SCORE: Enhancing Soccer Commentary Generation via Contextual Expansion and Information Retrieval

Anonymous ACL submission

Abstract

Automatic soccer commentary generation aims to bridge the gap between raw visual content and professional, tactical commentary. However, existing datasets often lack semantic richness and detailed scene analysis. They also fail to capture the continuity between events, resulting in fragmented and contextually disconnected commentaries. To address these issues, we propose two manually curated datasets: SN-Short and SN-Long. SN-Short focuses on enhancing the semantic description of scene details, while SN-Long captures event continuity to enable coherent, context-aware commentary. In this paper, we also introduce SCORE (Soccer Commentary Generation via Contextual Expansion and Information Retrieval), a novel framework designed to address both detailed scene understanding and global context awareness. SCORE employs a commentary expansion pipeline that integrates visual features with sparse annotations to generate detailed scene descriptions, and it utilizes a retrieval-augmented generation model that incorporates contextual cues from previous events to produce coherent commentary aligned with the visual flow of the game. The experimental results show that SCORE significantly outperforms existing baselines in the proposed datasets.

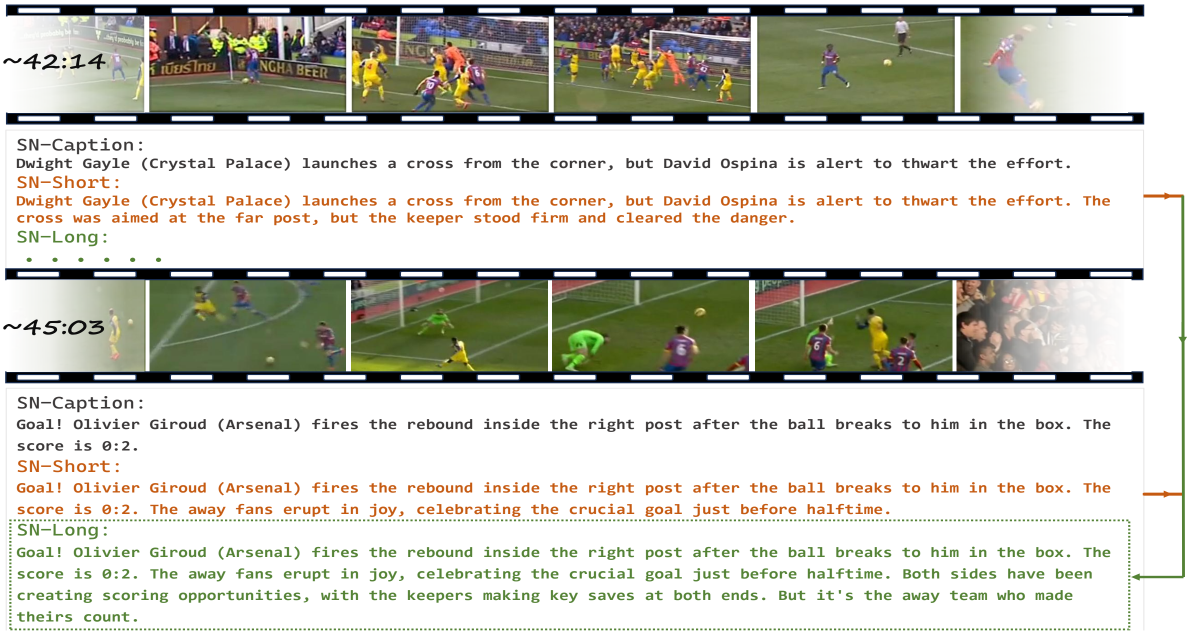
1 Introduction

Recent advances in large language models (LLMs) and vision-language models (VLMs) have sparked growing interest in the automatic generation of soccer commentary. To support this research direction, the SoccerNet dataset (Giancola et al., 2018), has been proposed for soccer video analysis tasks such as action recognition (Silvio Giancola, 2021), camera calibration (Anthony Cioppa, 2021), etc. Based on SoccerNet, the SoccerNet-Caption dataset (Mkhallati et al., 2023) was introduced for soccer video captioning. It contains match videos annotated with timestamped brief commentaries sourced from live text broadcast websites. More recent corpora like GOAL (Qi et al., 2023) and SoccerNet-Echoes (Gautam et al., 2024) extract continuous commentator transcripts from match audio, offering richer prose that spans entire games. However, those audio-based transcripts suffer from background noise, colloquial phrasing, making them suboptimal for fine-grained captioning. A follow-up study identified temporal misalignments in SoccerNet-Caption and proposed a multimodal alignment pipeline to produce MatchTime, accurately timestamped annotations (Rao et al., 2024).

Despite their contributions, existing datasets remain insufficient for practical use in the generation of high-quality soccer commentary. SoccerNet-Caption and MatchTime provide only brief, isolated annotations that lack contextual grounding and do not reflect the complexity of in-game dynamics. GOAL and SoccerNet-Echoes, while richer in linguistic content, suffer from noisy audio, informal phrasing that undermine their reliability for fine-grained video understanding. As a result, none of these datasets adequately captures the semantic richness, tactical nuance, or narrative coherence that characterize expert human commentary.

Furthermore, these data sets fail to model temporal continuity between events. By treating each play or video clip separately, they produce fragmented descriptions that lack connective structure and contextual awareness of the flow of the game. This disconnect prevents models from learning how events build upon one another over time—a crucial element of natural, expert-level sports narration.

To address the above limitations, we introduce two manually curated datasets: SN-Short and SN-Long. SN-Short focuses on enhancing the semantic detail of individual scenes, while SN-Long emphasizes event continuity to support coherent, context-aware commentary generation, as illustrated in Figure 1. In addition, we propose SCORE (Soccer Commentary Generation via Contextual Expansion



SN-Caption:
Dwight Gayle (Crystal Palace) launches a cross from the corner, but David Ospina is alert to thwart the effort.
SN-Short:
Dwight Gayle (Crystal Palace) launches a cross from the corner, but David Ospina is alert to thwart the effort. The cross was aimed at the far post, but the keeper stood firm and cleared the danger.
SN-Long:
. . . .

SN-Caption:
Goal! Olivier Giroud (Arsenal) fires the rebound inside the right post after the ball breaks to him in the box. The score is 0:2.
SN-Short:
Goal! Olivier Giroud (Arsenal) fires the rebound inside the right post after the ball breaks to him in the box. The score is 0:2. The away fans erupt in joy, celebrating the crucial goal just before halftime.
SN-Long:
Goal! Olivier Giroud (Arsenal) fires the rebound inside the right post after the ball breaks to him in the box. The score is 0:2. The away fans erupt in joy, celebrating the crucial goal just before halftime. Both sides have been creating scoring opportunities, with the keepers making key saves at both ends. But it's the away team who made theirs count.

Figure 1: Examples of different dataset contents. Our manually constructed SN-Short dataset contains more detailed and semantically dense commentaries, while SN-Long provides match trends and analytical content.

and Information Retrieval), a novel framework that addresses both fine-grained semantic understanding and global context modeling through the retrieval of relevant past events. Our main contributions are summarized as follows:

- We propose a new task for generating semantically rich and context-aware soccer commentary, and construct two high-quality datasets, **SN-Short** and **SN-Long**[1], to support depth scene understanding and event continuity.

- We introduce **SCORE**, a novel framework that consists of two components: **MatchText**, an automatic commentary expansion pipeline, and **MatchAware**, a retrieval-augmented generation model that integrates visual features and retrieves historical context from a memory bank to produce coherent, high-quality commentary.

- We conducted extensive experiments showing that **SCORE** significantly outperforms existing baselines on both proposed datasets across multiple evaluation metrics.

2 Related Works

Sports Commentary Generation. The goal of sports commentary generation is to produce natural language descriptions that summarize or explain sports content with precision and insight. Early work relied on unimodal models using structured or semi-structured data, often employing template-based or rule-based methods to generate commentary from statistics or play-by-play data (Taniguchi et al., 2019; Kumano et al., 2019; Sadikov et al., 2006). These approaches focused more on describing numeric or tabular data than interpreting visual content. With advances in neural networks and large language models (LLMs), more sophisticated methods have emerged. For example, Chess Commentary (Jhamtani et al., 2018) introduced a dataset of chess games paired with expert commentary and explored end-to-end neural models. LLM-Commentator (Cook and Karakuş, 2024) fine-tuned OpenLLaMA-7B (Geng and Liu, 2023) to generate commentary based on textual game logs. In multimodal settings, early video-to-text systems extracted visual cues and used modular pipelines to fill predefined commentary templates (Kim and Choi, 2020). SoccerNet-Caption (Mkhallati et al., 2023) marked a key advancement by introducing over 37k textual commentaries paired with 471 soccer match videos, using a frozen visual encoder and an LSTM decoder (Hochreiter and Schmidhuber, 1997). MatchTime (Rao et al., 2024) later addressed temporal misalignment in these annotations. More recent datasets such as GOAL (Qi et al., 2023) and SoccerNet-Echoes (Gautam et al., 2024) contain transcripts from real match commentators, offering rich context but posing challenges

[1] Both datasets will be publicly available.

Dataset	Games	Manual Verification	Event Anchored	Historical	Avg Len
SN-Caption	471	✗	✓	✗	23.18
MatchTime	422	✗	✓	✗	24.01
GOAL	20	✓	✗	**✗**	–
SN-Echoes	471	✗	✗	**✗**	–
SN-Short	47	✓	✓	✗	35.10
SN-Long	47	✓	✓	✓	57.81

Table 1: Statistics of different datasets for soccer commentary. **Games** indicates the number of games included in each dataset. **Manual Verification** indicates whether the dataset has been manually verified. **Anchored** denotes whether the captions are anchored to specific events in the match timeline. **Historical** (✓ = Yes, **✗**= Partially, ✗ = No) indicates whether the dataset contains historical or contextual information. For *GOAL* and *SN-Echoes*, such information is included when mentioned by commentators, while *SN-Long* is explicitly designed to provide comprehensive historical and contextual narrations. **Avg Len** shows the average number of words per event. '–' indicates that average length cannot be computed as the data is not event-anchored.

due to their colloquial style, fragmentation, and alignment issues.

Sports Video Understanding. As the availability of broadcast sports content grows, understanding sports videos has become a critical area in computer vision (Thomas et al., 2017). Traditional tasks such as action recognition (Kay et al., 2017) and video classification (Caba Heilbron et al., 2015) were initially developed using datasets focused on everyday activities, limiting their applicability to domain-specific tasks like sports analysis (Wu et al., 2022). In response, sport-specific datasets and methods have emerged (Liu et al., 2020; Giles et al., 2020; Hendry et al., 2020; Li et al., 2021). For soccer, early datasets such as Soccer-ISSIA (D'Orazio et al., 2009) and Soccer Player (Lu et al., 2017) supported player tracking. Larger-scale datasets like SoccerNet (Giancola et al., 2018) and SSET (Feng et al., 2020) offer extensive annotations, with follow-ups like SoccerNet-v2 (Deliege et al., 2021) and SoccerDB (Jiang et al., 2020) expanding event categories. These resources have enabled research into tasks such as action spotting (Cioppa et al., 2020; Anthony Cioppa, 2021), player detection (Vandeghen et al., 2022), and video understanding (Held et al., 2023; Mkhallati et al., 2023).

Memory Network. Memory networks aim to enhance model performance by enabling long-term storage and retrieval of contextual knowledge. Originally proposed for question answering (Weston et al., 2014; Sukhbaatar et al., 2015), early memory models focused on architectural improvements, while later work introduced explicit memory banks for long-range information retrieval (Kumar et al., 2016; Xiong et al., 2016). In dialogue generation, MGMR (Cai et al., 2022) integrated both short- and long-term memory to improve coherence in multi-turn conversations. Memory-based mechanisms have also been successfully applied in computer vision and materials science (Sang et al., 2022; Cao et al., 2022).

Existing soccer commentary models (Mkhallati et al., 2023; Rao et al., 2024) typically generate isolated textual descriptions based only on the current video clip, lacking the ability to model causal or temporal relationships between events. This limits their capacity to offer deeper tactical insights, such as the flow of the game or the causes behind key moments. To address this, we propose a retrieval-augmented generation model that stores and retrieves historical match information, enabling the generation of insightful, context-aware soccer commentary.

3 Benchmark Curation

We manually curate SN-Short and SN-Long datasets to address issues mentioned before by providing detailed, semantically rich, and context-aware commentaries for 47 soccer matches. All commentaries are carefully checked by experienced soccer enthusiasts to ensure accuracy and consistency. Table 1 summarizes key characteristics of soccer commentary datasets in comparison with existing datasets.

3.1 SN-Short

SoccerNet-Caption is an event-anchored dataset that provides brief commentary paired with timestamped events throughout the match, while SoccerNet-Echoes consists of rich, human-

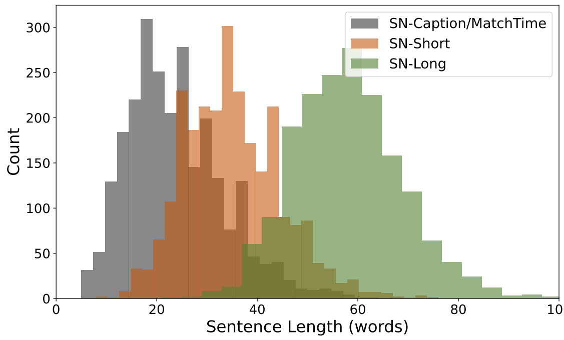


Figure 2: Distribution of sentence lengths in different datasets. The statistics are computed based on the 47 matches shared across all four datasets. *MatchTime* and *SN-Caption* share the same textual content, so their distributions are identical.

transcribed commentary with detailed narrative content. These datasets are complementary in nature. We therefore construct SN-Short leveraging SoccerNet-Caption and SoccerNet-Echoes as the source of our dataset. Specifically, we anchor each event in SoccerNet-Caption and locate all corresponding transcripts in SoccerNet-Echoes within a 30-second time window around the event timestamp. Given that the transcripts are fragmented and highly colloquial, we first manually remove irrelevant content. The remaining texts are then aligned and integrated using LLaMA3(Dubey et al., 2024). After manual review and revision, they are appended to the end of original caption text. Since SoccerNet-Caption contains a great number of visually irrelevant events, such as attendance and ball possession statistics, which cannot be inferred from videos, we remove such events based on strict string matching.

3.2 SN-Long

Considering that the soccer commentary can consist of multiple events to describe the flow of the game, we construct SN-Long, a multi-event, context-aware dataset based on SN-Short. For each target event in SN-Short, we retrieve several semantically similar events that occurred earlier within the same half of the match. Using a few-shot prompting approach with LLaMA3, each retrieved event is individually aggregated with the target event to generate a tactical commentary, providing an in-depth analysis of the match. To ensure the quality and analytical depth of the outputs, we then carefully review and refine each generated commentary to ensure its tactical relevance and consistency.

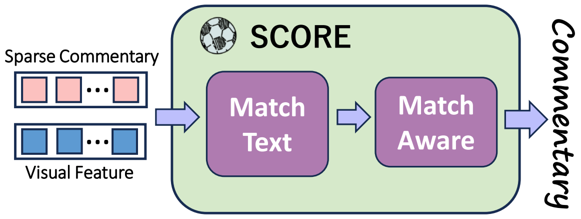


Figure 3: Overview of our proposed SCORE framework for soccer commentary generation. SCORE consists of two main parts: **MatchText** and **MatchAware**. **MatchText** performs commentary expansion, transforming the sparse commentary into enriched, dense commentary. Subsequently, **MatchAware** generates commentary while retrieving relevant historical events, producing dense and context-aware outputs.

3.3 Data Statistics

After careful manual editing and checking, SN-Short contains a total of 2777 video-text pairs, covering key events such as crosses, shots, corner kicks, free kicks, goals, fouls, etc., with most of the visually irrelevant events removed. SN-Long is a multi-event dataset consisting of 1765 events, each associated with an average of 2.84 related events. For every pair, we create a summarizing commentary that captures tactical content and game trends, resulting in more analytical and context-aware commentaries.

Figure 2 shows the commentary length distribution across datasets. SN-Caption and MatchTime are short (10 to 30 words), reflecting sparse, event-focused descriptions. SN-Short is longer and denser, while SN-Long further expands these commentaries, mostly ranging from 50 to 70 words.

4 SCORE

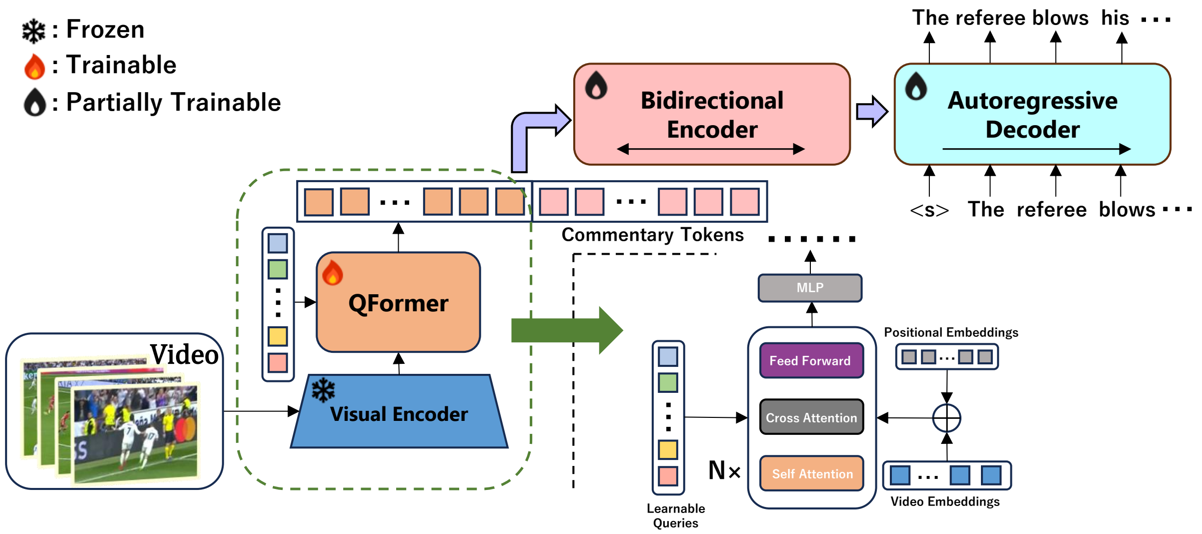
In this section, we present our novel framework, **SCORE**, as illustrated in Figure 3. SCORE consists of two main components: **MatchText**, an automatic commentary expansion pipeline described in Section 4.1, and **MatchAware**, a generation model with an integrated retrieval module, detailed in Section 4.2.

4.1 MatchText

MatchText is a commentary expansion pipeline that bridges the granularity mismatch between visual content and sparse textual commentary in existing datasets.

4.1.1 Problem Formulation

We consider a soccer match video extracted from the SoccerNet-Caption, and denote it as $\mathcal{D} =$



(a) Commentary Expansion Pipeline **(b)Visual Feature Extraction**

Figure 4: MatchText, our proposed automatic commentary expansion pipeline. **(a)** Overview of our proposed pipeline. We use a Q-Former to bridge visual features and the encoder-decoder model, allowing visual features and sparse commentary to guide the generation of dense commentary. **(b)** Details of the Q-Former module.

$\{\mathcal{V}, \mathcal{C}\}$, where $\mathcal{V} = \{V_{1,t_1}, \ldots, V_{n,t_n}\}$, V_{i,t_j} represents the video clip i at timestamp j, and $\mathcal{C} = \{C_{1,t_1}, \ldots, C_{n,t_n}\}$, C_{i,t_j} denotes the commentaries aligned to the timestamps. Our goal is to extract additional semantic features from the visual content that are not explicitly described in the commentary. We apply a Q-Former module(Li et al., 2023) to obtain learnable visual representations from $\mathcal{V}$, i.e., $\mathcal{F} = \{F_{1,t_1}, \ldots, F_{n,t_n}\}$, where F_{i,t_j} denotes the visual feature aligned to the timestamp. The combined input is represented as $\mathcal{I} = \{[F_{1,t_1}; C_{1,t_1}], \ldots, [F_{n,t_n}; C_{n,t_n}]\}$. Output is an enriched version of the commentary, represented as $\mathcal{O} = \{C_{1,t_1} + S_{1,t_1}, \ldots, C_{n,t_n} + S_{n,t_n}\}$. Here, S_{i,t_j} denotes the additional language representation learned from visual features via a set of learnable queries. In summary, the overall pipeline is formalized as:

$$\mathcal{O} = \Phi(Q(\mathcal{V}), \mathcal{C}),$$

where $Q(\mathcal{V})$ represents the Q-Former outputs based on soccer match video $\mathcal{V}$, $\mathcal{C}$ is the original sparse commentary, subsequently expanded by the commentary expansion pipeline Φ.

4.1.2 Architecture

As depicted in Figure 4, we develop our commentary expansion pipeline based on an encoder-decoder architecture. This pipeline enriches the original sparse commentary by integrating learnable visual features.

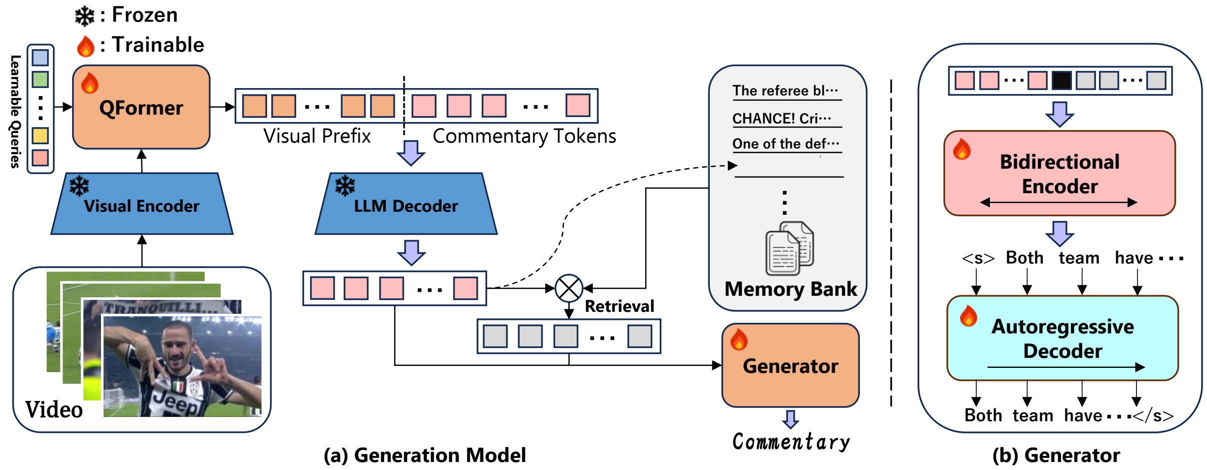
Commentary Expansion Pipeline. Given a sequence of video clips, a pre-trained, frozen visual encoder first extracts low-level visual features. These features are then passed into a trainable Q-Former module equipped with learnable queries to generate compact visual representations. The resulting visual features are concatenated with the original textual commentary and subsequently processed by the encoder-decoder model. Our proposed pipeline generates enriched outputs based on the original commentaries and combined with complementary information from visual features.

Visual Feature Extraction. Each query in the Q-Former interacts with low-level visual features through a multi-layer Transformer architecture that leverages cross-attention mechanisms. The output representations are then passed through an MLP layer to match the dimensionality of the encoder input.

By applying MatchText, we conduct extensive experiments with different visual features to enrich the sparse commentary. Specifically, we expand the remaining part of SoccerNet-Caption, covering 424 soccer match videos and resulting in 27,207 dense video-text pairs.

4.2 MatchAware

MatchAware is designed to generate dense and context-aware commentary by first offering an initial description of the current event and then retrieving relevant historical events from a memory bank to enrich the output.



(a) Generation Model

(b) Generator

Figure 5: The overview of our proposed commentary generation model, MatchAware. **(a)** The initial generated commentaries are stored in the memory bank for further usage. **(b)** The context-aware commentary generator is based on an encoder-decoder model. By combining the retrieved event from the memory bank and the current generated commentary, the generator produces analytical content.

4.2.1 Problem Formulation

Given a video-text pair denoted as $\mathcal{D} = \{\mathcal{V}, \mathcal{C}\}$, where $\mathcal{V} = \{V_{1,t_1}, \dots, V_{n,t_n}\}$ and $\mathcal{C} = \{C_{1,t_1}, \dots, C_{n,t_n}\}$ represent the video clips and commentaries, and their corresponding timestamps, separately. Our goal is to develop a video-language model augmented with a retrieval module. For each video clips V_{i,t_j}, the model generates an initial commentary commentary $\hat{C}_{i,t_j}$, then retrieve relevant event from memory bank $B = \{\hat{C}_{1,t_1}, \dots, \hat{C}_{j,t_j}\}$ where $j < i$, thus obtain a commentary $\hat{C}_{l,t_i} = R(\hat{C}_{i,t_i}, \hat{C}_{j,t_j})$, where R denotes the Retrieval function, that summarizes both match trends and long-term game content. The final output is represented as $\mathcal{O} = [\hat{C}_{i,t_i}; \hat{C}_{l,t_i}] = [\phi(V_{i,t_i}); R(\hat{C}_{1,t_1}, \dots, \hat{C}_{j,t_j})]$

4.2.2 Architecture

As shown in Figure 5, our proposed architecture contains two main components. A video-language generation module that produces commentary based on the current video clip, retrieval-augmented generator that generates long-term historical context by leveraging previous events.

We extract frame-wise visual features from the input video clips using a pre-trained frozen visual encoder. These visual features are then processed by Q-Former, where a set of learnable queries interact with the visual inputs through a Transformer to extract high-level semantic representations. For the commentary generation, we adopt an architecture similar to MatchVoice (Rao et al., 2024), where these representations are used to guide the initial soccer generation.

Meanwhile, the retrieval-augmented generator maintains a memory bank that stores previously generated commentaries and their corresponding timestamps. Upon generating initial commentary $\hat{C}_i$, it searches semantically relevant past events in the memory bank. In order to simulate the situation of real human commentary, we adopt a time-sensitive similarity score. If the time difference between the current timestamp and the retrieved event is within a certain threshold, i.e., $\Delta t \leq \mathcal{T}$, the final score is calculated as $Score = sim(\hat{C}_i, \hat{C})$. Otherwise, the final score is penalized by a temporal decay factor to reduce the influence of events that are temporally distant, i.e., $Score = sim(\hat{C}_i, \hat{C}) \cdot e^{-\alpha\cdot\left(\frac{\Delta t - \mathcal{T}}{\rho}\right)}$, where α stands for the decay rate, ρ is the scaling factor that adjusts the sensitivity of the decay.

5 Experiments

In this section, we present our experiments and results. We first explain how we chose the visual features used in our MatchText, as detailed in Section 5.1. Based on the selected features, we then conduct extensive experiments and ablation studies for SCORE, which are reported in Section 5.2.

5.1 Experiment 1

The goal of this experiment is to compare different visual features that contribute to better commentary expansion in the MatchText pipeline, and to select the one that yields the best performance for use in subsequent experiments.

Dataset	Visual F	B@1	B@4	M	R-L	C
SN-C	-	48.66	45.45	57.37	**68.95**	1.16
SN-S	Baidu	59.94	**51.80**	**62.32**	64.79	**3.75**
	ResNet(2)	59.71	50.76	62.27	63.51	3.71
	ResNet(5)	59.68	51.44	62.17	64.02	3.74
	CLIP	**60.01**	51.70	62.10	64.48	3.74

Table 2: Performance comparison of different visual features on our commentary expansion pipeline. For clarity, we also directly compare SoccerNet-Caption with SN-Short on corresponding parts. Best scores are in **red**, second best in blue.

Implementation Details. We adopt three off-the-shelf visual feature extractors: CLIP (Radford et al., 2021), Baidu (Zhou et al., 2021), and ResNet (He et al., 2016). CLIP features are extracted at 2 FPS, Baidu at 1 FPS, and ResNet at both 2 FPS and 5 FPS. We set the number of learnable queries in the Q-Former architecture to 32 and use BART (Lewis et al., 2019) as the backbone generation model, which is fine-tuned on the training set of SN-Short for 20 epochs using a learning rate of 5e-6. We also freeze its lower encoder and decoder layers during training. All experiments are conducted on a single NVIDIA RTX A100 GPU. We adopt SoccerNet-Caption as the textual input and evaluate the results on the test set of SN-Short.

Results. As shown in Table 2, integrating visual features improves BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015) scores except ROUGE-L (Lin, 2004). In particular, the significant improvement in CIDEr suggests that the outputs contain more distinctive and semantically dense content. Based on the result, we adopt Baidu features to expand the sparse commentary dataset.

5.2 Experiment 2

In this experiment, we compare our SCORE with baselines to validate its effectiveness. We adopt SN-Short and SN-Long as benchmarks and conduct two separate evaluations accordingly.

Implementation Details. In addition to the visual features described in Section 5.1, we also include C3D (Tran et al., 2015). We use LLaMA-3-8B as the LLM decoder and BART for the backbone of the retrieval-augmented generator. The 32 of learnable queries from Q-Former are projected to a 4096-dimensional space to match the hidden size of the LLM decoder layers. The initially generated commentaries are stored in the memory bank for context-aware content generation. The commentary generation model is trained for 40 epochs with a learning rate 1e-5. The retrieval-augmented generator is trained on the training set of SN-Long for 20 epochs with a learning rate 1e-5. The time-sensitive similarity function is parameterized with $\mathcal{T} = 600$, $\alpha = 0.3$, and $\rho = 900$. All experiments are conducted on a single NVIDIA RTX A100 GPU.

Baseline. We adopt the MatchVoice (Rao et al., 2024) as our baseline architecture and train it separately on the SoccerNet-Caption and training set of SN-Short.

Results. We adpot BLEU, METEOR, ROUGE and CIDEr score to evaluate the quality of generated commentaries. Specifically, we conduct experiments with two settings: with and without the retrieval-augmented generator. The results are evaluated on both SN-Short and SN-Long to present a comprehensive view of the performance. Table 3 compares the results with different visual features used.

We first conduct experiments using the SCORE model without the retrieval-augmented generator, and compare it against MatchVoice trained on SoccerNet-Caption and SN-Short. The results show that *SCORE without Retrieval* outperforms the baselines across all of the metrics. This shows the effectiveness of MatchText in expanding commentary. Next, we investigate the proposed retrieval-augmented generator and conduct extensive comparisons on the SN-Long dataset. The results show that *SCORE* achieves the best performance across all evaluation metrics and all types of visual features.

Our extensive experiments demonstrate that (i) MatchTime, our proposed automatic expansion pipeline, produces more semantically dense and informative commentaries compared to the SoccerNet-Caption. (ii) The retrieval-augmented generator significantly enhances performance across all metrics, demonstrating its effectiveness in generating context-aware commentary. (iii) The proposed SN-Short and SN-Long are very challenging datasets, containing more analytical and context-aware commentaries. This leads to a notable performance drop in models without retrieval enhancement, which also reveals the lack of research on event memory in the current soccer commentary generation approaches.

5.3 Ablation studies

Retrieval-Augmented Generator. We conduct the ablation study on our retrieval-augmented genera-

Method	Visual Features	BLEU-1	BLEU-4	METEOR	ROUGE-1	ROUGE-L	CIDEr
			SN-Short				
MatchVoice (Trained on Soccernet-Caption)	C3D	14.55	2.90	8.42	21.67	16.18	5.35
	Baidu	15.78	3.68	9.07	23.90	18.20	7.36
	ResNet(2)	16.67	2.49	9.08	22.12	16.03	6.55
	ResNet(5)	15.49	3.34	8.95	22.12	16.82	8.09
	CLIP	13.79	1.48	7.84	19.21	14.04	4.21
MatchVoice (Fine-tuned on SN-Short)	C3D	19.79	1.85	9.45	22.74	15.94	5.06
	Baidu	21.11	2.44	10.18	24.77	17.50	7.26
	ResNet(2)	19.82	2.53	9.89	23.50	16.82	6.44
	ResNet(5)	20.20	2.32	9.89	23.50	16.87	5.99
	CLIP	19.66	2.12	9.78	23.11	16.62	5.86
SCORE without Retrieval	C3D	19.63	3.12	10.23	24.09	17.79	9.71
	Baidu	**23.43**	**4.72**	**11.48**	**27.47**	**20.49**	**14.35**
	ResNet(2)	21.27	3.35	10.67	24.83	18.08	11.10
	ResNet(5)	20.34	2.69	10.60	25.29	18.41	10.81
	CLIP	18.49	2.27	9.77	22.88	16.97	9.00
			SN-Long				
MatchVoice (Trained on Soccernet-Caption)	C3D	9.10	1.75	18.86	17.24	10.46	0.38
	Baidu	9.75	2.17	19.66	17.63	11.18	0.27
	ResNet(2)	11.92	1.57	19.44	17.55	11.31	0.51
	ResNet(5)	10.34	1.99	19.08	17.37	11.05	0.72
	CLIP	8.79	1.00	16.51	14.89	9.33	0.75
MatchVoice (Fine-tuned on SN-Short)	C3D	16.67	1.54	19.52	17.50	12.38	1.10
	Baidu	17.25	1.90	20.83	18.67	13.21	1.35
	ResNet(2)	16.24	1.98	20.29	18.48	12.74	1.70
	ResNet(5)	17.13	1.84	20.16	18.15	12.84	1.45
	CLIP	17.76	1.93	20.00	18.43	12.70	1.66
SCORE	C3D	30.80	6.34	26.55	24.17	21.84	6.90
	Baidu	**35.38**	**8.45**	**29.44**	**27.01**	**24.66**	**11.87**
	ResNet(2)	33.07	6.06	27.78	25.18	23.19	8.87
	ResNet(5)	31.95	5.95	27.46	25.02	22.62	7.85
	CLIP	29.03	4.77	25.30	23.05	20.09	4.83

Table 3: Evaluation results of different visual features on SN-Short and SN-Long. Best scores are in **red**, second best in blue.

Method	Visual F	B@1	B@4	M	R-1	R-L	C
SCORE w/o Retrieval	C3D	15.13	2.19	19.83	18.19	12.69	1.66
	Baidu	19.12	3.67	23.01	21.19	15.23	1.94
	ResNet(2)	17.55	2.41	21.53	19.63	13.97	1.76
SCORE	C3D	30.80	6.34	26.55	24.17	21.84	6.90
	Baidu	**35.38**	**8.45**	**29.44**	**27.01**	**24.66**	**11.87**
	ResNet(2)	33.07	6.06	27.78	25.18	23.19	8.87

Table 4: Ablation study of the retrieval module in our SCORE framework on the SN-Long dataset.

tor. We compare the performance of the full model with a variant that removes the retrieval module across different visual features.

As depicted in Table 4, compared to *SCORE w/o Retrieval*, *SCORE* shows significant improvements across all metrics with all visual features. This highlights the effectiveness of our proposed retrieval-augmented generator, which enables the model to produce more context-aware and semantically rich commentary by leveraging relevant historical information.

6 Conclusion

In this paper, we propose two manually curated datasets, SN-Short and SN-Long. SN-Short focuses on detailed scene-level descriptions, while SN-Long captures the continuity of events throughout a match. To address the challenges in both datasets, we introduce SCORE, which comprises two key components: MatchText and MatchAware. MatchText leverages visual features and sparse commentaries to generate dense training commentaries. MatchAware then produces context-aware outputs by incorporating historical event information. Our experimental results demonstrate that integrating visual features and historical events enhances the analytical depth of soccer commentaries. Furthermore, our retrieval-augmented generator in MatchAware achieves consistent improvements across all evaluation metrics, highlighting the critical role of temporal and contextual information in sports commentary generation.

Limitations

First, we observe that the model struggles to distinguish between visually similar events in certain situations (e.g., free kicks and corner kicks, goals and saved shots) which is consistent with what is reported in previous studies. Fine-tuning on more specific training data may help solve this issue.

Second, due to the absence of player tracking and identification module, the model could not correctly identify the players and output their names.

Third, due to computing resources and dataset limitations, our memory bank can only store textual information. As a result, the retrieval-augmented generator performs tactical analysis based solely on text, and its contextual range is limited to within a single half of the match. Ideally, the system should extract relevant information from video and perform cross-match retrieval, to achieve deeper and longer-term analysis. We hope that larger-scale soccer datasets will become available in the future, which could support further research in automatic commentary generation.

References

Floriane Magera Silvio Giancola Olivier Barnich Bernard Ghanem Marc Van Droogenbroeck Anthony Cioppa, Adrien Deliège. 2021. Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Xiaoyu Cai, Yao Fu, Hong Zhao, Weihao Jiang, and Shiliang Pu. 2022. Memory graph with message rehearsal for multi-turn dialogue generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 108–117, New York, NY, USA. Association for Computing Machinery.

Yong Cao, Xin Sha, Xianwei Bai, Yan Shao, Yuanhong Gao, Yu-Ming Wei, Lingqiang Meng, Ni Zhou, Jin Liu, Bo Li, and 1 others. 2022. Ultralow light-power consuming photonic synapses based on ultrasensitive perovskite/indium-gallium-zinc-oxide heterojunction phototransistors. *Advanced Electronic Materials*, 8(3):2100902.

Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B Moeslund. 2020. A context-aware loss function for action spotting in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13126–13136.

Alec Cook and Oktay Karakuş. 2024. Llm-commentator: Novel fine-tuning strategies of large language models for automatic commentary generation using football event data. *Knowledge-Based Systems*, 300:112219.

Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. 2021. 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.

T. D'Orazio, M. Leo, N. Mosca, P. Spagnolo, and P.L. Mazzeo. 2009. A semi-automatic system for ground truth generation of soccer video sequences. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 559–564.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Na Feng, Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, Yizhu Zhao, Yunfeng He, and Tao Guan. 2020. Sset: a dataset for shot segmentation, event detection, player tracking in soccer videos. *Multimedia Tools and Applications*, 79:28971–28992.

Sushant Gautam, Mehdi Houshmand Sarkhoosh, Jan Held, Cise Midoglu, Anthony Cioppa, Silvio Giancola, Vajira Thambawita, Michael A. Riegler, Pål Halvorsen, and Mubarak Shah. 2024. Soccernet-echoes: A soccer game audio commentary dataset. *Preprint*, arXiv:2405.07354.

Xinyang Geng and Hao Liu. 2023. Openllama: An open reproduction of llama.

Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. 2018. 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.

Brandon Giles, Stephanie Kovalchik, and Machar Reid. 2020. A machine learning approach for automatic detection and classification of changes of direction from player tracking data in professional tennis. *Journal of sports sciences*, 38(1):106–113.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Jan Held, Anthony Cioppa, Silvio Giancola, Abdullah Hamdi, Bernard Ghanem, and Marc Van Droogenbroeck. 2023. 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 5086–5097.

Danica Hendry, Kevin Chai, Amity Campbell, Luke Hopper, Peter O'Sullivan, and Leon Straker. 2020. Development of a human activity recognition system for ballet tasks. *Sports medicine-open*, 6:1–10.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, Graham Neubig, and Taylor Berg-Kirkpatrick. 2018. Learning to generate move-by-move commentary for chess games from large-scale social forum data. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Yudong Jiang, Kaixu Cui, Leilei Chen, Canjin Wang, and Changliang Xu. 2020. 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.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and 1 others. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Byeong Jo Kim and Yong Suk Choi. 2020. Automatic baseball commentary generation using deep learning. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, page 1056–1065, New York, NY, USA. Association for Computing Machinery.

Tadashi Kumano, Manon Ichiki, Kiyoshi Kurihara, Hiroyuki Kaneko, Tomoyasu Komori, Toshihiro Shimizu, Nobumasa Seiyama, Atsushi Imai, Hideki Sumiyoshi, and Tohru Takagi. 2019. Generation of automated sports commentary from live sports data. In *2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–4.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, pages 1378–1387. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, pages 19730–19742. PMLR.

Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. 2021. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13536–13545.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Shenglan Liu, Xiang Liu, Gao Huang, Hong Qiao, Lianyu Hu, Dong Jiang, Aibin Zhang, Yang Liu, and Ge Guo. 2020. Fsd-10: A fine-grained classification dataset for figure skating. *Neurocomputing*, 413:360–367.

K. Lu, J. Chen, J. J. Little, and H. He. 2017. Light cascaded convolutional neural networks for accurate player detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–13.

Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. 2023. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics*, pages 311–318.

Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Bin Xu, Lei Hou, Juanzi Li, and 1 others. 2023. Goal: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5391–5395.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and

1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

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

Aleksander Sadikov, Martin Možina, Matej Guid, Jana Krivec, and Ivan Bratko. 2006. Automated chess tutor. In *International Conference on Computers and Games*, pages 13–25. Springer.

Hongrui Sang, Rong Jiang, Zhipeng Wang, Yanmin Zhou, and Bin He. 2022. A novel neural multi-store memory network for autonomous visual navigation in unknown environment. *IEEE Robotics and Automation Letters*, 7(2):2039–2046.

Julia Georgieva Johsan Billingham Andreas Serner Kerry Peek Bernard Ghanem Marc Van Droogenbroeck Silvio Giancola, Anthony Cioppa. 2021. Towards active learning for action spotting in association football videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, and 1 others. 2015. End-to-end memory networks. *Advances in neural information processing systems*, 28.

Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. 2019. Generating live soccer-match commentary from play data. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Graham Thomas, Rikke Gade, Thomas B Moeslund, Peter Carr, and Adrian Hilton. 2017. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding*, 159:3–18.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the International Conference on Computer Vision*, pages 4489–4497.

Renaud Vandeghen, Anthony Cioppa, and Marc Van Droogenbroeck. 2022. Semi-supervised training to improve player and ball detection in soccer. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3481–3490.

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

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Fei Wu, Qingzhong Wang, Jiang Bian, Ning Ding, Feixiang Lu, Jun Cheng, Dejing Dou, and Haoyi Xiong. 2022. A survey on video action recognition in sports: Datasets, methods and applications. *IEEE Transactions on Multimedia*, 25:7943–7966.

Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2397–2406. JMLR.org.

Xin Zhou, Le Kang, Zhiyu Cheng, Bo He, and Jingyu Xin. 2021. Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. *arXiv preprint arXiv:2106.14447*.