arXiv:2511.12606v3 [cs.CV] 21 Apr 2026

Pixels or Positions? Benchmarking Modalities in Group Activity Recognition

Drishya Karki[1] Merey Ramazanova[1] Anthony Cioppa[2] Silvio Giancola[1] Bernard Ghanem[1]

[1]King Abdullah University of Science and Technology (KAUST), Saudi Arabia
[2]University of Liège, Belgium

{drishya.karki, merey.ramazanova, silvio.giancola, bernard.ghanem}@kaust.edu.sa
anthony.cioppa@uliege.be

Abstract

*Group Activity Recognition (GAR) is well studied on the video modality for surveillance and indoor team sports (*e.g.*, volleyball, basketball). Yet, other modalities such as agent positions and trajectories over time,* i.e. *tracking, remain comparatively under-explored despite being compact, agent-centric signals that explicitly encode spatial interactions. Understanding whether pixel (video) or position (tracking) modalities leads to better group activity recognition is therefore important to drive further research on the topic. However, no standardized benchmark currently exists that aligns broadcast video and tracking data for the same group activities, leading to a lack of apples-to-apples comparison between these modalities for GAR. In this work, we introduce* SoccerNet-GAR*, a novel multimodal dataset curated from the* 64 *matches of the football World Cup 2022 by cleaning, synchronizing, and transforming raw broadcast video, player tracking, and event annotation sources into a unified, research-ready benchmark. Specifically, the broadcast videos and player tracking modalities for* 87,939 *group activities are synchronized and annotated with* 10 *categories. Furthermore, we define a unified evaluation protocol to benchmark two strong unimodal approaches: (i) competitive video-based classifiers and (ii) tracking-based classifiers leveraging graph neural networks. In particular, our novel role-aware graph architecture for tracking-based GAR directly encodes tactical structure through positional edges connecting players by their on-pitch roles. Our tracking baseline achieves* 77.8% *balanced accuracy and* 57.0% *macro F1 compared to* 60.9% *balanced accuracy and* 50.1% *macro F1 for the strongest video model, while training with* 7× *less GPU hours and* 479× *fewer parameters (*180*K* vs. 86.3*M). This study provides new insights into the relative strengths of pixels and positions for group activity recognition in sports.*

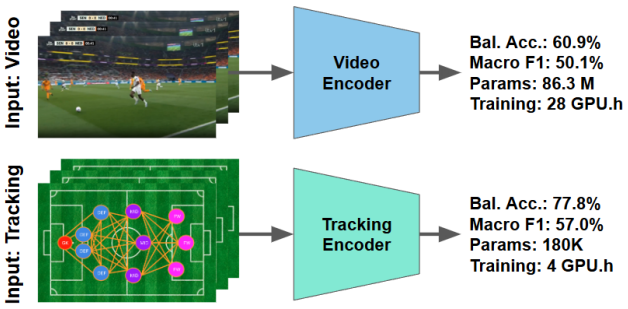


Figure 1. **Overview of Our Group Activity Recognition Benchmark.** Broadcast video and agent tracking modalities are processed through modality-specific backbones. The resulting representations are evaluated for group activity recognition.

1. Introduction

Group Activity Recognition (GAR) aims to identify collective behaviors that arise from multiple interacting agents rather than isolated individual actions. In team sports, this task is central to tactical analysis, automated coaching, and real-time decision support, where understanding both *who* acts and *how* players coordinate over space and time is essential. Effective GAR requires modeling individual dynamics together with their spatio-temporal relations at the scene level. Most contemporary GAR models operate exclusively on RGB video, which offers appearance, context, and rich semantics. This choice also introduces practical and methodological limitations such as high computational cost, sensitivity to occlusions and editorial cuts, and a tendency to over-rely on appearance cues instead of explicitly modeling interactions. Structured representations based on agent positions and trajectories promise an alternative representation that is compact, agent-centric, and efficient to process. Because positions abstract away appearance, they may also generalize better across varying visual conditions. Progress, however, has been hampered by the absence of a standardized benchmark that aligns videos and agent posi-

tions per group activity, making apples-to-apples comparisons difficult. Furthermore, it remains unclear which edge connectivity and temporal pooling strategy are most appropriate for tracking based GAR.

In this paper, we address these gaps with a unified dataset and evaluation protocol for fair cross-modality evaluation. To do so, we introduce *SoccerNet-GAR*, a novel multimodal dataset and benchmark built from the 64 matches of the football World Cup 2022 by cleaning, merging, and aligning raw broadcast video, player tracking, and event annotation files from PFF FC (now Gradients Sports) into a synchronized, research-ready benchmark. The resulting dataset contains 87,939 group activity instances, each pairing broadcast video and agent positions at the event level across 10 football action classes. To support this new benchmark, we study two strong *unimodal* approaches: a competitive video-based and a novel role-aware tracking-based classifier leveraging graph neural networks. On the 10-class group activity recognition task, our tracking baseline achieves 77.8% balanced accuracy and 57.0% macro F1 *vs*. 60.9% balanced accuracy and 50.1% macro F1 for the strongest video baseline, while training 7× faster and using 479× fewer parameters.

Contributions. We summarize our contributions as follows: **(i)** We present *SoccerNet-GAR*, a novel multimodal benchmark we curated from 64 football World Cup matches by cleaning, merging, and temporally aligning separate raw files for broadcast video, player tracking, and event annotations. The resulting benchmark provides synchronized broadcast video and tracking data for 87,939 events across 10 classes, with match-level splits for fair, reproducible evaluation. **(ii)** We propose a *novel role-aware, graph-based classifier for GAR* from tracking data, where tactical roles extracted from the source data define the graph connectivity, and conduct a comprehensive study of spatial connectivity and temporal aggregation design choices. **(iii)** We establish a *pixels* vs. *positions* benchmark with identical protocols, report performance and efficiency comparisons, and distill practical design principles for tracking-based GAR. The dataset, protocol, and code is available on https://github.com/drishyakarki/pixels_vs_positions.

2. Related Work

2.1. Group Activity Recognition

Group Activity Recognition (GAR) studies collective behaviors that arise from coordinated interactions among multiple agents, going beyond single-actor recognition [6]. Early works relied on hand-crafted relational cues and hierarchical temporal models [2, 6, 23]. The field then shifted toward deep learning approaches with graph-based architectures that explicitly encode inter-person dependencies [13, 49, 52], followed by transformer-based methods that model long-range spatial-temporal relationships through self-attention [16, 31]. Popular benchmarks include "Collective Activity" and "Volleyball", which focus on indoor team sports with controlled camera viewpoints [6, 23]. Contemporary methods predominantly operate on RGB clips using 3D CNNs or vision transformers [37, 39, 42, 43], often combined with person detection, region pooling, or attention over actor tokens to model interactions. Recent work has advanced relational modeling further: higher-order interactions through factor graphs [46] and bidirectional causality between human relations and human-object interactions [53]. While these approaches achieve strong performance, they remain video-centric, and their relational reasoning is mediated through appearance features extracted from pixels. Several works explored structured alternatives other than raw videos. Pose- and skeleton-based approaches explicitly model spatial relationships between joints and have shown that structured representations can reduce reliance on background and scene bias while remaining competitive [32, 54]. More broadly, tracking data from sports analytics systems offers a complementary modality by providing agent-centric trajectories and role metadata in a calibrated field coordinate frame, enabling explicit modeling of team structure and inter-agent geometry. Despite being compact, naturally encoding interactions, and robust to visual distribution shifts, tracking-based representations have received limited attention for GAR, and no existing benchmark evaluates these modalities under synchronized, comparable conditions. In our work, we address this gap by providing a fair, apples-to-apples evaluation of *unimodal* pixels versus *unimodal* positions under identical training protocols and matched event-level synchronization.

2.2. Sports Video & Tracking Analytics

Sports video understanding has advanced significantly with broadcast analysis spanning many sports including handball [22], basketball [24], and football [18]. The SoccerNet [17] and SoccerNet-v2 [11] datasets established large-scale benchmarks for action spotting, expanding from 3 to 17 action classes with precise temporal annotations for every event during a game. Recent extensions include SoccerNet-v3 [8] with multi-view player localization annotations, SoccerNet Ball Action Spotting [10], SoccerNet Game State Reconstruction [40], as well as many others [9, 15, 29, 38]. In parallel to broadcast video analysis, structured tracking signals such as agent centroids and trajectories have been widely used in sports analytics for forecasting, pass modeling, and event detection [5, 7, 14, 51]. Graph-based methods naturally capture coordinated play by representing players and the ball as nodes with relational edges [1, 41, 45, 50]. However, these studies typically address trajectory forecasting [50], pass and defensive performance modeling [14, 41], tactical pattern detection [45]

rather than multi-class group activity recognition under synchronized video-tracking alignment. The closest work is NETS [21], which recognizes group activities from NBA tracking data using a Transformer-LSTM architecture, but it operates on a single modality without a matched video counterpart for cross-modality comparison. Our work addresses this gap by proposing the first synchronized dataset where broadcast video and tracking windows correspond to the same annotated group activities, and leveraging it to directly compare pixel-based and position-based representations under identical evaluation settings.

2.3. Graph and Relational Modeling

Graph neural networks enable relational reasoning for multi-agent scenes by propagating messages between agents [4, 20, 27, 47]. When applying such models to multi-agent activity recognition, two design axes are particularly consequential: graph connectivity, which determines how agents exchange information, and temporal aggregation, which governs how frame-level representations are consolidated into a sequence-level prediction. EdgeConv [44] enables learned, adaptive edges through dynamic neighborhood construction rather than fixed graphs. Training deeper GCNs has been challenging due to oversmoothing, but DeepGCNs [30] show that residual connections enable much deeper graph networks. Temporal modules range from efficient temporal convolutional networks [28] to recurrent [19] or attention-based aggregators for long-range context [16]. While GNNs have been applied to sports tasks such as trajectory forecasting and action rating [12, 34, 51], prior studies typically fix parts of the modeling pipeline or evaluate on limited data, making it difficult to extract modality-specific best practices. We complement this literature with a novel role-aware graph classifier for tracking-based GAR and conduct ablations across connectivity schemes (fixed, distance-based, positional edges) and temporal aggregation strategies (pooling, TCN, attention) to identify which design choices yield optimal accuracy-efficiency trade-offs.

3. SoccerNet-GAR Dataset

We present **SoccerNet-GAR**, a new multimodal dataset for Group Activity Recognition constructed from the 64 matches of the football World Cup 2022 tournament. The raw data originate from PFF FC [36], which provides broadcast videos, player tracking data, and event annotations for each match. These modalities are distributed as separate files per match, not directly aligned for computer vision tasks. We transform this raw sports analytics data into a structured benchmark suitable for Group Activity Recognition by performing data cleaning, temporal synchronization, quality control, and window extraction around annotated events.

3.1. Data Modalities

Our dataset provides two input modalities: broadcast video and player tracking. Both originate from untrimmed full-game broadcast videos of the football World Cup 2022 matches, made available through PFF FC.

Video Modality. PFF FC provides the edited broadcast footage at 720p resolution. Each frame $I_t \in \mathbb{R}^{H\times W\times 3}$ is part of a temporal sequence $\mathbf{X}^V = \{I_t\}_{t=1}^{T}$ sampled within a temporal window. The video captures appearance cues, scene context, and visual motion patterns.

Tracking Modality. The tracking data consists of 2D player positions and 3D ball coordinates sampled at 30 fps, automatically extracted from broadcast footage using computer vision tools and manually refined by PFF FC annotators. Player positions span $x \in [-60, 60]$m, $y \in [-42, 41]$m; ball positions include height $z \in [-8, 25]$m. Formally, the tracking input is represented as a temporal sequence of multi-agent states $\mathbf{X}^P = \{\mathcal{S}_t\}_{t=1}^{T}$, where each state $\mathcal{S}_t = \{s_t^1, \ldots, s_t^N\}$ describes N entities (players and ball). Each entity state $s_t^i \in \mathbb{R}^D$ encodes spatial coordinates $(x, y) \in \mathbb{R}^2$ for players and $(x, y, z) \in \mathbb{R}^3$ for the ball, its identity, *i.e.* one-hot encoding indicating entity type, and motion dynamics, *i.e.* displacement vectors $(\Delta x, \Delta y)$ capturing movement between consecutive frames. In addition, PFF FC provides positional role metadata for each player. We extract these roles and group them into four tactical categories (goalkeeper, defender, midfielder, forward) that we use as fixed role labels to define graph connectivity. These role categories are used to define structured connectivity patterns in our role-aware graph model.

3.2. Annotation & Modality Alignment

The PFF FC dataset [36] provides event annotations with precise timestamps for each action occurrence. Each event is labeled with one of 10 football group activities (*PASS*, *TACKLE*, *OUT*, *HEADER*, *THROW IN*, *CROSS*, *FREE KICK*, *SHOT*, *GOAL* and *HIGH PASS*) and temporally marked at the moment of occurrence. Annotations were created by trained annotators and verified through quality control procedures by PFF FC using both video and tracking views. To create our SoccerNet-GAR dataset, we align event annotations with both input modalities by merging them with tracking streams using UTC timestamps. We then apply three successive filters to ensure data quality: (1) we remove events where no tracking frame falls within a 10 ms tolerance of the event timestamp, discarding temporally misaligned entries, (2) we discard events that lack corresponding data in either modality, ensuring every retained sample has both video and tracking coverage, and (3) we resolve duplicate labels, where a single timestamp is annotated with more than one action class (*e.g.*, a goal also labeled as a shot), by retaining only the most semantically specific label based on a predefined priority ordering. To-

Dataset	Year	Domain	Events	Cls.	Mod.
CAD [6]	2009	Pedestrian	2,511	5	V
Volleyball [23]	2016	Volleyball	4,830	8	V
SoccerNet [17]	2018	Football	6,637	3	V
NBA [48]	2020	Basketball	9,172	9	V
SoccerNet-v2 [11]	2021	Football	110,458	17	V
NETS [21]	2022	Basketball	61,053	3	T
SoccerNet-BAS [10]	2024	Football	11,041	12	V
Café [25]	2024	Indoor	10,297	6	V
FIFAWC [35]	2024	Football	5,196	12	V
SoccerNet-GAR	**2026**	**Football**	**87,939**	**10**	**V + T**

Table 1. **Comparison of Group Activity Recognition Datasets.** 'V' and 'T' indicate Video and Tracking, respectively. SoccerNet-GAR is the second largest dataset in terms of events and the only one providing synchronized video and tracking modalities.

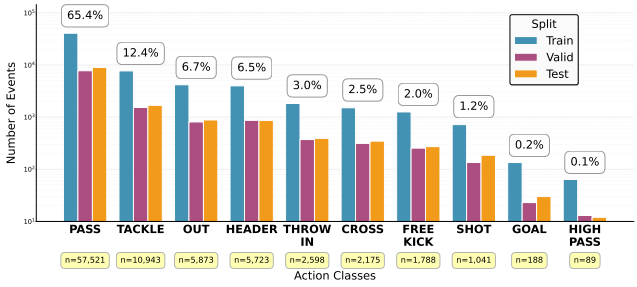


Figure 2. **Class Distribution Across Each Split.** The dataset is heavily skewed toward *PASS* (65.4%) while *GOAL* and *HIGH PASS* are rare (0.2% and 0.1%). The y-axis shows event counts; percentages denote proportions.

gether, these filters remove 6,346 events (6.8% of the raw annotations), yielding a final dataset of 87,939 annotated group activities. For each annotated event at timestamp t_e, we extract a temporal window of 4.5 seconds centered at t_e from both modalities, picking $T = 16$ samples at 30 fps with a 9-frame interval (effectively sampled at ≈ 3.3 fps over 4.5 seconds), producing synchronized observations $\mathbf{X}^V = \{I_t\}_{t=1}^T$ and $\mathbf{X}^P = \{\mathcal{S}_t\}_{t=1}^T$, labeled with a single football group activity.

3.3. Statistics

The dataset contains 87,939 annotated events across 10 action classes, averaging 1,374 events per match. To avoid leakage, data were split by match: 45 training matches (62,159 events, 70.7%), 9 validation matches (12,091 events, 13.7%), and 10 test matches (13,689 events, 15.6%).

Class Distribution. Figure 2 shows the distribution across classes and splits. Our dataset is highly imbalanced across classes, with *PASS* events representing 65.4% of annotations while rare events like *GOAL* and *HIGH PASS* account for only 0.2% and 0.1%. This severe imbalance (646:1 ratio) reflects the natural distribution of football events and poses significant challenges for recognition models.

Object Visibility and Tracking Quality. The tracking data maintain high completeness for players (99.9% of 1,485,008 frames contain all 11 players per team) whereas ball is visible in 93.4% of frames, due to occlusions, camera transitions, and extreme zoom levels. At the event level, 85.9% of annotated events have complete ball tracking within their temporal window, while the remaining 14.1% contain at least one frame with missing ball data.

3.4. Comparison with Existing Datasets

Table 1 compares our dataset with existing Group Activity Recognition benchmarks. With 87,939 annotated events, SoccerNet-GAR is the second largest GAR dataset after SoccerNet-v2 (110,458 events), and substantially larger than other football datasets: 8× larger than SoccerNet-BAS (11,041 events) and 16.9× larger than FIFAWC (5,196 events). Our 10 action classes overlap with the 12 classes in SoccerNet-BAS, covering the same core football events. SoccerNet-GAR is the only dataset providing synchronized tracking and video modalities for the same action instances, enabling direct apples-to-apples comparison between tracking-based and video-based approaches. This dual-modality design distinguishes our dataset from prior work such as SoccerNet-v2 [11] and SoccerNet-BAS [10] which provide only video, while existing tracking datasets (*e.g.* NETS [21]) lack corresponding broadcast video.

4. Methodology

4.1. Problem Formulation

We formulate the Group Activity Recognition (GAR) problem as a multi-class classification task over temporal windows of length T, centered on annotated event timestamps. Given input from one modality, either video frames $\mathbf{X}^V = \{I_t\}_{t=1}^T$ or tracking data $\mathbf{X}^P = \{\mathcal{S}_t\}_{t=1}^T$, the task is to learn a function $f_\theta : \mathcal{X} \to \mathcal{Y}$ that predicts the activity label $y \in \mathcal{Y}$.

4.2. Recognition Framework

Our recognition framework follows a standard backbone-neck-head architecture [3], where modality-specific backbones extract spatial features, the neck temporally aggregates features to capture dynamics, and a classification head produces final predictions.

Video Backbones. For the video modality, we employ pretrained vision transformers [42, 43] that encode spatiotemporal information from broadcast footage. Given an input clip $\mathbf{X}^V = \{I_t\}_{t=1}^T$, the backbone processes each frame I_t through transformer layers with self-attention mechanisms, producing contextualized token embeddings. We extract frame-level representations $\mathbf{z}_t^V \in \mathbb{R}^{d_V}$ at each timestep t, yielding a temporal sequence $\{\mathbf{z}_t^V\}_{t=1}^T$ for subsequent temporal aggregation.

Tracking Backbone. For the tracking modality, we propose a graph-based architecture that explicitly models spatial relationships among entities at each timestep. At each frame t, we construct a graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ from the tracking data $\mathbf{X}^P = \{\mathcal{S}_t\}_{t=1}^T$, where nodes $\mathcal{V}_t = \{v_t^1, \ldots, v_t^N\}$ represent N entities and edges $\mathcal{E}_t$ encode their spatial interactions. Each node carries an 8-dimensional feature vector comprising its pitch coordinates (x, y), instantaneous velocity $(\Delta x, \Delta y)$, ball height z (zero for players), and a one-hot indicator of entity type (home, away, ball). A graph neural network processes these graphs to produce frame-level embeddings $\mathbf{z}_t^P \in \mathbb{R}^{d_P}$.

Classification Task. Given frame-level representations $\{\mathbf{z}_t\}_{t=1}^T$, the temporal aggregation module synthesizes information across the observation window to produce a unified representation $\hat{\mathbf{z}} = f_{\text{temp}}(\mathbf{z}_1, \ldots, \mathbf{z}_T)$, where f_{temp} is a temporal aggregation function. We evaluate various temporal aggregation strategies spanning from parameter-free pooling operations to learnable sequential models. A Multi-Layer Perceptron (MLP) maps the aggregated representation to class logits through a softmax activation. We follow a standard training protocol across modalities, minimizing a cross-entropy loss and randomly sampling $4,000$ actions per class and per epoch to address class imbalance.

4.3. Graph Construction

The edge set $\mathcal{E}_t$ defines which entities exchange information during message passing. We evaluate several connectivity schemes encoding different spatial interaction priors: *no edges* (independent node processing), *full connectivity* (all nodes linked), *distance-based* (edges within radius r), *KNN* (k nearest neighbors), *ball-centric* variants of distance and KNN (anchored on ball position), and our proposed *positional* connectivity described below. These variants enable identifying which graph topologies effectively capture coordinated play.

Positional Roles. Our proposed connectivity scheme respects tactical structure based on playing positions. Players are grouped by role (goalkeeper, defender, midfielder, forward), as shown in Figure 3, using positional metadata from the tracking data. Within each team, edges connect adjacent tactical lines (goalkeeper to defenders, defenders to midfielders, midfielders to forwards), and the ball connects to all entities when present. This hierarchical connectivity encodes football-specific knowledge about team formations and plausible passing lanes.

Missing Data. Missing entities are assigned sentinel coordinates and excluded from graph message passing, allowing the model to leverage available spatial configurations and motion patterns even when ball tracking is unavailable.

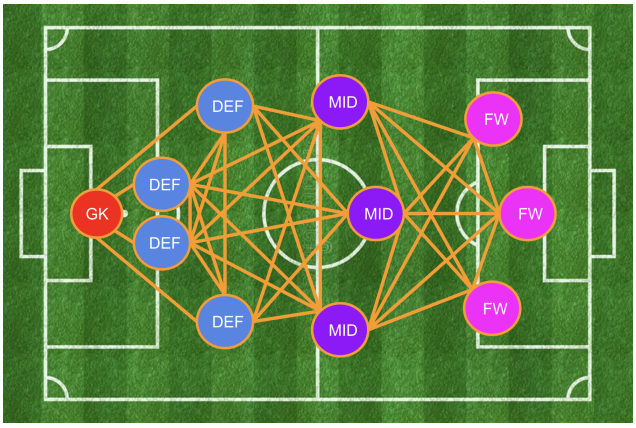


Figure 3. **Positional Graph Representation at a Single Frame.** Nodes represent players at their pitch coordinates, colored by tactical role (red: goalkeeper, blue: defenders, purple: midfielders, pink: forwards). Edges follow formation structure, connecting positionally adjacent roles within each team.

5. Experiments

5.1. Experimental Setup

We benchmark various models on SoccerNet-GAR.

Tracking-based models: As a baseline tracking model, we employ a 20-layer DeepGCN using GIN [47] layers with sum aggregation, ReLU, residual connections, and layer normalization. Each node projects 8-dimensional entity features to a 128-dimensional embedding. Frame-level graph representations are obtained via mean pooling and temporally aggregated with MaxPool. Edges follow a positional role-based scheme connecting players by tactical roles (goalkeeper, defender, midfielder, forward). The classifier is a two-layer MLP (hidden dimension 256, output dimension 10), totaling 180K parameters. Missing entities use sentinel coordinates (-2.0 after normalization) and are excluded from graph message passing.

Video-based models: The best video model is VideoMAEv2-Base [43], a large transformer pretrained via masked autoencoding and finetuned end-to-end. The classifier is a two-layer MLP (hidden dimension 512, output dimension 10), totaling 86.3M parameters.

Training. Tracking models are trained with Adam [26] (LR=10^{-3}), a ReduceLROnPlateau scheduler, and BS=32. For the video models, we use AdamW [33] with weight decay 10^{-4}. Training configurations are defined based on whether the backbone is frozen or fine-tuned, with batch sizes chosen to satisfy memory constraints on a single V100 GPU. Frozen backbones are trained with BS=64 and LR=10^{-4}. Fine-tuned models are trained with BS=16 and LR=10^{-4}, except for VideoMAEv2-B which uses BS=8 and LR=5×10^{-5}. Weighted random sampling draws $M = 4{,}000$ samples per class per epoch to address class

Modality	Params	Bal. Acc.	F1	Training
Video	86.3M	60.9	50.1	28 GPU hours
Tracking	180K	**77.8**	**57.0**	4 GPU hours

Table 2. **Best Video vs. Baseline Tracking Model on SoccerNet-GAR.** Tracking outperforms video by 16.9 percentage points in balanced accuracy and 6.9 percentage points in macro F1 while having 479× fewer parameters and 7× faster training time.

Backbone	Frozen	Params	Bal. Acc.	F1
VideoMAE-B	✓	86.3M	34.6	20.6
VideoMAE-B	✗	86.3M	55.2	49.4
VideoMAEv2-B	✓	86.3M	49.3	30.7
VideoMAEv2-B	✗	**86.3M**	**60.9**	**50.1**

Table 3. **Video Backbone Ablation.** Finetuning yields substantial gains for both backbones, with VideoMAE showing the largest improvement (+20.6%). VideoMAEv2-B achieves the best performance across both frozen and finetuned regimes.

imbalance. Data augmentation includes team flip, spatial flips, and temporal jittering for tracking; horizontal flip and color jitter for video. Gradient clipping ($max_norm = 1.0$) stabilizes training.

Metrics. Following common practices, we report balanced accuracy (also corresponding to the average per-class recall) and macro F1 score as primary benchmarking metrics.

5.2. Main Results: Pixels vs Positions

Table 2 presents our primary comparison between the tracking and video modalities. Our baseline tracking model (GIN + Maxpool + Positional edges) achieves 77.8% balanced accuracy and 57.0% macro F1 using only 180K parameters, training in approximately 4 hours on a single V100 GPU. On the other hand, the best video baseline (VideoMAEv2-B finetuned) achieves 60.9% balanced accuracy and 50.1% macro F1 but requires 86.3M parameters and 28 GPU hours of training. The tracking model outperforms the video baseline by 16.9 percentage points in balanced accuracy and 6.9 percentage points in macro F1 while using 479× fewer parameters and training with 7× less GPU hours. This efficiency stems from the compact representation of tracking data, where each frame consists of 23 nodes with 8-dimensional features, compared to $224 \times 224 \times 3$ RGB tensors per video frame. All tracking results report mean ± standard deviation over 5 random seeds; video results are from single run due to computational constraints.

Per-class Analysis. Figure 4 compares per-class metrics for balanced accuracy (per-class recall) and macro F1 (per-class F1 score) across modalities. Tracking outperforms video on nine of ten classes, with particularly large margins on events that occur in characteristic pitch regions and induce structured player formations: *GOAL* (73.3% vs 16.7%), *HIGH PASS* (83.3% vs 41.7%), *TACKLE* (54.0% vs 32.2%), and *OUT* (94.2% vs 75.8%). Video shows a slight advantage only on *HEADER* (66.3% vs 65.2%). *TACKLE* remains challenging for both modalities (54.0% tracking, 32.2% video), as tackles involve rapid, localized interactions between two players that are difficult to distinguish from other close-contact situations in either representation. Both modalities exhibit confusion between *PASS* with *FREE KICK* (498 and 470 misclassifications of *FREE KICK* as *PASS* respectively for the tracking and video modality), likely because both involve a single player initiating ball distribution from a stationary position, producing similar spatial configurations despite the visual distinction of a set-piece setup. Similarly, *PASS* and *TACKLE* as well as *TACKLE* and *HEADER* exhibit similar inter-class confusion for both modalities. Finally, the results for *GOAL* (30 test samples) and *HIGH PASS* (12 test samples) demonstrate that tracking models can learn discriminative features even in severely data-scarce regimes (73.3% and 83.3% respectively), whereas video models collapse on these classes (16.7% and 41.7%).

5.3. Ablation Studies

We evaluate design choices for both modalities. For tracking, we compare graph layer types, temporal aggregation strategies, and edge connectivity. For video, we compare different backbone architectures, frozen and finetuned.

Video Backbone Variants. Table 3 compares two video-pretrained backbones, VideoMAE-B and VideoMAEv2-B, under two training regimes: frozen backbone with a trainable classification head versus full end-to-end finetuning. Among frozen backbones, VideoMAEv2-B performs best (49.3%), compared to VideoMAE-B (34.6%). Finetuning substantially improves both models: VideoMAE-B gains +20.6 pp (34.6% → 55.2%) and VideoMAEv2-B gains +11.6 pp (49.3% → 60.9%). VideoMAEv2-B consistently achieves the best performance across both regimes, likely due to its larger and more diverse pretraining data compared to VideoMAE-B.

Graph Layer Variants. Table 4 compares six graph convolutional operators with MaxPool temporal aggregation and positional edges in 20-layer DeepGCN architectures with residual connections and layer normalization. GIN achieves the highest balanced accuracy ($77.8 \pm 0.7\%$, 180K parameters), very similar to GraphConv ($76.3 \pm 1.1\%$, 174K). The near-identical performance suggests that positional edge structure provides most of the discriminative signal, with the choice of message-passing operator having minimal impact. More expressive operators underperform: EdgeConv ($55.6 \pm 12.1\%$) and GATv2 ($61.8 \pm 5.4\%$) show both lower accuracy and higher variance. GEN (72.8%) and Graph-

True Label

Class	#	TRACKING (Predicted) P	T	O	H	TI	C	FK	S	G	HP	VIDEO (Predicted) P	T	O	H	TI	C	FK	S	G	HP
Pass	9,009	**7309**	364	195	167	110	128	498	109	5	124	**6994**	585	170	419	150	142	470	53	2	24
Tackle	1,690	295	**912**	171	169	19	33	29	49	5	8	463	**545**	129	442	13	45	13	36	3	1
Out	884	1	17	**833**	4	4	0	14	1	10	0	64	60	**670**	37	9	2	9	23	9	1
Header	867	20	152	41	**565**	24	5	10	24	9	17	156	53	25	**575**	8	14	8	25	3	0
Throw In	392	0	1	10	9	**330**	4	37	0	0	1	17	4	6	2	**308**	3	52	0	0	0
Cross	347	7	8	4	4	1	**301**	4	13	3	2	22	9	5	12	0	**268**	8	21	2	0
Free Kick	272	4	2	3	5	8	1	**246**	2	0	1	27	2	4	0	22	1	**216**	0	0	0
Shot	186	5	2	3	4	0	3	1	**142**	26	0	5	10	5	18	0	23	1	**118**	6	0
Goal	30	1	0	2	1	0	0	1	3	**22**	0	0	0	5	4	0	2	1	13	**5**	0
High Pass	12	0	0	0	2	0	0	0	0	0	**10**	2	0	1	1	0	0	3	0	0	**5**
Bal. Acc.		81.1	54.0	94.2	65.2	84.2	86.7	90.4	76.3	73.3	83.3	77.6	32.2	75.8	66.3	78.6	77.2	79.4	63.4	16.7	41.7
F1		88.9	59.1	81.4	61.9	77.1	75.4	60.3	58.4	47.0	15.6	83.5	36.8	70.4	48.4	68.3	63.3	41.0	49.7	16.7	23.3

Figure 4. **Confusion Matrices and Per-class Metrics for our Baseline Tracking (GIN + MaxPool + Positional) and the Best Video (VideoMAEv2-B Finetuned) Models.** Green cells indicate the best modality per action class. Tracking dominates on 9 of 10 classes in balanced accuracy, with its largest gains on less occurring classes like *GOAL* (+56.7 pp), and *HIGH PASS* (+41.7 pp).

Backbone	Params	Bal. Acc.	F1
GraphConv	174K	76.3 ± 1.1	56.5 ± 1.8
EdgeConv	174K	55.6 ± 12.1	35.3 ± 11.9
GATv2	177K	61.8 ± 5.4	42.7 ± 4.5
GIN	**180K**	$\mathbf{77.8 \pm 0.7}$	$\mathbf{57.0 \pm 0.9}$
GEN	341K	72.8 ± 3.0	54.7 ± 3.1
GraphSAGE	421K	75.9 ± 1.1	56.9 ± 1.7

Table 4. **Graph Layer Ablation.** All models use temporal maxpool and positional edges. We report mean ± std over 5 seeds. GIN achieves the highest balanced accuracy and macro F1 with the lowest variance across both metrics (bold). GraphConv and GraphSAGE perform comparably but with higher variance.

Edge Type	Params	Bal. Acc.	F1
No Edges	180K	68.9 ± 2.7	47.0 ± 3.1
Fully Connected	180K	71.4 ± 2.4	49.8 ± 1.6
Ball Distance ($r = 20$m)	180K	68.6 ± 1.0	45.3 ± 1.3
Ball KNN ($k = 8$)	180K	67.0 ± 0.9	45.9 ± 1.0
Distance ($r = 15$m)	180K	68.0 ± 1.1	45.4 ± 1.3
KNN ($k = 8$)	180K	66.7 ± 1.4	46.4 ± 1.6
Positional	**180K**	$\mathbf{77.8 \pm 0.7}$	$\mathbf{57.0 \pm 0.9}$

Table 5. **Edge Connectivity Ablation.** All models use GIN backbone and temporal MaxPool. We report mean ± std over 5 seeds. Positional edges substantially outperform all alternatives in both metrics, leading the second-best edge type (Fully Connected) by +6.4% in balanced accuracy and +7.2% in F1.

SAGE (75.9%) are competitive but require, respectively, 1.9× and 2.3× more parameters. GIN and GraphSAGE also achieve the best F1 (57.0% and 56.9%) with the lowest variance across both metrics for GIN, confirming its suitability as a strong tracking baseline.

Edge Connectivity. Table 5 compares seven connectivity patterns with GIN layers and MaxPool aggregation. Positional edges achieve the best balanced accuracy ($77.8 \pm 0.7\%$) and F1 ($57 \pm 0.9\%$), with the lowest variance on both metrics. No-edges processing ($68.9 \pm 2.7\%$ balanced accuracy, $47.0 \pm 3.1\%$ F1) performs comparably to alternatives, confirming that message passing with poorly chosen edges provides little benefit over independent node processing. All geometric patterns cluster between 66.7%–77.8% balanced accuracy, suggesting that unrestricted or proximity-based edges dilute the discriminative structure that posi-

Temporal	Params	Bal. Acc.	F1
AvgPool	180K	66.6 ± 0.9	42.5 ± 1.5
MaxPool	180K	$\mathbf{77.8 \pm 0.7}$	57.0 ± 0.9
Attention	197K	77.3 ± 0.9	58.3 ± 1.2
TCN	205K	75.5 ± 0.3	$\mathbf{58.4 \pm 0.6}$
BiLSTM	350K	75.3 ± 2.6	57.9 ± 1.0

Table 6. **Temporal Aggregation Ablation.** All models use GIN backbone and positional edges and we report mean ± std over 5 seeds. MaxPool achieves the highest balanced accuracy (77.8%) while Attention and TCN achieve higher macro F1 (58.3% and 58.4%). The strong MaxPool performance indicates that simple temporal pooling remains competitive with learnable alternatives.

tional connectivity preserves.

Temporal Aggregation. Table 6 further extends our baseline model with more advanced temporal aggregation strategies still using GIN layers and positional edges over the 4.5-second observation window at the cost of higher number of parameters. MaxPool achieves the highest balanced accuracy ($77.8 \pm 0.7\%$, 180K parameters), outperforming the learnable alternatives: Attention ($77.3 \pm 0.9\%$, 197K), TCN ($75.5 \pm 0.3\%$, 205K), and BiLSTM ($75.3 \pm 2.6\%$, 350K), while AvgPool lags behind ($66.6 \pm 0.9\%$, 180K).

However, TCN, Attention, and BiLSTM yield slightly higher macro F1 scores (58.4%, 58.3% and 57.9% vs. 57.0%), suggesting that learnable temporal modules better balance per-class recall despite lower overall balanced accuracy. We retain MaxPool as the default temporal aggregation for its simplicity, parameter efficiency, and strong balanced accuracy. Nonetheless, the competitive F1 gains from learnable modules indicate that the tracking modality backbone has room for improvement. Exploring more expressive temporal and spatio-temporal architectures intertwining the tracking backbone encoders with the temporal neck represents a promising direction for future work.

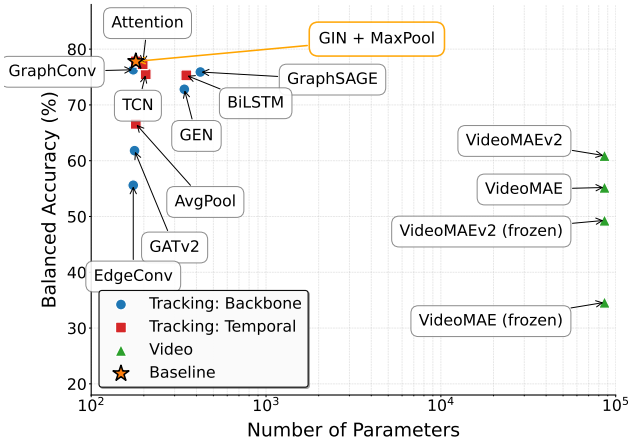


Figure 5. **Balanced Accuracy *vs*. Parameter Count.** Tracking models (174K–421K parameters) consistently outperform video models (86.3M parameters) in balanced accuracy despite using 200-496× fewer parameters, highlighting that lightweight graph-based representations of player dynamics capture more discriminative signal for group activity recognition than large-scale video encoders.

6. Analysis

Model Parameters vs. Balanced Accuracy. Figure 5 compares parameter efficiency across all evaluated models. Tracking models achieve 55.6%–77.8% balanced accuracy with 174K–421K parameters, while video models span 34.6%–60.9% with 86.3M parameters. Tracking models exhibit clear clustering by component type. Backbone variants span 174K–421K parameters with 55.6%–77.8% accuracy, where GraphSAGE (421K) uses 2.3× more parameters than GIN (180K) but underperforms by 1.9%. Temporal aggregation methods range from 180K to 350K parameters with 66.6%–77.8% accuracy, where MaxPool (180K) achieves the highest accuracy at 77.8%, surpassing BiLSTM (350K, 75.3%) despite using nearly half the parameters. Video models cluster at ∼86.3M parameters with accuracy varying by 26.3 percentage points (34.6%–60.9%). This large spread stems from pretraining strategy and frozen *vs*. finetuned regimes rather than architectural differences. VideoMAEv2-B frozen (86.3M, 49.3%) and VideoMAE frozen (86.3M, 34.6%) have identical parameter counts but differ by 14.7 percentage points in accuracy, demonstrating that parameter count alone does not determine performance for video models.

7. Conclusion

We introduced SoccerNet-GAR, a novel dataset and benchmark for group activity recognition that aligns broadcast video with tracking data for 87,939 football events across 10 action classes. On this benchmark, our baseline tracking model reaches 77.8% balanced accuracy and 57.0% macro F1 with 180K parameters, while the best video model reaches 60.9% balanced accuracy and 50.1% macro F1 with 86.3M parameters. Tracking therefore improves performance by 16.9 and 6.9 percentage points, while using 479× fewer parameters and training 7× faster. Our ablation studies demonstrate that domain-informed graph construction (positional edges encoding tactical structure) substantially outperforms geometric proximity patterns in both balanced accuracy and macro F1, and that exploring more expressive temporal architectures represent a promising direction for future work. Tracking excels on spatially distinctive events (*OUT*, *CROSS*, *GOAL*) but slightly struggles with *HEADER*, where video outperforms by +1.1 pp. Furthermore, tracking models can learn discriminative features even in severely data-scarce regimes, whereas video models collapse. These findings indicate that explicit spatial relationships provide stronger signal than appearance cues for recognizing coordinated actions in sports. The complementary failure modes across modalities motivate future work on multimodal fusion, alongside dynamic role assignment, spatio-temporal backbone-neck modeling and causal (past-only) temporal windows for online deployment. We will release the dataset, code, and evaluation protocol.

Acknowledgment. The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST) - Center of Excellence for Generative AI, under award number 5940.

References

[1] Gabriel Anzer, Pascal Bauer, Ulf Brefeld, and Dennis Faßmeyer. Detection of tactical patterns using semi-supervised graph neural networks. In *16th MIT sloan sports analytics conference*, pages 1–15, 2022. 2

[2] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *CVPR*, pages 4315–4324, 2017. 2

[3] Yassine Benzakour, Bruno Cabado, Silvio Giancola, Anthony Cioppa, Bernard Ghanem, and Marc Van Droogenbroeck. Osl-actionspotting: A unified library for action spotting in sports videos, 2024. 4

[4] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks?, 2022. 3

[5] Alejandro Cartas, Coloma Ballester, and Gloria Haro. A graph-based method for soccer action spotting using unsupervised player classification. In *ACM International Workshop on Multimedia Content Analysis in Sports (MMSports)*, pages 93–102, 2022. 2

[6] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*, pages 1282–1289. IEEE, 2009. 2, 4

[7] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Floriane Magera, Olivier Barnich, Bernard Ghanem, and Marc Van Droogenbroeck. Camera calibration and player localization in SoccerNet-v2 and investigation of their representations for action spotting. pages 4532–4541, 2021. 2

[8] Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Scaling up soccernet with multi-view spatial localization and re-identification. *Scientific data*, 9(1):355, 2022. 2

[9] Anthony Cioppa, Silvio Giancola, Adrien Deliège, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3490–3501, 2022. 2

[10] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Floriane Magera, Xin Zhou, Hassan Mkhallati, Adrien Deliège, Jan Held, Carlos Hinojosa, Amir M Mansourian, et al. SoccerNet 2023 challenges results. *Sports Engineering*, 27(2): 24, 2024. 2, 4

[11] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *CVPR Workshops*, 2021. 2, 4

[12] Dawei Ding and Hsiangsheng Huang. A graph attention based approach for trajectory prediction in multi-agent sports games. *arXiv preprint arXiv:2012.10531*, 2020. 3

[13] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[14] Javier Fernández, Luke Bornn, and Dan Cervone. Wide open spaces: A statistical technique for measuring space creation in professional soccer. In *MIT Sloan Sports Analytics Conference*, 2019. 2

[15] Sushant Gautam, Mehdi Houshmand Sarkhoosh, Jan Held, Cise Midoglu, Anthony Cioppa, Silvio Giancola, Vajira Thambawita, Michael Alexander Riegler, P. Halvorsen, and Mubarak Shah. Soccernet-echoes: A soccer game audio commentary dataset. *2024 International Symposium on Multimedia (ISM)*, pages 71–78, 2024. 2

[16] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 839–848, 2020. 2, 3

[17] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A scalable dataset for action spotting in soccer videos. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1711–1721, 2018. 2, 4

[18] Silvio Giancola, Anthony Cioppa, Bernard Ghanem, and Marc Van Droogenbroeck. Deep learning for action spotting in association football videos. *ArXiv*, abs/2410.01304, 2024. 2

[19] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005. 3

[20] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1024–1034, 2017. 3

[21] Sandro Hauri and Slobodan Vucetic. Group activity recognition in basketball tracking data–neural embeddings in team sports (nets). *arXiv preprint arXiv:2209.00451*, 2022. 3, 4

[22] Kristina Host and Marina Ivasic-Kos. An overview of human action recognition in sports based on computer vision. *Heliyon*, 8, 2022. 2

[23] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1980, 2016. 2, 4

[24] Maxime Istasse, Vladimir Somers, Pratheeban Elancheliyan, Jaydeep De, and Davide Zambrano. Deepsportradar-v2: A multi-sport computer vision dataset for sport understandings. *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, 2023. 2

[25] Dongkeun Kim, Youngkil Song, Minsu Cho, and Suha Kwak. Towards more practical group activity detection: A new benchmark and model. *arXiv preprint arXiv:2312.02878*, 2023. 4

[26] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[27] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. 3

[28] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 156–165, 2017. 3

[29] Arnaud Leduc, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-depth: a scalable dataset for monocular depth estimation in sports videos. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3280–3282, 2024. 2

[30] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. DeepGCNs: Can GCNs go as deep as CNNs? In *IEEE International Conference on Computer Vision (ICCV)*, pages 9267–9276, 2019. 3

[31] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In *ICCV*, pages 13668–13677, 2021. 2

[32] Zhengcen Li, Xianxiang Chang, Yueran Li, and Jing Su. Skeleton-based group activity recognition via spatial-temporal panoramic graph. In *ECCV*, pages 254–270. Springer, 2024. 2

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[34] Abdul Majeed, Mohammad Farukh Hashmi, Muhammad Umar Ashraf, Gitanjali Srivastava, Zong Woo Geem, and Neeraj Dhanraj Bokde. Real-time analysis of soccer ball-player interactions using graph convolutional networks for enhanced game insights. *Scientific Reports*, 15(1):1–19, 2025. 3

[35] Duoxuan Pei, Di Huang, and Yunhong Wang. Fifawc: a dataset with detailed annotation and rich semantics for group activity recognition. *Frontiers of Computer Science*, 18(6): 186351, 2024. 4

[36] PFF FC. Data-driven exploration of the 2022 fifa world cup. https://www.blog.fc.pff.com/blog/enhanced-2022-world-cup-dataset, 2023. Accessed: November 12, 2025. 3

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[38] Mehdi Houshmand Sarkhoosh, Sushant Gautam, Cise Midoglu, Thu Nguyen, Jan Held, Anthony Cioppa, Silvio Giancola, Vajira Thambawita, Michael A. Riegler, and Pål Halvorsen. Beyond audio: Enhancing soccernet-echoes with multimodal event extraction using llms. *Int. J. Semantic Comput.*, 19:589–613, 2025. 2

[39] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 2

[40] Vladimir Somers, Victor Joos, Anthony Cioppa, Silvio Giancola, Seyed Abolfazl Ghasemzadeh, Floriane Magera, Baptiste Standaert, Amir Mohammad Mansourian, Xin Zhou, Shohreh Kasaei, et al. SoccerNet game state reconstruction: End-to-end athlete tracking and identification on a minimap. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 2

[41] Michael Stöckl, Thomas Seidl, Daniel Marley, and Paul Power. Making offensive play predictable-using a graph convolutional network to understand defensive performance in soccer. In *Proceedings of the 15th MIT sloan sports analytics conference*. MIT sloan sports analytics conference United States, 2021. 2

[42] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2, 4

[43] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023. 2, 4, 5

[44] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph CNN for learning on point clouds. In *ACM Transactions on Graphics (TOG)*, pages 1–12, 2019. 3

[45] Zhe Wang, Petar Veličković, Daniel Hennes, Nenad Tomašev, Laurel Prince, Michael Kaisers, Yoram Bachrach, Romuald Elie, Li Kevin Wenliang, Federico Piccinini, William Spearman, Ian Graham, Jerome Connor, Yi Yang, Adrià Recasens, Mina Khan, Nathalie Beauguerlange, Pablo Sprechmann, Pol Moreno, Nicolas Heess, Michael Bowling, Demis Hassabis, and Karl Tuyls. Tacticai: an ai assistant for football tactics, 2023. 2

[46] Zhao Xie, Chang Jiao, Kewei Wu, Dan Guo, and Richang Hong. Active factor graph network for group activity recognition. *IEEE Transactions on Image Processing*, 33:1574–1587, 2024. 2

[47] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks?, 2019. 3, 5

[48] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition, 2020. 4

[49] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 2

[50] Raymond A Yeh, Alexander G Schwing, Jonathan Huang, and Kevin Murphy. Diverse generation for multi-agent sports games. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4610–4619, 2019. 2

[51] Raymond A Yeh, Alexander G Schwing, Jonathan Huang, and Kevin Murphy. Forecasting basketball trajectories and player intentions using graph neural networks. In *ICML Workshop on Computer Vision for Autonomous Vehicles*, 2019. 2, 3

[52] Hangjie Yuan and Dong Ni. Learning visual context for group activity recognition. In *AAAI*, pages 3261–3269, 2021. 2

[53] Youliang Zhang, Wenxuan Liu, Danni Xu, Zhuo Zhou, and Zheng Wang. Bi-causal: Group activity recognition via bidirectional causality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[54] Honglu Zhou, Asim Kadav, Aviv Shamsian, Shijie Geng, Farley Lai, Long Zhao, Ting Liu, Mubbasir Kapadia, and Hans Peter Graf. Composer: Compositional reasoning of group activity in videos with keypoint-only modality. In *ECCV*, pages 249–266. Springer, 2022. 2

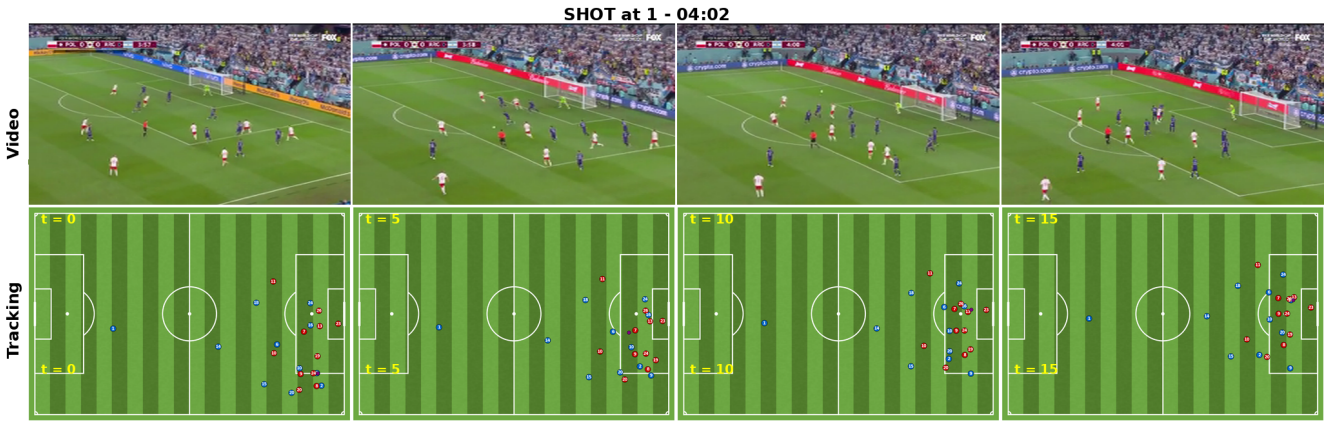


Figure 6. **Video vs Tracking Input for a SHOT Event.** Top: broadcast frames at four timesteps showing camera motion and partial field coverage. Bottom: synchronized tracking data on a standardized pitch, providing complete spatial context for all players regardless of camera position.

8. Visualization of data synchronization

Figure 6 illustrates the two input modalities for a *SHOT* event at match time 04:02. The top row shows four broadcast frames sampled at timesteps $t \in \{0, 5, 10, 15\}$ within the observation window. The camera pans and zooms to follow the play, introducing viewpoint changes, partial occlusions, and background clutter that the video encoder must contend with. The bottom row shows the corresponding tracking data projected onto a standardized pitch template, where red and blue markers denote players from opposing teams and the purple marker indicates the ball position. The tracking representation abstracts away visual noise and captures the spatial configuration of all 22 players in a canonical coordinate frame, making tactical patterns such as defensive compactness and attacking movement directly readable. This example highlights a key advantage of tracking modality: while the broadcast view loses several players outside the frame at each timestep, the bird's-eye tracking view preserves the full team-level spatial context throughout the observation window.

9. Data scaling

Figure 7 shows performance versus training set size for both modalities, using GIN + MaxPool + Positional edges for tracking and VideoMAEv2-B finetuned for video (mean over 5 seeds for tracking). Tracking achieves 67.0% with only 5 matches and plateaus around 35 matches (78.5%), while video achieves 41.6% with 5 matches and also plateaus around 35 matches (62.4%), after which both modalities fluctuate slightly. The gap narrows from 25.4% at 5 matches to 16.1% at 35 matches, indicating that video benefits more from additional data but remains substantially behind. Tracking maintains superior performance across all data regimes, suggesting that structured positional representations require less training data to reach strong performance.

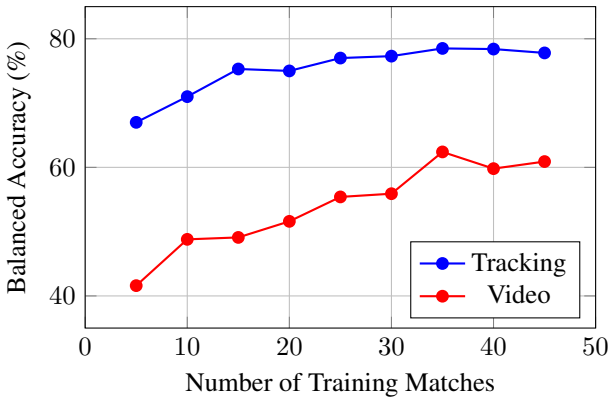


Figure 7. **Performance *vs*. Number of Training Matches for Video and Tracking Modalities.** Both modalities plateau around $\approx$ 35 matches, but tracking reaches strong performance (67.0%) with as few as 5 matches, whereas video starts at 41.6%. The gap narrows from 25.4% (5 matches) to 16.1% (35 matches).