Few-Shot Precise Event Spotting via Unified Multi-Entity Graph and Distillation

Zhaoyu Liu[1], Kan Jiang[1], Murong Ma[1], Zhe Hou[2], Yun Lin[3], Jin Song Dong[1]

[1]National University of Singapore   [2]Griffith University   [3]Shanghai Jiao Tong University
{liuzy, jiangkan}@nus.edu.sg, murongma@u.nus.edu
z.hou@griffith.edu.au, lin_yun@sjtu.edu.cn, dcsdjs@nus.edu.sg

Abstract

Precise event spotting (PES) aims to recognize fine-grained events at exact moments and has become a key component of sports analytics. This task is particularly challenging due to rapid succession, motion blur, and subtle visual differences. Consequently, most existing methods rely on domain-specific, end-to-end training with large labeled datasets and often struggle in few-shot conditions due to their dependence on pixel- or pose-based inputs alone. However, obtaining large labeled datasets is practically hard. We propose a Unified Multi-Entity Graph Network (UMEG-Net) for few-shot PES. UMEG-Net integrates human skeletons and sport-specific object keypoints into a unified graph and features an efficient spatio-temporal extraction module based on advanced GCN and multi-scale temporal shift. To further enhance performance, we employ multimodal distillation to transfer knowledge from keypoint-based graphs to visual representations. Our approach achieves robust performance with limited labeled data and significantly outperforms baseline models in few-shot settings, providing a scalable and effective solution for few-shot PES. Code is publicly available at https://github.com/LZYAndy/UMEG-Net.
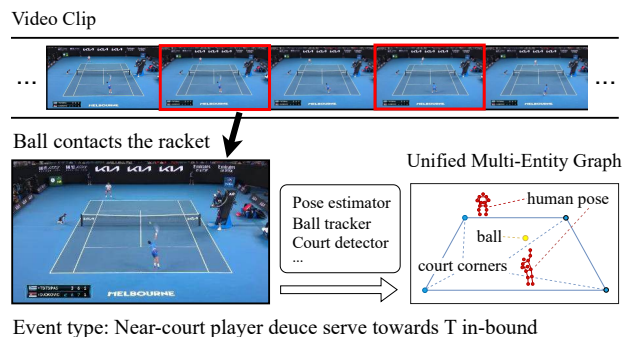
Figure 1: Precise event spotting in sports videos, with event timestamps highlighted in red. Each scene can be represented by a unified graph including human poses and sport-related entity keypoints (e.g., ball, court corners).

Introduction

Precise Event Spotting (PES) is a trending problem that aims to identify events and their class from long, untrimmed videos, particularly in sports (Hong et al. 2022; Xarles et al. 2024; Xu et al. 2025). The main objective is to accurately detect sequences of fine-grained, rapidly occurring sports events, such as a hitting event in racket sports, within a tight tolerance window (1-2 frames). This capability is essential for sports analytics applications such as match forecasting (Wang et al. 2022; Liu and Su 2025), strategic and tactical analysis (Dong et al. 2023; Liu et al. 2023, 2024a,b,c, 2025b), and player performance evaluation (Decroos et al. 2019; Pappalardo et al. 2019). Most existing PES methods are trained end-to-end using raw RGB images as input (Hong et al. 2022; Xarles et al. 2024; Liu et al. 2025c). One main challenge with these approaches is that they rely on large-scale datasets with dense, frame-level annotations (Shao et al. 2020; Xu et al. 2022; Wang et al. 2023b; Liu et al. 2025c), which leads to substantial label-

ing costs and significantly limits their scalability to new domains (Xu et al. 2025).

These limitations underscore the significance of *few-shot learning for PES*, an area that remains under-explored in existing research. Traditional few-shot methods, such as contrastive (Chen et al. 2020) and meta-learning (Wanyan et al. 2025) approaches, typically excel at coarse-grained, action-level recognition but do not directly address the frame-level accuracy required by PES. Recent advancements in human pose estimation (Sun et al. 2019) and sports object tracking (Huang et al. 2019; Jiang et al. 2020, 2023; Chen and Wang 2023; Jiang et al. 2024) have facilitated transforming raw RGB videos into compact, keypoint-based representations suitable for few-shot scenarios. Although numerous studies leverage human pose data for action recognition (Duan et al. 2022; Zhou et al. 2024; Liu et al. 2025a), these methods usually neglect critical contextual cues from objects (e.g., the ball) and the environment (e.g., the court), resulting in substantial information loss and compromised event detection performance. Some works have attempted to include additional context (Ibh et al. 2023; Li et al. 2024), yet they predominantly focus on coarse-grained recognition tasks and thus lack the fine-grained temporal precision essential for PES, making direct comparisons difficult. Additionally, keypoint-based approaches depend on accurate de-

tections, which are often unreliable in fast-moving sports due to motion blur and occlusion. Thus, relying solely on keypoints is suboptimal for robust PES.

To address these limitations, we propose a graph-based method for precise event spotting that flexibly incorporates human skeletons and object-level keypoints (e.g, ball positions and court corners). We first use pose estimation and object detection to extract keypoints from all relevant entities. These keypoints are integrated into a unified multi-entity graph (see Figure 1), capturing interactions among multiple players, objects, and contextual landmarks, thus overcoming the limitations of standard skeleton-only representations. To efficiently encode spatial-temporal relationships, we introduce the Unified Multi-Entity Graph Network (UMEG-Net), a novel graph-based module explicitly designed for few-shot PES. UMEG-Net employs an advanced Graph Convolutional Network (GCN) backbone for spatial modeling, replacing traditional temporal convolutions with a parameter-free multi-scale temporal shift mechanism. This temporal shift operation effectively captures temporal dynamics without introducing extra parameters, ensuring computational efficiency and robust few-shot performance. Furthermore, to mitigate inaccuracies from keypoint detection, we utilize multimodal knowledge distillation, transferring learned features from the graph-based teacher model to an RGB-based student network (denoted as UMEG-Net $_{distill}$). By leveraging abundant unlabeled videos from the target domain, the student network learns complementary visual representations, enhancing robustness and generalization in few-shot PES scenarios.

UMEG-Net improves few-shot PES across diverse sports domains. We validate its effectiveness on five sports video datasets with fine-grained event types and precise timestamps: F^{3}Set (Liu et al. 2025c), ShuttleSet (Wang et al. 2023b), FineGym (Shao et al. 2020), Figure Skating (Hong et al. 2021), and SoccerNet Ball Action Spotting (Cioppa et al. 2024). Under few-shot conditions, UMEG-Net consistently outperforms baseline methods, improving F1 scores by 1.3% to 5.5% and edit scores by 1.3% to 16.4%.. Additionally, incorporating multimodal distillation, our visual-based student model UMEG-Net$_{distill}$ achieves an additional average gain of 5.8% in F1 score and 6.7% in edit score, highlighting the robustness gained from complementary RGB features. The key contributions are as follows:

- We introduce and investigate the *few-shot* precise event spotting (PES) task, targeting frame-level event recognition with limited labeled data.

- We designed a unified multi-entity graph that incorporates human skeletons, objects (e.g., ball), and contextual landmarks (e.g., court corners) to represent sports events.

- We propose UMEG-Net, a new graph-based framework for few PES. It combines spatial graph convolution with a parameter-free multi-scale temporal shift mechanism, and enhances robustness through multimodal distillation.

- We conduct extensive experiments and ablation studies across five sports datasets, demonstrating that UMEG-Net achieves state-of-the-art performance in both few-shot and fully supervised settings.

Related Work

Precise Event Spotting

Precise event spotting in sports video analysis, initially introduced by (Einfalt et al. 2019) for athletics (long and triple jumps), involves identifying exact timestamps of specific actions within strict frame-level tolerances. Subsequently, the SoccerNet dataset (Giancola et al. 2018) expanded this research by providing extensive soccer video annotations for temporal action localization. Recent advances (Hong et al. 2022; Xarles et al. 2024) have broadened the applicability of PES across various sports. E2E-Spot (Hong et al. 2022), employing a CNN backbone with Gated Shift Modules (Sudhakaran, Escalera, and Lanz 2020) and GRUs (Dey and Salem 2017) for temporal modeling, performed effectively in tennis and figure skating scenarios but struggled with the complex temporal dynamics of SoccerNet V2 (Cioppa et al. 2024). Addressing these limitations, T-DEED (Xarles et al. 2024) improved temporal precision for fast-paced sports, achieving state-of-the-art accuracy on Figure Skating and Fine Diving benchmarks. More recently, Liu et. al. introduced a large-scale dataset spanning multiple sports and their proposed F^{3}ED further improved PES performance through efficient visual encoding and contextual sequence refinement (Liu et al. 2025c). Despite these advances, PES methods often suffer from limited supervision due to the high cost of fine-grained annotation. Our method addresses these limitations, demonstrating superior event spotting accuracy, particularly in few-shot scenarios.

Skeleton-Based Action Understanding

Skeleton-based human pose representations have been extensively utilized in sports analytics due to their efficiency, reduced complexity, and robustness in fine-grained action recognition, especially under few-shot conditions. Recent methods leveraging skeleton data have demonstrated effectiveness on single-athlete sports datasets, such as gymnastics (FineGym), achieving strong performance through dedicated skeleton-based frameworks (Duan et al. 2022; Liu et al. 2025a). Hong et al. (Hong et al. 2021) further highlighted the potential of pose features by distilling pose knowledge into RGB-based networks, illustrating how pose alignment can enhance recognition accuracy, yet they still focus primarily on single-athlete scenarios and remain reliant on accurate pose detection. In team sports such as volleyball and basketball, skeleton-based methods have explored group activity recognition through modeling multi-person interactions (Perez, Liu, and Kot 2022; Zhou et al. 2022). Recent studies have additionally integrated non-human entities, such as the shuttlecock trajectory and court locations in badminton (Liu and Wang 2022; Ibh et al. 2023), to enrich contextual information. Liu et al. (Li et al. 2024) improved upon these efforts by constructing panoramic graphs that integrate multiple players and objects but overlooked critical spatial relationships involving court locations. Furthermore, these approaches typically address action recognition tasks without considering precise temporal localization in PES. Additionally, our graph-based module efficiently supports precise temporal spotting and few-

shot learning. By further incorporating distillation from raw RGB inputs, our framework remains robust against inaccuracies in pose and object detection, significantly outperforming previous approaches under strict temporal localization constraints and limited data settings.

Proposed Method

This section presents our method in detail, starting with the problem statement and then presenting the modules.

Problem Statement

We define the few-shot PES task as follows: Given the input video clip $X \in \mathbb{R}^{T \times H \times W \times 3}$ consisting of T frames of RGB image size $H \times W$ with channel size of 3, our goal is to detect a sequence of M event-timestamp pairs $((E_1, t_1), \ldots, (E_M, t_M))$. Here, E_i denotes the event type with C possible classes, and t_i is the corresponding timestamp for $i \in \{1, \ldots, M\}$. We consider a target dataset D composed of $|D|$ video *clips*, each containing a certain number of events. Let $D_{\text{label}} \subset D$ denote the subset of D that contains labels, and let $D_{\text{unlabel}} \subset D$ be the unlabeled subset, such that $D = D_{\text{label}} \cup D_{\text{unlabel}}$. The number of clips in each subset is $|D_{\text{(un)label}}|$. We specifically address the **few-shot** scenario, where the number of labeled samples $k = |D_{\text{label}}|$ is small, referred to as the "k-clip" setting.

Our Proposed Framework

To tackle the problem described above, we propose UMEG-Net, a unified multi-entity graph network explicitly designed for few-shot PES. The overall architecture is depicted in Figure 2. A unified panoramic graph is first constructed by integrating keypoints representing human skeletons and sport-related entities (e.g., ball, court). This graph is then processed by a series of UMEG Blocks, each combining advanced spatial graph convolutional networks (GCNs) with a parameter-free multi-scale temporal shift mechanism for efficient spatio-temporal feature extraction. Finally, the feature maps are forwarded to the event localizer and classifier for precise event spotting. To further enhance model robustness and generalization, we introduce a multimodal distillation approach, transferring knowledge from the graph-based model to an RGB-based student network through weak-supervision even on large amounts of unlabeled videos.

Unified Multi-Entity Graph Construction Given a sports video, we first convert each frame into a structured graph of interacting entities. Specifically, for each frame t, we define a graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$, where $\mathcal{V}_t$ is the set of nodes that includes all detected human joints and keypoints of sports-related entities and $\mathcal{E}_t$ is the set of edges. $\mathcal{V}$ including all detected human joints and keypoints of sports-related entities, denoted $\mathcal{V}_t = \{V_p^t, V_b^t, V_c^t\}$, where $V_p^t = \{P_i^t | i = 1, ..., N\}$ is the set of N persons each $P_i^t = (j_{i,1}^t, \ldots, j_{i,K}^t)$ represented by joints K, V_b^t are ball keypoints and V_c^t are court keypoints. The number of nodes is denoted as $|\mathcal{V}_t| = N * K + |V_b^t| + |V_c^t|$. The edge set $\mathcal{E}_t$

captures both intra-entity structures and cross-entity interactions. It contains four components:

$$\mathcal{E}_t = \mathcal{E}_t^{\text{intra}} \cup \mathcal{E}_t^{\text{p–b}} \cup \mathcal{E}_t^{\text{p–c}} \cup \mathcal{E}_t^{\text{c–c}},$$

where $\mathcal{E}_t^{\text{intra}}$ encodes skeletal connections within each player (intra-person edges follow standard human-joint topology); $\mathcal{E}_t^{\text{c–c}}$ connects 4 court corners as a rectangle to model the field boundary; $\mathcal{E}_t^{\text{p–b}}$ connects human joints to the ball object (for racket sports the wrist joints connect to the ball, while for soccer the ankle and shoulder joints connect to the ball to approximate lower- and upper-body ball control); $\mathcal{E}_t^{\text{p–c}}$ links human foot joints to court corners (positional context). All edges are undirected. See full connection rules in our code.

Our proposed *Unified Multi-Entity Graph* flexibly integrates multiple players with objects and environmental cues. Compared to conventional skeleton-only graphs that ignore objects and context (resulting in loss of critical event information), our unified graph provides a richer and more holistic representation of the scene. By explicitly including these elements as nodes and edges, UMEG-Net can represent fine-grained event cues that would otherwise be overlooked.

Graph-based Spatio-Temporal Encoding Given a sequence of unified multi-entity graphs $G = \{\mathcal{G}_t\}_{t=1}^{T} \in \mathbb{R}^{T \times |\mathcal{V}| \times 2}$ over a video clip of T frames, we extract discriminative features via a specially designed spatio-temporal encoder, named **UMEG Block**, tailored for precise event spotting under few-shot conditions. The UMEG Block consists of a **spatial graph convolution network** that operates on the unified graph topology, combined with a novel, parameter-free **temporal multi-scale shift module** that enables efficient and effective temporal feature extraction. In the following, we describe the spatial GCN module and the temporal multi-scale shift mechanism.

(1) Spatial GCN. This spatial GCN layer updates node features by aggregating information from their neighbors in the graph. In matrix form, the output of one GCN layer is:

$$\mathcal{H}^{(\ell+1)} = \text{ReLU}(A^{(\ell)} \mathcal{H}^{(\ell)} W^{(\ell)}), \tag{1}$$

where $A(\ell) \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ is the adjacency matrix employed for spatial aggregation, $\mathcal{H}^{(\ell)} \in \mathbb{R}^{|\mathcal{V}| \times T \times d}$ symbolizes the hidden representation, and $W^{(\ell)} \in \mathbb{R}^{d \times d}$ is the weight matrix utilized for feature projection. Here, $|\mathcal{V}|$, T, and d denote the number of nodes, frames, and hidden feature dimension, respectively. ReLU$(\cdot)$ is the ReLU activation function, and the superscript ℓ indicates the layer number.

Unlike traditional GCNs in skeleton-based action recognition (Duan et al. 2022; Zhou et al. 2024; Liu et al. 2025a) that process each individual independently, our approach performs graph convolutions over the entire multi-entity graph. This allows for joint modeling of both human–human and human–entity interactions, capturing intra- and inter-entity relationships simultaneously. Therefore, it can exploit richer contextual cues essential for precise event spotting.

(2) Temporal Multi-Scale Shift. To model temporal correlations, most skeleton-based action recognition models (Yan, Xiong, and Lin 2018; Shi et al. 2020; Chen et al. 2021; Duan et al. 2022; Zhou et al. 2024; Liu et al. 2025a)
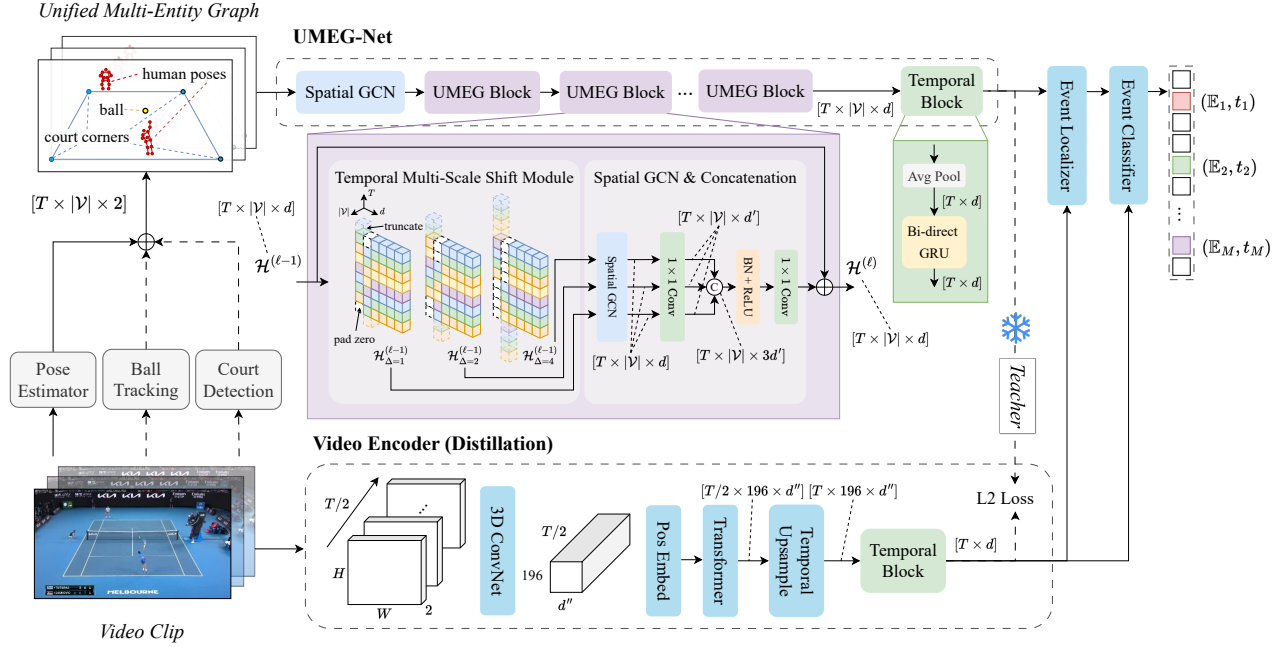
Figure 2: The framework of our proposed method, including UMEG-Net and multimodal distillation. Each frame is converted to a unified multi-entity graph and processed by stacked UMEG Blocks to produce features for precise event spotting. A transformer-based RGB student is trained via knowledge distillation from the frozen graph-based teacher.

apply multi-scale temporal convolution modules after each spatial GCN to capture frame-to-frame dynamics. However, in few-shot settings, such modules increase the number of trainable parameters, making training more difficult and prone to overfitting. To address this, we draw inspiration from TSM (Lin, Gan, and Han 2019) and introduce a Temporal Multi-Scale Shift Module that efficiently captures temporal dynamics by shifting feature vectors along the temporal axis at multiple scales, without adding additional trainable parameters. Let $\mathcal{H}_t^{(\ell)} \in \mathbb{R}^{|\mathcal{V}| \times d}$ be the input features at frame t and layer ℓ. We split channels into static, forward-shift, and backward-shift parts with fraction α (we use $\alpha = 1/8$)

$$\mathcal{H}_t^{(\ell)} = \big[\, \mathcal{H}_{t,\text{static}}^{(\ell)} \,\|\, \mathcal{H}_{t,\text{fwd}}^{(\ell)} \,\|\, \mathcal{H}_{t,\text{bwd}}^{(\ell)} \,\big], \qquad (2)$$

where $\mathcal{H}_{t,\text{static}}^{(\ell)} \in \mathbb{R}^{|\mathcal{V}| \times (1-2\alpha)d}$, and $\mathcal{H}_{t,\text{fwd}}^{(\ell)}, \mathcal{H}_{t,\text{bwd}}^{(\ell)} \in \mathbb{R}^{|\mathcal{V}| \times \alpha d}$. $\|$ is the notation for (channel-wise) concatenation. For temporal offsets $\Delta \in \{1, 2, 4\}$, define the *bidirectional* shift operator

$$\tilde{\mathcal{H}}_t^{(\ell,\Delta)} = \big[\, \mathcal{H}_{t,\text{static}}^{(\ell)} \,\|\, \mathcal{H}_{t-\Delta,\text{fwd}}^{(\ell)} \,\|\, \mathcal{H}_{t+\Delta,\text{bwd}}^{(\ell)} \,\big], \quad (3)$$

with boundary handling by zero padding. Each shifted stream is then passed to the spatial GCN to update the node embeddings with temporal context

$$Z_t^{(\ell,\Delta)} = \text{ReLU}(A^{(\ell)} \tilde{\mathcal{H}}_t^{(\ell,\Delta)} W^{(\ell)}), \quad \Delta \in \{1,2,4\}, \tag{4}$$

where A is the adjacency matrix, W is the weight for spatial GCN, and $Z_t^{(\ell,\Delta)} \in \mathbb{R}^{|\mathcal{V}| \times d}$. We then fuse multi-scale

contexts by

$$U_t^{(\ell)} = \big\|_{\Delta \in \{1,2,4\}} \text{F}_1(Z_t^{(\ell,\Delta)}), \tag{5}$$

where F_1 is a linear projection to down-scale the channel size to $\mathbb{R}^{|\mathcal{V}| \times \lfloor d/|\Delta| \rfloor}$. The output is

$$\mathcal{H}_t^{(\ell+1)} = \text{F}_2(\text{ReLU}(U_t^{(\ell)})) + \mathcal{H}_t^{(\ell)}, \tag{6}$$

where F_2 is a linear projection to scale the channel size to $\mathbb{R}^{|\mathcal{V}| \times d}$, and $\mathcal{H}_t^{(\ell)}$ is added via a residual connection. This multi-scale temporal shift mechanism preserves frame-level resolution, expands the temporal receptive field across short, mid, and long ranges (via $\Delta \in \{1, 2, 4\}$), and introduces no additional trainable parameters.

Multimodal Knowledge Distillation While the graph-based model excels at leveraging structured keypoint information, its performance can degrade if the pose or object detections are unreliable (e.g. due to motion blur or occlusions). To enhance robustness, we propose a multimodal distillation framework where knowledge is transferred from the graph domain to the raw visual domain. We introduce a student network that operates directly on the RGB video frames and learn it under the guidance of the trained graph-based teacher (our UMEG-Net).

Let the frozen graph-based teacher encoder ϵ_{tch} map the unified graph sequence $G = \{\mathcal{G}_t\}_{t=1}^T$ to per-frame embeddings $\mathbf{F}_{\text{tch}} = \epsilon_{\text{tch}}(G) \in \mathbb{R}^{T \times d}$. The RGB student encoder ϵ_{stu} maps the video $X \in \mathbb{R}^{T \times H \times W \times 3}$ to $\mathbf{F}_{\text{stu}} = \epsilon_{\text{stu}}(X) \in \mathbb{R}^{T \times d}$. ϵ_{stu} consists of a visual feature extractor followed by a bidirectional GRU to capture long-term dependencies and project the extracted features into the same

dimensional space d as the teacher encoder. On all unlabeled clips D_{unlabel}, we minimize a feature-matching L2 loss

$$\mathcal{L}_{\text{feat}} = \frac{1}{T} \sum_{t=1}^{T} \left\| \mathbf{F}_{\text{tch}}^{(t)} - \mathbf{F}_{\text{stu}}^{(t)} \right\|_2^2, \tag{7}$$

The teacher encoder, initialized from the pretrained weights in the previous stage, remains frozen during distillation, while the student networks are trainable. A few-shot adaption is adopted to fine-tune the event localizer and classifier on D_{label}. At inference time, the student alone performs PES directly from RGB, inheriting robustness from the teacher's structured graph supervision.

Implementation Details

In UMEG-Net, we employ the unit GCN from (Zhou et al. 2024) as the spatial GCN layer for its strong performance and efficiency. The multi-scale temporal shift module applies shifts of $\Delta \in \{1, 2, 4\}$ frames. For distillation, we use VideoMAEv2 (Wang et al. 2023a), an advanced transformer-based visual encoder, as the backbone for feature extraction (smaller variants also show consistent gains). Event localization and classification are performed by linear layers that output event probabilities and event types, respectively. The training protocol processes 96-frame sequences with a stride of 2. RGB frames are resized to 224 pixels in height, then randomly cropped to 224×224 to preserve essential visual information. Standard data augmentation (cropping, color jittering) enhances data diversity and model robustness during training but is omitted in testing. Models are optimized with AdamW (initial learning rate 0.001 for UMEG-Net, 0.0001 for VideoMAEv2 in distillation), using three linear warm-up steps followed by cosine annealing. Training is conducted on an RTX 4090 GPU. Further implementation details are provided in the Appendix.

Experimental Results

This section presents evaluation details and ablation studies.

Datasets

To evaluate the effectiveness of our method, we conduct experiments on several PES datasets, including racket sports F[3]Set-Tennis (Liu et al. 2025c) and ShuttleSet (Wang et al. 2023b), individual sports FineGym (Shao et al. 2020) and Figure Skating (Hong et al. 2021), and team sports SoccerNet-BAS (Cioppa et al. 2024). Detailed descriptions of these datasets are provided in the supplementary material.

We extract 2D poses using off-the-shelf pose estimators. For 2D pose estimation, we use HRNet (Sun et al. 2019) to detect athletes' poses through top-down estimation. For sports-specific object detection and tracking, we fine-tune the pretrained YOLOv8 (Jocher, Chaurasia, and Qiu 2023) model on corresponding public datasets from Roboflow (Dwyer et al. 2025) to detect and track target players and sports balls in F3Set-Tennis, ShuttleSet, and SoccerNet-BAS. For two racket sports datasets, we also employ deployed models for court detection to identify the four corners. Please refer to Appendix for more details.

Evaluation Metrics

Following Liu et al. (Liu et al. 2025c), we evaluate our method using two metrics that assess temporal precision and classification accuracy: Edit score and mean F1 with temporal tolerance. *(1) Edit Score* (Lea et al. 2017) measures the structural similarity between predicted and ground-truth event sequences using Levenshtein distance, accounting for missing, redundant, and misordered predictions. It is suitable for tasks requiring accurate event ordering and completeness. *(2) Mean F1 Score with Temporal Tolerance* evaluates both event classification and localization accuracy. A prediction is correct if it matches the event class and occurs within a temporal window of δ frames around the ground-truth timestamp. We report the average F1 across all event types, denoted as F1$_{\text{evt}}$. Unless otherwise specified, we use a strict tolerance of $\delta = 1$ frame; for SoccerNet-BAS, we follow prior work (Cioppa et al. 2024) and use $\delta = 1$ second.

Baselines

We compare our approach with SOTA PES methods that operate on RGB inputs and support end-to-end training, including E2E-Spot (Hong et al. 2022) and T-DEED (Xarles et al. 2024), each evaluated with RegNet-Y (Radosavovic et al. 2020) 200MF and 800MF backbones, as well as F[3]ED (Liu et al. 2025c). Furthermore, we construct skeleton-based PES variants of these baselines by replacing their visual encoders with existing skeleton-based GCN architectures. These variants take 2D human poses as input and employ F[3]ED's head module for event spotting. Specifically, we adopt graph-based models including MSG3D (Liu et al. 2020), AAGCN (Shi et al. 2020), CTRGCN (Chen et al. 2021), STGCN++ (Duan et al. 2022), BlockGCN (Zhou et al. 2024), and ProtoGCN (Liu et al. 2025a). Per-frame representations are computed by aggregating individual skeleton features via averaging across all detected persons.

Few-Shot Setting

We define the few-shot scenario as training with a limited number of annotated clips (k-clip), following (Hong et al. 2021). In F[3]Set-Tennis and ShuttleSet, each clip represents a rally spanning several seconds and comprising a sequence of shots. In FineGym-BB and Figure Skating, each clip corresponds to a routine with a series of technical and artistic movements. In SoccerNet-BAS, each clip captures a phase of play featuring actions such as pass, drive, and shoot.

Unlike the traditional k-shot approach, which segments videos into independent samples containing single events with backgrounds (Yang et al. 2020; Nag, Zhu, and Xiang 2021), our k-clip strategy offers a more practical and domain-adapted setting. *First*, due to the large number of event types in sports, many of which are rare, sampling k instances per type is often impractical; annotating a small set of clips per sport domain is more efficient and scalable. *Second*, sports actions are typically brief (1–2 frames) and occur rapidly in succession, thus isolated events lack essential temporal context. *Third*, consecutive events frequently exhibit strong dependencies (Hong et al. 2022; Liu et al. 2025c); therefore, training on k-clips rather than k-shots enables the model to better capture long-term event relationships.

Table 1: Experimental results and ablation studies for fine-grained sports event detection across five datasets using *100-clip* training data are reported with evaluation metrics $F1_{evt}$ and edit score. "Params (M)" refers to the number of model parameters. Top-performing results are **bolded**, while the best within each method category are <u>underlined</u>. Our method UMEG-Net outperforms all the assessed state-of-the-art (SOTA) methods.

	F^3Set-Tennis		ShuttleSet		FineGym-BB		Figure Skating		SoccerNet-BAS		Params (M)
	$F1_{evt}$	Edit	$F1_{evt}$	Edit	$F1_{evt}$	Edit	$F1_{evt}$	Edit	$F1_{evt}$	Edit	
(a) SOTA PES methods											
E2E-Spot $_{200MF}$ (Hong et al. 2022)	2.6	11.1	35.6	50.3	40.1	50.8	29.9	34.8	15.3	35.4	<u>4.5</u>
E2E-Spot $_{800MF}$ (Hong et al. 2022)	3.1	13.3	42.7	54.6	<u>44.7</u>	<u>53.1</u>	35.9	42.7	22.1	<u>43.1</u>	12.6
T-DEED $_{200MF}$ (Xarles et al. 2024)	1.0	4.2	33.8	41.7	44.1	48.4	36.6	33.5	4.7	8.8	16.4
T-DEED $_{800MF}$ (Xarles et al. 2024)	1.5	6.4	30.7	38.6	43.6	48.3	<u>37.9</u>	<u>40.6</u>	6.7	14.3	46.2
F^3ED (Liu et al. 2025c)	<u>3.9</u>	<u>15.3</u>	<u>44.1</u>	<u>55.1</u>	43.8	52.1	36.1	34.4	<u>22.7</u>	34.5	4.7
(b) Skeleton-based PES variants											
MSG3D (Liu et al. 2020)	5.2	15.4	39.7	56.3	45.8	50.1	18.1	32.8	<u>22.3</u>	39.5	4.6
AAGCN (Shi et al. 2020)	4.9	15.4	44.7	55.6	42.8	47.8	24.8	40.0	22.1	43.1	5.4
CTRGCN (Chen et al. 2021)	5.5	16.9	40.3	55.1	<u>46.6</u>	50.7	28.2	44.4	<u>22.3</u>	42.4	3.1
STGCN++ (Duan et al. 2022)	6.4	18.0	45.0	57.7	44.6	50.0	29.5	46.8	17.4	42.7	3.0
ProtoGCN (Liu et al. 2025a)	6.6	18.1	46.8	58.3	41.4	<u>51.1</u>	25.3	43.7	17.8	41.0	5.6
BlockGCN (Zhou et al. 2024)	<u>6.9</u>	<u>18.3</u>	<u>47.1</u>	<u>59.4</u>	44.5	49.1	<u>29.8</u>	<u>48.2</u>	19.9	<u>43.3</u>	<u>2.5</u>
(c) Our approach											
UMEG-Net	**9.4**	**31.7**	**49.2**	**64.0**	**49.2**	**54.4**	**39.2**	**49.6**	**27.0**	**44.8**	**<u>2.2</u>**
UMEG-Net $_{distill}$	<u>**12.5**</u>	<u>**40.7**</u>	<u>**59.1**</u>	<u>**69.0**</u>	<u>**58.4**</u>	<u>**61.2**</u>	**45.9**	**56.2**	**27.1**	<u>**50.8**</u>	67.8
Ablation studies											
(d) pose $* N$	5.6	23.9	47.4	61.5	<u>49.2</u>	<u>54.4</u>	<u>39.2</u>	<u>49.6</u>	20.7	39.6	–
pose $* N$ + court	6.6	26.1	46.7	61.5	–	–	–	–	–	–	–
pose $* N$ + ball	8.6	30.2	48.1	62.5	–	–	–	–	<u>27.0</u>	<u>44.8</u>	–
pose $* N$ + ball + court	<u>9.4</u>	<u>31.7</u>	<u>49.2</u>	<u>64.0</u>	–	–	–	–	–	–	–
(e) $\Delta \in \{1\}$	8.8	30.4	46.5	61.2	39.0	50.3	26.2	36.8	21.1	38.9	–
$\Delta \in \{1, 2\}$	<u>9.6</u>	<u>33.2</u>	47.4	61.6	42.9	49.8	32.1	45.3	23.1	40.1	–
$\Delta \in \{1, 2, 4\}$	9.4	31.7	<u>49.2</u>	<u>64.0</u>	<u>49.2</u>	<u>54.4</u>	<u>39.2</u>	<u>49.6</u>	<u>27.0</u>	<u>44.8</u>	–
(f) Self-supervise (Chen et al. 2020)	3.0	29.1	50.2	62.4	54.5	56.8	34.6	41.3	26.0	42.9	–
UMEG-Net $_{distill}$	<u>12.5</u>	<u>40.7</u>	<u>59.1</u>	<u>69.0</u>	<u>58.4</u>	<u>61.2</u>	<u>45.9</u>	<u>56.2</u>	<u>27.1</u>	<u>50.8</u>	–
(g) E2E-Spot (full supervision)	44.6	71.1	71.2	<u>76.1</u>	<u>72.9</u>	<u>73.0</u>	58.0	63.9	<u>46.2</u>	<u>72.9</u>	–
UMEG-Net (full supervision)	<u>47.5</u>	<u>71.2</u>	<u>71.4</u>	<u>76.1</u>	59.8	64.8	<u>61.8</u>	<u>71.8</u>	36.1	55.7	–

Result Analysis with Few-Shot Supervision

We evaluate performance in the few-shot setting across all datasets. For a fair comparison, all methods are trained on the same k-clip samples, with five random splits per dataset, and the results are averaged. We test $k \in \{15, 25, 50, 100\}$ across SOTA PES methods, skeleton-based PES variants, and our methods. Table 1 presents the results for the 100-clip setting. From the table, we observe that UMEG-Net consistently achieves strong performance across all datasets.

(1) Comparison with SOTA PES methods. UMEG-Net achieves substantial improvements over these methods as shown in Table 1(a). Specifically, it improves $F1_{evt}$ by 1.3% to 5.5% and Edit score by 1.3% to 16.4% over the best-performing PES baselines across five datasets, respectively. Despite strong performance on large-scale labeled data, these PES methods generally struggle in few-shot settings, highlighting the limitation of relying solely on RGB inputs.

(2) Compare with skeleton-based PES variants. UMEG-Net shows noticeable advantages over skeleton-based variants in Table 1(b), especially on datasets containing fine-grained events such as F^3Set-Tennis and ShuttleSet. For instance, it outperforms the best-performing variant BlockGCN by +2.5% $F1_{evt}$ and +13.4% Edit on F^3Set-Tennis, and +2.1% $F1_{evt}$ and +4.6% Edit on ShuttleSet. This advantage stems from UMEG-Net's incorporation of object and environmental cues, as well as its compact architecture, with the fewest parameters (2.2M) among all baselines.

(3) Multimodal distillation. UMEG-Net $_{distill}$ further enhances performance and robustness as shown in Table 1(c). Compared to UMEG-Net, it improves on average 5.8% in F1evt and 6.7% in Edit score. These improvements demonstrate the effectiveness of multimodal distillation in capturing complementary visual representations from RGB inputs.

(4) Various k-clip settings. Figure 3 illustrates performance trends for representative PES baselines and our proposed models across *various k-clip samples*. The figure clearly indicates that UMEG-Net consistently outperforms existing methods under all supervision levels, and
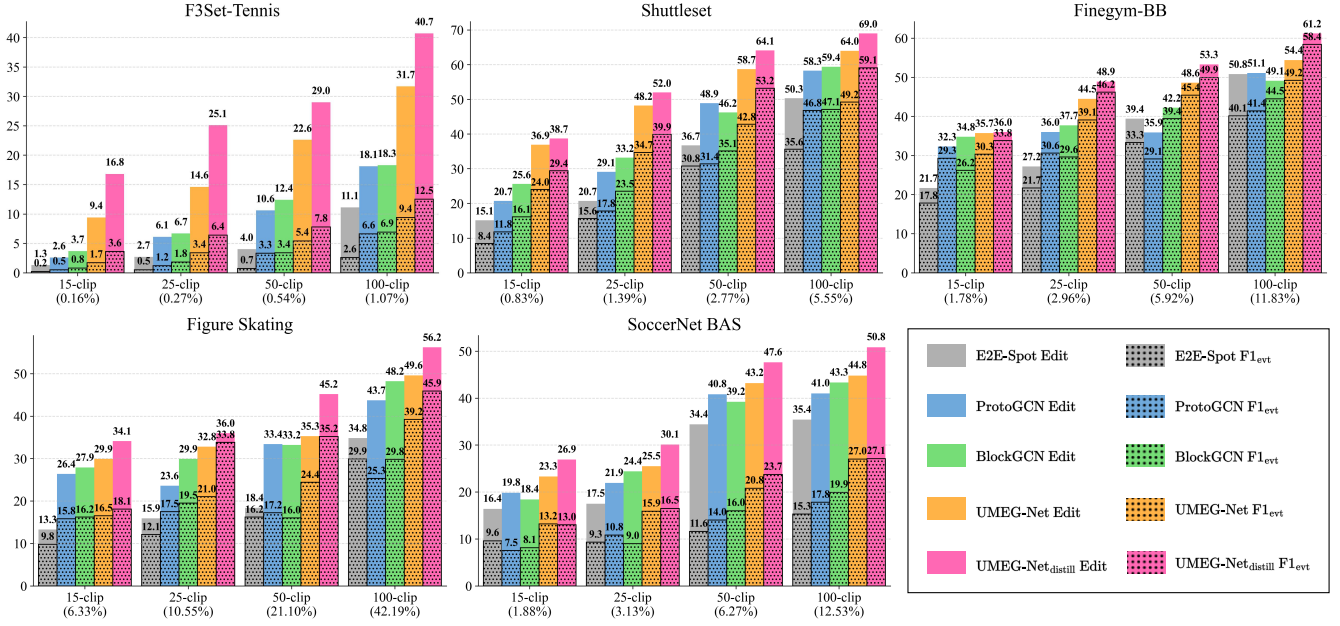
Figure 3: $F1_{evt}$ and Edit scores under few-shot (k-clip) training. Percentages indicate the fraction of the full training set.

UMEG-Net$_{distill}$ provided further improvement. These results confirm the robustness and effectiveness of our approach across diverse few-shot scenarios.

Ablation Studies

We conduct ablation studies to evaluate the effectiveness of our proposed framework in few-shot settings, primarily focusing on the 100-clip setup unless otherwise specified.

Effect of Graph Entities To assess the impact of different entity types, we analyze graph structures incorporating various object configurations, as shown in Table 1(d). Specifically, "pose × N", "ball", and "court" represent $N \times K$ human joint keypoints, a single keypoint for the ball, and four keypoints for court corners, respectively. The results indicate that incorporating ball or court information improves performance over using pose keypoints alone, while combining all entity types yields the best performance.

Temporal Module Configuration We examine the choice of Δ in UMEG-Net to see how it affects overall performance. We evaluate the effectiveness of different scales in temporal module by shifting $\Delta \in \{1\}$, $\Delta \in \{1, 2\}$, and compare with our default setting of $\Delta \in \{1, 2, 4\}$. As presented in Table 1(e), the performance drops with less temporal shift scales. However, further increasing the scales does not necessarily increase the performance.

Self-Supervise vs. UMEG-Net$_{distill}$ To validate the effectiveness of our multimodal distillation from UMEG-Net to visual encoders, we compare it with a visual-based self-supervise alternative approach, where the visual models is pretrained on unlabeled domain data using contrastive learning (Chen et al. 2020) and fine-tuned on k-clip labeled samples. As shown in Table 1(f), although the contrastive learning approach show notable improvement compared to visual-based method trained only on labeled data, our UMEG-Net$_{distill}$ substantially outperforms it.

Full Supervision Performance We also compare our UMEG-Net with existing RGB-based PES methods (E2E-Spot) under full supervision, where all of labeled clips in the training set are available. As shown in Table 1(g), our model is competitive even in the high-data regime. It shows better performance in 3 out of 5 datasets. This shows that UMEG-Net is not limited to few-shot regimes.

Conclusion

In this paper, we introduced the task of few-shot Precise Event Spotting (PES), addressing the critical challenge of frame-accurate event detection in sports videos with limited labeled data. To tackle this, we proposed a unified multi-entity graph representation that flexibly integrates human skeletons, sport-specific object keypoints, and contextual landmarks, enabling richer modeling of complex interactions within diverse sports scenes. Leveraging this representation, we developed UMEG-Net, a lightweight yet powerful graph-based architecture utilizing spatial graph convolutions combined with a parameter-free multi-scale temporal shift mechanism, enhanced further by multimodal knowledge distillation. Extensive experiments across five fine-grained sports event datasets demonstrated that our approach significantly outperforms existing methods in few-shot scenarios. These results confirm the effectiveness of structured graph representations and multimodal learning strategies in addressing annotation scarcity and generalization challenges in precise event spotting. Future work can extend our approach to handle events with weak or non-entity cues, improving robustness beyond entity-driven scenarios.

Acknowledgments

References

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709*.

Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13359–13368.

Chen, Y.-J.; and Wang, Y.-S. 2023. Tracknetv3: Enhancing shuttlecock tracking with augmentations and trajectory rectification. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, 1–7.

Cioppa, A.; Giancola, S.; Somers, V.; and et al. 2024. SoccerNet 2024 Challenges Results. arXiv:2409.10587.

Decroos, T.; Bransen, L.; Van Haaren, J.; and Davis, J. 2019. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 1851–1861.

Dey, R.; and Salem, F. M. 2017. Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, 1597–1600. IEEE.

Dong, J. S.; Jiang, K.; Liu, Z.; Dong, C.; Hou, Z.; Hundal, R. S.; Guo, J.; and Lin, Y. 2023. Sports analytics using probabilistic model checking and deep learning. In *2023 27th International Conference on Engineering of Complex Computer Systems (ICECCS)*, 7–11. IEEE.

Duan, H.; Wang, J.; Chen, K.; and Lin, D. 2022. Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, 7351–7354.

Dwyer, B.; Nelson, J.; Hansen, T.; et al. 2025. Roboflow (Version 1.0) [Software]. Computer vision.

Einfalt, M.; Dampeyrou, C.; Zecha, D.; and Lienhart, R. 2019. Frame-level event detection in athletics videos with pose-based convolutional sequence networks. In *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*, 42–50.

Giancola, S.; Amine, M.; Dghaily, T.; and Ghanem, B. 2018. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1711–1721.

Girdhar, R.; Gkioxari, G.; Torresani, L.; Paluri, M.; and Tran, D. 2018. Detect-and-track: Efficient pose estimation in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 350–359.

Hong, J.; Fisher, M.; Gharbi, M.; and Fatahalian, K. 2021. Video pose distillation for few-shot, fine-grained sports action recognition. In *ICCV*.

Hong, J.; Zhang, H.; Gharbi, M.; Fisher, M.; and Fatahalian, K. 2022. Spotting temporally precise, fine-grained events in video. In *ECCV*, 33–51. Springer.

Huang, Y.-C.; Liao, I.-N.; Chen, C.-H.; İk, T.-U.; and Peng, W.-C. 2019. TrackNet: a deep learning network for tracking high-speed and tiny objects in sports applications. In *16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–8. IEEE.

Ibh, M.; Grasshof, S.; Witzner, D.; and Madeleine, P. 2023. TemPose: a new skeleton-based transformer model designed for fine-grained motion recognition in badminton. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5199–5208.

Jiang, K.; Izadi, M.; Liu, Z.; and Dong, J. S. 2020. Deep Learning Application in Broadcast Tennis Video Annotation. In *2020 25th International Conference on Engineering of Complex Computer Systems (ICECCS)*, 53–62. IEEE.

Jiang, K.; Li, J.; Liu, Z.; and Dong, C. 2023. Court Detection Using Masked Perspective Fields Network. In *2023 IEEE 28th Pacific Rim International Symposium on Dependable Computing (PRDC)*, 342–345. IEEE.

Jiang, K.; Liu, Z.; Wu, Q.; Ma, M.; and Dong, J. S. 2024. Tracking Small and Fast Moving Ball in Broadcast Videos Using Transfer Learning and the Enhanced Interactive Multi-motion Model. In *International Sports Analytics Conference and Exhibition*, 81–96. Springer.

Jocher, G.; Chaurasia, A.; and Qiu, J. 2023. Ultralytics YOLOv8.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Lea, C.; Flynn, M. D.; Vidal, R.; Reiter, A.; and Hager, G. D. 2017. Temporal convolutional networks for action segmentation and detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 156–165.

Li, Z.; Chang, X.; Li, Y.; and Su, J. 2024. Skeleton-based group activity recognition via spatial-temporal panoramic graph. In *European Conference on Computer Vision*, 252–269. Springer.

Lin, J.; Gan, C.; and Han, S. 2019. Tsm: Temporal shift module for efficient video understanding. In *The IEEE/CVF international conference on computer vision*, 7083–7093.

Liu, H.; Liu, Y.; Ren, M.; Wang, H.; Wang, Y.; and Sun, Z. 2025a. Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29248–29257.

Liu, P.; and Wang, J.-H. 2022. MonoTrack: Shuttle trajectory reconstruction from monocular badminton video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3513–3522.

Liu, Z.; Dong, C.; Chen, J. W.; Jiang, A. M. J.; Chen, G.; Shaikh, A. F.; Dong, T. Y.; Wang, C.; Jiang, K.; and Dong, J. S. 2025b. Analyzing the Formation Strategy in Tennis Doubles Game. *SN Computer Science*, 6(2): 100.

Liu, Z.; Dong, C.; Wang, C.; Dong, T. Y.; and Jiang, K. 2024a. Exploring team strategy dynamics in tennis doubles matches. In *International Sports Analytics Conference and Exhibition*, 104–115. Springer.

Liu, Z.; Durrani, M.; Xuan, L. Y.; Simon, J.-F.; and Deon, T. Y. F. 2024b. Strategy Analysis in NFL Using Probabilistic Reasoning. In *International Sports Analytics Conference and Exhibition*, 116–128. Springer.

Liu, Z.; Jiang, K.; Hou, Z.; Lin, Y.; and Dong, J. S. 2023. Insight analysis for tennis strategy and tactics. In *2023 IEEE International Conference on Data Mining (ICDM)*, 1175–1180. IEEE.

Liu, Z.; Jiang, K.; Ma, M.; Hou, Z.; Lin, Y.; and Dong, J. S. 2025c. F^3Set: Towards Analyzing Fast, Frequent, and Fine-grained Events from Videos. *arXiv preprint arXiv:2504.08222*.

Liu, Z.; Ma, M.; Jiang, K.; Hou, Z.; Shi, L.; and Dong, J. S. 2024c. Pcsp# denotational semantics with an application in sports analytics. In *The Application of Formal Methods: Essays Dedicated to Jim Woodcock on the Occasion of His Retirement*, 71–102. Springer.

Liu, Z.; and Su, S. 2025. Analyzing Basketball Lineups with MDP Using NBA Statistics and Player Tracking. In *Sports Analytics: Second International Conference, ISACE 2025, Shanghai, China, September 26–27, 2025, Proceedings*, 142. Springer Nature.

Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; and Ouyang, W. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, 143–152.

Nag, S.; Zhu, X.; and Xiang, T. 2021. Few-shot temporal action localization with query adaptive transformer. *arXiv preprint arXiv:2110.10552*.

Pappalardo, L.; Cintia, P.; Ferragina, P.; Massucco, E.; Pedreschi, D.; and Giannotti, F. 2019. PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5): 1–27.

Perez, M.; Liu, J.; and Kot, A. C. 2022. Skeleton-based relational reasoning for group activity analysis. *Pattern Recognition*, 122: 108360.

Radosavovic, I.; Kosaraju, R. P.; Girshick, R.; He, K.; and Dollár, P. 2020. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10428–10436.

Roboflow. 2024. SoccerNET. https://universe.roboflow.com/socccer-net/soccernet-uoejr. Visited on 2025-08-05.

Shao, D.; Zhao, Y.; Dai, B.; and Lin, D. 2020. Finegym: A hierarchical video dataset for fine-grained action understanding. In *The IEEE/CVF conference on computer vision and pattern recognition*, 2616–2625.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2020. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29: 9532–9545.

Sudhakaran, S.; Escalera, S.; and Lanz, O. 2020. Gate-shift networks for video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1102–1111.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5693–5703.

Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; and Qiao, Y. 2023a. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14549–14560.

Wang, W.-Y.; Huang, Y.-C.; Ik, T.-U.; and Peng, W.-C. 2023b. ShuttleSet: A Human-Annotated Stroke-Level Singles Dataset for Badminton Tactical Analysis. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5126–5136.

Wang, W.-Y.; Shuai, H.-H.; Chang, K.-S.; and Peng, W.-C. 2022. Shuttlenet: Position-aware fusion of rally progress and player styles for stroke forecasting in badminton. In *AAAI*, volume 36, 4219–4227.

Wanyan, Y.; Yang, X.; Dong, W.; and Xu, C. 2025. A Comprehensive Review of Few-shot Action Recognition. arXiv:2407.14744.

Xarles, A.; Escalera, S.; Moeslund, T. B.; and Clapés, A. 2024. T-deed: Temporal-discriminability enhancer encoder-decoder for precise event spotting in sports videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3410–3419.

Xu, H.; Baniya, A. A.; Well, S.; Bouadjenek, M. R.; Dazeley, R.; and Aryal, S. 2025. Action Spotting and Precise Event Detection in Sports: Datasets, Methods, and Challenges. arXiv:2505.03991.

Xu, J.; Rao, Y.; Yu, X.; Chen, G.; Zhou, J.; and Lu, J. 2022. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2949–2958.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yang, P.; Hu, V. T.; Mettes, P.; and Snoek, C. G. 2020. Localizing the common action among a few videos. In *ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, 505–521. Springer.

Zhou, H.; Kadav, A.; Shamsian, A.; Geng, S.; Lai, F.; Zhao, L.; Liu, T.; Kapadia, M.; and Graf, H. P. 2022. Composer: Compositional reasoning of group activity in videos with keypoint-only modality. In *European Conference on Computer Vision*, 249–266. Springer.

Zhou, Y.; Yan, X.; Cheng, Z.-Q.; Yan, Y.; Dai, Q.; and Hua, X.-S. 2024. Blockgcn: Redefine topology awareness for skeleton-based action recognition. In *CVPR*, 2049–2058.

Appendix

Implementation Details

Architecture For RGB feature extraction, we employ VideoMAEv2 (Wang et al. 2023a) as the backbone, chosen for its superior classification performance. The encoder is pretrained on Kinetics-710 (Kay et al. 2017) and subsequently finetuned on the target sports datasets. VideoMAEv2 divides long video clips into slices and processes each slice independently, with a default slice length of 16 frames. However, in precise event spotting (PES), frame-level granularity is essential due to the short duration of sports events. Therefore, we set the slice length to 2 frames. Each slice, shaped $H \times W \times 2$ ($H = 224$, $W = 224$), is first passed through a 3D convolutional layer, resulting in 16×16 feature patches per slice. Each patch encodes the spatial features of its corresponding region. The sequence of patches is then concatenated and processed by a transformer-based network, which models relationships among spatial regions within the slice. The temporal features of the video are further modeled using the spatial features from all slices. After spatial modeling, we obtain features of shape $T/2 \times 196 \times d''$, which are temporally upsampled by a factor of 2 to $T \times 196 \times d''$. These features are subsequently fed into a temporal block to capture long-term dependencies and are used by the event localizer and classifier to predict frame-wise events.

Training For UMEG-Net, we conduct dense, per-frame classification to detect event types and precise timestamps. Due to severe class imbalance—event frames account for less than 3% of the data—the loss for foreground classes is increased fivefold to mitigate this disparity. The models are optimized using AdamW with a cosine annealing learning rate schedule, and training is performed on an RTX 4090 GPU. UMEG-Net is trained end-to-end on the limited labeled data for 50 epochs (30 epochs for ShuttleSet due to its smaller size), with an initial learning rate of 0.001 and three linear warm-up steps before cosine decay. Model selection is based on the best validation performance.

For the multimodal distillation stage, the trained UMEG-Net serves as the teacher to "distill" knowledge to the video encoder using a large set of unlabeled clips under weak supervision. UMEG-Net$_{distill}$ is trained on both labeled and unlabeled data for 50 epochs (30 for ShuttleSet) with an initial learning rate of 0.0001. Following distillation, the video encoder parameters are frozen, and only the temporal block, event localizer, and classifier are finetuned on the labeled data to adapt to the PES task. This finetuning is conducted for 10 epochs with a learning rate of 0.001.

Dataset Details

This section provides additional information on the datasets used in our experiments, as well as details on the extraction of relevant keypoints.

F^{3}Set-Tennis is a large-scale PES dataset proposed by Liu et. al. (Liu et al. 2025c). It consists of 11,584 clips from 114 professional tennis matches, with each clip containing 1 to 34 shots. Each shot is annotated with the exact frame of racket-ball contact and its corresponding event type. We focus on all 8 sub-classes and 1,108 event types. Detailed statistics are shown in Table2.

For this dataset, we extract keypoint information comprising players' 2D poses, ball positions, and court corners. The near-side player is identified using a detect-and-track algorithm (Girdhar et al. 2018), selecting the trajectory closest to the bottom of the video frame. To distinguish the far-side player from line judges, we apply YOLOv8 (Jocher, Chaurasia, and Qiu 2023) to detect individuals above the net and assign as the far-side player the lowest detected person within three meters of the court boundary. Both players' 2D poses and bounding boxes are estimated with YOLOv8. To account for camera motion, we detect court lines using the method of Jiang et al. (Jiang et al. 2023) and compute a homography to map the image plane to the court's ground plane. Assuming the player's feet are on the ground, we localize each player by projecting the bottom center of their bounding box onto the court. Ball positions in each frame are detected via (Jiang et al. 2024).

ShuttleSet is a publicly available badminton singles dataset featuring stroke-level annotations (Wang et al. 2023b). It comprises 104 sets, 3,685 rallies, and 36,492 strokes across 44 matches played between 2018 and 2021 by 27 top-ranking men's and women's singles players. While originally intended for tactical analysis, the dataset also provides detailed stroke types, precise stroke timestamps, and corresponding videos, making it suitable for the PES task. For our study, we construct the ShuttleSet dataset, which includes 3,685 clips (rallies), with an average clip length of 10.9 seconds and 10.5 shots per rally. Each shot is annotated with the exact racket–shuttle contact frame and its event type (36 categories). Detailed statistics are shown in Table3.

We extract keypoint information in ShuttleSet similar to F^{3}Set-Tennis, including players' 2D poses, shuttlecock positions, and court corners. Court corners are detected using the approach described in (Liu and Wang 2022), and shuttlecock tracking is performed with TrackNetV3 (Chen and Wang 2023). For player tracking, we use YOLOv8 (Jocher, Chaurasia, and Qiu 2023) for human detection, filtering out any detected players whose feet are not within the court, and distinguish near and far players based on their distance to the camera. The 2D poses of detected players are estimated using a pre-trained HRNet (Sun et al. 2019).

FineGym is a gymnastics dataset designed for fine-grained action understanding (Shao et al. 2020). The original annotations specify the start and end times of each action, which we treat as discrete events (e.g., "balance beam dismount start" and "balance beam dismount end") in line with (Hong et al. 2022). Our study focuses on the balance beam subset, referred to as FineGym-BB, which consists of 1,112 routines from 142 matches, with an average clip duration of 92 seconds and 24.8 events per clip. Event distributions are detailed in Table 4. For FineGym-BB, we extract 2D human poses of the single gymnastic athlete in each video clip using (Sun et al. 2019).

Table 2: Distribution of elements across sub-classes in the F^3Set-Tennis.

Sub-Class	Element	Count	Proportion (%)
sc_1	near	21,467	50.1%
	far	21,362	49.9%
sc_2	deuce	14,474	33.8%
	ad	16,310	38.1%
	middle	12,045	28.1%
sc_3	forehand	27,802	64.9%
	backhand	15,027	35.1%
sc_4	serve	11,584	27.0%
	return	8,216	19.2%
	stroke	23,029	53.8%
sc_5	T	4,428	10.3%
	Body	2,241	5.2%
	Wide	4,915	11.5%
	cross-court	11,933	27.9%
	down the line	3,521	8.2%
	down the middle	11,040	25.8%
	inside-in	608	1.4%
	inside-out	4,143	9.7%
sc_6	ground stroke	38,287	89.4%
	slice	3,358	7.8%
	volley	497	1.2%
	lob	334	0.8%
	drop	236	0.5%
	smash	117	0.3%
sc_7	approach	964	2.3%
	non-approach	41,865	97.7%
sc_8	in-bound	31,245	73.0%
	winner	3,734	8.7%
	forced error	2,808	6.5%
	unforced error	5,042	11.8%

Table 3: Distribution of event types in ShuttleSet.

Event type	Count
far-end player back-court-drive	169
far-end player clear	905
far-end player cross-court-net-shot	479
far-end player defensive-return-drive	99
far-end player defensive-return-lob	82
far-end player drive	233
far-end player driven-flight	18
far-end player drop	711
far-end player lob	1,710
far-end player long-service	159
far-end player net-shot	2,157
far-end player passive-drop	455
far-end player push	1,014
far-end player return-net	1,180
far-end player rush	190
far-end player short-service	946
far-end player smash	937
far-end player wrist-smash	586
near-end player back-court-drive	153
near-end player clear	925
near-end player cross-court-net-shot	467
near-end player defensive-return-drive	173
near-end player defensive-return-lob	102
near-end player drive	251
near-end player driven-flight	19
near-end player drop	681
near-end player lob	1,860
near-end player long-service	201
near-end player net-shot	2,091
near-end player passive-drop	434
near-end player push	990
near-end player return-net	1,277
near-end player rush	133
near-end player short-service	978
near-end player smash	773
near-end player wrist-smash	534
Total	24,072

Figure Skating includes 11 videos covering 371 short program performances. Following (Hong et al. 2022), the dataset defines 20 event types corresponding to the take-off and landing frames of 10 jump and flying spin classes (e.g., "axel take-off," "flip landing"). Each program lasts 170.7 seconds on average, containing approximately 10 events per performance. Event distributions are detailed in Table 5. Similarly, we only extract 2D human poses of the single figure skating athlete in each video clip using (Sun et al. 2019).

SoccerNet Ball Action Spotting (BAS) focuses on identifying both the timing and type of ball-related actions across 12 classes (Cioppa et al. 2024), with each action annotated by a single timestamp. The classes include Pass, Drive, Header, High Pass, Out, Cross, Throw In, Shot, Ball Player Block, Player Successful Tackle, Free Kick, and Goal. The original dataset consists of seven untrimmed broadcast videos of full English Football League matches, which contain many irrelevant scenes beyond key ball action events. To address this, we segment the videos to retain only segments containing ball action events.

For the resulting SoccerNet BAS clips, we extract 2D human poses and soccer ball positions. Soccer ball detection is performed using a YOLOv8 model pre-trained on an annotated Roboflow dataset (Roboflow 2024), while human 2D poses are estimated in a top-down manner using HRNet (Sun et al. 2019).

Baseline Methods

This section provides additional implementation details for the baseline methods discussed in *Experimental Results*.

E2E-Spot (Hong et al. 2021) E2E-Spot is an end-to-end deep learning framework for temporally precise spotting of fine-grained events in video, defined as predicting the ex-

Table 4: Distribution of event types in FineGym-BB.

Event type	Count
BB-dismounts start	1,112
BB-dismounts end	1,112
BB-flight-handspring start	2,714
BB-flight-handspring end	2,714
BB-flight-salto start	4,123
BB-flight-salto end	4,123
BB-leap-jump-hop start	4,602
BB-leap-jump-hop end	4,602
BB-turns start	1,265
BB-turns end	1,265
Total	27,632

Table 5: Distribution of event types in Figure Skating.

Event type	Count
axel takeoff	371
axel landing	371
flip takeoff	184
flip landing	184
flying-camel takeoff	216
flying-camel landing	216
flying-sit takeoff	151
flying-sit landing	151
flying-upright takeoff	6
flying-upright landing	6
loop takeoff	94
loop landing	94
lutz takeoff	248
lutz landing	248
salchow takeoff	61
salchow landing	61
toe-loop takeoff	504
toe-loop landing	504
Total	3,670

Table 6: Distribution of event types in Figure Skating.

Event type	Count
Pass	4,955
Drive	4,274
Head	707
High Pass	756
Out	550
Cross	260
Throw In	359
Shot	168
Ball Player Block	222
Player Successful Tackle	74
Free Kick	19
Goal	13
Total	12,357

act frame (within 1–2 frames tolerance) when an event occurs. It integrates a per-frame CNN (e.g. RegNet-Y (Radosavovic et al. 2020) with GSM (Sudhakaran, Escalera, and Lanz 2020)) to efficiently process hundreds of consecutive frames and a lightweight bidirectional GRU (Dey and Salem 2017) to model long-term temporal context. Unlike prior two-stage methods (feature extraction followed by separate head training), E2E-Spot jointly learns spatial–temporal representations directly from raw pixels under end-to-end supervision, enabling both fine-grained motion sensitivity and global temporal reasoning.

TDEED (Xarles et al. 2024) is a PES model that improves upon baseline methods (i.e., E2E-Spot) by explicitly enhancing frame-level discriminability and preserving high temporal resolution across multiple scales. The architecture integrates a feature extractor (e.g., RegNet-Y with local temporal modules) to produce per-frame tokens, followed by a temporal encoder-decoder that downsamples and then restores the original temporal resolution via skip-connections. Within this module, Scalable-Granularity Perception (SGP) layers, and in particular the SGP-Mixer variants, boost token discriminability by reducing similarity among adjacent frames. The combined architecture allows T-DEED to model both local and global temporal context while retaining precision.

F^{3}ED (Liu et al. 2025c) is an end-to-end deep learning framework devised to detect and timestamp sequences of fast, frequent, fine-grained (F^3) events, particularly suited to sports domains such as tennis, where events are brief, rapidly occurring, and highly detailed. The model begins by encoding consecutive video frames with a visual backbone, yielding dense per-frame features. A contextual refinement module then processes this sequence, producing precise temporal predictions for event sequences while preserving positional accuracy at the frame level.

MSG3D (Liu et al. 2020) extends STGCN (Yan, Xiong, and Lin 2018) by introducing multi-scale graph convolutions, enabling hierarchical feature extraction across different joint neighborhoods. It learns both local and global skeletal dependencies by applying multiple graph convolutions at varying scales. MSG3D enhances pose-based action recognition by capturing complex motion structures across multiple granularities.

AAGCN (Shi et al. 2020) incorporates adaptive adjacency learning to dynamically refine graph connections based on input features. Unlike STGCN, which uses a predefined skeleton topology, AAGCN learns data-driven spatial dependencies, allowing more flexible representation of human motion. This improves robustness to variations in pose estimation noise and enhances action recognition performance.

CTRGCN (Chen et al. 2021) introduces channel-wise topology refinement by modeling multi-channel dependencies within the graph structure. Instead of treating each joint's features independently, CTRGCN applies cross-channel interactions to capture co-occurring motion pat-

terns. This improves feature expressiveness and enables better generalization in skeleton-based action detection.

STGCN++ (Duan et al. 2022)   is an optimized version of STGCN that enhances efficiency and representation capacity. It refines spatial-temporal graph convolutions by introducing lightweight architectural modifications, improving performance while maintaining a compact parameter size. STGCN++ is chosen as our primary skeleton-based feature extractor due to its efficiency and strong baseline performance.

BlockGCN (Zhou et al. 2024)   addresses two key limitations of standard GCNs in skeleton-based action recognition: (1) the decay of bone connectivity information when adjacency matrices are jointly optimized with network weights, and (2) the inefficiency in multi-relational modeling using ensemble or attention-based convolutions. To mitigate topology forgetting, BlockGCN incorporates a static topological encoding based on graph distances (e.g. shortest path distances between joint pairs) and a dynamic topological encoding via persistent homology analysis to capture action-specific skeletal dynamics. Additionally, it introduces BlockGCN, a refined graph convolution module that partitions feature channels into groups and applies spatial aggregation and projection within each group via block-diagonal weight matrices.

ProtoGCN (Liu et al. 2025a)   introduces a novel GCN-based approach for skeleton-based action recognition that explicitly enhances the model's capacity to distinguish actions with subtly different joint dynamics. At its core is the Prototype Reconstruction Network (PRN), which learns a set of motion prototypes encoding prototypical joint-relationship patterns. Input skeleton representations are reconstructed as a weighted combination of these prototypes, thereby emphasizing fine-grained motion cues relevant for differentiating similar actions. ProtoGCN further incorporates a Motion Topology Enhancement (MTE) module that refines the graph representation via self-attention across joints and pairwise comparisons, enhancing feature richness. A class-specific contrastive learning objective encourages separation between prototype responses across action classes, reinforcing discriminative representation learning.