Live Football Commentary (LFC): A Large-Scale Dataset for Building Football Commentary Generation Models

Taiga Someya[1,2] **Tatsuya Ishigaki**[2] **Hiroya Takamura**[2]
[1]The University of Tokyo
[2]National Institute of Advanced Industrial Science and Technology (AIST)
taiga98-0809@g.ecc.u-tokyo.ac.jp
{ishigaki.tatsuya, takamura.hiroya}@aist.go.jp

Abstract

Live football commentary brings the atmosphere and excitement of matches to fans in real time, but producing it requires costly professional announcers. We address this challenge by formulating commentary generation from player- and ball-tracking coordinates as a new language–generation task. To facilitate research on this problem we compile the *Live Football Commentary (LFC)* dataset, 12,440 time-stamped Japanese utterances aligned with tracking data for 40 J1 League matches (≈ 60 h). We benchmark three LLM-based baselines that receive the tracking data (i) as plain text, (ii) as pitch-map images, or (iii) in both modalities. Human evaluation shows that the text encoding already outperforms image and multimodal variants in both accuracy and relevance, indicating that current LLMs exploit structured coordinates more effectively than raw visuals. We release the LFC transcripts and evaluation code to establish a public test bed and spur future work on tracking-based commentary generation, saliency detection, and cross-modal integration.

1 Introduction

Live match commentary is a vital medium for conveying the atmosphere and excitement of sports directly to spectators. By describing the match situation and players' movements in real time, commentators help viewers both grasp on-field situations accurately and heighten their immersion. However, providing such commentary requires specialized personnel and production infrastructure, leading to substantial costs. Consequently, delivering live commentary for lower leagues, amateur matches, and youth categories remains challenging.

In this paper, we tackle the problem of *automatic generation of football commentary*. Because analysing raw video is non-trivial, and wearable tracking devices (e.g., GPS vests) and optical tracking systems have become widespread (Thomas et al., 2017), we condition generation on the spatiotemporal coordinates of the players and the ball instead. The generated utterances are intended to be read or converted to speech in synchrony with the match.[1] Therefore, automatically producing immersive, natural commentary further demands identifying the most salient parts of each scene and the overall flow of play.

To establish this task, we first construct *Live Football Commentary (LFC)*, a dataset of timestamped commentary text covering 40 J1 League matches from the 2021 season. We then build and evaluate GPT-4o based models that take the corresponding tracking data as input and generate commentary. To examine how best to expose tracking data to large language models (LLMs), we experiment with three input encodings: (i) coordinates serialized as text, (ii) pitch-map images, and (iii) a naive text+image fusion. Human evaluation shows that the plain-text encoding already outperforms the image and hybrid variants in both accuracy and relevance, suggesting that simply appending visual frames does not necessarily yield better commentary and that effective multimodal fusion remains an open challenge. Together, the LFC corpus and our baselines provide the first publicly available benchmark for automatic football commentary, laying a foundation for future work on this task.

2 Related Work

Prior text-generation research on football data has focused mainly on **play-by-play text commentaries**—often called live text feeds— (Mkhallati et al., 2023; Rao et al., 2024). These brief textual updates allow fans to follow match developments without video. For example, a play-by-play log might read, "45' Corner to Liverpool." By con-

[1]Our focus is language generation. Coupling the output with an off-the-shelf text-to-speech system would enable fully spoken commentary, but we leave this engineering step for future work.

Proceedings of the 18th International Natural Language Generation Conference, pages 195–201
October 29–November 2, 2025.

Statistic	Before preprocessing	After preprocessing
Total utterances	35,375	12,440
Average utterances per match	884.4 ± 137.5	311.0 ± 106.6
Average duration per utterance (s)	4.7 ± 4.1	2.9 ± 1.1
Average characters per utterance	29.9 ± 27.4	18.5 ± 9.2

Table 1: Statistics of the commentary data described in Section 3 and after the preprocessing steps in Section 4.1.

trast, a synchronized spoken commentary would say, "Liverpool win a corner on the stroke of half-time!". The former is intended to be *read* after the fact or in parallel, whereas the latter is *heard* (or read as subtitles) in real time and must match the match tempo.

Earlier work has generated these play-by-play text commentaries from two main input sources: (i) structured event logs that list passes, shots, tackles and their locations (Taniguchi et al., 2019), and (ii) raw broadcast video of the match itself (Mkhallati et al., 2023; Rao et al., 2024). In addition, the PASS system (van der Lee et al., 2017) generated *post-match* summaries in Dutch from structured match statistics (e.g., possession percentage, pass success rate, etc.).

In contrast, work on *live commentary* generation—spoken or subtitle-style text delivered in sync with play—remains limited. Most prior efforts focus on e-sports titles or generic activity videos (Ishigaki et al., 2021; Saito et al., 2020; Marrese-Taylor et al., 2022; Wang and Yoshinaga, 2024). Very recently, SCBench (Ge et al., 2024) introduced a multi-sport commentary benchmark that also covers football, yet its systems consume *raw video* and the football portion totals roughly one hour of footage. By contrast, our LFC corpus covers 40 full matches ($\sim$60 hours) and, in this study, we focus on using the accompanying *structured tracking data* as input. To our knowledge, LFC is the first publicly available benchmark of timestamped football commentary that can be aligned with proprietary tracking data, even though only the commentary itself is distributed.

3 Live Football Commentary (LFC)

In this study, we created *Live Football Commentary (LFC)*, a dataset of Japanese commentary transcripts for 40 matches from the 2021 season of the J1 League, Japan's professional league.[2] To prepare the commentary, we started with official J League broadcast footage supplied by the Japan Professional Football League.[3] Because the original commentary audio could not be used for research due to copyright restrictions, we commissioned professional announcers to record fresh commentary for every match. They were given the following guidelines:

- Base each utterance strictly on observable facts so listeners can understand what is happening from the audio alone.
- Avoid unnaturally long periods of silence.
- In addition to formal football terminology, freely use commonly employed expressions (e.g., "minus cut-back," "pocket," "through pass," "advantage").
- Only commentators with prior experience calling live football matches were selected, and each match was assigned a single announcer.
- While commentating, refer to a pre-supplied roster and use player names whenever feasible.

The recorded commentary was transcribed using Adobe Premiere Pro's speech-to-text feature followed by manual corrections by a native speaker.[4] Each utterance was annotated with start and end timestamps; utterances separated by less than 0.1 second were merged into a single segment. Basic statistics and match information are summarized in Tables 1 and 3.

[2]Available at: https://kirt.airc.aist.go.jp/corpus/en/LFC

[3]https://www.jleague.co

[4]https://www.adobe.com/products/premiere.html

4 Experiments

4.1 Data Pre-processing

We first link the commentary data created in Section 3 with the corresponding tracking data.

The tracking data were purchased for research purposes from DataStadium Inc.[5] The linkage procedure is as follows. For each commentary segment, we take the timestamp two seconds *before* the utterance ends as the endpoint and extract the 50 consecutive frames (equivalent to five seconds) immediately preceding it. If the commentary start time precedes the first of these 50 frames, the segment is discarded because it likely describes earlier events. This yields 12,440 commentary–tracking pairs (50 frames each), from which we sampled 20 examples for human evaluation (Section 4.3).

4.2 Models

We build commentary generators based on OpenAI's GPT-4o (OpenAI, 2024) and evaluate their performance.[6] We compare three input formats for the tracking data—`text`, `image`, and `text+image`—plus a rule-based baseline, for a total of four model types.

`Nearest Player`. This simple rule-based baseline outputs the name of the player closest to the ball in the input frames. The heuristic rests on the observation that live commentators frequently mention the ball carrier or the player about to receive a pass; therefore, proximity to the ball is a strong cue. If the closest player changes within the five–second window, all such names are output in order of first appearance.

`text`. In this variant, the tracking data are serialized as text. For each player, we include position, shirt number, and velocity, together with ball position and velocity, team attacking direction, and kit colors. Positions are normalized to $[0, 1]$ pitch coordinates (Figure 4) and listed for 25 frames at 0.2-second intervals (five seconds total). If a kit has multiple colors, the most prominent color is shown. The prompt additionally contains instructions and cautions to steer generation[7], current match time and score, and four sample utterances drawn from training data outside the test set (Figure 3; the original Japanese prompt is provided in Appendix C).

`image`. In this variant, the tracking data are rendered as an animation and supplied to the model as a sequence of still frames (Figure 4).[8] The plotting color for each team matches the kit color specified in the text prompt. To enable the model to mention player names, we also provide textual information—names, positions, and shirt numbers—of players near the ball. As in `text`, the prompt includes additional guidelines and cautions, current match time and score, and four sample utterances drawn from outside the test set.

Instructions:
The following input provides ball and player position data for a single scene.
Using this data, generate natural, fluent *Japanese* live commentary.
Your commentary should appropriately include team and player names and accurately describe the match situation.

Example:
They pushed deep into the opposition half but play it back for now.
Nakano chases the ball.
Yamamoto uses his body well and wins possession!
From Fan Sotco it goes to Matsuoka, who has dropped deeper.

Cautions:
- Reflect passes and overall play flow precisely.
- Do not confuse home and away teams.
- Pay special attention to play location and direction.
- Output *exactly one* commentary sentence—nothing else.
- Keep the commentary roughly 10–30 Japanese characters (or shorter).
- Note that Frame 1, 2, 3 are chronological.

First half 21:16
Score: Yokohama FC 0 – 1 Sanfrecce Hiroshima

Frame 1
Ball: (0.32, 0.98) speed 3.27 m/s
Sanfrecce Hiroshima (attacking left, kit white):
Shunki Higashi (DF 24): (0.33, 0.98) speed 0.84 m/s *(nearest to ball)*
Tsukasa Morishima (MF 10): (0.34, 0.78) speed 2.07 m/s
Ezequiel (MF 14): (0.24, 0.81) speed 1.41 m/s
Hayao Kawabe (MF 8): (0.43, 0.82) speed 3.39 m/s
Yuta Imazu (DF 33): (0.52, 0.67) speed 0.59 m/s
Douglas Vieira (FW 9): (0.23, 0.49) speed 1.54 m/s
Kosei Shibasaki (MF 30): (0.27, 0.46) speed 2.27 m/s
Hayato Araki (DF 4): (0.48, 0.40) speed 0.31 m/s
Keisuke Osako (GK 38): (0.75, 0.51) speed 0.32 m/s
Yuki Nogami (DF 2): (0.38, 0.18) speed 1.48 m/s
Yuya Asano (MF 29): (0.22, 0.15) speed 1.22 m/s

Yokohama FC (attacking right, kit sky-blue):
Ryo Germain (FW 14): (0.32, 0.88) speed 4.25 m/s
Reo Yasunaga (MF 15): (0.28, 0.79) speed 0.93 m/s
Katsuya Iwatake (DF 22): (0.21, 0.82) speed 0.70 m/s
Kazuma Watanabe (FW 39): (0.38, 0.74) speed 1.88 m/s
Kohei Tezuka (MF 30): (0.28, 0.62) speed 0.93 m/s
Masatomi Tashiro (DF 5): (0.21, 0.63) speed 1.83 m/s
Shunsuke Nakamura (MF 10): (0.37, 0.56) speed 1.73 m/s
Hyokang Han (DF 26): (0.21, 0.50) speed 1.44 m/s
Yutaro Hakamata (DF 3): (0.21, 0.36) speed 0.98 m/s
Yusuke Matsuo (FW 37): (0.31, 0.32) speed 0.81 m/s
Yuta Minami (GK 18): (0.04, 0.49) speed 0.52 m/s

Frame 2
... (continue as needed)

Figure 1: English translation of the in-context prompt used for the `text` input. The model receives the Japanese prompt shown in Section C; this translation is provided here for demonstration purposes.

[5] https://datastadium.co.jp

[6] All experiments use `gpt-4o-2024-08-06`, the latest model version available at the time of writing.

[7] We set the desired commentary length to approximately (mean ± SD) of the dataset: mean 18.5 characters, SD 9.2.

[8] Because the API allows at most ten input images, we divide the animation into ten frames sampled every 0.5 seconds (five seconds total).

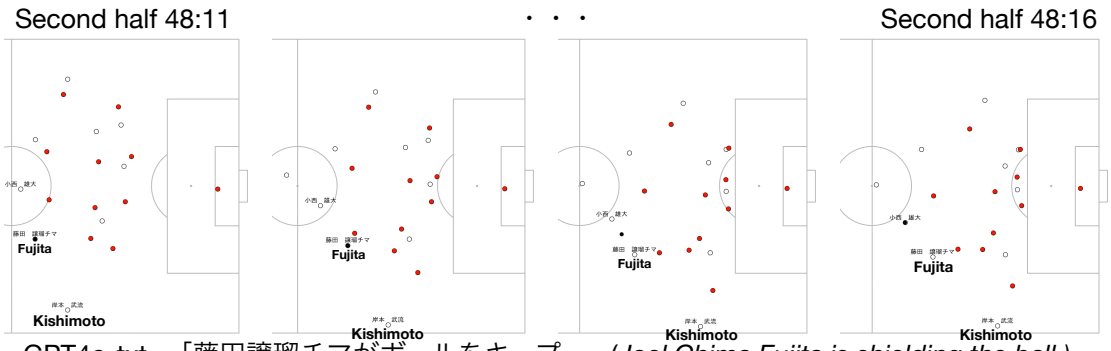

Figure 2: Examples of generated commentary (model) versus professional commentary (gold).

`text+image`. This hybrid prompt combines the information used in `text` and `image`. We merge the two prompts, remove the near-ball player data from the `image` part (because it duplicates the `text` content), and add explicit instructions that both images and coordinate data are provided.

4.3 Evaluation Method

We perform human evaluation of the generated commentary by two of the authors. We assess three aspects—*Accuracy*, *Relevance*, and *Ease of Understanding*—defined in this work, and we adopt a 3-point rating scale following prior practice in live text commentary evaluation (Taniguchi et al., 2019).

Accuracy The commentary correctly describes the animation or the match situation inferred from it. (Kit color may vary within the same color family; e.g., "blue" vs. "light-blue" is acceptable.)
3-point scale:

- 1: Major error(s).
- 2: Minor error(s).
- 3: No errors.

Examples

- High score: "Suzuki scores with a header!" (scene actually shows Suzuki heading the ball into the net)
- Low score: "Tanaka's volley finds the net!" (scene actually shows Suzuki's header)

Relevance The commentary directly relates to the key play or situation in the clip.
3-point scale:

- 1: Describes something that should not be commentated.
- 2: Relates to play but not the most salient moment.
- 3: Describes the most salient moment.

Examples

- High score: "Fantastic save by the goalkeeper!"
- Low score: "Suzuki is jogging." (not a highlight)
- Low score: "There are 22 players and a ball on the pitch." (no specific scene/action)

Ease of Understanding The text can be readily understood and evokes a concrete match situation (regardless of factual correctness).
3-point scale:

- 1: Unintelligible.
- 2: Partly unclear yet understandable.
- 3: Clear and vivid.

Examples

- High score: "Nakamura swings in a cross from the right flank."
- Low score: "He does that, and the ball goes there." (no concrete image)
- Low score: "There was a soccer match." (poor grammar, incomprehensible)

In this paper, we chose not to use automatic evaluation metrics, commonly used in the evaluation of text generation tasks. Because more than one commentaries can be good ones for a given scene, measuring deviation from a reference utterance offers limited insight. Instead, we judge how well each generated commentary describes and enhances the enjoyment of the scene in question.

5 Results

Human–evaluation scores and generation examples are presented in Table 2 and Section 4.2. For *Ease of Understanding*, every LLM-based variant (`text`, `image`, and `text+image`) scores essentially at the ceiling, confirming that the generated sentences are fluent and easy to read. In *Accuracy*, the `text` prompt outperforms both the `image` and `text+image` versions. *Relevance* shows the same ordering. Although the gaps are modest, they reinforce the finding of Ishigaki et al. (2021) that

	Accuracy	Relevance	Ease of Understanding
Nearest Player	1.28	1.30	1.45
text	**2.18**	**1.85**	**3.00**
image	1.53	1.48	**3.00**
text+image	1.55	1.60	2.98
Gold	2.53	2.55	2.78

Table 2: Human-evaluation scores for each model (3-point scale, averaged over two raters).The best scores among the models are in bold.

naïvely concatenating modalities (e.g., appending images to a text prompt) does not automatically lead to better commentary generation and can even dilute the useful signal in the structured text.

All variants, however, remain below 2.0 in *Relevance*. In other words, while the models often produce linguistically correct descriptions that capture *some* aspect of the scene, they frequently miss the most salient play that a human commentator would highlight. Taken together, the results suggest that (i) textual serialization of tracking data already conveys most of the information the model can readily exploit, (ii) richer cross-modal integration will be required to extract additional benefit from images, and (iii) future work should focus less on surface fluency and more on mechanisms for detecting which moments truly warrant commentary.

6 Conclusion

We release *Live Football Commentary (LFC)*, a large-scale dataset for training football-commentary models. Our GPT-4o baseline experiments show that text serialization beats image input, and a naive text+image fusion adds no benefit—mirroring Ishigaki et al. (2021). While outputs are fluent, *Relevance* remains below 2.0, highlighting the need for methods that better detect truly noteworthy moments. We hope LFC spurs work on saliency-aware and cross-modal approaches.

Acknowledgements

This work was supported by the New Energy and Industrial Technology Development Organization (NEDO), project JPNP20006, and AIST policy-based budget project, "R&D on Generative AI Foundation Models for the Physical Domain".

References

Kuangzhi Ge, Lingjun Chen, Kevin Zhang, Yulin Luo, Tianyu Shi, Liaoyuan Fan, Xiang Li, Guanqun Wang, and Shanghang Zhang. 2024. Scbench: A sports commentary benchmark for video llms. *CoRR*, abs/2412.17637.

Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating racing game commentary from vision, language, and structured data. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 103–113, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Edison Marrese-Taylor, Yumi Hamazono, Tatsuya Ishigaki, Goran Topić, Yusuke Miyao, Ichiro Kobayashi, and Hiroya Takamura. 2022. Open-domain video commentary generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7326–7339, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-10-17.

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

Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2020. SMASH corpus: A spontaneous speech corpus recording third-person audio commentaries on gameplay. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6571–6577, Marseille, France. European Language Resources Association.

Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. 2019. Generating live soccer-match commentary from play data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7096–7103.

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

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.

Qiong Wang and Naoki Yoshinaga. 2024. Tbd: Automatic commentary generation dataset for generic sports videos. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*. To appear.

A Matches Included in the Dataset

Table 3 lists all matches covered by LFC.

Round	Date	Home	Away
7	2021-04-02	C Osaka	Tosu
7	2021-04-03	Yokohama FM	Shonan
7	2021-04-03	Sendai	Kobe
7	2021-04-03	Nagoya	FC Tokyo
7	2021-04-03	Hiroshima	G Osaka
7	2021-04-03	Fukuoka	Sapporo
7	2021-04-03	Urawa	Kashima
7	2021-04-03	Yokohama FC	Kashiwa
7	2021-04-03	Kawasaki F	Oita
7	2021-04-04	Shimizu	Tokushima
8	2021-04-06	Yokohama FM	C Osaka
8	2021-04-07	Kashima	Kashiwa
8	2021-04-07	FC Tokyo	Sapporo
8	2021-04-07	Kawasaki F	Tosu
8	2021-04-07	Yokohama FC	Hiroshima
8	2021-04-07	Shonan	Nagoya
8	2021-04-07	Shimizu	Urawa
8	2021-04-07	G Osaka	Fukuoka
8	2021-04-07	Kobe	Oita
8	2021-04-07	Tokushima	Sendai
9	2021-04-10	Hiroshima	Shonan
9	2021-04-10	C Osaka	Fukuoka
9	2021-04-11	Sapporo	Kashima
9	2021-04-11	Sendai	Yokohama FM
9	2021-04-11	FC Tokyo	Kawasaki F
9	2021-04-11	Tosu	Yokohama FC
9	2021-04-11	Oita	Nagoya
9	2021-04-11	Urawa	Tokushima
9	2021-04-11	Kashiwa	G Osaka
9	2021-04-11	Kobe	Shimizu
10	2021-04-16	Sapporo	Yokohama FM
10	2021-04-17	Yokohama FC	Sendai
10	2021-04-17	Tokushima	Kashima
10	2021-04-17	Fukuoka	FC Tokyo
10	2021-04-17	Oita	Kashiwa
10	2021-04-17	Shonan	Kobe
10	2021-04-18	Kawasaki F	Hiroshima
10	2021-04-18	Nagoya	Tosu
10	2021-04-18	C Osaka	Urawa
10	2021-04-18	G Osaka	Shimizu

Table 3: Matches covered in LFC

B Annotator and Ethics Details

Participant instructions. The gist of the instruction sheet given to the professional announcers is given in Section 3.

Recruitment and payment. We hired a professional sports-broadcast agency via open bidding. One experienced announcer covered each match. The total fee was 2,248,400 JPY for the 40 games.

Consent. The agency provided written consent for the commentary to be used and redistributed for academic research on automatic commentary generation.

Ethics review. Under Japanese regulations the work does not require IRB approval because it involves no personal data or intervention with individuals.

Demographics. All announcers are adult native speakers of Japanese residing in Japan; no sensitive attributes were collected.

C Original Japanese Prompt

指示:
以下に入力されるのは，あるシーンに対応するボールと選手の位置データです.
このデータを基に，自然で流暢な日本語の実況を生成してください.
生成された実況には，チーム名や選手名を適切に含め，試合の状況を的確に表現してください.

例：
ここは，相手陣内深くまで入っていきましたけれども，一旦戻します
中野ボールを追いかける
しかし，体を使ってボールを奪いました山本です
ファンソッコから，少し下がった位置に来た松岡

注意：
・パスやプレーの流れを正確に反映してください.
・敵味方を混同しないように注意してください.
・プレーの場所や方向は特に注意してください.
・実況 1 つのみを生成しそれ以外のことは何も言わないでください.
・実況は，10文字から30文字程度かそれより短いものを生成してください.
・Frame1, 2, 3と時系列になっていることに注意してください.

前半 21分16秒
スコア: 横浜ＦＣ 0 - 1 サンフレッチェ広島

Frame 1
ボール: (0.32, 0.98) 速度: 3.27 m/s
サンフレッチェ広島 (攻撃方向: 左ユニフォーム: 白):
東　俊希 (DF 24): (0.33, 0.98) 速度: 0.84 m/s (最もボールに近い選手)
森島　司 (MF 10): (0.34, 0.78) 速度: 2.07 m/s
エゼキエウ (MF 14): (0.24, 0.81) 速度: 1.41 m/s
川辺　駿 (MF 8): (0.43, 0.82) 速度: 3.39 m/s
今津　佑太 (DF 33): (0.52, 0.67) 速度: 0.59 m/s
ドウグラス　ヴィエイラ (FW 9): (0.23, 0.49) 速度: 1.54 m/s
柴崎　晃誠 (MF 30): (0.27, 0.46) 速度: 2.27 m/s
荒木　隼人 (DF 4): (0.48, 0.40) 速度: 0.31 m/s
大迫　敬介 (GK 38): (0.75, 0.51) 速度: 0.32 m/s
野上　結貴 (DF 2): (0.38, 0.18) 速度: 1.48 m/s
浅野　雄也 (MF 29): (0.22, 0.15) 速度: 1.22 m/s

横浜ＦＣ (攻撃方向: 右ユニフォーム: 水):
ジャーメイン　良 (FW 14): (0.32, 0.88) 速度: 4.25 m/s
安永　玲央 (MF 15): (0.28, 0.79) 速度: 0.93 m/s
岩武　克弥 (DF 22): (0.21, 0.82) 速度: 0.70 m/s
渡邉　千真 (FW 39): (0.38, 0.74) 速度: 1.88 m/s
手塚　康平 (MF 30): (0.28, 0.62) 速度: 0.93 m/s
田代　真一 (DF 5): (0.21, 0.63) 速度: 1.83 m/s
中村　俊輔 (MF 10): (0.37, 0.56) 速度: 1.73 m/s
韓　浩康 (DF 26): (0.21, 0.50) 速度: 1.44 m/s
袴田　裕太郎 (DF 3): (0.21, 0.36) 速度: 0.98 m/s
松尾　佑介 (FW 37): (0.31, 0.32) 速度: 0.81 m/s
南　雄太 (GK 18): (0.04, 0.49) 速度: 0.52 m/s

Frame 2
...

Figure 3: Original Japanese prompt supplied to the model for the `text` input (included here for reproducibility).

D Example Input Image Frame

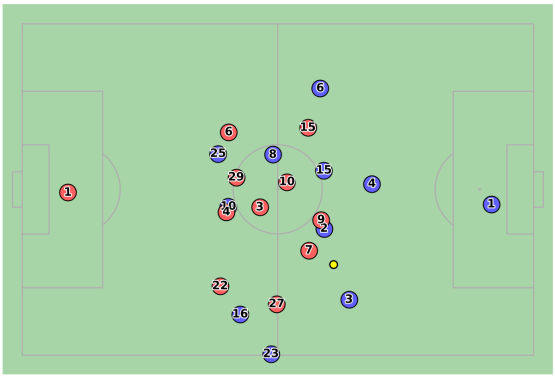

Figure 4: Example input frame. Player and ball coordinates and player shirt numbers are displayed.