Data-Knowledge-Driven Automatic Discovery of Teacher Classroom Teaching Behavior Indicator Categories

Ting Cai
Chongqing University of Posts and Telecommunications
caiting@cqupt.edu.cn

Qingyuan Tang
Chongqing University of Posts and Telecommunications
1549973104@qq.com

Yu Xiong*
Chongqing University of Posts and Telecommunications
xiongyu@cqupt.edu.cn

Lu Zhang
Chongqing University of Posts and Telecommunications
d240101037@stu.cqupt.edu.cn

ABSTRACT

Teacher classroom teaching behavior indicators serve as a crucial foundation for guiding instructional evaluation. Existing indicator system suffers from limitations such as strong subjectivity and weak contextual generalization capabilities. Generalized category discovery (GCD) enables automatic data clustering to identify known categories and discover novel ones. Drawing inspiration from GCD mechanisms, this paper proposes a data-knowledge-driven framework for automatic category discovery of classroom teaching behavior indicators (DKD-TBICAD). The framework utilizes partially labeled data as constraint guidance and leverages extensive unlabeled data as pattern mining carriers to achieve automatic discovery and classification of teaching behavior categories. Specifically, the framework enhances spatiotemporal feature discriminability through supervised contrastive learning and spatiotemporal neighborhood aggregation contrastive learning. Additionally, we design a dynamic domain feature aggregation strategy to optimize the adaptability of feature learning, further enhancing the framework's capabilities in feature aggregation and novel class discovery. Experimental results on the proprietary TBU dataset and public UCF101 dataset demonstrate that the proposed method achieves 4% higher overall accuracy than baseline models. On UCF101, it surpasses baselines by 8.9% in old-class accuracy, while on the TBU dataset, it achieves 10% higher accuracy in new-class recognition. We believe this study provides valuable insights for indicator generation research driven by bidirectional integration of expert knowledge and data knowledge.

Keywords

Teacher classroom Teaching behavior indicators, Generalized category discovery, Automatic discovery, Contrastive learning

Ting Cai, Qingyuan Tang, Yu Xiong, and Lu Zhang. Data-Knowledge-Driven Automatic Discovery of Teacher Classroom Teaching Behavior Indicator Categories. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 388–395. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870294

1. INTRODUCTION

The enhancement of classroom teaching quality relies on teachers' evidence-based reflective practice [19]. As the core carrier of the teaching process, the classroom teaching behaviors of teachers directly affect the efficiency of knowledge construction and the cognitive development of students [9]. Effective teaching behavior indicators should simultaneously meet the requirements of dynamic adaptability and interpretability to ensure they can flexibly reflect the diversity of teachers' behaviors in the classroom [23].

The current construction of the indicator system for teachers' classroom teaching behaviors mainly proceeds from two aspects: expert-driven and data-driven. The expert-driven method, based on literature analysis and expert experience, employs the Analytic Hierarchy Process (AHP) for optimization and weight determination. However, this approach tends to overly rely on expert experience, making it difficult to fully reflect the complexity and dynamism of teaching behaviors [1, 7, 21]. This behavioral difference of spatial dimension is the typical embodiment of the lack of dynamic adaptability of the existing index system. The data-driven method, on the other hand, integrates data knowledge with expert insights, mining behavioral patterns from data to refine the indicator system, thereby ensuring its scientific rigor and interpretability [2, 23]. However, this approach may rely too much on historical data and ignore behavioral patterns that are not highly relevant to existing topics but contain new information, especially in complex classroom environments and dynamic data [24].

Teacher-student interaction is often simplified into a single dimension in traditional indicator systems, overlooking spatial differences such as interactions on vs.outside the podium. Studies show that students feel stronger emotional support when teachers engage with them outside the podium [12]. While data-driven methods can capture such spatial distinctions, they depend on historical data. Without an automatic discovery mechanism, these differences may be averaged out or ignored, reducing the model's ability to distinguish behaviors.

Therefore, in the face of the existing indicator system's strong subjectivity and weak situational generalization ability, the automatic discovery mechanism of teacher classroom teach-

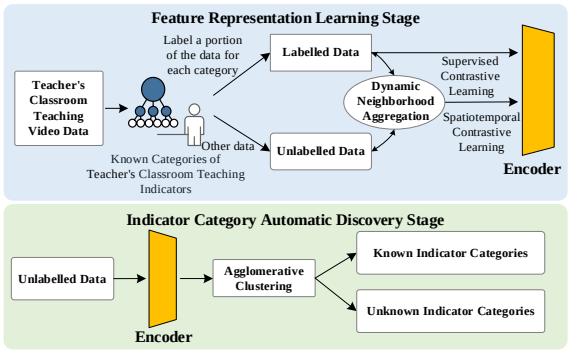

Figure 1: A Framework for Data-Driven Automatic Discovery of Teacher Classroom Behavior Indicator Categories.

ing behavior indicator categories came into being. This mechanism aims to integrate expert and data knowledge, dynamically identify behavioral patterns in diverse classroom scenarios through data mining technology, and supplement and improve the existing evaluation system.

Generalized category discovery (GCD) is a mechanism suitable for discovering categories in the real world [18]. It can automatically identify unknown categories and classify known ones using labeled category information in an open-world scenario, without the need to predefine the number of categories. This mechanism aligns closely with the requirements for automatic discovery of teacher classroom behavior indicator categories. Currently, GCD research primarily focuses on the field of image recognition, where contrastive learning is used to construct feature representation spaces, combined with clustering methods to achieve the discovery of unknown categories[18, 5, 14]. Despite significant progress in the image domain, research on GCD for behavior category discovery in video data is still in its early stages. Video data contains rich spatiotemporal information. Studies have shown that spatiotemporal contrastive learning can effectively uncover dynamic behavioral features[22, 6], providing a potential breakthrough for GCD to discover new classes in video data[13]. However, existing methods have the problem of insufficient intra-class feature aggregation when processing video data. Therefore, combining the GCD clustering idea with contrastive learning is expected to improve the discovery of new categories. Based on the above analysis, combined with the mechanism of GCD, this paper proposes a data-driven framework for the automatic discovery of teacher classroom behavior indicator categories (DKD-TBICAD), which aims to dynamically discover potential new categories. The framework can be seen in Figure 1. DKD-TBICAD consists of two stages: feature representation learning and indicator category automatic discovery. The feature representation learning stage uses a small amount of labeled data based on expert knowledge and a large amount of unlabeled data to conduct supervised contrastive learning and spatiotemporal aggregation contrastive learning. At the same time, dynamic neighborhood aggregation is designed to further enhance the spatiotemporal aggregation capability of video data. The indicator category automatic discovery stage aims to use the trained encoder to cluster features of unlabeled data, enabling the recognition of known categories and the discovery of unknown behavior categories. In this framework, known categories guide the model's sensitivity to unknown categories through expert knowledge, while the dynamic neighborhood fusion of known and unknown category data guides the model to focus on the features themselves. This process combines expert knowledge with data-driven, dynamically adapts to changes in teaching scenarios, optimizes indicator adjustments, enriches the category label library, and alleviates the deviation problem in the design of education evaluation indicators. Our contributions are summarized as follows:

- We propose a data-driven framework based on the GCD for automatic discovery of teacher classroom teaching behavior indicator categories (DKD-TBICAD). The framework uses a small number of known category data as guidance and a large amount of unknown category data as feature patterns to achieve automatic recognition and discovery of teaching behavior category indicators.
- We propose a spatio-temporal neighborhood aggregation contrastive learning method (STNA). This method optimizes feature consistency through a dynamic neighborhood aggregation strategy, generating representative minimal pseudo-category prototypes, enhancing the model's ability to learn discriminative features in complex classroom environments.
- Experimental results on proprietary and public datasets demonstrate that the proposed method achieves 4% higher overall accuracy and 10% superior new-class recognition accuracy compared to baseline models. Outcomes validate the method's effectiveness and generalizability in automatic discovery of classroom teaching behavior indicator categories.

2. RELATED WORK

2.1 Design of Teaching Behavior Indicator

Currently, there are mainly two approaches to constructing the teaching behavior indicator system: the expert-driven approach and the data-driven approach. The expert-driven approach relies on theoretical analysis and expert experience. For instance, Atapattu et al. [1] constructed the behavioral indicators of students' cognitive engagement based on theoretical analysis, emphasizing the combination of quantitative and qualitative methods. Ding et al.[7] used association rules to explore the relationships among indicators. Chong et al.[20] determined the indicator weights based on the Analytic Hierarchy Process. The data-driven approach determines key indicators through data analysis [16]. Yu et al. [23] constructed an evaluation indicator system for teachers' teaching reflection based on data. However, this study holds that this indicator system needs to be verified in a more complex environment. Shravya et al.[2] automatically generated indicators by integrating generative models. Zhang et al.[24] discovered that existing indicators tend to overlook data containing new information.

Current teaching behavior indicator systems have made progress, but two key challenges remain. First, expert-defined indicators struggle to keep up with dynamic classroom changes,

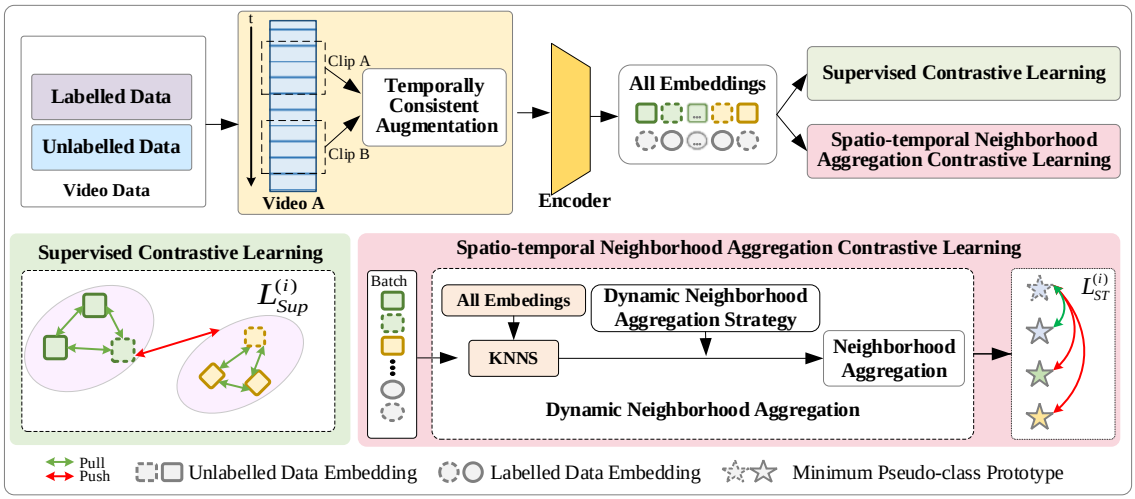

Figure 2: Overview of the STAN.

making it hard to capture new behavior patterns. Second, data-driven methods rely heavily on historical data, which may miss behaviors that are less correlated with existing indicators but still educationally valuable. This highlights the need for an automatic discovery framework to better adapt to evolving classroom environments.

2.2 Generalized Category Discovery

Generalized Category Discovery (GCD) is a method designed to automatically identify and explore unknown categories by leveraging existing known category information in open-world learning environments, thereby dynamically expanding the category system [18]. The general practice of GCD involves optimizing feature space representations through clustering algorithms. Zhao et al. [25] proposed a semi-supervised variant of Gaussian mixture models to examine the compactness and separation of clusters, dynamically determining feature prototypes, and further optimized representation learning through prototype refinement. Pu et al.[14] employed the Infomap clustering algorithm to generate dynamic concept prototypes, achieving highly discriminative feature representations through dual-layer contrastive learning at both instance and concept levels. To better integrate features between labeled and unlabeled data, Choi et al.[5] combined mean-shift clustering with contrastive learning, attaining enhanced feature representation performance in GCD. Although contrastive learning has demonstrated promising results in GCD, neighborhood features often suffer from interference caused by noisy data when processing complex datasets. Consequently, further filtering and refinement of neighborhood features have become critical.

In this context, this study applies key technologies such as contrastive learning and clustering in GCD to the analysis of time-series teaching behavior characteristics, solves key problems such as dynamic aggregation of time-series features and cross-scene generalization, and uses labeled data to design a neighborhood noise filtering strategy to improve the ability of the model to automatically discover indicator categories.

2.3 Video Representation Based on Spatiotemporal Contrastive Learning

Spatiotemporal contrastive learning is a widely used technique in video action recognition, aiming to enhance the expressive power of video features through joint learning in temporal and spatial dimensions. Qian et al. [15] proposed a temporal consistency data augmentation strategy that treats different clips of the same video as positive samples for contrastive learning, simultaneously capturing spatial and temporal characteristics of videos. Feichtenhofer et al. [10] encouraged temporal consistency in video features by utilizing different temporal segments of the same video as positive samples. Han et al. [11] used the RGB stream features and optical flow features of the same video data as positive sample pairs for self-supervised contrastive learning to learn spatiotemporal representations. However, the aforementioned works merely focus on the temporal-spatial contrastive learning of individual instances, while neglecting the influence of neighboring samples in the temporal-spatial domain.

3. PROBLEM FORMULATION

This study introduces the GCD mechanism that enhances the dynamic adaptability and interpretability of indicator systems by automatically identifying novel behavioral categories in classroom teaching through the integration of expert knowledge and data-driven strategies.

Specifically, GCD can automatically discover novel categories in the unlabeled dataset D_u without requiring predefined knowledge of the true number of categories. Let the training dataset be $D = D_l \cup D_u$, where $D_l = \{(x_i, y_i)\}_{i=1}^{N_l} \in X \times \mathcal{Y}_l$ represents the labeled subset (known categories), and $D_u = \{x_j\}_{j=1}^{N_u} \in X \times \mathcal{Y}_u$ represents the unlabeled subset (containing both known and unknown categories). Here, $\mathcal{Y}_l$ denotes the known categories. And $\mathcal{Y}_u$ denotes all possible categories in the data, including the known categories $\mathcal{Y}_l$ and the unknown categories. The terms x_i and x_j denote data samples, y_i represents data labels, X denotes the complete data space. Let K be the total number of categories in D,

which remains unknown during model training and needs to be predicted by the model. To estimate K, we introduce a validation set D_v disjoint from the training set. During the validation process, the model dynamically adjusts the number of clusters in the agglomerative clustering procedure and computes the clustering accuracy for samples from the known classes. When the accuracy for known classes reaches its maximum, we hypothesize that the corresponding number of clusters equals K, which represents the total number of both known and unknown categories.

4. METHODOLOGY

4.1 Model Framework

Figure 2 shows the model architecture (STNA) of the representation learning stage in the proposed framework. The model first inputs labeled and unlabeled teacher classroom teaching behavior video clips. Then, two clips are randomly selected for temporally consistent data augmentation. Finally, the augmented data is passed through an encoder to extract features, followed by supervised contrastive learning and spatio-temporal neighborhood aggregation contrastive learning. Among them, supervised contrastive learning makes the features of the same class more compact in the feature space and distinguishes the features of known and new classes. Spatio-temporal neighborhood aggregation contrastive learning combines the semantic similar features of labeled and unlabeled data in the feature space by selecting neighborhood samples through KNN and aggregating them. This process learns high-quality feature representations and enhances the ability to discover new classes.

4.2 Temporally Consistent Augmentation

In the framework of this study, we rely on contrastive learning to identify old classes and discover new teaching behavior categories. Therefore, a temporal consistency data augmentation method is adopted to provide effective positive sample pair generation for contrastive learning. Specifically, two clips $\{ClipA, ClipB\}$ are randomly selected from each video, and the same enhancement operation is applied to each frame to ensure the consistency of temporal cues between video clips. This operation prevents the time consistency from being disrupted, which is often caused by conventional spatial augmentation methods. It ensures that the sample pairs before and after augmentation $\{v_i, v_i^+\}$ exhibit different features in both spatial and temporal dimensions, helping the model learn richer spatiotemporal information.

4.3 Supervised Contrastive Learning

In GCD, the classification head tends to overfit to known categories, causing the features of new category data to align closely with those of known categories, which hinders the discovery of new categories. To mitigate this issue, this study employs supervised contrastive learning to directly learn feature representations of known category data. Specifically, in the feature space, the features of data from the same category are pulled closer as positive sample pairs, while the features of data from different categories are pushed apart as negative sample pairs. Through this approach, supervised contrastive learning not only clusters the features of data from the same category in the feature space but also increases the inter-class distance between different categories, making the feature space distribution more structured and reducing the difficulty of discovering new categories.

In supervised contrastive learning, the supervised contrastive loss for a single data sample is calculated as follows:

$$L_{\text{Sup}}^{(i)} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(v_i \cdot v_p / \tau_s\right)}{\sum_{j \notin P(i)} \exp\left(v_i \cdot v_j / \tau_s\right)} \quad (1)$$

$P(i)$ represents the set of data samples within the same batch that share the same label category as the data sample v_i. Here, v_p denotes the feature of the data or augmented sample that shares the same label as v_i, and τ_s represents the temperature coefficient.

4.4 Spatio-temporal Neighborhood Aggregation Contrastive Learning

This module is designed to improve the category discrimination ability of feature representation in contrastive learning. This module consists of Dynamic Neighborhood Aggregation and Minimal pseudo-class prototype contrastive learning.

4.4.1 Dynamic Neighborhood Aggregation

Traditional spatiotemporal contrastive learning usually relies on a single data instance and ignores the global semantics at the category level. To optimize feature space learning, steps such as neighborhood sample selection, dynamic neighborhood feature aggregation strategy and minimum pseudo-class prototype generation are designed to improve the performance.

KNN-based neighborhood sample selection. In the feature embedding space V of all data, the KNN algorithm is used to find the n nearest neighbor sample features $N(v_i)$ and $N(v_i^+)$ for each sample pair $\{v_i, v_i^+\}$. Specifically, during the KNN neighborhood sample selection process, the sample features are derived from all video embeddings, where the labels of some sample embeddings are known. Therefore, the feature aggregation representation of a single sample after KNN is as follows:

$$N(v_i) = \text{argmax}_{v_j \in V}^{n}\, v_i \cdot v_j \quad (2)$$

V represents the feature embeddings of all data, and n is the number of neighborhood samples.

Dynamic neighborhood aggregation strategy. To enhance the consistency of these neighborhood sample features, a dynamic neighborhood aggregation strategy to refine the neighborhood features is designed to obtain the optimized feature representation $refine(N(v_i))$, which better represents the category to which the sample belongs. The strategy is as follows:

(1) Labeled Sample Processing: If the data sample v_i in the training batch is itself a labeled sample, the known labeled samples in the neighborhood $N(v_i)$ must belong to the same category as the sample itself; otherwise, they are filtered out.

(2) Unlabeled Sample Processing: If the sample v_i is an unlabeled sample, only the labeled samples with the most frequent category in the neighborhood $N(v_i)$ are retained,

and other samples are filtered. If more than half of the data in the neighborhood $N(v_i)$ are of the same category label, only these labeled data are retained.

Generation of minimum pseudo-class prototype. After obtaining the refined feature representation $refine(N(v_i))$, a mean aggregation method is used to merge the neighborhood data. And an adjustment factor $\alpha \in [0, 1]$ is introduced to control the weight distribution between the neighborhood samples and the original features. Finally, the neighborhood aggregation method yields the minimal pseudo-category prototype for the sample, as shown in Formula 3.

$$z_i = (1 - \alpha)v_i + \alpha\, Mean(refine(N(v_i)) \tag{3}$$

4.4.2 Minimal Pseudo-class Prototype Contrastive Learning

Based on the generated minimal pseudo-category prototypes, a spatiotemporal neighborhood aggregation contrastive loss function is designed to optimize the feature learning ability of the encoder. The specific loss function is as follows:

$$L_{\mathrm{ST}}^{(i)} = -\log \frac{\exp(z_i, z_i^+/\tau_u)}{\sum_{j \neq i} \exp(z_i, z_j/\tau_u)} \tag{4}$$

τ_u is the temperature coefficient, which is used to adjust the similarity range of positive and negative sample pairs in contrastive learning. This loss function encourages pseudo-category prototypes of the same category to cluster together in the feature space while pushing apart prototypes of different categories, thereby enhancing the discriminative power of the features.

During the feature representation learning stage (training process), supervised contrastive learning and spatiotemporal neighborhood aggregation contrastive learning are jointly optimized to enhance the feature space, thereby improving the model's ability to recognize known categories and discover new ones. The total loss for model optimization is as follows:

$$L = \frac{1}{|B|} \sum_{i \in B} L_{ST}^{(i)} + \beta \frac{1}{|B_L|} \sum_{i \in B_L} L_{Sup}^{(i)} \tag{5}$$

β is a hyperparameter, and B_L represents the labeled data in each batch.

5. EXPERIMENT

5.1 Dataset

To comprehensively evaluate the effectiveness of STAN, two datasets are used: TBU dataset [3] and UCF101 dataset [17].

TBU: TBU is a large-scale proprietary multi-task video dataset for teacher classroom behaviors, containing video tasks such as classification, detection and description. Specifically, its action classification subset comprises 13 classes with 37,026 video samples. For this study, we select a subset of 11 classes (excluding the "teacher bowing" and "erasing blackboard" categories with small sample sizes), resulting in 15,638 video samples. The detailed sample distribution is presented in Table 1.

UCF101: UCF101 is a general-purpose action recognition benchmark containing 101 classes with 13,320 video samples.

Dataset Splitting: During dataset division, 20% of all data is first allocated as the validation set D_v. Next, a subset of categories is selected as the known classes $\mathcal{Y}_l$. Among the known class data, 50% is used to form the labeled dataset D_l, which contains the categories $\mathcal{Y}_l$. The remaining data constitutes the unlabeled dataset D_u, which contains the categories $\mathcal{Y}_u$. In addition to the known categories $\mathcal{Y}_l$, $\mathcal{Y}_u$ also includes unknown categories that do not appear in the labeled data. During model training, the labels of the data in D_u are invisible. The model is trained jointly on the labeled dataset D_l and the unlabeled dataset D_u. The entire D_u serves as the test set. The division of the number of categories and samples is shown in Table 2.

Selection of new and old classes: In the UCF10 dataset, due to the balanced distribution of classes, we randomly select 50 of them as unknown classes to verify the model performance. In the TBU dataset, considering that the category distribution is often unbalanced in real teaching scenarios, we include categories with more and less data in the division of known classes and unknown classes to be closer to the actual application situation. Specifically, we select 4 categories from the 7 categories with a relatively large amount of data and 2 categories from the 4 categories with a relatively small amount of data. In total, 6 categories are designated as the old classes, and the remaining categories are regarded as the new classes.

5.2 Experimental Implementation

STAN uses ViT-B-16 [8] pretrained with DINO [4] as the encoder. The training epoch is 50, the learning rate is 0.001, the batch size is 128, and the Adam optimizer is used. The experiment is completed on 4 3090 GPUs. After each training epoch, the clustering accuracy and the number of estimated categories of the model are evaluated using the agglomerative clustering method on the validation set. The best model is selected after multiple rounds of training. The performance of the model is evaluated on the test set. Since GCD in the video domain is still in its infancy, we adaptively modify the GCD [18] and CMS [5] models from the image domain to build two powerful baseline models.

5.3 Comparison with the Baselines

Table 1: TBU Dataset

Category	Quantity
Lecture on the Podium	3079
Multimedia Teaching	2755
Lecture Underneath the Podium	2361
Interact with Students Outside the Podium	2289
Interacting with Students On the Podium	1314
Board Writing	1440
Classroom Inspection Underneath the Podium	1362
Displaying Teaching Aids	468
Pointing To the Blackboard	335
Operating Multimedia	125
Classroom Inspection Around On the Podium	110

Table 2: Dataset Splitting

	TBU	UCF101
$\mathcal{Y}_l$	6	50
$\mathcal{Y}_u$	11	101
D_l	4336	2638
D_u	8174	8018

Table 3: Results on Different Datasets

	TBU			UCF101		
Classes	All	Old	New	All	Old	New
agglomerative	0.3026	0.2593	0.3427	0.6126	0.5559	0.6419
GCD	0.4836	0.7159	0.2735	0.5925	0.7605	0.5070
CMS	0.6449	**0.7905**	0.5131	0.7603	0.8632	0.7083
STNA	**0.6859**	0.7639	**0.6154**	**0.8109**	**0.9518**	**0.7392**

In table 3, we respectively report the comparative experimental results on the TBU dataset and the UCF101 dataset. Overall, STNA achieves a 4% improvement in overall clustering accuracy and 10% higher accuracy for novel classes compared to the baseline models. This demonstrates that the features encoded by STAN better capture the semantic information of the data itself, effectively balancing the representation of novel and known classes in the feature space, thereby providing reliable support for the dynamic discovery of teacher behavior indicators.

In terms of model comparison, STAN significantly outperforms GCD and Agglomerative in clustering accuracy for novel classes and surpasses the CMS model by 10%. This discrepancy arises because CMS excessively compresses the features of known classes through mean shift, limiting the learning space for novel classes. In contrast, STAN enhances the neighborhood relationships between known and novel class features in the feature space through a dynamic neighborhood aggregation strategy. This allows novel class features to maintain discriminability while co-evolving with known class features, making it more suitable for behavior discovery in open classroom scenarios.

The results on the UCF101 dataset show that STNA improves the accuracy for known classes by 8.9% compared to the baseline, validating the model's strong representation capability for general video behavior features. For the TBU dataset, the proposed method slightly underperforms CMS in known class recognition. This may be due to the challenges posed by the real-world classroom setting of TBU, such as long-tailed distribution and multi-view perspectives, which introduce noise during training and feature selection, affecting the performance on known classes. Nevertheless, STNA still achieves effective discovery of novel classes in the long-tailed TBU dataset, demonstrating its practicality in real-world educational scenarios.

5.4 Estimating the Number of Classes

Table 4 presents the true number of categories K for different datasets, along with the estimated K values from STAN, GCD, and CMS. GCD estimates the number of categories after training, while STAN and CMS estimate the number of categories at the end of each training epoch. The experimental results demonstrate that STAN provides the closest estimation to the true number of categories for unlabeled data, with a maximum error of only 9.1%, significantly outperforming GCD and CMS. This is because GCD suffers from overfitting to known classes, which limits its ability to discover novel classes and results in larger estimation errors. STAN outperforms CMS due to the more balanced distribution of novel and known class features in the feature space, leading to more accurate category estimation.

Table 4: Estimation of the Number of Classes

Method	TBU		UCF101	
	K	Err(%)	K	Err(%)
Ground truth	11	-	101	-
GCD	7	36.4	91	9.9
CMS	9	18.2	92	8.9
STNA	10	9.1	94	6.9

Figure 3 illustrates the changes in the accuracy of novel and old classes, as well as the estimated number of categories K, on the validation set across different epochs during the training process of the STNA and CMS models. It can be observed that as the training progresses, STNA significantly outperforms CMS in novel class discovery and category quantity estimation. The old class recognition rate of CMS is relatively stable, but the new class discovery performance is poor (as shown in Table 4). STNA, on the other hand, performs stably in new class discovery and old class recognition, making it more advantageous in category number estimation. This is primarily attributed to STNA's dynamic neighborhood aggregation design, which effectively mitigates the impact of background complexity and behavioral similarities, reducing the interference of noise on representation learning. This also proves STNA's ability to adapt to behavioral differences in complex classroom environments.

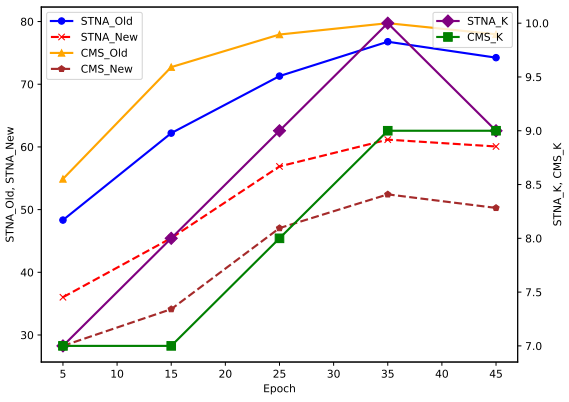

Figure 3: Comparison of STNA and CMS across Epochs on TBU

6. CONCLUSIONS

Drawing inspiration from the generalized category discovery mechanism, this study proposes a data-knowledge-driven

framework for the automatic discovery of teacher classroom instruction behavior indicators. The framework aims to dynamically uncover implicit, non-predefined behavior patterns from classroom teaching video data, providing technical support for building intelligent educational evaluation indicators. Supervised contrastive learning is designed to constrain the feature distribution of known categories, reserving semantic space for novel class discovery. Spatiotemporal neighborhood aggregation contrastive learning is designed to enable self-organization of novel class features into compact clusters while optimizing the discriminability of known class features through dynamic neighborhood feature fusion. To estimate the number of unknown categories, clustering algorithms are used during training to classify validation set data. This approach discovers both novel and known classes, eliminating the need for predefined category counts required by traditional methods. Experimental results demonstrate that the model achieves significant performance on both the proprietary dataset TBU and the general behavior dataset UCF101. Under the challenge of long-tailed distribution in the TBU dataset, STNA improves novel class discovery accuracy by 10% compared to baseline models. Additionally, it outperforms other baselines in estimating the number of unknown categories, showcasing the significant potential of the proposed model.

This study focuses on the automatic discovery of teaching behavior indicators in classroom, driven by both expert knowledge and data knowledge, under the constraints of limited annotated data and leveraging large-scale unlabeled data as feature carriers. This method not only overcomes the limitations of traditional manually set indicators, but also dynamically reflects the diversity and complexity of teaching behaviors, providing support for building a more objective education evaluation system. Specifically, in the educational evaluation scenario, the framework can automatically discover key teaching behaviors, supplement and extend the existing indicator system. By enabling the dynamic evolution of the index system of human-computer collaborative education, this study provides technical support for the two-way cooperation of expert knowledge and data knowledge to build a perfect, comprehensive and real-time index system. Although the method proposed in this paper has achieved good results, it has certain limitations. Mainly in two aspects: first, the model's ability to discover low-frequency behavior categories is limited, and when the sample size of new categories is too small, the feature representation learning effect will significantly decline; second, the validation data may have incomplete coverage of new categories, which will affect the reliability of model parameter updates. These limitations mainly stem from the inherent imbalance and dynamic evolution characteristics of behavior data in the educational scenario.

As educational data continue to grow in diversity and complexity, future research can explore the following directions. First, addressing the issue of class imbalance in educational data, particularly the imbalance in novel class data, by exploring data generation or intelligent sampling strategies to mitigate obstacles in novel class discovery. Second, address the issue of the model having an overly strong bias towards known categories. Disentangled representation learning or attention mechanisms can be considered to reduce the influence of known category data on the representation of new category data.

7. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No.62377007), the Chongqing Key Project for Higher Education Teaching Reform Research (No.232073), and the Major Science and Technology Research Projects of the Chongqing Municipal Education Commission (No.KJZD-M202400606, No.KJZD-M202300603),the Chong-qing Key Project for Natural Science Foundation Innovation and Development Joint Fund(No.CSTB2024NSCQ-LZX0133), the Chongqing Municipal Education Commission Science and Technology Research Youth Project (No.KJQN-202400634).

8. REFERENCES

[1] T. Atapattu, M. Thilakaratne, R. Vivian, and K. Falkner. Detecting cognitive engagement using word embeddings within an online teacher professional development community. *Computers & Education*, 140:103594, 2019.

[2] S. Bhat, H. A. Nguyen, S. Moore, J. C. Stamper, M. Sakr, and E. Nyberg. Towards automated generation and evaluation of questions in educational domains. In *EDM*, 2022.

[3] T. Cai, Y. Xiong, C. He, C. Wu, and S. Zhou. Tbu: A large-scale multi-mask video dataset for teacher behavior understanding. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[5] S. Choi, D. Kang, and M. Cho. Contrastive mean-shift learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23094–23104, 2024.

[6] I. Dave, R. Gupta, M. N. Rizve, and M. Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, 2022.

[7] Y. Ding. Research on the evaluation algorithm of english teaching indicators based on data mining technology. In *Proceedings of the International Conference on Modeling, Natural Language Processing and Machine Learning*, pages 368–374, 2024.

[8] A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] B. Fauth, W. Wagner, C. Bertram, R. Göllner, J. Roloff, O. Lüdtke, M. S. Polikoff, U. Klusmann, and U. Trautwein. Don't blame the teacher? the need to account for classroom characteristics in evaluations of teaching quality. *Journal of Educational Psychology*, 112(6):1284, 2020.

[10] C. Feichtenhofer, H. Fan, B. Xiong, R. Girshick, and K. He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 3299–3309, 2021.

[11] T. Han, W. Xie, and A. Zisserman. Self-supervised co-training for video representation learning. *Advances in neural information processing systems*, 33:5679–5690, 2020.

[12] L. Hu and Y. Wang. The predicting role of efl teachers' immediacy behaviors in students' willingness to communicate and academic engagement. *BMC psychology*, 11(1):318, 2023.

[13] X. Jia, K. Han, Y. Zhu, and B. Green. Joint representation learning and novel category discovery on single- and multi-modal data. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 590–599, 2021.

[14] N. Pu, Z. Zhong, and N. Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7579–7588, 2023.

[15] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6964–6974, 2021.

[16] S. Z. Salas-Pilco, K. Xiao, and X. Hu. Artificial intelligence and learning analytics in teacher education: A systematic review. *Education Sciences*, 12(8):569, 2022.

[17] K. Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[18] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501, 2022.

[19] Y. Wang. Teaching journal: An effective tool for reflective practice in teaching among english language teachers. *Journal of Contemporary Educational Research*, 5(8):73–79, 2021.

[20] C. Wei, X. Zeng, Z. Wang, S. Li, Y. Tong, H. Xu, and L. Cao. Construction and research on the evaluation system of university curriculum teaching quality based on analytic hierarchy process. *Curriculum and Teaching Methodology*, 5:10–17, 2022.

[21] X. Wu. A study on data-driven cluster analysis of teaching quality in civic and political education in colleges and universities. *Applied Mathematics and Nonlinear Sciences*, 9, 07 2024.

[22] T. Yao, Y. Zhang, Z. Qiu, Y. Pan, and T. Mei. Seco: Exploring sequence supervision for unsupervised representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10656–10664, 2021.

[23] N. Yu. Research on the construction of evaluation index system of teachers' teaching reflection and intelligent evaluation based on artificial intelligence. In *Proceedings of the 2023 International Conference on Information Education and Artificial Intelligence*, pages 799–803, 2023.

[24] S. Zhang, Q. Gao, Y. Wen, M. Li, and Q. Wang. Automatically detecting cognitive engagement beyond behavioral indicators. *Educational Technology & Society*, 24(2):58–72, 2021.

[25] B. Zhao, X. Wen, and K. Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16623–16633, 2023.