Master's Thesis

Visual and Memory–Augmented Soccer Commentary Generation

SUN Haoran

Supervisor KERTKEIDKACHORN Natthawut

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

March, 2026

Abstract

Automatic soccer commentary generation aims to close the gap between raw broadcast video and the fluent, tactical, and context-aware narrations produced by human commentators. Recent advances in Large Language Models (LLMs) and Vision-Language Models (VLMs) have shown promise in general video captioning; however, the domain of sports commentary presents unique challenges that remain unaddressed. Existing resources and methods are constrained by two fundamental limitations: (1) available annotations are typically concise and event-focused, offering only a single short sentence per clip (averaging around 24 words in datasets like SoccerNet-Caption). These annotations lack the semantic richness necessary to describe the full visual content of a broadcast segment. Crucially, they are not transcriptions of authentic human commentary but are collected from text-based live reporting platforms designed to log key events rather than narrate visual details. Consequently, such captions often omit contextual cues, visual dynamics, and tactical interpretations central to professional broadcast commentary; and (2) most approaches ignore temporal continuity, treating clips as independent moments and failing to model how past events shape the narrative and tactical interpretation of current plays. To address these issues, we first propose a commentary augmentation pipeline that transforms incomplete event-level annotations into a semantically enriched and structurally standardized dataset tailored to video clips. Building on this supervision, we then develop a generation model that produces informative and context-aware commentary by explicitly leveraging visual features and historical event context.

First, we introduce two manually curated datasets: SN-Short and SN-Long. These datasets are designed to capture the richness of language and contextual continuity essential for expert-level sports narration. SN-Short is a scene-level corpus that enriches the single-sentence annotations commonly used in prior work by integrating additional visual details extracted from corresponding audio transcripts and video context; After filtering out irrelevant metadata, SN-Short contains 2.8k verified video–text pairs covering core soccer events such as crosses, shots, set pieces, goals, and fouls. Building on SN-Short, SN-Long connects events across time by linking each target event with semantically related prior events from the same match half. After careful manual selection and aggregation, SN-Long contains 1,765 target events paired with an average of 2.84 historical context segments (5,006 contextual segments in total), producing longer, context-aware commentaries with substantially greater narrative and tactical depth. Both datasets were

manually verified by experienced annotators to ensure accuracy, fluency, and consistency.

Second, we design a commentary augmentation pipeline to tackle the scarcity of high-quality training data. Specifically, this pipeline transforms incomplete, moment-level captions into richly descriptive and structurally standardized texts. This commentary augmentation pipeline leverage pre-processed visual features together with original brief captions as input, to generate an augmented commentary that appends missing visual semantics while preserving stylistic consistency. By applying this pipeline to an extensive corpus of video segments from SoccerNet-Caption, we produce MatchText, a semantically complete and standardized dataset containing 27,207 video–text pairs from 424 matches. This dataset is intended to provide a high-quality supervision signal for downstream commentary generation.

Third, we propose MatchAware: a memory-augmented, retrieval-enhanced commentary generation model that explicitly models both the current visual event and relevant historical events. MatchAware is structured in three components: (i) an event-level video–language generator that produces an initial description of the current clip by projecting Q-Former visual prefixes into a frozen LLM decoder; (ii) a visual event retriever that maintains a memory bank of past event embeddings and retrieves semantically relevant historical events using a time-aware embedding objective; and (iii) a retrieval-augmented generator that conditions on the current feature, the retrieved historical feature, and the temporal offset embedding to produce a context-aware, analytically richer commentary. The retrieval objective encourages semantically meaningful matches while the temporal embedding ensures the model accounts for recency and temporal distance when integrating historical events.

To identify optimal visual representations and validate the effectiveness of both the commentary augmentation pipeline and MatchAware, we conduct extensive experiments on both of them. We first evaluate different visual feature sources (CLIP, Baidu soccer embeddings, ResNet at multiple frame rates, and C3D) within the commentary augmentation pipeline, finding that multimodal fusion improves standard metrics (BLEU, METEOR, CIDEr) and that Baidu features offered favorable information for MatchText construction in our setup. We then compare MatchAware to strong baselines across the SN-Short and SN-Long benchmarks. The comparison includes Video-LLaMA3 in zero- and few-shot settings, along with the MatchVoice architecture trained with different datasets. Models trained on MatchText consistently outperform those trained on raw SoccerNet-Caption or SN-Short; importantly, adding the retrieval-augmented generator yields substantial gains on SN-Long, showing the value of historical context in generating coherent,

context-aware commentary. Across feature variants, MatchAware trained on MatchText achieves the best scores (e.g., large relative improvements in BLEU, METEOR, and CIDEr over non-retrieval baselines), and also in retrieval recall metrics, indicating effective memory matching.

We further validate MatchAware via human evaluation. On a sampled set of clips from held-out games, experienced soccer annotators rated generated outputs along Accuracy (factual match to visual evidence), Completeness (coverage of the main event and its consequences), and Depth (tactical analysis and narrative coherence). While both MatchAware and baseline models achieve comparable accuracy for core events, MatchAware shows statistically significant improvements in Completeness and Depth — consistent with our goals of producing semantically rich and contextually informed commentaries. Ablations confirm that the retrieval mechanism is the primary driver of these improvements on SN-Long, while the augmented MatchText supervision is critical to improving moment-level generation quality.

Finally, we discuss practical limitations and future directions. MatchAware does not include an explicit player localization/tracking module, which constrains its ability to consistently produce correct player identity mentions; the retrieval mechanism is currently limited within a match half and relies on pre-anchored events, restricting cross-match retrieval or fully autonomous segmentation from raw streams; and, despite being the largest manually verified soccer commentary resources to date, the model's overall capability remains limited by the relatively small scale of the data. Our datasets (SN-Short, SN-Long, and MatchText), commentary augmentation pipeline, and MatchAware model together aim to provide a practical foundation for generating professional, analytically rich soccer commentary that aligns with broadcast visual flow.

In conclusion, this thesis presents a comprehensive method designed to bridge the semantic and contextual gaps in automated soccer commentary. By establishing the SN-Short and SN-Long benchmarks, we provide the first manually verified standards for scene-level detail and event continuity. Furthermore, our Commentary Augmentation Pipeline and the resulting MatchText dataset successfully address the scarcity of high-quality training data, transforming fragmented captions into semantically complete narratives. Finally, the MatchAware model demonstrates that explicitly leveraging historical visual context via memory augmentation is key to producing coherent, context-aware commentary. Together, these contributions demonstrate improvements over existing baselines and validate the effectiveness of integrating historical visual context for more coherent sports narration.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In recent years, the development of Computer Vision (CV) and Natural Language Processing (NLP) has become a focal point of artificial intelligence research. Specifically, the rapid advancement of Large Language Models (LLMs) and Vision-Language Models (VLMs) has sparked growing interest in multimodal generation tasks. Within this domain, automatic soccer commentary generation has emerged as a distinct and challenging research direction. Unlike generic video captioning, which often focuses on describing static scenes or simple, isolated actions, professional sports commentary requires the ability to bridge the gap between raw, informative visual content and fluent, tactical narration.

The release of the original SoccerNet dataset [1] marks the first large-scale effort in soccer broadcast video analysis, providing 500 full-length matches annotated with sparse temporal labels for key events. This benchmark enables research on action spotting in long untrimmed videos. SoccerNet-v2 [2] further enriches this benchmark with significantly denser and more diverse temporal annotations, extending supervision beyond sparse event-level labels toward more comprehensive broadcast video understanding.

Building upon SoccerNet, several datasets have further explored language supervision for soccer video understanding. SoccerNet-Caption [3] and SoccerReplay-1988 [4] both provide timestamped textual commentaries aligned with video clips, where annotations are collected from live text broadcasting platforms. While SoccerNet-Caption constitutes an early attempt at dense video captioning for soccer, SoccerReplay-1988 significantly expands the scale of such supervision by offering a substantially larger corpus.

MatchTime [5] mitigates the timestamp misalignment in SoccerNet-Caption

by releasing an automatic alignment pipeline. It also proposes MatchVoice, a generation model to evaluate commentary quality under improved temporal aligned data.

In parallel, GOAL [6] and SoccerNet-Echoes [7] provide human-transcribed commentaries extracted from match audio, offering more natural and expressive prose that often spans longer temporal contexts within a game. However, such audio-based transcripts frequently suffer from background noise, speaker overlap, and colloquial or incomplete phrasing, which limits their suitability for fine-grained, clip-level captioning and structured supervision.

Alongside dataset construction, several generative approaches have been proposed to automate this task. Dominant methods such as MatchVoice [5], SoccerRAG [8], and UniSoccer [4] formulate commentary generation as a Single-anchor Dense Video Captioning (SDVC) task. These approaches typically operate on short, isolated video clips (typically 30˜60s), leveraging paired video–text annotations from the aforementioned datasets to learn the mapping from visual features to textual descriptions.

1.2 Challenge

Despite the growing number of soccer video–text datasets and the progress of commentary generation models, existing approaches remain limited in their ability to produce informative and context-aware soccer commentary. Here, we describe two fundamental challenges for current methods.

Challenge 1: Annotation Semantic Incompleteness

Most existing datasets for soccer commentary generation rely on annotations collected from live text broadcasting platforms, which are primarily designed for event logging rather than descriptive narration. Consequently, the textual annotations associated with each video clip are typically short and event-focused, often restricted to a single sentence with limited semantic coverage.

While such concise descriptions are sufficient for indicating *what* event has occurred (e.g., a goal, foul, or substitution), relying on them introduces two critical mismatches. First, they frequently suffer from temporal misalignment, as the timestamps of live text logs often lag behind or drift from the actual occurrence of the event in the video stream. Second, in the standard Single-anchor Dense Video Captioning (SDVC) task, input video clips typically span 30 to 60 seconds. A brief, single-sentence caption is fundamentally insufficient to cover the dense visual information unfolding over such a long duration. This leads to a severe semantic imbalance, where the text annotation captures

only a fraction of the content or even deviates significantly from the rich visual details present in the full clip.

As a result, the imbalance in granularity between rich visuals and sparse text prevents the model from learning distinct features for similar clips. This lack of discriminative supervision causes the generation quality to deteriorate, yielding generic outputs that fail to capture the depth, comprehensiveness, or tactical nuances of the specific visual scene.

Challenge 2: Lack of Temporal Continuity

Professional soccer commentary is inherently contextual: commentators regularly refer to previous plays, tactical patterns, or recent incidents to interpret the current event. However, existing datasets and generation frameworks largely treat video clips as independent samples, without explicitly modeling temporal relationships between events.

By ignoring event continuity, current methods produce fragmented, moment-level commentary that lacks connective structure and narrative coherence across a match. Generated descriptions are confined to the immediate clip, lacking both the necessary data support and the architectural mechanisms to reference past events or maintain awareness of the evolving match context.

This limitation is particularly problematic for complex sequences where the significance of an event is deeply rooted in the game's history. Without temporal context, models fail to capture shifts in momentum, recurring tactical patterns, or the broader narrative—elements that are essential for expert-level sports narration.

Summary. In summary, these challenges reveal a dual necessity: we require richer data supervision to resolve the granularity mismatch at the clip level, and memory-augmented modeling strategies to explicitly reason about temporal context across events.

1.3 Contributions

To address the aforementioned challenges and advance the field of automatic sports narration, this thesis proposes a comprehensive method spanning dataset curation, data augmentation, and model architecture. The main contributions of this work are summarized as follows:

- **Manually Curated Benchmarks:** To address the challenges of semantic incompleteness and temporal discontinuity, we introduce two high-quality, manually verified datasets.

SN-Short focuses on scene-level semantic richness. It bridges the gap between rich visual content and concise annotations by enriching original incomplete descriptions with detailed human-transcribed narratives from authentic commentators. This effectively resolves the issue of semantic sparsity in existing supervision, offering dense and visually grounded annotations that align text complexity with visual information.

SN-Long targets event continuity and contextual flow. Building upon SN-Short, it explicitly links target events with semantically related historical events. This breaks the constraint of isolated clip processing, providing the necessary historical context to enable models to learn causal relationships and tactical evolution throughout a match.

- **Commentary Augmentation Pipeline and MatchText:** Addressing the scarcity of large-scale, high-quality training data, we design a commentary augmentation pipeline. Taking the original concise textual content as the backbone, this pipeline utilizes visual features to extract relevant match information and translates it into natural language to enrich the incomplete backbone sentences. This process transforms brief, moment-level captions into semantically complete and structurally standardized narratives. By applying this pipeline to the SoccerNet-Caption corpus, we construct **MatchText**, a standardized dataset containing over 27k video-text pairs. This dataset provides a robust supervision signal that is both semantically complete and visually grounded, effectively mitigating the granularity mismatch problem.

- **Memory-Augmented Generation Model (MatchAware):** We propose MatchAware, a novel architecture designed to capture the temporal flow of a match. Unlike standard captioning models, MatchAware incorporates a visual event retriever that maintains a memory bank of historical events. By retrieving and integrating relevant past visual cues with the current event, the model generates context-aware commentary that reflects tactical depth and game dynamics.

- **Comprehensive Experiments:** We conduct comprehensive experiments on the SN-Short and SN-Long benchmarks. Quantitative results demonstrate that MatchAware significantly outperforms existing baselines across standard metrics.Furthermore, human evaluation confirms that our approach yields statistically significant improvements in commentary *Completeness* and *Depth*, validating its ability to produce semantically complete and context-ware commentaries for soccer match.

1.4 Thesis Outline

The remainder of this thesis is organized as follows:

Chapter 2 reviews related work in sports commentary generation, multimodal models, retrieval-augmented multimodal generation, and visual encoders.

Chapter 3 details the benchmark curation process. It describes the construction of the SN-Short and SN-Long datasets, analyzes the content and statistics of these curated datasets, and provides a comprehensive comparison with existing benchmarks along with an annotation quality evaluation.

Chapter 4 presents the proposed methodology. This chapter first introduces the Commentary Augmentation Pipeline used to construct the MatchText dataset by enriching incomplete annotations. Subsequently, it details the MatchAware model, elaborating on its core components: the video-language generator, the visual event retriever, and the retrieval-augmented generator.

Chapter 5 provides the experiments and results. It covers the visual feature selection for the commentary augmentation pipeline and presents a evaluation of the MatchAware model.

Chapter 6 presents an in-depth analysis and ablation study. This chapter investigates retrieval performance, conducts ablation studies on the retrieval-augmented generator, and details the human evaluation criteria and results. It concludes with qualitative examples of the generated commentary.

Chapter 7 concludes the thesis by summarizing the main contributions, discussing the limitations of the current research and potential directions for future work.

Chapter 2

Related Work

This chapter provides a comprehensive review of the technical foundations relevant to automatic soccer commentary generation. The review is organized into four main sections. **Section 2.1** traces the evolution of sports commentary generation systems, ranging from early template-based approaches to recent deep learning frameworks, and analyzes the datasets that underpin these advancements. **Section 2.2** explores the paradigm of Retrieval-Augmented Multimodal Generation, which is central to our proposed MatchAware model for modeling historical context. Finally, **Section 2.3** and **Section 2.4** discuss the foundational multimodal architectures and specific visual encoders that serve as the backbone for extracting semantic representations from video content.

2.1 Sports Commentary Generation

Sports commentary generation aims to automatically produce natural language descriptions of sports matches by interpreting structured or unstructured representations of game progress. Early research in this area focused primarily on converting symbolic match representations, such as event logs or manually annotated statistics, into textual commentary. With the increasing availability of broadcast videos and multimodal datasets, recent work has shifted toward learning-based approaches that jointly model visual, temporal, and linguistic information. This section reviews representative methods for sports commentary generation, beginning with template- or rule-based systems and then moving to neural multimodal approaches and the datasets that support them.

2.1.1 Template- or Rule-based Systems

Early approaches to sports commentary generation are predominantly template- or rule-based, largely developed within simulated environments or structured data domains. These systems typically operate on event logs or discrete state vectors, converting them into textual commentary through predefined linguistic templates.

A foundational example in the soccer domain is the MIKE system [9]. MIKE operates on structured state information from the RoboCup simulator, such as player and ball positions. For each candidate event, the system assigns an importance score based on several factors, including the event type, its distance to the goal, and the current match situation. Events with low importance scores are filtered out, allowing the system to generate commentary that focuses on key moments rather than routine actions.

Similar ideas can be found in other domains where expert knowledge is explicitly encoded. For example, in strategy games, the Automated Chess Tutor [10] applies an argument-based approach to explain the reasons behind players' moves. Although designed for a different task, such systems illustrate the strong interpretability and rule-driven nature of rule-based generation methods.

In the broader NLP context, foundational work [11] extend these concepts to American football, proposing a collective content selection model that learns to generate summaries from database records, thereby bridging the gap between rigid rules and statistical generation.

In more recent applications involving real-world data, live sensing data is also utilized for automated sports commentary [12]. A hybrid framework is employed in such approaches, where detection rules first identify key moments from raw sensor inputs, and a subsequent template-filling mechanism generates the corresponding utterances. This ensures factual consistency and allows for precise control over the output content.

To reduce the limitations of fixed templates, later work explores more flexible generation mechanisms. For example, a neural approach is proposed to generate live soccer-match commentary directly from play data [13]. Instead of relying on fixed templates, this method uses a probabilistic model to select content and sentence structures according to the current game context. Similarly, in basketball, the Rotowire dataset [14] shows that models can generate summaries from statistics without using fixed templates. By learning the relationships between game situations and commentary patterns, the system is able to produce more natural commentary than purely rule-based approaches.

Despite these advancements, these systems still share several limitations.

They rely on structured play data and require substantial manual effort to define domain-specific representations. Moreover, their generation remains constrained by predefined event descriptions, making it difficult to capture fine-grained visual details or the evolving narrative flow found in human commentary. Most importantly, because these approaches operate solely on symbolic or textual inputs, they cannot directly process raw visual information from broadcast videos.

2.1.2 Multimodal-based Systems

With the increasing availability of broadcast sports videos and the rapid progress in computer vision, recent research has shifted toward multimodal-based systems that generate commentary by modeling visual data or expert knowledge. Advances in video understanding, including action recognition [15] and camera calibration [16], have made it possible to extract meaningful visual representations from raw match footage. These approaches typically rely on video clips aligned with textual annotations and employ neural architectures to learn correspondences between visual events and natural language descriptions. Compared to play-data-based methods, multimodal systems are able to directly process visual content and capture information that is not explicitly available in structured event logs, such as player movements, spatial configurations, and visual context.

SoccerNet-Caption

The **SoccerNet-Caption** framework [3] extends the paradigm of Dense Video Captioning (DVC) [17, 18] to the sports domain, establishing the first large-scale benchmark for soccer. While standard DVC approaches focus on detecting and describing concurrent events in open-domain videos, SoccerNet-Caption introduces the *Single-anchor Dense Video Captioning (SDVC)* task, where each commentary sentence is generated for a short video segment centered around a single annotated event. This formulation casts soccer commentary generation as a set of temporally localized captioning problems, enabling systematic evaluation of video-grounded captioning models.

In the proposed baseline, visual features are extracted from video clips using pre-trained Convolutional Neural Networks (CNNs). The extracted features are then pooled over time and fed into a Long short-term memory (LSTM) model, to generate the corresponding caption. This architecture provides a straightforward pipeline for mapping event-centered video segments to textual descriptions and serves as a reference point for subsequent multimodal commentary systems.

MatchTime

A critical challenge in training commentary models is the temporal misalignment between visual events and the corresponding commentaries. MatchTime [5] addresses this issue by manually correcting timestamps for a subset of matches to create a high-quality evaluation benchmark named SN-Caption-test-align and then proposes a multimodal temporal alignment pipeline to automatically correct and filter existing annotations at scale. Based on this pipeline, the authors curate a refined training corpus, named MatchTime, which aims to provide better video–text synchronization for commentary generation.

The paper also develops a commentary generation model, MatchVoice. MatchVoice leverages frozen pre-trained visual encoders together with a Perceiver-like temporal aggregator and a lightweight projection layer to produce prefix tokens for a decoder-only language model. This design allows the model to combine spatiotemporal visual representations with an autoregressive text decoder for generating soccer commentary.

TimeSoccer

TimeSoccer [19] presents the first end-to-end multimodal large language model for Single-anchor Dense Video Captioning (SDVC) on full-match soccer videos. Instead of relying on ground-truth timestamps as external controls, the model jointly predicts event timestamps and generates captions in a single pass, which allows it to capture global context across 45-minute match segments. To handle long video inputs, the authors introduce MoFA-Select, a training-free, motion-aware frame compression module that adaptively selects representative frames through a coarse-to-fine strategy. Complementary training paradigms are also used to strengthen the model's ability to learn long-range temporal dependencies. Together, these components enable TimeSoccer to produce timestamped, context-aware commentary without requiring pre-aligned annotations.

SoccerRAG

SoccerRAG [8] presents a retrieval-augmented framework for querying multimodal soccer data through natural language. The system combines large language models with retrieval over structured records and multimodal representations derived from soccer matches, such as event annotations, textual descriptions, and visual features. Unlike approaches that use retrieval to introduce additional external knowledge, SoccerRAG employs retrieval mainly to guide and constrain the generation process.

Retrieved results act as references to previously generated outputs or verified records, defining factual boundaries and preferred output patterns for the current response. By conditioning generation on these retrieved representations, the model is encouraged to produce more accurate and consistent answers, particularly for complex queries that require precise reasoning over match events.

2.1.3 Resources for Soccer Commentary Generation

A range of datasets has been developed to support research on automatic soccer commentary. Historically, early resources were symbolic or event logs designed for template-based systems; more recently, datasets have shifted toward multimodal corpora that pair video with natural-language commentary or transcripts. Below we group representative resources into two categories: event-anchored datasets, which link short video segments to event-centered captions, and human-transcribed datasets, which provide longer or spoken commentaries aligned to broadcasts.

Event-Anchored Datasets

SoccerNet-Caption [3]. SoccerNet-Caption establishes a large benchmark for dense video captioning in the soccer domain. Built on the SoccerNet collection, it contains broadcast videos from 471 full matches, paired with approximately 37,000 short natural language commentaries aligned to annotated events. Each commentary describes a temporally localized video segment centered around a single match event, making the dataset suitable for event-anchored captioning and dense video captioning tasks. The dataset is released together with a standard baseline architecture for supervised training and evaluation.

MatchTime [5]. MatchTime is a curated version of SoccerNet-Caption that focuses on improving temporal alignment between video segments and commentary. The dataset covers 422 full matches and contains approximately 33,000 aligned clip–caption pairs, with a manually verified test set to ensure high alignment quality. By correcting timestamp noise present in the original annotations, MatchTime provides a cleaner benchmark for training and evaluating video-grounded commentary generation models that require precise temporal synchronization.

SoccerReplay-1988 [4]. SoccerReplay-1988 is a large-scale multimodal dataset consisting of 1,988 full soccer matches with automatically generated

annotations. The dataset includes rich event labels, timestamped commentary alignments, and additional derived metadata covering diverse in-game situations. Owing to its scale and annotation coverage, SoccerReplay-1988 is well suited for pretraining and for a wide range of soccer understanding tasks, such as event classification, commentary generation, and retrieval.

Human-Transcribed Datasets

GOAL [6]. The GOAL dataset consists of human-transcribed live commentary extracted from broadcast audio. It is built from 20 matches and provides high-quality, time-aligned spoken commentary segments. Unlike event-anchored caption datasets, GOAL is not originally designed for training commentary generation models. Its primary goal is to analyze and evaluate the amount of factual and contextual knowledge expressed in natural commentary sentences. Moreover, the absence of explicit event or temporal anchors makes this format challenging for direct use in supervised video-to-text training pipelines.

SoccerNet-Echoes [7]. SoccerNet-Echoes augments the SoccerNet collection with speech transcripts automatically derived from broadcast audio using ASR, followed by normalization and translation. The dataset covers the full set of SoccerNet matches and provides time-aligned spoken commentary segments. Compared to manually transcribed resources, SoccerNet-Echoes offers much broader coverage but also introduces noise inherent to automatic transcription. Moreover, while the commentary is temporally aligned to the video, it is not explicitly anchored to discrete match events, which poses challenges for direct use in supervised video-to-text generation pipelines.

Summary. Event-anchored datasets (e.g., SoccerNet-Caption, MatchTime, and SoccerReplay-1988) are well suited for localized captioning and event-centric generation. However, the associated commentaries are typically short and factual, and often provide limited contextual information for the corresponding video clips. In contrast, human-transcribed datasets (e.g., GOAL and SoccerNet-Echoes) contain much richer linguistic content, including professional commentary, tactical analysis, and spontaneous reactions. At the same time, such data is often noisy, fragmented, and conversational in nature, and usually lacks explicit anchoring to discrete match events. In conclusion, these resources reflect the field's shift from symbolic event logs toward large-scale multimodal corpora.

2.2 Retrieval-Augmented Multimodal Generation

In this section, we review representative studies on retrieval-augmented multimodal generation. These works provide useful insights into how knowledge and memory mechanisms can be integrated into multimodal generation models.

2.2.1 Foundations in NLP

Retrieval-Augmented Generation (RAG) was originally proposed to enhance Large Language Models (LLMs) by addressing their limitations in long-tail knowledge and hallucination. Pioneering works such as RAG [20] and REALM [21] introduced mechanisms to retrieve relevant documents from external non-parametric corpora and condition the generation on these retrieved contexts. This paradigm has since been adapted to computer vision, where retrieved visual or multimodal examples serve as prompts to guide generation.

2.2.2 REVEAL

REVEAL [22] proposes an end-to-end retrieval-augmented visual–language pretraining framework that stores multi-source multimodal knowledge in a large memory and learns to retrieve from it during generation. The method jointly learns four components—memory, encoder, retriever and generator, so that retrieval is integrated into the visual-language modeling pipeline. By encoding diverse knowledge sources (image–text pairs, QA examples, and unstructured knowledge) into a unified memory and training retrieval together with generation, REVEAL improves performance on knowledge-intensive vision–language tasks.

2.2.3 SoccerAgent

SoccerAgent [23] is a multi-agent framework for comprehensive soccer understanding, built on a large multimodal knowledge base (SoccerWiki) that encodes structured domain information about players, teams, referees, and venues. The system decomposes complex queries into subtasks handled by specialized agents, enabling collaborative reasoning over visual and textual inputs.

The paper also introduces SoccerBench, an evaluation suite with approximately 10,000 multimodal question–answer pairs spanning text-, image-, and

video-based understanding tasks. Unlike our commentary generation task, SoccerAgent primarily targets reasoning-oriented tasks.

2.3 Multimodal Models

Multimodal models learn joint representations across different modalities, most commonly vision and language, through large-scale pretraining. The field has evolved from contrastive representation learning, which focuses on aligning visual and textual feature spaces, to the recent surge of Multimodal Large Language Models (MLLMs) capable of open-ended generation and reasoning.

Foundational architectures like CLIP [24] pioneered scalable alignment, while recent frameworks such as Flamingo [25], BLIP [26], BLIP2 [27] and LLaVA [28] have demonstrated how to effectively bridge powerful visual encoders with Large Language Models to handle complex multimodal tasks. This section reviews representative architectures that underpin current advancements.

2.3.1 CLIP

CLIP (Contrastive Language–Image Pretraining) [24] learns a shared image–text embedding space by contrastive pretraining on a large corpus of image–caption pairs. The architecture is an image encoder maps an image to a fixed vector, and a text encoder maps a caption to a vector; both vectors are projected into a common latent space and trained with a symmetric contrastive loss so that matching image–text pairs are close while non-matching pairs are distant. A key characteristic of CLIP is that its large-scale contrastive pretraining produces generalizable visual representations. As a result, the pretrained image encoder can be reused as a fixed or lightly adapted visual encoder in downstream multimodal systems.

2.3.2 BLIP-2

BLIP-2 [27] is vision-language model that enables efficient multimodal learning by reusing frozen pretrained image encoders and large language models. Its core design introduces a lightweight Querying Transformer (Q-former), which serves as an interface between the two frozen components. The Q-former employs a small set of learnable query tokens to attend to visual features and extract compact, language-aligned representations, allowing cross-modal

interaction without updating the vision encoder or the language model. BLIP-2 introduces a parameter-efficient design paradigm that has influenced many later multimodal models to reuse frozen pretrained components and focus training on lightweight cross-modal adapters.

2.4 Visual Encoders

Visual encoders are responsible for transforming raw visual inputs into compact feature representations that can be consumed by multimodal generation models. In sports commentary generation, the quality of visual encoding directly affects the model's ability to recognize actions, capture spatial context, and align visual events with language. Recent work has therefore emphasized strong pretrained visual encoders, often reused as frozen backbones, to provide reliable visual representations for downstream multimodal tasks.

2.4.1 Baidu

In the SoccerNet ecosystem, Baidu features [29] have become a standard visual representation for downstream tasks such as action spotting and commentary generation. Originally developed for the SoccerNet Challenges in 2021 and 2022, these features are extracted using a robust pipeline designed to capture both static appearance and dynamic motion in broadcast soccer videos. Owing to their high discriminative power and low computational overhead, Baidu features are widely adopted as the primary visual input in state-of-the-art(SOTA) baselines for SoccerNet-Caption and action spotting tasks.

2.4.2 Other Visual Encoders

Beyond benchmark-specific features, various general-purpose architectures have been employed to encode sports video data. These can be broadly categorized into spatiotemporal networks and vision-language pre-trained models.

Convolutional Neural Networks. Early video understanding relied heavily on CNN architectures. ResNet [30] serves as the standard 2D backbone for extracting spatial features from individual frames. To capture temporal dynamics, C3D [31] utilizes 3D convolution kernels to process video volumes directly, explicitly modeling motion. Bridging these approaches, I3D (Inception-3D) [32] inflates pre-trained 2D kernels into 3D, allowing the

model to leverage robust spatial representations from ImageNet while learning temporal evolution.

Vision Transformers and CLIP. Recent advancements have shifted towards Transformer-based architectures. The Vision Transformer (ViT) [33] divides images into fixed-size patches and processes them via self-attention mechanisms, enabling the model to capture global context and long-range dependencies more effectively than CNNs. Furthermore, CLIP [24] typically employs a ViT-based encoder trained via large-scale contrastive learning. Unlike traditional models trained on closed-set labels, CLIP aligns visual representations directly with the text space.

Chapter 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.

Figure 3.1 provides a visual comparison demonstrating how our datasets bridge the semantic and temporal gaps in existing benchmarks.

As shown in the dashed orange boxes, existing captions often overlook key visual cues. For instance, at timestamp ∼45:03, the baseline caption only mentions the score change, ignoring the visual context of the fans' reaction. SN-Short captures these missing details , making the description more distinctive and visually grounded. Similarly, at ∼42:14, it supplements the specific action of the keeper clearing the danger, which is absent in the original text.

The green arrow illustrates the construction of SN-Long. By connecting the current goal event (∼45:03) with the historical context of a missed chance (∼42:14), the model generates a high-level summary. Instead of treating the goal in isolation, it synthesizes a narrative that reflects the game's flow and tactical evolution, offering a deeper level of match analysis that simple captioning cannot achieve.

Figure 3.1: 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.

3.1 SN-Short

Construction of SN-Short. To address the limitations of existing datasets in providing informative and detailed commentary within short 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 broadcast narratives without explicit event anchoring. These two datasets are complementary in nature: SoccerNet-Caption supplies reliable temporal anchors, whereas SoccerNet-Echoes provides rich contextual information.

Specifically, we select 47 matches from SoccerNet-Caption as the foundation of SN-Short. For each timestamped event, we retrieve the corresponding audio transcripts from SoccerNet-Echoes within a ± 15-second temporal window around the anchor. Since SoccerNet-Echoes does not undergo thorough human curation, the extracted transcripts are often fragmented, highly colloquial, and contain a substantial amount of irrelevant or noisy content. We therefore manually remove invalid segments and correct noisy transcriptions before further processing.

To enrich the original event descriptions with relevant contextual details,

we employ LLaMA-3.1-405B [34] as an auxiliary generation model. Given the cleaned transcripts and the anchor description, the model is prompted to expand the original caption by appending factually coherent and contextually relevant information, while preserving the wording and structure of the anchor as much as possible. This ensures the output remains grounded in the event while being enriched with professional narration.

Finally, all generated commentaries are manually reviewed and refined by annotators to ensure fluency, factual consistency, and stylistic coherence. In addition, we discard visually irrelevant events (e.g., attendance announcements or generic ball possession descriptions) using strict string-matching rules. Through this process, SN-Short provides semantically complete and more informative event-anchored commentaries than existing benchmarks.

Prompt Design. We use the following prompt to guide the auxiliary generation model:

System Instruction: 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.

Input 1 (Anchor Description):
`[ANCHOR_DESCRIPTION]`

Input 2 (Audio Transcript):
`[AUDIO_TRANSCRIPT]`

Task: 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:

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 designed to capture the narrative evolution of a match.

The construction process begins with the rigorous curation of historical context. For each target event in SN-Short (grouped by match-half), we manually select semantically related prior events to serve as the historical background. This manual curation is crucial to ensure that the retrieved history is logically connected to the current action, avoiding unrelated noise often found in purely temporal windows. To further elevate the generation quality to professional standards, we distill 17 summary paradigms from authentic human commentary transcripts. These paradigms, formulated into concise and standardized language, serve as high-quality few-shot exemplars that cover diverse soccer scenarios. Leveraging these resources, we employ LLaMA-3.1-405B [34] in a few-shot prompting setup. By conditioning the model on the extracted exemplars, the curated historical context, and the target event description, we generate coherent tactical commentaries that reflect the match's rhythm and evolution. All generated outputs subsequently undergo a rigorous human review to guarantee factual coherence, stylistic fluency, and the accuracy of the tactical analysis.

Prompt Design. We use the following prompt template to guide the model, utilizing the extracted paradigms as few-shot demonstrations:

System Instruction: 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.

Examples (Guidance):
`[FEW_SHOT_1]`
`[FEW_SHOT_2]`
...

Input 1 (Current Event):
`[CURRENT_DESCRIPTION]`

Input 2 (Historical Context):
`[HISTORY_DESCRIPTION]`

Output:

3.3 Content of Curated Datasets

This section provides detailed examples of the data structures for our constructed datasets: **SN-Short** and **SN-Long**. We present the JSON schema for typical entries in both datasets and define the specific fields.

3.3.1 SN-Short

The SN-Short dataset focuses on providing semantically complete descriptions for individual video clips. It augments the original concise captions with manually curated details. A representative data sample is shown in Figure 3.2.

Data Sample from SN-Short

```
{
  "annotations": [
    {
      "gameTime": "...",
      "game_time": "...",
      "query": "...",
      "short-term": "...",
      "query_ano": "...",
      "short-term_ano": "..."
    }
  ]
}
```

Figure 3.2: Structure of an annotation entry in SN-Short.

The fields are defined as follows:

- `gameTime`: The refined timestamp aligned using the MatchTime to ensure alignment with the video stream.
- `game_time`: The original timestamp provided by the SoccerNet-Caption dataset.
- `query`: The original, concise event description from SoccerNet-Caption.
- `short-term`: Our manually constructed, semantically complete description that captures detailed actions and outcomes.
- `_ano`: The anonymized versions of the descriptions (replacing specific names with `[PLAYER]`, `[TEAM]` and `[REFEREE]`).

3.3.2 SN-Long

The **SN-Long** dataset extends SN-Short by incorporating historical context. Each entry includes a `history` field containing retrieved relevant events to support context-aware generation. A representative sample is shown in Figure 3.3.

Data Sample from SN-Long


```
{
  "annotations": [
    {
      "gameTime": "...",
      "game_time": "...",
      "query": "...",
      "short-term": "...",
      "query_ano": "...",
      "short-term_ano": "...",
      "history": [
        {
          "history_time": "...",
          "short-term": "...",
          "long-term": "..."
        }
        ...
      ]
    }
  ]
}
```


Figure 3.3: Structure of an annotation entry in SN-Long.

In addition to the fields defined in SN-Short, SN-Long includes:

- `history`: A list of relevant historical events retrieved from the memory bank.
- `history_time`: The occurrence time of the retrieved historical event.
- `long-term`: A high-level summary capturing match trends and tactical analysis derived from the historical context.

3.4 Benchmark Comparison

In this section, we provide a comprehensive comparison between our proposed datasets and existing benchmarks. Table 3.1 presents a comprehensive comparison of our proposed benchmarks against existing soccer commentary datasets.

Dataset	**Com.(Games)**	**Man.**	**Evt.**	**Hist.**	**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 3.1: Comparison of representative datasets for soccer commentary generation. The table details the scale, annotation quality, and content richness of each dataset. Column abbreviations are defined as follows: **Com. (Games)** represents the total count of **Commentaries** and the number of unique source **Games**; **Man.** denotes **Manual** Verification, indicating whether annotations are manually checked; **Evt.** stands for **Event**-anchored Alignment, specifying if commentaries are strictly aligned with specific match events; **Hist.** refers to **Historical** context, indicating inclusion of contextual information; **Avg Len** reports the **Average Length** of the commentary in words. '–' for SoccerReplay-1988, which is not publicly available. '–' for GOAL and SN-Echoes, since these two datasets are not event-archored, the average length can not be calculated. Symbols ✓, ✗, and ✗ denote fully supported, partially supported, and unsupported features, respectively.

First, most existing large-scale datasets, rely primarily on the automated alignment of broadcast logs. While scalable, this approach is often limited by temporal misalignment. In contrast, SN-Short and SN-Long utilize manual verification to ensure factual consistency between the text and visual content. Additionally, unlike GOAL or SoccerNet-Echoes, which lack strict event boundaries, our benchmarks are event-anchored, facilitating the learning of precise associations between specific actions and their corresponding descriptions.

Another critical limitation of prior datasets is the semantic density and contextual depth of the commentary. SoccerNet-Caption has an average length of only 23.18 words, which tends to result in generic descriptions. Our datasets significantly elevate this semantic richness. Specifically, SN-Short achieves an average length of 35.10 words by explicitly capturing visual content present in the video clips but absent in SoccerNet-Caption, thereby yielding finer-grained descriptions and more discriminative commentary. Furthermore, SN-Long reaches the highest average length of 57.81 words and uniquely incorporates historical context. Unlike other datasets that treat events in isolation, SN-Long supports long-term narrative modeling, enabling the generation of commentary that references previous game events.

3.5 Annotation Quality Evaluation

To assess the reliability and linguistic quality of the constructed datasets, we conduct a manual annotation quality evaluation on both SN-Short and SN-Long. We randomly sample approximately 3% of each dataset, resulting in 81 video–text pairs from SN-Short and 55 current commentary–context pairs from SN-Long. All samples are independently annotated by three annotators, each of whom is a soccer enthusiast with over six years of regular match-watching experience.

3.5.1 SN-Short

For SN-Short, we evaluate each commentary along three dimensions: *Accuracy*, *Fluency*, and *Consistency*. Each dimension is annotated using a binary label, Y (yes) or N (no). Accuracy assesses whether the commentary correctly corresponds to the visual event at the given timestamp; Fluency measures grammatical correctness and naturalness; Consistency evaluates whether the style remains coherent and free of abrupt shifts.

Based on the original timestamps provided by SoccerNet-Caption [3], we observe that *Accuracy* is noticeably lower than the other two dimensions. To better understand the source of these errors, we further analyze the nine samples that receive at least two "N" labels on the Accuracy dimension.

Dimension	Individual Y Prop. (A/B/C)	Perfect Agreement (Y/N)	Fleiss's κ
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%	97.5% / 0.0%	0.33

Table 3.3: Quality evaluation results of the SN-Short dataset. **Individual Y Prop.** denotes the proportion of "Y" judgments from each annotator. **Perfect Agreement** reports the percentage of samples on which all annotators agree. **Fleiss's** κ [35] measures inter-annotator agreement beyond chance.

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

Table 3.2: Error analysis of samples receiving at least two "N" labels on the Accuracy dimension.

As shown in Table 3.2, the majority of accuracy failures (7 out of 9) stem from timestamp misalignment in the original SoccerNet-Caption annotations, rather than from errors introduced during dataset construction. One additional failure is caused by premature truncation when an event occurs near the end of a match, and one is attributed to a rare manual annotation mistake in SN-Short.

To mitigate the impact of timestamp misalignment, we further adopt the refined timestamps provided by MatchTime [5]. After this correction, six out of the seven misaligned samples are fixed, while one remains incorrectly aligned. As shown in Table 3.3, the final Accuracy of SN-Short reaches an average of 95.5%. The remaining errors are primarily caused by residual temporal misalignment and incomplete video content in the original broadcasts. Both *Fluency* and *Consistency* remain consistently high, indicating that SN-Short commentaries are generally well-formed and stylistically stable.

3.5.2 SN-Long

Each SN-Long sample consists of three components: a commentary on the current event, a commentary describing a previous relevant event, and an

analytical commentary that explicitly links the two to reflect tactical or strategic aspects. This structure introduces additional complexity, as the commentary must remain accurate not only at the event level but also at the narrative and analytical level.

Dimension	**Individual Y Prop.** (A/B/C)	**Perfect Agreement** (Y/N)	**Fleiss's κ**
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% / 96.4%	94.5% / 0.0%	−0.02

Table 3.4: Quality evaluation results of the SN-Long dataset. Column definitions (Individual Y Prop. , Perfect Agreement, and Fleiss's κ) follow the same metrics as in Table 3.3.

The quality of the SN-Long dataset is evaluated along three dimensions. *Accuracy* measures whether the commentary correctly captures tactical and analytical relationships across multiple events. *Fluency* evaluates whether the generated sentences are natural and fluent. *Consistency* assesses whether the commentary maintains a coherent style without drift.

As reported in Table 3.4, SN-Long achieves an average Accuracy of 92.7%, demonstrating that most commentaries successfully capture and relate multiple events in a tactically meaningful manner. Compared to SN-Short, the slightly lower Accuracy reflects the increased difficulty of generating multi-event, history-aware descriptions. Nevertheless, *Fluency* and *Consistency* remain high, suggesting that the inclusion of historical context does not significantly degrade linguistic quality or stylistic coherence.

3.6 Data Statistics

After manual verification, **SN-Short** contains 2,777 video–text pairs covering key soccer events, including crosses, shots, set pieces, goals, and fouls. Compared to existing benchmarks, the commentaries in SN-Short are more informative, as they integrate event descriptions with locally relevant broadcast context, while visually irrelevant events (e.g., attendance announcements or generic ball possession updates) are explicitly removed.

Based on SN-Short, we further construct **SN-Long** by filtering out context-independent events and retaining only those that benefit from historical or narrative grounding. This process results in 1,765 core video-text pairs, each

augmented with an average of 2.84 preceding historical events, yielding a total of 5,006 contextual segments.

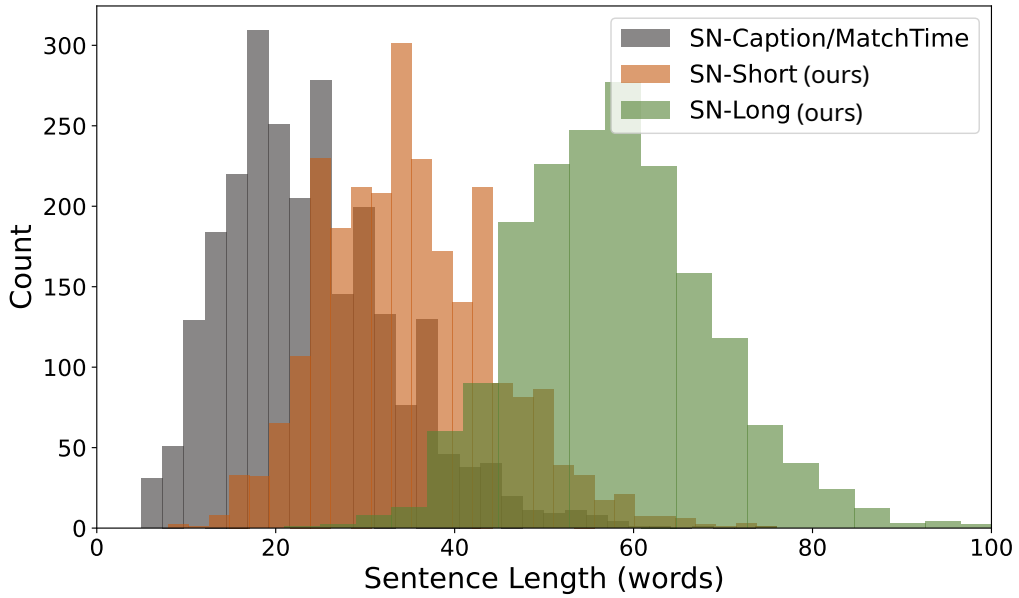


Figure 3.4: Distribution of sentence lengths across different datasets. Statistics are computed on the 47 matches shared by all datasets. *MatchTime* and *SN-Caption* share identical textual content and therefore exhibit the same distribution.

Figure 3.4 compares the distributions of commentary length across datasets. *SN-Caption* [3] and *MatchTime* [5] primarily consist of short, event-centric descriptions, typically ranging from 10 to 30 words, which often capture only the core action without broader context. In contrast, *SN-Short* produces noticeably longer commentaries by incorporating nearby broadcast information, resulting in more complete and descriptive sentences. *SN-Long* further extends this paradigm by explicitly incorporating historical context, with most commentaries ranging between 50 and 70 words.

Chapter 4

Proposed Method

In this chapter, we present our comprehensive approach for automatic soccer commentary generation.

We first introduce the **Commentary Augmentation Pipeline**. This pipeline is designed to bridge the granularity gap between visual content and textual descriptions. By applying this pipeline to large-scale data, we construct **MatchText**, a high-quality, structurally standardized dataset enriched with fine-grained visual semantics.

Subsequently, we propose **MatchAware**, a retrieval-augmented generation model. MatchAware leverages a memory bank to align current visual perceptions with relevant historical context, enabling the generation of insightful and context-aware commentary.

4.1 Commentary Augmentation Pipeline

4.1.1 Motivation

A primary bottleneck in sports commentary generation is the granularity mismatch between visual content and available textual commentary. Prior work indicates that many event-anchored commentaries in existing datasets provide only minimal descriptions and frequently omit visual information that is readily observable from video clips [5, 3]. Such omissions limit the semantic richness of the data and hinder the ability of downstream models to learn fine-grained visual-language grounding.

Furthermore, although we manually curated the SN-Short dataset to ensure high-quality visual alignment, its scale remains limited by the cost of manual annotation and is insufficient for training a robust domain-specific Vision-Language generation model.

To address these challenges, we design a commentary augmentation pipeline that targets visual information present in the video clips but absent from the original textual descriptions. The goal of this pipeline is to enrich the commentary with these missing semantics, thereby enabling the construction of datasets that are content-detailed, closely aligned with visual evidence, and structurally standardized.

4.1.2 Problem Formulation

The task of commentary augmentation is formulated as a multimodal conditional text generation problem. We are given a dataset of soccer match clips $\mathcal{V} = \{V_1, \ldots, V_n\}$ aligned with their corresponding, yet semantically sparse, textual commentaries $\mathcal{C} = \{C_1, \ldots, C_n\}$. Each C_i provides the basic event anchor but lacks the fine-grained visual details present in the video V_i.

Our goal is to learn a mapping function Φ that projects the sparse input pair (V_i, C_i) to an enriched commentary sequence $C_i' = \{y_1, \ldots, y_L\}$. The enriched output C_i' should preserve the factual correctness of C_i while incorporating explicit visual details derived directly from V_i.

Formally, we aim to model the conditional probability distribution:

$$P(C_i' \mid V_i, C_i; \theta),$$

where θ represents the learnable parameters of the model.

To make the visual content computationally accessible, we extract latent visual features F_i using a visual encoder and adapter:

$$F_i = \texttt{VisualEncoder}(V_i).$$

Consequently, the augmentation model Φ instantiates the generation process by jointly conditioning on these dense visual features and the original structural text:

$$C_i' = \Phi(F_i, C_i) \sim P(C_i' \mid F_i, C_i; \theta).$$

In the following section, we describe the concrete architecture used to parameterize Φ and the training strategy to optimize θ.

4.1.3 Architecture

As depicted in Figure 4.1, we develop our commentary augmentation pipeline based on an encoder–decoder architecture, fine-tuned on SN-Short training set. Taking textual descriptions as a structural backbone and integrating visual features as complementary semantic prompts, the pipeline enriches

commentary with additional semantics while preserving the standardized, structured format of the original data.

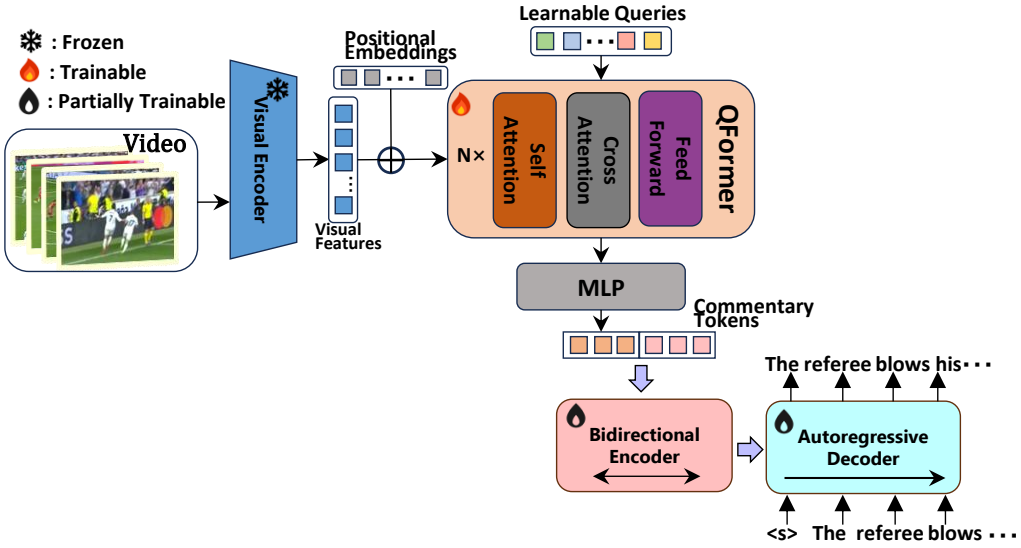


Figure 4.1: Overview of the commentary augmentation pipeline. Given pre-processed video features and concise textual descriptions, the pipeline generates detailed and structurally standardized commentary that captures fine-grained and discriminative visual information from each video clip.

Visual Feature Extraction and Adaptation. The goal of this module is to distill high-level semantic concepts from the raw video stream. Given a video clip V_i, we first extract dense frame-level features using a frozen, pre-trained visual encoder, denoted as $f_i \in \mathbb{R}^{T \times D_v}$, where T represents the temporal sequence length and D_v is the feature dimension.

Directly feeding these high-dimensional and temporally redundant features into a language model is computationally inefficient and may introduce noise. To address this, we employ a Q-Former [27] as a visual-language adapter. The Q-Former is initialized with K learnable query tokens $\{q_k\}_{k=1}^{K}$ (where $K \ll T$). These queries interact with the frozen video features f_i, selectively aggregating the most relevant visual information:

$$F_i = \texttt{QFormer}(f_i, \{q_k\}), \quad F_i \in \mathbb{R}^{K \times D_q}$$

Finally, to align the visual representation with the textual embedding space, the output query tokens F_i are projected through a Multi-Layer Perceptron

(MLP), ensuring that the visual signals are dimensionally compatible with the text encoder:

$$P_i = \texttt{MLP}(F_i), \quad P_i \in \mathbb{R}^{K \times D_{\text{text}}}$$

where D_{text} denotes the dimension of the text encoder.

Multimodal Fusion and Commentary Augmentation. For each event-aligned pair (V_i, C_i), the projected visual representations P_i are concatenated with the token embeddings of the original commentary $\texttt{Embed}(C_i)$ to to form a unified input sequence:

$$input = [P_i; \texttt{Embed}(C_i)].$$

The encoder processes this joint input using a multi-layer Transformer architecture. Within the attention layers, the textual tokens are allowed to attend to the visual tokens. The resulting context-aware hidden states are computed as:

$$h_i = \texttt{Encoder}(input),$$

where h_i captures both the original textual structure and the complementary visual semantics. Subsequently, the decoder operates auto-regressively conditioned on h_i. It generates the enriched commentary C_i' token by token:

$$C_i' = \texttt{Decoder}(h_i),$$

aiming to preserve the structure and wording of the original commentary while inserting missing, visually grounded details.

Training Objective. The model is fine-tuned on our manually curated dataset, SN-Short. We optimize the model parameters θ by minimizing the negative log-likelihood (NLL) loss:

$$\mathcal{L}_{NLL} = -\sum_{t=1}^{L} \log P(y_t | y_{<t}, C_i, V_i; \theta)$$

By training on SN-Short, the model is encouraged to capture the characteristics of professional and visually grounded commentary, which facilitates generalization to unseen and noisier data.

4.1.4 Usage: Construction of MatchText

Applying the proposed pipeline to large-scale soccer data, we construct a new dataset, **MatchText**, consisting of 27,207 video-text pairs from 424 games.

Compared to existing event-anchored datasets, MatchText provides more semantically complete and visually grounded commentaries, while maintaining a standardized format.

4.2 MatchAware: Context-Aware Commentary Generation

4.2.1 Motivation

Standard video captioning models typically process video clips in isolation, failing to capture the long-term match dynamics essential for professional analysis. To bridge this gap, we propose **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.

Specifically, MatchAware first offers an initial description of the current event and then retrieves relevant historical contexts from a memory bank of video features to enrich the output.

4.2.2 Problem Formulation

We aim to enhance commentary generation with memory-based context modeling. Let a soccer match video be segmented into a sequence of event-centric clips $\mathcal{V} = \{V_1, \ldots, V_N\}$ with corresponding timestamps $\mathcal{T} = \{t_1, \ldots, t_N\}$. For any given timestamp 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 define a history memory bank $\mathcal{M}_i$ at step $\rangle$ as:

$$\mathcal{M}_i = \{F_{1,t_1}, \ldots, F_{j,t_j} \mid 1 \leq j < i\}$$

A retrieval function $R(\cdot)$ selects the most relevant historical visual feature $F_l \in \mathcal{M}_\rangle$ 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 $\Delta t = |t_i - t_l|$ as joint inputs. Finally, the context-aware commentary is denoted as:

$$\hat{C}_{i,t_i} = \phi_{\text{ctx}}(F_i, F_l, \Delta t)$$

The system output is denoted as:

$$\mathcal{O}_{i,t_i} = [\tilde{C}_{i,t_i}; \hat{C}_{i,t_i}]$$

which provides both immediate description and historical context.

4.2.3 Architecture

As shown in Figure 4.2, 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 semantic of the current video clip; and (iii) a retrieval-augmented generator that incorporates long-term match context using the retrieved visual events.

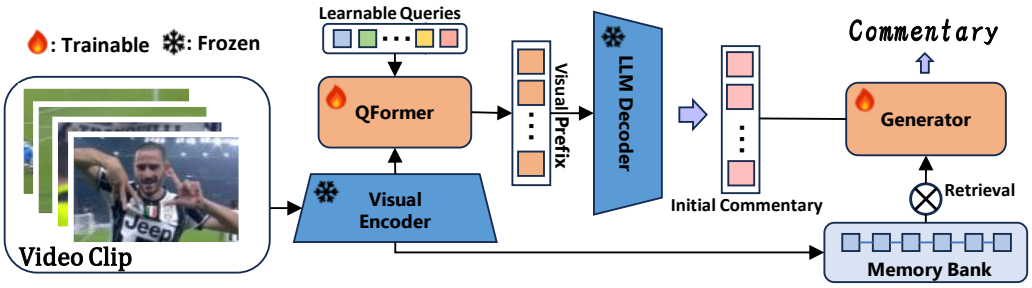


Figure 4.2: Overview of our proposed commentary generation model, MatchAware. It generates an initial commentary and a context-aware commentary through an LLM-based decoder and a generator, respectively, enabling more detailed and in-depth soccer commentary generation.

(i) Video–language Generator. The foundation of our model is a robust visual encoder bridged to a frozen Large Language Model (LLM). For an input clip V_i, we extract dense frame-level features $f_i \in \mathbb{R}^{T \times D_v}$ using a pre-trained visual backbone. To distill these high-dimensional features into a compact semantic representation, we employ a Q-Former. It utilizes K learnable query tokens $\{q_k\}_{k=1}^{K}$ to interact with f_i via cross-attention, yielding high-level visual tokens:

$$F_i = \texttt{QFormer}(f_i, \{q_k\}), \quad F_i \in \mathbb{R}^{K \times D_q}.$$

To align these visual tokens with the LLM's textual embedding space, we project them via a learnable MLP. This transformation maps the visual

dimension D_q to the LLM's hidden dimension D_{llm}, producing a sequence of soft visual prompts:

$$P_i = \texttt{MLP}(F_i), \quad P_i \in \mathbb{R}^{K \times D_{\text{llm}}}.$$

For commentary generation, we adopt an architecture that is similar to MatchVoice [5], treating P_i as a continuous prefix. These soft prompts are prepended to the input embeddings of the frozen LLM decoder. Conditioned on this visual prefix, the decoder generates the initial event-level commentary $\tilde{C}_i = \{y_1, \ldots, y_L\}$ auto-regressively:

$$p(\tilde{C}_i|V_i) = \prod_{t=1}^{L} p(y_t|y_{<t}, P_i).$$

(ii) Visual Event Retriever. The visual event retriever aims to identify historical events from the memory bank $\mathcal{M}_i$ that are semantically associated with the current video clip. Since relevance in sports depends not only on visual similarity but also on temporal proximity, we leverage a time-aware embedding function $g(\cdot)$ to project both visual content and temporal offsets into a unified metric space. Specifically, for a feature F and a temporal offset Δt, the embedding is computed as $g(F, \Delta t) = \texttt{MLP}_{\text{proj}}([F; \texttt{TimeEmb}(\Delta t)])$, where $\texttt{TimeEmb}(\cdot)$ is a sinusoidal positional encoding.

To optimize this space, we construct triplets (F_i, F_i^+, F_i^-) comprising an anchor, a semantically related positive event, and an unrelated negative event from the memory bank $\mathcal{M}_i$.

we optimize:

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

where $d(\cdot, \cdot)$ denotes the Euclidean distance and m is a margin hyperparameter.

Based on this learned metric, we retrieve the optimal historical context F_l for the current query F_i by selecting the candidate with the minimum distance:

$$F_l = \underset{F_j \in \mathcal{M}_i}{\text{argmin}}\ d\big(g(F_i, 0),\ g(F_j, t_i - t_j)\big)$$

This retrieved information is subsequently used by the Retrieval-Augmented Generator to produce context-aware commentary.

(iii) Retrieval-Augmented Generator. The generation module is instantiated as a Transformer-based encoder–decoder network, designed to synthesize analytical commentary by reasoning over the joint multimodal context. Given the retrieved historical visual feature F_l and the current event feature F_i, we first compute the time difference Δt and project it into a continuous temporal embedding $E_{\Delta t}$ to explicitly model temporal causality. To fuse these diverse signals, we concatenate them along the feature dimension and map them to the encoder's input space via a learnable linear projection. Formally, the fused context sequence H_{ctx} is computed as:

$$H_{\text{ctx}} = W_{\text{proj}}[F_i; F_l; E_{\Delta t}] + b_{\text{proj}}$$

where $[\cdot;\cdot]$ denotes the concatenation operation, and W_{proj} and b_{proj} are the weight matrix and bias of the projection layer, respectively. The resulting H_{ctx} serves as the input to the encoder. Subsequently, the decoder generates the commentary sequence auto-regressively.

Finally, the decoder produces the final context-aware commentary:

$$p(\hat{C}_{i,t_i}|H_{\text{ctx}}) = \prod_{t=1}^{L} p(y_t|y_{<t}, \texttt{Encoder}(H_{\text{ctx}}))$$

Chapter 5

Experiments and Results

In this chapter, we present a comprehensive evaluation of our proposed architecture, organized into two primary experimental phases. We aim to validate both the commentary augmentation pipeline and the effectiveness of the generation model.

Feature Selection for Commentary Augmentation Pipeline. We begin by evaluating different visual feature backbones within the Commentary Augmentation pipeline (Experiment 1).

The goal is to identify the most robust visual representations that can effectively capture fine-grained actions missing from the original text. We compare various pre-trained encoders to determine the optimal setup for constructing the high-quality MatchText dataset.

Evaluation of MatchAware. Building upon the optimal features identified in Phase 1, we conduct extensive experiments on the MatchAware model (Experiment 2).

5.1 Visual Feature Selection for Commentary Augmentation Pipeline

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

Parameter	Pipeline	MatchAware		
		VLG	**RET**	**RAG**
Gen. Backbone	BART	LLaMA-3-8B	–	BART
Q-Former K	32	32	–	32
Textual Input	SN-Cap → 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× NVIDIA RTX A100			

Table 5.1: 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.

5.1.1 Implementation Details

We extract features from 30s video clips using CLIP (2 FPS), Baidu (1 FPS), and ResNet (2 and 5 FPS) [24, 29, 30]. BART [36] 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 Table 5.1.

5.1.2 Results and Analysis

The quantitative results of the feature selection experiment are summarized in Table 5.2. We compare the performance of the baseline (text-only SN-Caption) against variants augmented with different visual features.

Textual F	Visual F	Bleu-1	Bleu-4	Meteor	Rouge-L	Cider
SN-C	–	48.66	45.45	57.37	**68.95**	1.16
	Baidu	<u>59.94</u>	**51.80**	**62.32**	<u>64.79</u>	**3.75**
	ResNet(2)	59.71	50.76	<u>62.27</u>	63.51	3.71
	ResNet(5)	59.68	51.44	62.17	64.02	<u>3.74</u>
	CLIP	**60.01**	<u>51.70</u>	62.10	64.48	<u>3.74</u>

Table 5.2: Comparison of different visual features on MatchText. Best results are shown in **bold**, and second-best results are <u>underlined</u>. SN-C: SN-Caption

Integrating visual features yields consistent improvements across most semantic metrics. Notably, the CIDEr score, which measures the consensus with human references and captures distinctiveness, triples from 1.16 to 3.75 (Baidu). This significant improvement shows that the visual features successfully ground the generation, allowing the model to hallucinate less and describe specific actions that are visually present but absent in the source text.

We also observe a slight drop in ROUGE-L. We attribute this to the fact that the text-only baseline tends to conservatively copy the input caption structure, resulting in high overlap with the reference's longest common subsequence. In contrast, our multimodal approach rewrites and expands the sentence with new visual details. While this lowers structural overlap, it significantly enhances semantic information density.

Consequently, we select **Baidu** features as the visual foundation for constructing the **MatchText** dataset.

5.2 Evaluation of MatchAware

In this experiment, we compare **MatchAware** with several baselines to validate its effectiveness. We aim to validate two key hypotheses: (1) that our constructed MatchText dataset provides superior supervision for commentary generation compared to existing captions; and (2) that our retrieval-augmented architecture significantly enhances the generation of long-term, context-aware narratives. 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.

5.2.1 Baseline

We compare our approach against two categories of baselines:

SOTA Vision Language Model. We employ Video-LLaMA3-7B [37] (VLLaMA) to assess the zero-shot and few-shot capabilities of off-the-shelf foundation models. Specifically, we evaluate the model in a zero-shot setting using direct inference, and an 8-shot setting where video-text exemplars are provided to guide the generation style. Further prompting details are available in Section 5.2.3.

Domain-Specific Baselines. We adopt MatchVoice [5] as the representative baseline architecture for soccer commentary generation. To rigorously analyze the impact of data quality, we train this architecture on three datasets: (1) the original SoccerNet-Caption; (2) our manually curated SN-Short; and (3) our proposed large-scale MatchText. The model trained on MatchText is denoted as **MatchAware**† (†: without Retrieval), which serves as an ablation baseline to explicitly isolate the contribution of the data augmentation pipeline from the retrieval mechanism.

5.2.2 Implementation Details

We extract features from 30s video clips using CLIP (2 FPS), Baidu (1 FPS), ResNet (2 and 5 FPS) and C3D [24, 29, 30, 31]. Further details are provided in Table 5.1.

5.2.3 Details on the Usage of Video-LLaMA3 for Commentary Generation

We employ Video-LLaMA3 to generate commentaries using both zero-shot and few-shot prompting strategies. The evaluation is conducted based on the SN-Short dataset.

For the zero-shot setting, we directly input a 30-second video clip centered around each anchor point into the model.

For the few-shot setting, we augment the prompt with exemplar event labels (e.g., corner kick, offside) and corresponding commentary descriptions to guide generation. The specific prompting format is structured as follows:

System Instruction: You are a professional football commentator.
Here are some examples to guide your generation:

Examples (Few-shot):
Label i: `[EVENT_LABEL]`

Commentary i: `[EVENT_DESCRIPTION]`

...

Task:
Now describe the following video:

5.2.4 Results and Analysis

The quantitative results on SN-Short and SN-Long are summarized in Table 5.3. We employ standard metrics including BLEU, METEOR, ROUGE-L, and CIDEr to evaluate generation quality.

Impact of Data Quality. We first analyze the effectiveness of the proposed data construction pipeline by comparing models without the retrieval module. As shown in Table 5.3 (SN-Short section), **MatchAware**† significantly outperforms the MatchVoice baselines trained on SoccerNet-Caption and SN-Short across all metrics. Specifically, compared to the event-centric SoccerNet-Caption, SN-Short provides more complete semantic information and shows moderate performance improvements.

We observe consistent improvements for MatchAware† (trained on MatchText) across all evaluation metrics. In particular, the CIDEr score increases noticeably compared to the SN-Short baseline. These results indicate that the scale and density of the augmented MatchText are essential for generating more distinctive and informative commentaries. Overall, these results show that our commentary augmentation pipeline effectively connects visual cues with textual descriptions by providing large-scale and semantically complete commentaries.

Comparison with Vision-Language Models. Next, regarding Video-LLaMA3, although the few-shot (8-shot) setting improves performance over the zero-shot setting, both lag significantly behind domain-specific models.

We attribute this performance gap primarily to the scarcity of domain-specific training data and the inherent trade-off between generalization and specialization. General VLMs are pre-trained on diverse, open-domain corpora to prioritize broad generalization capabilities; however, they lack sufficient exposure to the specialized linguistic patterns, terminology, and narrative structures essential for professional soccer commentary.

Impact of Retrieval-Augmented Generation. Finally, we evaluate the contribution of the retrieval mechanism, particularly on the challenging SN-Long benchmark which requires modeling long-range context. Results show that MatchAware 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.

Summary. 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.

Method	Visual F	B-1	B-4	M	R-1	R-L	C
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
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	<u>11.33</u>	<u>15.52</u>	<u>35.38</u>	27.32	13.24
	Baidu	**35.58**	**17.43**	**18.06**	**40.22**	**33.68**	**15.73**
	ResNet(2)	<u>31.83</u>	9.62	13.16	30.74	26.30	<u>14.32</u>
	ResNet(5)	30.03	9.15	15.21	34.56	<u>27.80</u>	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	<u>15.57</u>	16.18	41.37	30.83	19.03
	Baidu	**47.05**	**20.16**	**18.41**	**45.13**	**35.50**	<u>20.26</u>
	ResNet(2)	42.03	12.88	15.06	39.67	28.28	20.02
	ResNet(5)	<u>44.20</u>	15.00	<u>16.21</u>	<u>42.14</u>	<u>30.98</u>	**22.42**
	CLIP	40.43	11.13	14.28	37.50	26.35	9.76

Table 5.3: Evaluation results of different visual features on SN-Short and SN-Long. Best results are shown in **bold**, and second-best results are <u>underlined</u>. 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. Abbreviations: Visual F: Visual Features; B-n: BLEU-n; M: METEOR; R-n: ROUGE-n; C: CIDEr.

Chapter 6

Ablation Study and Analysis

6.1 Retrieval Performance

To determine the optimal visual representation for retrieving historical events, we evaluate the performance of the **Visual Event Retriever** using different visual backbones.

We evaluate retrieval performance using different visual feature representations. The results, measured by Recall@K (R@K), are summarized in Table 6.1.

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

Comparing the ResNet variants, we observe that ResNet(5) consistently outperforms ResNet(2) across all metrics, achieving the highest R@1 (36.75), R@3 (61.97) and R@10 (89.32). This suggests that higher temporal resolution captures more fine-grained visual cues, which are critical for distinguishing between visually similar but distinct match events.

Meanwhile, CLIP also demonstrates highly competitive performance, particularly achieving the best R@5 (74.36) and second-best R@10 (88.46).

This indicates that CLIP's strong semantic understanding ensures robust retrieval recall, effectively keeping the correct historical event within the candidates.

6.2 Ablation Study on Retrieval-Augmented Generator

To verify the necessity of integrating historical context, we analyze the contribution of the retrieval module by comparing the full *MatchAware* model against its retrieval-free variant ($MA^{\dagger}$). The results on the SN-Long dataset are presented in Table 6.2.

Model	**Visual F**	**B-1**	**B-4**	**M**	**R-1**	**R-L**	**C**
$MA^{\dagger}$	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	<u>15.57</u>	16.18	41.37	30.83	19.03
	Baidu	**47.05**	**20.16**	**18.41**	**45.13**	**35.50**	<u>20.26</u>
	ResNet(2)	42.03	12.88	15.06	39.67	28.28	20.02
	ResNet(5)	<u>44.20</u>	15.00	<u>16.21</u>	<u>42.14</u>	<u>30.98</u>	**22.42**

Table 6.2: Ablation study of the retrieval module in MatchAware on the SN-Long dataset. MA: MatchAware. $MA^{\dagger}$: MatchAware without retrieval. The full model significantly outperforms the baseline, highlighting the importance of historical context.

As shown in the table, incorporating the retrieval module leads to substantial performance gains across all visual backbones and evaluation metrics. Most notably, the CIDEr score, which reflects the degree to which the generated commentary contains informative and distinctive content compared to reference texts, shows a significant improvement. This improvement suggests that visual information from the current clip alone is insufficient for generating high-quality match commentaries.

6.3 Human Evaluation

To validate the perceptual quality of the generated commentaries, we conducted a human evaluation on 97 randomly sampled video-text pairs from three distinct games. Three experienced soccer fans served as annotators to compare the commentaries generated by the baseline **MatchVoice** (trained on SoccerNet-Caption) and our proposed **MatchAware** (trained on MatchText). The evaluation was performed in a blind setting, where annotators were unaware of the model sources.

6.3.1 Human Evaluation Criteria

Annotators rated each generated commentary on a five-point Likert scale across three dimensions: **Accuracy**, **Completeness**, and **Depth**. The detailed scoring criteria are defined as follows:

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.

6.3.2 Results and Statistical Analysis

The average scores and statistical significance test results are summarized in Table 6.3. We utilized the Wilcoxon signed-rank test [38] to determine the statistical significance of the differences between the two models.

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

Table 6.3: Human evaluation results on 97 video-text pairs. Scores are reported on a 1–5 Likert scale.

Analysis. As shown in Table 6.3, MatchAware outperforms the baseline across all three dimensions.

Regarding **Accuracy**, the results indicate no statistically significant difference between the two models. This parity suggests that since both models utilize similar visual and generation backbones, they have similar ability to recognize basic visual events such as goals or fouls.

However, a distinct differences appear in the other dimensions. MatchAware shows statistically significant improvements in both **Completeness** (+0.89, $p = 0.012$) and **Depth** (+0.97, $p = 0.008$). The baseline MatchVoice is constrained by the concise nature of its training data and therefore often produces generic descriptions. On the other hand, MatchAware leverages the semantically enriched MatchText and historical context retrieval to generate comprehensive narratives containing player details and tactical insights, thereby aligning much closer to professional commentary.

6.4 Qualitative Examples on Commentary Generation

To provide a clearer illustration of the richness of our constructed dataset and the generation capabilities of **MatchAware** compared to the baseline **MatchVoice**, Figure 6.1 presents additional examples sampled from the same match. These qualitative results further validate the superiority of our proposed method in three key aspects: visual grounding, narrative coherence, and tactical depth.

As shown in the first row, the models are tasked with describing a long-range shooting attempt. The baseline MatchVoice generates a generic and somewhat 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," which aligns perfectly with the Ground Truth (GT). Furthermore, our model adds qualitative descriptors such as "lacked precision" and "missing the target by inches," making the commentary more vivid and human-like.

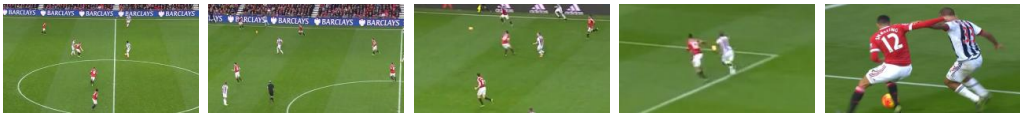
The second and third rows demonstrate the effectiveness of the retrieval-augmented mechanism in capturing the match's atmosphere. In complex foul events, a simple description of the physical contact is insufficient. While MatchVoice produces repetitive templates regarding the referee's signal, MatchAware successfully contextualizes these fouls within the broader match dynamics. For instance, in the second row, it notes that "Both teams have shown moments of aggression," and in the third row, 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.

In the fourth row, 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 professional-grade tactical commentary.

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 6.1: More examples from the same match. **MatchAware**: our proposed model. **MatchVoice**: baseline. **GT**: ground truth.

Chapter 7

Conclusion

In this thesis, we presented a comprehensive approach for automatic soccer commentary generation, addressing the limitations of existing methods in visual grounding and long-term context modeling.

First, we constructed two hierarchically complementary benchmarks: **SN-Short**, focusing on fine-grained, visually grounded event descriptions, and **SN-Long**, targeting narrative-driven tactical summarization. By employing an LLM-in-the-loop curation strategy, we ensured high semantic density and factual consistency in these datasets.

Second, to overcome the noise and sparsity in existing broadcast logs, we developed a **Commentary Augmentation Pipeline** based on a Visual-BART architecture. This pipeline effectively bridged the modality gap, resulting in the creation of **MatchText**, a large-scale, semantically complete, and structurally standardized dataset containing over 27,000 video-text pairs.

Third, based on these data foundations, we proposed **MatchAware**, a novel retrieval-augmented generation model. By introducing a dynamic memory bank, MatchAware explicitly aligns current visual perceptions with relevant historical context.

Finally, extensive quantitative experiments demonstrated that our approach significantly outperforms state-of-the-art baselines, particularly in metrics reflecting semantic richness. Qualitative evaluations further confirmed that MatchAware produces commentary that is not only visually accurate but also structurally coherent and tactically insightful, closely mimicking professional human commentators.

7.1 Limitations

Despite the promising results, our current framework has several limitations that differentiate it from human-level capability.

Absence of Explicit Entity Grounding. Our model focuses on event-level semantics and does not explicitly track individual players. As a result, it struggles to distinguish between players with similar appearances or accurately name players in crowded scenes. Incorporating a dedicated player identification module remains a challenging but necessary step for generating broadcast-level commentary.

Constraints on Retrieval Flexibility. The current retrieval mechanism is bounded to intra-match history, specifically within a single half, and depends on pre-segmented event anchors. This dependency precludes the system from performing cross-match retrieval (e.g., referencing statistics from previous games) or autonomously localizing relevant context directly from unsegmented raw video streams, which restricts the depth of tactical analysis.

Data Scalability. While SN-Short and SN-Long represent the largest manually curated datasets in this domain to date, their absolute scale remains modest compared to general-domain captioning benchmarks. Although our Commentary Augmentation Pipeline effectively mitigates this data scarcity by synthesizing the large-scale MatchText dataset, the field still faces a bottleneck due to the lack of massive-scale, high-quality human annotations required to fully unlock the potential of large foundation models.

7.2 Future Work

Based on the identified limitations, we propose several promising directions for future research.

First, to address the lack of fine-grained player identification, future work should integrate a Player Re-identification system into the visual encoder. By explicitly linking visual tracks to specific player metadata, such as names and positions, the model could generate commentary that accurately identifies individuals involved in complex interactions, thereby significantly enhancing the realism and utility of the broadcast.

Second, to generate more professional commentary, it is important to extend the retrieval bank beyond the current match. Future models could make use of an external knowledge base containing season-level statistics,

player biographies, and historical match records. This additional information would help generate commentary that is richer and more informative.

Third, future work should reduce the reliance on pre-defined event timestamps. Instead, an end-to-end system could be developed to both detect events and generate commentary directly from continuous raw video streams. This could be achieved by replacing the current pipeline with a dense video captioning architecture that can detect the start and end times of significant events, making the system applicable to real-time live streaming scenarios.

Finally, most existing methods mainly use visual and textual information, neglecting the rich information contained in the audio track. Auditory cues, such as the crowd reactions, referee whistles, and the emotion in commentators' voices, provide strong signals for event importance and emotional atmosphere. Incorporating these audio features into the model could help the model better capture the tempo of the match and generate more emotionally resonant commentary.

Bibliography

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

[2] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 4508–4519, 2021.

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

[4] Jiayuan Rao, Haoning Wu, Hao Jiang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards universal soccer video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8384–8394, 2025.

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

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

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

[8] Xiang Li, Yangfan He, Shuaishuai Zu, Zhengyang Li, Tianyu Shi, Yiting Xie, and Kevin Zhang. 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, 2025.

[9] Kumiko Tanaka-Ishii, Itsuki Noda, Ian Frank, Hideyuki Nakashima, Kôiti Hasida, and Hitoshi Matsubara. Mike: an automatic commentary system for soccer. *Proceedings International Conference on Multi Agent Systems (Cat. No.98EX160)*, pages 285–292, 1998.

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

[11] Regina Barzilay and Mirella Lapata. Collective content selection for concept-to-text generation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 331–338, 2005.

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

[13] Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. Generating live soccer-match commentary from play data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7096–7103, 2019.

[14] Sam Wiseman, Stuart Shieber, and Alexander Rush. Challenges in data-to-document generation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, 2017.

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

[16] Floriane Magera Silvio Giancola Olivier Barnich Bernard Ghanem Marc Van Droogenbroeck Anthony Cioppa, Adrien Deliège. 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*, 2021.

[17] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the International Conference on Computer Vision*, pages 706–715, 2017.

[18] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7190–7198, 2018.

[19] Ling You, Wenxuan Huang, Xinni Xie, Xiangyi Wei, Bangyan Li, Shaohui Lin, Yang Li, and Changbo Wang. Timesoccer: An end-to-end multimodal large language model for soccer commentary generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3418–3427, 2025.

[20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[21] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[22] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 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, 2023.

[23] Jiayuan Rao, Zifeng Li, Haoning Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Multi-agent system for comprehensive soccer understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3654–3663, 2025.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

[25] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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

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

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

[32] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[33] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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

[35] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[36] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880, 2020.

[37] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

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