Precise Event Spotting in Sports Videos: Solving Long-Range Dependency and Class Imbalance

Supplementary Material

This supplementary presents the following details which we could not include in the main paper due to space constraints:

Contents

S1. Dataset & Implementation Details

Dataset Description

We have utilized event spotting datasets like Tennis [10], Figure Skating (FS) [3], FineGym [6] and the SoccerNet V2 [1] action spotting dataset to evaluate our method. Below, we provide the details of these datasets.

SoccerNet V2 [1] is a large-scale dataset of soccer videos containing 764 hours of data from 500 games, annotated for tasks like action spotting, camera shot segmentation and boundary detection, and replay grounding. We have utilized the action spotting dataset, which designates 17 different actions as events. This data is processed at 2 FPS. Each action of this dataset is annotated with a single timestamp as per the well-established soccer rules. There are 110,458 annotations, averaging one action every 25 seconds. Due to the nature of events, there is an inherent imbalance among the classes. For example, card events are much less likely to occur than other events like fouls or throw-ins. As it can be seen, Figure S1 (a), "Red Card" and "Yellow→Red Card" have only a few samples, while the "Ball out of play" class has thousands of samples.

Tennis dataset, compiled by [4], is an extension of the dataset proposed in Vid2Player [10]. It contains 3345 clips from 28 tennis matches (9 original + 19 new) from Wimbledon and US Open tournaments. The videos are either 25 or 30 FPS frames. 19 videos were used for training and validation, whereas the remaining nine were kept for testing. The events are categorized into six classes: "Player serve ball contact", "regular swing ball contact" and "ball bounce" for near- and far-court. Out of the 1.3M frames in the dataset, only 33,791 frames (2.6%) contain precise temporal events. Imbalance can also be seen in this dataset; the "Serve" event (both far court and near court) has significantly fewer samples than other actions (Figure S1 (b)).

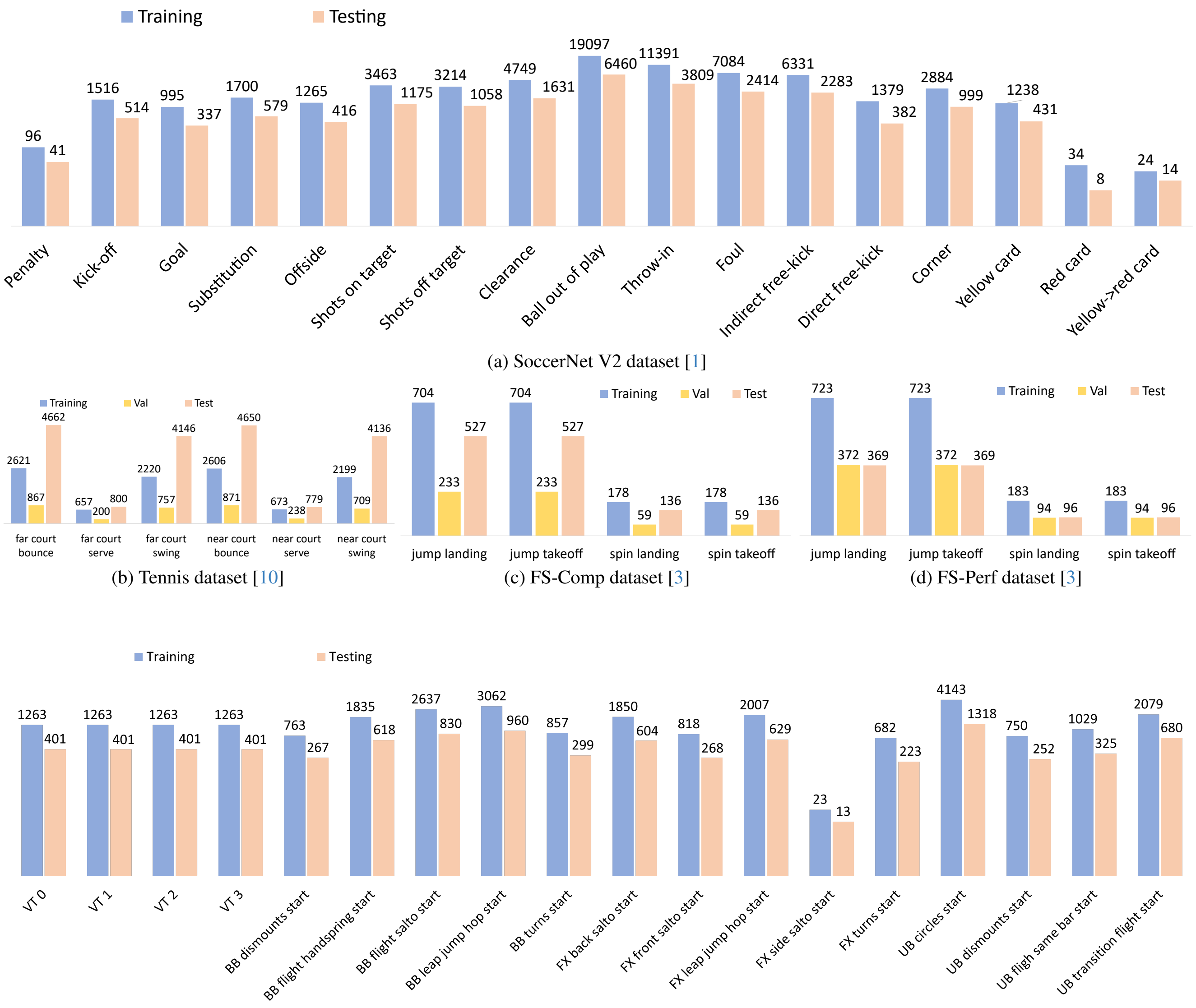
Figure Skating (FS) [3] dataset contains 11 videos featuring performances from the Winter Olympics (2010-2018) and World Championships (2017-2019). All videos are 25 FPS. The original labels have been re-annotated by [4] considering four actions: take-off and landing frames of jump and flying spins. In this dataset also, the sample count is not uniform; both "Spin" events have significantly less number of samples compared to the "Jump" event (Figure S1 (c-d)). Two splits of this dataset are considered for evaluation:

- **Competition Split (FS-Comp)**: All the videos from the 2018 season are kept for testing. So, the generalization capability of the methods to unseen videos (for example, change in background) could be evaluated.
- **Performance Split (FS-Perf)**: In this split, each competition is stratified across train, validation and test. This mainly evaluates the performance of the method when the skater changes without the unseen background situation.

FineGym [6] dataset contains 5,374 gymnastics performances, each treated as an untrimmed video. It has 32 classes, derived from a hierarchy of action categories (e.g., balance beam dismount; balance beam turns). The original annotations denote the start and end of the actions, but here, these boundaries are considered as events: "balance beam dismount start" and "balance beam dismount end". Original splits are designed for action recognition, so we use the split proposed by the [4] for the action spotting task. There are variations in the input video frame rates, so 50 and 60 FPS videos are resampled to 25 and 30, respectively. In this dataset, only the "FX side salto" event has less number of samples. In contrast, all other events have a sufficiently large number of samples (Figure S1 (e)).

Implementation Details

In addition to the implementation details provided in Subsection 4.1 in the main manuscript, we have provided additional details of the proposed model applied to training on different datasets. In one training epoch, we sample a fixed number of clips from each video. During testing, samples are taken using a sliding window of 128 frames with a 50%



(a) SoccerNet V2 dataset [1]

(b) Tennis dataset [10]

(c) FS-Comp dataset [3]

(d) FS-Perf dataset [3]

(e) FineGym dataset [6]. This dataset has total 32 classes, apart from 'VT' classes, all other classes have same number of start and end events. So, here we show the 'start' events only.

Figure S1. Class-wise distributions of SoccerNet V2, Tennis, Figure Skating (i.e., FS-Comp & FS-Perf) and FineGym datasets.

overlap. Due to that, the sample size of each epoch varies from dataset to dataset. The training configuration differs for each dataset. Below are the specific changes in the training configuration for each dataset:

- SoccerNet V2: The training data is sampled uniformly at random without overlap. Fifty clips are sampled from each video in one epoch. During training, the proposed model is trained up to 120 epochs. The frames are processed at 398 × 224 without cropping, as cropping may result in the loss of events occurring on the edges of the frame.
- Tennis: The training data is sampled at uniform random. Four clips are sampled from each video in one epoch. Like the SoccerNet V2 dataset, the frames are processed at 398 × 224 without cropping. Here, the model is trained for 100 epochs.
- Figure Skating: The training data is sampled at uniform random. Ten clips are sampled from each video in one epoch. Training is conducted using frames cropped to 224 × 224. We have observed that using non-cropped frames results in increased computation without improvement in the spotting accuracy. Here, the model is trained for 300 epochs. The same configuration has been utilized in both FS-Comp and FS-Perf.

Table S1. Analysis of various Temporal networks on Tennis [10] dataset test set. The metric of comparison is **mAP**.

Method	Tennis		
	$\delta = 0$	$\delta = 1$	$\delta = 2$
Baseline with Bi-GRU	45.34	96.10	97.70
Proposed with Bi-GRU	**61.01**	**96.21**	**97.75**
with Deformable Attention	53.71	88.50	97.33
with Bi-GRU 2 Layers	51.22	88.42	97.44
with Transformer (L1H8)	52.83	89.23	97.10
with Transformer (L2H8)	52.89	90.49	96.95
with Bi-LSTM	52.54	88.06	97.63
with MSTCN	59.90	95.26	97.37

- FineGym: In FineGym, the training data is sampled at uniform random, taking 10 clips at a time from each video. The model is trained with random crops of 224 × 224, while during testing, we center crop the video to 224 × 224. The model is trained for 100 epochs.

The model is trained on multiple GPUs with a batch size of 2 at each GPU in all datasets. The best-performing model is chosen based on the score in the validation dataset, and the same model generates the results. Additionally, we use the Soft-NMS with a window size of 20 to process the results.

We reproduce the results from the author-provided checkpoint of E2E-Spot[1], COMEDIAN[2], UGLF[3] and T-DEED[4] methods only when the corresponding results are not provided in their respective paper.

S2. Analysis on Different Temporal modules

In our proposed approach, we opted for the bidirectional GRU (Bi-GRU) as the long-range dependency module in the temporal block. The main paper shows the results of using different networks on the SoccerNet V2 [1] dataset. In Table S1, we present the results obtained from the Tennis [10] dataset. The results align with what we have observed in the main paper. There is a significant improvement in the scores across all the tolerances. Compared to the baseline method of E2E-Spot [4], which also uses Bi-GRU as the temporal module, there is an improvement of 15.67% in the $\delta = 0$ setting on the Tennis dataset. This reiterates the importance of the proposed ASTRM module and SoftIC Loss function.

S3. Further analysis on the effect of clip length

Table S2. Study of various clip lengths on Tennis dataset

Method	$\delta = 0$	$\delta = 1$	$\delta = 2$	$\delta = 4$
Clip length = 100	59.61	95.39	97.49	97.86
Clip length = 128	61.01	96.21	97.75	98.05
Clip length = 144	60.39	96.08	97.65	97.99

Table S3. Analysis with different random-seeds.

Proposed	mAP-Tight	mAP-Loose
ASTRM	67.99± 0.23	74.70± 0.40
ASTRM + ASAM	72.65± 0.01	78.40± 0.03
ASTRM + ASAM + Soft-IC loss	73.41± 0.06	78.83± 0.08

In the main paper, we analyzed the effect of clip length on the SoccerNet V2 dataset. To further study its effect, we performed additional experiments on the Tennis dataset, and the results are shown in Table S2. Clip length = 128 provides the best results, as observed in the main paper ablations. Given that some videos in the Tennis dataset contain only 160 frames, we consider clip length up to 144 for this study.

S4. Effect of Randomness

Training a neural network involves several steps where randomness plays a significant role. This starts with the initial values assigned to the weights and includes various augmentations applied to the training data. Consequently, the outcomes are influenced by these random initial conditions. To illustrate the impact of different initial conditions, we conducted experiments using three different random seeds and recorded the mean and variance of the mean Average Precision (mAP) values, as presented in Table S3. Our results indicate that using ASAM reduces the variance, supporting its optimization effectiveness.

S5. Efficiency Comparison

This paper demonstrates that our proposed method outperforms existing SOTA methods, especially in tight settings with a simpler network architecture. Here, we have quantitatively validated this claim in detail. In Table S4, we provide a quantitative analysis of its efficiency, specifically in terms of the number of parameters and the computational complexity measured in GFLOPs. For a fair comparison, all the calculations are done from the models on their respective repos with an input size of 3 × 100 × 224 × 398. Our proposed method requires 77.49 GFLOPs and 6.46 million parameters, showcasing a substantial reduction in both computational complexity and model size compared to recent SOTA methods [2, 4, 7–9] except E2E-Spot (RegNet-Y 200MF) [4] and T-DEED (RegNet-Y 200MF) [9]. Among them, only COMEDIAN (ViSwin) [2] achieves comparable results but with a significantly larger number of parameters and higher computational requirements. Similarly, Spi-

[1] https://github.com/jhong93/spot

[2] https://github.com/juliendenize/eztorch

[3] https://github.com/Fsoft-AIC/UGLF

[4] https://github.com/arturxe2/T-DEED

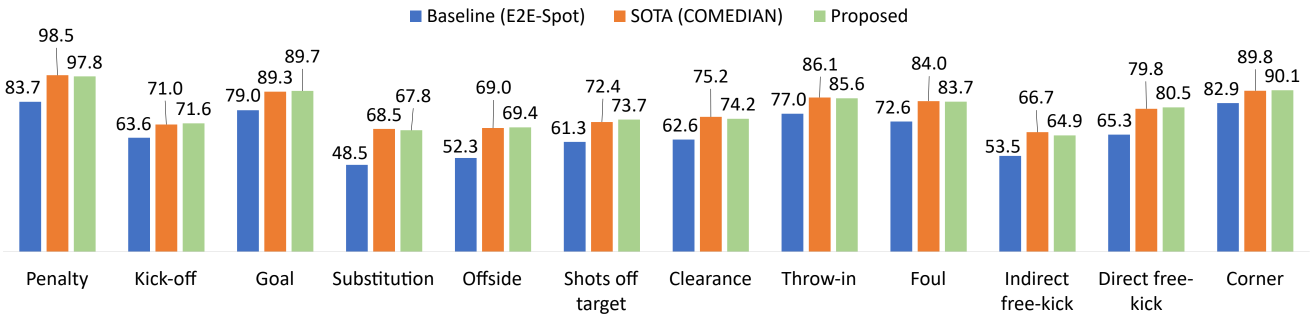
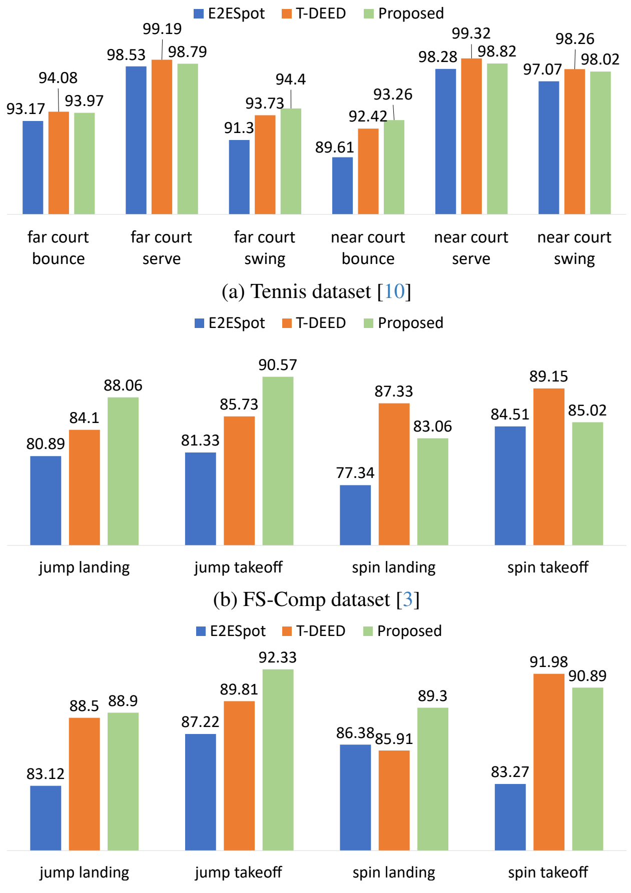


Figure S2. Per-class score comparison on tight setting in terms of **mAP** on few classes of SoccerNet V2 dataset.



(a) Tennis dataset [10]

(b) FS-Comp dataset [3]

(c) FS-Perf dataset [3]

Figure S3. Per-class score comparison on $\delta = 1$ in **mAP** on Tennis and Figure Skating (i.e., FS-Comp & FS-Perf) datasets.

Table S4. Efficiency Comparison in terms of GFLOPs and number of parameters of the proposed and recent SOTA methods. * indicate that the GFLOPs value is calculated from the temporal network only without feature extractors. Here, ASTRA model utilized the features extracted from Baidu model which is made up of 5 large networks.

Methods	GFLOPs	# of Parameters (in Millions)
E2E-Spot (RegNet-Y 200MF)	39.61	4.46
E2E-Spot (RegNet-Y 800MF)	151.4	12.64
ASTRA	8.83*	44.33
Spivak	461.89	17.46
COMEDIAN (ViSwin)	222.76	70.12
T-DEED (RegNet-Y 200MF)	21.96	16.36
T-DEED (RegNet-Y 800MF)	85.58	46.22
Proposed	60.25	6.46

vak [7] and ASTRA [8] achieve similar results only in the loose-mAP setting, even with increased parameters.

E2E-Spot (RegNet-Y 200MF), T-DEED (RegNet-Y 200MF), and our proposed method utilize similar structured networks, but some modules differ. Consequently, the proposed method has more GFLOPs than both methods. Meanwhile, the T-DEED (RegNet-Y 200MF) has a significantly larger number of parameters than the proposed method. Nonetheless, our proposed method outperforms both methods significantly. This balance between efficiency and performance is crucial for practical applications, particularly in environments with limited computational resources. This makes it an appealing choice for real-world deployments where both computational efficiency and high performance are essential.

S6. Per Class Score Comparison

In addition to the per-class score analysis shown in Figure 1 and Figure 4 in the main manuscript, we have included some additional analyses. Specifically, in Figure S2, we have presented the tight-mAP scores of the SoccerNet V2 dataset for classes that were not covered in Figure 1 of the main paper. Additionally, Figure S3 presents the per-class score analysis on $\delta = 1$ setting for the Tennis and Figure Skating (FS-Comp and FS-Perf) datasets. The per class score of the FineGym dataset is presented in Figure S4.

In Figure S2, it is evident that the proposed method achieves comparable performance with the COME-

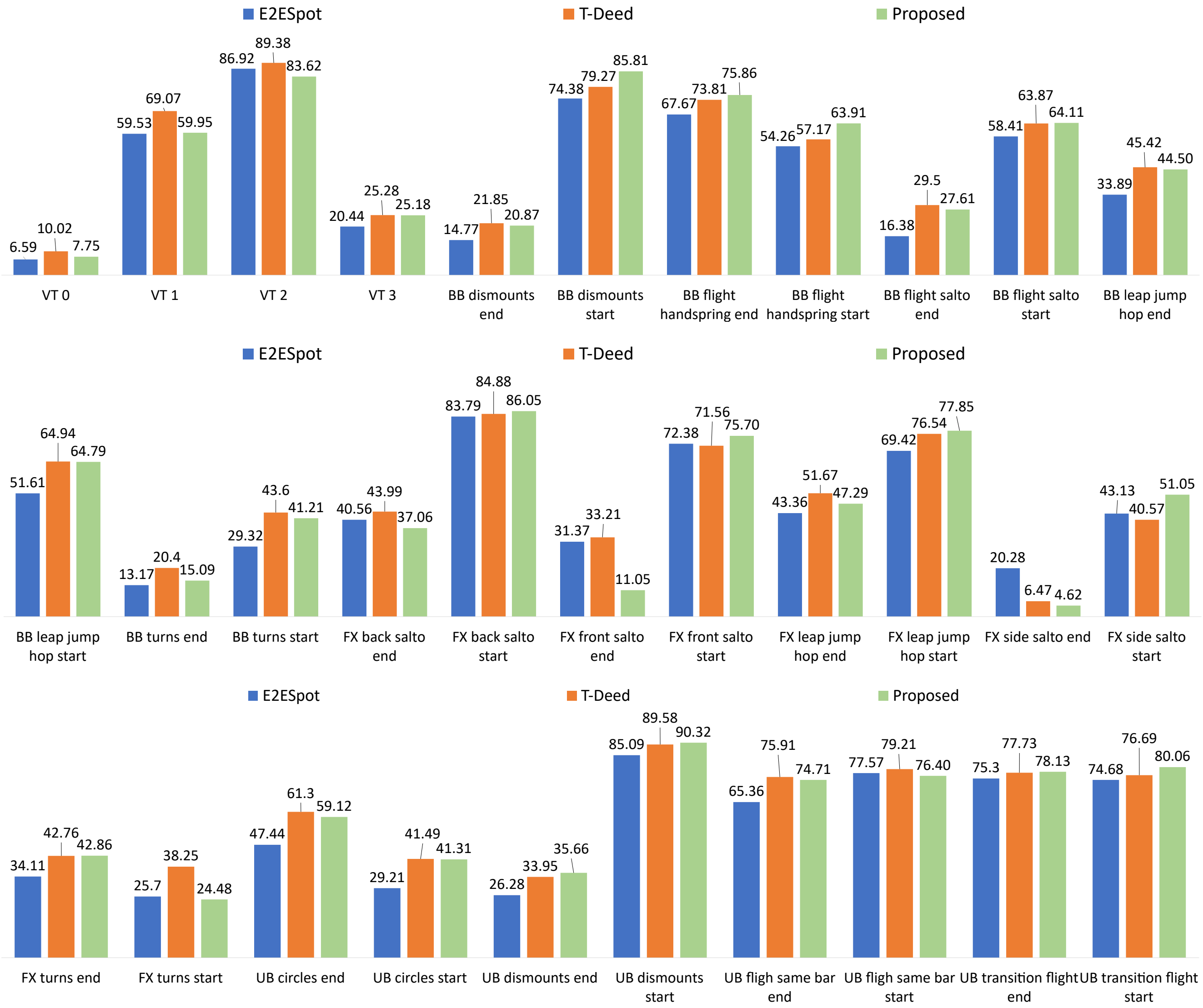


Figure S4. Per-class score comparison on $\delta = 1$ in **mAP** on FineGym [6] dataset. Graph split on multiple rows for better visualization.

DIAN [2] method despite having significantly fewer parameters and lower computational complexity. While from Figure S3 and Figure S4, it can be observed that the proposed method achieves comparable performance in most of the classes while outperforming the T-DEED [9] in many instances.

S7. Result on non-sports dataset

In this work, we primarily focused on event spotting within sports videos. However, our proposed method is not limited to sports and can be adapted for non-sports datasets. To evaluate this hypothesis, we conducted experiments using the FineAction [5] benchmark dataset.

FineAction [5] is a large-scale temporal action localization dataset with fine-grained labels. It contains 106 action classes with three levels of granularity: four coarse-level actions (Household, Personal care, Socializing-Relaxing, and Sports-Exercise) and 14 middle-level actions, in addition to the 106 fine-level actions. For our experiments, we focused on a few middle-level action classes, as using the fine-level classes would complicate the training process. We tested three scenarios: activities based on plants and two mixed environments (comprising of both plant and outdoor classes).

Similar to FineGym experiments, we adapted the actions for precise event-spotting. Specifically, we identified the start and end of each event as two separate events and noted their corresponding timestamps as the occurrence time of

Table S5. Analysis on non-sports videos under diff categories

Model	Plants		Mix (2 Classes)		Mix (3 Classes)	
	$\delta = 10$	$\delta = 50$	$\delta = 10$	$\delta = 50$	$\delta = 10$	$\delta = 50$
T-DEED	24.82	38.58	27.42	34.67	18.85	25.71
Proposed	26.45	39.36	33.98	44.09	18.84	30.95

the event. The videos in the dataset vary in resolution; however, we only used the 720p landscape videos for the training and inference, resizing them to 224p while maintaining the same frames per second (FPS). Under those conditions, we selected the data and split it into training, validation, and testing sets in a 60-20-20 ratio. We kept other training hyperparameters consistent with those used in the SoccernetV2 experiments.

For a fair comparison, we retrained the T-DEED model on the same dataset. Table S5 presents the results in terms of mAP for different tolerance levels (δ=10, 50). Our method, with minimal hyperparameter tuning, outperformed T-DEED in most cases.

S8. Limitations

In the main manuscript, we noted that our proposed method is focused on sports events, which might give the impression of limited applicability. However, it is important to clarify that none of the components of our method are specifically tailored for sports videos, as validated in the previous section. The concept of precise event spotting has primarily been defined in relation to sports, which is why the existing precise event spotting datasets predominantly consist of sports videos. On the other hand, the scarcity of datasets featuring non-sports videos has also forced precise event-spotting methods to focus on sports videos only. Moreover, everyday events are typically characterized by specific starting and ending times and rarely occur instantaneously. While we could approach the problem as detecting the precise start and end times of events, as done for FineGym and FineAction dataset, this may not always be the most appropriate solution for the intended application.

References

[1] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4508–4519, 2021. 1, 2, 3

[2] Julien Denize, Mykola Liashuha, Jaonary Rabarisoa, Astrid Orcesi, and Romain Hérault. Comedian: Self-supervised learning and knowledge distillation for action spotting using transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 530–540, 2024. 3, 5

[3] James Hong, Matthew Fisher, Michaël Gharbi, and Kayvon Fatahalian. Video pose distillation for few-shot, fine-grained sports action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9254–9263, 2021. 1, 2, 4

[4] James Hong, Haotian Zhang, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Spotting temporally precise, fine-grained events in video. In *European Conference on Computer Vision*, pages 33–51. Springer, 2022. 1, 3

[5] Yi Liu et al. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE TIP*, 31:6937–6950, 2022. 5

[6] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 5

[7] Joao VB Soares, Avijit Shah, and Topojoy Biswas. Temporally precise action spotting in soccer videos using dense detection anchors. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2796–2800. IEEE, 2022. 3, 4

[8] Artur Xarles, Sergio Escalera, Thomas B Moeslund, and Albert Clapés. Astra: An action spotting transformer for soccer videos. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pages 93–102, 2023. 4

[9] Artur Xarles, Sergio Escalera, Thomas B. Moeslund, and Albert Clapés. 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 (CVPR) Workshops*, pages 3410–3419, 2024. 3, 5

[10] Haotian Zhang, Cristobal Sciutto, Maneesh Agrawala, and Kayvon Fatahalian. Vid2player: Controllable video sprites that behave and appear like professional tennis players. *ACM Trans. Graph.*, 40(3), 2021. 1, 2, 3, 4