arXiv:2508.01802v2 [cs.CV] 5 Oct 2026

SoccerTrack v2: A Full-Pitch Panoramic Video Dataset for Game State Reconstruction and Ball Action Spotting

Atom Scott [1*], Ikuma Uchida [2], Kento Kuroda [2], Yufi Kim[2], Keisuke Fujii [1]

[1*]Nagoya University, Japan.
[2]University of Tsukuba, Japan.

*Corresponding author(s). E-mail(s): atom.james.scott@gmail.com;

Abstract

Soccer analytics draws on two kinds of information: spatio-temporal data describing where players and the ball are, and event data describing what they do. Public datasets offer them apart, or together only on broadcast footage that leaves players outside the frame unobserved. SoccerTrack v2 combines continuous full-pitch video, long player trajectories and actor-linked events in one resource: ten university-level matches, 932 minutes of fixed-camera 4K panoramic video, annotated per frame with metric pitch coordinates, jersey numbers and persistent identities, roles and team sides for all players, and with ball action events in twelve classes, linked to the acting players through the same identifiers used in the trajectories. We fix a match-level split and report baselines for two tasks. For game state reconstruction, we run a full pipeline over all twenty halves and find that GS-HOTA scores degrade as sequence length increases. For ball action spotting, we train a model on the player trajectories, with and without the ball track. The data, the split and the evaluation tooling are released so that both tasks can be developed and compared at match length on the same footage.

Keywords: Soccer analytics, Multi-object tracking, Game state reconstruction, Ball action spotting, Dataset

1 Introduction

Most quantities in soccer analytics are defined over positions and actions together. Off-ball movement is valued relative to where the ball is and where it can next be played [1, 2], and the value of a possession asks how a spatial configuration changed as a consequence of a specific action [3, 4]. Positions come from tracking data and actions from event data, and both are in principle recoverable from video, for which the vision stack has matured: detectors and multi-object trackers assign consistent boxes to players [5–7], pitch line and keypoint estimators project those boxes into metric coordinates [8–10], appearance models and jersey number readers attach durable identities [11–16], and video backbones with temporal heads localise actions in time [17–21]. Game state reconstruction unifies the geometric half of that pipeline into one task and one metric [22, 23].

Video plays two roles in this programme. It is the input from which such systems must recover positions and events, and it is the reference against which their output, and the annotations used to train and score them, can be inspected: a position or an event released together with the footage it describes can be rewound and looked at, whereas the same record released without footage supports statistical modelling but not inspection. How much of the scene the footage shows, and for how long, therefore shapes what a dataset can be used for.

Existing public datasets each supply some of these ingredients. The largest are broadcast-based. SoccerNet releases full broadcast matches with action timestamps [24–26], automatically estimated camera calibration and positions of the visible persons [27], and image-space player tracking on 30-second clips, with one 45-minute half held out for its challenge [28]; SoccerNet-GSR defines game state reconstruction on 30-second broadcast clips [22]; SportsMOT tracks players in broadcast clips of three sports [29]; MultiSports localises the actions of individual players in broadcast clips of four sports, including football [30]; SoccerNet-GAR pairs broadcast event clips with all-player tracking and group-activity labels [31]; and FOOTPASS aligns 54 full broadcast matches with player-linked play-by-play events and all-player game state [32]. SoccerNet's ball action spotting task has since added attribution to the acting team [33] and, on FOOTPASS, to the acting player [34]. In all of these the camera follows play, so players outside the frame are either unannotated or have their state reconstructed rather than observed (FOOTPASS reports 9 to 12 of the 22 players visible on average and imputes the remainder), and the tracking benchmarks among them (SoccerNet tracking, SoccerNet-GSR, SportsMOT and SoccerNet-GAR) are defined on clips rather than matches. Tracking and event data are also released without video: Bassek et al. [35] synchronise the optically tracked positions of every player with hierarchically structured events across seven matches of the first and second German divisions but cannot release the footage for licensing reasons; data providers publish samples of the same kind, broadcast-derived tracking with events for all 64 matches of the 2022 World Cup [36] and for twenty A-League matches [37], and tracking with events for three anonymised matches [38]; and large public event collections carry event locations but no player tracking or video [39]. Instrumented full-pitch captures pair imagery with a separate measurement: the Alfheim dataset combines a stitched panorama with body-worn sensors for the home team [40], and WorldPose recovers

3D poses from a static multi-view rig and releases broadcast clips [41]; neither carries events. Earlier full-pitch video releases, including our own SoccerTrack v1 and TeamTrack [42, 43], provide positions on short clips without events, jersey numbers or roles, and synthetic generators render video, trajectories and events together for simulated rather than recorded play [44]. Section 2.2 tabulates the soccer releases among these.

SoccerTrack v2 adds the combination: continuous full-pitch video, long player trajectories and actor-linked events in one resource, at match length. It comprises ten university-level matches, 932 minutes of 4K panoramic video recorded by fixed camera systems whose field of view spans the entire field of play and its surroundings, so that players are not lost to camera motion as play moves. Each frame carries game state labels for the players on the pitch: metric pitch coordinates on a canonical 105×68 m pitch, jersey numbers and a persistent player identifier, role, and team side, with the median track annotated for 45 continuous minutes. The same footage carries ball action labels over twelve event classes, and each event names the acting team and, where the actor is identifiable, the actor in the *same* identity space as the tracks, so that an event resolves to a specific tracked player at a metric pitch location. A curated bounding-box subset supports multi-object tracking and was released with the SoccerTrack Challenge 2025 [45].

The combination supports questions its parts do not. Because identities are annotated for whole halves, a reconstruction method can be evaluated on whether it holds identity over 45 minutes rather than 30 seconds, and the effect of sequence length on its accuracy can be measured on identical footage. Because events and state are synchronised, how much of the event layer is determined by the configuration of players and how much by the ball, and how a reconstruction system and an event spotter compose from video to actions, can be measured on the same matches; because events name their actors in the trajectory identity space, actor attribution can be scored against a tracked player. And because the field of view covers the pitch, a predicted or annotated position can be inspected against the footage at the instant it is claimed, which is what failure analysis and label checking require.

We make two contributions:

- **Dataset.** Ten matches, two 45-minute halves each, of full-pitch panoramic video with per-frame game state for all players and ball actions linked to the players who performed them, released with a fixed match-level split, loaders and evaluation tools: one resource carrying a game state reconstruction benchmark and a ball action spotting benchmark on the same full-pitch footage.
- **Baselines.** Reference systems for both tasks: a game state reconstruction pipeline for fixed panoramic footage, run end to end over all twenty 45-minute halves, with a sequence-length study that rescores every half over growing windows and reruns one half on nested prefixes; and a ball action spotting model trained on the supplied trajectories and evaluated by cross-match cross-validation. In every half the reconstruction baseline's full-length score lies below its 30-second score, evidence that clip-scale evaluation can overstate match-length accuracy.

2 Results

2.1 The SoccerTrack v2 dataset

SoccerTrack v2 comprises ten university-level matches, each released as two 45-minute periods, referred to below as halves, recorded by fixed panoramic camera systems whose field of view spans the entire pitch. Play is not lost to camera motion as it moves, as it is when a broadcast camera pans. The corpus totals 1.4 million annotated frames, 932 min of play at 25 fps, and 30.7 million annotated player-frame records, one record per player annotated in a frame, about 22 per frame. Recording conditions vary across venues, weather and lighting, including night fixtures (Fig. 1); Table 1 gives the duration and event count of each match.

Each frame carries game state labels: metric pitch coordinates, jersey numbers and persistent player identities, roles, and team sides. The same footage carries ball actions over twelve classes, 21,428 events, and 99.5% of them name the acting player in the identity space of the tracks, so that an event resolves to a tracked player at a pitch location (Fig. 2). This synchronisation of the two layers is what the trajectory-based spotter of Section 2.4 consumes. Fig. 3 summarises both layers: the class distribution spans a factor of 300, the positions cover the full pitch, and identities are held for whole halves, the median track persisting for 45 minutes where existing game state benchmarks are defined on 30-second clips [22].

Table 1: The ten SoccerTrack v2 matches, labelled M1 to M10 throughout the paper. *Duration* is the annotated play time of each match at 25 fps. Every frame carries game state annotations for all players on the pitch: metric pitch coordinates on a canonical 105×68 m system, jersey number and persistent player identity, role, and team side. *Events* counts the twelve-class ball actions. A curated tracking subset, one four-minute clip per match with every player boxed in every frame (59,842 frames and 1.3 million boxes in total), is described in Section 4.5. $*$ marks the two matches whose recorded ball track is clamped to the pitch rectangle and therefore never leaves play.

Match	Duration (min)	Events
M1	90	2,085
M2	92	2,252
M3	97	2,117
M4	97	2,129
M5	97	1,945
M6	94	2,106
M7	92	1,932
M8	91	2,144
M9*	91	2,440
M10*	91	2,278
Total	932	21,428



(a) Daylight fixture.



(b) Daylight fixture at a second venue.



(c) Night fixture under floodlights.

Fig. 1: Example panoramic frames from SoccerTrack v2: the released field of view spans the pitch across venues and lighting conditions.

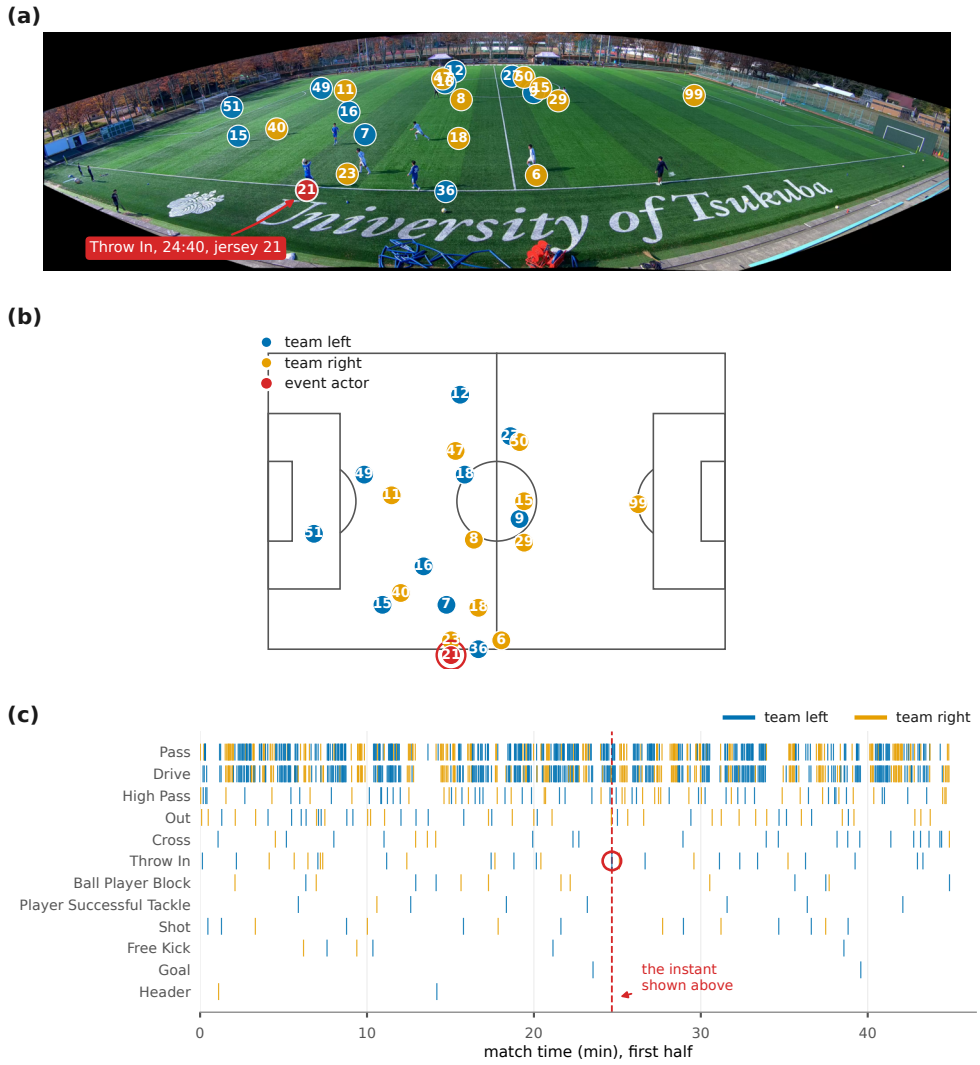


Fig. 2: One match at a glance: M2, first half. (a) A frame of the released panoramic video at the instant of a *Throw In* event, 24:40 into the half: all twenty-two players are annotated, drawn as markers carrying their jersey numbers, with the acting player named by the event (jersey 21) in red. (b) The game state at the same frame, drawn from the game state annotations: all twenty-two players in metric pitch coordinates, coloured by team side, the event's actor at $(-10.5, 35.4)$ m, beyond the touchline to take the throw. (c) The ball action layer of the same half: all 1,138 events over 45 continuous minutes, one row per class, coloured by acting team, with the pictured instant marked. The events name their actors in the same identity space as the tracks, so the three views resolve to the same players at the same instants.

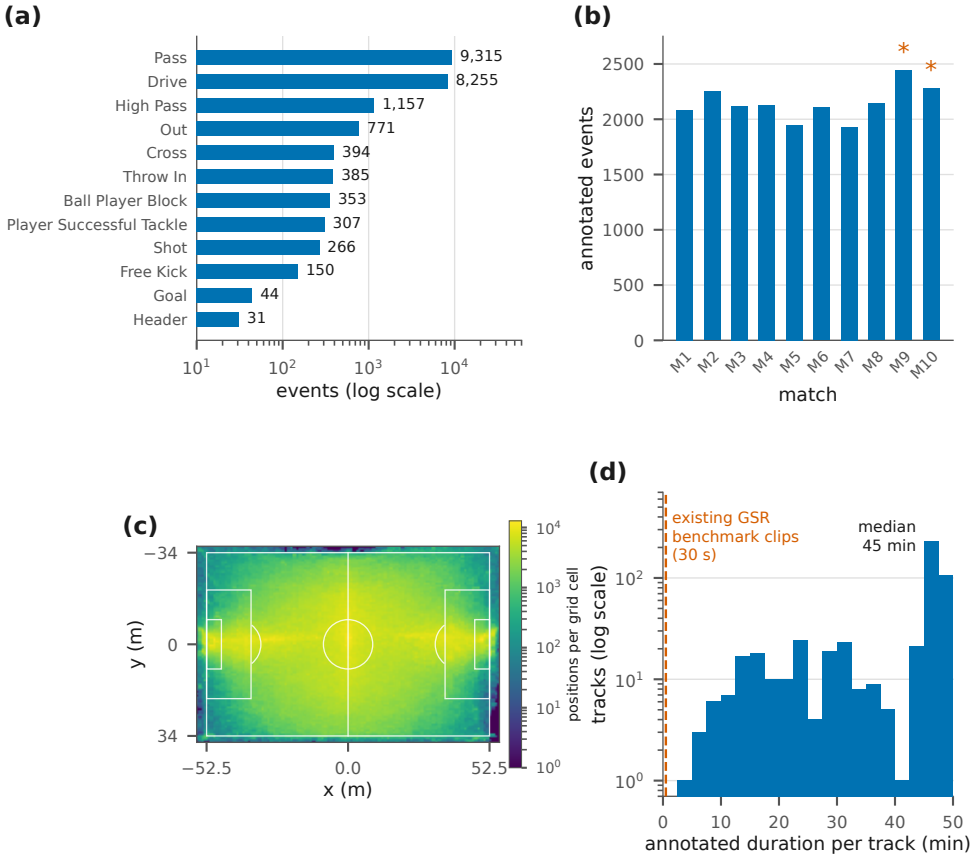


Fig. 3: Dataset statistics. (a) Ball action class distribution over the 21,428-event benchmark (log scale): the two most frequent classes account for 82% of events and the support spans a factor of 300. (b) Ball action events per match; asterisks mark the two matches whose recorded ball track is clamped to the pitch rectangle (Table 1). (c) Density of all 30.7 million annotated player positions of the game state layer in metric pitch coordinates, pooled over the ten matches (log colour scale). Positions are annotated on a 100×100 grid of the pitch, so each histogram cell measures 1.05×0.68 m; 0.2% of positions fall outside the displayed margin. (d) How long identities stay annotated. Each of the 522 ground-truth tracks follows one player through one half; the histogram counts tracks by their annotated duration (log count scale). The median track is annotated for 45 continuous minutes, while the dashed line marks the 30-second clips on which existing game state benchmarks are defined [22].

2.2 Positioning against existing datasets

Table 2 places the release beside the datasets of Section 1, grouped by the relation between the released video and the game state. The broadcast-based group releases video and, in FOOTPASS, actor-linked events with all-player state, but the state of players outside the frame is imputed rather than observed: FOOTPASS reports 9 to 12 of 22 players visible on average [32], SoccerNet-GSR annotates the visible persons of 30-second clips [22], and SoccerNet-GAR refines all-player tracking from 4.5-second event clips and labels group activities rather than actors [31]. The instrumented group observes every player with a measurement separate from the released camera: Bassek et al. [35] withhold the video, the Alfheim dataset [40] releases a full-pitch panorama with sensor positions for the home team only and no events, and WorldPose [41] releases poses and broadcast clips without events. Providers' tracking and event samples [36–38] and event feeds [39] release no video, and the broadcast-derived samples estimate the players off screen, as FOOTPASS does; SoccER [44] renders the scene it annotates. Among full-pitch video releases, SoccerTrack v1 [42] and TeamTrack [43] provide positions on clips without events. SoccerTrack v2 is the entry that carries match-length full-pitch video, all-player state within the released field of view and actor-linked events together.

Table 2: Positioning of SoccerTrack v2. The first four groups are defined by where the game state comes from relative to the released video: inferred from the released broadcast, measured by an instrument separate from the released camera, released without video, or rendered. The final group is defined by footage instead: releases of full-pitch video, with this release last; whether its provider measured the positions on the released view is not documented (Section 4.1). *Events* counts released event annotations and states whether each names the acting player. *Length* is as each release reports it.

Dataset	Video released	Pitch state	Events	Length
SoccerNet [24, 25, 27, 28]	broadcast, full matches	estimated positions of visible persons, no identities; image-space boxes on 30 s clips	110,458, not actor-linked	500 matches, 764 h
SoccerNet-GSR [22]	broadcast, 30 s clips	visible persons only	none	200 clips, 100 min
SportsMOT [29]	broadcast clips, three sports	no	none	240 clips, 150 k frames
SoccerNet-GAR [31]	broadcast, 4.5 s event clips	all 22 and ball, refined from broadcast	87,939, group-level	64 matches, event windows only
FOOTPASS [32]	broadcast, full matches (NDA)	all 22, partly imputed (9 to 12 visible)	102,992, actor-linked	54 matches
Bassek et al. [35]	withheld	all 22, observed at 25 Hz	11,137, actor-linked	7 matches
Alfheim [40]	full-pitch panorama	home team only, sensors at 20 Hz	none	3 matches, ∼220 min
WorldPose [41]	broadcast clips	3D poses, visible players	none	88 clips, 49.7 min
Provider samples [36–38]	no	all 22 and ball; broadcast-derived with off-screen players estimated in [36, 37]	actor-linked, counts not reported	64, 20 and 3 matches
Event feeds [39]	no	event locations only	3.25 M, actor-linked	1,941 matches
SoccER [44]	rendered, full matches	all 22, simulated	simulated, actor-linked	8 matches, 500 min
SoccerTrack v1 [42]	full-pitch fish-eye and drone, 30 s clips	all 22, GNSS at 10 Hz	none	63 min of clips
TeamTrack [43]	full-pitch fish-eye and drone, three sports	projected from boxes	none	156 min, 280 k frames
SoccerTrack v2 (ours)	**full-pitch panorama, full matches**	**all players on the pitch, within the released field of view**	**21,428, actor-linked (99.5%)**	**10 matches, 932 min, 1.4 M frames**

2.3 Game state reconstruction baseline

The reconstruction baseline is a game state reconstruction (GSR) pipeline assembled from published components for this footage (Section 4.7): its calibration stage is fitted per match from the hand-annotated pitch keypoints supplied with the dataset (Fig. 5), the jersey-region gate was relaxed and a detection-level team-assignment variant added, both set by us on released footage, and implementation changes let it run at match length (Section 4.8). The scores below describe an adapted reference system on this footage; because the adaptation drew on released data, including prefixes of a test half, they are not a held-out transfer measurement.

For each of the twenty halves we run the pipeline independently, under the same configuration, on the whole half as annotated (45 to 49 minutes) and on its opening 30 seconds, the clip length of the SoccerNet-GSR benchmark (Table 3). A full-half run costs about 37 GPU-hours (Section 4.8). Each result is scored both with the official attribute-matching requirements and with attribute matching disabled; the latter measures geometry and association alone.

Over the twenty halves the full-half GS-HOTA is 12.25 (DetA 8.53, AssA 18.24, LocA 82.71), and 21.66 with attributes off; the four halves of the released test split average 10.96. On the opening 30 seconds the same twenty halves average 31.35, and 55.72 with attributes off (Table 3). In every half the full-half score lies below the 30-second score of the same half, by a factor of 1.8 to 3.9 under the official metric and 1.7 to 4.2 with attributes off (Table 3). The 30-second windows are anchored at kickoff rather than sampled throughout each half. Both lengths in Table 3 use the per-detection team-assignment variant; the comparison of team-assignment variants below concerns only M7, first half. The sequence-length study below examines the gap by rescoring every half over growing windows and by rerunning the pipeline on one half.

The spread across halves is itself a finding. Over whole halves the official score ranges from 18.23 (M5, first half) to 2.94 (M9, second half), and the twenty halves divide by match: the four halves of M1 and M9 score between 2.94 and 9.92, the sixteen halves of the other eight matches between 10.22 and 18.23, and localisation under the official metric separates the two groups without overlap (70.2 to 79.0 against 82.6 to 87.0). With attributes off, detection accuracy (DetA, averaged over the metric's localisation thresholds, which span 0.65 to 5 m) lies between 15.9 and 41.4 on the four halves against 42.5 to 51.2 on the sixteen (Supplementary Table 1 lists the attributes-off components). On two of the four, the second halves of M1 and M9, attribute matching imposes a substantial additional penalty (official DetA 1.15 and 1.23), taking the official score to 3.60 and 2.94 where geometry and association alone score 7.44 and 17.91. On M7, the attributes-off score changes less between halves (26.04 against 25.08) than the official score (18.09 against 12.89). A related separation is visible in the first 30 seconds: three of the four halves of M1 and M9 score between 27.37 and 31.31 with attributes off, and their localisation is about ten points lower on average (73 to 78 against 79 to 90, Supplementary Table 1); the other seventeen, among them M9, second half (52.58), score between 43.92 and 77.45. The four lowest full-half scores belong to two matches, and several factors vary together across them: capture conditions, calibration quality (the misaligned stitch seam of M1, Section 4.2), attribute legibility, and the ground-truth property noted in Methods (player and ball positions clamped

Table 3: Game state reconstruction on all twenty halves, each run and scored over the whole half and, independently, on its opening 30 seconds, the clip length of the SoccerNet-GSR benchmark: the game state reconstruction pipeline of Section 4.7, configured by the authors on released footage that includes a test-split half and run end to end under one configuration, the per-detection team assignment. GS-HOTA is the official metric, which classes detections by (role, team, jersey); DetA, AssA and LocA are its detection, association and localisation components; *Attrs off* rescores the same predictions with attribute matching disabled, so that only geometry and association count (components in Supplementary Table 1). The test split of the released match-level split is marked †. Means are unweighted means of the per-half scores in each column, computed from unrounded scores.

	Opening 30 s		Full half (45 min)				
Half	GS-HOTA	Attrs off	GS-HOTA	DetA	AssA	LocA	Attrs off
M1, 1st	23.10	27.37	6.68	4.29	10.50	71.77	8.46
M1, 2nd	6.49	31.31	3.60	1.15	11.25	72.06	7.44
M2, 1st	55.20	66.12	14.28	10.04	20.37	83.02	22.14
M2, 2nd	26.49	49.15	12.13	8.01	18.38	83.00	22.08
M3, 1st	36.49	61.75	16.59	10.78	25.73	82.73	26.09
M3, 2nd	32.49	63.71	12.21	6.83	21.84	85.71	23.71
M4, 1st	24.60	48.57	12.36	7.11	21.58	86.42	24.56
M4, 2nd	41.14	58.80	11.03	8.39	14.55	82.55	20.05
M5, 1st	39.89	72.09	18.23	15.68	21.20	86.67	25.10
M5, 2nd	39.74	62.52	16.54	12.26	22.34	87.00	24.68
M6, 1st	29.59	54.83	13.36	10.47	17.08	85.33	23.00
M6, 2nd	31.60	54.20	10.22	8.16	12.80	84.88	18.49
M7, 1st†	37.15	77.45	18.09	14.67	22.32	86.84	26.04
M7, 2nd†	33.51	66.37	12.89	7.75	21.48	86.67	25.08
M8, 1st	41.80	72.72	12.73	9.60	16.91	85.74	25.89
M8, 2nd	34.29	60.62	12.26	7.80	19.34	83.17	21.65
M9, 1st†	21.56	29.53	9.92	9.69	10.16	79.04	16.15
M9, 2nd†	8.92	52.58	2.94	1.23	7.10	70.19	17.91
M10, 1st	38.43	60.79	15.27	8.71	26.79	85.32	29.26
M10, 2nd	24.41	43.92	13.64	8.07	23.07	86.18	25.40
Mean, all twenty halves	31.35	55.72	12.25	8.53	18.24	82.71	21.66
Mean, test split†	25.29	56.48	10.96	8.34	15.26	80.68	21.29

to the pitch rectangle in M9). M10, whose positions are clamped in the same way, scores within the stronger group, so clamping alone does not account for the divide. We do not separate these potential contributors here, which is one reason we report every half rather than a single aggregate.

With attributes off, association is the larger bottleneck in every half.

The association component is the smaller of the two in every half: AssA lies between 3.3 and 18.6 against DetA between 15.9 and 51.2, a ratio of 2.5 to 6.0 (Supplementary Table 1), and the pipeline produces 123 to 422 tracklets per half against 22 to 35 annotated identities, where the same halves at 30 seconds produce 23 to 37 tracklets for 22. Table 4 decomposes the score of M7 first half, the best half of the test split, in detail. Disabling attribute matching gives 26.04, the ceiling for attribute correction with the current detections and associations held fixed: perfect role, team and jersey

would score exactly that, so all attribute work together is worth at most 7.94 points. Within that budget the successive costs, in the displayed order, are role −0.40, team −3.27 and jersey −4.27, and the per-attribute accuracies on 36,665 geometrically matched pairs are role 97.92%, team 76.17% and jersey 45.95%. Jersey numbers are wrong rather than missing: only 3.4% of predictions carry a null jersey.

The ceiling itself is set by detection in pitch space and by association, and these are far from balanced. With attributes off the pipeline places the players it detects accurately (LocA = 84.7), reaches a detection accuracy of DetA = 51.2 and fails to hold identity over time (AssA = 13.4): perfect detection would raise the ceiling to 36.6, perfect association to 71.6. DetA here confounds missed and spurious detections with projection error between the provider's calibration and the keypoint homography, which are consistent only to 2 to 3 m (Section 4.2); we have not separated the two. Nor is the association loss simple fragmentation. The run produces 345 predicted tracklets for 23 identities, but most are short fragments: discarding every tracklet shorter than 100 detections leaves 65 and changes GS-HOTA by 0.01, while 41 of the 154 tracklets with at least ten detections span more than three times as many frames as they contain detections, so identities are reused across gaps rather than following one player. On this half, then, the association loss is identity reuse across gaps more than fragmentation.

Table 4: Marginal cost of each game state attribute, computed on the first half of match M7 (the best-performing half of the test split in Table 3). GS-HOTA classes detections by the triple (role, team, jersey), so a detection with any incorrect attribute cannot match ground truth at all; each row switches one further attribute on, and *Marginal* is the incremental cost of enabling that attribute after those in the preceding row, computed from unrounded scores, so it may differ by 0.01 from the difference of the displayed values. DetA and AssA are the detection and association components of the metric. The first row, with no attributes scored, is the ceiling available to the official metric: perfect attributes would score exactly this, so it carries no marginal cost.

Configuration	GS-HOTA	DetA	AssA	Marginal
Geometry and association only	**26.04**	51.21	13.41	ceiling
+ role	25.64	49.36	13.48	−0.40
+ role, team	22.36	35.12	14.37	−3.27
+ role, team, jersey (official)	**18.09**	14.67	22.32	−4.27

Clip-scale scores exceed full-half scores on this footage.

Two experiments separate how far this gap belongs to the length scored and how far to running the pipeline on less footage. In the first, the whole-half predictions of every half are held fixed and scored over growing windows from kickoff, from the first 30 seconds to the whole half (Fig. 4; Supplementary Table 2 lists every half and window). The mean over the twenty halves falls monotonically, from 23.21 at 30 seconds to 12.25

over the whole half under the official metric and from 55.33 to 21.66 with attributes off. Individual halves wobble, and on two, M6 first half and M10 second half, the whole half scores above its 30-second window (13.36 against 11.90 and 13.64 against 11.09); on every other half the whole half scores below it. The independent 30-second runs of Table 3 score higher still on the same clips, 31.35 against 23.21 under the official metric, while with attributes off the two agree (55.72 against 55.33): running the pipeline on a 30-second clip rather than on the whole half gains about eight points, almost entirely in the attributes, which the pipeline assigns per tracklet and by clustering over the whole sequence (Section 4.7).

In the second experiment the pipeline is rerun end to end on nested prefixes of the first half of match M7, the half on which the adaptation was also explored, under the per-detection team-assignment configuration, a design the fixed cameras and continuous footage permit, so that footage, calibration and configuration are shared across prefixes and each prefix is a separate run. Accuracy peaks at one minute, at a GS-HOTA of 47.59, then falls monotonically through every longer prefix (39.88 at two, 30.97 at five and 25.60 at ten minutes), reaching 18.09 over the full half: 38% of the one-minute figure from the same half. From the 30-second prefix to the full half, predicted tracklets grow from 26 to 345 against 22 real identities in the shorter prefixes and 23 over the full half, a substitute entering late.

Two properties of the design bear on both experiments: the windows are anchored at kickoff, and the metric's association term is computed over whole trajectories, so some decline with length is intrinsic to scoring longer horizons. The first experiment measures that intrinsic decline directly, since its predictions do not change; the second adds the effect of running on less footage. The increase from 26 to 345 predicted tracklets is consistent with fragmentation but does not isolate its contribution from either. A method characterised on thirty-second or one-minute excerpts, the lengths at which existing game state benchmarks are defined [22, 29], cannot therefore be assumed to retain that accuracy over a half: full-match evaluation is a distinct protocol, not a longer clip.

The best configuration depends on the length evaluated.

Team assignment makes the same point from a different direction, on the same prefixes. Clustering per-detection appearance embeddings, rather than tracklet-mean embeddings, raises team accuracy at five minutes from 51.6%, chance on two classes, to 86.6%, and GS-HOTA from 19.48 to 30.97. At two minutes and below the ranking reverses: tracklet-mean clustering scores 40.35 against 37.15 at thirty seconds, 57.82 against 47.59 at one minute and 44.81 against 39.88 at two minutes. No single variant is best at every length, so Methods describes both; every score of Table 3 uses the per-detection variant, and the crossover is a concrete instance of a design choice that looks right on clips but no longer holds by five minutes, the longest prefix at which both variants were scored.

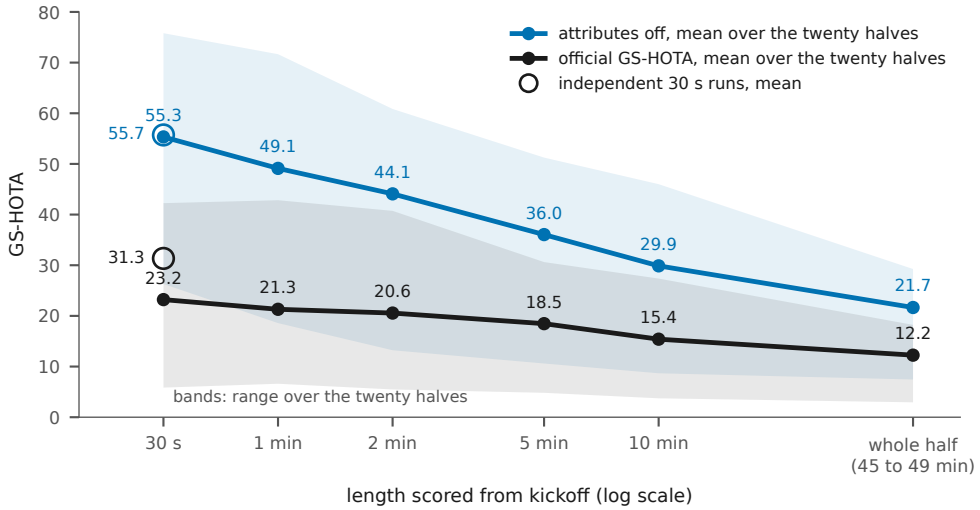


Fig. 4: Game state reconstruction accuracy against the length scored. Filled markers: the whole-half predictions of every half (Table 3) scored over their first 30 seconds, 1, 2, 5 and 10 minutes and the whole half, mean over the twenty halves (Supplementary Table 2), under the official metric (black) and with attribute matching disabled (blue); the bands span the twenty halves. Hollow markers: the mean of the independent 30-second runs of Table 3, in which the pipeline saw only the clip.

2.4 Ball action spotting baseline

The second baseline spots ball actions (ball action spotting, BAS) from the released game state rather than from pixels, which the synchronised event and state layers permit (Section 2.1). It is an offline temporal convolutional network over per-frame features of the supplied player trajectories, with or without the supplied ball track (Section 4.9, Fig. 6b). Events have been detected from tracking data with rule-based algorithms, classical classifiers and learned models [46–50], and group activities from trajectories and, separately, from video, with most events detectable from either [51]; SoccerNet-GAR compares the two inputs under one protocol and finds tracking the stronger for group activity recognition [31]. Here the trajectories come with full-pitch video of the same matches and are scored under the ball action spotting metric. Because the input is the supplied ground-truth state, the figures measure how much of the event layer the released state carries under one offline model; spotting from video works from a different input, and its obstacles on this footage are set out in Section 3. The supplied ball track is itself a reconstruction rather than a per-frame measurement: the ball moves in straight lines at constant speed between anchors, most of which are annotated events, so most of its changes of velocity fall within about 0.1 s of an annotated event instant. The with-ball condition therefore sees information tied to the event annotation, which may inflate its scores relative to an observed ball, and we report it as a reference condition rather than as the value of ball tracking.

Table 5 reports the five-fold cross-match evaluation and Table 6 the per-class breakdown pooled over all ten matches. Player positions alone yield a cross-validated mean average precision of 0.222 ± 0.021 at $\tau = 1\,\text{s}$ and 0.524 ± 0.048 at $\tau = 5\,\text{s}$, against 0.008 and 0.064 for a uniform-cadence chance baseline. Adding the supplied ball track raises these to 0.599 ± 0.064 and 0.666 ± 0.045; on the Challenge fold alone, the held-out pair of the match-level split, the with-ball figure is 0.617. The chance baseline is not a formality: its support-weighted score at $\tau = 5\,\text{s}$ is 0.279, because two classes account for 82% of events and, over the 932 minutes of Table 1, an event occurs every 2.6 s and a pass every 6 s, so the margin over chance rather than the absolute value is the quantity of interest.

The two conditions separate far more at the tight tolerance than at the loose one. Without the ball, tightening τ from 5 s to 1 s costs 0.302 mean average precision; with the ball it costs 0.067. At five-second tolerance the ball adds 0.14 mean average precision (0.524 to 0.666); at one second it adds 0.38, a factor of 2.7. A larger improvement at the tighter tolerance is what a track whose changes of velocity coincide with the annotated events would produce, so it does not measure the temporal precision an observed ball would provide.

Duel outcomes are the weakest classes in both conditions.

Events defined by the outcome of a duel are the two weakest classes with and without the ball: `Ball Player Block` reaches 0.070 from trajectories and 0.181 with the ball, and `Player Successful Tackle` 0.046 and 0.121, against 0.286 and 0.853 for `Pass`. Both classes ask which of two converging players won a contact, and the pattern is consistent with that outcome not being represented in ground positions and ball location at this resolution. The classes the ball transforms are those defined by the ball rather than by the players: `Throw In` gains 0.66, `Drive` 0.60, `Pass` 0.57 and `Free Kick` 0.53. `Out` gains only 0.22, reaching 0.364: it is a synthesised restart marker (Section 4.4) rather than an observed boundary crossing, and in two of the ten matches the recorded ball never crosses the lines. `Header` ($n = 31$) and `Goal` ($n = 44$) are reported for completeness; at that support the figures are indicative only, which is why both an unweighted and a support-weighted mean are given.

The with-ball figures pool two regimes.

The ball track also differs between matches. In M7 the ball leaves the field of play on 3% of frames, as it must; in M9 and M10 the recorded position is clamped to the pitch rectangle and never does. On the Challenge fold, adding the ball raises mean average precision at $\tau = 1\,\text{s}$ from 0.267 to 0.734 on M7 and from 0.191 to 0.530 on M9. The pooled with-ball figure on that fold (0.617) therefore averages an unclamped and a clamped match; Table 5 gives both.

Table 5: Five-fold cross-match evaluation of the spotting model on the supplied trajectories. Each match appears in the test set exactly once; fold 0 is the SoccerTrack Challenge 2025 pair. Validation for each fold is the following fold's test pair, so no fold's decoding or stopping epoch is chosen on its own test matches; the network settings shared by all folds were fixed once on M2 and M10 (Section 4.9). Each entry is a single training run. The last block gives the two matches of the Challenge fold separately: the ball track is unclamped in M7 and clamped to the pitch rectangle in M9, so the value of the ball differs sharply between them and the fold's pooled row averages two regimes.

		$\tau = 1\,\mathrm{s}$		$\tau = 5\,\mathrm{s}$	
Input	Test fold	mAP	mAP_w	mAP	mAP_w
trajectory	M7, M9 (Challenge)	0.207	0.202	0.472	0.637
	M1, M2	0.222	0.272	0.501	0.676
	M3, M4	0.231	0.259	0.600	0.709
	M5, M6	0.199	0.258	0.516	0.678
	M8, M10	0.252	0.275	0.531	0.664
	Mean	**0.222**	**0.253**	**0.524**	**0.673**
	SD across folds	0.021	0.030	0.048	0.026
trajectory + ball	M7, M9 (Challenge)	0.617	0.788	0.679	0.826
	M1, M2	0.633	0.827	0.685	0.846
	M3, M4	0.624	0.819	0.701	0.843
	M5, M6	0.636	0.808	0.679	0.832
	M8, M10	0.486	0.737	0.588	0.777
	Mean	**0.599**	**0.796**	**0.666**	**0.825**
	SD across folds	0.064	0.036	0.045	0.028
Challenge fold, per match					
trajectory	M7	0.267	0.275	0.550	0.669
	M9	0.191	0.176	0.437	0.632
trajectory + ball	M7	0.734	0.829	0.785	0.852
	M9	0.530	0.742	0.604	0.815
Uniform cadence	Mean	0.008	0.033	0.064	0.279

Table 6: Per-class average precision at $\tau = 1\,\mathrm{s}$, pooled over the five cross-match folds and therefore over all ten matches. n is the number of ground-truth events across the benchmark. Rows marked † have $n < 100$. Classes are ordered by support. Because AP is pooled over matches, the column means differ slightly from the fold-averaged means of Table 5.

Class	n	chance	trajectory	+ball
Pass	9315	0.042	0.286	0.853
Drive	8255	0.036	0.227	0.830
High Pass	1157	0.007	0.320	0.745
Out	771	0.004	0.148	0.364
Cross	394	0.001	0.327	0.794
Throw In	385	0.006	0.254	0.910
Ball Player Block	353	0.000	0.070	0.181
Player Successful Tackle	307	0.002	0.046	0.121
Shot	266	0.002	0.317	0.735
Free Kick	150	0.001	0.171	0.701
Goal†	44	0.000	0.336	0.718
Header†	31	0.000	0.189	0.347

3 Discussion

SoccerTrack v2 places continuous full-pitch video, match-length trajectories with persistent identities and actor-linked events in one resource. Its value lies in that combination: methods that reason over positions and events together can be trained and scored on the same matches, a reconstruction system can be asked to hold identity for 45 minutes rather than 30 seconds, and what either produces can be inspected against the footage at the moment it is claimed to occur. The two baselines provide initial reference results and identify limitations of the evaluated systems.

What the baselines reveal.

Over whole halves, attributes-off detection accuracy (DetA) ranges from 15.9 to 51.2 (Supplementary Table 1), while official localisation exceeds 70 in every half. With attributes off, association is the smaller component in every half (AssA 3.3 to 18.6 against DetA 15.9 to 51.2); on M7, first half, perfect association would nearly triple the attributes-off score, whereas perfect detection would add about eleven points. The four lowest full-half scores belong to two matches; capture and calibration conditions, jersey legibility and ground-truth properties vary together across them, and their contributions have not been separated, although clamping alone does not explain the gap, since M10 is clamped like M9 and scores with the stronger group. Three of the four weakest full halves also form the lowest-scoring attributes-off group at 30 seconds. Scored over growing windows from kickoff, the twenty-half mean of the whole-half runs falls from 23.21 at 30 seconds to 12.25 over the whole half, and independent 30-second runs score higher still (31.35); on nested prefixes of one half rerun end to end, accuracy peaks at one minute and falls to 38% of that value over the full half, and the better team-assignment variant changes with the length evaluated. Together, these results show that opening-clip scores overstate full-half scores for this adapted baseline on this corpus, supporting full-half evaluation alongside clip-scale evaluation. The spotting baseline shows the complementary pattern on the supplied trajectories: at five-second tolerance player positions alone recover most of what the with-ball model reaches (0.524 against 0.666), while at a one-second tolerance adding the supplied ball track increases mAP by a factor of 2.7, a gain that the track's reconstruction around the annotated events may inflate; both models struggle with duel outcomes. These offline results on ground-truth state quantify how well events can be predicted from the supplied trajectories.

Limitations.

The annotations originate from a commercial tracking and event system. We inspected the positions against the video by projecting them onto the panorama and onto a bird's-eye view of the pitch (Figs. 2 and 5), in spot checks and throughout the development of the baselines, and the event timelines were checked and corrected against the video during curation. We have not conducted a quantitative audit: there is no inter-annotator agreement, no correction-rate statistic and no positional residual against independently placed reference points. The full-pitch view is what makes such inspection possible; it does not by itself establish accuracy. What can be stated is that the

release itself fixes a scale of a few metres: positions are quantised to a 1.05×0.68 m grid, and the hand-annotated calibration correspondences are self-consistent only to roughly 2 to 3 m on the pitch (Section 4.2); neither figure measures the provider's accuracy, and positional claims finer than a few metres are not supported. Player identity is carried by the jersey number and by a provider identifier that persists across the halves of a match and, for squads that appear in several fixtures, across matches (Section 4.3); how the provider maintained identity where a number was occluded is not documented, identity consistency was not audited, and the pseudonymised release does not resolve identifiers to persons. The field of view covers the pitch; it does not follow that a given player is unoccluded at every instant, and occlusion by other players was not audited. All ten matches are university-level fixtures from a single league, recorded by one provider's fixed panoramic systems, so playing style, tempo and camera conditions may not transfer to professional or broadcast settings, and exhaustive bounding-box annotation was beyond the resources of this release, so box-level supervision is released as a curated subset (Section 4.5).

The baselines are bounded in turn. Reconstruction is evaluated on all twenty halves, over the whole half and on the opening 30 seconds; the attribute decomposition and the nested-prefix reruns were computed on one half. The adapted pipeline's jersey gate and team-assignment variant were chosen on released footage that includes a test half, so its scores characterise the adapted system on this data, and no variance estimate accompanies any single-half figure. The spotting figures assume the supplied trajectories and an offline spotter. The with-ball condition uses a ball track reconstructed around the annotated events rather than measured per frame (Section 4.9), and its cross-validated figures include M9 and M10, whose recorded ball never leaves the pitch.

Relation to broadcast datasets.

SoccerTrack v2 complements the broadcast-derived corpora that dominate the field rather than replacing them. Models intended for television footage are best trained and evaluated on television footage, and the SoccerNet family [22, 24, 25, 28] remains the right resource for that setting. What a fixed panoramic recording adds is a field of view that spans the pitch: players are not lost to camera motion, every annotated position has footage at the instant it is claimed, identities are held for whole halves (the median track spans 45 minutes), and events resolve to those identities. Whether the provider measured each position on this view is not documented (Section 4.1). The two settings answer different questions, and the game state annotations use the SoccerNet GameState format while the event annotations follow the SoccerNet ball action spotting convention, so tooling and methods can move between them.

Spotting events from pixels.

The spotting baseline consumes trajectories rather than video because of a property of the footage. A panoramic frame covering the whole pitch at 4K presented to a video backbone at its usual input resolution requires an order-of-magnitude downscale, at which a 22 cm ball spans about one pixel, or else a crop. A crop centred on the annotated ball position is an oracle a deployed system would not have; a crop centred on a detected ball must first solve ball detection in panoramic footage, where the ball

spans only a few pixels, and compounds its input's failures at exactly the contested moments where events concentrate. The released ball track cannot serve as ground truth for such a detector: it is anchored at the annotated events and interpolated between them rather than tracked, and it is clamped to the pitch rectangle in two matches (Table 1). A ball track in pitch coordinates is in any case a demanding artefact, since an airborne ball projected through a ground-plane homography lands far from where it is. The remaining alternatives, a play-centred crop derived from player positions or a full-frame downscale, are natural first experiments.

Outlook.

The most informative next experiment is the composition of the two baselines. The spotting figures assume the supplied game state, so running the same spotter on tracks predicted by the reconstruction baseline would show how much of the end-to-end gap is attributable to reconstruction and how much to spotting, and would give the first measurement from panoramic video to ball actions on this data. The fixed match-level split and the released evaluation tooling are intended to make such results directly comparable across contributions. The SoccerTrack Challenge leaderboard [45], rather than this paper, carries the reference tracking results on the curated bounding-box subset. Because the panoramic recording contains every narrower framing within it, the cost of ball-following framing can also be studied on this data, by rendering a viewport and comparing against the full view; we leave that ablation to future work.

4 Methods

4.1 Match recording and camera systems

Matches were captured by fixed, stitched panoramic camera systems (BePro) whose field of view covers the complete pitch; the released panorama is about 4K wide (3840 or 4096 px) and 1080 to 1906 px tall. The ten matches involve university-level amateur teams, selected for their standard of play and for their willingness to participate in data collection. Positional data for players were provided by the BePro system, synchronised to video timestamps; the provider's processing chain, and whether it derives the positions from the same panoramic feed that is released or from a separate pass over its cameras, is not part of the release, so the positions are described here as lying within the released field of view rather than as measured on the released view.

4.2 Pitch coordinates and calibration

Pitch coordinates are metric. The origin is the centre circle ($x = 0$, $y = 0$); x increases towards the right half of the pitch as seen from the main camera, and y increases towards the touchline nearer the camera. We assume FIFA-standard dimensions of $105\,\text{m} \times 68\,\text{m}$, so that goal lines are at $x = \pm 52.5$, touchlines at $y = \pm 34.0$, and the halfway line at $x = 0$. Positions outside the rectangle are legal, since balls and players can leave the field of play, and are not clipped before metric computation, with one exception: in matches M9 and M10 the recorded positions of the players as well as the ball (Section 4.9) are clamped to the pitch rectangle at source, so no position beyond

the lines occurs in either. Fig. 5 illustrates the geometry: a released panoramic frame, the same frame after distortion correction with the pitch template and the frame's annotated positions projected into it, and the same instant in the metric coordinate system of the annotations.

The release ships, for each match, 65 hand-annotated correspondences between pitch coordinates and the distortion-corrected panorama; they are what Fig. 5 and the pipeline's calibration stage (Section 4.7) are fit from. A planar homography refit to these correspondences reproduces them with a median residual of 2.9 to 6.9 px in nine of the ten matches; the exception is M1, a training-split match whose stitch has a misaligned central seam, which no single-view distortion model rectifies (median 34.5 px). Expressed on the pitch through the inverse mapping, the same correspondences show a median consistency of roughly 2 to 3 m, dominated by the far half of the pitch where a pixel subtends the most metres; this is the scale against which the localisation component of GS-HOTA, whose thresholds span 0.65 to 5 m, should be read. Because the ground truth and the pipeline's predictions reach pitch coordinates through different calibrations, the provider's and a homography on these keypoints, part of the localisation error is a mismatch between the two frames that we have not quantified beyond this consistency check; on M2, ground-truth positions projected through the keypoint homography land within a median of 14 px of their annotated image positions. The released positions themselves are quantised: every pitch coordinate lies on a fixed grid of one hundredth of each pitch dimension, 1.05 m in x and 0.68 m in y, machine-exact across all twenty halves, so the grid sets the effective positional resolution of the release. Beyond these consistency checks, the provider's own positional accuracy cannot be audited from the release.

4.3 Game state annotation

GSR annotations comprise 2D pitch locations for each player, persistent player identities with their jersey numbers, team sides, and player roles. Positions and identities were supplied by the BePro system synchronised to the video. We inspected them visually against the footage: the annotated positions were projected onto the panorama and onto a bird's-eye view of the pitch (Figs. 2 and 5) and checked frame by frame in spot checks and throughout the development of the baselines. No quantitative audit of positional accuracy or identity consistency was performed; the consistency checks of Section 4.2 give the scale at which the positions should be read.

GSR annotations are shipped in the SoccerNet GameState format [22], one file per half, so that SoccerNet tooling reads them directly. Each object annotation describes one player or goalkeeper in one frame; the files contain no ball or referee annotations, and each frame also carries normalised pitch-line points. An annotation carries a per-half track identifier; a role (player or goalkeeper); a team side, fixed per half from the attacking direction at kickoff, so one squad may carry different sides in its two halves; the jersey number, present on every record and constant along a track, so that it describes the player rather than what is legible in the frame; the provider's integer player identifier, present on every record, which persists across the two halves of a match and, for squads that appear in several fixtures, across matches, so it links a

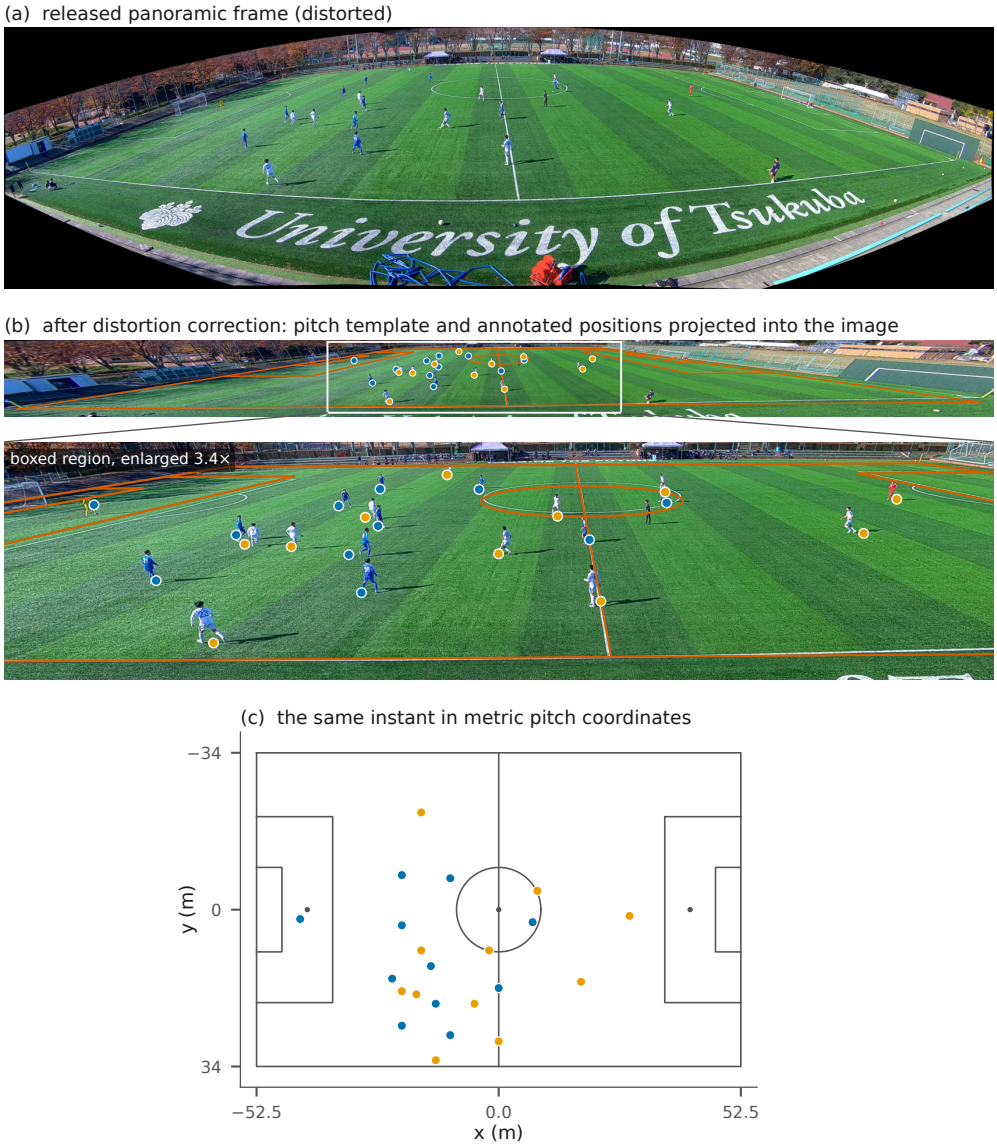


Fig. 5: From released pixels to metric pitch coordinates, match M2, first half. (a) A frame of the released panoramic video: the stitched view is distorted and the pitch lines are curved. (b) The same frame after distortion correction, with the FIFA-standard pitch template and the frame's annotated player positions (coloured by team side) projected into the image through a homography fitted on the hand-annotated pitch keypoints, shown over the pitch region with the boxed region enlarged below. (c) The same instant in the metric coordinate system of the annotations: a 105×68 m pitch with the origin at the centre circle, the touchline nearer the camera at the bottom.

player's appearances across the corpus without resolving them to a person; an image-space box on the released panorama, placed from the pitch position with a position-dependent size rather than drawn around each player, and distinct from the human-annotated boxes of Section 4.5; and the pitch position, a single point at the player's feet rather than a box. Jersey number and team side are the identity the reconstruction benchmark scores; the event annotations name their actor by the same player identifier.

4.4 Event annotation

Event data were obtained from the BePro system and converted to the twelve classes used for ball action spotting. Classes are a many-to-one, hand-built mapping of the provider's event tags; tags that carry no spottable ball action (possession spells, failed tackles and duels, fouls, substitutions) are discarded. `Out`, which no BePro tag encodes, was synthesised from the event stream itself: one `Out` is inserted before each restart tag (throw-in, corner kick, free kick or successful goal kick), placed one second after the preceding record when that record names an actor and at its timestamp otherwise, and attributed to the last actor-bearing action before the restart. It therefore marks play going dead rather than a measured boundary crossing, and because free kicks are among the triggers, roughly a fifth of `Out` events follow fouls rather than the ball leaving the field. The mapped timelines were then manually checked and corrected against the synchronised video. Each event carries the acting team and, when identifiable, the actor's player identifier, the same provider identifier carried by the game state annotations, which is what allows events to be cross-referenced with tracks.

BAS annotations are stored per match in the SoccerNet BAS convention [26], with the half, a timestamp in milliseconds from the first decoded frame of that half's video, and the nearest video frame, which is the same frame index in the video and in the game state annotation of that half. The twelve classes are *Pass*, *Drive*, *Header*, *High Pass*, *Out*, *Cross*, *Throw In*, *Shot*, *Ball Player Block*, *Player Successful Tackle*, *Free Kick*, and *Goal*. Dual actions at the same timestamp (for example a shot that becomes a goal) are emitted as two separate annotations.

4.5 Bounding-box subset curation

Comprehensive bounding boxes are not part of the release. A curated subset with bounding boxes and player identities is released instead, and was used for the SoccerTrack Challenge at MMSports 2025: one four-minute panoramic clip per match, 59,842 clip frames in total, each annotated with the 22 players on the pitch, for 1,316,524 boxes.

4.6 Benchmark split

We adopt the SoccerTrack Challenge 2025 split, holding out M7 and M9 for test. How the eight training and validation matches are divided between training and validation differs between the two protocols the released tooling implements, and the difference matters (see below). Every headline spotting figure in this paper comes from the five-fold cross-match protocol of Section 4.9, under which the Challenge fold (fold 0) trains on six matches and validates on M1 and M2. The released single-split protocol instead

reserves M2 and M10 for validation, a pair that mirrors the test pair: each contains one of the two matches whose ball track is clamped to the pitch rectangle. That pair enters the results through the validation-sensitivity check below and through the earlier sweep that fixed the network settings shared by all folds (Section 4.9).

We deliberately do not split within matches. The two halves of a match share the two squads, both squads' shape, the pitch, the camera and the weather, and every per-match data property, so a within-match split would place all of those on both sides of the partition and conceal exactly the variation a held-out match exposes.

The choice of *validation* matches is equally consequential and less obvious. With the test pair held at the Challenge matches, validating on {M2, M10} rather than {M1, M2} lowers the with-ball mean average precision at $\tau = 1\,\mathrm{s}$ from 0.617 to about 0.40 (three seeds, 0.395 to 0.408), almost entirely on M9 (0.53 against 0.23); M7 moves by at most 0.03. Both configurations select the same decoding threshold and suppression radius, so the difference does not come from decoding. Changing the validation pair also changes the training set: validating on {M2, M10} trains on M1 in place of M10, the other match whose ball track is clamped, and the loss falls on the clamped test match. We therefore fix the validation rule as part of the protocol, each fold validating on the following fold's test pair, and report the cross-validated mean rather than any single configuration.

4.7 Game state reconstruction baseline

Our reference system is a game state reconstruction pipeline assembled in the TrackLab framework [22] from published components; Fig. 6a diagrams it and marks our contributions. The stages are RF-DETR [5] person detection, ViTPose [52] pose estimation, PRTReID [12] appearance features and role classification, BoT-SORT [7] tracking, pose-guided extraction of a torso region followed by ViTSTR [15] jersey-number recognition, calibration by thin-plate spline from the hand-annotated pitch keypoints to calibrated pixels and then a homography to pitch coordinates (learned broadcast-view field registration [9] is trained on broadcast framing rather than fixed panoramas), global tracklet association [53], and per-tracklet majority voting for role, team and jersey number. Beyond the published components, we contribute the calibration stage, fitted per match from the hand-annotated correspondences of Section 4.2; the relaxed jersey-region gate and the per-detection team assignment described below; and the match-length fixes of Section 4.8. The adaptation drew on released footage, including a test-split half: the relaxed gate was first explored on prefixes of the first half of M7 and the team-assignment comparison was made on the same prefixes. Evaluation uses the official GS-HOTA metric [22, 23] in pitch space with a 5 m Gaussian tolerance. GS-HOTA partitions detections into classes by the triple (role, team, jersey), so a detection carrying any incorrect attribute cannot match ground truth at all rather than losing partial credit.

Jersey-region gating.

The jersey stage crops a torso region from each detection using pose keypoints, then gates that region before recognition on a minimum area and a minimum oriented-box aspect ratio. Those thresholds are consequential at panoramic resolution. The median

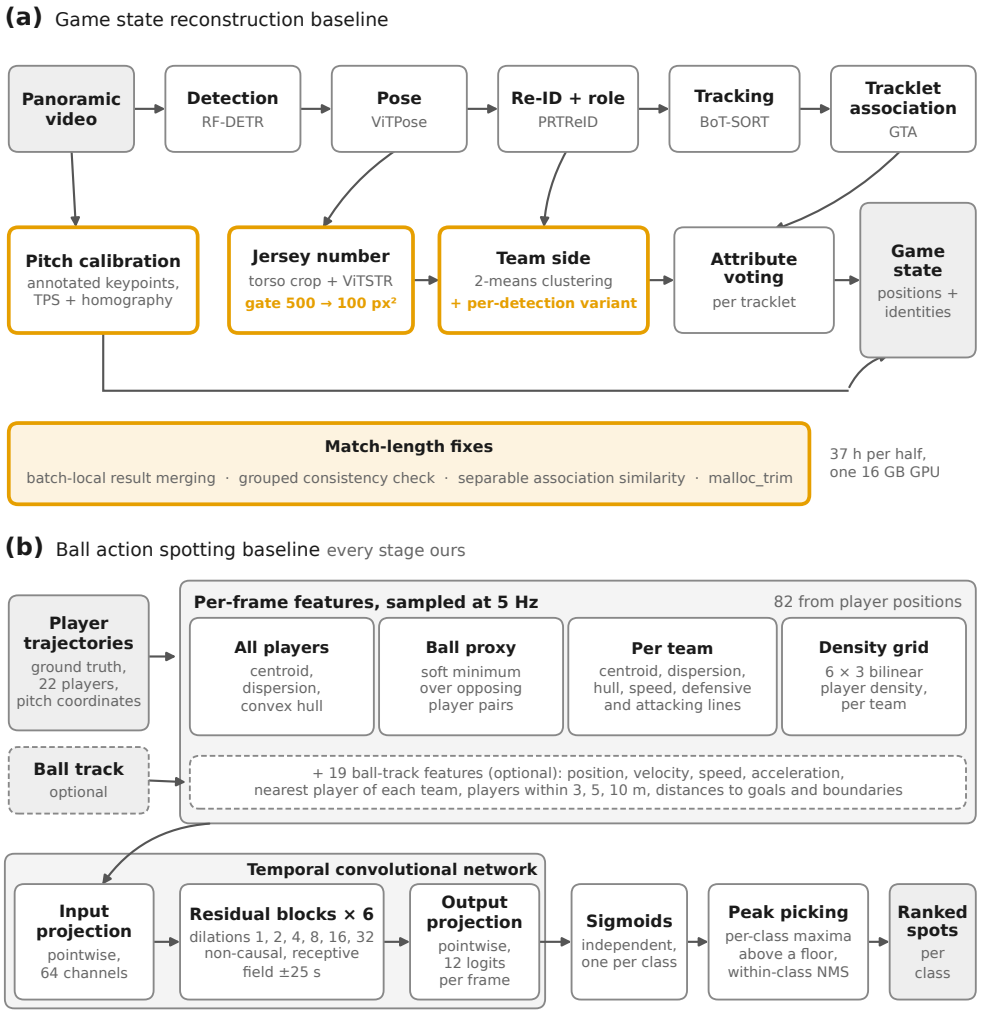


Fig. 6: The two baselines. (a) Game state reconstruction. White boxes are published components; highlighted elements mark our contributions: the calibration stage fitted from the hand-annotated keypoints, the relaxed jersey-region gate, the per-detection team-assignment variant compared with the original on nested prefixes (Section 4.7), and the four fixes that let the pipeline run at match length (Section 4.8). (b) Ball action spotting, every stage of which is ours (Section 4.9). Per-frame features computed from the supplied ground-truth player trajectories and sampled at 5 Hz, 82 in four groups and optionally 19 more from the ball track, pass through a non-causal temporal convolutional network: a pointwise projection to 64 channels, six residual blocks with dilations 1 to 32 (receptive field ±25 s) and a pointwise projection to twelve per-frame logits. Independent sigmoids score each class, and per-class peak picking above a floor with within-class non-maximum suppression yields ranked spots per class.

player box here is 25×55 px and the median proposed torso region 12×11 px ($132\,\text{px}^2$), so the default gate of $500\,\text{px}^2$ at aspect ratio 0.6 admits only 6.9% of detections. The recogniser is therefore starved of candidates rather than failing on those it sees. We relax the area threshold to $100\,\text{px}^2$, keeping the aspect ratio at 0.6. A subsequent sweep on M2, one of the two validation matches of the match-level split (Section 4.6), ranks the same setting first; no sweep row is reported as a held-out result.

Team assignment.

The default team assignment clusters tracklet-mean appearance embeddings with 2-means and labels the two sides by mean pitch position. Because tracklets fragment as sequences lengthen, tracklet means become unreliable, so we also implement clustering over per-detection embeddings with each tracklet taking the majority label of its own detections. Goalkeepers, referees and non-person tracks are excluded from clustering in both variants. Team accuracy is measured on prediction and ground-truth pairs matched geometrically in pitch space by the Hungarian assignment at a 2 m threshold. Neither variant dominates on the nested prefixes of Section 2.3, so we report both there; every score of Table 3 uses the per-detection variant.

4.8 Running the pipeline over a full half

The pipeline's implementation is implicitly clip-scale. On a full half it does not finish, because of four scaling problems that are not apparent at clip length. Each was individually blocking, the fixes are documented with the released tooling so that the pipeline is runnable at the scale the benchmark defines, and outputs were verified unchanged, exactly or to float32 rounding, in every case.

- **Quadratic result merging.** Per-batch result merging cost $O(N)$ per batch, so quadratic overall; restricting the merge to the batch's own rows and columns removes the superlinear term with score-identical output.
- **Quadratic consistency check.** A consistency check iterated the track-id column's values rather than its unique ids, giving $O(N^2)$; a single grouped pass has bit-identical output.
- **Association memory.** Global tracklet association materialised full $n_1 \times n_2$ similarity tensors on the GPU to compute a scalar mean, exceeding GPU memory over a 45-minute half; the mean pairwise cosine equals the dot product of mean unit embeddings, an exact separable identity.
- **Retained arena memory.** glibc retained freed per-thread arena memory across pipeline stages, so resident memory grew to tens of gigabytes while live data was under 1 GB and the process was killed by the kernel; periodic `malloc_trim` bounds the resident set.

With these changes a full half completes in approximately 37 h on one 16 GB GPU. Every half was run over its whole length and is reported in Table 3.

Scoring long sequences needs one further robustness change. Calibration yields non-finite pitch coordinates for a small and length-dependent fraction of detections near the horizon, at most 0.78% of object predictions over a full half. These cannot be

matched in pitch space; our released scorer drops them and reports the count, where the official evaluator aborts on the first such value.

4.9 Trajectory-based ball action spotting

The ball action spotting baseline predicts, for each of the twelve event classes, a ranked list of timestamps at which an event of that class occurs. We report temporal mean average precision at a tight tolerance of $\tau = 1\,\mathrm{s}$ and a loose tolerance of $\tau = 5\,\mathrm{s}$, following the SoccerNet ball action spotting protocol [26, 54]. We follow their implementation exactly, which differs from the obvious reading in three ways that together change the score by a factor of two to four: the tolerance τ is the *total* width of the acceptance window centred on the ground-truth event, so $\tau = 1\,\mathrm{s}$ accepts a prediction within $\pm 0.5\,\mathrm{s}$ of it, not $\pm 1\,\mathrm{s}$; assignment runs from each ground-truth event to the highest-scoring unmatched prediction inside the window, not from predictions to the nearest ground truth; and the precision-recall curve is sampled at 200 fixed confidence thresholds before 11-point interpolation. Our evaluator is checked against a port of their source and agrees to 10^{-9}. Because the class distribution spans a factor of 300 between the most and least frequent class, we report both the unweighted mean over the twelve classes and a mean weighted by class support, and we give the support alongside every per-class figure. Average precision is computed over predictions pooled across the test matches rather than averaged per match, so that a rare class is scored on all of its instances at once.

Benchmark definition.

The benchmark comprises the 21,428 events of Table 1, of which 17,056 fall in the eight training and validation matches and 4,372 in the two test matches.

Input representation.

The model consumes the released ground-truth player positions on the pitch and no pixels; predicted tracks are not used, so every figure in Section 2.4 is an offline result on the supplied trajectories. Positions are quantised to 1.05 m in x and 0.68 m in y, so a single-frame finite difference carries $26\,\mathrm{m\,s^{-1}}$ of quantisation noise; velocities are therefore computed at two scales, 0.48 s and 1.52 s, and both are supplied to the model. From each frame we derive 82 features in four groups: the centroid, dispersion and convex hull of all 22 players; a soft minimum over opposing player pairs, which serves as a proxy for the location of the ball, since player positions constrain the ball strongly enough that its trajectory can be inferred from them [55]; per-team centroid, dispersion, hull, speed and defensive and attacking lines; and a 6×3 bilinearly-binned player-density grid per team, which encodes configuration without depending on any player ordering. Features are sampled at 5 Hz, comfortably finer than the tighter evaluation tolerance.

We report two input conditions. The *trajectory* condition uses the 82 features above and is reproducible from the released game-state annotations alone. The *trajectory and ball* condition adds 19 features derived from the ball track, which is released with the dataset as an auxiliary annotation. The ball track is a reconstruction rather than a per-frame measurement: the ball moves in straight lines at constant speed between

anchors, with no deceleration, and most anchors are annotated events, so most of its changes of velocity fall within about 0.1 s of an annotated event instant, so this condition has access to information tied to the event annotation. In matches M9 and M10 the recorded position is also clamped to the pitch rectangle, as marked in Table 1, and Table 5 gives the per-match consequence on the Challenge fold. The features are: position, velocity, speed, acceleration, distance to the nearest player of each team, the number of players within 3, 5 and 10 m, and distances to each goal and to the boundary lines. The per-frame ball-status flag is deliberately excluded: its out-of-play value is close to a direct label for the *Out* class on eight matches, and it is constant on the remaining two.

Model, training and decoding.

A temporal convolutional network [56] maps the feature sequence to twelve per-frame logits (Fig. 6b). Its layout follows the single-stage network of MS-TCN [57], with residual blocks of two dilated convolutions [58]: a pointwise input projection to 64 channels, six residual blocks with dilations 1 to 32 giving a receptive field of ± 25 s, and a pointwise output projection. The network is non-causal, seeing 25 s on either side of each prediction, which is standard for offline action spotting and does not describe a real-time system. The twelve outputs are independent sigmoids rather than a softmax over classes, as in event heads that allow simultaneous events [59], because simultaneous events are part of the annotation convention. Targets are Gaussian of standard deviation 300 ms centred on each event, and the loss is binary cross-entropy with per-class positive weighting capped at 50. Training uses random 102 s crops with reflection augmentation in both pitch axes, which is label-preserving because no class is handed. Models are selected on validation mean average precision, computed with a looser $\pm\tau$ matching window than the reported metric, rather than on validation loss, which the positive weighting renders non-monotone in the reported metric. At inference, per-class local maxima above a floor are thinned by within-class non-maximum suppression, and each surviving peak is emitted with its height as its confidence. The floor, the suppression radius and the stopping epoch within a fixed 24-epoch schedule are chosen on the validation matches of the fold being evaluated, never on its test matches. The network width, dropout, learning rate and schedule were fixed once, for all folds and both inputs, by an earlier sweep of the trajectory model validated on M2 and M10, so for the two folds that test on those matches one test match informed these shared settings.

Split.

We evaluate with five-fold cross-match cross-validation. Each fold holds out two matches for test and two for validation, every match appears in the test set exactly once, and fold 0 is the SoccerTrack Challenge 2025 pair, so that fold 0 coincides with the held-out matches of the Challenge and future work on that split is directly comparable with these figures. Validation for each fold is the following fold's test pair, so fold 0 validates on M1 and M2; the rule is arbitrary but fixed, and no fold's decoding or stopping epoch is tuned on its own test matches. One consequence is that fold 4

validates on the Challenge pair, so the cross-validated mean is not fully independent of the Challenge fold, as in any k-fold design.

We do not split within matches, for the reasons given in Section 4.6; for this release the shared per-match properties are especially consequential, including the clamping of ball and player positions in M9 and M10 described above. A fully team-disjoint split is not achievable with these ten fixtures: one squad appears in four matches and another in three, so some folds share a team between training and test.

Chance baseline.

Because two classes account for 82% of events, and over the 932 minutes an event occurs on average every 2.6 s and a `Pass` every 6 s (Table 1), evenly spaced guesses achieve a non-trivial score. We therefore report a uniform-cadence chance baseline that emits spots at each class's mean rate, estimated once on the eight non-test matches of the Challenge split and used in every fold, and knows nothing else about the match; estimating the rates on each fold's own training matches instead changes the chance row by at most 0.002.

4.10 Evaluation metrics

For GSR we use the official SoccerNet GSR metric [22], with the non-finite-coordinate handling described in Section 4.8, and report Game-State Higher Order Tracking Accuracy (GS-HOTA) [23] with its standard decomposition into DetA, AssA and LocA. Scores are reported per half at both lengths in Table 3, whose means are over all twenty halves, with the mean over the four halves of the released test split alongside. Supplementary Table 1 lists the DetA, AssA and LocA components of the attributes-off scores at both lengths, so that every attributes-off component quoted in Section 2.3 has a display item. DetA and AssA are averaged over the metric's localisation thresholds, which under its 5 m Gaussian kernel correspond to pitch distances from 0.65 to 5 m, so DetA is not an accuracy at a single 5 m gate.

Evaluation windows.

To measure how the score of one run depends on the length scored, the whole-half predictions of every half are rescored against the first 30 seconds, 1, 2, 5, 10 and 20 minutes and against the whole half; predictions and annotations outside the window are removed before scoring, and non-finite coordinates are handled as above. The predictions are those of Table 3, so the whole-half window reproduces that table (Supplementary Table 2). This is distinct from the nested-prefix runs of Section 2.3, which rerun the pipeline on each prefix of one half.

For BAS, a prediction at time $\hat{t}$ with class c counts as a true positive against a ground-truth event at time $t^\star$ with the same class if it falls within a window of total width τ centred on $t^\star$, that is $|\hat{t} - t^\star| \leq \tau/2$ (Section 4.9). We report a tight $\tau = 1$ s (maximum offset 0.5 s) and a loose $\tau = 5$ s (maximum offset 2.5 s). Each ground-truth event matches at most one prediction; remaining predictions count as false positives and remaining events as false negatives. Dual events at the same timestamp are treated as two independent targets. The reference evaluator implements the 11-point interpolated AP computation.

The released evaluation tooling computes every metric reported here from prediction files.

4.11 Ethics and consent

All recordings were carried out under informed written consent agreements approved by the ethical committee at the University of Tsukuba, and the consent covers public release and redistribution of the recordings. The release is pseudonymised: personal names are replaced by jersey numbers, which are provided for re-identification and tracking purposes, and the players' real-world identities are not included; because the video and the jersey numbers are released together, individuals may remain identifiable to those who know the teams.

Declarations

Funding. This work was financially supported by JST SPRING Grant Number JPMJSP2125 and JSPS KAKENHI Grant Number 23H03282.

Competing interests. The authors declare no competing interests.

Ethics approval and consent to participate. All recordings were carried out under informed written consent agreements approved by the ethical committee at the University of Tsukuba (see Section 4.11). The public release is pseudonymised: personal names are replaced by jersey numbers, and the players' real-world identities are not included. Because video and jersey numbers are released together, the data are pseudonymised rather than anonymised. The written consent covers public release and redistribution of the recordings.

Data availability. The full package of videos, game state and ball action annotations, the ball track used by the with-ball spotting condition, the hand-annotated calibration keypoints and the curated bounding-box subset (Section 4.5) is hosted on Hugging Face at https://huggingface.co/datasets/atomscott/soccertrack-v2, available after accepting the dataset terms; the code and documentation are at https://github.com/AtomScott/SoccerTrack-v2. The release names matches by the provider's identifiers; its dataset card maps them to the labels M1 to M10 used here. All results in this paper use release v1.2 of the annotations, whose changes from v1.0 are listed in the dataset. The data are released under the CC BY 4.0 licence, which permits academic and commercial use with attribution. Researchers are asked to cite this paper when publishing results.

Code availability. The dataset loader, the evaluation command-line tools for both metrics, and the baseline starter kits are available at https://github.com/AtomScott/SoccerTrack-v2 under the MIT licence. The evaluation tools compute every metric reported here from prediction files, and the modifications to the reconstruction pipeline (Sections 4.7 and 4.8) are documented in the same repository.

Author contributions. A.S.: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing (original draft) and

project administration. I.U., K.K. and Y.K.: data curation. K.F.: conceptualization, supervision, funding acquisition and writing (review and editing). All authors read and approved the final manuscript.

References

[1] Fernández, J. & Bornn, L. *Wide open spaces: A statistical technique for measuring space creation in professional soccer. Proceedings of the 12th MIT Sloan Sports Analytics Conference* (Boston, MA, 2018).

[2] Spearman, W. *Beyond expected goals. Proceedings of the 12th MIT Sloan Sports Analytics Conference* (Boston, MA, 2018).

[3] Decroos, T., Bransen, L., Van Haaren, J. & Davis, J. *Actions speak louder than goals: Valuing player actions in soccer. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 1851–1861 (ACM, 2019).

[4] Fernández, J., Bornn, L. & Cervone, D. A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning* **110**, 1389–1427 (2021).

[5] Robinson, I., Robicheaux, P. & Popov, M. RF-DETR. https://github.com/roboflow/rf-detr (2025). SOTA Real-Time Object Detection Model.

[6] Zhang, Y. *et al. ByteTrack: Multi-object tracking by associating every detection box. European Conference on Computer Vision (ECCV)*, 1–21 (Springer, 2022).

[7] Aharon, N., Orfaig, R. & Bobrovsky, B.-Z. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651* (2022).

[8] Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

[9] Gutiérrez-Pérez, M. & Agudo, A. *No bells, just whistles: Sports field registration by leveraging geometric properties. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2024).

[10] Magera, F., Hoyoux, T., Barnich, O. & Van Droogenbroeck, M. *BroadTrack: Broadcast camera tracking for soccer. 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6177–6187 (IEEE, 2025).

[11] Li, S., Sun, L. & Li, Q. *CLIP-ReID: exploiting vision-language model for image re-identification without concrete text labels. Proceedings of the AAAI conference on artificial intelligence*, Vol. 37, 1405–1413 (2023).

[12] Mansourian, A. M., Somers, V., De Vleeschouwer, C. & Kasaei, S. *Multi-task learning for joint re-identification, team affiliation, and role classification for sports visual tracking. Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, 103–112 (2023).

[13] Somers, V., Alahi, A. & Vleeschouwer, C. D. *Keypoint promptable re-identification. European Conference on Computer Vision*, 216–233 (Springer, 2024).

[14] Bautista, D. & Atienza, R. *Scene text recognition with permuted autoregressive sequence models. European conference on computer vision*, 178–196 (Springer, 2022).

[15] Atienza, R. *Vision transformer for fast and efficient scene text recognition. Document Analysis and Recognition – ICDAR 2021*, Vol. 12821 of *Lecture Notes in Computer Science*, 319–334 (Springer, 2021).

[16] Koshkina, M. & Elder, J. H. *A general framework for jersey number recognition in sports video. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 3235–3244 (2024).

[17] Carreira, J. & Zisserman, A. *Quo vadis, action recognition? A new model and the kinetics dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6299–6308 (2017).

[18] Li, Y. *et al. MViTv2: Improved multiscale vision transformers for classification and detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4804–4814 (2022).

[19] Hong, J., Zhang, H., Gharbi, M., Fisher, M. & Fatahalian, K. *Spotting temporally precise, fine-grained events in video. Computer Vision – ECCV 2022*, Vol. 13695 of *Lecture Notes in Computer Science*, 33–51 (Springer, 2022).

[20] Xarles, A., Escalera, S., Moeslund, T. B. & Clapés, A. *T-DEED: Temporal-discriminability enhancer encoder-decoder for precise event spotting in sports videos. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 3410–3419 (2024).

[21] Xarles, A., Escalera, S., Moeslund, T. B. & Clapés, A. *ASTRA: An action spotting transformer for soccer videos. Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports (MMSports)*, 93–102 (2023).

[22] Somers, V. *et al. SoccerNet game state reconstruction: End-to-end athlete tracking and identification on a minimap. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3293–3305 (2024).

[23] Luiten, J. *et al.* HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision* **129**, 548–578 (2021).

[24] Giancola, S., Amine, M., Dghaily, T. & Ghanem, B. *SoccerNet: A scalable dataset for action spotting in soccer videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1711–1721 (2018).

[25] Deliege, A. *et al. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. 7th International Workshop on Computer Vision in Sports (CVsports) at IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR' 21)*, 4508–4519 (2021).

[26] Cioppa, A. *et al.* SoccerNet 2024 challenges results. *arXiv preprint arXiv:2409.10587* (2024).

[27] Cioppa, A. *et al. Camera calibration and player localization in SoccerNet-v2 and investigation of their representations for action spotting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 4532–4541 (2021).

[28] Cioppa, A. *et al. SoccerNet-Tracking: Multiple object tracking dataset and benchmark in soccer videos. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 3491–3502 (2022).

[29] Cui, Y. *et al. SportsMOT: A large multi-object tracking dataset in multiple sports scenes. Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9921–9931 (2023).

[30] Li, Y. *et al. MultiSports: A multi-person video dataset of spatio-temporally localized sports actions. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 13516–13525 (2021).

[31] Karki, D., Ramazanova, M., Cioppa, A., Giancola, S. & Ghanem, B. Pixels or positions? Benchmarking modalities in group activity recognition. *arXiv preprint arXiv:2511.12606* (2025).

[32] Ochin, J., Chekroun, R., Stanciulescu, B. & Manitsaris, S. FOOTPASS: A multi-modal multi-agent tactical context dataset for play-by-play action spotting in soccer broadcast videos. *Computer Vision and Image Understanding* **269**, 104790 (2026).

[33] Giancola, S. *et al.* SoccerNet 2025 challenges results. *arXiv preprint arXiv:2508.19182* (2025).

[34] Cioppa, A. *et al.* SoccerNet 2026 challenges results. *arXiv preprint arXiv:2607.07320* (2026).

[35] Bassek, M., Rein, R., Weber, H. & Memmert, D. An integrated dataset of spatiotemporal and event data in elite soccer. *Scientific Data* **12**, 195 (2025).

[36] PFF FC. PFF FC 2022 FIFA World Cup dataset: Broadcast tracking and event data. https://www.blog.fc.pff.com/blog/enhanced-2022-world-cup-dataset (2024). All 64 matches; released September 2024, restructured February 2025. Accessed 24 September 2026.

[37] SkillCorner. SkillCorner open data. https://github.com/SkillCorner/opendata (2025). Broadcast tracking data and derived dynamic events for 20 matches of the 2024/2025 Australian A-League (10 before 8 September 2026). Accessed 24 September 2026.

[38] Metrica Sports. Metrica Sports sample data. https://github.com/metrica-sports/sample-data (2020). Synchronised tracking and event data for three anonymised matches. Accessed 24 September 2026.

[39] Pappalardo, L. *et al.* A public data set of spatio-temporal match events in soccer competitions. *Scientific Data* **6**, 1–15 (2019).

[40] Pettersen, S. A. *et al. Soccer video and player position dataset. Proceedings of the 5th ACM Multimedia Systems Conference (MMSys)*, 18–23 (ACM, 2014).

[41] Jiang, T. *et al. WorldPose: A world cup dataset for global 3D human pose estimation. Computer Vision – ECCV 2024*, Vol. 15077 of *Lecture Notes in Computer Science*, 343–362 (Springer, 2024).

[42] Scott, A. *et al. SoccerTrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. 8th International Workshop on Computer Vision in Sports (CVsports) at IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR' 22)*, 3569–3579 (2022).

[43] Scott, A. *et al. TeamTrack: A dataset for multi-sport multi-object tracking in full-pitch videos. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3357–3366 (2024).

[44] Morra, L. *et al.* SoccER: Computer graphics meets sports analytics for soccer event recognition. *SoftwareX* **12**, 100612 (2020).

[45] SoccerTrack Challenge 2025 Organisers. SoccerTrack Challenge 2025. Challenge at the International Workshop on Multimedia Content Analysis in Sports (MMSports), ACM Multimedia 2025, Dublin. https://sites.google.com/g.sp.m.is.nagoya-u.ac.jp/stc2025; evaluation server https://codalab.lisn.upsaclay.fr/competitions/22532 (2025).

[46] Richly, K., Bothe, M., Rohloff, T. & Schwarz, C. *Recognizing compound events in spatio-temporal football data. Proceedings of the International Conference on*

Internet of Things and Big Data (IoTBD), 27–35 (SciTePress, 2016).

[47] Link, D. & Hoernig, M. Individual ball possession in soccer. *PLOS ONE* **12**, e0179953 (2017).

[48] Khaustov, V. & Mozgovoy, M. Recognizing events in spatiotemporal soccer data. *Applied Sciences* **10**, 8046 (2020).

[49] Fassmeyer, D., Anzer, G., Bauer, P. & Brefeld, U. Toward automatically labeling situations in soccer. *Frontiers in Sports and Active Living* **3**, 725431 (2021).

[50] Vidal-Codina, F., Evans, N., El Fakir, B. & Billingham, J. Automatic event detection in football using tracking data. *Sports Engineering* **25**, 18 (2022).

[51] Sanford, R., Gorji, S., Hafemann, L. G., Pourbabaee, B. & Javan, M. *Group activity detection from trajectory and video data in soccer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3932–3940 (2020).

[52] Xu, Y., Zhang, J., Zhang, Q. & Tao, D. ViTPose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484* (2022).

[53] Sun, J., Huang, H.-W., Yang, C.-Y., Jiang, Z. & Hwang, J.-N. *GTA: Global tracklet association for multi-object tracking in sports. Proceedings of the Asian Conference on Computer Vision*, 421–434 (2024).

[54] Cioppa, A. *et al.* SoccerNet 2023 challenges results. *arXiv preprint arXiv:2309.06006* (2023).

[55] Kim, H., Choi, H.-J., Kim, C. J., Yoon, J. & Ko, S.-K. *Ball trajectory inference from multi-agent sports contexts using set transformer and hierarchical Bi-LSTM. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 4296–4307 (ACM, 2023).

[56] Lea, C., Flynn, M. D., Vidal, R., Reiter, A. & Hager, G. D. *Temporal convolutional networks for action segmentation and detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1003–1012 (2017).

[57] Farha, Y. A. & Gall, J. *MS-TCN: Multi-stage temporal convolutional network for action segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3570–3579 (2019).

[58] Bai, S., Kolter, J. Z. & Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

[59] Voeikov, R., Falaleev, N. & Baikulov, R. *TTNet: Real-time temporal and spatial video analysis of table tennis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3866–3874 (2020).

Supplementary Information

Supplementary Tables 1 and 2 carry the run-level detail behind Sections 2.3 and 4.10: the DetA, AssA and LocA components of every attributes-off score of Table 3, and the score of every whole-half run over growing evaluation windows.

Supplementary Table 1: Components of the attributes-off scores of Table 3: the same predictions rescored with attribute matching disabled, so that DetA, AssA and LocA measure geometry and association alone; *Tracklets* is the number of predicted identities and *Identities* the number of annotated identities in the sequence, substitutes included. The upper block is the opening 30 seconds of every half; the lower block is the full half of the same halves. Configuration is as in that table; the released test split is marked †. Means are per column within each block, computed from unrounded scores.

Half	DetA	AssA	LocA	Tracklets	Identities
First 30 s, all twenty halves					
M1, 1st	21.60	35.39	73.05	24	22
M1, 2nd	25.33	39.19	73.59	27	22
M2, 1st	60.06	73.36	85.98	26	22
M2, 2nd	41.67	58.57	82.69	33	22
M3, 1st	57.79	66.38	85.65	25	22
M3, 2nd	55.21	74.29	85.31	37	22
M4, 1st	42.75	56.00	82.22	29	22
M4, 2nd	49.24	71.02	83.66	25	22
M5, 1st	66.28	78.77	88.34	29	22
M5, 2nd	56.65	69.48	85.95	29	22
M6, 1st	50.74	59.57	84.68	27	22
M6, 2nd	50.30	58.85	84.67	28	22
M7, 1st†	73.68	81.63	89.56	26	22
M7, 2nd†	59.42	74.55	87.74	29	22
M8, 1st	68.32	77.64	89.13	23	22
M8, 2nd	57.54	64.63	85.09	23	22
M9, 1st†	26.37	33.67	77.84	28	22
M9, 2nd†	47.53	58.30	79.02	27	22
M10, 1st	53.34	69.67	83.74	30	22
M10, 2nd	33.63	58.19	81.17	29	22
Mean, all twenty halves	49.87	62.96	83.45		
Full half (45 min), all twenty halves					
M1, 1st	15.89	4.55	72.45	127	23
M1, 2nd	17.01	3.27	72.63	137	33
M2, 1st	45.98	10.84	82.28	146	23
M2, 2nd	47.09	10.49	82.84	224	27
M3, 1st	44.90	15.48	82.82	335	22
M3, 2nd	46.01	12.40	83.20	422	29
M4, 1st	46.26	13.24	83.01	141	23
M4, 2nd	42.50	9.71	80.80	142	28
M5, 1st	49.69	12.77	84.01	168	24
M5, 2nd	46.05	13.39	83.64	351	30
M6, 1st	49.43	10.80	84.15	123	22
M6, 2nd	45.71	7.59	82.28	195	30
M7, 1st†	51.21	13.41	84.68	345	23
M7, 2nd†	50.82	12.52	84.84	293	23
M8, 1st	49.15	13.76	84.30	133	23
M8, 2nd	46.16	10.36	82.35	151	32
M9, 1st†	37.53	6.97	78.14	240	22
M9, 2nd†	41.42	7.77	77.61	273	28
M10, 1st	46.43	18.63	83.30	240	22
M10, 2nd	45.55	14.31	84.14	201	35
Mean, all twenty halves	43.24	11.11	81.67		

Supplementary Table 2: GS-HOTA of the whole-half predictions of Table 3 scored over growing windows from kickoff: the first 30 seconds, 1, 2, 5, 10 and 20 minutes, and the whole half (Methods). The predictions are fixed and only the scored window changes, so the last column reproduces Table 3. The upper block is the official metric, the lower block the same windows with attribute matching disabled. Means are per column, computed from unrounded scores; the released test split is marked †.

Half	30 s	1 min	2 min	5 min	10 min	20 min	Whole half
Official GS-HOTA							
M1, 1st	14.28	10.70	8.18	8.60	6.32	6.39	6.68
M1, 2nd	9.90	7.08	5.48	5.77	3.92	4.14	3.60
M2, 1st	29.15	21.02	17.24	17.44	16.87	17.01	14.28
M2, 2nd	25.45	25.33	23.47	26.12	20.70	16.99	12.13
M3, 1st	32.19	29.11	29.82	21.39	17.31	17.16	16.59
M3, 2nd	25.36	26.86	23.33	17.97	16.95	13.62	12.21
M4, 1st	34.05	31.15	28.14	22.10	16.61	13.56	12.36
M4, 2nd	14.24	12.07	11.42	13.42	11.47	10.25	11.03
M5, 1st	42.26	42.84	40.74	30.63	25.23	21.82	18.23
M5, 2nd	29.34	30.77	30.72	23.19	18.63	21.35	16.54
M6, 1st	11.90	13.32	17.28	14.89	11.90	12.87	13.36
M6, 2nd	18.02	13.45	14.28	15.62	13.22	10.46	10.22
M7, 1st†	36.91	30.84	30.07	25.11	20.19	18.78	18.09
M7, 2nd†	24.90	24.64	21.87	26.05	20.12	16.33	12.89
M8, 1st	24.68	19.21	16.22	15.53	11.84	12.61	12.73
M8, 2nd	31.52	29.38	29.95	24.67	18.89	18.39	12.26
M9, 1st†	15.26	13.51	17.97	18.14	14.71	11.73	9.92
M9, 2nd†	5.84	6.60	6.34	4.81	3.73	3.64	2.94
M10, 1st	27.91	28.26	27.51	27.82	27.37	19.64	15.27
M10, 2nd	11.09	9.94	11.13	10.11	12.08	12.56	13.64
Mean, all twenty halves	23.21	21.30	20.56	18.47	15.40	13.97	12.25
Attribute matching disabled							
M1, 1st	26.30	18.58	13.54	12.53	9.75	8.53	8.46
M1, 2nd	31.88	19.71	13.22	10.60	8.67	7.95	7.44
M2, 1st	64.50	39.62	32.07	32.23	29.46	25.93	22.14
M2, 2nd	49.16	52.22	48.28	42.51	35.45	28.05	22.08
M3, 1st	61.77	53.21	53.68	39.81	32.27	28.90	26.09
M3, 2nd	63.89	62.84	57.62	39.41	38.59	29.58	23.71
M4, 1st	48.80	46.50	45.60	37.89	31.49	27.53	24.56
M4, 2nd	58.32	51.66	38.49	31.05	26.37	20.30	20.05
M5, 1st	71.79	71.66	60.85	45.30	37.16	32.67	25.10
M5, 2nd	62.35	65.43	53.94	39.05	31.64	32.20	24.68
M6, 1st	55.68	54.28	57.36	36.43	27.27	23.93	23.00
M6, 2nd	53.17	39.57	38.29	30.78	24.25	21.93	18.49
M7, 1st†	75.81	60.53	54.81	37.18	30.86	29.08	26.04
M7, 2nd†	65.51	61.41	47.03	51.25	38.78	33.67	25.08
M8, 1st	70.38	54.96	44.80	46.51	33.57	27.48	25.89
M8, 2nd	61.06	55.20	54.60	42.95	35.98	31.46	21.65
M9, 1st†	29.54	23.25	23.87	24.52	21.14	18.13	16.15
M9, 2nd†	51.95	46.73	41.16	28.80	24.22	23.07	17.91
M10, 1st	60.83	55.87	51.02	49.13	46.00	38.12	29.26
M10, 2nd	43.87	49.15	51.53	43.05	34.67	28.88	25.40
Mean, all twenty halves	55.33	49.12	44.09	36.05	29.88	25.87	21.66