Natural Language-Driven Teacher Gesture Recognition

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

Shengyi Chen
Chongqing University of Posts
and Telecommunications
835044065@qq.com

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

Lulu Chen
Chongqing University of Posts
and Telecommunications
chenll@cqupt.edu.cn

Jun Li
Chongqing University of Posts
and Telecommunications
22070811637@qq.com

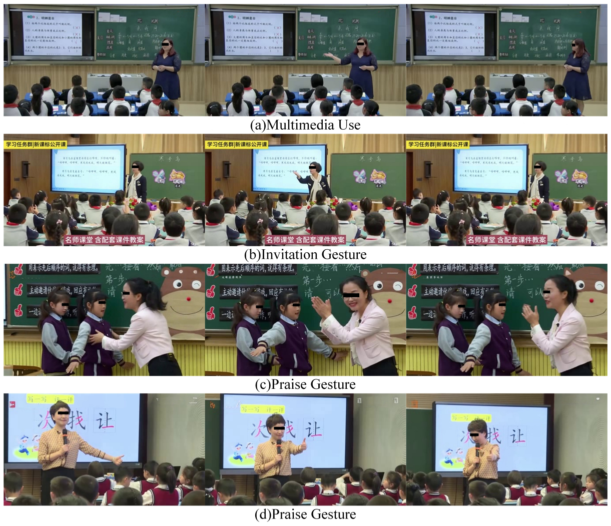
ABSTRACT

Teacher gesture recognition aims to identify and interpret teacher gestures within academic settings. It has been applied in domains such as teaching performance evaluation, the optimization of online education, and special needs education. However, the background similarity of teacher gestures, the inter-class similarity, and the intra-class variability limit the recognition capabilities of visual neural networks. In this paper, a Natural Language-Driven Teacher Gesture Recognition (NLD-TGR) framework is proposed. To mitigate the effects of background similarity, textual descriptions for each frame are generated using GPT-4o, guided by prompts specifically designed to describe the teacher's hand posture in the frames. Then, we combine video features with text features mapped to a high-dimensional space to create semantically-enhanced fused features. To overcome the limitations of one-hot labels in capturing inter-class and intra-class relationships, we embed semantically interpreted category names into a textual feature space. Gesture classification is then performed by computing the similarity between these textual embeddings and the fused feature representations. The experimental results validate the effectiveness of the proposed method, which achieves state-of-the-art performance with an accuracy of 93.7% on the TBU-G teacher gesture benchmark.

Keywords

Teacher Gesture Recognition, Classroom Scenario, Natural Language

1. INTRODUCTION



(a)Multimedia Use

(b)Invitation Gesture

(c)Praise Gesture

(d)Praise Gesture

Figure 1: Figures a and b have similar teacher hand gestures, but due to the differences in their interaction objects, they belong to different categories. Conversely, the gestures in Figures c and d, although different, share the same meaning. Additionally, the backgrounds of each gesture are highly similar. Please zoom in for the best view.

Teacher gestures are widely regarded as an effective tools to improve teaching efficiency and foster a positive learning atmosphere[19]. When used in coordination with verbal communication, gestures improve the clarity and precision of delivering teaching content, directing students' attention to key points highlighted by teachers [14]. Thus, teacher gestures serve as important indicators of instructional attitudes and pedagogical skills. With the continuous advancement of deep learning technologies, the integration of artificial intelligence in education is becoming increasingly profound [24, 12]. AI technologies can be utilized for evaluating educational quality, thereby enabling an objective analysis of teaching issues [4]. To explore how teacher gestures influence the effectiveness and quality of teaching, a detailed analysis of classroom gestures is necessary [22]. The key to success in this analysis lies in efficiently and accurately recognizing these gestures.

Extensive studies[7, 13, 17, 18, 26, 28] have confirmed the effectiveness and reliability of deep learning techniques in teacher gesture recognition. However, current studies still lack a comprehensive understanding and in-depth analysis of teacher gestures. Most approaches [7, 17, 28] focus on extracting hand shape features and tracking changes in hand postures, directly adapting techniques from keypoint-based

Yu Xiong, Shengyi Chen, Ting Cai, Lulu Chen, and Jun Li. Natural Language-Driven Teacher Gesture Recognition. 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. 482–488. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870296

sign language recognition. Unfortunately, these methods exhibit significant limitations when specifically applied to teacher gesture recognition. Keypoint-based gesture recognition methods[3, 7, 23] demonstrate high robustness to challenges such as variations in illumination, complex backgrounds, and occlusions. These methods effectively track hand skeletal points and reliably identify hand shapes and motion trajectories. However, their primary limitation lies in isolating hand gestures from the overall teaching context by focusing exclusively on hand postures. In classroom settings, hand gestures are closely linked to the surrounding environment and teaching activities, and their meanings can only be accurately interpreted when analyzed in context. As illustrated in Figure 1a and 1b, identical hand gestures may convey completely different meanings depending on the specific interaction context. Meanwhile, appearance-based methods[29, 31, 32] for recognizing teacher gestures predominantly focus on temporal modeling. These methods aim to capture and analyze subtle hand movement variations within video sequences to achieve robust gesture recognition. However, these methods often overlook the issue of inter-class similarity and intra-class variability in teacher gestures, Which leads to frequent misclassification of teacher gesture tasks.

To address the aforementioned issues, we propose a natural language-driven teacher gesture recognition (NLD-TGR) model designed to accurately recognize gestures of teachers in real classroom teaching scenarios. In contrast to conventional natural language-assisted methods, our approach utilizes GPT to generate detailed descriptions of gesture actions for individual frames, offering a significantly higher level of specificity and granularity. Specifically, given a video of teacher gestures, a generative model GPT-4o[1] is employed to produce textual descriptions for the extracted frames based on the prompt “Please describe the hand posture of the teacher in the image”. The textual descriptions are transformed into high-dimensional features and subsequently integrated with video features to mitigate the problem of information loss within visual features arising from background similarity. Moreover, semantic analysis is applied to each class name, followed by the generation of corresponding textual feature vectors using a text encoder. The final prediction is obtained by calculating the cosine similarity between the fused features and the class name features.

Our contributions can be summarized as follows:

(1) We propose a fusion method that combines textual information from hand movements with video features to effectively compensate for the insufficiency of visual information, addressing the challenges of teacher gesture recognition in highly similar backgrounds.

(2) We propose a classification head network based on class name definitions, which performs classification by computing the similarity between visual features and textual features derived from class name definitions. The proposed method effectively smooths inter-class relationships and enhances intra-class feature consistency.

(3) Our proposed NLD-TGR method achieves state-of-the-art performance on the most representative TUB-G dynamic teacher gesture dataset, with an accuracy of 93.7%.

2. RELATED WORK

Teacher gesture recognition (TGR) is a fundamental task in intelligent educational. The performance of TGR models depends heavily on feature extraction. Recent studies in TGR have attempted to reduce the impact of background noise through skeletal point-based feature analysis. Chen[7] employed RTMPose[9] to extract teachers’ skeletal keypoint coordinates. The extracted skeletal sequences were then input into the MoGRU action recognition network for gesture classification. By leveraging keypoints from object detection and pose estimation algorithms, Wu[18] constructed a graph convolutional network for automatically identifying teachers’ gestures. However, these methods often overlook a significant characteristic of teacher gestures: their interaction with the environment. The same gesture, when associated with different entities such as multimedia resources, students, or teaching tools, may convey different meanings as shown in Figure 1a and 1b. Yet, methods based on skeletal points inevitably ignore this critical characteristic of teacher gestures.

To address the aforementioned issues, Wu[28] explored an gesture recognition method that leverages both RGB video and skeletal information, integrating them to enhance recognition accuracy. However, these methods also show clear limitations. Models based on ImageNet pre-trained weights perform poorly in extracting teacher gesture features. This is because their feature representation capabilities are not optimized for gesture characteristics in educational settings. Moreover, the backgrounds of teacher gestures in classrooms are generally highly similar, a point that has not been effectively considered. In response to these challenges, this paper attempts to use the large-scale pre-trained CLIP[20] model to extract spatial features of teacher gestures. GPT-4o is used to generate textual descriptions of hand movements based on prompts of the teacher’s hand postures. The extracted video frame features are combined with the encoded textual description features, allowing for a more comprehensive understanding and representation of the teacher’s hand gestures by integrating visual features and textual descriptions.

Inter-class similarity and intra-class variability are a noteworthy challenge in teacher gesture recognition. Teacher gestures frequently comprise numerous similar movements or shapes, posing significant challenges to model discrimination. Song[15] proposes using a deep neural network to learn feature embeddings that minimize intra-class distance while maximizing inter-class distance, generating class prototypes through feature mean calculation for improved comparison. Furthermore, Zuo[32] introduces a language-aware label smoothing technique. This generates soft labels for each training sample, effectively alleviating inter-class similarity issues. In this study, class names are interpreted and mapped into a high-dimensional text feature space. They are then compared with video features for similarity calculation to achieve classification prediction. The approach utilizes the multimodal alignment features of CLIP, allowing class names to not only indicate categories, but also reveal diverse gestures within the class. By using similarity classification, the boundaries between classes are effectively soften.

3. METHODOLOGY

The overview of our NLD-TGR recognition framework is illustrated in Figure 2. Our framework primarily comprises three components: 1) a video-text fusion network which fuses video features with descriptive text features generated from video frames. 2) a temporal information mining decoder encompasses the extraction of both short-term and long-term temporal information. 3) a head network contains the definition of the classification head and the implementation methods for classification.

3.1 Video-Text Fusion Network

In this study, we observed that teacher gestures exhibit a high degree of similarity in their background. Therefore, we aim to leverage the semantic information from hand movements to enhance the discriminability among different gestures. Considering that concatenation has the ability to fully conserve the initial information of visual and textual features, it does not cause the inhibition or augmentation of specific features that might be brought about by the weight distribution in the attention mechanism, a video-text fusion network is proposed, which integrates visual features with textual representations of hand movements to obtain a more comprehensive feature representation, as illustrated in Figure 2.

To generate robust spatial features, the CLIP model based on contrastive learning is employed. Meanwhile, GPT-4o is utilized to generate detailed text descriptions for video frames. Specifically, given a teacher gesture video $V \in \mathbb{R}^{T \times H_K \times W_K}$ with T frames and a spatial resolution $H_K = W_K$. Initially, video features $\boldsymbol{F_{video}} \in \mathbb{R}^{T \times N_1}$ are extracted using CLIP, where T represents the number of frames and N_1 represents the feature dimension. Then using the generative capabilities of GPT-4o, a prompt like "describe the hand posture of teacher in the image" is provided, resulting in T descriptions of the teacher's hand movements. For instance, one praise gesture frame description reads: "The palms of teacher are together, seemingly clapping". Subsequently, the text is encoded using the sentence-level embedding model, Sentence-BERT[21], to obtain descriptive features $\boldsymbol{F_{text}} \in \mathbb{R}^{T \times N_2}$, where T represents the number of frames and N_2 represents the feature dimension. The final fusion feature $\boldsymbol{F_{fusion}} \in \mathbb{R}^{T \times (N_1+N_2)}$ representation is formed by combining $\boldsymbol{F_{video}} \in \mathbb{R}^{T \times N_1}$ and $\boldsymbol{F_{text}} \in \mathbb{R}^{T \times N_2}$, thereby further enhancing the expressiveness and distinctiveness of the teacher gesture features. This approach allows the model to extract critical but often overlooked information, significantly improving its discriminative performance in teacher gesture recognition with highly similar backgrounds.

3.2 Temporal Information Mining in Spatial Features

Teacher gestures, a distinct form of visual language in educational contexts, primarily convey semantic information through various hand shapes and movements. Although the CLIP model demonstrates excellent performance in representing spatial features, it lacks the necessary capability to capture temporal dependencies. Although the Transformer decoder has been proven effective in aggregating global temporal information via weighted feature fusion [25, 11], it is still necessary to capture temporal features at both global and local levels to accurately model teacher gestures. This dual-level strategy is crucial for accurately capturing the intrinsic temporal characteristics of dynamic gestures. In our framework design, the extraction of temporal information is performed both before and after feature fusion to capture hierarchical temporal features more comprehensively.

Temporal decoding before fusion: 1D temporal convolution effectively captures short-term and local patterns. GRU, a variant of recurrent neural networks, excels in capturing long-range dependencies in sequential data. It retains and extracts long-term contextual information. Through weighted summation, the integration of these two methods combines temporal features at different scales. This fusion enables the model to comprehensively understand time series data. Formally, the features encoded by this temporal decoder are denoted as Y_{time}:

$$Y_{\text{time}} = \alpha(t) \cdot \text{GRU}(X) + \beta(t) \cdot \text{Conv1D}(X) + \gamma \cdot R(X) \quad (1)$$

Where $\alpha(t) = \frac{e^{\lambda_1 t}}{e^{\lambda_1 t} + e^{\lambda_2 t}}$ and $\beta(t) = \frac{e^{\lambda_2 t}}{e^{\lambda_1 t} + e^{\lambda_2 t}}$ are time-dependent dynamic weight functions, with λ_1 and λ_2 as learning parameters, allowing the model to adaptively adjust the contributions of GRU and Conv1D based on the time point t in the sequence. $R(X)$ is a regularization term, used to control model complexity and prevent overfitting. γ is the coefficient that controls the impact of the regularization term. The decoder is inserted between each transformer block of the main CLIP backbone.

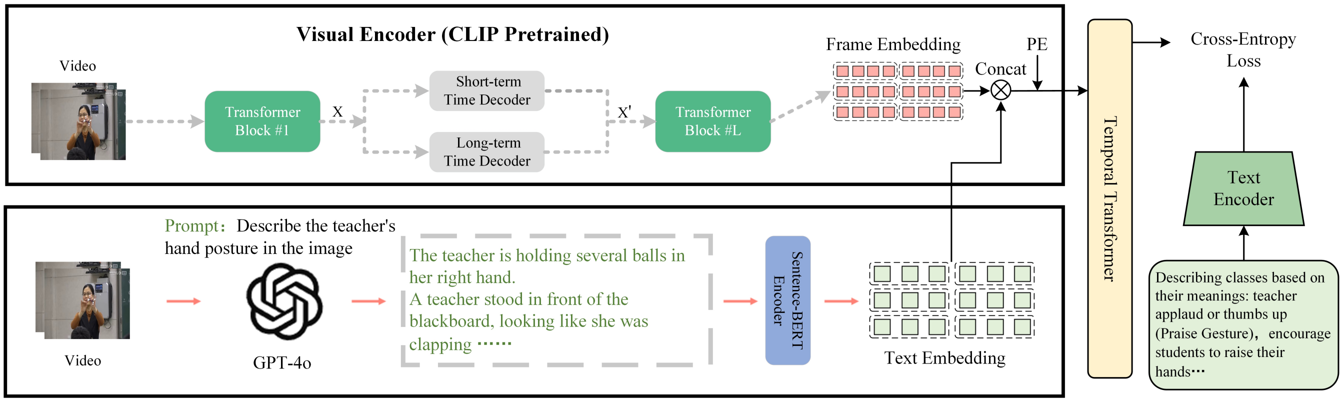
Temporal decoding after fusion: Positional embeddings are integrated into the fused feature matrix $\boldsymbol{F_{fusion}}$ to compensate for the self-attention mechanism's lack of intrinsic positional discrimination. These embeddings enable precise modeling of the temporal dynamics by providing spatial context, which is critical for distinguishing the sequential order of frames in the video. Formally, the computation of the video features post fusion and positional encoding can be denoted by the following enhanced expression:

$$Y_{\text{video}} = \text{MHA}\left(\sum_{i=1}^{T} \boldsymbol{F}^{(i)}_{\boldsymbol{fusion}} + \text{PE}(i, P)\right) \quad (2)$$

where MHA signifies the MultiHead Attention mechanism. $\boldsymbol{F}^{(i)}_{\boldsymbol{fusion}}$ denotes the i-th frame feature vector from the fused feature matrix. $\text{PE}(i, P)$ represents the positional encoding for frame i based on its position P in the sequence, enhancing the model's ability to understand temporal placement.

3.3 Head Network

In studies on teacher gesture recognition in classroom environments, traditional fully connected models that rely on one-hot encoding[3, 7, 28] for class label representation face two significant limitations. (1) the association between different classes is overlooked. (2) the issue of intraclass variability has not been adequately addressed. To address these limitations, class names are redefined and then embedded into textual features, and classification prediction is performed by calculating the similarity between video feature embeddings and textual feature embeddings.



Visual Encoder (CLIP Pretrained)
Video
Transformer Block #1
X
Short-term Time Decoder
Long-term Time Decoder
X'
Transformer Block #L
Frame Embedding
PE
Concat
Temporal Transformer
Cross-Entropy Loss
Text Encoder
Describing classes based on their meanings: teacher applaud or thumbs up (Praise Gesture)，encourage students to raise their hands···
Prompt：Describe the teacher's hand posture in the image
Video
GPT-4o
The teacher is holding several balls in her right hand.
A teacher stood in front of the blackboard, looking like she was clapping ······
Sentence-BERT Encoder
Text Embedding

Figure 2: Natural Language-Driven Teacher Gesture Recognition(NLD-TGR). We first use GPT-4o to generate textual descriptions of gestures based on prompts "describe the teacher's hand gestures in the frame". These textual descriptions are then mapped into high-dimensional features and combined with video features to extract information that might be overlooked in the visual data. Next, the interpreted category names are encoded into the text feature space and compared with the previously fused features through similarity computation to achieve classification.

Specific, we first define class names based on the guidelines below: Teacher gestures exhibit distinct expressive styles across different individuals, leading to intraclass variability. Thus, it is essential to reflect this variability in the textual descriptions. Take the category "explanation gesture" as an example. Teachers often use spontaneous and specific gestures during teaching to enhance instructional effectiveness or convey emotions. For instance, some teachers wave their arms while explaining complex concepts or support their chin in thought. Such actions can be characterized as "A video of a teacher waving their arms or resting their chin". The definitions of class names can be flexibly adapted to specific behaviors within each category. This approach effectively distinguishes between classes, without completely severing the connections between classes like one-hot labels do. Incorporating specific behaviors into class name definitions highlights the unique characteristics of each category and reflects potential inter-category relationships.

For the definition of class names for each category, we employ a sentence representation learning framework to map them into high-dimensional features. Specifically, Sentence-BERT processes the class name annotations for N categories to derive a D-dimensional feature for each. This approach facilitates a more intuitive representation of the semantic features of the categories within the feature space. Consequently, we obtain a feature matrix $Y_{class} \in \mathbb{R}^{N \times D}$. The n-th row of Y_{class}, denoted as Y^n, represents the textual features of the class name for the n-th category.

Finally, we use the cross-entropy as the loss function, given the final video feature Y_{video} and N class name textual features $Y_{class} \in \mathbb{R}^{N \times D}$, we compute the cosine similarity, apply softmax, and then compute the cross-entropy loss.

$$s_i = \frac{Y_{\text{video}} \cdot Y_{\text{class}_i}}{\|Y_{\text{video}}\| \|Y_{\text{class}_i}\|}, \quad \text{for } i = 1, 2, \ldots, N \tag{3}$$

$$L = -\sum_{i=1}^{N} y_i \log \left(\frac{e^{s_i}}{\sum_{j=1}^{N} e^{s_j}} \right) \tag{4}$$

where y_i is the ground truth label, p_i is the predicted label, and L is the cross entropy loss. The proposed classification method softens inter-class boundaries effectively. It also utilizes class name definitions to enhance the understanding of intrinsic variability within categories.

4. EXPERIMENTS

4.1 Dataset and Implementation Details

Dataset: We evaluate our method on the TBU-G dataset, which is an extended version of the Teacher Behavior Understanding (TBU) dataset [5]. To the best of our knowledge, TBU-G is the largest publicly available dynamic teacher gesture dataset available as of now. In contrast to previous teacher gesture datasets that primarily emphasized hand posture variations, TBU-G is classified based on the actual meaning and interactive nature of teacher gestures, making it better aligned with the requirements of real-world teaching environments, as illustrated in Figure 1. This dataset comprises 8 categories, with a total of 2,908 video clips ranging in duration from 1 to 10 seconds. The dataset provides a comprehensive representation of the complexity and diversity of teacher gestures in authentic classroom settings.

Implementation Details: Our experiments are conducted using a machine equipped with four NVIDIA RTX 4090 GPUs, each with 24 GB of memory. For a given video, we begin by uniformly sampling T frames (e.g., 8, 16, 32) throughout its duration. We then use a Vision Transformer (ViT) as the video encoder and GPT-4o as the video frame description generator. To enhance training efficiency, the generated descriptions are stored as weight files, allowing subsequent training processes to utilize these pre-saved weights directly, thereby minimizing redundant resource consumption. Additionally, sentence-BERT is employed as the text encoder. During training, we set the learning rate to 5×10^{-5} and utilize the AdamW optimizer. We use Top-1 Accuracy as the primary metric to assess the model performance on the test set. To balance accuracy and speed, we evaluate using only one clip per video. For efficiency, a center crop is applied during the evaluation process.

4.2 Comparison with State-of-the-art Methods

Table 1: Comparison with state-of-the-arts on TBU-G. *Keypoint-based* models refer to models that perform gesture recognition by extracting skeleton keypoints through publicly available models. *Appearance-based* models are those that rely on video features for recognition. *CLIP-based* models refer to those that first extract spatial feature information from videos using the CLIP model, and then perform gesture classification through temporal decoding.

Method	Input	Backbones	Pre-training	Top-1
Keypoint-based				
Spoter[3]	—	ViT-B/14	ImageNet	79.4%
VTN-PF[8]	—	ViT-B/14	ImageNet	77.7%
SLGTformer[23]	—	ViT-B/14	ImageNet	81.3%
Appearance-based				
TDN[27]	16×256^2	ResNet-50	ImageNet	83.4%
ViViT[2]	16×256^2	ViT-B/14	ImageNet	85.5%
I3D[6]	16×224^2	ResNet-50	ImageNet	69.9%
MvfNet[29]	16×224^2	ResNet-50	ImageNet	83.3%
S3D[31]	16×256^2	ResNet-50	ImageNet	74.3%
CLIP-based				
Text4Vis[30]	8×224^2	ViT-L/16	CLIP	89.6%
ST-Adapter[16]	16×256^2	ViT-L/16	CLIP	89.4%
EVL[11]	8×224^2	ViT-L/16	CLIP	90.2%
Ours	$\mathbf{8\times224^2}$	**ViT-L/16**	**CLIP**	**93.7%**

Table 1 presents the experimental results on the TBU-G dataset, comparing our approach with mainstream methods based on keypoint features, appearance features, and the large language model CLIP. The results clearly demonstrate that our method achieves state-of-the-art performance. Compared to keypoint-based methods, our approach shows significant advantages, with accuracy improvements of 12.4% over SLGTformer and a substantial 14.3% increase over Spoter. Skeleton-based recognition models do not achieve the same level of excellence in performance as observed in sign language recognition tasks. This is consistent with our expectation that relying solely on posture keypoint information is inadequate to fully reveal the true intentions behind the teacher's gestures. To recognize teacher gestures, it is necessary to consider the objects they interact with as part of the analysis. Furthermore, the proposed method clearly outperforms appearance-based recognition approaches in various aspects. For instance, it achieves a 10.3% increase in Top-1 accuracy compared to TDN, which utilizes ImageNet pre-trained weights. Moreover, even with a smaller spatial resolution (224 vs. 336), our method achieves a substantial improvement over ViViT, with a Top-1 accuracy of 93.7% compared to 85.5%, demonstrating its superior performance. Moreover, compared to the large-model methods based on CLIP, such as EVL and Text4Vis, our approach achieves performance improvements of 3.4% and 4.1%, respectively. This significant enhancement demonstrates that our model more effectively leverages the transfer learning capabilities of CLIP, thus improving its adaptability and performance in the task of teacher gesture recognition.

4.3 Ablation Studies

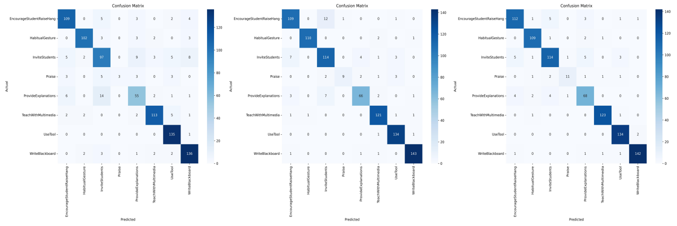
Table 2: Ablation studies on NLD-TGR. The *Base* model utilizes only positional encoding and multi-head attention mechanisms to capture temporal information, ultimately employing a fully connected layer for gesture classification. The *Text* variant incorporates a fusion network and uses class name interpretation as the classification head. Meanwhile, the *Time* model integrates a time decoder into the CLIP backbone, extracting both global and local temporal information.

Base	Base+Text	Base+Text+Time	Top-1%
✓	✗	✗	86.4%
✓	✓	✗	92.8%
✓	✓	✓	93.7%

To thoroughly analyze the contributions of each design component, we performed an extensive ablation study. As reported in Table 2, the base model utilizes only positional encoding and multi-head attention mechanisms to capture temporal information, ultimately employing a fully connected layer for gesture classification. Text-based assistance resulted in a 6.4% improvement in model performance. The combination of the class name definitions head network and text-video fusion network effectively enhanced gesture classification. These components leveraged the transfer learning capabilities of large-scale models more efficiently. Moreover, extending the CLIP backbone with global and local temporal decoders resulted in an additional 0.9% performance boost. These outcomes clearly demonstrate that the extraction of multi-level temporal information plays a vital role in advancing the model's overall performance.

Meanwhile, based on the confusion matrix shown in Figure 3, we analyzed the contributions of the model components in addressing challenges related to teacher gestures. In the base model, label misclassifications exhibit an approximately uniform distribution. This can be attributed to the high degree of similarity in background characteristics across various gesture categories. Additionally, the one-hot labels ignore the potential relationships between categories, making it difficult for the model to effectively extract key differences between categories. Consequently, the misclassification of the model approximates a random distribution. Taking the gesture "encourage student to raise hand" as an example, the traditional one-hot encoding method ignores the intrinsic relationships among categories, which often leads to misclassification into significantly different gesture categories. However, by incorporating text-assisted tasks, misclassifications are primarily concentrated within the "invitation" gesture category. This is due to the significant similarities in gesture between the two categories. The proposed model effectively identifies and exploits these underlying relationships, facilitating the extraction of more discriminative and representative features. The observed regularity in misclassification patterns underscores a significant enhancement in the model's learning capability.

Certain categories with lower classification accuracy, such as “praise” and “explain” gestures, exhibit considerable intra-class variability stemming from individual differences in teachers’ expression styles. This characteristic poses significant challenges for traditional models in accurately classifying complex and dynamic features. With the integration of text-assisted information, the accuracies for these two gesture categories improved respectively by 35.3% and 13.9%. This improvement can be attributed to the proposed head network design, which effectively learns and generalizes the individualized variations in these gestures, thereby significantly enhancing classification accuracy and overall performance. Furthermore, the integration of a multi-level temporal decoder into the backbone network resulted in measurable improvements in the recognition accuracy of temporal-context-dependent gestures, such as “praise” and “raise hand”. These findings further substantiate the efficacy of the proposed framework.



Base Base+Text Base+Text+Time

Figure 3: Confusion Matrix of the Ablation Study. Please zoom in for the best view.

5. LIMITATIONS AND FUTURE WORK

In this paper, class names are revised according to the observed characteristics of the dataset. These redefined class names are represented as high-dimensional text vectors and aligned with video feature vectors via similarity-based computation to achieve classification. However, this redefinition-based approach requires detailed dataset analysis, making it highly time-consuming, especially for large-scale datasets with numerous gesture categories, such as WLASL[10]. Additionally, optimizing class redefinitions demands extensive experiments and iterative fine-tuning, which further adds to the complexity and cost.

Future work will focus on the potential applications of natural language in improving gesture recognition. Analysis of the generated frame textual descriptions reveals that certain teacher gestures can be directly categorized based on their textual descriptions. Building on this observation, future work plans to incorporate a text classification branch as an important auxiliary module for gesture recognition tasks. Meanwhile, there is a remarkable correlation between teachers’ audio and gestures, so it can serve as another core modality and be integrated into the gesture recognition system. In addition, considering the computational costs of current generative models, future research will prioritize the use of a predefined dictionary of textual descriptions. Through a classification model, the most relevant textual descriptions for each frame will be selected from the dictionary and integrated with video features. Furthermore, subsequent studies aim to optimize the fusion mechanism between textual and video features to significantly enhance the effectiveness of natural language as a support in gesture recognition tasks. To address the bottleneck in the efficiency of class interpretation, an automated class definition tool will be developed. By leveraging natural language processing techniques and integrating them with the image or video features of the dataset, this tool will generate representative textual descriptions as class names, eliminating the need for manual and detailed analysis of the dataset.

In terms of the application in practical teaching scenarios, the gesture recognition model enables teachers to transcend spatial constraints and achieve interactive effectiveness equivalent to that of in-person classrooms. When specific gestures such as thumbs-up or pause gestures are executed, the system’s pre-programmed algorithms automatically trigger visual feedback on student terminals and regulate the teaching process, thereby ensuring the immediacy and synchronization of remote teaching interactions.

6. CONCLUSIONS

In this study, we introduce a natural language-driven teacher gesture recognition (NLD-TGR) framework, which leverages natural language information with the aim of enhancing the performance of teacher gesture recognition. Specifically, a video-text fusion network is designed. Initially, a generative model is utilized to derive textual descriptions of gesture postures from video frames based on prompts. These descriptions are then mapped into high-dimensional features, which are fused with the video features. Subsequently, a multi-dimensional temporal decoder is proposed to extract temporal information from the spatial features. Ultimately, accurate predictions are achieved by calculating the cosine similarity between the fused features and the textual features representing category names. Experimental results demonstrate that this approach surpasses the state-of-the-art methods on the teacher gesture dataset.

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] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[3] M. Boháček and M. Hrúz. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF winter conference on*

applications of computer vision, pages 182–191, 2022.

[4] R. Bojorque and F. Pesántez-Avilés. Academic quality management system audit using artificial intelligence techniques. In *Advances in Artificial Intelligence, Software and Systems Engineering*, pages 275–283. Springer, 2020.

[5] 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.

[6] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[7] Z. Chen, W. Huang, H. Liu, Z. Wang, Y. Wen, and S. Wang. St-tgr: Spatio-temporal representation learning for skeleton-based teaching gesture recognition. *Sensors*, 24(8):2589, 2024.

[8] M. De Coster, M. Van Herreweghe, and J. Dambre. Isolated sign recognition from rgb video using pose flow and self-attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3441–3450, 2021.

[9] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023.

[10] D. Li, C. Rodriguez, X. Yu, and H. Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020.

[11] Z. Lin, S. Geng, R. Zhang, P. Gao, G. De Melo, X. Wang, J. Dai, Y. Qiao, and H. Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, pages 388–404. Springer, 2022.

[12] Y. Liu, L. Chen, and Z. Yao. The application of artificial intelligence assistant to deep learning in teachers' teaching and students' learning processes. *Frontiers in Psychology*, 13:929175, 2022.

[13] Meng. Research on intelligent recognition of teaching gestures for smart classroom. Master's thesis, Central China Normal University, 2022.

[14] T. Ngo, L. Unsworth, and M. Herrington. Teacher orchestration of language and gesture in explaining science concepts in images. *Research in Science Education*, 52(3):1013–1030, 2022.

[15] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

[16] J. Pan, Z. Lin, X. Zhu, J. Shao, and H. Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022.

[17] S. Pang, S. Lai, A. Zhang, Y. Yang, and D. Sun. Graph convolutional network for automatic detection of teachers' nonverbal behavior. *Computers and Education: Artificial Intelligence*, 5:100174, 2023.

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

[19] Z. Peng, Z. Yang, J. Xiahou, and T. Xie. Recognizing teachers' hand gestures for effective non-verbal interaction. *Applied Sciences*, 12(22):11717, 2022.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[21] N. Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[22] W.-M. Roth. Gestures: Their role in teaching and learning. *Review of educational research*, 71(3):365–392, 2001.

[23] N. Song and Y. Xiang. Slgtformer: An attention-based approach to sign language recognition. *arXiv preprint arXiv:2212.10746*, 2022.

[24] L. Valenzeno, M. W. Alibali, and R. Klatzky. Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology*, 28(2):187–204, 2003.

[25] A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[26] J. Wang, T. Liu, and X. Wang. Human hand gesture recognition with convolutional neural networks for k-12 double-teachers instruction mode classroom. *Infrared Physics & Technology*, 111:103464, 2020.

[27] L. Wang, Z. Tong, B. Ji, and G. Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1895–1904, 2021.

[28] D. Wu, J. Chen, W. Deng, Y. Wei, H. Luo, and Y. Wei. The recognition of teacher behavior based on multimodal information fusion. *Mathematical Problems in Engineering*, 2020(1):8269683, 2020.

[29] W. Wu, D. He, T. Lin, F. Li, C. Gan, and E. Ding. Mvfnet: Multi-view fusion network for efficient video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2943–2951, 2021.

[30] W. Wu, Z. Sun, and W. Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2847–2855, 2023.

[31] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[32] R. Zuo, F. Wei, and B. Mak. Natural language-assisted sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14890–14900, 2023.