Towards Universal Soccer Video Understanding

Appendix

Contents

A. SoccerReplay-1988 Dataset Details

In this section, we provide additional details of our **SoccerReplay-1988** dataset. Specifically, Sec. A.1 elaborates on the structure and format of the dataset; Sec. A.2 presents statistical information and analyses of the dataset; and Sec. A.3 describes the methodology to automatically generate event labels within the dataset.

A.1. Dataset Format

The SoccerReplay-1988 dataset consists of match videos, descriptions of events, and related game information of 1988 soccer matches. Each match includes two `mkv` video files (1-half and 2-half), covering the match from the initial kick-off to the final whistle. Additionally, a `json` file is accompanied by encapsulating detailed information, including event descriptions and comprehensive match backgrounds, structured as follows:

Match Information provides background details of the match, including competing teams, final results, and match contexts, such as start time, team formations, and venue details, as illustrated below:

```
{
    "timestamp": "2022-08-07 21:00:00",                                # Match start time
    "score": "1 - 2",                                                  # Final score
    "home_team": "Manchester Utd",                                     # Home team name
    "away_team": "Brighton",                                           # Away team name
    "home_formation": "4 - 3 - 3",                                     # Home team formation
    "away_formation": "3 - 4 - 2 - 1",                                 # Away team formation
    "venue": "Old Trafford (Manchester)",                              # Venue and city
    "capacity": "75 635",                                              # Stadium capacity
    "attendance": "73 711",                                            # Number of attendees
}
```

Referee Information includes details about the primary referee officiating the match, which is formatted as follows:

```
{
    "country": "Eng",                                                  # Referee’s nationality
    "name": "Paul Tierney"                                             # Referee’s name
}
```

Player Information contains details about various types of individuals involved in the match, including starting players, substitutes, absent players, and coaches. All these types are stored in a unified list, with the following format:

```
{
    "players_name": "Caicedo M.",                                      # Player’s abbreviated name
    "players_number": "25",                                            # Jersey number
    "Full Name": "Moises Caicedo",                                     # Player’s full name
    "players_rating": 7.6,                                             # Post-match rating
    "Country": "Ecuador",                                              # Player’s nationality
    "Role": "Midfielder",                                              # On-field role
    "Age and Birthdate": "22, (02.11.2001)",                           # Age and birth date
    "Market Value": "€89.4m"                                           # Player’s market value
}
```

Event Descriptions is a list that records all key events during the match, including their types and detailed commentary. A typical example of an event entry is shown below:

```
{
    "half": 1,                                                         # Match half (1 or 2)
    "time_stamp": "00:16",                                             # Timestamp within the half
    "comments_type": "shot off target",                                # Event type
    "comments_text": "A mistake by Leandro Trossard (Brighton)...",    # Commentary text
    "comments_text_anonymized": "A mistake by [PLAYER]([TEAM])..."     # Commentary after anonymization
}
```

A.2. Additional Dataset Statistics

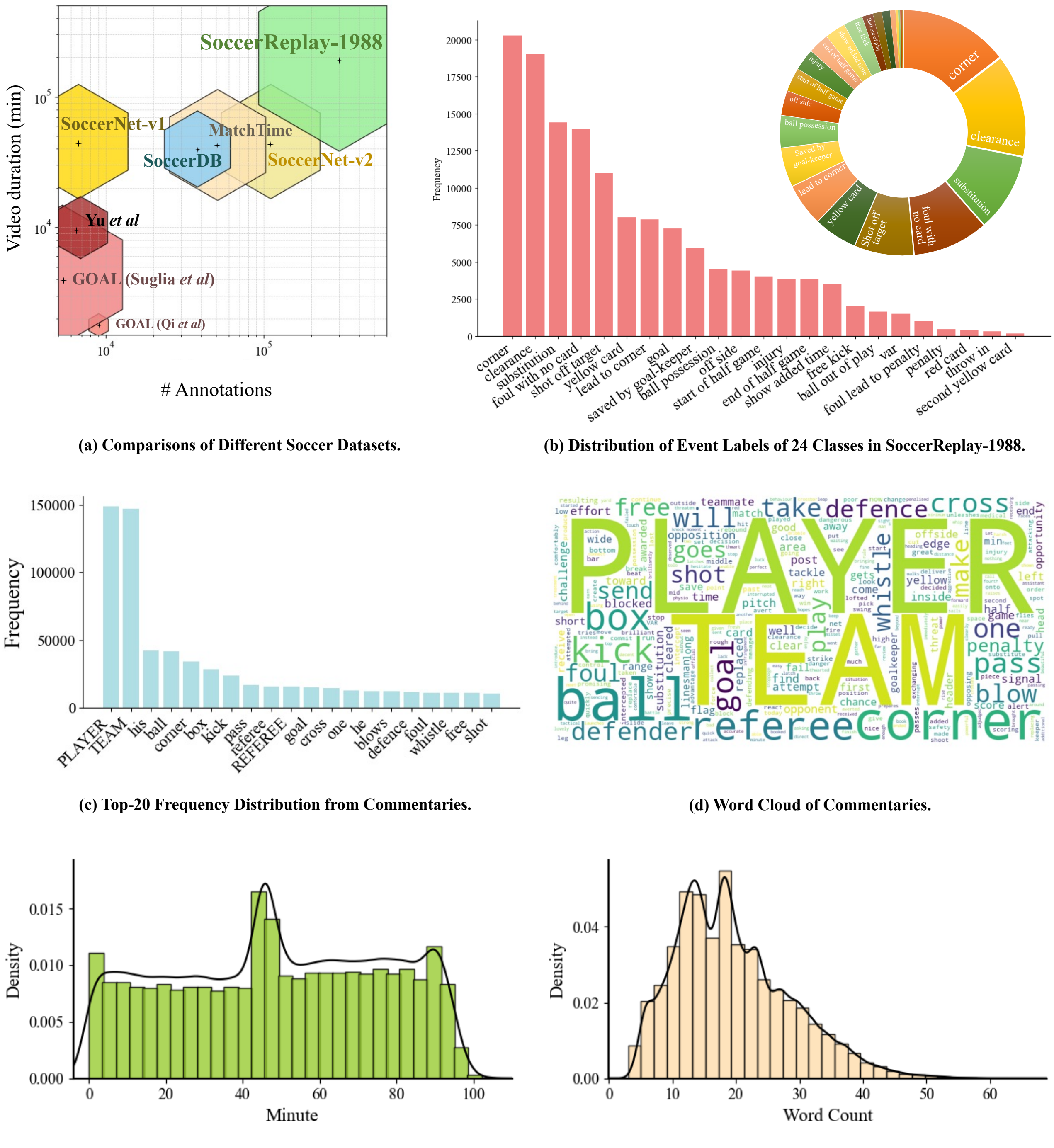
League	# Match
Italy Serie-a	367
England Premier League	552 z
UEFA Champions League	469
France Ligue 1	123
Spain LaLiga	235
Germany Bundesliga	242

Table 1. **League-wise Match Statistics.**

Season	# Match
2017-2018	172
2018-2019	325
2019-2020	300
2020-2021	323
2021-2022	330
2022-2023	416
2023-2024	122

Table 2. **Season-wise Match Statistics.**

To provide a comprehensive analysis of **SoccerReplay-1988**, we present statistics and visualizations in the following tables and figures. Concretely, Table 1 and 2 illustrate the distribution of the 1988 matches across different leagues and seasons. Figure 1 (a) compares SoccerReplay-1988 with other soccer datasets [2, 4, 7, 9–11, 13], highlighting its unique scale. Figure 1 (b) depicts the distribution of event labels for the 24 newly defined categories. Finally, Figures 1 (c), (d), (e), and (f) present detailed analyses of commentary data, including frequency distributions, timestamps, and word counts.



(a) Comparisons of Different Soccer Datasets.

(b) Distribution of Event Labels of 24 Classes in SoccerReplay-1988.

(c) Top-20 Frequency Distribution from Commentaries.

(d) Word Cloud of Commentaries.

(e) Distribution of Temporal Occurrences of Commentaries.

(f) Commentary Word Count Distribution.

Figure 1. **Comprehensive Visualizations of SoccerReplay-1988 Dataset.**

A.3. Event Summarization

Our dataset expands the original 17 event categories in SoccerNet [2] to 24 types. For MatchTime [10] and SoccerReplay-1988, LLaMA-3 (70B) [3] is employed to analyze commentaries and generate corresponding event labels. This is guided by a carefully refined prompt, iteratively improved through manual checks and enriched with comprehensive definitions and examples. Notably, text commentaries unrelated to visual content (*e.g., 'possession ratio is 55:45'*) are categorized as *'statistics and summary'*, and excluded from model training and testing. Finally, a random sample of 2% of the data yields 98% manual verification accuracy, confirming the high quality of automated labeling. The entire prompt is presented below:

`<|begin_of_text|>`
`<|start_header_id|>system<|end_header_id|>`

You are an expert in soccer, you have a very important task to summarize a soccer commentary into certain types of events. The accuracy of your classification is the most emergency thing. I will give you a commentary sentence. You need to select one type of event that can best describe this event from the following 24 types: *'corner', 'goal', 'injury', 'own goal', 'penalty', 'penalty missed', 'red card', 'second yellow card', 'substitution', 'start of game(half)', 'end of game(half)', 'yellow card', 'throw in', 'free kick', 'saved by goal-keeper', 'shot off target', 'clearance', 'lead to corner', 'off-side', 'var', 'foul (no card)', 'statistics and summary', 'ball possession', 'ball out of play'*.

Here are some rules you have to obey when summarizing types, you should consider it strictly following these steps:

1. Firstly, you need to find if there is any evidence of foul in commentary, if yes, it can only be *'foul (no card)'*, *'yellow card'*, *'red card'* or *'second yellow card'* according to the situation, even though it introduces the result *'free kick'* or *'penalty'*. For example: 'Per Mertesacker (Arsenal) commits a rough foul. Michael Dean stops the game and makes a call. That's a free kick to Manchester Utd.' can ONLY be *'foul (no card)'* since there is a foul in commentary, even though the result is *'free kick'*.
2. Secondly, only if the word *'corner'* is in the commentary, you need to select it from *'lead to corner'*. *'lead to corner'* means the process of how the *'corner'* occurs, which is before the *'corner'* kick. For type *'lead to corner'*, there will always be words like 'award a *corner*', 'will have a *corner*', 'point at *corner* flag' and so on. For example: 'Victor Wanyama (Southampton) goes on a solo run, but he fails to create a chance as an opposition player blocks him. The referee signals a *corner* kick to Southampton.' is *'lead to corner'*.
3. Thirdly is the most easy-confused part, you need to be cautious: only if the word *'free-kick'/'free kick'* is in the commentary will it be a *'free kick'*. According to the first rule, if there is foul in the sentence, it cannot be *'free kick'*. *'free kick'* can only be selected when *'free-kick'/'free kick'* occurs in commentary and is describing the process of a *'free kick'* attack. For example: 'Olivier Giroud (Arsenal) gets on the ball and beats an opponent, but his run is stopped by the referee Michael Dean who sees an offensive foul. It's a *free kick* to Burnley, but they probably won't attempt a direct shot on goal from here.' is *'foul (no card)'*; 'Ander Herrera (Manchester United) makes a slide tackle, but referee Michael Dean blows for a foul. *Free kick*. Arsenal will probably just try to cross the ball in from here.' is *'foul (no card)'*; 'Marcos Rojo (Manchester United) connects with the *free kick* and produces a header goalwards which is well blocked. The goalkeeper doesn't have to worry about that one.' is *'free kick'*.
4. Similarly, *'penalty'* and *'penalty missed'* only describe things that happen during a *'penalty'* kick. If it is introducing the reason that leads to a *'penalty'*, you should return the type describing the reason, like *'foul (no card)'*, *'yellow card'*, and so on.
5. The type *'statistics and summary'* includes all the commentaries that are not introducing visually evidential events, but those statistics or overviews of the game. These sentences won't concentrate on certain events, but on the overall game.
6. *'ball possession'* represents those commentaries that describe any of the teams controlling the *'ball possession'*.
7. You need to be sensitive about the type *'shot off target'*; if there is an event of a shot happening in the commentary, it is a shot. If it's not a *'goal'*, didn't make a score, and was not saved by the goalkeeper, it would probably be a *'shot off target'*. Normally there will be keywords like 'wide of the right post', 'over the crossbar', 'crashes against the crossbar' and so on. You have to judge it sensitively about the situation after the shot.
8. An important type after a shot: *'saved by goal-keeper'* describes that the shot is saved by the goalkeeper; there would be words like 'blocked', 'saved', and so on. Especially when *'goal-keeper'/'goal keeper'* occurs in the commentary sentence!!! it will probably be *'saved by goal-keeper'*. You need to find it carefully!!!
9. If a player lofts or swings a pass to a penalty area/dangerous area, they might be *'shot off target'*, *'clearance'*, *'saved by goal-keeper'*, and so on. It should NOT be identified as *'corner'* or *'free kick'* if there is no obvious evidence in commentary! For example: 'Tomas Rosicky (Arsenal) fails to find any of his teammates inside the box as his pass is blocked.' should be *'clearance'* rather than *'corner'* or *'free kick'*.
10. *'clearance'* means those good performances in defense; they stop the offense of opponents. If such a successful defense happens in the commentary, it can only be *'clearance'*. In these commentaries, there are always some words like 'opponent's defense', 'intercepts the ball', 'clear the ball', and so on.
11. *'offside'* is an obvious event; there are always the words 'flag', 'linesman', 'too fast to defense' in the commentary since *'offside'* is the player running forward the defense line, and the linesman will raise the flag.
12. *'ball out of play'* means any player kicks the ball out of boundary lines. These commentary sentences will mostly end up with throw-ins or goal kicks.
13. *'throw-in'* means exactly the process of *'throw-in'* balls.
14. Most *'goals'* are normal *'goals'*. If you see a scoring event, you can only identify the score as *'own goal'* when there is obvious evidence.

`<|eot_id|>`
`<|start_header_id|>user<|end_header_id|>`

With the classification rules, you should tell me the type of a commentary from above candidate options: *'corner', 'goal', 'injury', 'own goal', 'penalty', 'penalty missed', 'red card', 'second yellow card', 'substitution', 'start of game(half)', 'end of game(half)', 'yellow card', 'throw in', 'free kick', 'saved by goal-keeper', 'shot off target', 'clearance', 'lead to corner', 'off-side', 'var', 'foul (no card)', 'statistics and summary', 'ball possession', 'ball out of play'*. The commentary sentence you need to define type is:

> [COMMENTARY TEXT HERE (before anonymization)]

You need to carefully consider the rules in order and make your final decision. Now, you must return me the name of its type from candidate options (in lower case, only return the name of type, answer it right away after my prompt without any other words).

`<|eot_id|>`

B. SoccerNet-pro Dataset Details

As discussed in the main text, alongside the SoccerReplay-1988 dataset, we also incorporate two existing datasets, SoccerNet-v2 [2] and MatchTime [10] to enrich the training data. These datasets undergo the following preprocessing strategies and are then unified into the SoccerNet-pro dataset, ensuring format consistency with SoccerReplay-1988.

B.1. SoccerNet-v2

The SoccerNet-v2 [2] dataset comprises over 110k event labels across 500 matches, categorized into 17 distinct classes. Based on soccer rules and specific domain knowledge, these labels are systematically reclassified into 24 categories with our proposed automated data curation pipeline, as detailed in Table 3.

Original Label	Processed Label	Reference
Penalty	Penalty Penalty Missed	Scored penalties are categorized as “Penalty.” Missed penalties are categorized as “Penalty Missed.”
Kick-off	Start of Game (Half)	Matches the start of a half after goals.
Shots off target	Shot Off Target	No change.
Throw-in	Throw In	No change.
Ball out of play	Ball Out of Play	No change.
Foul	Foul (No Card)	Refers to fouls without cards for only.
Yellow card	Yellow Card	No change.
Yellow→red card	Second Yellow Card	No change.
Red card	Red Card	No change.
Direct free-kick Indirect free-kick	Free Kick	Both direct and indirect free kicks are grouped.
Substitution	Substitution	No change.
Goal	Goal	No change.
Clearance	Clearance	No change.
Offside	Off-Side	No change.
Corner	Corner	No change.

Table 3. **Processing Strategy for SoccerNet-pro.** The **Reference** column details the specific processing applied to the original labels.

B.2. MatchTime

The MatchTime dataset [10], curated from SoccerNet-Caption [8], contains a substantial amount of commentary, with only a small portion accompanied by event labels. To bridge this gap, we apply the prompt-based approach described in Sec. A.3 to summarize commentaries into event labels, assigning each commentary a corresponding label.

B.3. Data Split Strategy

As described in the manuscript, SoccerReplay-1988 is divided into 1,488 matches for training, 250 for validation, and 250 for testing. For the processed SoccerNet-pro dataset (including SoccerNet-v2 and MatchTime), we adhere to the original partitioning strategies and match distributions of its source datasets as detailed in Table 4.

Dataset	Train	Valid	Test	Total
SoccerNet-v2 [2]	300	100	100	500
MatchTime [10]	373	49	49	471
SoccerReplay-1988	1488	250	250	1988

Table 4. **Dataset Splits for Training, Validation, and Testing.**

C. Implementation Details

In this section, we provide additional implementation details about MatchVision. Sec. C.1 presents more information on data preprocessing strategies; Sec. C.2 elaborates on the evaluation strategies used during model training; and Sec. C.3 discusses several hyperparameter choices inspired by prior works.

C.1. Data Preprocessing

Our automated data curation pipeline filters out video clips with missing annotations, incorrect cropping, or invalid timestamps. In all experiments, video frames are resized to 224 × 224 and preprocessed using the image preprocessor of SigLIP [14], which normalizes frames to a mean of 0.5 and a standard deviation of 0.5 before serving as inputs. For overlapping video content between SoccerNet-v2 [2] and MatchTime [10], we prioritize using event labels from SoccerNet-v2.

C.2. Validation Strategy during Training

We select the best-performing checkpoints on the validation set with the evaluation strategies detailed below:

During pretraining: (i) For **supervised classification**, we adopt top-1/3/5 event classification accuracy on the validation set to select the best model; (ii) For **visual-language contrastive learning**, video-to-text retrieval is performed, with top-1/3/5 accuracy of event classification (comparing retrieved texts' event labels to ground truth) as the validation metric.

During downstream tasks training: (i) In **event classification** and **foul recognition**, classification accuracy on the validation set is used as the evaluation metric; (ii) For **commentary generation**, the CIDEr [12] score of the model's predictions on the validation set is employed to select the best checkpoint.

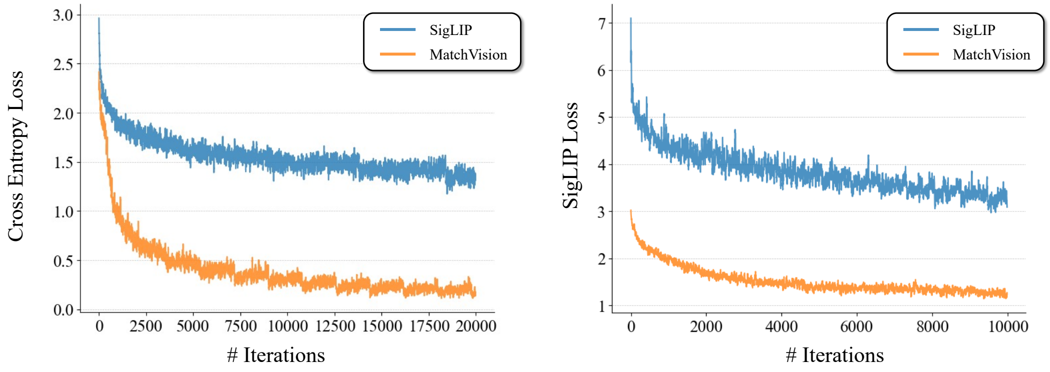
C.3. Hyperparameter Selection

Here, we provide further explanations about the hyperparameters in our model, inspired by prior works, as detailed below:

Temporal Window Size. We adopt a 30-second temporal window to extract video clips. This is inspired by MatchTime [10], which demonstrates that a 30-second window is sufficient to capture adequate visual information for optimal performance, outperforming the 45-second window used in SoccerNet-Caption [8].

LoRA Rank. For finetuning the commentary generation head, we use LoRA [5] with a rank of 16, following [10].

Query Length of Perceiver. For the Perceiver [6] module in the commentary generation head, we utilize a query length of 32 for temporal information aggregation, consistent with the optimal configuration reported in [10].



(a) Supervised Training

(b) Video-Language Contrastive Learning

Figure 2. **Training Loss Curves of Visual Encoders Pretraining.**

D. Experiments

In this section, we provide additional details to offer deeper insights into our model and its performance. Specifically, Sec. D.1 presents training curves to clearly illustrate the training process; Sec. D.2 and Sec. D.3 showcase more quantitative and qualitative results, respectively, demonstrating the model's capability to effectively understand soccer dynamics.

D.1. Training Curves

We present the loss curves for visual encoder pretraining in Figure 2. Our MatchVision demonstrates significantly better convergence compared to SigLIP [14] backbone, indicating that it effectively leverages spatiotemporal attention to utilize temporal information, learning representations better suited for highly dynamic soccer videos.

D.2. More Quantitative Results

Here, we compare present comparisons with more advanced methods in the SoccerNet Foul Recognition challenge [1], where MatchVision remains competitive even with **a frozen visual encoder**.

Top-1 Accuracy	**Ours**	Baseline	zyz	PD_PS_GSN	Redsox	xiao_he_shang
Foul	0.44	0.36	0.58	0.44	**0.60**	0.46
Severity	**0.58**	0.54	0.58	0.50	0.05	0.47

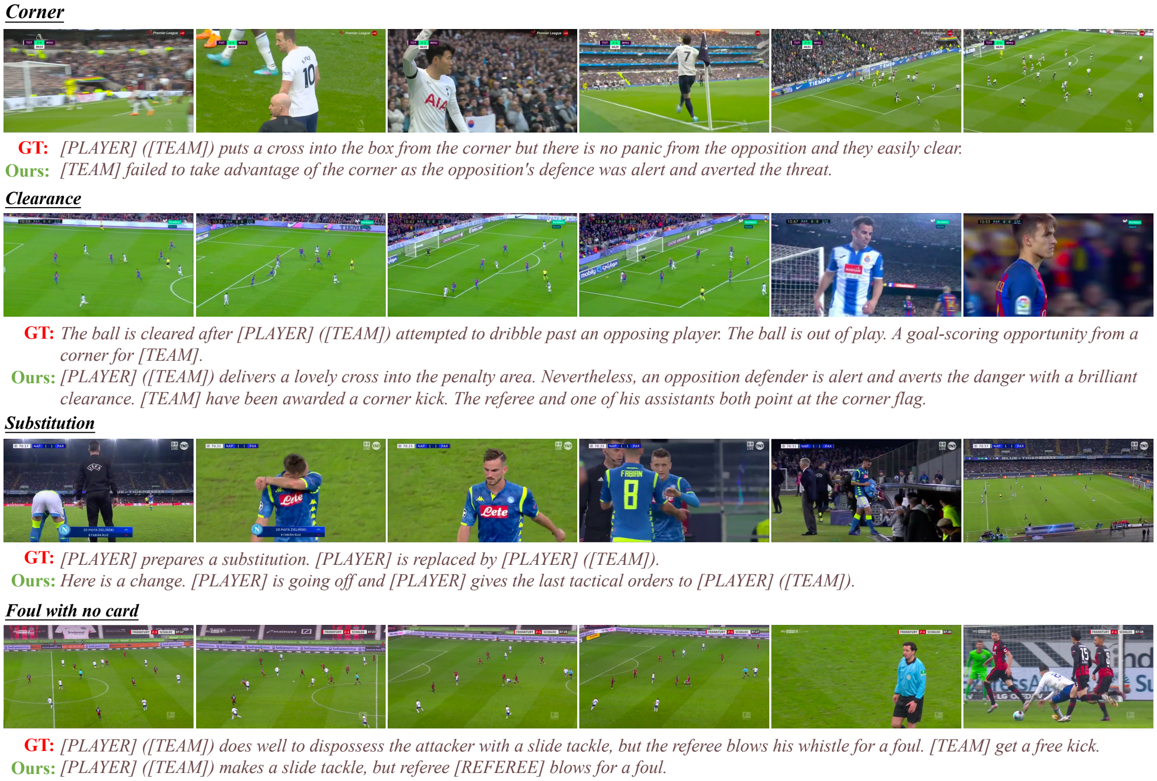
Table 5. **More Quantitative Results on Multi-view Foul Recognition.**

D.3. More Qualitative Results

More qualitative visualizations of commentary generation across various events on the field are depicted in Figure 3, 4, and 5.

E. Limitations & Future Work

Although MatchVision explores establishing a soccer-specific visual encoder, it is not without its limitations: (i) Currently, MatchVision is adapted to event classification, commentary generation, and foul recognition tasks. In the future, we plan to further extend it to more challenging tasks such as player tracking and dense video captioning, aiming to develop a more comprehensive foundation model for soccer analysis. (ii) Given computational and annotation constraints, SoccerReplay-1988 primarily focuses on European league soccer data. We aim to leverage our scalable automated annotation pipeline to further expand the dataset, encompassing a more comprehensive range of soccer data. (iii) Following prior works, our commentary generation remains anonymized. This is left for future work, where we aim to fully leverage contextual information available in our SoccerReplay-1988 dataset to enable more vivid, accurate, and context-aware commentary generation.



Corner

GT: *[PLAYER] ([TEAM]) puts a cross into the box from the corner but there is no panic from the opposition and they easily clear.*
Ours: *[TEAM] failed to take advantage of the corner as the opposition's defence was alert and averted the threat.*

Clearance

GT: *The ball is cleared after [PLAYER] ([TEAM]) attempted to dribble past an opposing player. The ball is out of play. A goal-scoring opportunity from a corner for [TEAM].*
Ours: *[PLAYER] ([TEAM]) delivers a lovely cross into the penalty area. Nevertheless, an opposition defender is alert and averts the danger with a brilliant clearance. [TEAM] have been awarded a corner kick. The referee and one of his assistants both point at the corner flag.*

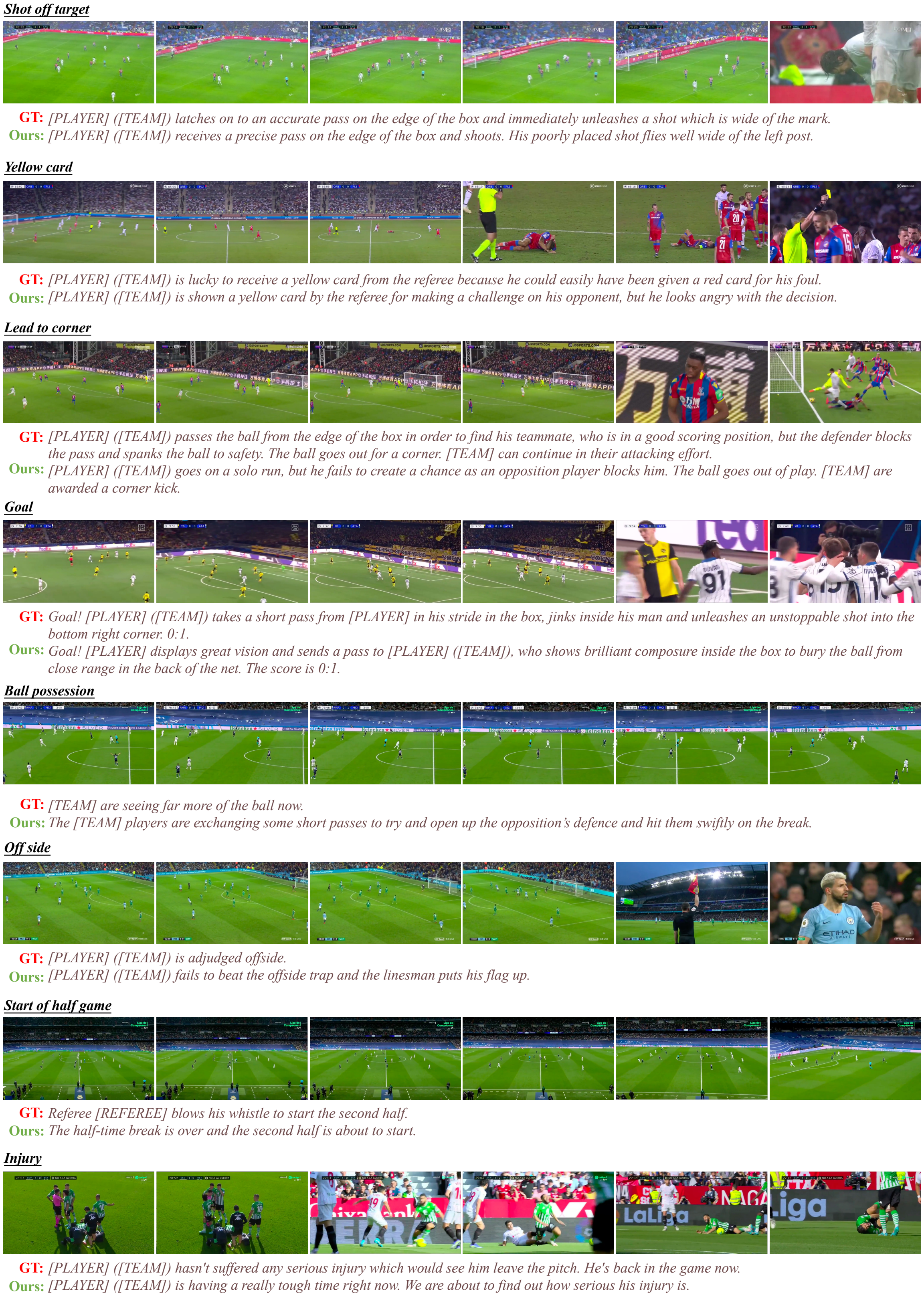
Substitution

GT: *[PLAYER] prepares a substitution. [PLAYER] is replaced by [PLAYER] ([TEAM]).*
Ours: *Here is a change. [PLAYER] is going off and [PLAYER] gives the last tactical orders to [PLAYER] ([TEAM]).*

Foul with no card

GT: *[PLAYER] ([TEAM]) does well to dispossess the attacker with a slide tackle, but the referee blows his whistle for a foul. [TEAM] get a free kick.*
Ours: *[PLAYER] ([TEAM]) makes a slide tackle, but referee [REFEREE] blows for a foul.*

Figure 3. **More Qualitative Results of Commentary Generation.**



Shot off target

GT: *[PLAYER] ([TEAM]) latches on to an accurate pass on the edge of the box and immediately unleashes a shot which is wide of the mark.*
Ours: *[PLAYER] ([TEAM]) receives a precise pass on the edge of the box and shoots. His poorly placed shot flies well wide of the left post.*

Yellow card

GT: *[PLAYER] ([TEAM]) is lucky to receive a yellow card from the referee because he could easily have been given a red card for his foul.*
Ours: *[PLAYER] ([TEAM]) is shown a yellow card by the referee for making a challenge on his opponent, but he looks angry with the decision.*

Lead to corner

GT: *[PLAYER] ([TEAM]) passes the ball from the edge of the box in order to find his teammate, who is in a good scoring position, but the defender blocks the pass and spanks the ball to safety. The ball goes out for a corner. [TEAM] can continue in their attacking effort.*
Ours: *[PLAYER] ([TEAM]) goes on a solo run, but he fails to create a chance as an opposition player blocks him. The ball goes out of play. [TEAM] are awarded a corner kick.*

Goal

GT: *Goal! [PLAYER] ([TEAM]) takes a short pass from [PLAYER] in his stride in the box, jinks inside his man and unleashes an unstoppable shot into the bottom right corner. 0:1.*
Ours: *Goal! [PLAYER] displays great vision and sends a pass to [PLAYER] ([TEAM]), who shows brilliant composure inside the box to bury the ball from close range in the back of the net. The score is 0:1.*

Ball possession

GT: *[TEAM] are seeing far more of the ball now.*
Ours: *The [TEAM] players are exchanging some short passes to try and open up the opposition's defence and hit them swiftly on the break.*

Off side

GT: *[PLAYER] ([TEAM]) is adjudged offside.*
Ours: *[PLAYER] ([TEAM]) fails to beat the offside trap and the linesman puts his flag up.*

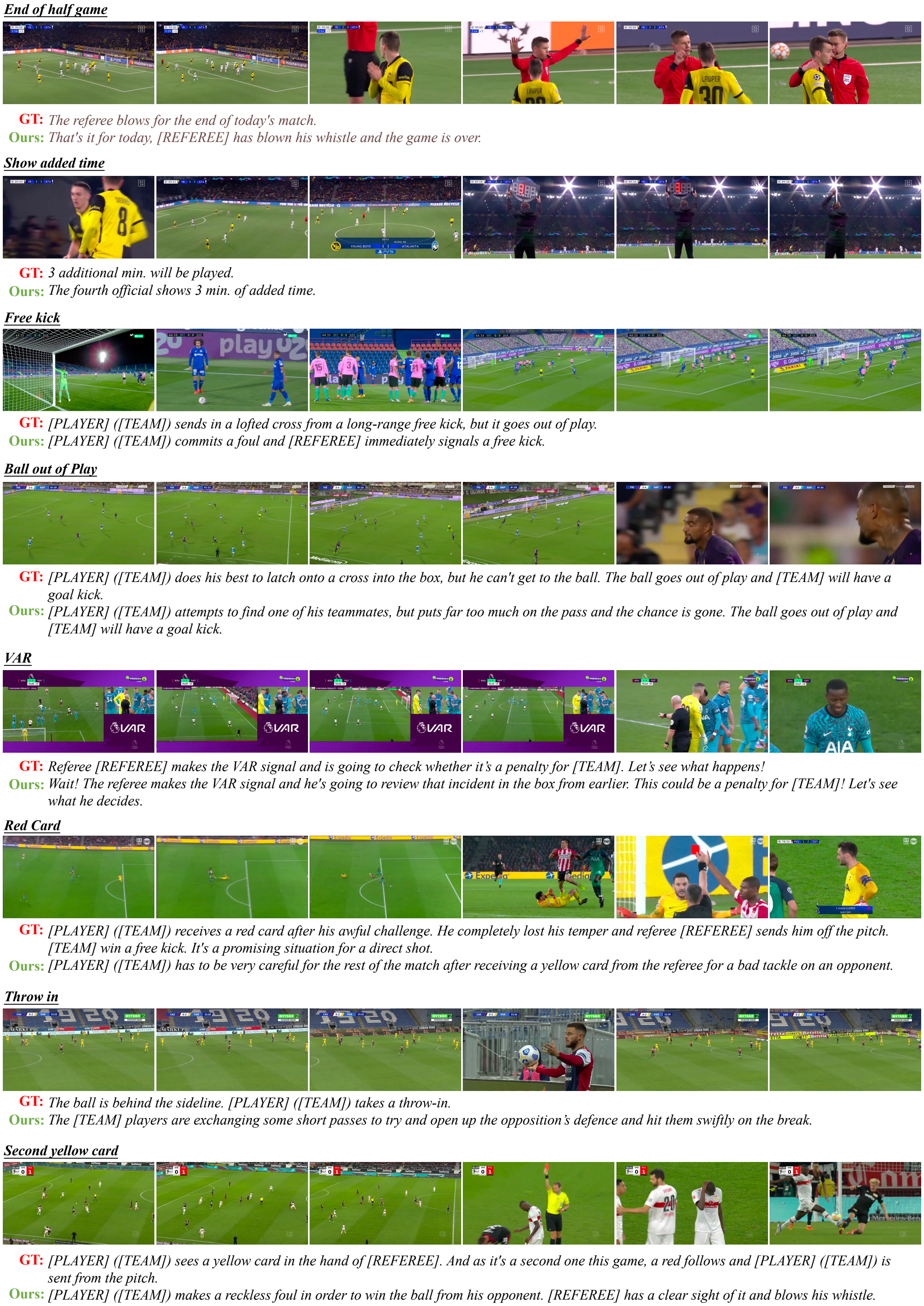
Start of half game

GT: *Referee [REFEREE] blows his whistle to start the second half.*
Ours: *The half-time break is over and the second half is about to start.*

Injury

GT: *[PLAYER] ([TEAM]) hasn't suffered any serious injury which would see him leave the pitch. He's back in the game now.*
Ours: *[PLAYER] ([TEAM]) is having a really tough time right now. We are about to find out how serious his injury is.*

Figure 4. **More Qualitative Results of Commentary Generation.**



End of half game

GT: *The referee blows for the end of today's match.*
Ours: *That's it for today, [REFEREE] has blown his whistle and the game is over.*

Show added time

GT: *3 additional min. will be played.*
Ours: *The fourth official shows 3 min. of added time.*

Free kick

GT: *[PLAYER] ([TEAM]) sends in a lofted cross from a long-range free kick, but it goes out of play.*
Ours: *[PLAYER] ([TEAM]) commits a foul and [REFEREE] immediately signals a free kick.*

Ball out of Play

GT: *[PLAYER] ([TEAM]) does his best to latch onto a cross into the box, but he can't get to the ball. The ball goes out of play and [TEAM] will have a goal kick.*
Ours: *[PLAYER] ([TEAM]) attempts to find one of his teammates, but puts far too much on the pass and the chance is gone. The ball goes out of play and [TEAM] will have a goal kick.*

VAR

GT: *Referee [REFEREE] makes the VAR signal and is going to check whether it's a penalty for [TEAM]. Let's see what happens!*
Ours: *Wait! The referee makes the VAR signal and he's going to review that incident in the box from earlier. This could be a penalty for [TEAM]! Let's see what he decides.*

Red Card

GT: *[PLAYER] ([TEAM]) receives a red card after his awful challenge. He completely lost his temper and referee [REFEREE] sends him off the pitch. [TEAM] win a free kick. It's a promising situation for a direct shot.*
Ours: *[PLAYER] ([TEAM]) has to be very careful for the rest of the match after receiving a yellow card from the referee for a bad tackle on an opponent.*

Throw in

GT: *The ball is behind the sideline. [PLAYER] ([TEAM]) takes a throw-in.*
Ours: *The [TEAM] players are exchanging some short passes to try and open up the opposition's defence and hit them swiftly on the break.*

Second yellow card

GT: *[PLAYER] ([TEAM]) sees a yellow card in the hand of [REFEREE]. And as it's a second one this game, a red follows and [PLAYER] ([TEAM]) is sent from the pitch.*
Ours: *[PLAYER] ([TEAM]) makes a reckless foul in order to win the ball from his opponent. [REFEREE] has a clear sight of it and blows his whistle.*

Figure 5. **More Qualitative Results of Commentary Generation.**

References

[1] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Victor Joos, Floriane Magera, Jan Held, Seyed Abolfazl Ghasemzadeh, Xin Zhou, Karolina Seweryn, Mateusz Kowalczyk, Zuzanna Mróz, Szymon Łukasik, Michał Hałoń, Hassan Mkhallati, Adrien Deliège, Carlos Hinojosa, Karen Sanchez, Amir M. Mansourian, Pierre Miralles, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Adam Gorski, Albert Clapés, Andrei Boiarov, Anton Afanasiev, Artur Xarles, Atom Scott, ByoungKwon Lim, Calvin Yeung, Cristian Gonzalez, Dominic Rüfenacht, Enzo Pacilio, Fabian Deuser, Faisal Sami Altawijri, Francisco Cachón, HanKyul Kim, Haobo Wang, Hyeonmin Choe, Hyunwoo J Kim, Il-Min Kim, Jae-Mo Kang, Jamshid Tursunboev, Jian Yang, Jihwan Hong, Jimin Lee, Jing Zhang, Junseok Lee, Kexin Zhang, Konrad Habel, Licheng Jiao, Linyi Li, Marc Gutiérrez-Pérez, Marcelo Ortega, Menglong Li, Milosz Lopatto, Nikita Kasatkin, Nikolay Nemtsev, Norbert Oswald, Oleg Udin, Pavel Kononov, Pei Geng, Saad Ghazai Alotaibi, Sehyung Kim, Sergei Ulasen, Sergio Escalera, Shanshan Zhang, Shuyuan Yang, Sunghwan Moon, Thomas B. Moeslund, Vasyl Shandyba, Vladimir Golovkin, Wei Dai, WonTaek Chung, Xinyu Liu, Yongqiang Zhu, Youngseo Kim, Yuan Li, Yuting Yang, Yuxuan Xiao, Zehua Cheng, and Zhihao Li. Soccernet 2024 challenges results. *arXiv preprint arXiv:2409.10587*, 2024. 7

[2] 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 Conference on Computer Vision and Pattern Recognition Workshops*, pages 4508–4519, 2021. 3, 4, 5, 6

[3] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 4

[4] 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. 3

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, 2022. 6

[6] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the International Conference on Machine Learning*, pages 4651–4664, 2021. 6

[7] Yudong Jiang, Kaixu Cui, Leilei Chen, Canjin Wang, and Changliang Xu. Soccerdb: A large-scale database for comprehensive video understanding. In *Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports*, pages 1–8, 2020. 3

[8] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 5074–5085, 2023. 5, 6

[9] Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Bin Xu, Lei Hou, Juanzi Li, et al. Goal: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 5391–5395, 2023. 3

[10] Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. Matchtime: Towards automatic soccer game commentary generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024. 4, 5, 6

[11] Alessandro Suglia, José Lopes, Emanuele Bastianelli, Andrea Vanzo, Shubham Agarwal, Malvina Nikandrou, Lu Yu, Ioannis Konstas, and Verena Rieser. Going for goal: A resource for grounded football commentaries. *arXiv preprint arXiv:2211.04534*, 2022. 3

[12] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015. 6

[13] Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, and Xiaokang Yang. Fine-grained video captioning for sports narrative. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[14] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the International Conference on Computer Vision*, pages 11975–11986, 2023. 6