International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2025): 8.036

STQNet-VL: Vision-Language Guided Multi-Modal Learning for Multi-View Soccer Foul Recognition

Liangyu Chen[1], Zhouyou Dong[2]

[1]North China Electric Power University, School of Control and Computer Engineering, Changping District, Beijing, China
Email:*120232227074[at]ncepu.edu.cn*

[2]North China Electric Power University, School of Control and Computer Engineering, Changping District, Beijing, China
Email:*120232227083[at]ncepu.edu.cn*

Abstract: ***This paper addresses the problem of multi-view soccer foul recognition, which aims to identify foul actions and classify their severity from broadcast videos. Existing methods rely solely on visual information and often suffer from ambiguous predictions under occlusion, motion blur, and controversial contact scenarios. To overcome these limitations, we propose STQNet-VL, a vision-language guided multi-modal framework built upon the spatial-temporal query network (STQNet). The model first employs a large vision-language model to generate fine-grained textual descriptions of player interactions and contact regions. The descriptions are encoded using a pre-trained CLIP text encoder to ensure alignment with visual features. A transformer-based multi-modal feature fusion module is then designed to enable deep cross-attention interaction between visual and textual representations. Experimental results on the SoccerNet-MVFouls dataset demonstrate that STQNet-VL achieves 53% BAact, 45% BAsev, and 49% overall balanced accuracy, outperforming prior state-of-the-art methods and improving balanced accuracy by 2% over STQNet-Large. These results validate the effectiveness of integrating language priors for robust multi-view sports action understanding.***

Keywords: multi-view action recognition; vision-language models; transformer-based fusion; cross-modal learning; sports video analysis; Soccer Net-MVFouls; spatial-temporal modeling; class imbalance learning

1. Introduction

In sports video analysis and understanding, the intelligent analysis and understanding of sports actions in complex scenarios stand as core research topics. As illustrated in Figure 1, soccer foul recognition aims to identify fouls in soccer matches using multi-view videos, including foul action recognition and foul severity classification. Multi-view videos are essential for soccer foul recognition, as they can provide comprehensive 3D scene information and capture fine-grained player action details from diverse perspectives. Existing multi-view-based methods like VARS[0] extract spatial-temporal features from multi-view soccer videos via pre-trained visual encoders and fuse these features through pooling strategies for foul action and severity classification. However, such methods solely rely on visual information, which easily leads to ambiguous recognition results in the presence of severe player occlusion, motion blur caused by high-speed movement, and controversial actions on the edge of refereeing standards.

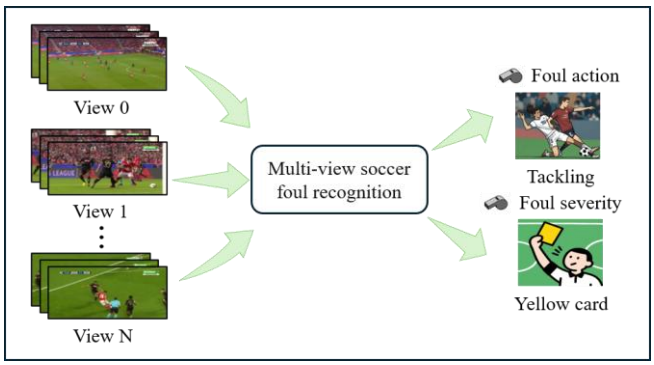

Figure 1: Multi-view soccer foul recognition task

In recent years, with the rapid development of Large Vision-Language Models (VLMs) and large-scale image-text contrastive pre-training techniques, cross-modal representation learning has demonstrated remarkable generalization ability and cognitive level in general computer vision understanding and reasoning tasks. Early vision-language pre-trained models were typified by CLIP[3]. Through pre-training on large-scale image-text contrastive datasets, CLIP constructed a unified visual-text feature space and achieved efficient semantic mapping between images and text, providing high-quality feature initialization for various subsequent cross-modal tasks. On this basis, researchers further optimized model architectures and pre-training strategies, proposing a series of high-performance VLMs. Among them, BLIP-2[4] adopts a distinctive two-stage pre-training strategy and realizes efficient connection between frozen visual encoders and large language models via the lightweight Q-Former Transformer module, effectively solving the problem of semantic alignment between visual and text features. In addition, LLaVA[5] strengthens model robustness and generalization by introducing an adaptive training mechanism that dynamically adjusts contrastive learning weights for different tasks. Qwen2.5-VL exhibits exceptional performance in capturing temporal dynamic information of videos and recognizing fine-grained visual details, owing to its advantages such as raw-resolution image processing and multi-modal Rotary Position Encoding (MRoPE), offering new technical insights for video-oriented cross-modal tasks.

As a highly condensed knowledge carrier abstracted by advanced human logic, the language modality can provide extremely rich and explicit prior contextual constraints in the form of natural language. Compared with visual pixel tensors with abundant background noise, textual descriptions can directly clarify the main interacting subjects and specific action contact details in video frames. Therefore, by using

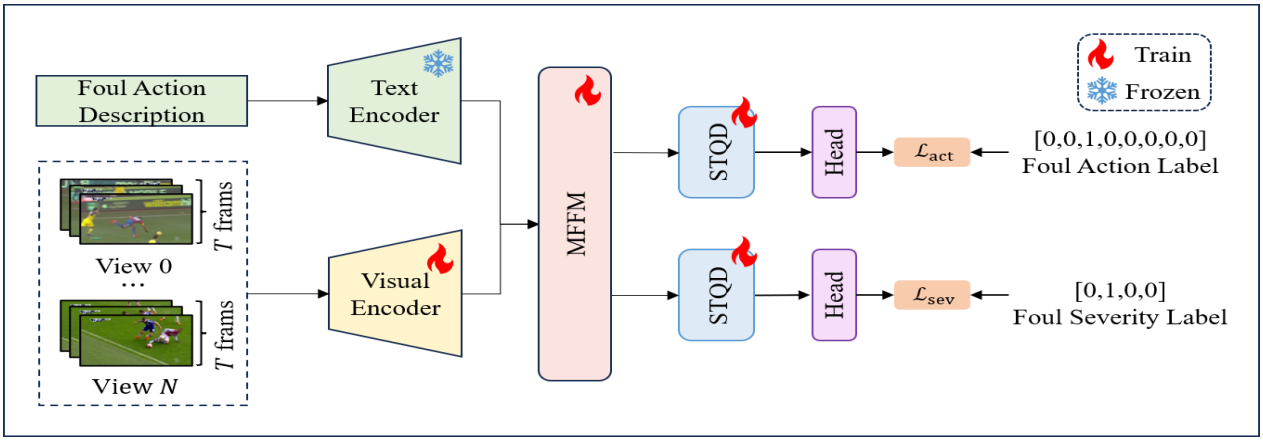

Figure 2: Overall framework of the proposed STQNet-VL

language models to parse complex sports action scenes, the explicit semantic information of text can be harnessed to guide, supplement and even rectify visual features, which can effectively compensates for the inherent limitations of pure visual recognition frameworks.

Based on the above analysis, this paper further explores the application of the text modality on the basis of STQNet[2], and proposes a vision-language model guided multi-modal multi-view soccer foul action recognition model named STQNet-VL. The main contributions of this paper are summarized as follows: a) Introduces a powerful Large Vision-Language Model to generate fine-grained text descriptions of player actions and behaviors in the dataset. b) Adopts the CLIP text encoder to map the unstructured natural language descriptions into text features, and designs a Transformer-based cross-modal feature fusion module (MFFM) to effectively fuse visual and text features. c) Experiments on the benchmark dataset demonstrate significant improvements over other methods.

2. Method

The framework of proposed STQNet-VL is shown in Figure 2. Based on the former state-of-the-art purely visual multi-view soccer foul recognition model STQNet, the proposed method introduces large-scale Vision-Language Models and a cross-modal feature fusion mechanism. The overall pipeline of the network architecture is mainly decoupled into three core stages: first, VLM-based foul action description generation, which generates foul action descriptions in textual form by leveraging the prior knowledge of large models; second, the feature extraction and multi-modal fusion stage, where a pre-trained text encoder is used to extract text features, and a Transformer-based fusion module (Multi-modal Feature Fusion Module, MFFM) is designed to achieve deep cross-attention interaction between visual and text features; and finally, dual-branch decoding and multi-task classification prediction based on the fused features.

2.1 VLM-based Foul Action Description Generation

To obtain high-quality textual modality for subsequent multimodal feature fusion, large-scale Vision-Language Models (VLMs) with strong generalization and complex-scenario reasoning capabilities are employed to generate detailed descriptive information of soccer foul actions.

In this paper, the Qwen2.5-VL-7B model is adopted to generate text descriptions of foul actions in soccer videos. The model's visual encoder reuses the weights of the Vision Transformer (ViT) pre-trained on large-scale image-text pairs in the Qwen2-VL model. The input to Qwen2.5-VL consists of two parts: one is the prompt, and the other is the video input. A prompt is a specific instruction that informs the large model of the task to perform and the required output format. The prompt is input into the model together with the video in textual form, guiding the model to quickly locate the target according to the prompt and generate responses corresponding to it. In soccer matches, foul actions often heavily depend on the player's body contact regions. For instance, pulling and pushing usually occur in the upper body, while tackling and standing tackle are more likely to take place in the lower body.

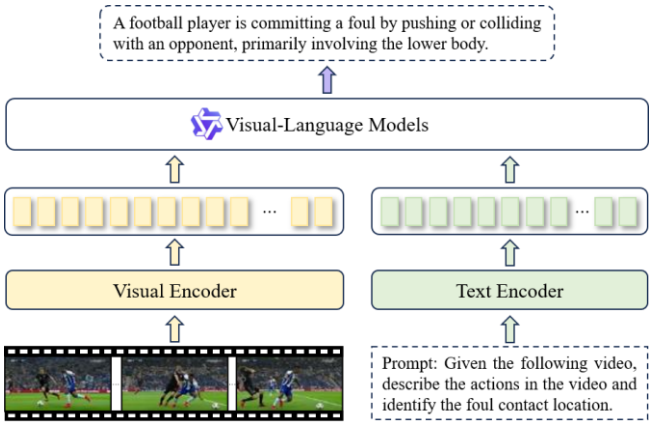

Figure 3: The workflow of Foul Action Description generation by the Qwen2.5-VL Model

Therefore, the prompt is designed as: Given the following video, describe the actions in the video and identify the foul contact location, requiring the language model to describe the actions in the video and explicitly specify whether the foul contact region is located in the upper body or lower body. The video input is processed through the visual encoder provided by Qwen2.5-VL to extract spatial-temporal features and visual semantic information between video frames from video clips. After concatenating the text feature encoded from the prompt and the visual features extracted from the video, the Qwen2.5-VL model fully leverages the feature decoding and

generation capabilities of the pre-trained large-scale vision-language model to generate fluent and contextually consistent text descriptions, as shown in Figure 3.

However, it is worth noting that while these generated descriptions provide rich linguistic priors, the VLM may occasionally produce semantic noise or "hallucinations" due to inherent visual ambiguities such as severe occlusion or motion blur in soccer matches. To address this, our framework ensures robustness through the subsequent Multi-modal Feature Fusion Module (MFFM). As detailed in Section 2.3, the MFFM utilizes an asymmetric cross-attention mechanism that allows the model to dynamically filter out inconsistent linguistic noise by querying the text features using reliable visual cues, thereby maintaining high recognition accuracy even with suboptimal textual inputs.

It should be explicitly noted that the generation of these textual descriptions is performed offline and the resulting text remains fixed before the model training begins. The Qwen2.5-VL-7B model serves as a pre-processing module rather than an end-to-end component of the training pipeline. Consequently, no gradients are backpropagated through the VLM during the optimization of STQNet-VL. This decoupled approach allows the framework to leverage advanced linguistic priors without the prohibitive computational cost of fine-tuning a large-scale vision-language model during the multi-modal fusion stage.

2.2 Text Feature Extraction

After extracting the natural language descriptions of actions in videos via Vision-Language Model, it is necessary to align and fuse these action descriptions with visual feature. How to bridge the gap between the visual and linguistic modalities constitutes a key and challenging problem. After obtaining the text descriptions of soccer foul actions, the most straightforward approach is to employ the text encoder provided by the same VLM to further extract text feature. Nevertheless, this work does not adopt such a scheme for text feature extraction; instead, the CLIP text encoder is selected for this purpose, which is mainly designed for the consideration of multi-modal feature space alignment. In the architecture of STQNet, the visual feature is extracted based on the CLIP text encoder. Therefore, if the text encoder provided by the VLM is adopted, the extracted text features will reside in the VLM-specific latent vector space, which is inherently different from the visual features extracted by the CLIP text encoder in the semantic space. As the features of the two modalities lie in distinct feature spaces, effective subsequent fusion becomes rather difficult.

To achieve simple and efficient alignment between text feature and visual feature, this work chooses the CLIP text encoder, which shares the same origin as the visual encoder, to extract the text features from the descriptions of soccer foul actions. To bridge the modality gap, we employ the pre-trained CLIP text encoder E_{CLIP} to map the natural language sequence input $S = \{w_1, w_2, ..., w_L\}$ into a high-level text feature representation $F_{text} \in \mathbb{R}^{B \times D}$. This encoding process yields a feature vector where the joint dimension D is strictly aligned with the visual latent space. By transforming unstructured descriptions into these aligned semantic embeddings, the module provides explicit prior constraints that are essential for guiding the subsequent multi-modal fusion stage.

2.3 Multi-modal Feature Fusion

The interactive fusion of independently extracted visual and textual features is the core step to improve the accuracy of soccer foul recognition. Visual feature contains abundant spatial-temporal dynamic details of match actions, while textual feature provides highly abstract global semantic summaries of foul behaviors. Simple feature fusion operations such as concatenation or element-wise addition cannot fully exploit the intrinsic correlations between the two modalities at different semantic levels. To address this issue, this paper designs and implements a Multi-modal Feature Fusion Module (MFFM) based on a transformer decoder architecture, which realizes dynamically weighted fusion of visual and textual features. MFFM is composed of K multi-modal feature fusion layers, the structure is shown in Figure 4.

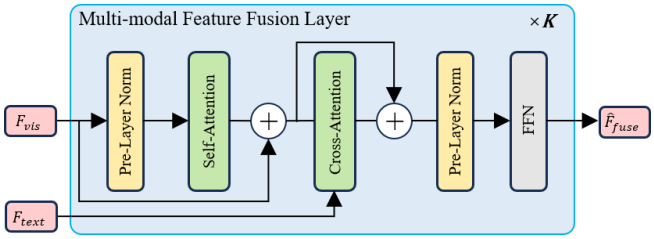

Figure 4: Structure of Multi-modal Feature Fusion Module (MFFM): A stack of K multi-modal feature fusion layers.

The Multi-modal Feature Fusion Module (MFFM) integrates visual features $F_{vis} \in \mathbb{R}^{VT \times B \times D}$ and expanded text features $F_{text} \in \mathbb{R}^{1 \times B \times D}$ through a stack of transformer-decoder layers. Each layer updates the representation by first modeling spatial-temporal dependencies via Multi-Head Self-Attention (MHSA) and then performing asymmetric cross-attention where the visual features query the text features. This operation produces a fused feature tensor $F_{fuse} \in \mathbb{R}^{1 \times B \times D}$ that effectively filters irrelevant background noise while rectifying ambiguous visual signals with explicit semantic guidance from the VLM. The process is formulated as:

$$F_{fuse} = \text{CrossAttn}(\text{MHSA}(F_{vis}), F_{text}, F_{text}) \quad (1)$$

Consequently, the resulting multi-modal representation provides a more robust and decisive evidence base for the subsequent classification heads.

2.4 Loss Function

Since the SoccerNet-MVFouls dataset used in this paper suffers from severe class imbalance—for instance, in the task of foul severity classification, the number of red card fouls is far fewer than that of non-card fouls—the model continues to adopt the Weighted Cross-Entropy (WCE)[6] loss from STQNet for optimization training to enhance the recognition capability for minority classes.

To alleviate the model's neglect of minority severe fouls caused by the long-tailed distribution in the dataset, a static

class weight coefficient w_c is introduced and the loss function is constructed as follows:

$$L_{WCE} = -\sum_{i=1}^{B}\sum_{c=0}^{C-1} w_c y_{i,c} \log\left(\frac{\exp(z_{i,c})}{\sum_{k=0}^{C-1}\exp(z_{i,k})}\right) \quad (2)$$

where B denotes the batch size, C is the total number of classes, $y_{i,c}$ represents the ground truth one-hot label of sample i belonging to class c, and $z_{i,c}$ is the Logits value output by the classification head. This loss function is applied to the action classification branch L_{act} and the severity classification branch L_{sev} respectively, and the final joint loss function L_{joint} can be obtained as follows:

$$L_{joint} = L_{act} + L_{sev} \quad (3)$$

3. Experiments

3.1 Dataset

SoccerNet-MVFouls[0] is a publicly available multi-view soccer foul classification dataset. It contains 3,901 annotated actions extracted from 500 European top-league matches spanning three seasons (2014–2017). Each action is provided with at least two broadcast camera views and one replay view, and is manually annotated by a professional referee with over six years of experience and more than 300 official matches officiated. The dataset defines eight foul action categories: *Standing Tackle*, *Tackle*, *Holding*, *Pushing*, *Challenge*, *Dive*, *High Leg*, and *Elbowing*. In addition, foul severity is categorized into four levels: *No Foul*, *Minor Foul*, *Yellow Card Foul*, and *Red Card Foul*. In this work, all training data are used to train the proposed model, and the evaluation is conducted on all test data.

3.2 Implementation Details

Our STQNet-VL follows the experiment settings of STQNet, the network is implemented in PyTorch and trained on a NVIDIA A800 GPU. We initialize the transformer encoder with AIM ViT-B/16[7] pretrained on Kinetics400[8]. Specifically, during the training process, the visual encoder is fine-tuned to capture domain-specific spatial-temporal features of soccer fouls, while the CLIP text encoder is kept frozen to maintain the stability of the pre-trained multi-modal semantic alignment. STQNet-VL is optimized using AdamW with a weight decay of 0.05, an initial learning rate of 1e-6, and 5 warm-up epochs. We uniformly sample 32 frames per video, with a batch size of 1, and train for 60 epochs. The starting and ending frames are set to 60 and 92, respectively. Data augmentation includes random affine and perspective transformations, random rotation, color jitter, and random horizontal flipping. For a standard multi-view soccer foul video clip of approximately 5 seconds, the inference time is roughly 1.67 seconds on a single NVIDIA A800 GPU, demonstrating its potential for near real-time assistant referee applications.

During evaluation, we report both standard classification accuracy and balanced accuracy (BA) following 0. Standard accuracy quantifies the overall proportion of correctly predicted samples, while BA, defined as the mean recall across all classes, mitigates the bias caused by class imbalance. In the results, we denote these metrics as Acc_{act}, Acc_{sev}, BA_{act}, and BA_{sev}, respectively. To ensure the reliability and reproducibility of the results, all reported metrics for STQNet-VL are obtained by averaging the results of three independent training runs with different random seeds, accompanied by the standard deviation.

3.3 Comparison with State-of-the-art methods

We compare our STQNet-VL with state-of-the-art foul recognition methods, including VARS[0], X-VARS[9], MatchVision[10] and STQNet[2]. Notably, the visual encoder of MatchVision is originally trained on large-scale soccer data with LLM-assisted supervision and kept frozen during its fine-tuning on SoccerNet-MVFouls. For a fairer comparison, we additionally trained the encoder first on single-view SoccerNet-MVFouls data without LLM-assisted supervision and then fine-tuned MatchVision following its official setup on SoccerNet-MVFouls, denoted as MatchVision*.

The quantitative evaluation results are shown in Table 1, where STQNet-VL achieves 53%±0.3%, 45%±0.4%, and 49%±0.3% in terms of BA_{act}, BA_{sev} and BA, respectively, which are superior to all state-of-the-art methods. Compared with STQNet-Large, our model yields a 2% improvement on each of the three metrics mentioned above. This result underscores that integrating linguistic priors is far more parameter-efficient for capturing complex foul semantics than simply scaling up the visual backbone. While MatchVision exhibits a relatively good performance on Acc_{sev}, it lags significantly behind other methods in the remaining metrics. Furthermore, STQNet-VL (187.7M) reaches 73%±0.2% on Acc_{act}, which is a 9% increase over STQNet-Base (148.6M) with an additional parameter overhead of only 39.1M.

Table 1: Quantitative results on SoccerNet-MVFouls

Method	#Para	Acc_{act}	ACC_{sev}	BA_{act}	BA_{sev}	BA
VARS-ResNet[11]	40.0M	0.31	0.34	0.28	0.25	0.26
VARS-R(2+1)D[12]	47.6M	0.31	0.36	0.34	0.30	0.32
VARS-MviT[13]	49.0M	0.40	0.38	0.45	0.31	0.38
X-VARS[9]	7008.2M	0.62	-	-	0.35	-
MatchVision[10]	8216.4M	0.44	0.58	-	-	-
MatchVision*	142.5M	0.46	**0.60**	0.29	0.31	0.30
STQNet-Base[2]	148.6M	0.64	0.53	0.40	0.44	0.42
STQNet-Large[2]	431.4M	0.50	0.55	0.51	0.43	0.47
STQNet-VL	187.7M	**0.73±0.2%**	0.50±0.4%	**0.53±0.3%**	**0.45±0.4%**	**0.49±0.3%**

3.4 Ablation Studies

3.4.1 Ablation on VLMs

Vision-Language Model plays an extremely crucial role in our model. The accuracy of the generated action description text directly determines the guidance quality in cross-modal feature interaction. Since the SoccerNet-MVFouls dataset used in this paper provides explicit ground truth labels for the contact regions (upper body/lower body) of fouls, this paper proposes a more intuitive and rigorous metric to count the accuracy of the generated text descriptions in judging the action contact location (Action Contact Location Accuracy, ACLA).

Based on the above quantification criterion, this paper selects three state-of-the-art open-source VLMs for ablation comparison, namely LLaMA-3, LongVA, and Qwen2.5-VL. The experiment is divided into two stages: first, the three models are respectively used to infer video samples in the dataset to generate corresponding action descriptions, and their accuracy in predicting whether the contact region of foul actions occurs in the upper body or lower body is counted; then, these three groups of different text descriptions are respectively extracted as text feature and fed into the model designed in this paper for training and testing, and BA of soccer foul recognition is counted.

Table 2: Ablation on VLMs

VLM	ACLA	BA_{act}	BA_{sev}	BA
LlaMA-3	0.59	0.36	0.42	0.39
LongVA	0.64	0.40	0.45	0.43
Qwen2.5-VL	**0.70**	**0.53**	**0.45**	**0.49**

The specific experimental comparison results are shown in Table 2. The results clearly demonstrate that Qwen2.5-VL achieves the best performance on both core metrics. Based on these experimental results, the final model in this paper uniformly adopts Qwen2.5-VL as the default text feature generation model.

3.4.2 Ablation on the number of layers in MFFM

We analyze the influence of the number of layers in MFFM. As shown in the Table 3, the variant stacking two layers achieves the best performance, improving BA by 3% compared to one layer and by 2% to four layers. In detail, fewer layers limit interactions between visual feature and text feature, while more layers may cause overfitting and increase computational cost. Therefore, the variant with two layers provides the best trade-off between modeling capability and efficiency.

Table 3: Ablation on the number of layers in MFFM

# Layer(s)	BA_{act}	BA_{sev}	BA
1	0.49	0.44	0.46
2	**0.53**	**0.45**	**0.49**
4	0.51	0.43	0.47
6	0.49	0.41	0.45

3.4.3 Ablation on Pre-trained Weights for Encoder

The feature extraction capabilities of the visual encoder and text encoder directly determine the model performance. In this work, the AIM model is adopted as the visual encoder and the official pre-trained CLIP text encoder is used to ensure visual-text feature alignment. However, pre-trained models with different parameter scales yield notably different feature dimensions. Since the fusion module in this paper requires visual and text features to perform dot product and attention interaction in the same dimensional space, unifying their dimensions becomes an unavoidable issue.

To explore the optimal combination of pre-trained weights, four encoder pairs are designed, with a linear adapter introduced to upsample low-dimensional text features. The first combination uses AIM-Base and CLIP-Large, both outputting a dimension of 768; the second combination uses AIM-Large and CLIP-Huge, both outputting a dimension of 1024; the third combination uses AIM-Base and CLIP-Base, and since the text feature dimension is only 512, an adapter is used to expand its dimension to 768; the fourth combination uses AIM-Large and CLIP-Large, and the 768-dimensional text features are similarly expanded to 1024 dimensions via an adapter. The experimental results are shown in Table 4, where the combination of AIM-Base weights and CLIP-Large weights achieves the best foul recognition performance among all combinations.

Table 4: Ablation on encoders' pre-train weights

Visual Encoder	Text Encoder	Aligment Strategy	BA
AIM-Base(768)	CLIP-Large(768)	Direct	**0.49**
AIM-Large(1024)	CLIP-Huge(1024)	Direct	0.44
AIM-Base(768)	CLIP-Base(512)	Adapter	0.42
AIM-Large(1024)	CLIP-Large(768)	Adapter	0.40

4. Conclusion

This paper presented STQNet-VL, a vision-language guided multi-modal framework for multi-view soccer foul recognition. By integrating VLM-generated action descriptions and transformer-based cross-modal fusion into STQNet, the proposed method achieved 49% overall balanced accuracy on SoccerNet-MVFouls, improving performance over prior state-of-the-art approaches. The results demonstrate that structured language priors effectively enhance spatial-temporal reasoning under ambiguous visual conditions.

Nevertheless, the model performance remains dependent on the accuracy and consistency of VLM-generated descriptions. Future work will focus on joint optimization of visual and language components and more effective multi-task collaboration for action and severity classification.

References

[1] Held J, Cioppa A, Giancola S, et al. VARS: Video assistant referee system for automated soccer decision making from multiple views[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 5086-5097.

[2] PU M, DONG Z, TENG J, et al. Multi-view soccer foul recognition using spatial-temporal query network[J]. IEICE Transactions on Information and Systems, 2026: 2025EDL8056.

[3] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763.

[4] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International conference on machine learning. PMLR, 2023: 19730-19742.
[5] Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2023, 36: 34892-34916.
[6] Verdier C, Perriot S, Pachot A. Weighted cross-entropy to tackle overlapping in fraud detection[C]//Proceedings of the 2023 15th International Conference on Machine Learning and Computing. 2023: 211-216.
[7] Yang T, Zhu Y, Xie Y, et al. Aim: Adapting image models for efficient video action recognition[J]. arXiv preprint arXiv:2302.03024, 2023.
[8] Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset[J]. arXiv preprint arXiv:1705.06950, 2017.
[9] Held J, Itani H, Cioppa A, et al. X-vars: Introducing explainability in football refereeing with multi-modal large language models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 3267-3279.
[10] Rao J, Wu H, Jiang H, et al. Towards universal soccer video understanding[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 8384-8394.
[11] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[12] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 6450-6459.
[13] Fan H, Xiong B, Mangalam K, et al. Multiscale vision transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 6824-6835.