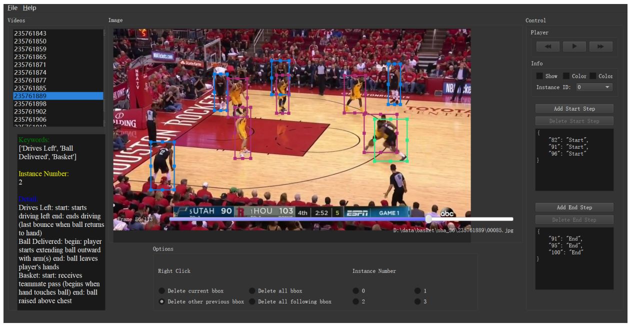
# FineSports: A Multi-person Hierarchical Sports Video Dataset for Fine-grained Action Understanding

Supplementary Material

## 1. Annotation System

Fig. 7 presents the annotation system's user interface (UI) and a detailed annotation process. We employed three professional athletes from the basketball association to help construct a lexicon for subsequent annotation. Given that the temporal boundaries of fine-grained actions are often ambiguous, professional athletes meticulously defined precise temporal boundaries for each sub-action type in the lexicon, such as "Keywords" and "Detail" in the bottom left corner of the UI. This enables crowdsourced annotators to accurately label fine-grained actions' beginning and ending frames. In addition, "Instance Number" indicates the number of target players being observed, especially for interaction between them, each described by the jersey number and color. The annotation process involves (1) localizing the temporal boundaries of the fine-grained sub-actions of the target player and (2) tagging the target player's spatial bounding box within each frame. Due to the rapid changes, unpredicted players' movements, and severe occlusions, most bounding boxes are inaccurate and require frequent manual adjustments, which is time-consuming and difficult. Therefore, we integrate MixSort-OC [4] into the annotation system to speed up the annotation process. Just click the buttons in the right part of the UI whenever the target action starts and ends.



Keywords:
['Drives Left', 'Ball Delivered', 'Basket']

Instance Number:
2

Detail:
Drives Left: start: starts driving left end: ends driving (last bounce when ball returns to hand)
Ball Delivered: begin: player starts extending ball outward with arm(s) end: ball leaves player's hands
Basket: start: receives teammate pass (begins when hand touches ball) end: ball raised above chest

Frame 86/113

D:\data\basket\nba_36\235761889\00085.jpg

Instance ID: 0

```
{
    "82": "Start",
    "91": "Start",
    "96": "Start"
}
```

```
{
    "91": "End",
    "95": "End",
    "100": "End"
}
```

○ Delete current bbox ○ Delete all bbox
● Delete other previous bbox ○ Delete all following bbox

○ 0 ○ 1
○ 2 ○ 3

Figure 7. User interface of the annotation system.

## 2. More Ablation Studies

We provide more experimental results of different ways of integrating the *text* in the PTA module, demonstrating the advantage of our PoSTAL. For all experiments, we utilize frame-mAP and video-mAP as evaluation metrics.

**Different Ways of Integrating Text in PTA.** In PTA, there are three ways of obtaining two descriptive words (i.e., Color and Number), including (1) introducing additional prediction heads (denoted as text-pred), (2) introducing the ground-truth text (denoted as text-gt), and (3) utilizing the visual features (denoted as text-visual). Concretely, (1) the text-pred way introduces two additional prediction heads, one of which predicts the descriptive word "Color" and the other of which predicts the descriptive word "Number". Each prediction head contains two MLP layers with a ReLU non-linearity, optimized by minimizing the cross-entropy loss, where "Color" has 6 categories and "Number" has 43 categories. (2) The text-gt way utilizes the description "a player wearing a [Color] jersey number [Number]" as the input both during training and testing, where "Color" and "Number" are the annotations. This way can be seen as the oracle of the PoSTAL performance. (3) The text-visual way utilizes the description "a player wearing a [Color] jersey number [Number]" as the input to guide the model encoding visual features with the characteristics of target players during training. During testing, this way utilizes the visual features of each testing video to obtain the embeddings of the descriptive words "Color" and "Number". This way, the annotations of "Color" and "Number" are used as implicit supervision to guide the model training in obtaining prompt-driven target action representations without introducing additional model parameters. In Tab. 5, we see that introducing additional prediction heads for two descriptive words (i.e., text-pred) can achieve better frame-mAP and video-mAP than utilizing the visual features (i.e., text-visual), but the text-pred way increases the model parameters. Our PoSTAL takes the text-visual way.

| Method | Metrics | | |
|---|---|---|---|
| | F@0.5 | V@0.2 | V@0.5 |
| text-pred | 21.87 | 32.38 | 24.54 |
| text-gt | 23.82 | 35.77 | 27.14 |
| **text-visual** | **21.54** | **31.18** | **24.31** |

Table 5. Ablation study on different ways of integrating the *text* in PTA. The results of our PoSTAL are highlighted in **bold** format.