Rugby Scene Classification Enhanced by Vision Language Model

Naoki Nonaka[1] Ryo Fujihira[1] Toshiki Koshiba[1] Akira Maeda[2] Jun Seita[1]
[1]Advanced Data Science Project, RIKEN Information R&D and Strategy Headquarters
[2] Hakata Knee & Sports Clinic

Abstract

This study investigates the integration of vision language models (VLM) to enhance the classification of situations within rugby match broadcasts. The importance of accurately identifying situations in sports videos is emphasized for understanding game dynamics and facilitating downstream tasks like performance evaluation and injury prevention. Utilizing a dataset comprising $18,000$ *labeled images extracted at* 0.2*-second intervals from* 100 *minutes of rugby match broadcasts, scene classification tasks including contact plays (scrums, mauls, rucks, tackles, lineouts), rucks, tackles, lineouts, and multiclass classification were performed. The study aims to validate the utility of VLM outputs in improving classification performance compared to using solely image data. Experimental results demonstrate substantial performance improvements across all tasks with the incorporation of VLM outputs. Our analysis of prompts suggests that, when provided with appropriate contextual information through natural language, VLMs can effectively capture the context of a given image. The findings of our study indicate that leveraging VLMs in the domain of sports analysis holds promise for developing image processing models capable of incorpolating the tacit knowledge encoded within language models, as well as information conveyed through natural language descriptions.*

1. Introduction

Identifying situations in sports videos is fundamental for understanding the dynamics of sports and is intricately linked to various downstream tasks. Properly capturing the context within the footage enables not only immediate evaluations of specific game situations but also facilitates longer-term assessments, such as performance over a season. For instance, in football [1], it becomes possible to quantitatively assess aspects like passing accuracy or ball possessions automatically [16, 44]. Furthermore, when considering injury prevention in sports, classifying situations prone to injuries (such as contact in rugby [32, 34] or specific movements in baseball [36]) could be vital. Analyzing sports footage in this manner contributes to a more objective understanding of sports, enhancing our ability to evaluate performances and potentially mitigate risks associated with injuries.

Deep Neural Networks (DNNs) have significantly improved performance in areas where manual feature design is challenging by automatically acquiring the necessary features from training data. For instance, tasks such as image classification [8, 22, 47], object detection [11, 26, 42, 43], and pose estimation [4, 37, 49, 59] have been successfully tackled in the field of image processing. Beyond image processing, applications like natural language processing [2, 46, 58] and speech recognition [1, 12, 13] have also benefited from DNNs. In sports-related research, studies predominantly utilize models from image processing, focusing on tasks such as player localization and tracking [53, 62], ball localization [50, 51] and pose analysis [17, 31, 34], showcasing various applications. However, despite these advances, DNN solely trained with image data faces challenges such as the difficulty of incorporating prior knowledge into models.

In the field of natural language processing, it has been demonstrated that using large language models (LLMs) can achieve high performance on various downstream tasks with fewer data than training DNNs from scratch [3, 7, 39, 40]. Particularly, autoregressive language models such as GPT [3, 39, 40] offer versatility and generality, enabling a wide range of applications. For example, they exhibit capabilities such as solving specific tasks following pre-specified prompts or generating context-aware responses through in-context learning [28, 56, 57]. Owing to these capabilities, and the impressive performance of LLMs on commonsense reasoning benchmarks, several works leverage LLMs as a source of commonsense knowledge assuming LLMs embed implicit knowledge of the world [63]. Furthermore, models integrating language and vision have shown utility in general-domain image classification [27, 41, 65] or object detection [5, 25, 29, 64].

Building upon these achievements, this study aimed to

[1]Often reffered to as "soccer" in North America.

validate the effectiveness of VLM (Vision Language Models) when classifying scenes within sports match broadcasts. Specifically, we conducted scene classification using a dataset comprising $18,000$ labeled images extracted at 0.2-second intervals from a total of 100 minutes of randomly sampled rugby match broadcast footage. The scene classification tasks included binary classifications of contact plays (scrums, mauls, rucks, tackles, lineouts), rucks, tackles, lineouts, as well as multiclass classification to predict one of the assigned labels. Experimental results revealed that incorporating VLM outputs improved classification performance across all tasks compared to using only image data for classification.

This paper is organized as follows. First, Sec. 2 describes the related studies, and Sec. 3 and Sec. 4 describes details of data and models used for our system. Then, in Sec. 5, we explain the experimental setting and in Sec. 6 we explain the obtained results. Finally, discussion are given in Sec. 7 and conclusions and limitions are given in Sec. 8.

2. Related Works

Large Language Models (LLMs) have become a cornerstone in natural language processing research, with a growing trend towards even larger architectures, demonstrating exceptional performance across a range of downstream tasks such as sentence classification, question answering, sentiment analysis and commonsense reasoning [7, 40]. The LLMs have further demonstrated strong performance in task with few data settings [39], and possess the capability of in-context learning, allowing tasks to be inputted with minimal examples and no parameter updates [3, 56]. Furthermore, they exihit that the perfomance can be improved by giving well designed prompts [21, 24, 28, 60]. The observed phenomena indicate that LLMs trained on vast corpora of data acquire implicit knowledge, which can be leveraged to generate outputs that integrate this tacit understanding through natural language prompting.

Based on the advances in natural language domain, some studies have proposed models to incorpolate LLMs in vision domain. Several works, such as CLIP [41], ALIGN [19] and Florence [61] have successfully connected the vision and natural language modalities. Additionally, studies such as LLaVa [27] and MiniGPT4 [65], which combine LLM with vision, enable linguistic interactions with images through LLM. Moreover, incorplation of LLMs improved the performance of open-world object detections [5, 25, 29, 64] Such advancements in VLM suggest the potential to extract information from images based on linguistically described or LLM embedded knowledge.

On the other hand, in the field of sports data analysis, the emergence of DNNs has led to significant advancements. Studies utilizing DNNs for analysis span a wide range of sports including football [15, 51], rugby [32, 35], basketball [38, 50], ice hockey [52], skiing [9], baseball [36], table tennis [23, 54], and canoeing [55]. These studies include efforts to acquire positional information such as player or ball location and tracking [51, 53], evaluations of game content such as receiver decision-making and pass success/failure determination [16, 48], as well as analyses of movements using estimated pose information [17]. Moreover, there are studies focused on injury prevention and improving the safety of sports through analysis [34, 36]. These advancements have been facilitated by the elimination of the need for feature extraction with DNNs and the availability of pretrained models in the general image domain.

DNNs require a large amount of labeled data for training, and the quality of the model obtained is greatly influenced by the scale and quality of the dataset. In the domain of football, where extensive manual annotation is available through initiatives like SoccerNet [6, 10], competitions have led to the development of high-performance models. However, obtaining such data in the sports domain is not always straightforward. Therefore, there are studies focused on constructing and providing sports-specific datasets [18, 33, 45] and developing methods to efficiently collect data [30]. While acquiring large-scale datasets represents a promising approach, the associated costs are often prohibitive. Thus, in this work, we investigate an alternative direction by examining whether leveraging sports-related knowledge encoded within LLMs can enhance the performance of DNN models on rugby analysis tasks.

3. Data

To examine the efficacy of training rugby scene classifier with VLM, we prepared labeled dataset of rugby image using rugby match videos of Japanese elite league. A total of 366 videos corresponding to matches from three seasons of the Top League, an elite rugby league in Japan, from 2016 to 2018 seasons were used to prepare dataset. The original videos obtained were edited for broadcast on TV, and we resized all videos to height of 720 pixels and width of 1280 pixels. We randomly selected five matches and further randomly extracted video clips corresponding to ten minutes length from first and second halves of selected matches respectively.

Subsequently, we manually annotated static images extracted at 0.2 second intervals from the ten video clips randomly extracted from selected five matches. For all extracted static image, we gave the scene label corresponding to the playing situation in the image. The play situations were categorized into eleven labels: goal kick, normal kick, restart kick, ruck, lineout, maul, scrum, tackle, general play, out of play and replay mark for broadcasting [2]. Resulting number of labels from each video clip is shown in Tab. 1 and

[2]"Normal kick" indicates situations where ball was kicked during the course of the match. "General play" indicates situations where no kicking

Table 1. The number of scene labels assigned to manually annotated randomly extracted 10-minute segments from the first and second halves of five randomly selected matches.

Match ID	Half	Normal kick	Goal kick	Restart kick	Tackle	Ruck	Lineout	Maul	Scrum	General play	Out of play	Replay mark
1	First	275	0	146	230	476	203	59	188	607	769	47
	Second	131	0	26	141	396	24	0	571	410	1292	9
2	First	168	91	50	171	289	218	0	269	486	1228	30
	Second	52	0	17	198	408	18	0	611	589	1077	30
3	First	146	0	71	302	412	118	46	432	608	848	17
	Second	71	263	189	239	287	110	0	0	667	1145	29
4	First	139	144	35	307	332	137	0	101	637	1131	37
	Second	109	335	27	151	297	78	0	143	344	1497	19
5	First	91	153	95	142	183	60	0	235	391	1640	10
	Second	184	0	114	242	298	127	43	125	499	1333	35

example images of contact related labels (tackle, ruck, lineout, maul and scrum) are shown in Fig. 1. The total number of labeled images amounted to $3,000$ per video clip, resulting in a total of $30,000$ labeled images.

4. Model

To examine whether the performance of scene classification could be enhanced by employing a VLM, we utilized a model shown in Fig. 2. The model comprises three fundamental components: the VLM, the Image Encoder, and the Head module. This model takes both the image and text prompt as inputs. The VLM processes both the image and text inputs, while the Image Encoder specifically handles the image input. The outputs from both the VLM and Image Encoder are fed into the Head module, which in turn generates predictions for scene labels. In this study, only the parameters of the Image Encoder and the Head module were updated. The parameters of VLM were kept fixed, utilizing pretrained weights, and were not updated during the training process of the scene classification model.

We employed the LLaVa-7B model [27] as the VLM for our experiments. Regarding the Image Encoder component, we conducted preliminary experiments across various ResNet architectures, namely ResNet 18, 34, 50, 101, and 152 [14], to determine the most suitable structure for each task. For the Head module, we concatenated the outputs from the Image Encoder and VLM, followed by a linear layer[3], ReLU activation function, dropout regularization, and an additional linear layer. This ensured that the final output dimension corresponded to the number of target labels.

or contact is happening, for example if ball carrier was carrying a ball without being tackled the image is labeled as "general play".

[3] The linear layer takes vector of $D_{ie} + D_{vlm}$ as an input, where D_{im} and D_{vlm} is a dimension of an output vector from the Image Encoder and VLM.

5. Experiment

To verify the utility of VLM in rugby scene classification, we conducted five image classification tasks. For each of the five targeted tasks, we first determined the optimal baseline conditions without using VLM. Subsequently, we compared and evaluated suitable prompts for each task before finally conducting a comparison based on the presence or absence of VLM.

5.1. Data split

The manually labeled dataset comprised of five rugby match videos was divided into three subsets for training and evaluation of the model. To split the dataset, we took following two steps. First, one match was randomly selected from the five matches, and the image-label pairs obtained from the first and second halves of that match were designated as the test set. Second, from the remaining four matches, one match was chosen for the validation set, and the other three matches were used for the train set. This process was repeated four times, ensuring that each of the four matches served as the validation set once. One of the four train/validation sets was used for the optimization of baseline, prompt selection and hyperparameter tuning. Three remaining train/validation sets were used to train models for the final comparison. For the final comparison, the training of the models was independently conducted three times using the remaining train/validation sets. Each of the three resulting models was then applied to a common test set, and the average performance across these three runs was taken as the final evaluation metric.

5.2. Classification tasks for the evaluation

Rugby is a contact-intensive sport, and player collisions are closely associated with the occurrence of injuries. Therefore, this study set up a scene classification task focusing on

(a) Tackle

(b) Ruck

(c) Lineout

(d) Maul

(e) Scrum

Figure 1. Examples of image for each contact related class labels.

contact scenes. Specifically, among the five labels related to contact—tackle, scrum, lineout, maul, and ruck—we formulated a binary classification task where tackles, lineouts, and rucks observed in all ten videos were considered positive instances, while other labels were considered negative instances. Additionally, we conducted a binary classification task where any instances labeled with one of the five contact-related labels were considered positive, and the remaining instances were treated as negative (referred to as "contact"). Furthermore, a multi-class classification targeting the ten labels excluding replay marks was carried out (referred to as "multi-class"). For the evaluation of the multi-class classification task, we employed the weighted F1 metric, while the remaining four binary classification tasks were evaluated using the F1 score of the positive class. We used softmax function to calculate the loss during the training.

5.3. Optimization of the baseline

To determine the optimal conditions for the model trained without the outputs from the VLM, we conducted three experiments. First, since the similarity between adjacent frames may have a negative impact on the classification performance, we explored the suitable interval for extracting data from the training set. Second, to determine the optimal model size and efficacy of the use of pretrained weights, we

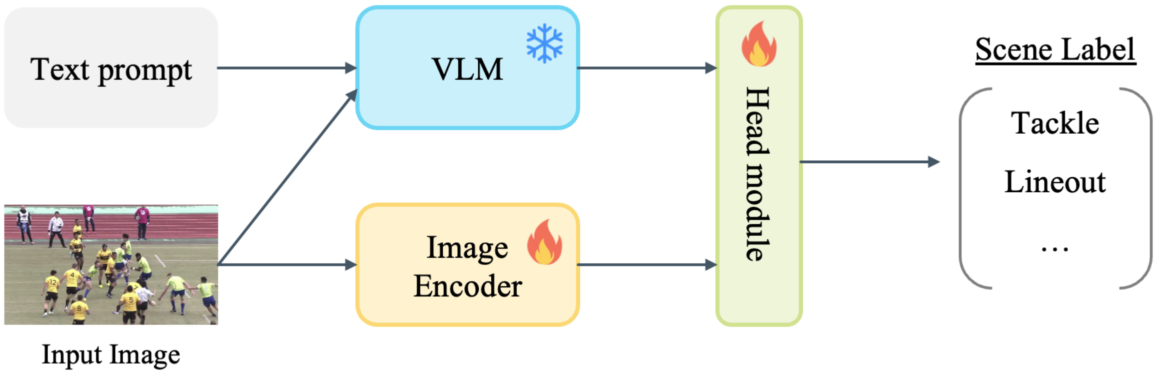

Figure 2. Rugby scene classification with VLM. Our model have three main component; 1) VLM: This component takes natural language prompt and image data as an input and outputs vector representations corresponding to a given input. For this component, we use pretrained model and do not update the parameters during training. 2) Image encoder: This component takes image data as an input and extract image features. We use standard ResNet model and update parameters during training. 3) Head module: This component takes output vectors of the VLM and image encoder as an input and outputs vectors corresponding to a number of classes for the task.

compared various ResNet architectures [14] with and without pretrained weights for each task. Third, based on the determined frame interval and model architecture, we examined the optimal combination of batch size and learning rate for each task. For each experiment, we utilized one set of training and evaluation pairs, and conducted performance comparison based on the F1 score computed on the validation set.

The original labeled data were extracted from videos at intervals of 0.2 seconds, resulting in high similarity between adjacent frames, which could potentially have a negative impact during training. Therefore, for each task, we conducted experiments using frames at intervals of 0.2 seconds, 1 second, 2 seconds, and 3 seconds during training. In this experiment, we used a ResNet-50 pretrained with the Imagenet-1K dataset, with a learning rate of 0.001 and a batch size of 256.

Subsequently, we investigated optimal model architectures and use of pretrained weight for each task, and then searched for learning rate and batchsize. As for the model architecture, we considered five types of model structures: ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152. For each model, we compared two scenarios: one without pretraining and one pre-trained on the Imagenet-1K dataset, resulting in a total of ten configurations. After examining optimal model architecture and the use of pretrained weight, we conducted a grid search to find the optimal combination of batch size and learning rate for each task respectively. The obtained optimal settings for each task were used throughout following experiments (further details are in Appendix 9.3).

5.4. VLM prompt selection

Since the output of VLM are affected by the given prompts, we explored suitable prompts for each task. We tested baseline prompt (#1), which simply asks to explain the given image, seven prompts (#2 - #8) which asks to explain the image with focus to rugby with simple instruction, and four prompts (#9 -#12) with relatively detailed information of specific situation of rugby, as shown in Tab. 2. We inserted each prompt into `<PROMPT>` part of "`<image>\nUSER: <PROMPT>\nASSISTANT:`" as recommended and input it into VLM along with the images. For the image encoder part, we adopted the conditions obtained from the exploration of the baseline, and for the VLM model, we used LLaVa-7B model [27]. Among the outputs of VLM, the output of the last hidden layer was passed into the head module along with the output of the image encoder to obtain predictions. Similar to the baseline investigation, we conducted training for each prompt using one set of the four training/evaluation sets and compared the results based on the F1 score on the validation set.

5.5. Evaluation of VLM efficacy

To evaluate the effectiveness of VLM outputs on rugby scene classification task, we compared the model with VLM output to the baseline model without VLM output for each of the five tasks. In conditions using VLM outputs, we used the output of the last hidden layer of VLM, as in the prompt comparison. Additionally, we examined the performance of the model when using the vectors converted using CLIP [41] from generated sentence of VLM. Model evaluation was conducted by training three independent models using the three sets of training/validation pairs which were not used for the baseline exploration and prompt comparisons.

Table 2. List of prompts tested in this work.

	Prompt
#1	Explain the image.
#2	Explain if contact happening in the image.
#3	Explain if tackle happening in the image.
#4	Explain if lineout happening in the image.
#5	Explain if ruck happening in the image.
#6	Explain the image briefly as an expert of rugby.
#7	Write a short, caption for this rugby image that captures its essence.
#8	You are looking at an image of rugby. Explain the situation in the image.
#9	You are looking at an image of rugby. Firstly, focus on the location of rugby ball, and then explain the situation in the image.
#10	You are looking at an image of rugby. Firstly, focus on the location of players, and then explain if contact is happening in the image.
#11	You are looking at an image of rugby. Is players coming together, pushing to restart play and contest possession?
#12	Are there any specific cues in this image that point towards a tackle (e.g., open arms, bent legs) or a scrum (e.g., three rows of players, bound together)?

We then applied each model to a common evaluation set and calculated the average F1 score based on the results.

Other training conditions were kept consistent across the baseline investigation, prompt comparison, and evaluation of VLM effectiveness. Specifically, we set the maximum number of epochs to 500 and applied early stopping if the metrics on the evaluation set did not improve for five consecutive evaluations. We used the Adam [20] as an optimizer. The data for the validation and test sets consisted of all labeled data, i.e., data extracted at 0.2-second intervals. To mitigate the class imbalance problem, we applied the inverse of the ratio of positive to negative samples in the training set as weights for the positive samples.

6. Result

Table 3. Scene classification with different sampling intervals. The **bold** number indicate the best setting for each task.

Interval [seconds]	0.2	1.0	2.0	3.0
Multi-class	0.508	0.539	**0.558**	0.402
Lineout	0.242	**0.468**	0.351	0.133
Ruck	**0.507**	0.068	0.000	0.010
Tackle	0.369	**0.468**	0.429	0.297
Contact	0.704	**0.711**	0.637	0.643

First, we conducted experiments to find the optimal baseline settings for each task. The optimal frame intervals for training were determined to be every 2 seconds for multi-class classification, every 1 second for lineout, tackle and contact classification, and every 0.2 seconds for ruck classification as shown in Tab. 3. Upon comparing model architectures, ResNet-18 exhibited the highest performance for multi-class classification, ResNet-152 for lineout and ruck classification, and ResNet-101 for tackle and contact scene classification, with consistently better performance when pretrained on Imagenet-1k dataset (see Appendix 9.2 for detailed results). Furthermore, upon examining the learning rate and batch size of the models, the optimal batch size was 512 for contact scenes, 128 for multi-class classification, 64 for lineout and tackle, and 32 for ruck, while the optimal learning rate was 0.0001 for multi-class classification, 0.00025 for ruck, tackle, and overall contact, and 0.00005 for lineout (see Appendix 9.3 for detailed results). Based on these experimental results, we selected the baseline conditions for following experiments.

Subsequently, to examine the impacts of varying prompts given to the VLM, we compared 12 prompts and evaluated the classification performance. The results are shown in Tab. 4. For multi-class classification and tackle classification, the prompt "Write a short, caption for this rugby image that captures its essence." (#7) showed the best performance. For lineout classification, the prompt "Are there any specific cues in this image that point towards a tackle (e.g., open arms, bent legs) or a scrum (e.g., three rows of players, bound together)?" (#12) performed best. The best prompt for ruck classification was "You are look-

Table 4. Results of prompt comparison for each task, showing F1 scores on the validation set. **Bold** indicates the best setting.

Prompt	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12
Multi-class	0.592	0.631	0.622	0.590	0.604	0.622	**0.631**	0.620	0.569	0.570	0.627	0.578
Lineout	0.393	0.643	0.571	0.429	0.618	0.377	0.438	0.437	0.548	0.668	0.365	**0.692**
Ruck	0.464	0.599	0.575	0.493	0.606	0.632	0.536	0.582	0.504	**0.687**	0.527	0.418
Tackle	0.375	0.449	0.449	0.484	0.499	0.424	**0.531**	0.470	0.409	0.465	0.471	0.463
Contact	0.665	0.657	0.728	0.652	0.761	0.705	0.763	0.682	0.686	0.646	**0.768**	0.716

ing at an image of rugby. Firstly, focus on the location of rugby ball, and then explain if contact is happening in the image." (#10) and for contact "You are looking at an image of rugby. Is players coming together, pushing to restart play and contest possession?" (#11). In terms of the average ranking prompt #7 showed the best performance. For each task, the prompt with the best results was used in a comparison experiment with the baseline.

After selecting the prompt for each task, we compared the classfication performance with and without VLM. The results are shown in Tab. 5. For all five tasks, the classification performance was improved when output from the last hidden layer was used compared to the baseline, with lineout classification showing largest gain of 95.1% and median improvement of 3.8%. When the model was trained with vectors converted from VLM generated text, the classification performance improved with four tasks. Comparing the results of the model using output of the last hidden layer of VLM and the model using VLM generated text, the former showed better performance on multi-class classification, lineout and ruck classification, while the latter was better on tackle and contact classification.

7. Discussion

The evaluation of frame intervals during training suggested that increasing the interval yielded improved performance, with the exception of ruck classification. Owing to the inherent nature of rugby gameplay, events such as the moments preceding lineouts or scrums involve minimal player movement as the game momentarily pauses, resulting in smaller interframe differences. Therefore, maintaining a small frame interval during training could negatively impact performance due to data similarity resulting from minimal interframe differences.

Subsequently, we examined optimal settings for the model size, learning rate, and batch size of the baseline. The result of architecture comparison exhibited a propensity to select larger models such as ResNet-101 and ResNet-152 for all tasks, except for multi-class classification where smaller models were preferred. Notably, the best performance was consistently achieved using models pre-trained on ImageNet-1k, regardless of the task. This finding suggests that the parameters acquired through pre-training on the ImageNet-1k dataset are beneficial even when dealing exclusively with domain-specific rugby images.

After exploring the baseline settings, we compared prompts given to the VLM. Comparing the simplest prompt (#1) with prompts containing the word "rugby" or rugby related terms (#2-8), performance improved in many cases when using prompts #2-8, suggesting that explicitly stating the image's subject matter as rugby may yield higher-quality results. However, when comparing prompts #2-5, the best-performing prompt for each task did not always match the rugby-specific terminology mentioned, indicating that the VLM or underlying LLM may not consistently process the nuances of rugby gameplay accurately. For the five tasks tested in this study, prompt #7 exhibited the highest average performance. It is worth noting that providing contextual information in the prompt regarding the image's relation to rugby, without explicitly specifying the play type, may have been advantageous.

The experimental result of comparing the classification performance with and without VLM output exhibited positive impact of using VLM output for all five tasks examined. In all cases, performance was enhanced when utilizing the VLM output, with the improvement being particularly pronounced for the lineout classification. A lineout is a distinctive situation in rugby where players from both teams form a perpendicular line along the touchline and contest for possession. While the characteristics of a lineout can be relatively easily described linguistically, learning solely from visual data requires capturing the spatial relationship with the touchline and player positioning. Consequently, the baseline lineout classification model, trained exclusively on images, exhibited poor performance, which was significantly improved by leveraging the VLM output. Conversely, rucks and tackles are situations where players are in physical contact, irrespective of location, suggesting that classification performance for this event is comparatively robust even when trained solely from image data.

Finally, we evaluated the sentence outputs when providing the simplest prompt (#1) and the prompt with the highest average performance (#7), along with the image. A representative example is shown in Fig. 3. The lack of contextual information in Prompt #1 regarding the given image be-

Table 5. The mean and standard deviation of F1 scores on the test set. "VLM-hidden": the model trained with output of last hidden layer. "VLM-text": the model trained with vectors obtained by converting output of generated text from VLM using CLIP.

VLM-hidden	×	✓	×
VLM-text	×	×	✓
Multi-class	0.615 ± 0.019	0.631 ± 0.022 (2.60%)	0.622 ± 0.049 (1.14%)
Lineout	0.263 ± 0.067	0.513 ± 0.150 (95.06%)	0.369 ± 0.291 (40.30%)
Ruck	0.526 ± 0.055	0.542 ± 0.010 (3.04%)	0.469 ± 0.050 (-10.84%)
Tackle	0.409 ± 0.062	0.428 ± 0.047 (4.65%)	0.441 ± 0.048 (7.82%)
Contact	0.602 ± 0.084	0.625 ± 0.118 (3.82%)	0.679 ± 0.026 (12.79%)

Prompt #1	Prompt #7
The image captures a group of men playing a game of rugby on a field. There are at least 12 players in the scene, with some of them actively participating in the game, while others are watching or waiting for their turn	A group of men playing rugby on a field, with one player falling to the ground.

(a) Tackle

Prompt #1	Prompt #7
The image captures a group of men playing a game of soccer on a field. There are at least 12 players visible in the scene, with some of them actively participating in the game, while others are standing or waiting for	A group of men playing rugby on a field, with one player holding a white sign.

(b) Lineout

Prompt #1	Prompt #7
The image captures a group of young men playing a game of soccer on a grassy field. There are at least 13 players visible in the scene, with some of them actively participating in the game, while others are standing	A group of men playing rugby on a field, with one player falling to the ground while others try to get him out of the way.

(c) Ruck

Prompt #1	Prompt #7
The image captures a group of young men playing a game of football on a grassy field. They are in the middle of a huddle, with some players standing closer to the center of the field while others are positioned around the	A group of rugby players in a huddle, with one player holding a ball, ready to play.

(d) Scrum

Figure 3. Examples of VLM outputs. We show the results of the simplest prompt (#1) and the best performing prompt (#7). While Prompt #1 frequently misidentified the sport depicted in the image as American or European football, Prompt #7 correctly recognized it as an image of rugby.

ing about rugby often results in the image being incorrectly explained as depicting American or European football. In contrast, prompt #7, which explicitly mentions rugby, accurately recognizes the sport, highlighting the beneficial effect of contextual information regarding the subject matter during the prompting process.

8. Conclusion and limitation

In this study, we evaluated the efficacy of utilizing the vector representations generated as output from the VLM for the task of rugby scene classification. A comparison of prompts showed that the optimal prompt for each task differed; however, when the prompt included the word "rugby" or related terminology, it outperformed prompts that did not contain such rugby specific words. Comparing the results with and without the VLM output revealed improved classification performance across all five tasks tested in this study when the VLM output was utilized. Additionally, the sentences generated from the VLM were coherent, suggesting that providing contextual information about the image depicting a rugby game may enable a correct understanding of the context. Overall, the results obtained in this study indicate that the performance of DNNs on the task of rugby scene classification can be enhanced by leveraging the knowledge encoded within LLMs, a component of LVMs, through the use of carefully designed prompts.

One notable limitation of this study is that the exploration of prompts tailored for classifying each distinct play type was not comprehensive. Moreover, the primary focus of verification in this study was the utility of the knowledge encoded within the LLMs, while the verification of whether linguistically representing insights through prompts can effectively facilitate task completion remained inadequately explored. For instance, although the significance of proper head positioning in mitigating the concussion risk from dangerous tackles is well-established, the potential benefits of incorporating such domain-specific knowledge into prompts have not been sufficiently investigated. These limitations underscore potential avenues for future research endeavors aimed at deepening our comprehension of the practical utility of leveraging VLMs in the domain of sports data analysis.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 1

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 1

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[5] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. *arXiv preprint arXiv:2401.17270*, 2024. 1, 2

[6] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4508–4519, 2021. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[9] Matteo Dunnhofer, Luca Sordi, and Christian Micheloni. Visualizing skiers' trajectories in monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5188–5198, 2023. 2

[10] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1711–1721, 2018. 2

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[12] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013. 1

[13] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 5, 1

[15] Jan Held, Anthony Cioppa, Silvio Giancola, Abdullah Hamdi, Bernard Ghanem, and Marc Van Droogenbroeck. Vars: Video assistant referee system for automated soccer decision making from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5085–5096, 2023. 2

[16] Yutaro Honda, Rei Kawakami, Ryota Yoshihashi, Kenta Kato, and Takeshi Naemura. Pass receiver prediction in soccer using video and players' trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3503–3512, 2022. 1, 2

[17] Magnus Ibh, Stella Grasshof, Dan Witzner, and Pascal Madeleine. Tempose: A new skeleton-based transformer model designed for fine-grained motion recognition in badminton. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5198–5207, 2023. 1, 2

[18] Christian Keilstrup Ingwersen, Christian Møller Mikkelstrup, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders Bjorholm Dahl. Sportspose-a dynamic 3d sports pose dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5218–5227, 2023. 2

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 2

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[23] Kaustubh Milind Kulkarni and Sucheth Shenoy. Table tennis stroke recognition using two-dimensional human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4576–4584, 2021. 2

[24] Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*, 2022. 2

[25] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1, 2
[26] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022. 1
[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2, 3, 5
[28] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021. 1, 2
[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2
[30] Yang Liu and Luiz G Hafemann. A scale-invariant trajectory simplification method for efficient data collection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5128–5137, 2023. 2
[31] Katja Ludwig, Julian Lorenz, Robin Schön, and Rainer Lienhart. All keypoints you need: Detecting arbitrary keypoints on the body of triple, high, and long jump athletes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5179–5187, 2023. 1
[32] Zubair Martin, Sharief Hendricks, and Amir Patel. Automated tackle injury risk assessment in contact-based sports - a rugby union example. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4594–4603, 2021. 1, 2
[33] William McNally, Kanav Vats, Tyler Pinto, Chris Dulhanty, John McPhee, and Alexander Wong. Golfdb: A video database for golf swing sequencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 2
[34] Monami Nishio, Naoki Nonaka, Ryo Fujihira, Hidetaka Murakami, Takuya Tajima, Mutsuo Yamada, Akira Maeda, and Jun Seita. Objective detection of high-risk tackle in rugby by combination of pose estimation and machine learning. In *JSAI International Symposium on Artificial Intelligence*, pages 215–228. Springer, 2022. 1, 2
[35] Naoki Nonaka, Ryo Fujihira, Monami Nishio, Hidetaka Murakami, Takuya Tajima, Mutsuo Yamada, Akira Maeda, and Jun Seita. End-to-end high-risk tackle detection system for rugby. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3550–3559, 2022. 2
[36] AJ Piergiovanni and Michael S. Ryoo. Early detection of injuries in mlb pitchers from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 1, 2
[37] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016. 1
[38] Julian Quiroga, Henry Carrillo, Edisson Maldonado, John Ruiz, and Luis M Zapata. As seen on tv: Automatic basketball video production using gaussian-based actionness and game states recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 894–895, 2020. 2
[39] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 1, 2
[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1, 2
[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 5
[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1
[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1
[44] Saikat Sarkar, Amlan Chakrabarti, and Dipti Prasad Mukherjee. Generation of ball possession statistics in soccer using minimum-cost flow network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 1
[45] Atom Scott, Ikuma Uchida, Masaki Onishi, Yoshinari Kameda, Kazuhiro Fukui, and Keisuke Fujii. Soccertrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3569–3579, 2022. 2
[46] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 1
[47] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1
[48] Rajkumar Theagarajan, Federico Pala, Xiu Zhang, and Bir Bhanu. Soccer: Who has the ball? generating visual analytics and player statistics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1749–1757, 2018. 2

[49] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[50] Gabriel Van Zandycke and Christophe De Vleeschouwer. 3d ball localization from a single calibrated image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3472–3480, 2022. 1, 2

[51] Renaud Vandeghen, Anthony Cioppa, and Marc Van Droogenbroeck. Semi-supervised training to improve player and ball detection in soccer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3481–3490, 2022. 1, 2

[52] Kanav Vats, William McNally, Pascale Walters, David A. Clausi, and John S. Zelek. Ice hockey player identification via transformers and weakly supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3451–3460, 2022. 2

[53] Kanav Vats, Pascale Walters, Mehrnaz Fani, David A Clausi, and John S Zelek. Player tracking and identification in ice hockey. *Expert Systems with Applications*, 213:119250, 2023. 1, 2

[54] Roman Voeikov, Nikolay Falaleev, and Ruslan Baikulov. Ttnet: Real-time temporal and spatial video analysis of table tennis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 884–885, 2020. 2

[55] Marie-Sophie von Braun, Patrick Frenzel, Christian Kading, and Mirco Fuchs. Utilizing mask r-cnn for waterline detection in canoe sprint video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 876–877, 2020. 2

[56] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 1, 2

[57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 1

[58] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. 1

[59] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 1

[60] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023. 2

[61] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[62] Ruiheng Zhang, Lingxiang Wu, Yukun Yang, Wanneng Wu, Yueqiang Chen, and Min Xu. Multi-camera multi-player tracking with deep player identification in sports video. *Pattern Recogn.*, 102(C), 2020. 1

[63] Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[64] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 1, 2

[65] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2