# FineSports: A Multi-person Hierarchical Sports Video Dataset for Fine-grained Action Understanding

Jinglin Xu[1], Guohao Zhao[2], Sibo Yin[2], Wenhao Zhou[1], Yuxin Peng[2†]
[1]School of Intelligence Science and Technology, University of Science and Technology Beijing
[2]Wangxuan Institute of Computer Technology, Peking University

PEKING UNIVERSITY
University of Science and Technology Beijing

GitHub MIPL's Website MIPL's GitHub Homepage

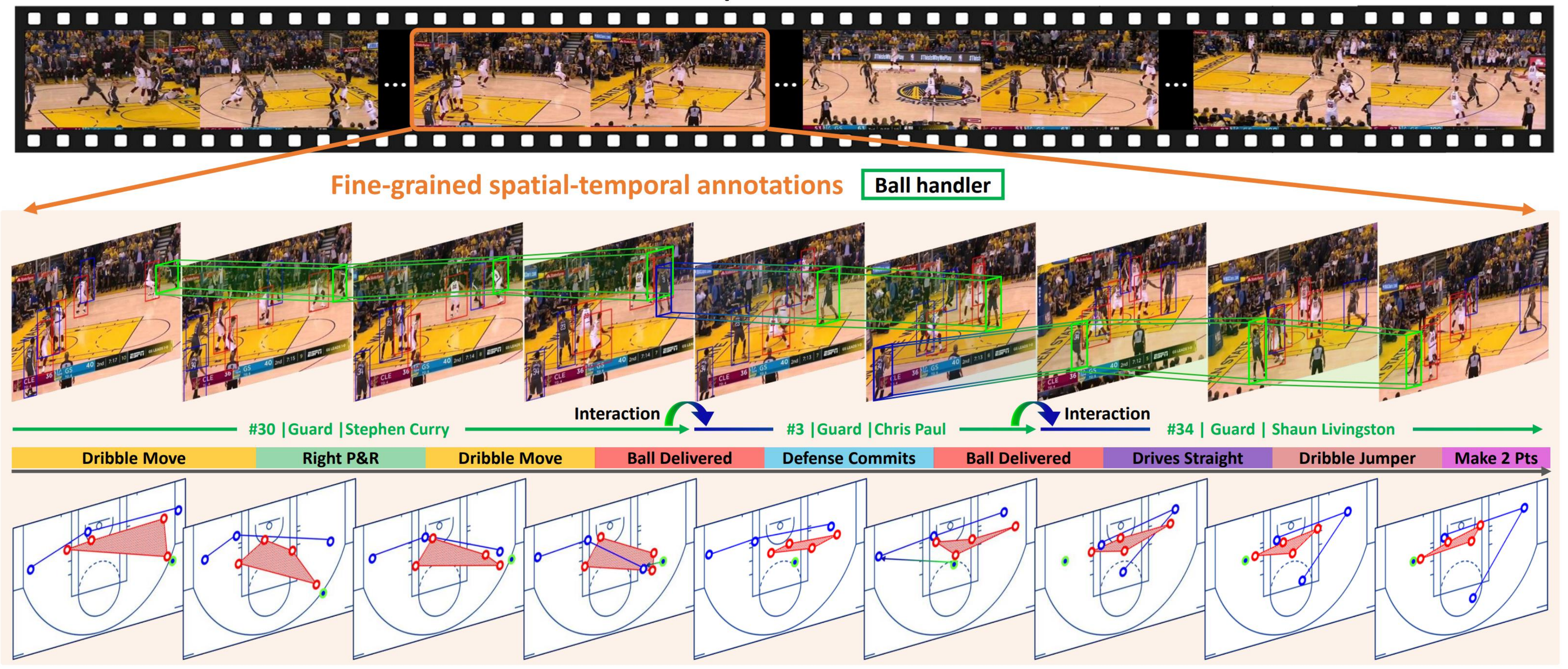
CVPR SEATTLE, WA JUNE 17-21, 2024

## Motivation

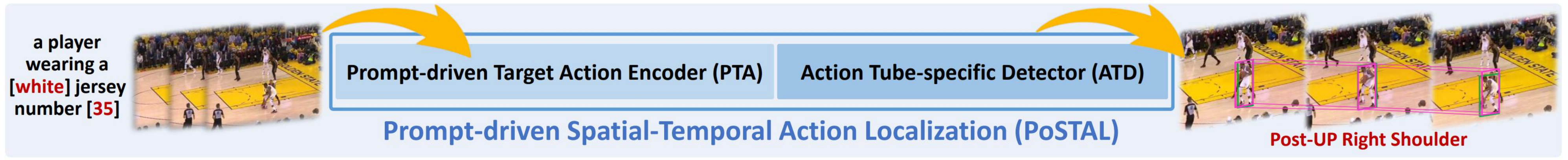
- Spatial-temporal action localization (STAL) aims to detect action tubes by a sequence of bounding boxes in space and time, as well as the corresponding action class.
- Existing video action datasets usually lack high-quality fine-grained annotations, leading to difficulties in fine-grained video understanding.
- Video understanding of team sports is challenging due to its chaotic nature. For instance, in NBA games, players' actions exhibit overlapping and rapid changing, making it difficult for fine-grained understanding of such videos.

## Contribution

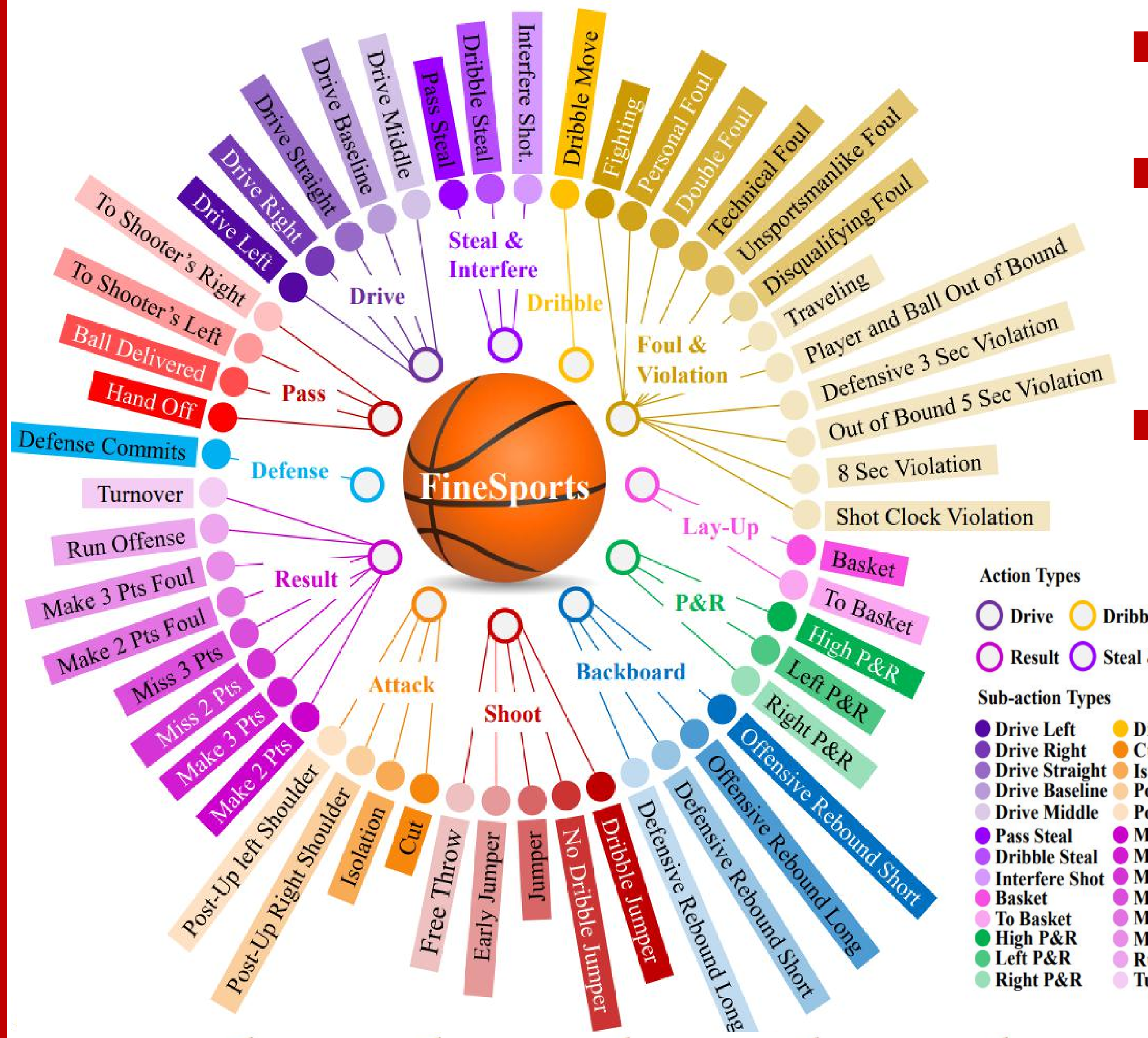
- A new multi-person sports video dataset with fine-grained annotations, **FineSports**.



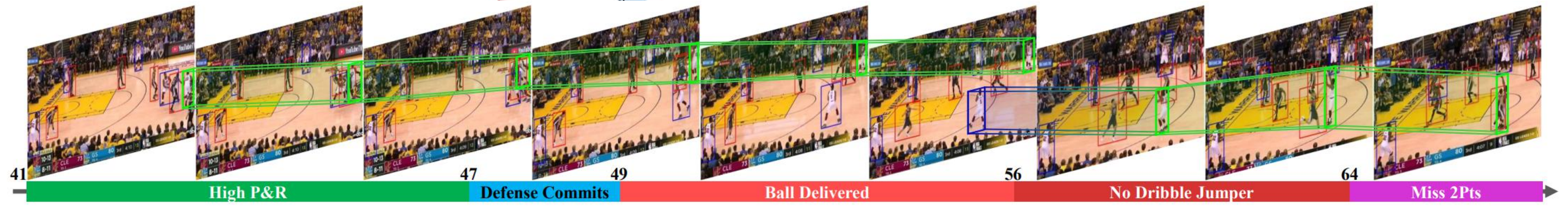

- We proposed a new prompt-driven spatial-temporal action location (STAL) approach, named **PoSTAL**.



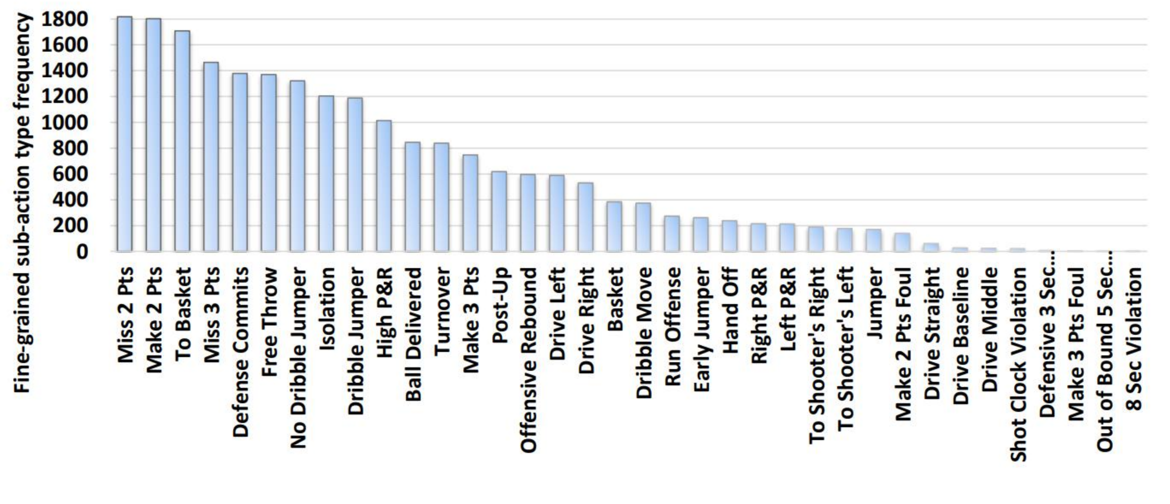

## The FineSports Dataset




- The two-level **semantic** structure of FineSports.
- An action type denotes the coarse-grained action categories the player performs.
- A sub-action type is the component of an action type.

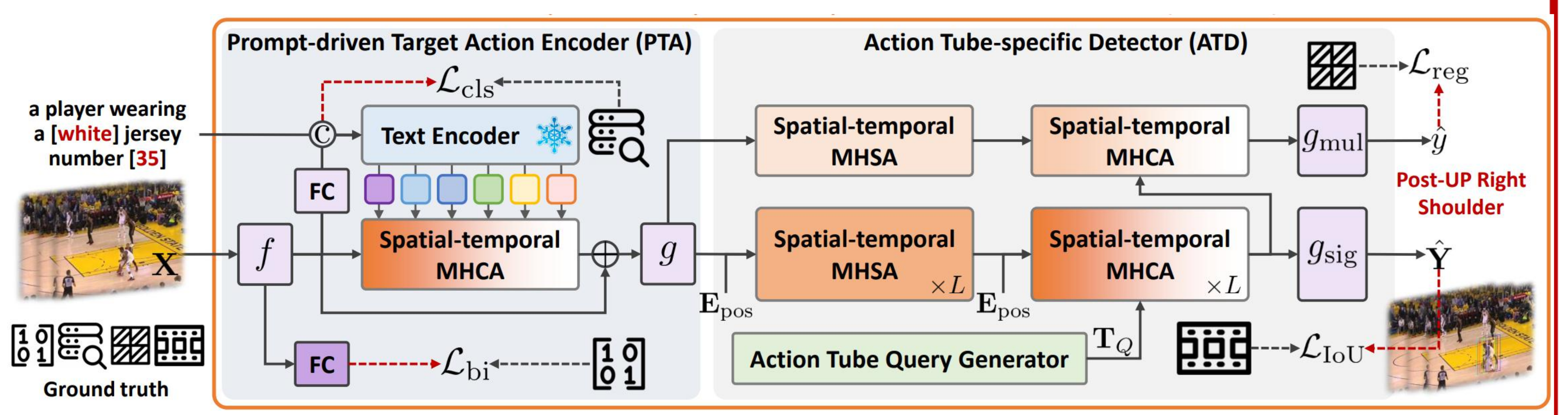



- The **spatial-temporal** structure of fine-grained action types of the target players (green bounding boxes).




- **Statistics**. FineSports contains **10,000** basketball video games, covering **12** action types and **52** sub-action types, providing **123,014** spatial bounding boxes and **32,096** temporal boundaries of associated fine-grained sub-actions.

## Method: PoSTAL




- **Prompt-driven Target Action Encoder (PTA).** Learns action representation via spatial-temporal vision-language cross-attention with the guidance of the appearance characteristics of the target player and the associated fine-grained sub-action type.

$$A_P^S = softmax\left(Q \otimes K^T/\sqrt{C'/H}\right),$$
$$X_P = g\left(X_P' + f(X)\right), X_P' = A_P^S \otimes V$$

where $X_P$ is the prompt-driven target action representation.

- **Action Tube-specific Detector (ATD).** Utilizes a single-level and a multi-level action tube-specific transformer to predict target action's spatial locations, temporal boundaries and fine-grained sub-action types.

$$X_P^E = \mathcal{E}_{sig}(X_P + E_{pos}), X_P^D = \mathcal{D}_{sig}(T_Q, X_P^E + E_{pos}),$$
$$\tilde{X}_P^D = \mathcal{D}_{mul}(\tilde{X}_P^E, X_P^D), \tilde{X}_P^E = \mathcal{E}_{mul}(\tilde{X}_P),$$
$$\hat{Y} = g_{sig}(X_P^D[-1]), \hat{y} = g_{mul}(\tilde{X}_P^D)$$

where $\hat{Y}$ is the predicted actions tubes and $\hat{y}$ is the predicted action type.

## Experiments

| Method | F@0.5 | V@0.2 | V@0.5 | Year |
|---|---|---|---|---|
| MOC [23] | 19.21 | / | / | ECCV'20 |
| TubeR [46] | 19.48 | 28.91 | 17.76 | CVPR'22 |
| **PoSTAL** (Ours) | **21.54** | **31.18** | **24.31** | |

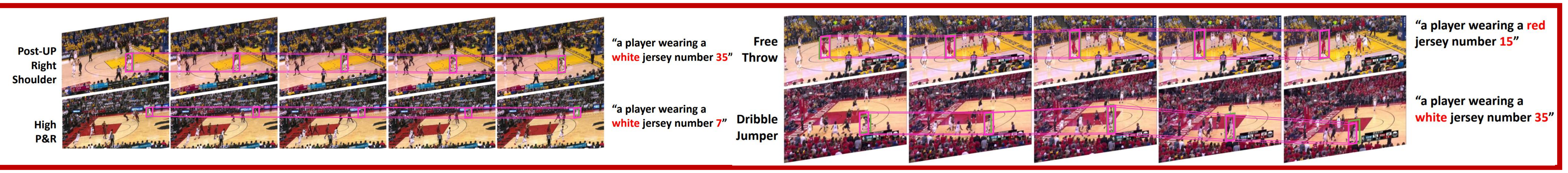
| # Tube Query ($N$) | F@0.5 | V@0.2 | V@0.5 |
|---|---|---|---|
| 2 | 20.16 | 32.72 | 21.34 |
| **6** | **21.54** | **31.18** | **24.31** |
| 10 | 20.41 | 30.54 | 19.21 |

| PTA Settings | F@0.5 | V@0.2 | V@0.5 |
|---|---|---|---|
| w/o Descriptive Words | 18.26 | 27.99 | 18.53 |
| w/o Learnable Embeddings | 18.13 | 27.91 | 17.60 |
| **PTA** | **21.54** | **31.18** | **24.31** |

**F@0.5**: frame-mAP with θ = 0.5
**V@0.2**: video-mAP with θ = 0.2
**V@0.5**: video-mAP with θ = 0.5

## Visualization

The left is action types and the right is descriptive words of target player.




[1] TubeR: Tubelet Transformer for Video Action Detection. CVPR 2022. [2] MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions. ICCV 2021.