Towards Automated Commentary Generation for Soccer Highlights

Chidaksh Ravuru
chidaksh@cs.unc.edu

1 Abstract

Automated soccer commentary generation has evolved from template-based systems to advanced neural architectures, aiming to produce real-time descriptions of sports events. While frameworks like SoccerNet-Caption laid foundational work, their inability to achieve fine-grained alignment between video content and commentary remains a significant challenge. Recent efforts such as MatchTime, with its MatchVoice model, address this issue through coarse and fine-grained alignment techniques, achieving improved temporal synchronization. In this paper, we extend MatchVoice to commentary generation for soccer highlights using the GOAL dataset, which emphasizes short clips over entire games. We conduct extensive experiments to reproduce the original MatchTime results and evaluate our setup, highlighting the impact of different training configurations and hardware limitations. Furthermore, we explore the effect of varying window sizes on zero-shot performance. While MatchVoice exhibits promising generalization capabilities, our findings suggest the need for integrating techniques from broader video-language domains to further enhance performance. Our code is available at https://github.com/chidaksh/SoccerCommentary.

2 Related Work

Automated soccer commentary generation has evolved from early template-based methods to sophisticated neural network approaches. Initial efforts focused on using predefined templates to convert structured match data into natural language summaries. Notable early works include (Barzilay and Lapata, 2005) which introduced a content selection framework to generate sports summaries from statistics, and (Bouayad-Agha et al., 2011) which improved upon this by emphasizing linguistic variation and content determination. However, such approaches lacked the ability to adapt dynamically to real-time events.

The development of neural network-based methods marked a significant shift. (van der Lee et al., 2017) introduced PASS, a data-to-text system for generating tailored soccer reports from match statistics. While this approach improved stylistic adaptability, it remained heavily dependent on predefined rules. (Wiseman et al., 2017) presented a neural model for generating basketball game summaries, highlighting the potential of deep learning for data-driven text generation. However, these efforts primarily focused on generating post-game summaries rather than live commentary.

More recent efforts have focused on large-scale datasets and multimodal learning. SoccerNet-Caption (Mkhallati et al., 2023) provided 37,000 timestamped commentaries from 471 complete games, enabling training of models capable of generating contextually relevant descriptions. The GOAL dataset (Qi et al., 2023) expanded on this by incorporating external knowledge such as player statistics and strategies, allowing for more accurate and informative descriptions. However, the lack of fine-grained event-commentary alignment remained a limitation.

To address these challenges, MatchTime and its MatchVoice model (Rao et al., 2024) introduced advanced alignment techniques, improving synchronization between commentary and video events. By leveraging multimodal representations and attention mechanisms, MatchTime achieved superior performance in generating real-time descriptions of ongoing events. Additionally, UniSoccer (Rao et al., 2025) presented the largest multimodal soccer dataset to date, demonstrating the effectiveness of foundation models for soccer-related tasks. Our work builds on these advancements by focusing on commentary generation for highlights using the GOAL dataset, enhancing quality through improved alignment techniques inspired

by MatchTime.

3 Methodology

The methodology for generating automatic soccer commentary consists of several steps including coarse and fine-grained temporal alignment, as well as commentary generation using the MatchVoice architecture.

3.1 Addressing Misalignment in SoccerNet-Caption

The SoccerNet-Caption dataset suffers from significant temporal misalignment between video clips and their corresponding textual commentary. This misalignment arises due to the delayed nature of commentators' descriptions during live matches, as well as inaccurate timestamp annotations. (Rao et al., 2024) highlighted this issue by analyzing the offset distribution between commentary and visual events, which reveals a substantial proportion of commentaries occurring with significant delays.

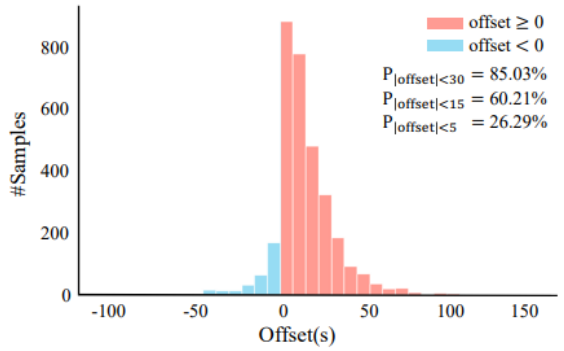

Figure 1: Misalignment identified in SoccerNet-Caption dataset where many samples are significantly delayed from the original event.

The misalignment issue must be addressed to enhance the performance of models trained on this dataset. To tackle this problem, the MatchTime framework implements a two-level alignment approach: coarse alignment and fine alignment.

3.2 Coarse Alignment using ASR and LLMs

The coarse alignment aims to approximately match textual commentary with video content. This is achieved by extracting narration from audio using WhisperX (Bain et al., 2023), an automatic speech recognition (ASR) system that produces timestamped text. Since soccer commentary is typically fragmented and colloquial, the transcriptions are then processed by LLaMA-3 (Grattafiori et al., 2024) to generate concise event descriptions for each 10-second video clip. The summarized text is used to infer a rough alignment with the original commentary.

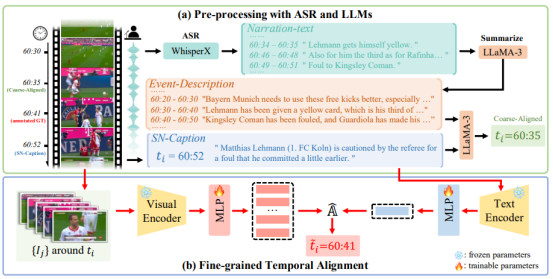

Figure 3: **Temporal Alignment Pipeline**. (a) Pre-processing with ASR and LLMs: We use WhisperX to extract narration texts and corresponding timestamps from the audio, and leverage LLaMA-3 to summarize these into a series of timestamped events, for data pre-processing. (b) Fine-grained Temporal Alignment: We additionally train a multi-modal temporal alignment model on manually aligned data, which further aligns textual commentaries to their best-matching video frames at a fine-grained level.

Figure 2: MatchTime: Coarse and Fine Alignment Pipeline.

3.3 Fine-grained Temporal Alignment

The fine-grained alignment process further refines the matching of commentary to visual events through contrastive learning. The alignment model uses a pre-trained CLIP model (Radford et al., 2021) to encode textual commentaries and video key frames, which are then projected using trainable MLPs $f(\cdot)$ and $g(\cdot)$:

$$C, V = f(\Phi_{\text{CLIP-T}}(C)), \quad g(\Phi_{\text{CLIP-V}}(V))$$

Where $C \in \mathbb{R}^{k \times d}$ and $V \in \mathbb{R}^{n \times d}$ represent the textual and visual embeddings, respectively.

The affinity matrix $\hat{A} \in \mathbb{R}^{k \times n}$ is computed as:

$$\hat{A}[i,j] = \frac{C_i \cdot V_j}{\|C_i\| \|V_j\|}$$

The alignment model is trained using a contrastive loss function:

$$\mathcal{L}_{\text{align}} = -\frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{\sum_j Y[i,j] \exp(\hat{A}[i,j])}{\sum_j \exp(\hat{A}[i,j])} \right)$$

Where $Y \in 0,1^{k \times n}$ is the ground-truth label matrix indicating whether commentary C_i corresponds to key frame V_j.

3.4 MatchVoice Architecture

The MatchVoice model is designed to generate natural language commentary from aligned video-text pairs. As illustrated in Figure 3, the model consists

of three main components: a frozen pre-trained visual encoder, a Perceiver-like temporal aggregator, and an LLM-based decoder.

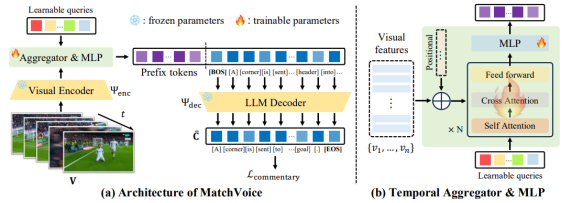

Figure 4: **MatchVoice Architecture Overview**. Our proposed MatchVoice model leverages a pretrained visual encoder to encode video frames into visual features. A learnable temporal aggregator aggregates the temporal information among these features. The temporally aggregated features are then projected into prefix tokens of LLM via a trainable MLP projection layer, to generate the corresponding textual commentary.

Figure 3: MatchVoice Architecture Overview.

Visual Encoding. The frozen visual encoder Ψ_{enc} processes video frames to obtain framewise features:

$$v_1, v_2, \ldots, v_n = \Psi_{\text{enc}}(V)$$

Temporal Aggregation. The Perceiver-like aggregator Ψ_{agg} applies transformer decoder layers with learnable queries over visual features to obtain temporally-aware features F:

$$F = \Psi_{\text{agg}}(v_1, v_2, \ldots, v_n)$$

Prefix Token Generation. The aggregated features are mapped to a set of prefix tokens using an MLP projection layer Ψ_{proj}:

$$\hat{C} = \Psi_{\text{dec}}(\Psi_{\text{proj}}(F))$$

Training Objective. The model is trained to minimize the negative log-likelihood loss:

$$\mathcal{L}\text{commentary} = -\sum_{i} \log P(\hat{C}_i | F)$$

This comprehensive approach ensures that the generated commentary is both temporally accurate and contextually relevant, enhancing the overall quality of the commentary generation process.

4 Dataset and Evaluation

To expand the applicability of MatchVoice for commentary generation in short videos such as highlights, we curated the GOAL dataset from YouTube. The dataset is divided into three subsets:

- **Training Data**: 358 videos
- **Validation Data**: 86 videos
- **Test Data**: 192 videos

Approximately 500 videos could not be downloaded due to regional restrictions.

4.1 Evaluation Metrics

We evaluated the performance of our model using several standard metrics commonly employed in video-language research, including BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr, and sBERT. These metrics are consistent with those used in the original MatchTime framework (Rao et al., 2024).

4.2 Hardware Configuration

All experiments were conducted using a single NVIDIA L-40 GPU on the Longleaf computing cluster, ensuring consistent and efficient training across various configurations.

5 Experiments

In this section, we present the experimental results for MatchVoice trained on SoccerNet-Caption and MatchTime datasets, along with our reproduced results and comparison. Additionally, we evaluate the impact of different window sizes on the performance of MatchVoice for the GOAL dataset.

5.1 Comparison of Original and Reproduced Results

We compare the results reported in the original paper with our reproduced results using the same dataset and evaluation metrics. The original results are summarized in Table 1, while our reproduced results are presented in Table 2. The absolute differences between these results are shown in Table 3.

5.2 Discussion

The difference in the reproduced results compared to the original paper can be attributed to several factors. Firstly, the original paper utilized A100 GPUs which have superior memory management and higher computation power compared to the L40 GPU used in our experiments. This hardware limitation likely affected the training stability and efficiency of the model, especially during fine-tuning. Additionally, subtle differences for example, ((Rao et al., 2024) used a batchsize of 64 with 16 workers, but due to resource constraints, we used a batch size of 32 with 8 workers) could have also contributed to the observed discrepancies.

Features	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDER	GPT-score
Baidu Aligned	**31.42**	**8.92**	**26.12**	**29.66**	**38.42**	**7.08**
Baidu Misaligned	30.32	8.45	25.25	29.40	33.84	7.07
CLIP Aligned	29.56	6.90	24.62	31.25	28.66	6.82
CLIP Misaligned	28.65	6.62	24.20	27.33	24.35	6.78

Table 1: Original Results Reported in the Paper - Aligned vs. Misaligned Models. Red denotes the best performance and the second-best performance in Blue.

Features	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDER	sBERT
Baidu Aligned	**29.643**	**8.227**	**26.491**	**26.316**	**34.452**	**68.981**
Baidu Misaligned	27.329	6.703	23.494	22.765	26.511	63.341
CLIP Aligned	27.433	6.219	25.255	24.456	24.965	66.464
CLIP Misaligned	24.625	4.250	22.701	20.623	15.053	60.882

Table 2: Reproduced Results. Red denotes the best performance and the second-best performance in Blue.

Features	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDER	sBERT
Baidu Aligned	0.677	0.223	-1.241	3.084	-0.612	-61.911
Baidu Misaligned	2.281	0.127	1.886	2.515	-5.901	-56.621
CLIP Aligned	1.217	0.561	-0.905	2.874	2.765	-59.684
CLIP Misaligned	-0.005	0.000	-0.001	-0.003	0.003	-0.802

Table 3: Difference Between Original and Reproduced Results (Original - Reproduced).

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDER	sBERT
GOAL Test (Window = 3)	6.832	0.056	14.932	8.235	1.079	33.736
GOAL Test (Window = 5)	6.905	0.107	14.889	8.312	1.130	33.868
GOAL Test (Window = 10)	**7.120**	0.095	**15.157**	8.449	**1.264**	**34.299**
GOAL Test (Window = 15)	7.024	0.091	15.105	8.493	1.232	34.010

Table 4: Window Size Comparison for MatchVoice (Zero-shot on GOAL Dataset). Red denotes the best performance and the second-best performance in Blue.

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDER	sBERT
SN-Caption-test-align before Fine Tuning	27.433	6.219	25.255	24.456	24.965	66.464
SN-Caption-test-align after Fine Tuning	20.12	4.75	23.89	20.35	21.23	63.421
GOAL Test before Fine Tuning	7.120	0.095	15.157	8.449	1.264	34.299
GOAL Test after Fine Tuning	8.24	0.401	16.138	8.92	2.18	38.76

Table 5: Comparison of MatchVoice Performance Before and After Fine-Tuning. The highest value for each metric is highlighted in Red, while the second-highest is highlighted in Blue.

5.3 Window Size Experiments

We also experimented with varying window sizes for MatchVoice in a zero-shot setting on the GOAL dataset. The results are presented in Table 4. Based on the experiments, we decided to use a window-size of 10 for rest of the experiments on GOAL dataset.

5.4 Fine-Tuning on GOAL

The results presented in Table 5 demonstrate the effectiveness of fine-tuning MatchVoice on the GOAL dataset. Despite being trained on only 100 videos due to resource constraints, the model shows noticeable improvements across all evaluation metrics for the GOAL test set. Specifically, BLEU-1 and BLEU-4 scores improved significantly from 7.120 to 8.24 and from 0.095 to 0.401, respectively. METEOR, ROUGE-L, CIDER, and sBERT scores also showed considerable improvements, indicating that the model has effectively learned from the fine-tuning process.

Importantly, the metrics for the SN-Caption-test-align dataset did not experience a substantial decrease after fine-tuning, particularly with respect to METEOR and sBERT. This observation suggests that the model is successfully retaining its ability to capture semantic similarity and paraphrasing, which is essential for commentary generation. The relative stability of these metrics underscores the model's capacity to generalize across different datasets, even with limited fine-tuning resources.

5.5 Ablation Study on Fine-tuning

To understand the contribution of different modalities in the fine-tuning process, we conducted an ablation study by selectively enabling vision and language components during training. The results are presented in Table 6.

From the table, we observe that:

- Fine-tuning only the vision encoder (row 2) yields improvements across all metrics compared to the frozen baseline (row 1), especially in CIDEr (+0.37) and BLEU-4 (+0.204).
- Fine-tuning only the language model (row 3) leads to even more pronounced gains in BLEU, METEOR, and semantic similarity (sBERT).
- Joint fine-tuning of both the vision encoder and the language model (row 4) produces the best overall performance, confirming that full end-to-end optimization leads to richer visual grounding and better linguistic fluency.

This ablation underscores the importance of synergistic tuning of both visual and linguistic components in multimodal models like ours.

6 Zero-shot Inference

To evaluate baseline generalization without task-specific supervision, we conducted zero-shot inference using two pretrained video-language models: **Video-ChatGPT** and **UniSoccer/MatchVision**. These models were tested on short 10–15 second video clips centered around events in our Goal dataset, created by cropping longer videos to satisfy the input constraints of the models (typically less than 30 seconds). This ensured compatibility and focused event relevance.

6.1 Video-ChatGPT Zero-shot Results

We first evaluated the Video-ChatGPT (Maaz et al., 2024) model, which follows a CLIP-based visual encoder and Vicuna-based LLM decoder architecture (Figure 4). As the model was primarily trained for general video summarization and lacks domain-specific grounding in soccer, its raw zero-shot performance was low.

The initial results, without any domain adaptation, are summarized below:

Post-processing involved anonymizing the generated commentaries to match the Goal dataset style by replacing names such as `player` and `team` with tokens like `[PLAYER]`, `[TEAM]`, etc., and removing subjective or filler language like "I am asked to..." or "Let me describe...". This led to noticeable improvements across all metrics. However, performance remained limited due to distribution shift: the model was not trained on soccer-specific language and lacked grounding in event terminology.

6.2 UniSoccer / MatchVision Zero-shot Results

We also evaluated the UniSoccer (Rao et al., 2025) model, which integrates spatiotemporal attention over video frames and supports downstream tasks like commentary generation, foul classification, and event recognition (Figure 5). Despite its specialization in soccer events, fine-tuning on the Goal dataset was infeasible due to lack of explicit event labels. Attempts at automatic weak-labeling (e.g., detecting "goal" or "corner" from commentaries)

Vision	Language	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	sBERT
✗	✗	7.125	0.105	14.889	8.462	1.469	33.373
✓	✗	7.491	0.309	15.300	8.691	1.840	34.170
✗	✓	8.102	0.401	15.780	8.776	2.050	36.896
✓	✓	**8.240**	**0.382**	**16.138**	**8.920**	**2.180**	**38.760**

Table 6: Ablation study on finetuning. Checkmarks indicate whether the vision encoder and/or the language model were fine-tuned.

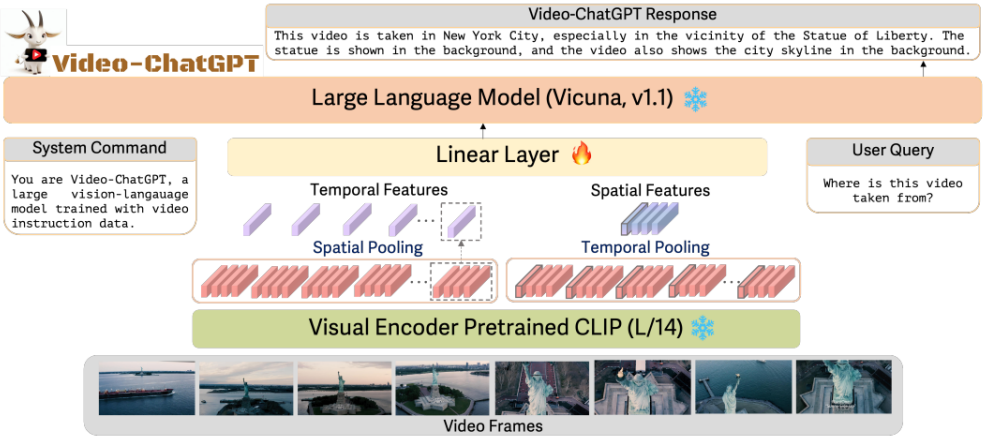

Figure 4: Architecture of Video-ChatGPT

were noisy and unreliable due to short or vague annotations.

The zero-shot and fine-tuned performance of UniSoccer on the SN-Caption-test-align subset is as follows:

Although fine-tuning on MatchTime improved scores, the results were still weaker than those of MatchVoice, which had direct supervision from the same dataset. We attribute this gap to dataset distribution shift: MatchVision was trained on SoccerNet-v2, a broader but stylistically different dataset from MatchTime.

The performaces of all the models is summarized in Figure 6. Post-processing (token anonymization and style cleaning) improves Video-GPT's zero-shot performance. Fine-tuning MatchTime on the GOAL dataset leads to the best results across all metrics, demonstrating the benefit of supervised domain alignment. Interestingly, MatchTime Before FT already outperforms Video-GPT post-processed, indicating a stronger inductive bias from soccer-specific pretraining.

7 Potential Issues in general Soccer Commentary Frameworks

- **Ground Truth Alignment:** One of the key challenges in soccer commentary generation is the misalignment between ground truth captions and actual human commentary. Future work should explore techniques for aligning temporally relevant commentary with specific game events.
- **Entity Identification and De-Anonymization:** The current datasets use anonymized tokens like `[PLAYER]`, `[TEAM]`, etc., which hinders the generation of personalized and context-rich commentary. A potential direction is the development of models capable of entity resolution and dynamic name generation.
- **Expanding Dataset Scale and Diversity:** To improve model generalization and robustness, larger and more diverse datasets are required. Incorporating extensive datasets such as *SoccerReplay-1988 (Rao et al., 2025)* could enable better training for event-rich, multi-view, and multilingual soccer scenarios.

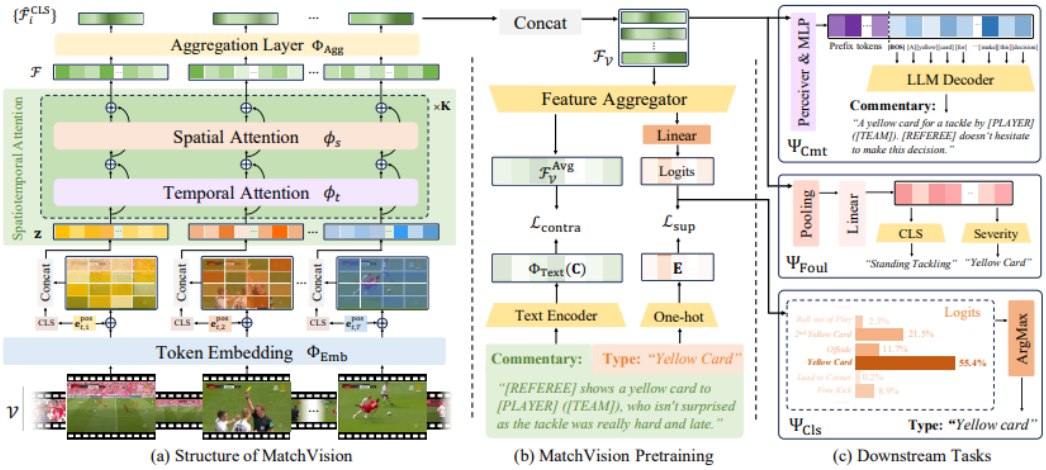

Figure 3. **Overview of MatchVision.** (a) The model architecture and its spatiotemporal feature extraction process; (b) Details of visual encoder pretraining, including supervised classification and video-language contrastive learning; (c) Implementation details of specific heads for various downstream tasks, including commentary generation, foul recognition, and event classification.

Figure 5: Architecture of MatchVision (UniSoccer)

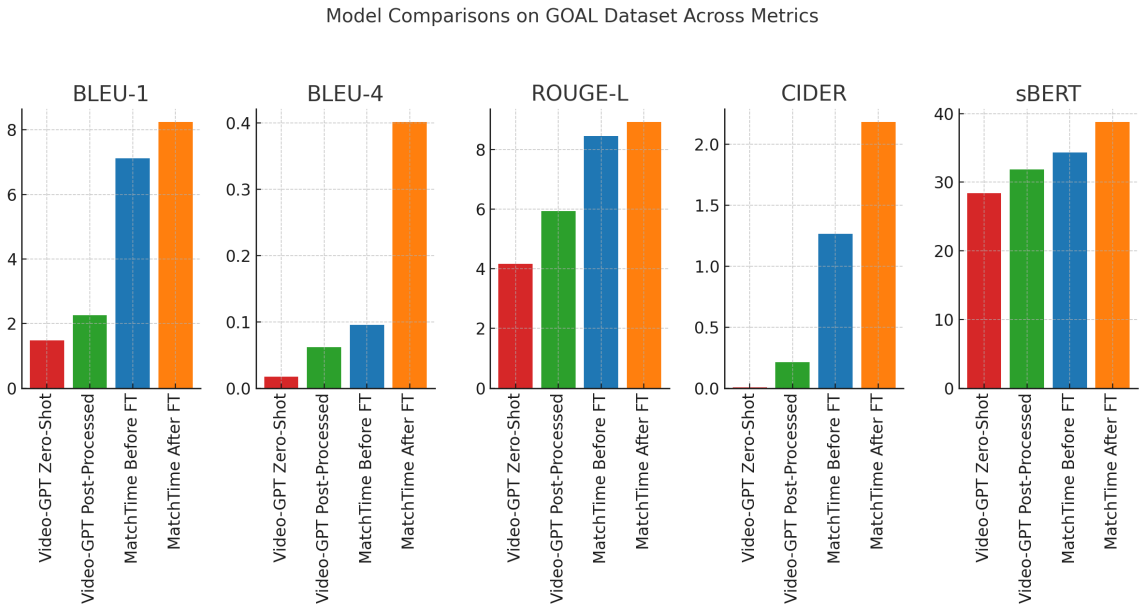

Figure 6: **Comparative Evaluation of Video-GPT and MatchTime on the GOAL Dataset.** The figure reports performance across five metrics: BLEU-1, BLEU-4, ROUGE-L, CIDER, and sBERT.

Setting	BLEU-1	BLEU-4	ROUGE-L	CIDER	sBERT
Zero-Shot	1.484	0.017	4.166	0.005	28.380
Zero-Shot + Post-Processing	2.260	0.062	5.935	0.214	31.842

Table 7: Video-ChatGPT zero-shot results before and after post-processing.

Model	BLEU-1	BLEU-4	ROUGE-L	CIDER	sBERT
UniSoccer (Zero-shot)	21.65	3.27	21.02	17.79	12.90
UniSoccer (Fine-tuned on MatchTime)	27.49	6.96	24.50	23.33	30.81

Table 8: UniSoccer performance on SN-Caption-test-align before and after fine-tuning.

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *Preprint*, arXiv:2303.00747.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan. Association for Computational Linguistics.

Nadjet Bouayad-Agha, Gerard Casamayor, and Leo Wanner. 2011. Content selection from an ontology-based knowledge base for the generation of football summaries. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 72–81, Nancy, France. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. *Preprint*, arXiv:2306.05424.

Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. 2023. SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries. abs/2304.04565.

Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Yuxiao Dong, Bin Xu, Lei Hou, Juanzi Li, Jie Tang, Weidong Guo, Hui Liu, and Yu Xu. 2023. Goal: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation. *Preprint*, arXiv:2303.14655.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Jiayuan Rao, Haoning Wu, Hao Jiang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Towards universal soccer video understanding. *Preprint*, arXiv:2412.01820.

Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. 2024. Matchtime: Towards automatic soccer game commentary generation. *Preprint*, arXiv:2406.18530.

Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104, Santiago de Compostela, Spain. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.