instructional behavior analysis. To facilitate transparency and reproducibility, samples from the TS-Interact dataset has been made publicly available. Interested researchers can access these examples via the following link: https://osf.io/5q7kj/overview?view_only=6847956cd787410aa72f9cd407a0da91

3.3 Dataset Validation

Baseline Models for Multimodal Behavior Recognition. To evaluate the adaptability and effectiveness of the TS-Interact dataset for fine-grained classroom instructional behavior recognition, three representative multimodal baseline models were selected. All chosen models are lightweight, open-source, and easy to reproduce, with strong multimodal fusion capabilities, making them suitable as general benchmarks for assessing dataset quality rather than pursuing model-level state-of-the-art performance. The Qwen-2.5-Omni-3B (Qwen-2.5) model was selected due to its efficient multimodal fusion design and moderate parameter scale, which allows effective processing of video, audio, and textual inputs. It adopts a three-stage architecture consisting of single-modality feature extraction, cross-modal fusion via attention mechanisms, and behavior classification through fully connected layers, making it well aligned with the multimodal characteristics of TS-Interact. MiniCPM-V-4.5 (MiniCPM-V) emphasizes enhanced visual–language alignment and logical reasoning capabilities, enabling it to capture fine-grained instructional interaction patterns such as teacher–student question–answer dynamics. Its strong generalization ability makes it suitable for evaluating the discriminability and pedagogical rationality of the proposed annotation system. Qwen3-VL-4B-Thinking (Qwen3) adopts a dual-branch fusion architecture, where the visual branch extracts spatiotemporal video features and the textual branch models semantic information from transcribed classroom discourse, with audio features mapped into the textual space. Adaptive fusion between the two branches enables robust behavior recognition under limited data conditions, making this model particularly appropriate for validating dataset representativeness. Together, these three models cover diverse multimodal fusion strategies and provide a comprehensive evaluation of the dataset's adaptability, annotation quality, and practical usability.

Experimental Setup. The TS-Interact dataset was split into training and test sets using subject-stratified sampling with an 80%/20% ratio to maintain disciplinary balance. All splits were performed at the segment level to prevent data leakage. Classification performance was evaluated using Accuracy, Precision, Recall, and Macro-F1 to reflect both overall recognition and robustness to imbalanced behavior categories. Generation quality was assessed with ROUGE-1/2/L and BLEU-4, measuring semantic overlap and content completeness between generated descriptions and gold annotations. All baseline models were evaluated under identical settings within the LlamaFactory framework (max input length 2048, batch size 4, max generation length 384, temperature 0.01, nucleus sampling 1.0), with a fixed random seed (42) to ensure reproducibility.

Evaluation Results and Analysis. Table 3 reports the quanti-

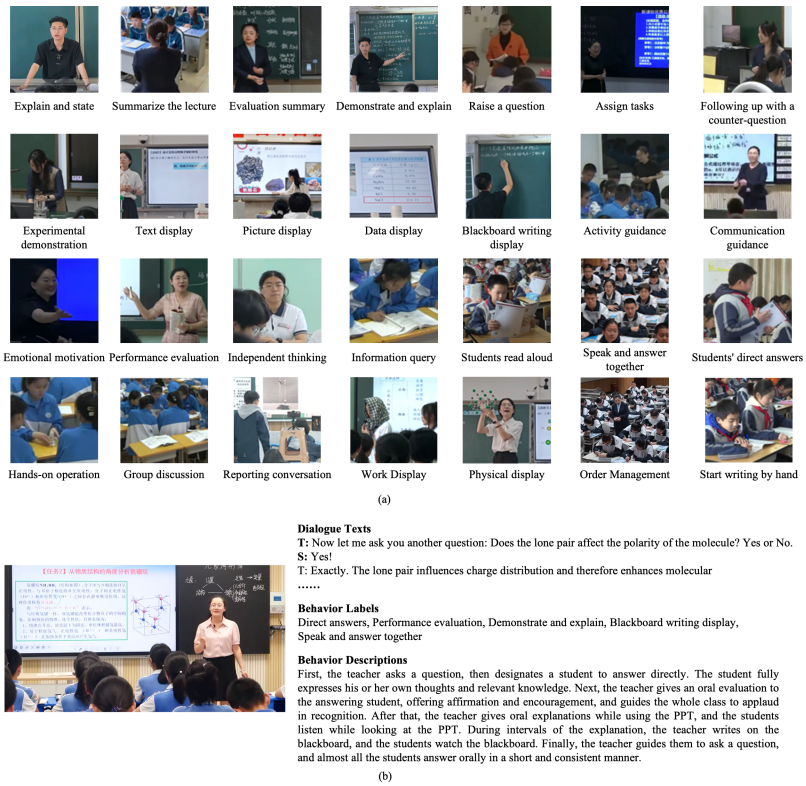

Figure 2: (a) Examples of Teacher-and-Student-led instructional behaviors in TS-interact (single-frame screenshot from a multimodal segment) (b) Annotated segment example in TS-Interact.

Table 3: Average Performance of Baseline Models on TS-Interact Across Seven Subjects

Model	F1	Accuracy	Precision	Recall	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4
Qwen3	**0.6582**	**0.1482**	**0.6525**	**0.6625**	**0.3572**	**0.1641**	**0.3330**	**0.0913**
Qwen-2.5	0.6275	0.1005	0.6208	0.6345	0.3315	0.1455	0.3068	0.0781
MiniCPM-V	0.5788	0.0615	0.5704	0.5860	0.2974	0.1239	0.2753	0.0647

tative results averaged over seven subjects. Qwen3 achieves the strongest overall performance, obtaining a Macro-F1 of 0.6582 and consistently outperforming other baselines across all generation metrics (e.g., ROUGE-L 0.3330, BLEU-4 0.0913). Compared with MiniCPM-V, Qwen3 yields a relative improvement of 13.7% in Macro-F1 and 21.8% in ROUGE-1, indicating superior alignment between multimodal inputs and fine-grained instructional behavior labels. Qwen-2.5 shows competitive results with reduced model size, suggesting a favorable trade-off between efficiency and accuracy. The consistent ranking of models across both classification and generation metrics demonstrates strong metric agreement and confirms the internal consistency of the evaluation protocol. Overall, the clear performance gaps among baselines highlight the discriminative power of TS-Interact and validate its suitability as a high-quality benchmark for multimodal classroom behavior recognition and instructional modeling.

4. DISCUSSION

This study presents TS-Interact as a multimodal classroom interaction dataset grounded in a structured and pedagogically informed teaching behavior taxonomy. From the perspective of the overall research design, several important insights emerge.

First, the results demonstrate that incorporating multimodal information is essential for fine-grained instructional behavior understanding. Across all baseline models, performance gains are consistently observed when visual, auditory, and textual cues are jointly leveraged, indicating that classroom teaching behaviors cannot be reliably inferred from a single modality alone. This finding aligns with educational theories that emphasize the embodied and situational nature of teaching, where gestures, visual materials, and classroom dynamics play a critical role alongside spoken language.

Second, the proposed teaching behavior annotation system contributes to more scientifically grounded behavior modeling. By explicitly distinguishing teacher-led and student-led behaviors and further decomposing them into pedagogically meaningful subcategories, TS-Interact avoids overly coarse

or purely action-level labels. This hierarchical structure enables models to capture both instructional intent and interaction patterns, which is reflected in improved classification stability and more coherent generation outputs. Compared with prior classroom datasets that rely on flat or loosely defined labels, the TS-Interact taxonomy provides clearer semantic boundaries and stronger interpretability.

Third, the performance gap among different multimodal large language models highlights both the promise and current limitations of general-purpose MLLMs in educational scenarios. While advanced models achieve reasonable results on behavior recognition and description, their accuracy remains relatively low, suggesting that classroom interaction understanding poses challenges beyond conventional vision-language tasks. This indicates opportunities for future work on education-specific adaptation, multimodal alignment, and reasoning mechanisms tailored to instructional contexts.

Finally, this study also has limitations. The dataset focuses on structured classroom settings and predefined academic subjects, which may limit direct generalization to informal or online learning environments. In addition, although the annotation framework is designed to be extensible, further validation across diverse educational cultures and instructional styles is still needed. Addressing these limitations will be an important direction for future research.

5. CONCLUSION

In this paper, we introduced TS-Interact, a multimodal classroom interaction dataset designed to support fine-grained analysis of teaching and learning behaviors. By integrating video, audio, and textual information with a pedagogically grounded annotation system, TS-Interact provides a more comprehensive and interpretable representation of classroom instructional dynamics. We conducted extensive experiments using state-of-the-art multimodal large language models, evaluating both classification and generation tasks. The results demonstrate that multimodal information is critical for understanding instructional behaviors and that the proposed teaching behavior taxonomy enables more structured and meaningful modeling of classroom interactions. Overall, TS-Interact offers a valuable benchmark for advancing multimodal educational AI research, particularly in areas such as classroom behavior understanding, instructional analysis, and intelligent tutoring systems. We hope this dataset will facilitate future studies on interpretable, scalable, and pedagogically informed multimodal learning models, and contribute to the development of AI systems that better support teaching and learning in real-world classrooms.

In future work, we plan to expand TS-Interact to cover a broader range of disciplines, extending from seven to nine subjects while maintaining a balanced distribution across categories. In addition, to complement the current validation based on lightweight multimodal models, we will incorporate larger-scale multimodal models to further examine the dataset's scalability and robustness. Beyond dataset expansion and validation, we aim to explore interdisciplinary instructional behavior analysis, dataset-driven model optimization, and the development of ethically grounded AI applications for educational contexts.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 62407042, 62277 045), Humanity and Social Science Foundation of Ministry of Education (No. 24YJC880004), and Natural Science Foundation of Shandong Province (No. ZR2024QF075)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

6. REFERENCES

[1] B. Cunling, L. Weigang, and F. Wei. Skeleton-based human action recognition: History, status and prospects. *Journal of Computer Engineering & Applications*, 60(20):1–29, 2024.

[2] N. A. Flanders. Analyzing teaching behavior. 1970.

[3] D. K. Geeganage, Y. Xu, and Y. Li. A semantics-enhanced topic modelling technique: semantic-lda. *ACM Transactions on Knowledge Discovery from Data*, 18(4):1–27, 2024.

[4] A. Geetha, T. Mala, D. Priyanka, and E. Uma. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. *Information Fusion*, 105:102218, 2024.

[5] C. Huang, J. Zhu, Y. Ji, W. Shi, M. Yang, H. Guo, J. Ling, P. De Meo, Z. Li, and Z. Chen. A multi-modal dataset for teacher behavior analysis in offline classrooms. *Scientific Data*, 12(1):1115, 2025.

[6] J. Huang, H. Hashim, H. Norman, M. H. Zaini, and X. Zhang. Automatic detection of teacher behavior in classroom videos using alphapose and faster r-cnn algorithms. *PeerJ Computer Science*, 11:e2933, 2025.

[7] A. Kapoor and R. W. Picard. Multimodal affect recognition in learning environments. In *Proceedings of the annual ACM international conference on Multimedia*, pages 677–682, 2005.

[8] Q. Liu, X. Jiang, and R. Jiang. Classroom behavior recognition using computer vision: A systematic review. *Sensors*, 25(2):373, 2025.

[9] X. Ma, Y. Xie, X. Yang, H. Wang, Z. Li, and J. Lu. Teacher-student interaction modes in smart classroom based on lag sequential analysis. *Education and Information Technologies*, 29(12):15087–15111, 2024.

[10] K. Mangaroska, R. Martinez-Maldonado, B. Vesin, and D. Gašević. Challenges and opportunities of multimodal data in human learning: The computer science students' perspective. *Journal of Computer Assisted Learning*, 37(4):1030–1047, 2021.

[11] C. Romero and S. Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics*, 40(6):601–618, 2010.

[12] C. Romero and S. Ventura. Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 10(3):e1355, 2020.

[13] J. Xiao, M. Chen, Y. Yang, and M. Liu. An exploratory multimodal study of the roles of teacher-student interaction and emotion in academic performance in online classrooms. *Education and Information Technologies*, 30(11):15507–15527, 2025.