令和 7 年度

静岡大学大学院　総合科学技術研究科

工学専攻　数理システム工学コース

修士論文

映像と位置情報時系列データを用いたマルチモーダル推論によるサッカーのラストタッチ判定

指導教員	岡部　誠　准教授
提出者	524c5016　廣瀬　有真
提出日	2026 年 2 月 2 日

目次

第 1 章　はじめに 5

第 2 章　関連研究 8

2.1　スポーツ映像解析と物体検出・追跡 8

2.2　サッカーにおける自動判定・支援システム 8

2.3　大規模言語モデルとマルチモーダル推論 9

第 3 章　提案手法 10

3.1　処理フロー 10

3.2　拡大映像の生成 11

3.2.1　モデル選定とファインチューニング 12

3.2.2　データセットとクラス定義 13

3.2.3　学習とパラメータ設定 14

3.2.4　追跡アルゴリズムと拡大映像の生成 15

3.3　静止画シーケンスの作成 16

3.3.1　抽出フレーム枚数の設定 16

3.3.2　抽出フレームの詳細 17

3.4　拡大映像の詳細検出と追跡 18
3.4.1　拡大領域における再検出 18
3.4.2　IoU に基づく簡易追跡 18
3.5　位置情報時系列データのノイズ除去・補正 19
3.5.1　欠損補完と光学的フローによる微修正 19
3.5.2　RTS スムーシングによる軌道推定 20
3.6　マルチモーダル推論 21
3.6.1　入力データの構成 21
3.6.2　プロンプト作成 22
3.6.3　推論モデル 22
第 4 章　結果と考察 23
4.1　実験評価 23
4.1.1　データセット 23
4.1.2　比較手法 23
4.2　定量評価 24
4.2.1　提案手法における構成要素の有効性 24

4.2.2　既存手法及び他モデルとの比較 25
4.3　定性評価 26
4.3.1　成功事例：位置情報時系列データが決定打になったケース 26
4.3.2　成功事例：位置情報時系列データが誤検出を防いだケース 28
4.3.3　失敗事例：検出器によるエラー 31
4.3.4　失敗事例：複雑な事象 31
第５章　まとめと今後の展望 33
謝辞　33
参考文献 35
付録　38

第1章　はじめに

サッカーとは，1 チーム 11 人の 2 チームが手や腕（ゴールキーパーを除く）以外を用いて得点数を競う競技であり，世界の競技人口は約 2 億 6500 万人[1]と報告されている．サッカーの勝敗は数少ない得点機会によって決まることが多く，審判員の判定は試合結果に大きな影響を与える．また近年は戦術の高度化に伴ってプレースピードが向上しており，審判員には高い身体能力に加えて，瞬時の事象を正確に見極める判断能力が継続的に求められている．

審判員が試合中に扱う判定のうち，特に頻繁に発生する事象としてボールアウト判定がある．図 1 に示すように，ボールがタッチラインまたはゴールラインを越えた際，最後にボールへ触れたチーム（Home/Away）を見極める必要がある．本研究では SoccerNet-v2[2]に含まれる 500 試合の判定記録を調査・分析し（図 2），全判定イベントのうち約 33．5%をボールアウト判定が占めることを確認した．これは，審判業務の中でボールアウト判定が最も頻出する意思決定の一つであることを示す．

一方で，ボールアウト判定は常に容易とは限らない．とりわけ，選手同士の密集やボールと身体部位の重なり（オクルージョン）により，現場の視点では接触の瞬間が視認しづらい場面が生じる．さらに，試合を通じて頻繁な判定を繰り返すこと自体が累積的な認知的負荷となり，判断の確信度に影響を与える要因となり得る．そして，審判員は高い走行強度などの身体的負荷がかかった状態で，知覚・認知的プロセスを連続的に遂行しなければならない．McEwan ら[3，4]は，高い身体的負荷（高心拍数や高強度走行）が審判員の注意リソースに負担を与え，事象への集中力を低下させる可能性を示唆している．このように，「頻発する判定」×「見えづらい局面」×「疲労」が重なると，判定ミスのリスクが高まることが懸念される．

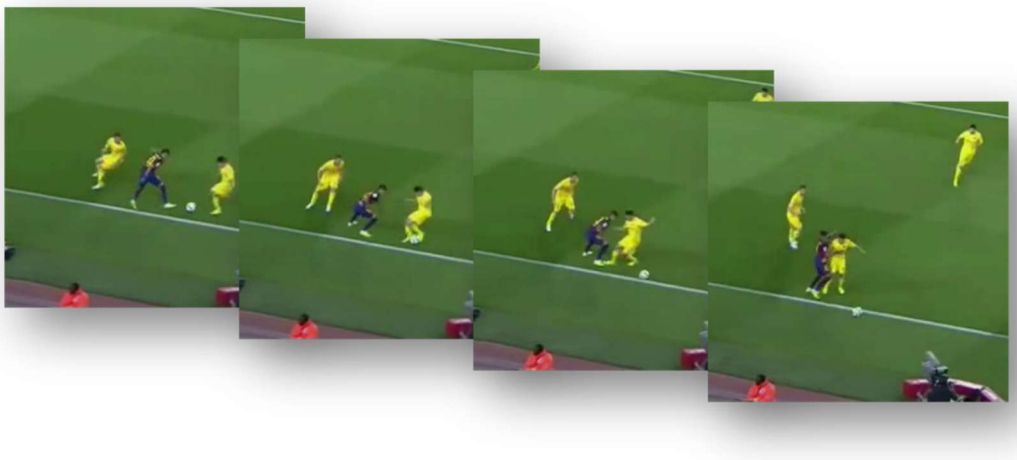


図１．ボールアウトの様子

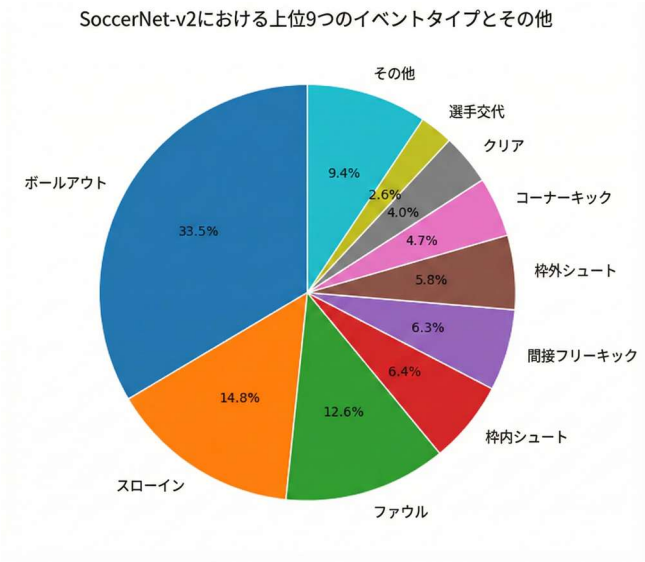


図２．大規模サッカー動画のデータセットである SoccerNet-v2 に含まれる 500 試合の判定記録を調査・分析した結果

近年，サッカーにおいて複数の高速度カメラと専任の審判団によってビデオ・アシスタント・レフェリー（VAR）システムや高価なセンサーを用いたゴールラインテクノロジーが導入され，誤審の減少に寄与している．[5]しかしながら，現代テクノロジーの恩恵を享受できるのは欧州五大リーグやワールドカップなどのトッププロの舞台に限られている．さらに，現在の VAR は「得点」「PK」「退場」といった試合の結末を左右する重大な事象のみを対象としており，審判の疲労の主因である日常的なボールアウト判定の多くは VAR の介入対象外である．そのため，少年サッカーや地域リーグ等のアマチュア・育成年代の現場でも導入可能で，頻発するボールアウト判定に対して広範に意思決定を支援できる安価で汎用性の高い技術が求められている．

従来の画像処理を用いたスポーツ解析や自動判定システム[6, 7]の研究は数多く行われている．しかし，その多くはスタジアムの設置された固定カメラによるカリブレーション（位置合わせ）を前提とするか，あるいは特定のシーンを学習させるために膨大な教師データを必要とするものであった．ディープラーニング，物体検出モデルを用いた手法[6, 8, 9]は高い検出精度を誇るが，サッカーのラストタッチ判定のようなタスクにおいては，「ボールと選手が重なっている」という静的な情報だけでは不十分である．「ボールが足に当たった瞬間に軌道がどう変化したのか」という物理的な因果関係や，「その選手が Home/Away どちらのチームに所属しているか」という文脈的情報を同時に処理する必要があるため，単純な物体検出だけでは解決困難な課題であった．

また，入力映像から直接判定結果を出力するようなエンドツーエンドのモデルを作成しようとすれば，あらゆる角度，天候，ユニフォームの組み合わせを網羅した巨大なデータセ

ットが必要となり，個別のチームや環境への適用コスト（ファインチューニングの手間）が高くなる．多様な撮影条件に対応し，かつ物理法則に基づいた納得感のある判定を行うためには，単なるパターン認識を超えた，高度な推論能力を持つアプローチが必要となる．

そこで本研究では，頻発し，かつ現場で見えづらい局面があるボールアウト判定に着目し，その中でも最終接触選手（ラストタッチ）の帰属チーム判定を対象として審判員の意思決定を補助することを目的とする．具体的には，単一視点の映像から，物体検出により得られる拡大画像と位置情報時系列データを抽出し，さらに高度な推論能力を有するLarge Vision-Language Model（LVLM）を組み合わせることで，ゼロショット推論によるラストタッチ判定を行う．ここでいうゼロショット推論とは，個別試合や特定チームに対する追加学習を行わず，拡大画像・位置情報時系列データ・ユニフォーム画像とプロンプトに基づいて判定する設定を指す．なお，本研究は判定の最終決定を自動化するものではなく，審判員が自身の判断と照合可能な根拠を提示することで意思決定を支援することを目的とする．

第 2 章　関連研究

2.1　スポーツ映像解析と物体検出・追跡

スポーツの映像解析において，選手やボールの位置を正確に特定することは，戦術分析や自動ハイライト生成などのあらゆる応用の基盤となる技術である．初期の研究では，背景差分法や色情報に基づく古典的な画像処理手法を用いられてきたが，照明条件の変化や複雑な背景，選手同士の重なり（オクルージョン）に対して脆弱である課題があった．しかし，近年，ディープラーニング（深層学習）の発展，特に畳み込みニューラルネットワーク（CNN）の登場により，物体検出技術の精度は飛躍的に上昇した．特に Redmon らによって提案された YOLO（You Only Look Once）シリーズ[8]は，高い検出精度とリアルタイム性を兼ね備えた画期的なモデルであり，スポーツ映像のような動きの速い対象の検出において広く用いられる手法となっている．

また，検出した物体を時間方向に紐づける「追跡（Tracking）」技術も重要である．Bewley らによる SORT[10]や，その改良版である DeepSORT[11]，ByteTrack[12]などの Multi-ObjectTracking（MOT）アルゴリズムは過去の位置座標から次フレームの物体位置を推定するカルマンフィルタと，前後のフレームで検出された物体同士を最適に紐付けるハンガリアンアルゴリズムを組み合わせることで，一貫した個体追跡を実現している．しかし，これらの技術はあくまで「画像上のどこに何があるのか」という座標情報を出力するものであり，その動きが競技ルール上どのような意味を持つのか（例：ラストタッチはどちらか）という高度な意味理解を行うものではない．

2.2　サッカーにおける自動判定・支援システム

サッカーにおける判定支援システムとして最も普及しているのが，ビデオ・アシスタント・レフェリー（VAR）および，ボールライン・テクノロジー（GLT）である．GLT の代表例である Hawk-Eye システム[13]は，スタジアムに設置された多数の高速度カメラからの映像を用い，三角測量の原理でボールの正確な 3 次元位置をミリ単位で特定する．これにより，ゴールラインを割ったか否かを物理的に判定することができる．また，VAR システムは，多視点の映像を専任の審判員が確認する Human-in-the-Loop のアプローチをとっている．これらのシステムは非常に高精度であるが，導入には多額のコストと大規模なインフラ設備が必要であり，プロリーグ以外での運用は困難[5]である．

一方でコンピュータービジョンを用いた安価な判定システムの自動化に関する研究も進められている．Giancola らは，大規模データセット SoccerNet[14]を公開し，アクションスポッティング（サッカーの試合中におけるゴール，カード交代などのイベント検知）のベンチマークを確立した．その後，Deliège らによって，より詳細なアノテーションを付与した拡張版である SoccerNet-v2[2]へと発展した．これに続く多くの研究は，映像全体の特徴量

から「いつイベントが起きたか」を分類することに焦点を当てており，「最後に誰が触ったのか」といった微細な物理的接触の判定や，チームごとの詳細は帰属判定までは踏み込んでいないものが多い．

また，より実践的な審判支援として，Held らは VARS（VideoAssistantRefereeSystem）[6]を提案している．彼らは SoccerNet データセットを拡張し，ファウルの有無や種類，その重要度（カードの提示が必要か否か）を自動分類するマルチタスクモデルを構築した．この研究は，コンピュータービジョンが複雑なルールの解釈や主観的な判定基準の学習にも適用可能であることを示している．ただし，VARS はファウルの有無やカード判断といった意味的イベントの分類を目的としており，本研究が焦点とするようなボールと選手の瞬間的な接触をミリ秒単位で特定するタスクは扱っていない．したがって，VARS の枠組みだけではラストタッチ判定に必要な精密な物理的推定（ボール・身体部位の時系列的トラッキングや接触瞬間の特定）は十分に行えない．さらに最新の研究では，Held らによって X-VARS[15]が提案されている．これは，マルチモーダル LLM を用いてサッカーのビデオを審判の視点から理解し，判定の根拠を言語で説明（Explainability）することを目的とした手法である．彼らは SoccerNet-XFoul という 2 万件以上の解説付きデータセットを構築し，モデルがルールの解釈に基づいた対話を行えることを示した．しかし，これらのモデルは依然として視覚情報のみに依存しており，本研究が提案する位置情報時系列データ（物理情報）を統合した精密な接触判定については十分に扱われていない．

2.3 大規模言語モデルとマルチモーダル推論

近年，自然言語処理分野における Transformer アーキテクチャ[16]の成功を受け，画像と言語の統合的に扱うマルチモーダル大規模言語モデル（Large Vision-Language Model:LVLM）が急速に発展している．ChatGPT や Gemini，Qwen2-VL[17]が代表とされる最新の LVLM は，画像を単なる数値行列としてではなく，意味的な文脈を含んだ情報として解釈し，ユーザーの自然言語による指示（プロンプト）に基づいて高度な推論を行うことができる．

これらのモデルは，従来の教師あり学習を必要とせず，未知のタスクに対してもプロンプトエンジニアリングのみで対応できるゼロショット推論の能力を有している．この特性は，ルールが複雑であり，かつ状況が多岐にわたるスポーツの判定タスクにおいて有効であると考えられる．しかし，LVLM は一般的な画像理解には優れているものの，スポーツ映像の「ボールの微妙な回転変化」や「数フレームの軌道変化」といった物理的な時系列情報の処理に関しては，依然として課題が残るとされている．そこで本研究では，LVLM に対して単に画像を見せるのではなく，物体検出によって得られた物理的な位置情報時系列データを言語的なコンテキストとして同時に与えることで，視覚情報と物理情報の双方を活用した高精度な推論を実現するアプローチをとる．

第 3 章　提案手法

本研究では，物体検出モデルとマルチモーダル大規模言語モデル（LVLM）を用いて，サッカーのラストタッチ判定を支援する手法を提案する（図 3）．

映像から物理情報の抽出を担う物体検出モデルには，最新のアーキテクチャであり，小物体検出と推論速度に優れた YOLOv11[9]を選定した．本研究では，これをサッカー映像に特化させるために，独自のデータセットを用いたファインチューニングを施して使用する．

一方で，抽出された情報をもとに最終的な判定を行う推論エンジンには，画像と言語をシームレスに理解し，複雑な事象に対する理論的推論が可能な GPT-4o を採用した．これは後述する比較実験（4.2.2 節）において他のモデルと比較して最も高い精度と安定性を示したためである．この 2 つのモデルを組み合わせることで，画像認識単体では困難な「文脈の理解」と，言語モデル単独では困難な「物理的な詳細把握」を相互に補完するシステムを実現する．

3.1　処理フロー

本提案手法における一連の処理フローを図 3 に示す．本システムは以下の 6 つのステップで構成される．

1．判定対象の指定：
入力として，ラストタッチが発生した瞬間（タイムスタンプ t），ラストタッチの含んだ 6 秒の映像（以下，全体映像），両チームのユニフォーム画像を受け付ける．

2．拡大映像の生成：
入力された全体映像に対して YOLOv11 を適用してボールの位置を特定し，その座標を中心とした 320×320 ピクセルの領域をクロッピングすることで，視認性を高めた拡大映像を作成する．

3．静止画シーケンスの作成：
生成された拡大映像から指定されたタイムスタンプtを中心として，前後 5 フレームを含む計 11 フレームの連続する静止画像として抽出する．これにより，接触の瞬間の前後における微細な変化を捉える．

4．拡大映像の詳細検出と追跡：
抽出された拡大映像に対して，再度 YOLOv11 を適用してボールと選手を検出・追跡する．これにより，フレームごとのバウンディングボックス（以下，BBox）情報を取得し，初期の検出データ(RawTracks)を生成する．

5．位置情報時系列データのノイズ除去・補正：
検出データに含まれるノイズや欠損を除去するため，カルマンフィルタ[20]および RTS(Rauch-Tung-Striebel)スムーシング[18]，EMA（指数移動平均）を用いた多段階の補正処理を適用し，ボールの位置情報時系列データを構築する．

6．マルチモーダル推論：
11 枚の画像, 両チームのユニフォーム画像, 平準化処理済みの位置情報時系列データ, および構造化されたプロンプトを LVLM(GPT-4o)に入力し，最終的な判定結果と根拠を出力する．

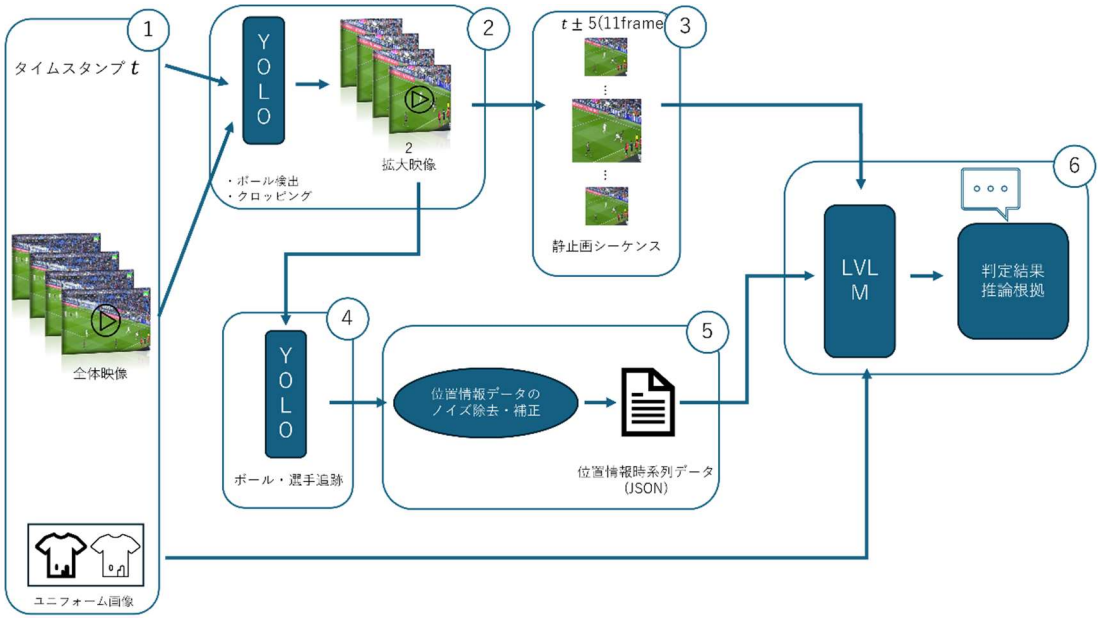


図 3．提案手法の全体パイプライン．①判定対象の指定，②拡大映像の生成，③静止画シーケンスの作成，④拡大映像の詳細検出と追跡，⑤位置情報時系列データのノイズ除去・補正，⑥マルチモーダル推論の 6 ステップで構成される．ユーザが指定したタイムスタンプ t を起点とし，視覚情報（t±5 の 11 枚）と物理情報（平準化済み位置情報時系列データ：JSON）を統合して LVLM が判定結果と推論根拠を出力する．

3.2　拡大映像の生成

入力された全体映像に対して YOLOv11 を適用してボールの位置を特定し，その座標を中心とした 320×320 ピクセルの領域をクロッピングすることで，視認性を高めた拡大映像を作成する．

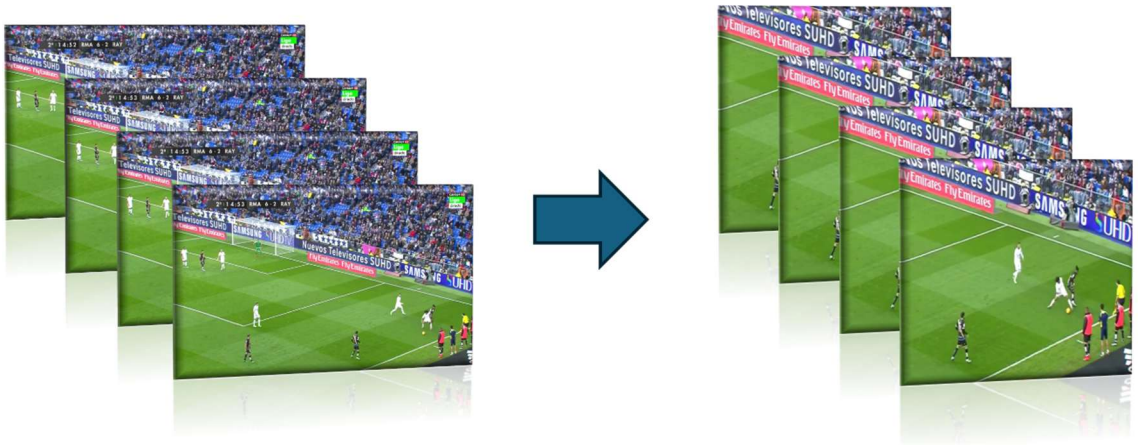


図４．全体映像と拡大映像の関係．全体映像（左）は入力された元映像であり，拡大映像（右）は YOLOv11 により検出したボール座標を中心に 320×320 ピクセルでクロッピングして生成した映像.

3.2.1　モデル選定とファインチューニング

本手法において物体検出フェーズでは，リアルタイム性と検出精度のバランスに優れた YOLOv11 アーキテクチャを採用した．YOLO シリーズは，画像全体を一度に処理する 1 段階の検出器であり，動画処理における高速な推論が可能である．しかし，一般的なデータセット（COCO 等）で事前学習されたモデルをそのままサッカーに映像に適用した場合以下の課題が生じる．

1．**小物体の検出限界：**
サッカーボールは画面全体に対して極めて小さく，特に全体映像では数ピクセル程度になる．図５（左）に示す通り，選手が密集する場面や引きの映像では，事前学習モデルはボールを背景やノイズとして処理し，検出に失敗する（未検出）傾向がある．拡大映像を正確に生成するためには，まず全体映像でボールを確実に捉える必要があり，この小物体への感度向上が必須である．

2．**オクルージョンへの脆弱性：**
ラストタッチの瞬間は選手の足とボールが密接しており，一般的なモデルではボールを「選手の一部」や「別の物体」として誤認し，検出が途切れることが多い．また，図６のように，選手の足先をボールと見間違えるなどの誤検出も発生しやすい．

そこで本手法では，YOLOv11 をファインチューニングしてサッカー映像に特化させ，使用することとした．

3.2.2　データセットとクラス定義

本研究では特定のスタジアムや撮影条件に過学習することを防ぎ，様々な環境下での検出精度を確保するため，コンピュータービジョン向けのデータセットプラットフォームである RoboflowUniverse 上で公開されている複数のサッカー関連データセットを収集，統合し学習データを構築した．

統合後のデータセットは合計 62,536 枚の画像から構築され，これを学習用，検証用，評価用に以下の通りに分割した．

・学習データ：57,118 枚
・検証データ：2,726 枚
・テストデータ：2,692 枚

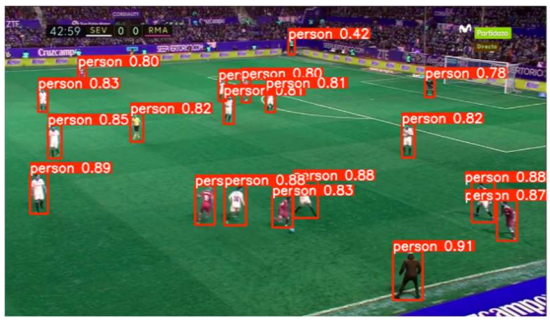
各データセット間ではアノテーションやクラス名や定義が異なる場合があったため，本研究の目的に合わせて以下の 2 クラスにラベルを統一・再構築する前処理を行った．

学習データの構築にあたっては，様々な環境のサッカーの映像からフレームを抽出し，アノテーションを行う必要がある．検出対象となるクラスは以下の 2 クラスを定義した．

1．Ball:
　サッカーボール
2．Player:
　フィールド上に存在するすべての人物（フィールドプレイヤー，ゴールキーパー，審判員を含む）

本手法では YOLO の段階で「Home/Away」や「審判員」といった属性の分類は行わない．これらは，検出された「Player」クラスの領域（BBox）内の画像と入力されているユニフォーム画像によって LVLM によって事後的に分類する設計とした．

このように大規模かつ多様なソースを統合することで，ボールが選手の陰に隠れるオクルージョンシーンや，遠距離からの全体映像，夜間照明下の映像など，実環境で想定されるあらゆるバリエーションをモデルに学習させることが可能となった．



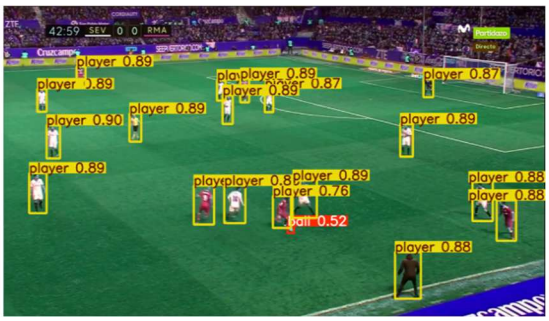


図 5. 全体映像におけるボール検出の比較（未検出の解消）（左）選手密集地帯において，事前学習モデルがボールを見失っている例．（右）ファインチューニングモデルにより，微小かつ遮蔽の多い状況下でもボールの検出に成功している例．

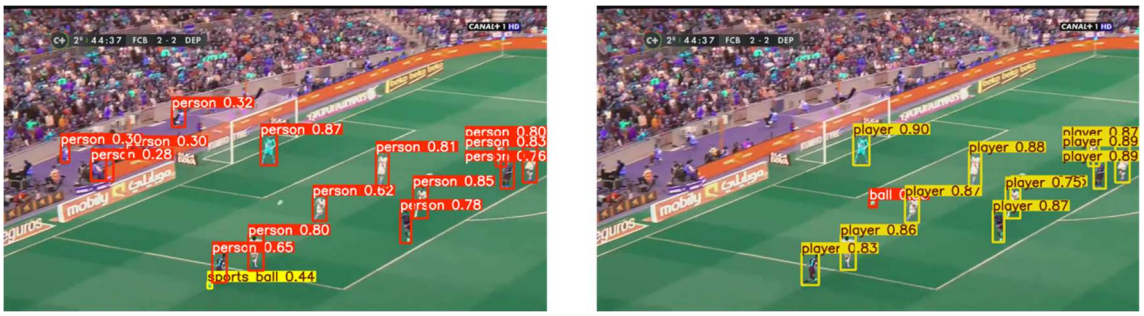


図 6. 事前学習モデルと提案モデルによる識別精度の比較（誤検出の抑制）（左）事前学習モデルによる出力．（右）ファインチューニングモデルによる出力．ファインチューニングモデルにより誤検出を防いでいる例．左の事前学習モデルでは選手の足をボール判定しているが右のファインチューニングモデルはボールを正確に検出できている．

3.2.3 学習とパラメータ設定

本研究では，全体映像における微小なボールの検出精度を最大化するため，YOLOv11 シリーズの中で最もパラメータ数が多く，高い特徴抽出能力を持つ YOLOv11x(ExtraLarge)をベースモデルとして採用した．学習におけるハイパーパラメータ設定を表 1 に示す．

特に重要な戦略として，入力画像サイズ（imgsz）を，YOLO の標準的な 640 ピクセルではなく，1088 ピクセルに拡張して設定した．全体映像においてサッカーボールは極めて小さな領域(数ピクセル四方)にしか映らないため，画像サイズを縮小(ダウンサンプリング)する過程で情報が消失するリスクが高い．高解像度のまま学習・推論を行うことで，ボールの微細な特徴量を保持し，検出漏れ（False Negative）を防ぐ設計とした．

学習は 100 エポック行い，GPU メモリの制約およびモデルサイズを考慮してバッチサイズは 4 に設定した．最適化手法には初期学習率 0．01，重み減衰（Weight Decay）0．0005 を適用し，Automatic Mixed Precision(AMP)を有効化することで学習の安定化と効率化を図った．

パラメータ	設定値	備考
モデルアーキテクチャ	YOLOv11x	シリーズ最大モデルを採用し精度を優先
入力解像度	1088	微小物体（ボール）検出のため高解像度化

パラメータ	設定値	備考
エポック	100	十分な収束を確認
バッチサイズ	4	高解像度・大型モデルによるメモリ制約のため
クラス	2	'ball'，'player'
オプティマイザー	SGD/AdamW	(YOLODefaultAuto)lr0=0．01

表 1．YOLO のファインチューニングの設定

3.2.4　追跡アルゴリズムと拡大映像の生成

全体映像からボールを中心とした拡大映像を生成する際，単一フレームの物体検出結果のみに依存してクロッピングを行うと，検出ノイズによる画面の微細な振動（ジッター）や，オクルージョン時のターゲットロストにより，視認性の低い映像となる課題がある．そこで本手法では，YOLO の検出結果に対して物理モデルに基づくカルマンフィルタとゲーティング処理を適用し，ボールの推定位置を滑らかに追従するように拡大映像を生成した．

１．状態空間モデルによる軌道推定

ボールの運動を等速直線運動と仮定し，時刻kにおける状態ベクトル$x_k = [x_k, y_k, v_{xk}, v_{yk}]^T$を定義する．全体映像上のボール検出座標$z_k$に対し，以下の状態遷移モデルと観測モデルを適用する．

$$\boldsymbol{x_k = Fx_{k-1} + w_k}$$

$$\boldsymbol{z_k = Hx_k + v_k}$$

ここで，$\boldsymbol{F}$は前の時刻から現在の状態を予測する状態遷移行列，$\boldsymbol{H}$は内部状態から観測される位置を取り出す観測行列である．また，w_kおよびv_kは風などの影響による予測できない物理的なズレ（プロセスノイズ）と観測の誤差（観測ノイズ）を表す．

実装においては，これらのプロセスノイズ，観測ノイズの大きさ$\boldsymbol{Q}$および$\boldsymbol{R}$を定義すると

$\boldsymbol{Q}$を10^{-4}，$\boldsymbol{R}$を 10 に設定した．これにより，「YOLO の瞬間的な検出位置のズレ（観測ノイ

ズ)」よりも「物理法則に基づいた予測軌道の滑らかさ（慣性)」を重視し，安定したクロッピング中心座標を算出する設計とした．

２．ゲーティング処理による誤検出の排除

YOLO が誤って別の物体（選手のスパイクや審判の旗など）をボールとして検出した場合，拡大映像の視点が大きく逸脱してしまうリスクがある．これを防ぐために，予測位置からの距離に基づくゲーティング（Gating）を実装した．予測位置$\widehat{z_k}$と観測位置z_kのユークリッド距離の閾値がT_{gate}を超える場合，その検出を「誤検出」とみなして棄却して，予測値のみを用いて更新を行う．

$$IF \|z_k - \widehat{z_k}\|_2 > T_{gate} \ then\ use\ \widehat{z_k}$$

本手法では画像サイズ 1280×720 に対して，$T_{gate} = 240$ピクセルに設定した．これにより，物理的に不自然な急加速を伴う移動をノイズとして排除している．

３．オクルージョン対策と欠損補完

選手とボールが交錯するラストタッチの局面では，ボールが数フレームに渡って遮蔽（オクルージョン）される頻度が高い．本手法では，連続欠損許容数を 5 フレームに設定した．検出が途切れた場合でも予測位置による追尾を継続する．また，5 フレームを超えて検出が復帰しない場合は，予測フレームの暴走を防ぐために速度成分(v_x, v_y)を 0 にリセットする安全策を講じている．これにより，ボールが一瞬隠れた場合でも，カメラ（クロップ領域）がその動きを予測して追従し続けるため，ボールが再出現した際にフレームアウトすることを防ぎ，常にボールを画面中央にとらえた拡大映像の生成が可能になった．

3.3　静止画シーケンスの作成

3.2 節で生成された拡大映像は，ボールを中心とした連続的な映像データである．しかし，LVLM を用いた推論においては，長時間の映像全体を入力することは，計算コストの増大だけではなく，無関係なフレームがノイズとなり判定精度を低下させる要因となる．そこで本手法では，ラストタッチ判定に必要な時間的情報のみを抽出し，図 7 に示すように連続的な映像を離散的な静止画シーケンスへと変換する処理を行う．

3.3.1　抽出フレーム枚数の設定

ユーザーによって入力されたタイムスタンプt（ラストタッチの瞬間）を基準として，前後 5 フレームを含む計 11 フレームを抽出対象とした．このフレーム数$(t \pm 5)$の設定には，以下の 3 つの理由がある．

1． 物理的接触の推移と網羅:
サッカーにおいて，足がボールに触れてから離れるまでの接触時間は極めて短く（通常数ミリから数十ミリ），30fps や 60fps の映像においては 1 から 3 フレーム程度に収まることが多い．しかし，ラストタッチの判定には接触の瞬間だけではなく「どちらのチームの選手の足がボールに向かっているのか」および「接触後にボールと足がどう離れたのか」という前後因果関係が不可欠である．よって前後 5 フレームを確保することで，接触前，接触中，接触後の 3 つのフェーズを確実にとらえ，かつ無関係なプレーを含まない最小限のフレーム数に設定した．

2． LVLM の認識能力とトークン効率の最適化：
LVLM は，入力情報が増えるほど「特定の微細な変化（接触の瞬間）」に対する認識精度が低下し，全体的なプレーの流れ（パスやドリブル）の記述に終始する特性を持つ．11 フレームという枚数は，モデルが各フレームの細部に集中しつつ，時間的な変化を論理的に追跡できる最適な情報量である．

3． ユーザ入力誤差へのロバスト性の確保:
本手法はユーザが手動でタイムスタンプを指定することを前提としている．指定したtに数フレーム程度の微細な誤差が生じた場合でも，前後 5 フレームのバッファがあることで，抽出範囲内に真の接触瞬間を確実に含めることができ，安定性を担保できる．

3.3.2 抽出フレームの詳細

抽出された 11 フレームは，MP4 等の動画形式ではなく，独立した 11 枚の静止画像（PNG）として LVLM に入力する設計とした．一般的な動画圧縮（フレーム間予測等）は，データ量削減と引き換えに，高速移動する物体に対してブラーやブロックノイズを生じさせやすい．特にボールと足が交錯する瞬間のような微細な領域において，これらのアーティファクトは判定精度を著しく低下させる要因となる．そのため本手法では，各フレームを独立した高解像度静止画として扱うことで圧縮影響を排除し，微細な特徴を鮮明に保持する設計とした．

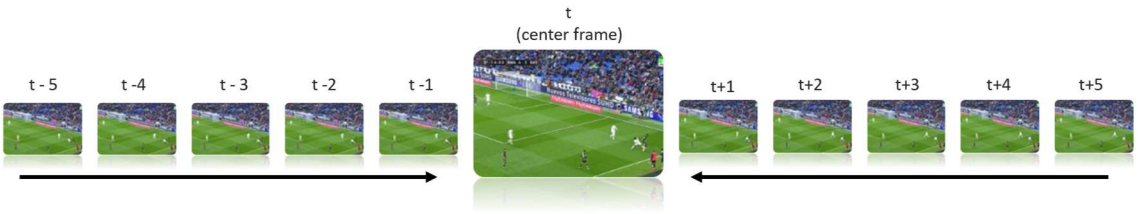


図 7. タイムスタンプ t を中心とした t±5 の計 11 フレーム

3.4 拡大映像の詳細検出と追跡

3.2 節で抽出された拡大映像に対して，画像内のオブジェクト検出および追跡をする処理を行う．

3.4.1 拡大領域における再検出

各フレームに対して，3.2 節と同様に拡大映像をファインチューニングした YOLOv11 モデルを適用し，ボールおよび選手の位置座標（BBox）を取得する．図 8 に，全体映像での検出結果（左）と，切り出した拡大映像領域での再検出結果（右）の比較を示す．

この段階での入力画像は，ボール周辺を 320×320 ピクセルで切り出した高解像度領域である．図 8（左）のような全体映像では，ボールのピクセルサイズが極めて小さく検出信頼度が不安定になりやすいが，図 8（右）のように拡大映像で再検出を行うことで，相対的なオブジェクトサイズが大きく保持され，特定精度が大幅に向上する．これにより，照明の変化や高速移動に伴うボールの変形に対しても，より堅牢な検出が可能となる．

3.4.2 IoU に基づく簡易追跡

各フレームに対して独立して検出されたオブジェクトを時系列に沿って同一の個体として紐づけるために，隣接するフレーム間における BBox の重なり度合い（IoU）を用いた簡易トラッキングアルゴリズムを実装した．

IoU を計算し，最大かつ閾値（本実験では 0．3）を超えるペアを同一 ID として関連付ける．図 8（右）の各ボックス上部に表示されている「id:44100」等の固有 ID は，この処理によって付与されたものである．これにより，11 フレームという短いシーケンス内において，特定の選手（ID）とボールの動きを一貫した軌跡として管理する初期データ（Raw Tracks）を生成する．

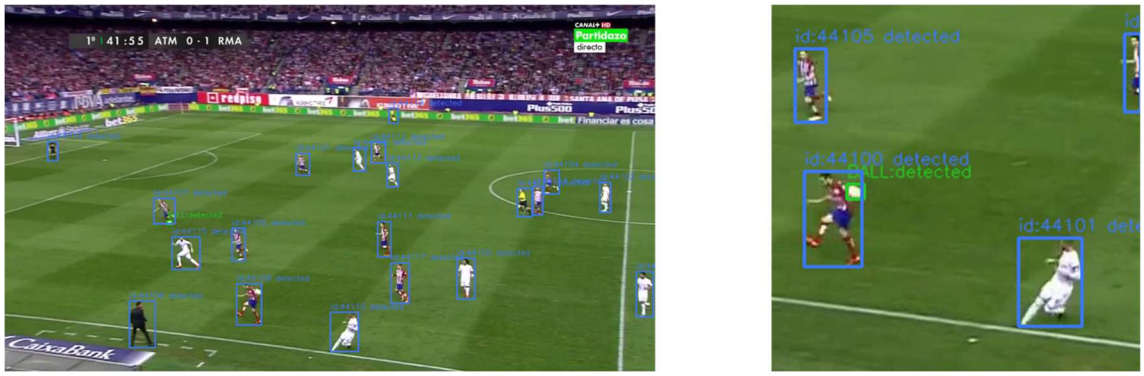


図 8．全体映像（左）と拡大映像（右）における検出精度の比較

全体映像（左）：ボールの占めるピクセルサイズが極めて小さく，背景ノイズとの区別が困難なため，検出信頼度が不安定になりやすい．（右）拡大映像：同一シーンをクロッピングにより拡大した結果．相対的なオブジェクトサイズが大きくなることで，ボールの形状や特徴が鮮明になり，安定した検出と ID 付与が可能となっている．

3.5　位置情報時系列データのノイズ除去・補正

前節で得られた初期の検出データには，映像特有の検出ブレや，選手とボールが重なることによる検出欠損（オクルージョン）が含まれる．図 9（上段）に示す通り，0 フレーム目の欠損や，座標の急激な変化による不連続な箇所が見受けられる．これらのノイズは，後段の LVLM による推論において誤った物理認識を引き起こす原因となるため，本手法では以下の多段階の補正処理を適用し，ボールが不自然に静止したりせず，慣性の法則や等速直線運動などの物理法則に矛盾しない高精度な軌道データを構築する．

3.5.1　欠損補完と光学的フローによる微修正

ボールの検出が途切れた区間に対して，欠損期間に応じた補完処理を行う．3 フレーム以下の短期期間の欠損に対しては，Lucas-Kanade 法を用いたオプティカルフロー（Optical Flow）[19]を適用する．直前のフレームの画像特徴点を追跡することで，単なる数値補完ではとらえきれない非線形的動きを画像情報から推定・補完する．一方でそれ以上の長期期間の欠損に対しては，前後の検出位置に基づく線形補完（Linear Interpolation）を適用し，大域的な軌道の連続性を担保する．

3.5.2　RTS スムーシングによる軌道推定

補完された軌道データに対し，観測ノイズを除去し，物体の運動モデルに基づいた滑らかな軌跡を得るためにカルマンフィルタを適用する．本手法ではボールの運動を等速直線モデルとして定義し，以下の時刻 k の状態ベクトル$\boldsymbol{x}$を推定する．

$$x_k = [x_k, y_k, v_{x,k}, v_{y,k}]^T$$

通常のカルマンフィルタは，開始フレームから順方向にのみ推定を行うため，急激な方向転換に対する追従に遅れ（タイムラグ）が生じやすい．そこで本手法では，順方向の処理に加え，終了フレームから逆方向への解析も行う RTS スムーシング（双方向スムーシング）を適用した．図 9（下段）に示す通り，前後両方の文脈を利用して軌道を最適化することで，欠損していた 0 フレーム目の座標を推定し，かつノイズを排した滑らかな軌道を得ることに成功した．前後両方の文脈を利用して軌道を最適化できるため，遅れを解消し，ラストタッチの瞬間における真の位置を正確に推定することを実現した．

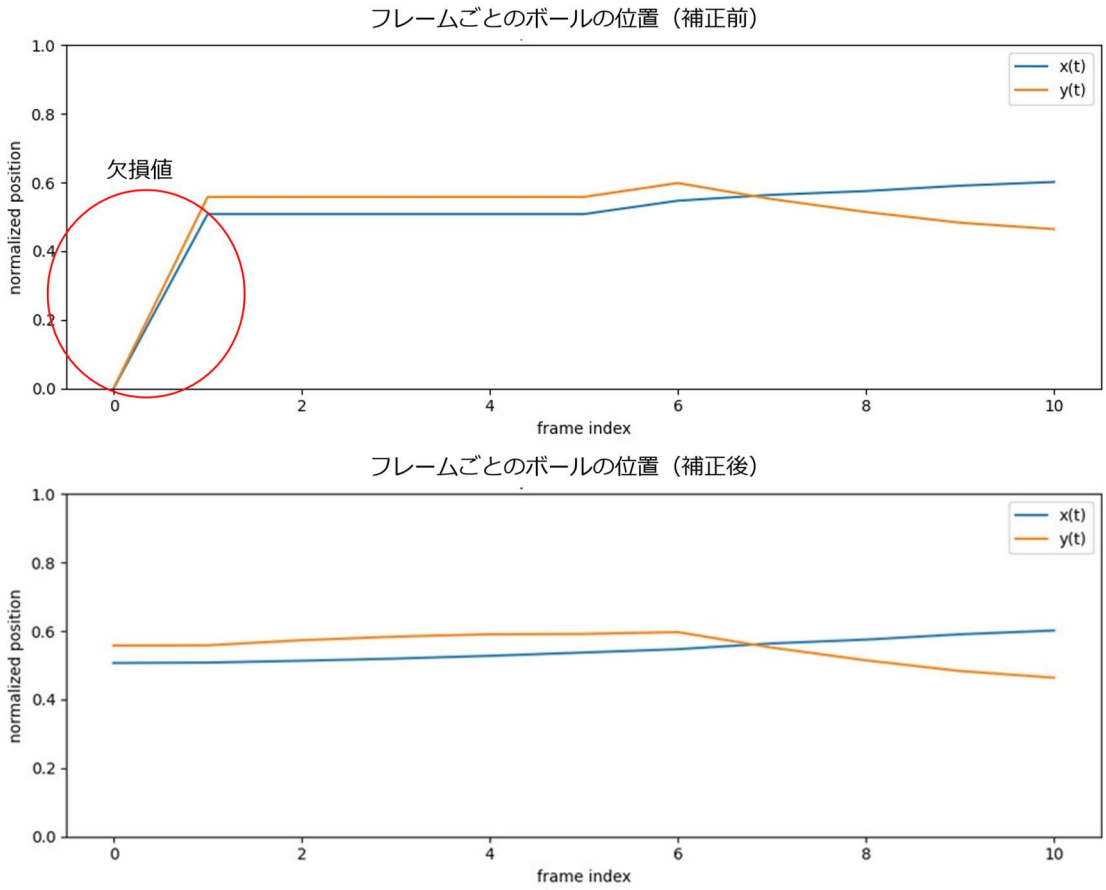


図 9. 位置情報データのノイズ除去・補正前後の比較（上段）補正前：YOLO によるボールの位置座標検出結果．フレーム 0 における未検出（欠損）により，ボール座標が(0，0)として処理され，急激な不連続点が生じている．（下段）補正後：RTS スムーシング適用後の軌道．前後フレームの文脈からフレーム 0 の座標を精度高く推定し，物理法則に即した滑らかな軌跡へと平準化されている．x(t)，y(t)は画像内におけるボールの正規化横座標を示す．

3.6　マルチモーダル推論

最終的な判定フェーズでは，生成された拡大映像（視覚情報）と位置情報時系列データ（物理情報）を統合し，マルチモーダル大規模言語モデル（LVLM）を用いた推論を行う．

3.6.1　入力データの構成

LVLM に入力する情報は，以下の 2 種類の形式で構成される．

1．視覚情報：

3.3 節で生成した 11 枚の拡大映像の静止画像シーケンス．これは選手の体の向きや足

の形状といった数値化が困難な情報をモデルに提供する．

2．物理情報（位置情報時系列データ）：
3.5 節で平準化された追跡データを構造化テキスト形式に変換したもの．ここには各フレームにおけるボールと選手の座標が含まれており，画像だけでは判断しにくい微細なボールの軌道の変化を客観的な数値として提供する役割を担う．

3．Home/Away のユニフォーム画像：
各チームの公式サイト等から取得したユニフォームの画像を入力情報として用いる．両チームのユニフォーム画像を提供することで，最終的にボールに接触した選手がどちらのチーム（Home または Away）に属しているかを，画像認識を通じて正確に判定するための視覚的コンテキストをモデルに提供する．

3.6.2　プロンプト作成

これらの入力データを適切に解釈させるために，以下の要素を含む構造化されたプロンプトを設計した．

1．役割の定義：
モデルに対して「プロのサッカー審判員（VAR 担当）」としての役割を与え，客観的かつ物理的な根拠に基づいた判定を要求する．

2．タスクの指示：
提供された 11 枚の静止画シーケンス，Home/Away のユニフォーム画像，位置情報時系列データをもとに「最後にボールに触れたチームを（Home または Away）」を特定するように指示する．

3．推論の制約：
単に結論を出すのではなく「フレーム T で（Home または Away）の選手がボールに接触し，この接触によりボールの軌道が変化した」といった具体的な根拠を記述させることでハルシネーション（幻覚）を抑制する．

3.6.3　推論モデル

推論エンジンには GPT-4o を採用した．本モデルは，長いコンテキストを扱えるため，複数の画像と詳細な位置情報時系列データを同時に処理することに適している．また，ゼロショット推論能力が高く，追加の学習を行わずに提示された情報のみから複雑な物理現象を論理的に推論することが可能である．

第 4 章　結果と考察

4.1　実験評価

4.1.1　データセット

評価実験には，大規模サッカー動画データセット SoccerNet-v2 を基盤として独自に構築したデータセットを用いた．本研究の目的に合致するシーンとして，286 件のイベントクリップを抽出した．SoccerNet-v2 には「ボールアウト」のイベントラベルが多数含まれているが，既存のラベルは「ボールが外に出た瞬間」を特定するのみであり，「どちらのチームが最後に触れたか」という属性情報が付与されていない．そこで本研究では，以下の 3 つの選定条件に基づき，全 286 件のシーンに対して独自にラストタッチの帰属チームに関するアノテーションを実施した．ラベル（Home/Away）は著者が付与し，判定が曖昧な事例については，サッカー経験 10 年以上を有する第三者 2 名と協議し，合意に基づいてラベルを確定した（合意が得られない場合は当該事例を除外した）．

1. 判定事象の多様性：
   ボールがタッチラインやゴールラインを越える直前の攻防など，「どちらのチームが最後にボールに触れたのか」の帰属判定が必要なシーンを選定した．対象とする事象は，一人の選手によるクリアのような「基本事象」から，複数の選手が密集している「複雑事象」まで，実用的な難易度の幅を広く網羅している．

2. 撮影条件の汎用性：
   実際の放送映像において最も一般的である，ピッチサイドからの映像を採用した．

3. 解析精度と動画品質:
   ラストタッチを行う選手が画面上で高さ 80 ピクセル以上のサイズで映っていることを条件とし，視認不可能な微小対象を除外した．入力動画の解像度は 1280×720 ピクセル，フレームレートは 25fps に統一し，ボールの軌道変化等の物理情報を精密に解析できる品質を確保した．

4.1.2　比較手法

提案手法の有効性を多角的に検証するために以下の比較実験を行った．

1. ベースライン：
   画像認識モデル（YOLOv11）の検出結果のみを用いたルールベースの手法

・判定ロジック：ボールと最もユークリッド距離が近い選手を「ラストタッチした選手」として判定する

・チーム分類：BBox 内の平均色情報を取得し，HSV 色空間における距離でチームを分類する．

なお，本研究のベースラインは，ボールと選手の単純な幾何（距離）および色情報のみに基づくルールベースであり，下限性能（lower bound）として設定している．

2. 視覚情報のみ：
   位置情報時系列データを使用せず，11 枚の静止画シーケンスとユニフォーム情報のみを LVLM に入力して推論を行う手法
   ・全体画像：切り抜きを行わず，全体画像を入力する．
   ・拡大画像：3.3 節の手法でボール周辺をクロップした画像を入力する．

3. 提案手法
   拡大画像に加え，平準化した位置情報時系列データをプロンプトに含めて推論を行う手法．推論モデルには GPT-4o を採用し，比較対象として Gemini-2.5Pro，Qwen2-VL についても同様の評価を行った．

4.2 定量評価

各手法における判定精度の評価結果を示す．各手法の判定性能を比較するため，判定結果が正解ラベルと一致する割合を Accuracy および Macro-F1 により評価した．本研究は最終判定の自動化を目的としないが，支援として機能するためには「誤ったチームを提示しない」ことが最低限必要であるため，まずは判定の一致度を定量的に確認する．

4.2.1 提案手法における構成要素の有効性

入力データの形式が判定精度に与える影響を検証するため，GPT-4o を用いたアブレーションスタディ（切除実験）の結果を表 2 に示す．なお，本項における「Improvement(%)」は，最も基本的な入力形式である「WideOnly」を基準とした精度の向上幅を示している．

手法名	VisualInput（視覚情報）	AuxiliaryInput（補助情報）	Accuracy（%）	Improvement(%)
WideOnly	Wide image	-	73.78	-
ZoomOnly	Zoom image	-	83.57	+9.79
Proposed	**Zoom image**	**Tracking data**	**87.06**	**+13.28**

表 2：アブレーション（入力構成）による精度比較（GPT-4o）

表 2 からわかるように入力情報をリッチにするにつれて着実に精度が向上しているのがわかる．

1. 拡大画像の有効性：
   全体画像から拡大画像へと切り替えることで約 9.8%の大幅な精度向上が見られた．これは高解像度のままボール周辺を切り出すことで，足元の接触やボールの軌道変化といった微細な特徴を LVLM が認識可能になったためである．

2. 位置情報時系列データの統合効果：
   視覚情報に加えて平準化された位置情報時系列データを追加することでさらに 3.5%の精度が向上し，最高精度の 87.06%を達成した．すでに高い画像認識能力を持つ GPT-4o であっても，視覚だけでは判断が難しい「オクルージョン」や「微細な接触」の判定においては数値データが重要な補完役割を果たしたといえる．

4.2.2 既存手法及び他モデルとの比較

次に，提案手法（拡大画像＋位置情報時系列データ）を適用した場合の各モデルの性能を比較し，ベースライン手法に対する優位性を検証する．本項の「Improvement(%)」は，ルールベースの「Base line」を基準としている．

Model	VisualInput（視覚情報）	AuxiliaryInput（補助情報）	Accuracy（%）	Macro-F1（%）	Improvement（%）
Base line	-	Tracking data	49.30	46.36	-
Qwen2-VL	Zoom image	Tracking data	55.24	43.13	＋5.94
Gemini-2.5-pro	Zoom image	Tracking data	76.57	76.55	＋27.27
GPT-4o(Ours)	**Zoom image**	**Tracking data**	**87.06**	**87.06**	**+37.76**

表 3：推論モデル別性能比較（Zoom+Tracking data）

表 3 に示す通り，提案手法を用いた GPT-4o は 87.06%と高い正解率を記録し全てのモデルを上回った．

1. Base line：
   単純な距離計測に基づくベースライン（49.30%）と比較して＋37.76%上回っている．これは，単なる座標の近接性だけでなく，AI による試合の文脈理解と物理的な因果推論の統合が，高精度な判定において不可欠であることを実証している．

2. 他モデル：
   Gemini-2.5pro も 76.57%と良好な結果を示したが GPT-4o はそれをさらに約 10%上回っている．これは複雑な位置情報時系列データを解釈し画像情報と論理的に結びつける「推論能力」において，GPT-4o が現時点で最も適していることを示唆している．また，GPT-4o は Accuracy と Macro-F1 がほぼ一致しており，Home/Away いずれかに偏った判定ではなく，両クラスでバランスよく高精度に判定できていることが示された．

4.3 定性評価

本節では，提案手法（GPT-4o+拡大画像＋位置情報時系列データ）が実際にどのようなプロセスを経て，正解を導き出したのかを分析する．成功事例として，「位置情報時系列データが決定打となったケース」,「位置情報時系列データが誤検出を防いだケース」を，失敗事例として「検出器由来のエラー」と「複雑な物理挙動によるエラー」の計 4 例を取り上げる．

4.3.1 成功事例：位置情報時系列データが決定打になったケース

図 10 は，画像情報のみでは誤判定が発生したが，提案手法によって正解できた事例である．本シーンでは，Home（白色のユニフォーム）チームと Away（黒色のユニフォーム）チームがボールを奪い合っている局面である．両選手がボールにきわめて接近しており，視覚的にはどちらの足がボールにインパクトを与えたのか判別が困難である．

1. 拡大画像のみ：
   画像のみを入力した手法では，視覚的な印象に引きずられ，以下のような間違った推論を行っている．以下はモデルが出力した推論理由である．

   「画像クリップでは，黒いユニフォームを着た Away チームの選手がボールに触れている様子が確認できます。白いユニフォームの Home チームの選手がボールに触れる前に, Away チームの選手がボールを蹴っているため, ラストタッチは Away チームです.」

これは典型的なハルシネーションであり，実際には接触していない Away 選手を，ボールの

軌道変化の原因であると誤認している．
確かに，近年の LVLM は高度な画像認識能力を持っており，映像のみでも判定可能であるとする見解もあるだろう．しかし，視覚情報のみに依存した場合，実際には接触していない Away 選手を軌道変化の原因と誤認するハルシネーションを引き起こし，誤った判定を下すこととなる

2. 提案手法（拡大画像＋位置情報時系列データ）：
   一方, 位置情報時系列データを用いた提案手法では Home チームによるラストタッチを正確に特定した．以下はモデルが出力した推論理由である．

   「フレーム 3 から 4 にかけて，Home チームの選手（白いユニフォーム）がボールに接触し，その後ボールの軌道が変化しました．特に，track_id 44103 の選手がボールに触れたと判断されます．Away チームの選手（黒いユニフォーム）はその後にボールに触れていないため，最後に触れたのは Home チームです．」

図 11 の速度グラフ（下段）を参照すると，黄色で強調したフレーム 3 から 6（Impact Zone）にかけてボールの速度が上昇し，明確なピークを迎えていることが確認できる．提案手法では，この「物理的な軌跡の急変」という客観的な事実を数値データから読み取り，そのタイミングでボールに近接していたのが Home 選手であることを視覚情報と照合して特定した．このように位置情報時系列データが「いつ接触が発生したのか」を確定させることで，視覚的な曖昧さを排除し，正しい判定へと導いた事例である．

図 10．成功事例 1：視覚的に曖昧な接触シーン（白：Home，黒/縞：Away）両チームの選手がボールに極めて接近しており，画像情報のみではどちらの選手が最後に接触したか，ハルシネーション（誤認）が発生しやすい状況を示している．

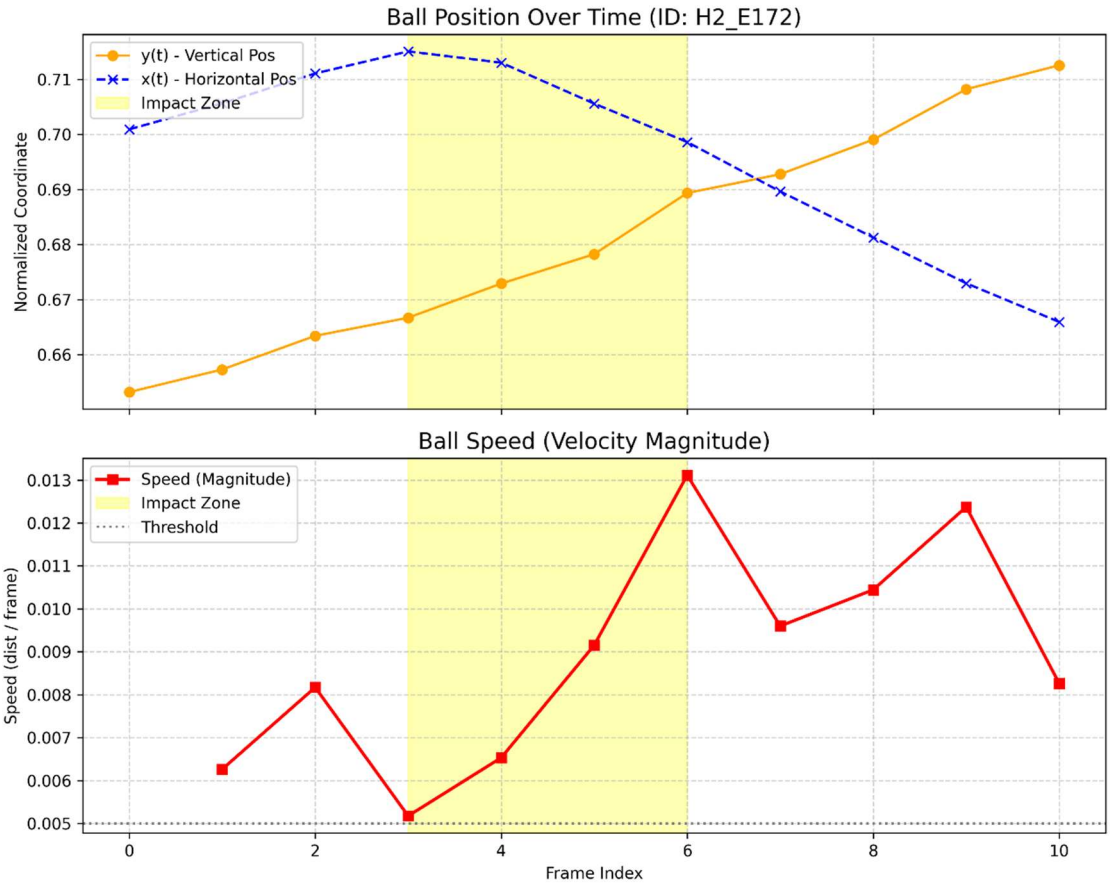


図 11．図 10 における各フレームのボールの軌道および速度変化の解析上段は正規化座標（x，y），下段は速度の変化を示す．黄色で示した Impact Zone（フレーム 3〜6）において明確な速度のピークが観測され，物理的な接触事象が発生したことを裏付けている．

4.3.2　成功事例：位置情報時系列データが誤検出を防いだケース

図 12 に画像のみを入力したモデルが誤判定を起こしたが，提案手法がそれを回避した正解した事例を示す．

本シーンは，Home チーム（赤紫のユニフォーム）の選手がボールをキックした直後，Away チーム（青白のユニフォーム）の選手がその軌道上で足を伸ばしてブロックを試みている場面である．視覚的には，Away 選手の足がボールの軌道と交差しており，あたかも「クリア（Kick）」や「ブロック」を行ったかのような強いアクションが見られる．しかし，実際にはボールに触れておらず（空振り/スルー），ボールはそのまま通過している．

1. 拡大画像のみ：

   画像のみのモデルは選手の動作という視覚的情報に過剰反応し以下のように誤判定した．以下はモデルが出力した推論理由である．

「動画クリップの最後の数フレームで，Away チームの選手がボールをクリアしようとする動作が見られ，その直後にボールの軌道が変化しているため，Away チームが最後にボールに触れたと判断しました.」

2. 提案手法（拡大画像＋位置情報時系列データ）：
一方，提案手法は以下のように推論し，Home チームによるラストタッチを正しく特定した．以下はモデルが出力した推論理由である．

「フレーム 5 で，Home チームのユニフォームを着た選手（track_id:44101）がボールに接触し，その後のフレーム 6 でボールの軌道が変化していることから，最後にボールに触れたのは Home チームの選手であると判断しました.」

ここで注目すべきは，提案手法が「Away 選手の派手な動き」というノイズを排除できている点である．これには，位置情報時系列データ上の物理情報が決定的な役割を果たしている．図 13 の下段（速度グラフ）を参照すると，黄色で示した Impact Zone 内のフレーム 5 から 6 にかけて，Home 選手のキックに伴う顕著な速度スパイクが確認できる．しかし，その後の Away 選手が足を伸ばしてボールに近接したタイミング（フレーム 8 から 10）では，有意な速度変化が観測されていない．提案手法はこの「物理的なイベントの不在（Absence of Event）」を検知することで，視覚的には接触に見えるシーンであっても，「物理的には触れていない」という客観的な結論を導き出すことに成功した．これは，物理情報が LVLM のハルシネーション（視覚的な思い込み）を訂正するための強力なフィルタリング効果を持つことを示している．

図 12. 成功事例 2：視覚的に接触と誤認しやすい「空振り」のシーン（図 12 右）Away 選手（青白）がボールの軌道上でブロックを試みる強い動作を見せているが，実際には接触が発生していない事例. 画像情報のみに依存した推論では，この動作をラストタッチと誤認するリスクが高い.

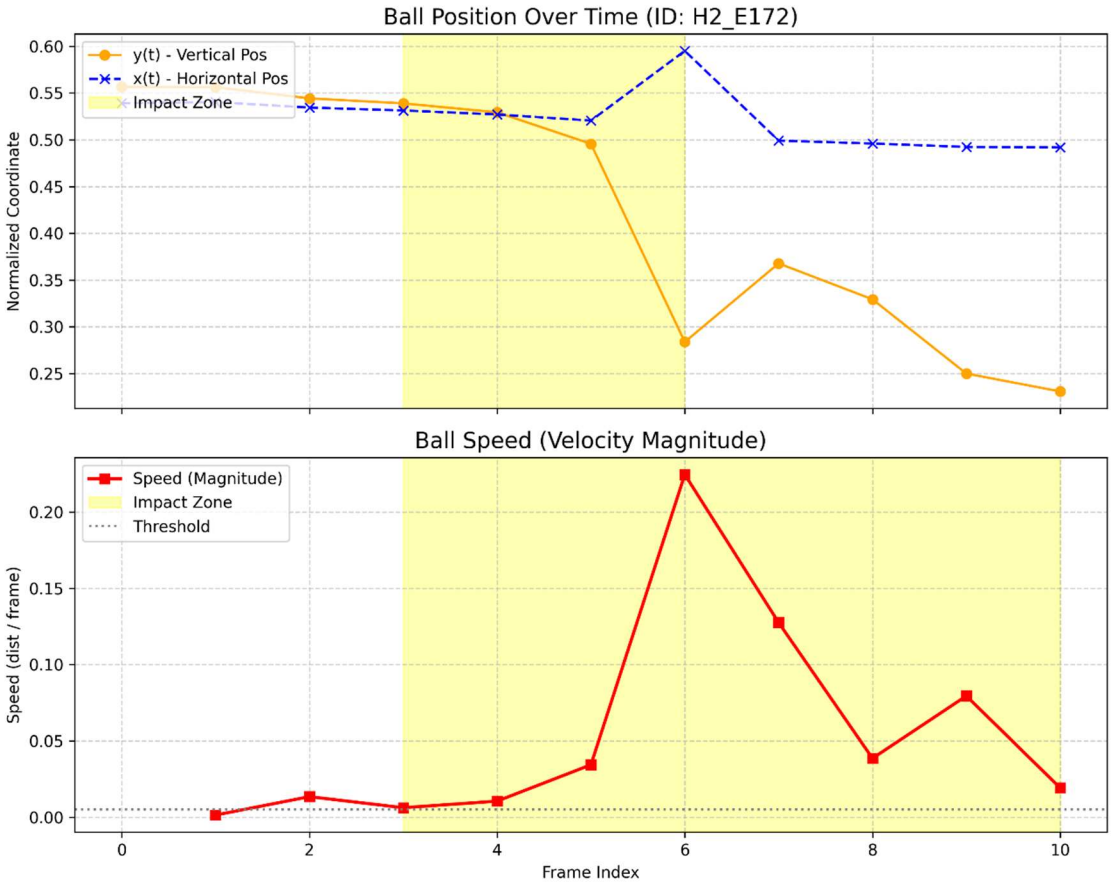


図 13. 図 12 におけるボールの物理データ解析上段は正規化座標 x(t)，y(t)の推移，下段は速度の変化を示す．フレーム 5 から 6 において Home 選手の接触に伴う速度スパイクが確認できる一方，その後の Away 選手の接近時には有意な物理変化が見られない．この「物理的イベントの不在」が，誤判定を回避する決定的な根拠となっている.

4.3.3　失敗事例：検出器によるエラー

検出器由来のエラーの例を図 14 に示す．これは，重度のオクルージョン（遮蔽）や照明条件の不備により，上流の物体検出モデル（YOLOv11）がボールの検出に失敗した事例である．本手法は，高精度な位置情報時系列データが入力されることを前提としている．そのため，ボールが数十フレーム以上にわたって未検出となったり，選手のスパイクなどをボールと誤認識したりした場合，LVLM に入力される位置情報時系列データ自体が大きなノイズとなる．この場合，LVLM がいかに高度な推論能力を有していても，誤った物理情報に基づいて誤判定を引き起こす結果となった．これは，各工程の精度が後続の処理に直結するパイプライン型手法の限界であり，検出モデルのさらなる堅牢性向上が不可欠であることを示唆している．

図 14．失敗事例：重度のオクルージョンによるボールの検出失敗 選手とボールが完全に重なり，物体検出モデルがボールの位置を特定できなかった事例．物理情報の欠損により，後段の LVLM へ正確なコンテキストを伝達できず，判定エラーを招く要因となる．

4.3.4　失敗事例：複雑な事象

図 15 は，Away（青白のユニフォーム）選手がボールをトラップしようとした瞬間，Home（赤紫のユニフォーム）選手が背後からアプローチし，双方がほぼ同時にボールに接触した事例を示す．本シーンでは，物理的解析と視覚的解析の両面での限界が見られた．

1. 物理的な解析の限界:
   通常，接触プレーは「ボールの軌道が変わる」ことで検知される．しかし本シーンでは，両選手が「進行方向と同じ向き」にボールを扱っており，かつ接触のタイミングが重なっている．その結果，位置情報時系列データ上で速度の変化（加速）が観測されたとしても，その物理的な力が「Away 選手のトラップによるもの」なのか，「Home 選手のプッシュによるもの」なのか，あるいは「両者の合成力」なのかを，速度ベクトルの向きだけでは分解することが不可能であった．

2. 視覚的な解析の限界:

視覚的にも両選手の足がボール位置で重なっており（図 15 中央），LVLM は「どちらの足が，最後の瞬間にボールに触れていたか」を画素情報から判別する術を持たなかった．このように，物理的情報（ベクトル）と視覚的情報（重なり）の双方が「どちらとも取れる」状態で拮抗した場合，提案手法であっても判定を誤る，あるいは確信度の低い推論しか行えないという限界が示された．

図 15．失敗事例：物理的・視覚的情報の双方が拮抗した複雑な接触事象 両選手がほぼ同時に，かつ同一方向へボールを扱っているシーン．位置情報時系列データ（速度ベクトル）による力の分離が困難であり，かつ画像上でも足元の重なりが激しいため，マルチモーダル推論を用いても判定の曖昧さを排除しきれない限界事例を示している．

第５章　まとめと今後の展望

本研究では，サッカーの試合映像における「ラストタッチ判定」という，審判にとって極めて認知的負荷の高いタスクを支援するために，物体検出モデルから得られる位置情報時系列データとマルチモーダル大規模言語モデル（LVLM）による文脈理解を統合した，新たなマルチモーダル推論手法を提案した．実験の結果，提案手法（GPT-4o+拡大画像+物理情報）は，既存のルールベース手法（49.30%）や画像情報のみを用いた LVLM（83.57%）を大きく上回る，87.06%の正解率を達成した．本手法の主要な成果は，物理情報を入力として統合することで，視覚的に曖昧な接触に対しては速度変化の検知により見落としを抑制し，一方で「空振り」のような場面では物理的イベントの不在を根拠に視覚的誤認（ハルシネーション）を抑制できる可能性を示した点にある．これにより，単一視点の汎用的な映像のみを用いても，審判員が自身の判断と照合可能な根拠を得ながら意思決定を行うという，根拠提示型の判定支援が成立し得ることを示した．

しかし，本手法にはいくつかの課題も残されている．第一に，システム全体の性能が上流の物体検出モデル（YOLOv11）の精度に依存する点である．重度のオクルージョンが発生し入力データが欠損した場合には，後段の LVLM による推論が困難となる．第二に，2 次元の位置情報時系列データに基づく物理的手がかりには限界があり，選手が密集し同方向へ同時に力が加わるような複雑な事象では，速度（スカラー量）の変化のみから因果関係を十分に特定できない場合が見られた．第三に，本研究ではラストタッチ近傍の時間区間を手動で指定して評価を行っており，実運用に向けてはボールアウト（out of play）検出を前段に導入し，タイムスタンプ入力を自動化する必要がある．また，判定支援としての有効性を厳密に示すためには，判断時間や確信度，判定の一貫性といった人間中心指標による評価が今後必要である．

今後の展望として，まずは判定精度のさらなる向上に向けて 3 次元姿勢推定（Pose Estimation）の導入を検討する．選手の四肢の 3 次元的な位置関係を解析に組み込むことで，ボールと身体部位の接触判定をより厳密に行い，現在の 2 次元解析では限界があった混戦状況への対応を可能にしたい．一方で，実用化に向けた重要な課題としてリアルタイム性の確立が挙げられる．高精度な 3 次元姿勢推定や LVLM の利用は計算コスト増大と処理時間延長のトレードオフを伴うため，軽量モデルの採用や推論アルゴリズムの効率化により，「精度とリアルタイム性の両立」を目指した最適化が必要である．将来的には，本手法を発展させ，アマチュアや育成年代の現場においても特別な設備なしで利用可能な判定支援システムの構築を目指す．

謝辞

本研究及び論文の作成にあたり，研究の着想や論文執筆等，多くのご指導，ご助言を頂きました静岡大学工学部の岡部誠准教授に心から感謝申し上げます．また，ご助力頂いた修士課程学生及び学部生の皆様に深く感謝致します．

参考文献

[1]FIFA, "FIFA Big Count 2006: 270 million people active in football," 2006.

[2] A. Deliège, A. Cioppa, S. Giancola, et al., "SoccerNet-v2: A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021.

[3] N. E. Brick, G. Breslin, M. Shevlin, and S. Shannon, "The impact of verbal and physical abuse on distress, mental health, and intentions to quit in sports officials," Psychology of Sport and Exercise, Vol. 63, 102274, 2022.

[4] G. P. McEwan, V. B. Unnithan, C. Easton, A. J. Glover, and R. Arthur, "Decision-making accuracy of soccer referees in relation to markers of internal and external load," European Journal of Sport Science, Vol. 24, No. 6, pp. 783-792, 2024.

[5] FIFA, "VAR Light concept taking shape," Inside FIFA, [https://inside.fifa.com/innovation/standards/video-assistant-referee/video-assistant-referee-technology], (参照 2026-01-19)

[6] J. Held, H. Ackermann, H. Blume, and M. Rosenhahn, "VARS: Video Assistant Referee System for Automated Soccer Decision Making," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023.

[7] B. T. Naik, M. F. Hashmi, and N. D. Bokde, "A Comprehensive Review of Computer Vision in Sports: Open Issues, Future Trends and Research Directions," Applied Sciences, Vol. 12, No. 9, 4429, 2022.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, 2016.

[9] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," arXiv preprint arXiv:2410.17725, 2024.

[10] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple Online and Realtime Tracking," In Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 3464-3468, 2016.

[11] N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," In Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 3645-3649, 2017.

[12] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," In Proceedings of the European Conference on Computer Vision (ECCV), 2022.

[13] B. S. Bal and G. Dureja, "Hawk Eye: A Logical Innovative Technology Use in Sports for Effective Decision Making," Journal of Sport and Health Research, Vol. 4, No. 1, pp. 13-20, 2012.

[14] S. Giancola, et al., "SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos," CVPR Workshops, 2018.

[15] J. Held, H. Itani, A. Cioppa, S. Giancola, B. Ghanem, and M. Van Droogenbroeck, "X-VARS: Introducing Explainability in Football Refereeing with Multi-Modal Large Language Models," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2024.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), pp. 5998–6008, 2017.

[17] P. Wang, S. Bai, S. Tan, et al., "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," arXiv preprint arXiv:2409.12191, 2024.

[18] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," AIAA Journal, Vol. 3, No. 8, pp. 1445-1450, 1965.

[19] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," In Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI), pp. 674-679, 1981.

[20] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," Journal of Basic Engineering, Vol. 82, No. 1, pp. 35-45, 1960.

[21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," In Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS), 2022.

付録

これは 3．6 節で使用したプロンプトテンプレートである．

```
{
  "prompt_configuration": {
    "system_message": {
      "role": "system",
      "description": "レフェリーとしての役割定義",
      "content": "あなたは，サッカーのビデオ・アシスタント・レフェリー（VAR）です．
あなたの任務は，提供された証拠を分析し，ボールに最後に触れたチーム（ラストタッチ）
を正確に判定することです．"
    },
    "user_message": {
      "role": "user",
      "description": "証拠資料の提示と推論ステップの指示",
      "content_structure": [
        {
          "section": "証拠資料",
          "details": [
            {
              "title": "1.　コンテキスト（チーム情報）",
              "description": "Home/Away 両チームのユニフォーム画像"
            },
            {
              "title": "2.　Zoom 動画クリップ",
              "description": "時系列順に並んだ 11 枚の連続静止画像シーケンス"
            },
            {
              "title": "3.　検出・追跡データ（JSON）",
              "description": "各フレームでのオブジェクト座標とステータス
（detected/smoothed）"
            }
          ]
        },
        {
          "section": "判定指示（推論ステップ）",
```

```
            "steps": [
              "1.  チーム分類: ユニフォーム画像と映像内の見た目を比較し，track_id の帰
属チームを特定する．",
              "2.  ラストタッチ判定: JSON の数値データ（速度・軌道）と視覚情報を照合
し，接触の瞬間を特定する．",
              "3.  最終判定: 上記の結果からラストタッチチームを決定する．"
            ],
            "logic_constraints": [
              "物理法則: ボールの軌道や速度が不自然に変化した瞬間を接触とみなす．",
              "禁止事項: 空間的な近接性（単にボールに近いだけ）で判定してはならない．
"
            ]
          },
          {
            "section": "回答フォーマット",
            "format": "JSON",
            "fields": {
              "last_touch_team": "Home または Away",
              "reason": "判定の根拠（フレーム，イベント，track_id）の詳細な説明"
            }
          }
        ]
      }
    }
}
```