Visual and Memory–Augmented Soccer Commentary Generation

Haoran Sun **Natthawut Kertkeidkachorn** **Kiyoaki Shirai**
Japan Advanced Institute of Science and Technology
magotsugeshikaru@gmail.com {natt, kshirai}@jaist.ac.jp

Abstract

Automatic soccer commentary generation aims to bridge the gap between raw visual content and professional, tactical commentary. However, existing datasets tend to produce incomplete commentary that lacks semantic richness and fails to convey the full visual information present in standard video clips. To address these limitations, we propose two manually curated datasets: SN-Short, which enhances scene-level semantic descriptions, and SN-Long, which captures event continuity for context-aware commentary. Based on these, we design a commentary augmentation pipeline that transforms incomplete annotations into **MatchText**, a semantically complete and structurally standardized dataset. Leveraging this supervision, we introduce **MatchAware**, a generation model that incorporates contextual cues from previous events to produce coherent commentary aligned with the visual flow of the game. Experimental results show that proposed approach significantly outperforms existing baselines on the constructed 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, both SoccerNet-Caption (Mkhallati et al., 2023) and SoccerReplay-1988 (Rao et al., 2025b) provide concise timestamped commentaries from live text websites, with the latter offering a substantially larger corpus. To address timestamp misalignment in SoccerNet-Caption, a follow-up study release SN-Caption-test-align and further automatically correct the full dataset as MatchTime. They also propose MatchVoice, a generation model to evaluate soccer commentary quality (Rao et al., 2024).

Datasets such as GOAL (Qi et al., 2023) and SoccerNet-Echoes (Gautam et al., 2024) provide human-transcribed commentaries from match audio, offering richer prose that spans entire games. However, those audio-based transcripts suffer from background noise and colloquial phrasing, making them suboptimal for fine-grained captioning.

Existing methods for soccer commentary generation typically rely on video–text pairs from short (30~60s) clips with brief annotations (Mkhallati et al., 2023; Rao et al., 2024; Li et al., 2025). These annotations are usually collected from live text websites and are designed to highlight key events within each clip. As a result, they are typically restricted to a single concise sentence, averaging around 24 words. While sufficient for event spotting, such brevity fails to capture the rich visual context, providing only coarse-grained supervision for commentary generation.

Furthermore, existing datasets 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. Consequently, current methods produce only fragmented, moment-level commentary, with no capacity to reference past events or connect ongoing plays to earlier developments—these crucial elements of expert-level sports narration that are still absent in current approaches.

To address the above limitations, we introduce two manually curated datasets: SN-Short and SN-Long. SN-Short enriches scene-level textual descriptions to bridge the gap between rich visual content and concise annotations. SN-Long, built upon SN-Short, connects related events to capture temporal continuity and contextual flow, as illustrated in Figure 1. Based on these datasets, we construct a new dataset, **MatchText**, obtained by aug-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10614–10629

Figure 1: Examples of different dataset contents. Our manually constructed SN-Short dataset contains more detailed and semantically dense commentaries, while SN-Long enhances coherence and tactical depth by leveraging prior event annotations.

menting incomplete scene-level annotations with fine-grained semantic information grounded in the video. Furthermore, we propose **MatchAware**, a commentary generation model that retrieves relevant historical visual events to produce fluent and temporally grounded commentary. 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 deep scene understanding and event continuity.
- We develop a commentary augmentation pipeline to construct **MatchText**, a large-scale dataset with semantically complete and structurally standardized soccer commentary.
- We propose **MatchAware**, a memory-augmented commentary generation model that retrieves relevant historical visual events to generate coherent and context-aware commentary.
- We conduct extensive experiments demonstrating that the proposed approach significantly outperforms existing baselines on both proposed datasets across multiple evaluation metrics.

[1]Data is available at https://github.com/JAIST-KnOWLab/Augmented_Soccer

2 Related Work

Sports Commentary Generation Early sports commentary generation focuses on structured data using template- or rule-based systems (Taniguchi et al., 2019; Kumano et al., 2019; Sadikov et al., 2006), which lack interpretability of visual content. With neural networks and LLMs, LLM-Commentator (Cook and Karakuş, 2024) fine-tunes OpenLLaMA on textual logs. In multimodal settings, early work adopts modular pipelines that extract visual cues to populate predefined templates (Kim and Choi, 2020). SoccerNet-Caption (Mkhallati et al., 2023) is a milestone with 37k short commentaries collected from text live website aligned to videos, later refined by MatchTime (Rao et al., 2024) for temporal precision. UniSoccer (Rao et al., 2025b) expands the dataset scale as SoccerReplay-1988 and unifies classification and generation tasks, while TimeSoccer (You et al., 2025) explores end-to-end match-level captioning. SoccerComment (Li et al., 2025) retrieves similar past scenes and leverages commentary paradigms to improve generation accuracy. Some corpora, including GOAL (Qi et al., 2023) and SoccerNet-Echoes (Gautam et al., 2024), provide long-form human-transcribed narratives, which are valuable for training but face challenges in alignment, consistency, and noise.

Retrieval-Augmented Multimodal Generation Retrieval-Augmented Generation (RAG) models

Dataset	Commentaries(Games)	Manual Verification	Event Anchored	Historical	Avg Len
GOAL	– (20)	✓	✗	✗	–
SN-Caption-test-align	3.2k (49)	✓	✓	✗	23.41
SN-Short (Ours)	2.8k (47)	✓	✓	✗	35.10
SN-Long (Ours)	5k (47)	✓	✓	✓	57.81
MatchText (Ours)	27k (424)	✗	✓	✗	34.97
SN-Caption	37k (471)	✗	✓	✗	23.18
MatchTime	33k (422)	✗	✓	✗	24.01
SN-Echoes	– (471)	✗	✗	✗	–
SoccerReplay-1988	150k (1988)	✗	✓	✗	–

Table 1: Summary of soccer commentary datasets. **Commentaries (Games)** is the number of video-text pairs ('–' means no event anchoring, so commentary count is unavailable). **Manual Verification** shows whether annotations are manually checked. **Event Anchored** shows whether commentaries are aligned with specific match events. **Historical** (✓ = Yes, ✗= Partial, ✗ = No) denotes inclusion of contextual information. **Avg Len** is the average word count per event. '–' for SoccerReplay-1988, which is not publicly available.

have been widely studied in natural language processing by retrieving relevant external knowledge to augment generation (Lewis et al., 2020). Recently, RAG has been adapted to multimodal tasks. Models like RA-CM3 (Yasunaga et al., 2023) and REVEAL (Hu et al., 2023) show that retrieving relevant image-text pairs can improve performance on visual question answering and multimodal generation. In the domain of soccer commentary generation, GOAL (Qi et al., 2023) incorporates external knowledge retrieval to enhance the informational richness of commentaries. SoccerAgent (Rao et al., 2025a) further employs retrieval techniques to construct a comprehensive soccer agent for question answering tasks. Different from these, we focus on retrieval over event-level visual context within the match itself. By retrieving historically relevant visual events, our method generates analytical and in-depth commentaries that reflect match dynamics and tactical interpretations.

3 Benchmark Curation

We manually curate SN-Short and SN-Long to address the issues mentioned above by progressively enhancing the quality of commentary for 47 soccer games. **SN-Short** provides detailed and semantically rich scene-level commentaries, while **SN-Long** builds upon SN-Short by linking related events to construct context-aware commentaries with temporal continuity. All annotations are manually verified and refined by three soccer fans with over 10 years of experience. Our datasets are the largest soccer commentary benchmarks with manual verification to date. Table 1 summarizes key characteristics of soccer commentary datasets in comparison with existing datasets. Details of dataset construction are provided in Appendix A.1.

3.1 SN-Short

To address the limitations of existing datasets in providing informative and detailed commentary within a video clips, we construct SN-Short by leveraging SoccerNet-Caption and SoccerNet-Echoes. The former provides brief event-timestamped commentaries, while the latter offers dense human-transcribed narratives without event-anchored alignment. These datasets are complementary in nature. For each timestamped event in SoccerNet-Caption, we retrieve transcripts from SoccerNet-Echoes(±15s window). Given that the transcripts are fragmented and highly colloquial, we first manually remove irrelevant/noisy segments. The remaining content is then consolidated using LLaMA3 (Dubey et al., 2024) as an additional context, appended to the original caption, subsequently reviewed and refined by annotators for fluency and consistency. We also discard visually irrelevant events (e.g., attendance, ball possession) using strict string-matching rules.

3.2 SN-Long

Existing soccer datasets primarily offer shallow, clip-level captions, lacking summaries of inter-event relationships or deeper tactical insights. To address this gap, we build **SN-Long** on top of SN-Short as a multi-event, context-aware dataset. For each target event, we retrieve semantically related prior events within the same match-half. We first extract 17 summary paradigms from transcripts of authentic human commentary, which cover diverse scenarios in soccer games. These paradigms are then formulated into concise and standardized language to serve as high-quality exemplars. Using these exemplars, we use LLaMA3 to aggregate the retrieved events and the target event into a tacti-

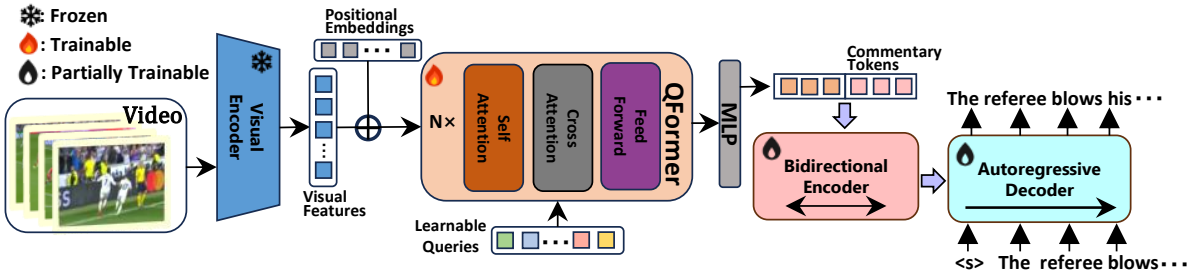

Figure 2: Overview of our commentary augmentation pipeline. Given pre-processed video features and concise textual descriptions, the pipeline generates detailed and structurally standardized commentary that captures fine-grained and discriminative visual information from each video clip.

cal commentary providing in-depth analysis. All outputs are manually reviewed for coherence, relevance, and quality.

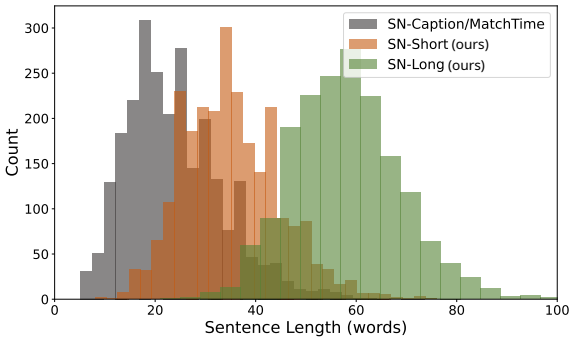

Figure 3: Distribution of sentence lengths in different datasets. The statistics are computed based on the 47 matches shared across all four datasets. *MatchTime* and *SN-Caption* share the same textual content, so their distributions are identical.

3.3 Data Statistics

After manual verification, SN-Short contains 2,777 video-text pairs covering key soccer events (e.g., crosses, shots, set pieces, goals, and fouls), with visually irrelevant events removed.

Based on SN-Short, we construct SN-Long by filtering out context-independent events, resulting in 1,765 video-text pairs, each paired with an average of 2.84 historical events, totaling 5,006 prior contextual segments. Details of the annotation quality evaluation are provided in Appendix A.2

Figure 3 shows the distribution of commentary lengths across datasets. SN-Caption (Mkhallati et al., 2023) and MatchTime (Rao et al., 2024) contain brief, incomplete descriptions (typically 10–30 words) focused on isolated events. SN-Short provides more detailed commentaries, while SN-Long further extends them with contextual information, with most entries ranging from 50 to 70 words.

4 Commentary Augmentation

We design a commentary augmentation pipeline to bridge the granularity mismatch between visual content and textual commentary. Existing commentary frequently misses important information observable from the video clips (Rao et al., 2024; Mkhallati et al., 2023). Our pipeline targets visual information that is present in the video clips but absent from the original textual descriptions, enriching the commentary with such missing semantics. Through this augmentation process, we construct a dataset that is both content-detailed and closely aligned with the visual evidence, while maintaining a standardized and consistent structure.

4.1 Problem Formulation

Given a soccer match video segmented into timestamped clips $\mathcal{V} = \{V_1, \ldots, V_n\}$ and their corresponding incomplete commentary $\mathcal{C} = \{C_1, \ldots, C_n\}$, the task is to generate detailed and complete commentary $\mathcal{O} = \{C'_1, \ldots, C'_n\}$. Each output C'_i is obtained by semantically augmenting the original commentary C_i with additional visual semantics S_i derived from the video content, which we abstractly denote as $C'_i = f(C_i, S_i)$, where $f(\cdot)$ represents a semantic augmentation.

The visual semantics S_i are implicitly captured by latent visual features extracted from the corresponding video clip. Specifically, we apply a Q-Former (Li et al., 2023) to obtain visual representations F_i from V_i, which are then combined with the textual input C_i. A commentary augmentation model Φ instantiates the function $f(\cdot)$ and generates the enriched commentary: $C'_i = \Phi(F_i, C_i)$.

4.2 Architecture

As depicted in Figure 2, we develop our commentary augmentation pipeline based on an encoder–decoder architecture, fine-tuned on SN-Short. Taking textual descriptions as the backbone

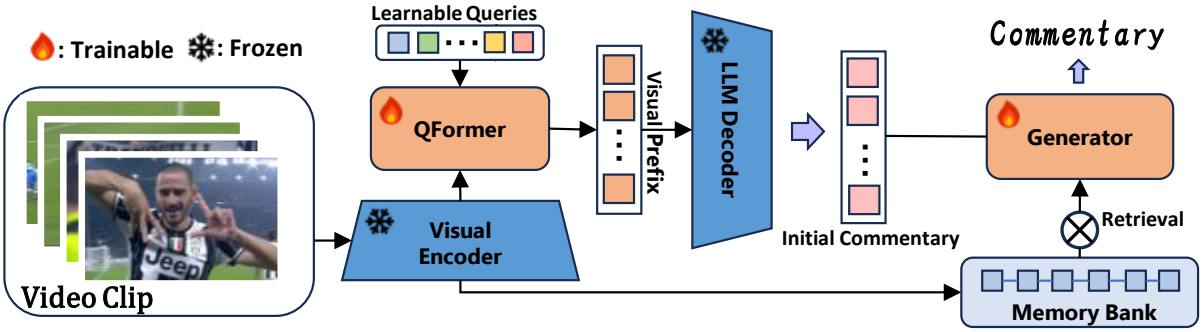

Figure 4: The overview of our proposed commentary generation model, MatchAware. It generates an initial commentary and a context-aware commentary through an LLM-based decoder and a generator, respectively, enabling more detailed and in-depth soccer commentary generation.

and integrating visual features as complementary inputs, the pipeline enriches commentary with additional semantics while preserving the standardized, structured format of the original data.

Visual Feature Extraction Given a video clip V_i, we extract frame-level features $f_i = \texttt{VE}(V_i)$ using a pre-trained, frozen visual encoder. These features are then passed to a Q-Former (Li et al., 2023) equipped with K learnable queries $\{q_k\}_{k=1}^{K}$, producing compact visual representations $F_i = \texttt{QFormer}(f_i, \{q_k\})$. The resulting $F_i \in \mathbb{R}^{K\times d}$ is projected via an MLP to match the encoder input dimension.

Commentary Augmentation For each event-aligned pair (V_i, C_i), we concatenate the projected visual features F_i and the token embeddings of C_i to form the encoder input $[F_i; \texttt{Embed}(C_i)]$. The encoder processes this joint representation to produce hidden states $h_i = \texttt{Encoder}([F_i; \texttt{Embed}(C_i)])$, which the decoder attends to in generating the enriched commentary $C_i' = \texttt{Decoder}(h_i)$.

Through this process, we develop a new dataset, named **MatchText**, which offers semantically complete and structurally standardized textual data. Containing 27,207 video–text pairs from 424 games, MatchText serves as the foundational supervision for training the MatchAware model.

5 MatchAware

In this section, we present our commentary generation model, **MatchAware**, which produces detailed and context-aware commentary by comprehensively describing the visual content of the current video clip and retrieving relevant historical visual events. MatchAware first offers an initial description of the current event and then retrieves relevant historical events from a memory bank of video features to enrich the output.

5.1 Problem Formulation

We aim to enhance commentary generation with memory-based context modeling. For each timestamp t_i, given the corresponding video clip V_{i,t_i}, we extract visual features F_i using a Q-Former. These features are projected to a frozen LLM decoder, which generates an initial commentary $\tilde{C}_{i,t_i} = \phi_{\text{init}}(F_i)$.

To incorporate historical context, we maintain a memory bank $\mathcal{M}_i = \{F_{1,t_1}, \ldots, F_{j,t_j} \mid 1 \le j < i\}$ of encoded video features from previous timestamps, where $j < i$. A retrieval function $R(\cdot)$ selects the most relevant historical visual feature $F_l \in \mathcal{M}$ based on event-level semantic association with the current clip feature F_i. The retrieved feature captures long-term match dynamics and related historical patterns. The generator then takes the current visual feature F_i, the retrieved historical feature F_l, and their temporal distance Δt as joint inputs. The context-aware commentary is denoted as $\hat{C}_{i,t_i} = \phi_{\text{ctx}}(F_i, F_l, \Delta t)$. The output is denoted as $\mathcal{O}_{i,t_i} = [\tilde{C}_{i,t_i}; \hat{C}_{i,t_i}]$.

5.2 Architecture

As shown in Figure 4, MatchAware consists of three components: (i) an event-level video–language generator that produces an initial commentary grounded in the current video clip; (ii) a visual event retriever that selects relevant historical video clips from a memory bank based on the current clip; and (iii) a retrieval-augmented generator that incorporates long-term match context using the retrieved visual events.

Video–language Generator. Following the same setup in Section 4.1, we obtain high-level visual

features $F_i = \texttt{QFormer}(f_i, \{q_k\})$ from video clip V_i and project them via an MLP to obtain prefix embeddings P_i. For commentary generation, we adopt an architecture similar to MatchVoice (Rao et al., 2024), where P_i is prepended to the input of a frozen LLM decoder. Conditioned on these prefix embeddings, the decoder generates an initial event-level commentary $\tilde{C}_{i,t_i}$ describing the visual content of the current clip.

Visual Event Retriever. The visual event retriever aims to identify historical events that are semantically associated with the current video clip. Given an anchor event feature F_i, positive events F_i^+ are selected from semantically related events, while negative events F_i^- are sampled from unrelated events in the memory bank $\mathcal{M}$. We optimize:

$$\mathcal{L}_{\text{ret}} = \max\Big(0, d\big(g(F_i,0), g(F_i^+, \Delta t^+)\big) - d\big(g(F_i,0), g(F_i^-, \Delta t^-)\big) + m\Big)$$

Here, $g(\cdot)$ denotes a time-aware embedding function that takes both visual features and relative temporal offsets as input.

Retrieval-Augmented Generator. Given the retrieved historical visual feature F_l, we combine it with the current feature F_i to model long-term contextual information. We also compute the time difference Δt and project it into a continuous temporal embedding $E_{\Delta t}$ to ensure the influence on the time offset. A linear projection layer maps these concatenated features to the input embedding space of the sequence-to-sequence model. The retrieval-augmented generator then produces a context-aware commentary $\hat{C}_{i,t_i}$ that integrates local event details from the current clip with long-term match dynamics.

6 Experiments

In this section, we present our experiments and results. We begin by evaluating visual features within the **Commentary Augmentation** to identify the most suitable representations for this task (Experiment 1). Based on the selected features, we then conduct extensive experiments and ablation studies for **MatchAware** (Experiment 2).

6.1 Visual Features in Commentary Augmentation Pipeline

This experiment compares different visual feature representations for commentary augmentation and select the best-performing one for constructing **MatchText**.

Textual F	Visual F	B-1	B-4	M	R-L	C
SN-C	–	48.66	45.45	57.37	**68.95**	1.16
	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: Comparison of different visual features on MatchText. Best results are shown in **bold**, and second-best results are underlined. SN-C: SN-Caption, B-1/B-4: BLEU-1/4, M: METEOR, R-L: ROUGE-L, C: CIDEr.

Details We extract features from 30s video clips using CLIP (2 FPS), Baidu (1 FPS), and ResNet (2 and 5 FPS) (Radford et al., 2021; Zhou et al., 2021; He et al., 2016). BART (Lewis et al., 2019) is adopted as the generation backbone and fine-tuned on the SN-Short training set, using SoccerNet-Caption as textual input and evaluating on the SN-Short test set. Further implementation details are provided in Appendix A.5.

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 for ROUGE-L (Lin, 2004). In particular, the significant improvement in CIDEr shows that the augmented commentaries contain more distinctive and semantically dense content extracted from visual details. Based on these results, we adopt Baidu features for commentary augmentation and construct the **MatchText** dataset.

6.2 Evaluation of MatchAware

In this experiment, we compare **MatchAware** with several baselines to validate its effectiveness. We adopt SN-Short and SN-Long as benchmarks and conduct two separate evaluations on each. In addition, we include off-the-shelf VLMs for zero-shot and few-shot comparison (see Appendix A.3 for details)

Implementation Details In addition to the visual features described in Section 6.1, we further include C3D (Tran et al., 2015) for feature extraction. Further details are provided in Appendix A.5.

Baseline We use Video-LLaMA3-7B (VLLaMA) in zero-shot and eight-shot settings as a naive baseline. We further evaluate MA-LMM (He et al., 2024) on SN-Short under zero-shot conditions for comparison. Additionally, we adopt

Method	Visual Features	BLEU-1	BLEU-4	METEOR	ROUGE-1	ROUGE-L	CIDEr
			SN-Short				
VLLaMA (0-shot)	ViT	15.56	0.46	8.27	20.16	14.70	1.69
VLLaMA (8-shot)	ViT	15.04	0.90	9.41	21.63	15.06	2.28
MA-LLM	ViT	3.66	0.13	3.37	14.16	11.24	0.23
MatchVoice (Trained on Soccernet-Caption)	C3D	19.56	3.16	7.53	21.11	17.35	5.36
	Baidu	19.82	3.99	8.15	23.13	19.08	8.27
	ResNet(2)	23.25	3.67	8.55	23.40	18.59	8.54
	ResNet(5)	20.29	3.15	7.81	21.65	18.11	6.42
	CLIP	22.48	3.65	8.39	22.03	17.78	7.45
MatchVoice (Trained on SN-Short)	C3D	23.80	2.45	8.50	22.62	18.27	8.30
	Baidu	27.26	4.31	9.84	25.94	20.28	11.79
	ResNet(2)	25.22	2.87	8.92	23.53	18.65	8.11
	ResNet(5)	24.31	3.47	8.60	23.39	18.76	9.92
	CLIP	23.88	3.45	8.69	22.85	19.25	7.85
MatchAware† (Trained on MatchText)	C3D	31.30	11.33	15.52	35.38	27.32	13.24
	Baidu	**35.58**	**17.43**	**18.06**	**40.22**	**33.68**	**15.73**
	ResNet(2)	31.83	9.62	13.16	30.74	26.30	14.32
	ResNet(5)	30.03	9.15	15.21	34.56	27.80	13.63
	CLIP	27.48	7.90	10.25	28.63	26.22	9.25
			SN-Long				
MatchVoice (Trained on Soccernet-Caption)	C3D	12.89	1.90	6.76	23.98	15.43	0.61
	Baidu	12.55	2.41	6.99	24.79	16.45	0.52
	ResNet(2)	17.96	2.36	7.18	25.02	15.51	0.85
	ResNet(5)	13.61	2.15	6.87	24.48	16.27	0.34
	CLIP	15.81	2.33	7.11	24.76	15.91	0.68
MatchVoice (Trained on SN-Short)	C3D	19.82	2.29	7.87	26.95	17.52	1.08
	Baidu	22.62	3.35	8.97	29.33	19.82	2.07
	ResNet(2)	21.15	2.26	8.28	27.50	17.97	1.94
	ResNet(5)	19.54	2.42	7.78	27.08	17.91	1.36
	CLIP	19.13	3.02	8.16	27.22	19.05	1.09
MatchAware (Trained on MatchText)	C3D	43.81	15.57	16.18	41.37	30.83	19.03
	Baidu	**47.05**	**20.16**	**18.41**	**45.13**	**35.50**	20.26
	ResNet(2)	42.03	12.88	15.06	39.67	28.28	20.02
	ResNet(5)	44.20	15.00	16.21	42.14	30.98	**22.42**
	CLIP	40.43	11.13	14.28	37.50	26.35	9.76

Table 3: Evaluation results of different visual features on SN-Short and SN-Long. Best results are shown in **bold**, and second-best results are underlined. MatchVoice is a general generation model that excludes the commentary augmentation pipeline and retrieval; therefore, we train it on both the original SoccerNet-Caption and the manually curated SN-Short as baselines for comparison. †: without retrieval, under the SN-Short test setting, which provides annotations for the current event only, and is used to evaluate the model's ability to generate current-event commentary in isolation.

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

Results We adopt 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.

We first evaluate *MatchAware*† (†: without retrieval, which can also be viewed as MatchVoice) against MatchVoice trained on SoccerNet-Caption, SN-Short, and MatchText. The results show that *MatchAware*† outperforms the baselines across all of the metrics. We observe a clear performance improvement when moving from SoccerNet-Caption to SN-Short and further to MatchText. SoccerNet-Caption provides limited supervision with coarse descriptions, while SN-Short introduces richer textual content but remains constrained by its scale, resulting in suboptimal performance. Models trained on MatchText achieve the best results, showing the effectiveness of our commentary augmentation pipeline in providing large-scale, informative narratives.

Next, we investigate the proposed retrieval-augmented generator and evaluate *MatchAware* on SN-Long. Results show that it achieves the best performance under all visual feature settings

and evaluation metrics, indicating that our retrieval mechanism effectively extracts relevant historical video features and enables the generation of more insightful, globally coherent commentary that enhances the depth of the descriptions. A case study is provided in Appendix A.6.

In addition, we directly input the video clips into two state-of-the-art Large Multimodal Models,Video-LLaMA3 (Boqiang Zhang, 2025) and MA-LMM (He et al., 2024), to evaluate their performance on SN-Short under different settings, as shown in Table 3, part 1. In both cases, both models significantly underperform compared to models specifically designed or fine-tuned for structured soccer commentary generation. We provide qualitative case studies across the zero-shot and few-shot settings in Appendix A.7.

Our extensive experiments demonstrate that (i) Our Commentary augmentation pipeline effectively augments incomplete commentary into detailed, informative narratives; the enhanced descriptions reduce the information gap between text and video, providing richer training data for downstream generation models. (ii) The proposed retrieval-augmented generator consistently improves performance across all evaluation metrics, demonstrating its effectiveness in generating context-aware commentary. (iii) SN-Short and SN-Long are more challenging due to their analytical, context-rich commentaries. Models trained on SoccerNet-Caption perform worse than those trained on the augmented dataset with retrieval.

6.3 Analysis

Retrieval Performance We evaluate retrieval performance using different visual feature representations, as shown in Table 4. ResNet(5) consistently performs best across most Recall@K metrics; this suggests that higher temporal resolution may be beneficial for matching historical events accurately.

Visual Feature	R@1	R@3	R@5	R@10
Baidu	34.19	61.11	68.80	85.47
C3D	27.35	51.71	68.80	85.04
ResNet(2)	36.32	58.55	70.51	85.90
ResNet(5)	**36.75**	**61.97**	73.50	**89.32**
CLIP	34.19	59.40	**74.36**	88.46

Table 4: Retrieval performance using different visual features. Recall@K: whether the ground-truth historical event appears in the top-K retrieved candidates.

Ablation Study on Retrieval-Augmented Generator We further analyze the contribution of retrieval by comparing *MatchAware* without retrieval.

Model	Visual F	B-1	B-4	M	R-1	R-L	C
MA†	C3D	22.95	8.33	10.87	34.50	24.31	5.45
	Baidu	26.70	12.14	13.16	39.33	29.49	9.67
	ResNet(2)	22.05	6.04	10.10	32.52	21.44	3.79
	ResNet(5)	23.55	7.62	11.06	34.85	24.00	5.27
MA	C3D	43.81	15.57	16.18	41.37	30.83	19.03
	Baidu	**47.05**	**20.16**	**18.41**	**45.13**	**35.50**	20.26
	ResNet(2)	42.03	12.88	15.06	39.67	28.28	20.02
	ResNet(5)	44.20	15.00	16.21	42.14	30.98	**22.42**

Table 5: Ablation study of the retrieval in MatchAware on the SN-Long dataset. MA: MatchAware

As shown in Table 5, incorporating retrieval consistently improves performance across all metrics and visual features, showing its effectiveness in leveraging relevant historical information for context-aware commentary generation.

6.4 Human Evaluation

We conduct a human evaluation on 97 video-text pairs (from 3 games), where three experienced soccer fans compare commentaries generated by **MatchVoice** (trained on SN-Caption) and **MatchAware**(trained on MatchText). Each output is rated on a five-point Likert scale along three dimensions. Table 6 shows the average scores. Appendix A.4 provides details of the human evaluation.

Model	Accuracy	Completeness	Depth
MatchVoice	3.27	2.87	2.77
MatchAware	**3.44**	**3.76**	**3.74**
p-value	0.39	0.012	0.008

Table 6: Human evaluation on 97 video-text pairs. Each is evaluated on three aspects: **Accuracy** (how well the description matches the video content), **Completeness** (whether the main event is fully described), and **Depth** (the degree of tactical depth and narrative coherence).

Both models achieve similar accuracy in capturing key events. However, MatchAware leverages semantically enriched commentaries and historical context, allowing it to better reflect the video content and provide more comprehensive soccer commentary with deeper tactical insights.

We also conduct statistical significance analysis on the human evaluation results using the Wilcoxon signed-rank test (Wilcoxon, 1945).

The results show that the difference in *Accuracy* between the two models is not statistically

significant ($p = 0.39$). In contrast, MatchAware achieves statistically significant improvements over the MatchVoice in both *Completeness* ($p = 0.012$) and *Depth* ($p = 0.008$). These results show the effectiveness of leveraging semantically enriched commentaries and historical context for soccer commentary generation.

7 Conclusion

In this paper, we introduced two curated datasets, **SN-Short** and **SN-Long**, which focused on scene-level descriptions and long-range event continuity for soccer commentary generation. To address the incompleteness of existing annotations, we developed a commentary augmentation pipeline to construct **MatchText**, a semantically complete and structurally standardized dataset. Based on this, we proposed **MatchAware**, a generation model that incorporates relevant historical events to produce context-aware and coherent soccer commentaries. Extensive experiments showed that our approach achieved promising results in soccer commentary generation.

Limitations

First, consistent with previous studies, our model lacks a player localization and tracking module, which limits its ability to correctly identify specific players or generate their correct names within the commentary.

Second, the retrieval mechanism is currently constrained to a single match half and strictly relies on pre-defined anchored events. This dependency prevents the system from performing cross-match retrieval or autonomously localizing relevant segments directly from the raw video stream, thereby restricting the richness and depth of the information available for tactical analysis.

Third, although experiments confirm that training on our SN-Short dataset improves performance, and while our two constructed datasets are the largest manually annotated ones to date, the model's overall capability remains limited by the relatively small scale of the data. This data insufficiency is a primary motivation for developing our Commentary Augmentation Pipeline, yet the need for larger-scale, high-quality manual datasets remains a bottleneck for the field.

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.

Zesen Cheng Zhiqiang Hu Yuqian Yuan Guanzheng Chen Sicong Leng Yuming Jiang Hang Zhang Xin Li Peng Jin Wenqi Zhang Fan Wang Lidong Bing Deli Zhao Boqiang Zhang, Kehan Li. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

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.

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*.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Sushant Gautam, Mehdi Houshmand Sarkhoosh, Jan Held, Cise Midoglu, Anthony Cioppa, Silvio Giancola, Vajira Thambawita, Michael A. Riegler, Pal Halvorsen, and Mubarak Shah. 2024. Soccernet-echoes: A soccer game audio commentary dataset. In *2024 International Symposium on Multimedia (ISM)*, pages 71–78.

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.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. *CVPR*.

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.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23369–23379.

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.

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*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and Sebastian Riedel. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*.

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.

Xiang Li, Yangfan He, Shuaishuai Zu, Zhengyang Li, Tianyu Shi, Yiting Xie, and Kevin Zhang. 2025. Multi-modal large language model with rag strategies in soccer commentary generation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6197–6206.

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

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, Zifeng Li, Haoning Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025a. Multi-agent system for comprehensive soccer understanding. In *ACM Multimedia 2025*.

Jiayuan Rao, Haoning Wu, Hao Jiang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025b. Towards universal soccer video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

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*.

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.

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.

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.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning (ICML)*.

Ling You, Wenxuan Huang, Xinni Xie, Xiangyi Wei, Bangyan Li, Shaohui Lin, Yang Li, and Changbo Wang. 2025. Timesoccer: An end-to-end multimodal large language model for soccer commentary generation. *arXiv preprint arXiv:2504.17365*.

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*.

A Appendix

A.1 Details on the Dataset Construction

A.1.1 Details on the Construction of the SN-Short Dataset

We select 47 matches from the SoccerNet-Caption dataset as the foundation for constructing our dataset. For each match, we use the timestamps provided in the dataset as anchor points and extract the corresponding 15-second audio transcripts before and after each anchor from Soccer-Echoes. Since Soccer-Echoes does not undergo thorough human curation, it contains a significant amount of irrelevant noise. To address this issue, we manually remove invalid information and correct noisy segments. Then, using LLaMA-3.1-405B (Dubey et al., 2024), we integrate relevant supplementary information based on the event descriptions at the anchor points. Finally, to ensure the overall quality of the dataset, we perform a round of manual review and correction to guarantee that the commentary texts are fluent and factually consistent.

We use the following prompt:

> You are a professional sports commentator. Your task is to expand the given event description into a more informative commentary by appending relevant contextual details from the surrounding transcript.
>
> Here is the event description (anchor): `[ANCHOR_DESCRIPTION]`
>
> Here is the transcript of the surrounding audio (15 seconds before and after the event): `[AUDIO_TRANSCRIPT]`
>
> Your output should preserve the wording and structure of the anchor as much as possible. Then, append natural and factually coherent details from the transcript to enrich the context. The result should read like a smooth and realistic commentary.
>
> Output:

A.1.2 Details on the Construction of the SN-Long Dataset

Building upon SN-Short, we construct SN-Long by expanding the contextual scope of each event. Specifically, for each event in SN-Short (grouped by half-time), we manually select earlier relevant events as historical context. We then use LLaMA-3.1-405B in a few-shot prompting setup to generate commentaries conditioned on the anchor event and historical context. All generated outputs undergo rigorous human verification and correction to ensure factual coherence and stylistic fluency.

Below is an example of the prompting template used:

> You are a professional sports commentator. Your task is to generate a coherent and informative commentary by incorporating the current description and relevant historical context. The commentary should reflect the overall rhythm and evolution of the match. Please refer to the following examples as guidance: `[FEW_SHOT_1]`, `[FEW_SHOT_2]` ...
>
> Here is the current event: `[CURRENT_DESCRIPTION]`
>
> Here is the historical context: `[HISTORY_DESCRIPTION]`
>
> Output:

A.2 Annotation Quality Evaluation for SN-Short and SN-Long

We randomly sampled 3% of the SN-Short (81 video-text pairs) and SN-Long (55 current commentary-relevant commentary pairs) datasets. Three annotators, all soccer enthusiasts with at least six years of experience watching matches, evaluate the quality across multiple dimensions.

A.2.1 SN-Short

For SN-Short, the evaluation focuses on three dimensions: *Accuracy*, *Fluency*, and *Consistency*, where each dimension is annotated with either Y (yes) or N (no).

Based on the timestamps from the SoccerNet-Caption dataset (Mkhallati et al., 2023), we observe that *Accuracy* is considerably lower compared to the other two dimensions. To further investigate, we examine the nine samples that received at least two "N" labels on the Accuracy dimension.

Error Type	#Samples
Timestamp misalignment	7
Premature truncation	1
Manual annotation error	1
Total	9

Table 7: Error analysis of samples with at least two "N" labels on Accuracy.

As shown in Table 7, we find that seven of these nine failures are due to timestamp misalignment in SoccerNet-Caption. Another error is caused by premature truncation because the timestamp falls near the end of a match, and another is a manual annotation error in SN-Short.

We further use the timestamps provided by MatchTime (Rao et al., 2024), which aim to address the temporal misalignment in SoccerNet-Caption. As a result, six of the seven misalignments are corrected, while one remains incorrectly aligned.

Table 8 shows the final results. For the SN-Short dataset, after correction, *Accuracy* reaches an average of 95.5%, with the few remaining errors mainly due to timestamp misalignment and incomplete video content in the source dataset. Both *Fluency* and *Consistency* remain high, showing that commentaries in SN-Short are mostly fluent and exhibit no style drift.

A.2.2 SN-Long

In SN-Long, each sample consists of a commentary on the current event, a commentary on a previous event, and a summarizing/analytical commentary linking the two to reflect tactical aspects. We evaluate each sample on *Accuracy*, *Fluency* and *Consistency*.

As shown in Table 9, the commentaries in SN-Long achieve an average *Accuracy* of 92.7% in capturing and analyzing relevant events in tactical aspects, while *Fluency* and *Consistency* remain at a high level, showing that the commentaries in SN-Long are generally natural, coherent, and stylistically consistent.

A.3 Details on the Usage of VLMs for Commentary Generation

As described in Section 6.2, we employ Video-LLaMA3 and MA-LLM to generate commentaries using different prompting strategies. The evaluation is conducted based on the SN-Short.

For the zero-shot setting of MA-LLM, we adopt the following prompt.

> You are a professional football commentator.
> Please describe the tactical event in the given video clip accurately and concisely in a broadcast style.

For the few-shot setting of Video-LLaMA3, we augment the prompt with exemplar event labels (e.g., corner kick, offside) and corresponding commentary descriptions to guide generation. The prompting format is detailed below.

> You are a professional football commentator.
> Here are some examples:
>
> > Label i:
> > `[EVENT_LABEL]`
> > Commentary i:
> > `[EVENT_DESCRIPTION]`
> > ...
>
> Now describe the following video:

A.4 Details of Human Evaluation

A.4.1 Human Evaluation Criteria

To better interpret the human evaluation results, we provide detailed definitions of each evaluation dimension. Annotators were asked to rate each generated commentary on a 1–5 Likert scale along three dimensions: **Accuracy**, **Completeness**, and

Dimension	Individual Y Proportion (A/B/C)	Perfect Agreement (Y/N)	Fleiss's kappa
Accuracy	95.1% / 95.1% / 96.3%	93.8% / 3.7%	0.81
Fluency	93.8% / 97.5% / 96.3%	91.4% / 0.0%	0.27
Consistency	98.8% / 97.5% / 100.0%	97.5% / 0.0%	0.33

Table 8: Quality evaluation of the SN-Short dataset. **Individual Y Proportion (A/B/C)** shows the proportion of "Y" judgments from each annotator separately. **Perfect Agreement (Y/N)** shows the percentage of items where all annotators labeled "Y" or all labeled "N" **Fleiss's kappa** (Fleiss, 1971) measures the inter-annotator chance-corrected agreement. The three evaluation dimensions are: **Accuracy**: whether the commentary reflects the main event shown in the video; **Fluency**: whether the sentence is natural and fluent; **Consistency**: whether the style remains coherent without drift.

Dimension	Individual Y Proportion (A/B/C)	Perfect Agreement (Y/N)	Fleiss's kappa
Accuracy	92.7% / 94.5% / 90.9%	87.3% / 0.0%	0.37
Fluency	96.4% / 98.2% / 98.2%	94.5% / 0.0%	0.23
Consistency	98.2% / 100.0% / 96.4%	94.5% / 0.0%	-0.02

Table 9: Quality evaluation results of the SN-Long dataset. The three evaluation dimensions are: **Accuracy**: whether the commentary correctly reflects the tactical and analytical aspects of the summarized events, linking multiple events coherently; **Fluency**: whether the sentence is natural and fluent; **Consistency**: whether the style remains coherent without drift. The column headers have the same meaning as in Table 8.

Depth. Below we provide the definitions and scoring criteria for each:

Accuracy Measures how well the generated commentary reflects the actual events in the video.

5: Completely accurate, with no factual errors; all actions, players, and results are consistent with the video.

4: Mostly accurate with minor inaccuracies, but overall understandable.

3: Contains 1–2 factual errors but the main event is still correctly conveyed.

2: Multiple factual mismatches that affect comprehension.

1: Severely incorrect or unrelated to the video content.

Completeness Assesses whether the key components of the event are sufficiently covered.

5: Comprehensive and covers all essential actions and involved players.

4: Covers most key details, though may miss minor elements (e.g., whether the shot was on target).

3: Mentions only the main action (e.g., shot) but lacks prior context or result.

2: Minimal description, missing several core elements.

1: Lacks informative content or unrelated to the actual event.

Depth Evaluates the level of tactical understanding or contextual coherence expressed in the commentary.

5: Shows clear connection to the broader match context with tactical/strategic analysis.

4: Includes moderate insights into causes or background of the event.

3: Describes surface-level facts without deeper explanation.

2: Mechanically written or logically incoherent.

1: Generic or irrelevant description.

A.5 Implementation details

Here we summarize the implementation details of both the commentary augmentation pipeline and the proposed MatchAware model, including backbone architectures, input modalities, and training configurations. Detailed settings are reported in Table 10.

A.6 More Qualitative Examples on commentary generation

To provide a clearer illustration of the richness of our constructed dataset and the generation capabilities of the MatchAware compared to the MatchVoice (Rao et al., 2024), Figure 5 presents additional examples from the same match.

Parameter	Pipeline	MatchAware		
		VLG	RET	RAG
Gen. Backbone	BART	LLaMA-3-8B	–	BART
Q-Former K	32	32	–	32
Textual Input	SN-Cap $\rightarrow$ SN-Short†	MatchText, SN-Caption, SN-Short	–	SN-Long†
Visual Input	CLIP, Baidu, ResNet‡	CLIP, Baidu, ResNet‡, C3D	Shared	Shared
Video Clip	30s	30s	30s	30s
Epochs	20	40	10	20
Learning Rate	5×10^{-6}	1×10^{-5}	1×10^{-5}	1×10^{-5}
Hardware		1$\times$ NVIDIA RTX A100		

Table 10: Implementation details for the **Commentary Augmentation Pipeline** and the **MatchAware** model. The MatchAware model consists of three components: Video-Language Generator (VLG), Visual Event Retriever (RET), and Retrieval-Augmented Generator (RAG). † denotes supervision targets used for training. ‡ indicates features extracted at both 2 fps and 5 fps. Note: "Shared" indicates the component uses the same visual features as VLG.

Visual Grounding As shown in the first event, the models are tasked with describing a long-range shooting attempt. The baseline MatchVoice generates a generic and hallucinated description (" goes just wide of the left post"), which contradicts the visual evidence. In contrast, **MatchAware** accurately perceives the ball's trajectory, stating that the effort "sails high over the bar," aligning perfectly with the Ground Truth (GT). Furthermore, our model incorporates expressive descriptors such as "lacked precision" and "missing the target by inches," making the commentary more vivid and human-like.

Contextual Coherence The second and third events demonstrate the effectiveness of the retrieval-augmented mechanism in capturing the match's atmosphere. In complex foul events, a simple description of physical contact is often insufficient. While MatchVoice produces repetitive templates regarding the referee's signal, MatchAware successfully contextualizes these fouls within the broader match flow. For instance, in the second event, it notes that 'Both teams have shown moments of aggression," and in the third event, it mentions "an increase in aggressive challenges." These observations echo the sentiments found in the GT ("intense battles for possession"), proving that our model effectively retrieves and utilizes historical context to construct a coherent narrative rather than describing isolated events.

Tactical Insight and Summarization In the fourth event, the example highlights the model's ability to summarize tactical outcomes. The baseline focuses on the immediate result (the linesman pointing at the corner flag). However, MatchAware provides a higher-level analysis, describing how the opposition's defense "thwarts the attack" and noting that the attacking team has been "struggling to create clear-cut chances." This level of insight mirrors the GT's description of "solid defensive plays," indicating that MatchAware goes beyond surface-level captioning to offer tactical commentary.

A.7 Case studies of General-Domain VLMs

We compare Video-LLaMA with MA-LMM on our SN-Short dataset. A case study is shown in Table 11.

Example Case	Model	Generated Commentary Content
Case 1: Foul	Ground Truth	*[PLAYER] ([TEAM]) commits a foul after unfairly battling his opponent for the ball. His physical approach to winning the ball was deemed excessive by the referee.*
	Video-LLaMA (8-shot)	The team in white is attacking the goal defended by the team in red. The player in white shoots the ball towards the goal, but it is blocked by the defender. The ball deflects off the defender's foot and goes out of play. The referee awards a corner kick to the team in white.
	Video-LLaMA (0-shot)	The video clip features a football match where a player kicks the ball towards the goal. The goalkeeper attempts to block the shot, but unfortunately misses. The ball enters the goal, and the crowd erupts in cheers. The player who scored celebrates by running towards the corner flag, while the opposing team's players look on in disappointment.
	MA-LMM (0-shot)	soccer.
Case 2: Passing	Ground Truth	*Some great controlled play and passing in the middle of the pitch from [TEAM]. They maintained possession effectively, looking for gaps in [TEAM]'s defensive setup.*
	Video-LLaMA (8-shot)	A player from the team in white crosses the ball into the box. The player from the team in red is running to intercept the ball.
	Video-LLaMA (0-shot)	The video features a football match between two teams, one in red and the other in white. The players are seen running up and down the field, passing the ball to each other. At one point, a player in red kicks the ball towards the goal, but it is blocked by the goalkeeper. The game continues with both teams trying to score, but neither team manages to do so.
	MA-LMM (0-shot)	The players are trying to score a goal by kicking the soccer ball into the opposing team's goal.
Case 3: Intercept	Ground Truth	*[PLAYER] ([TEAM]) slips past one challenge, but is unable to feed a low pass into the path of one of his teammates. His attempt to set up a scoring chance was intercepted by an alert [TEAM] defense.*
	Video-LLaMA (8-shot)	[PLAYER] crosses the ball into the box, but the ball is intercepted by [PLAYER]. The crowd cheers loudly.
	Video-LLaMA (0-shot)	The video clip features a football match between two teams, one in white and the other in red. The white team is on the offensive, attempting to score a goal. The goalkeeper, dressed in black, successfully blocks the shot from the red team's player. The white team then regains possession of the ball and continues their attack.
	MA-LMM (0-shot)	A soccer match is taking place between two teams.

Table 11: Case studies of Video-LLaMA3 (0/8-shot) and MA-LMM (0-shot) in comparison with Ground Truth.

MatchAware
[PLAYER] receives a pass from [PLAYER] and decides to strike from long range. His effort sails high over the bar. The shot had power but lacked precision, missing the target by inches.
MatchVoice
[PLAYER] latches on to a precise pass on the edge of the box and unleashes a shot which goes just wide of the left post.
GT
[PLAYER] collects a pass and strikes from about 23 metres out. It's a poor attempt which flies well over the bar. The chance was created by a well-placed pass from [PLAYER], who was looking to break the opponent's defensive line.

MatchAware
The referee blows his whistle for a foul [PLAYER]([TEAM]) tripped his opponent. The challenge was late, making it an easy call for the referee. Both teams have shown moments of aggression, with one committing a foul and the other attempting to create scoring opportunities from distance.
MatchVoice
[PLAYER] is penalised for a foul. It's a clear and obvious case of a foul. The referee points to the penalty spot.
GT
[PLAYER] ([TEAM]) commits a foul after unfairly battling his opponent for the ball. The foul occurred just after [PLAYER]'s contested shot, which came from a difficult position. They have been suppressed by their opponents during this period, allowing them to easily complete their attacks.

MatchAware
[REFEREE] blows his whistle after one of the players from [TEAM] commits an offensive foul. The challenge was late, leaving the referee with no choice but to penalize him. The match is seeing an increase in aggressive challenges, with both teams committing fouls in their attempts to gain possession.
MatchVoice
[PLAYER] ([TEAM]) brings his opponent down with a tackle and [REFEREE] immediately signals a free kick.
GT
[PLAYER] ([TEAM]) commits a foul during an attacking move. The opposition had just recovered the ball, prompting [PLAYER] to make a late challenge in an attempt to regain possession. The match has seen an increase in intense battles for possession, with both teams committing fouls in their attempts to gain control.

MatchAware
[PLAYER] ([TEAM]) attempts to find his teammate with a pass from outside the box, but the opposition's defence is quick to react and thwarts the attack. The ball goes out of play. [TEAM] are awarded a corner kick. The delivery lacked precision, allowing the defense to clear comfortably. They have been struggling to create clear-cut chances, with their attacks being consistently thwarted by the opposition's defense.
MatchVoice
[PLAYER] ([TEAM]) races towards goal but the defender gets back well to make a challenge. The ball is out of play and the linesman points at the corner flag.
GT
[PLAYER] ([TEAM]) whips the ball into the box and [PLAYER] takes a perfect touch to control it, but one of the defenders does well to intercept. The play developed from a quick transition, with [TEAM] exploiting space on the right before delivering into the area. Both teams have shown flashes of creativity in their attacking play, but ultimately struggled to capitalize on scoring opportunities due to solid defensive plays.

Figure 5: More examples from the same match. **MatchAware**: our proposed framework. **MatchVoice**: baseline. **GT**: ground truth.