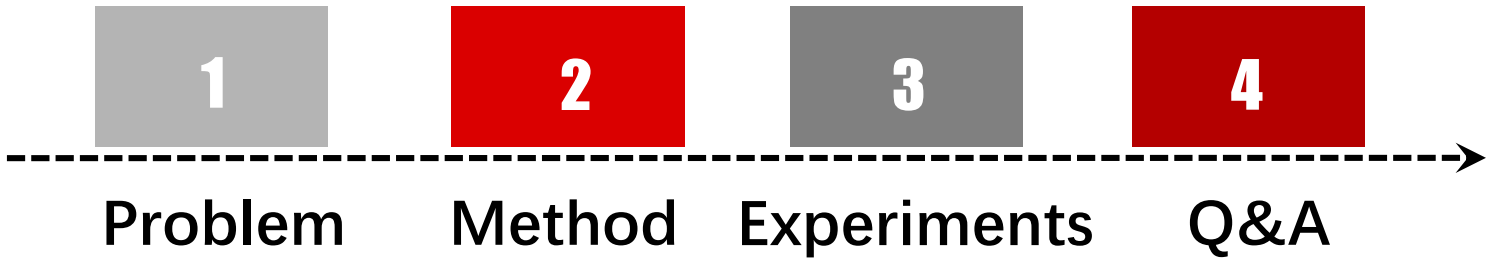
Segregated Temporal Assembly Recurrent

Networks for Weakly Supervised Multiple Action Detection

Yunlu Xu, Chengwei Zhang, Zhanzhan Cheng,
Jianwen Xie, Yi Niu , Shiliang Pu

Hikvision Research Institute, China

CONTENT



1. Problem

Task of Temporal Action Localization

frame-level annotations ?

In a fully-supervised approach, the training needs

- ✓ ~~precise start time point~~
- ✓ ~~precise end time point~~
- ✓ ~~frame-level~~ class label
  video-level

time-consuming, expensive

video-level annotations ?

Q1. What is the ideal representations?

Background

Cricket bowling

Cricket Shot

Background

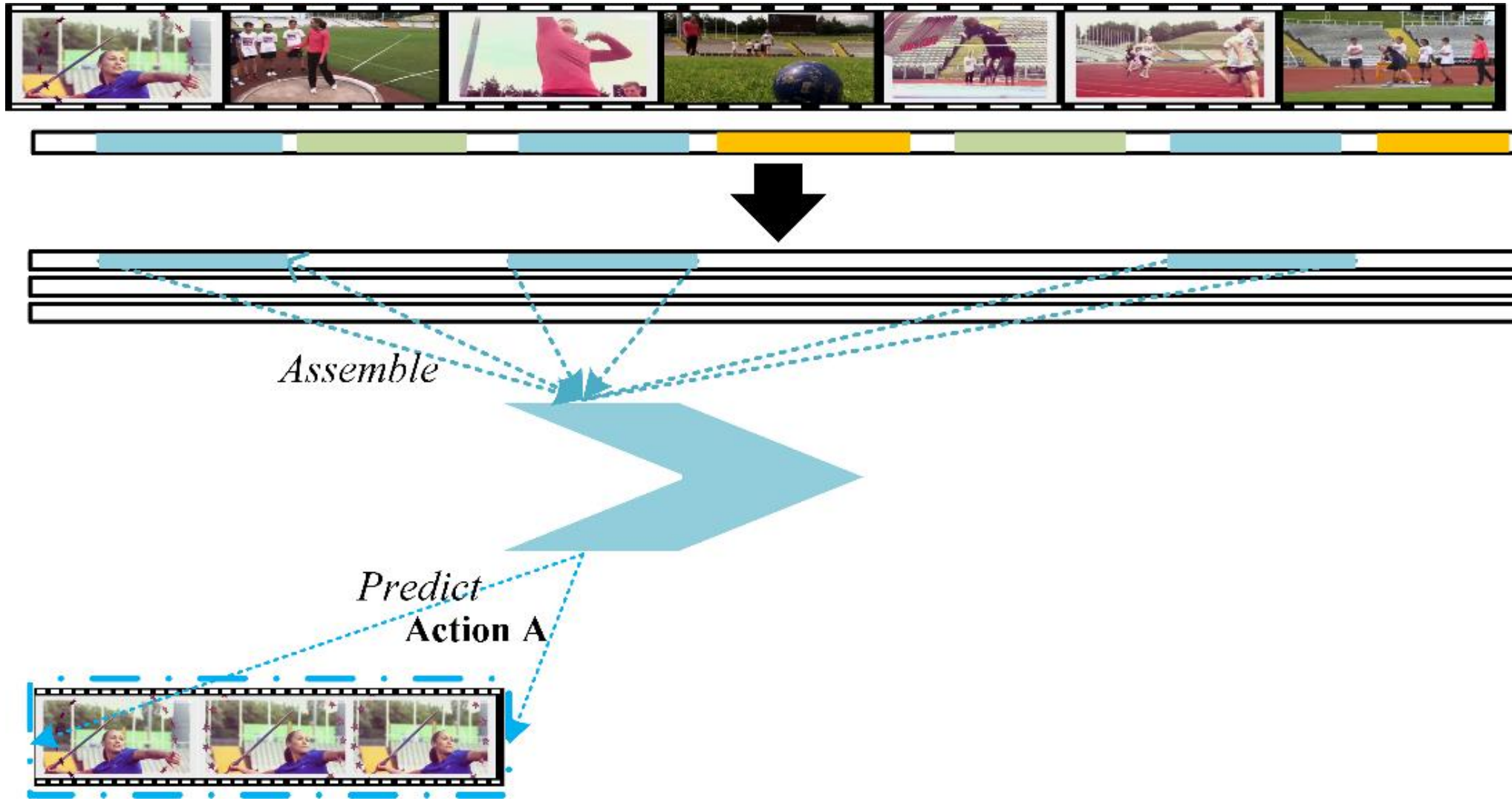
Q2. How to get the representation without frame-level annotation?

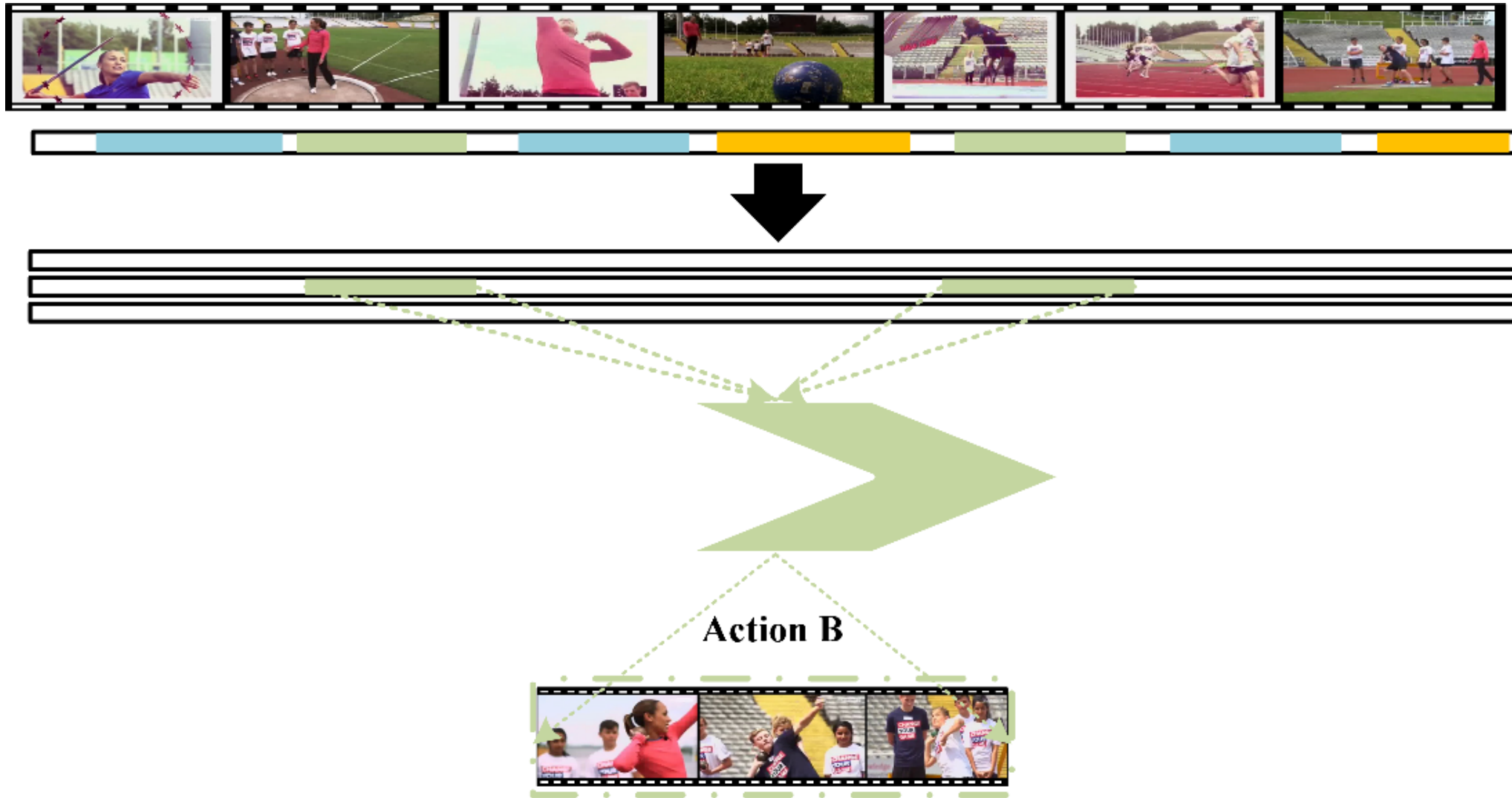
- no **interference** of backgrounds

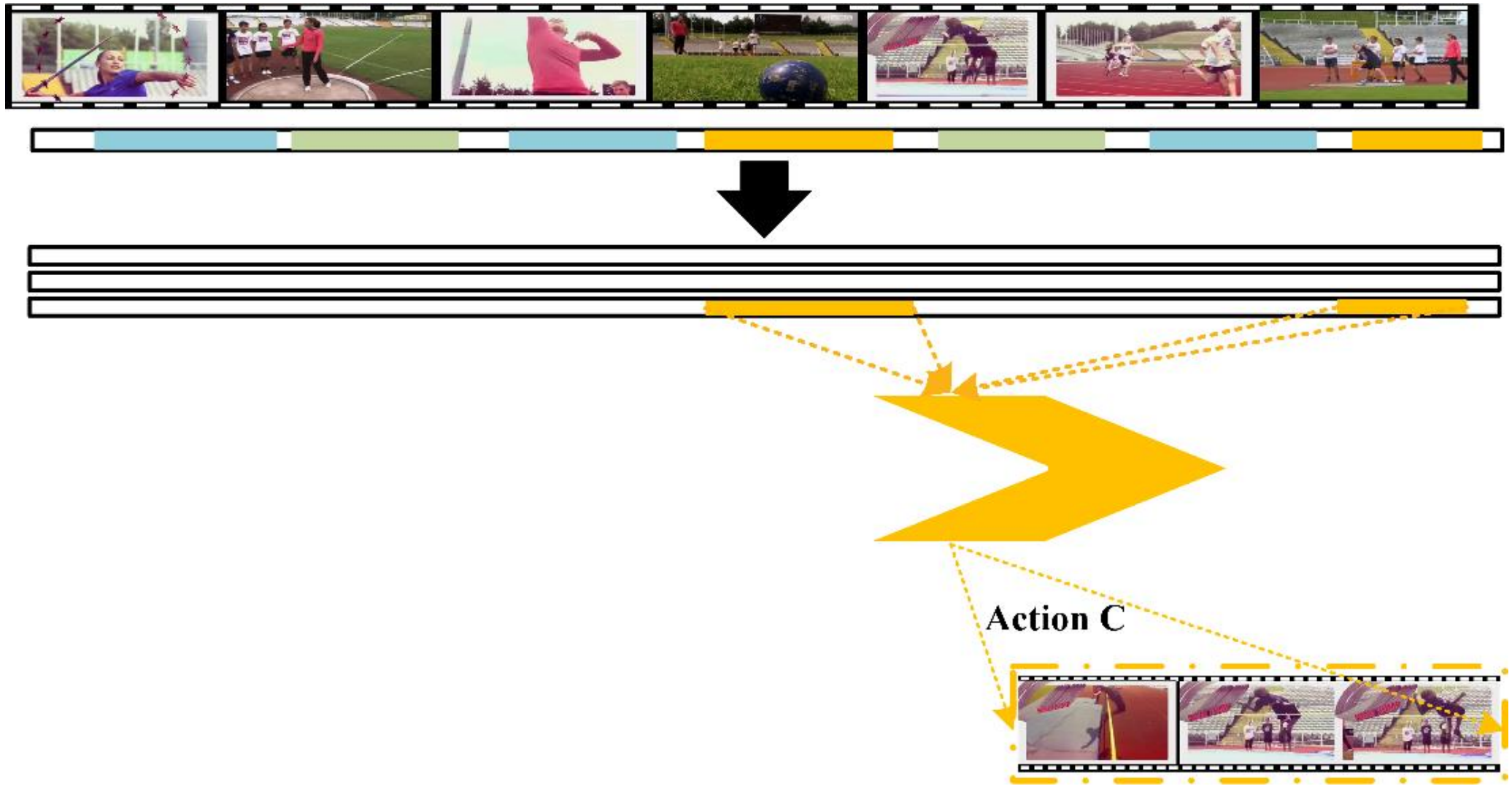
  ➡ Action assembling for separated instance-patterns

- **correlation** among actions

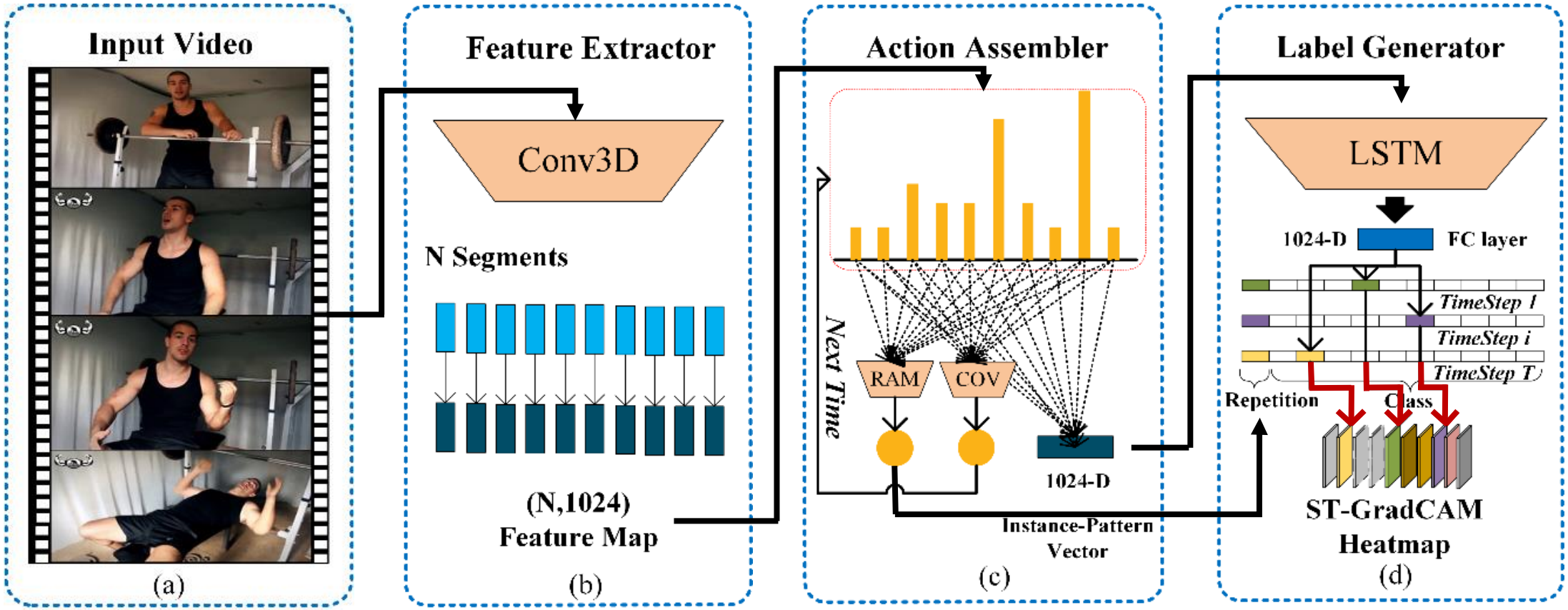
  ➡ RNN for multiple instance-pattern generation

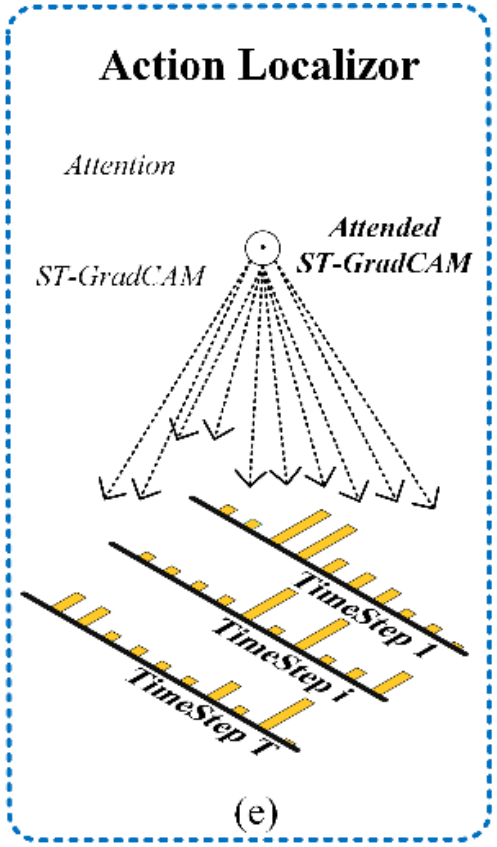

Assemble
Predict
Action A


Action B


Action C

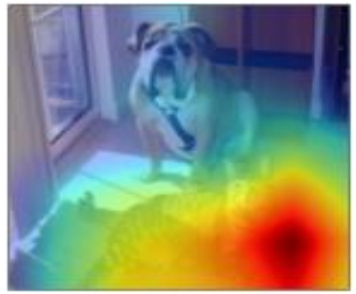
2. Method


Input Video
(a)
Feature Extractor
Conv3D
N Segments
(N,1024)
Feature Map
(b)
Action Assembler
Next Time
RAM
COV
1024-D
Instance-Pattern
Vector
(c)
Label Generator
LSTM
1024-D
FC layer
TimeStep 1
TimeStep i
TimeStep T
Repetition
Class
ST-GradCAM
Heatmap
(d)

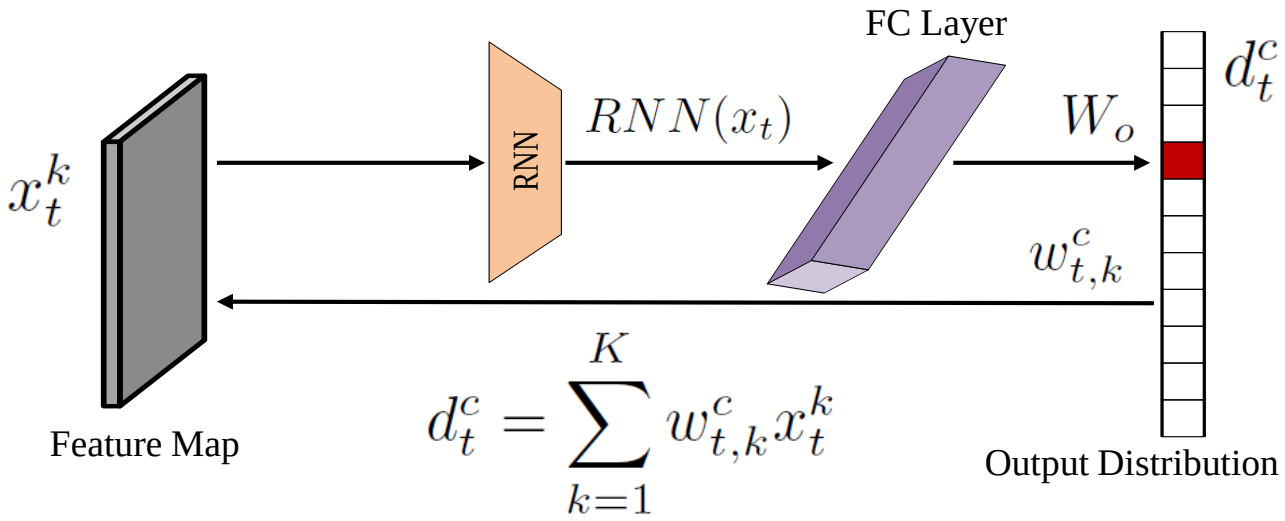


Generalization of Class Activation Mapping (CAM):

Original Image

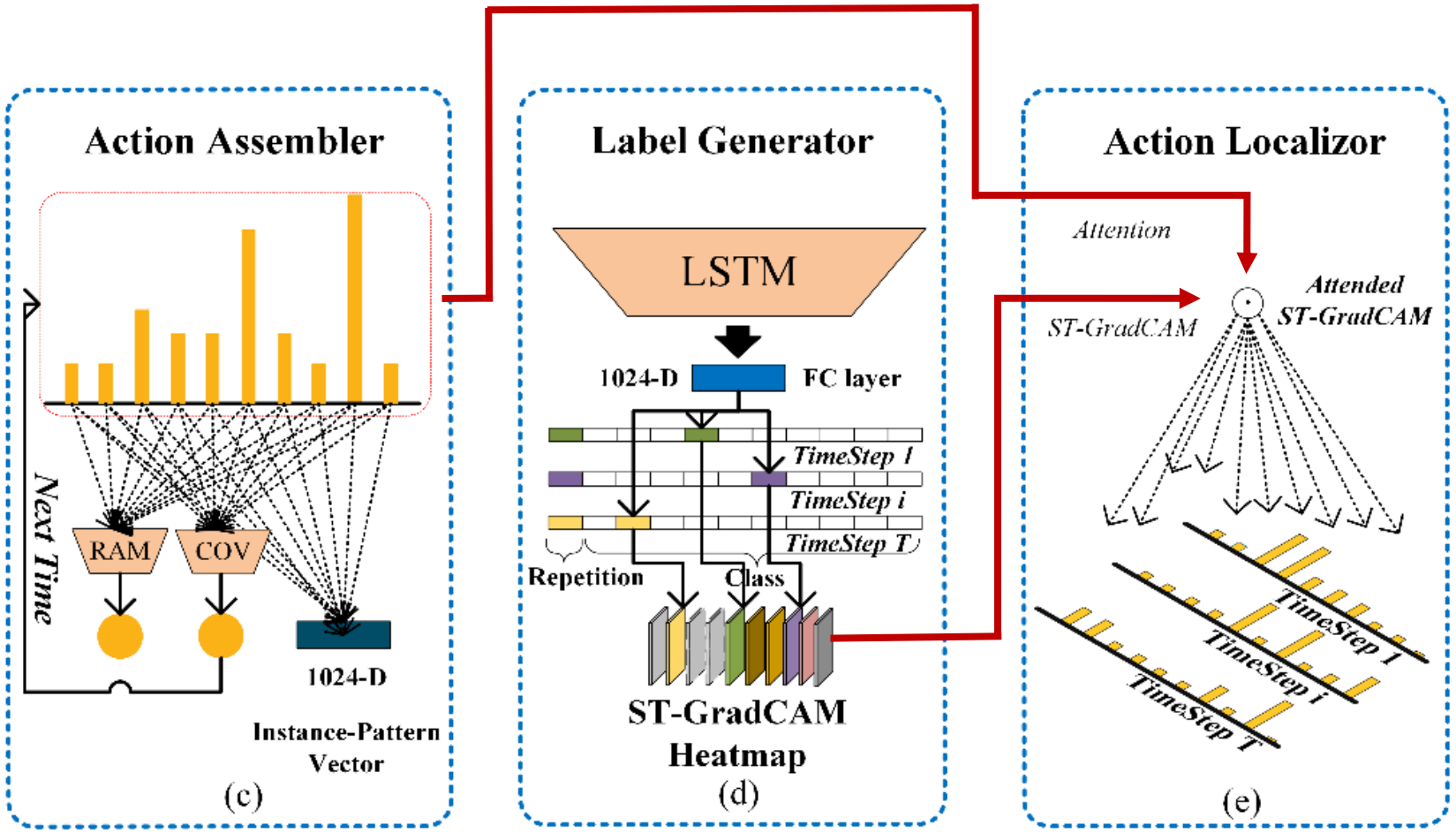


CAM ‘Cat’



$$d_t^c = \sum_{k=1}^{K} w_{t,k}^c x_t^k$$

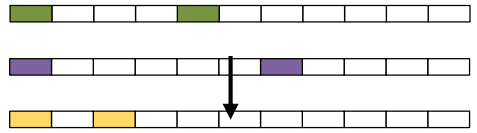
$$w_{t,k}^c = \frac{\partial d_t^c}{\partial x_t^k} = \frac{\partial d_t^c}{\partial h_t^c} \cdot \frac{\partial h_t^c}{\partial x_t^k} = W_o \cdot \frac{\partial RNN(x_t)}{\partial x_t^k}$$

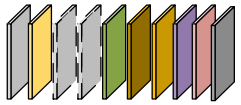


Attended ST-GradCAM:

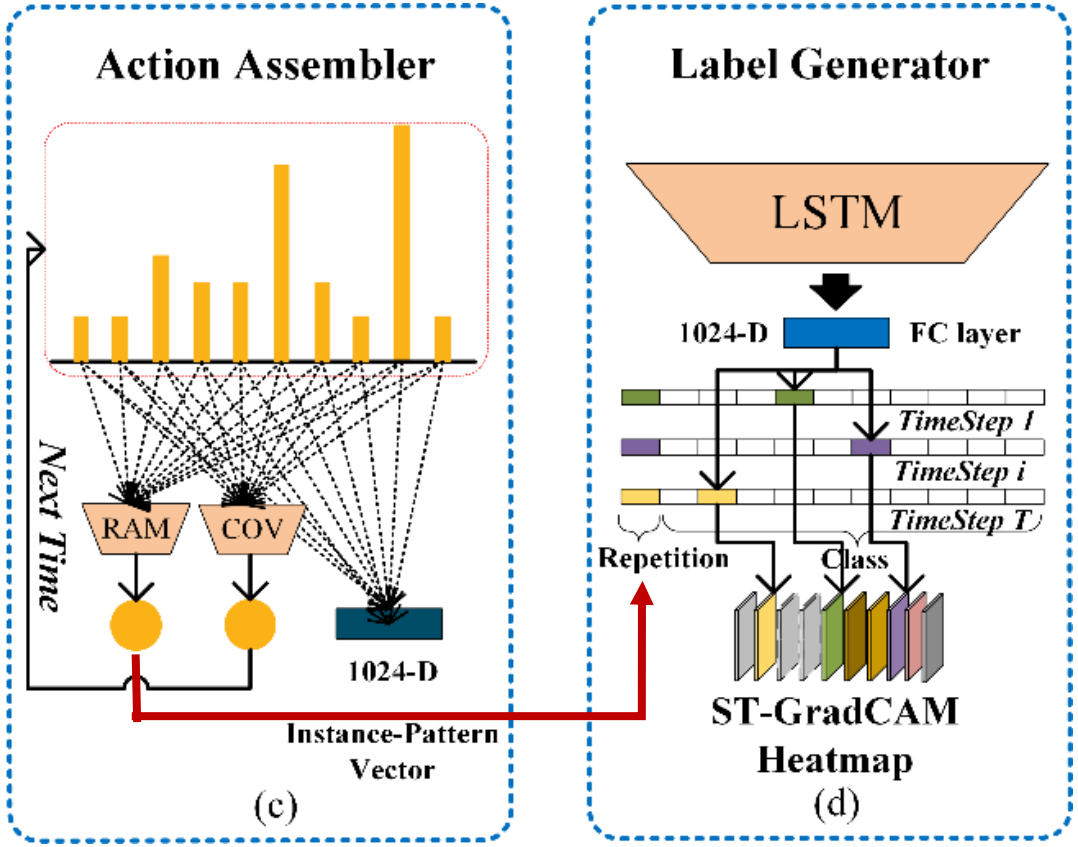
$$\xi_{t,i}^{c} = \sum_{k=1}^{K} w_{t,k}^{c} s_{i}^{k}$$

$$\alpha_{t,i} \cdot \sigma(\xi_{t,i}^{c})$$

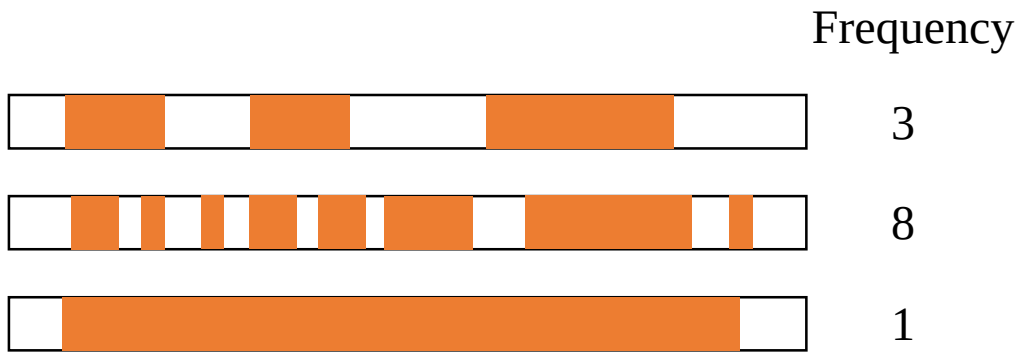




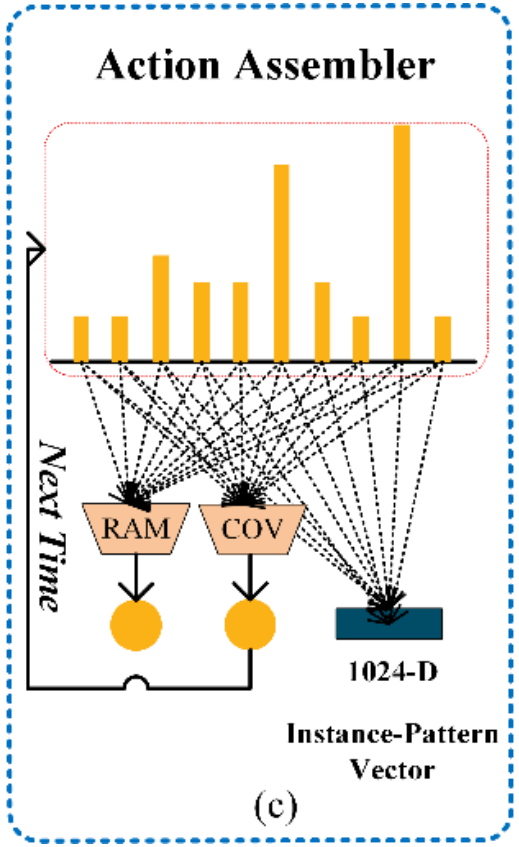
ST-GradCAM Heatmap



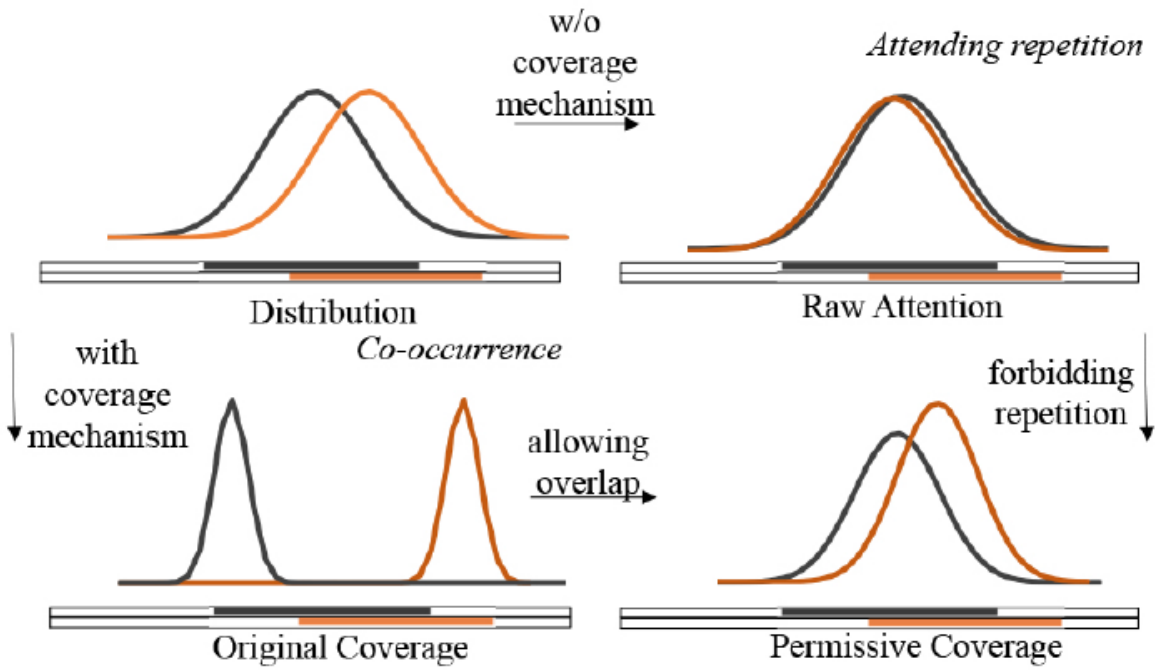
Repetition Alignment:



$$h_t = RNN(h_{t-1}, y_{t-1}, x_t, RAM_t)$$

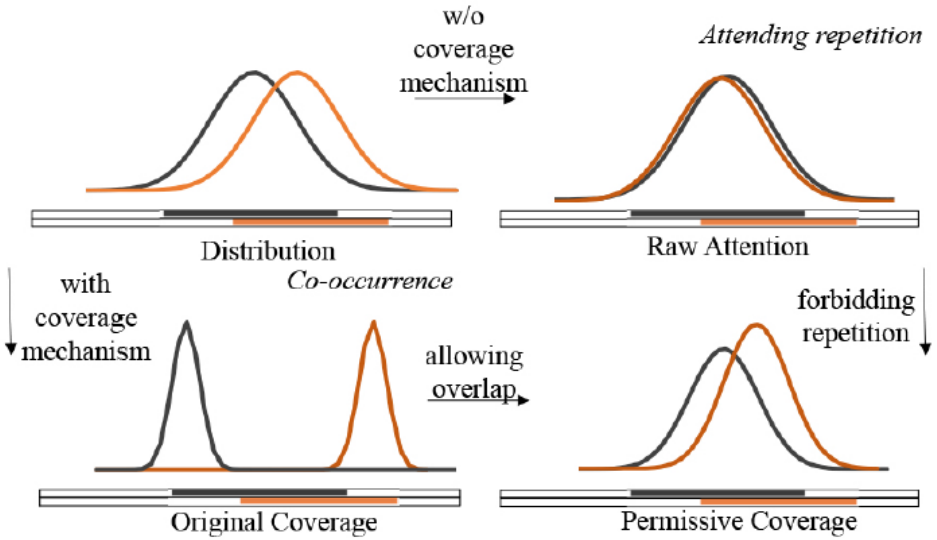


Permissive Coverage:

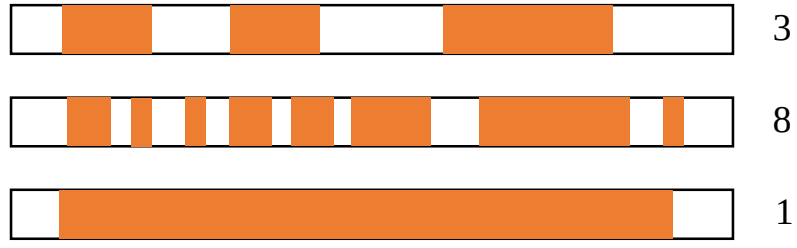


3. Experiment

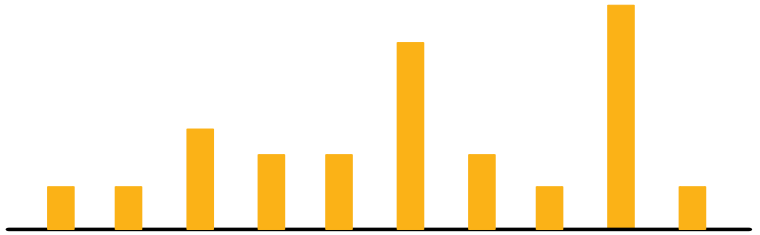
Sub-module Study:



COV



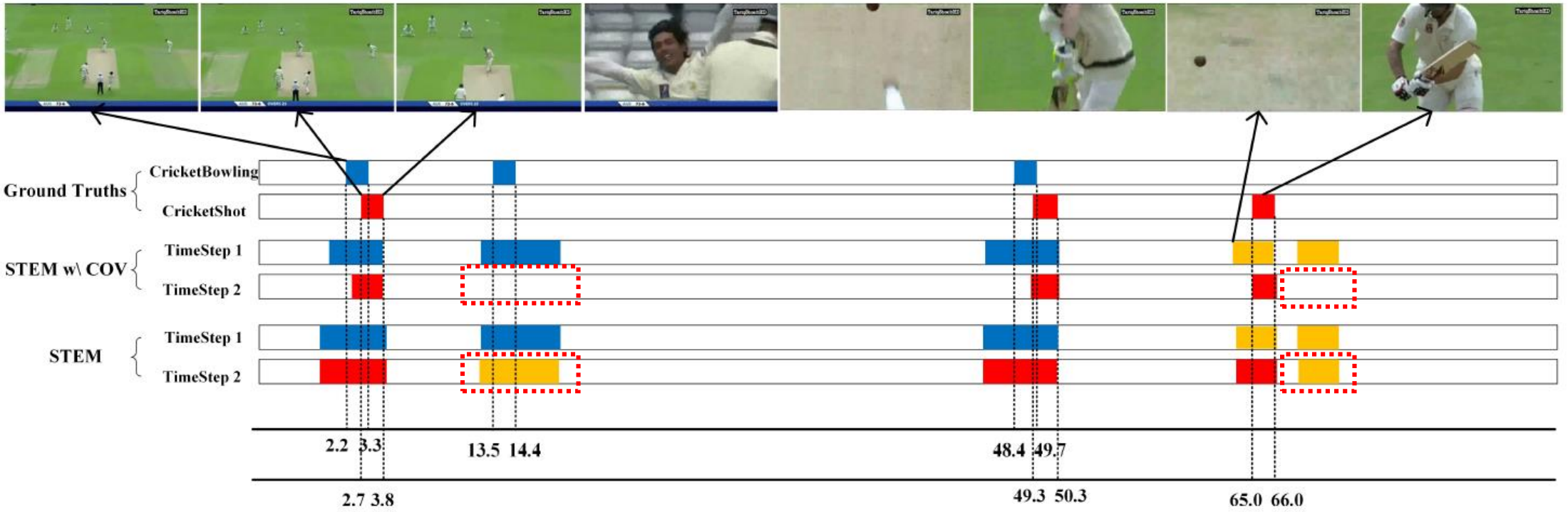
RAM



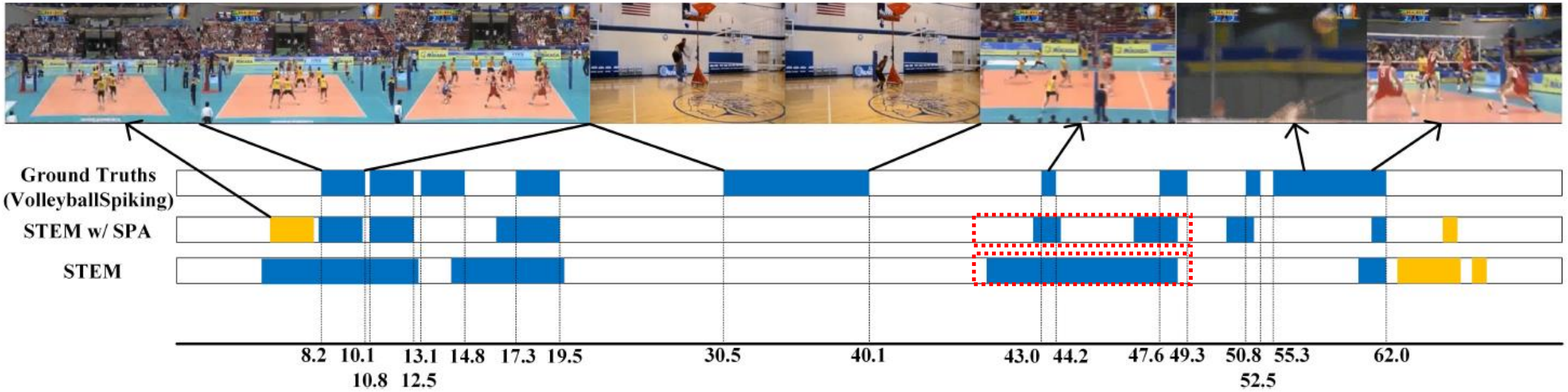
SPA

Table 1: Effects of sub-modules on THUMOS'14

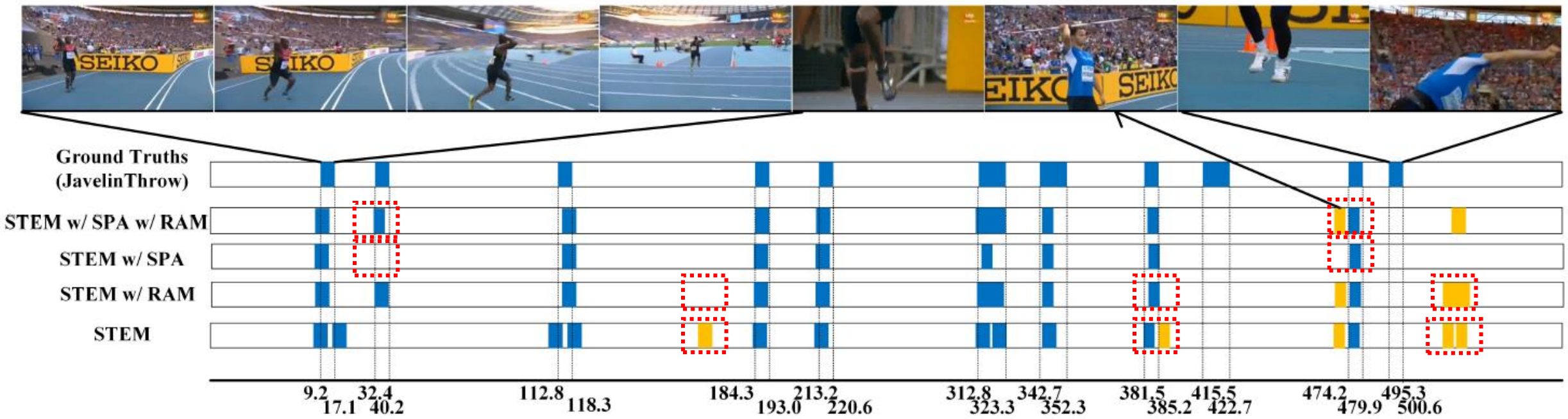
Stem	√	√	√	√	√	√	√	√
Sparsity		√			√	√		√
Coverage			√		√		√	√
RAM				√		√	√	√
Ave-mAP(%)	39.0	43.8	42.4	43.8	43.3	44.0	44.7	**47.0**



Effects of Coverage Sub-module

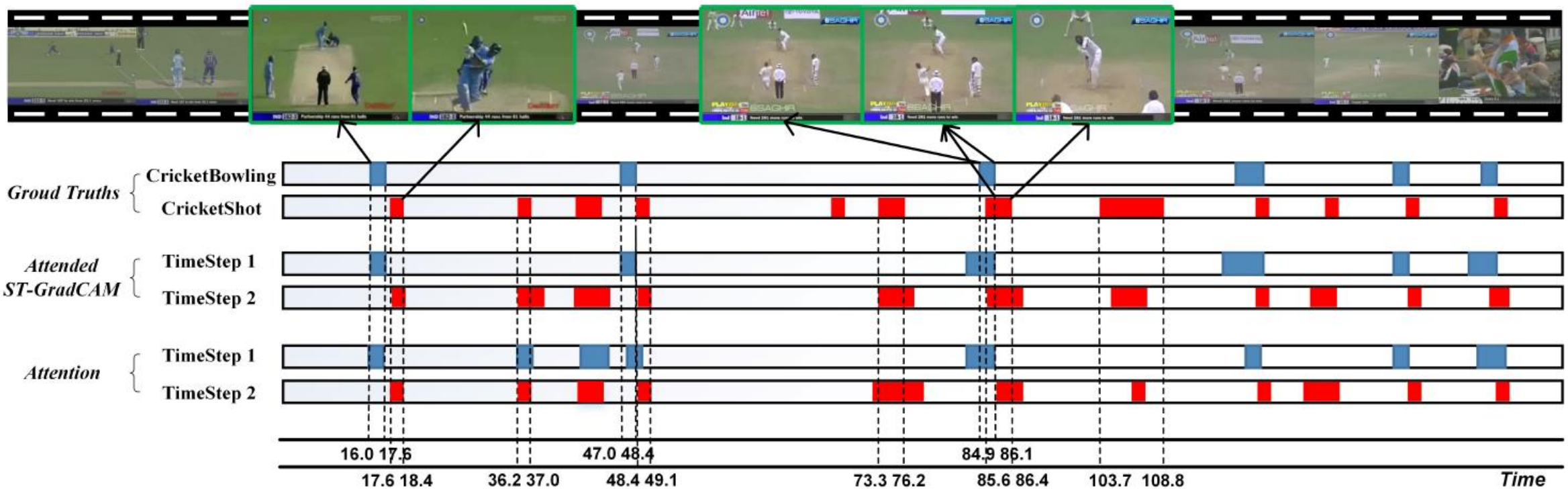


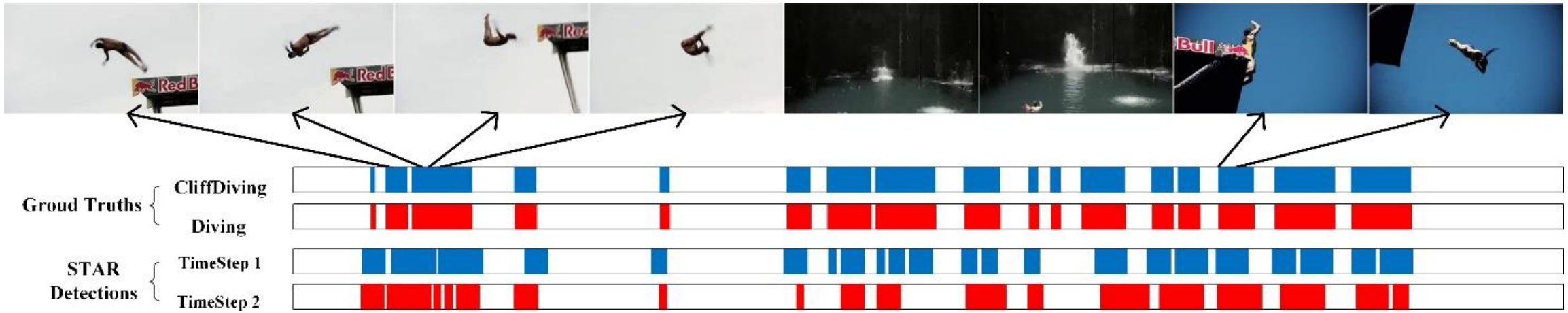
Effects of Sparsity Sub-module



Effects of RAM Sub-module

Overall Results:




Groud Truths
CliffDiving
Diving
STAR Detections
TimeStep 1
TimeStep 2

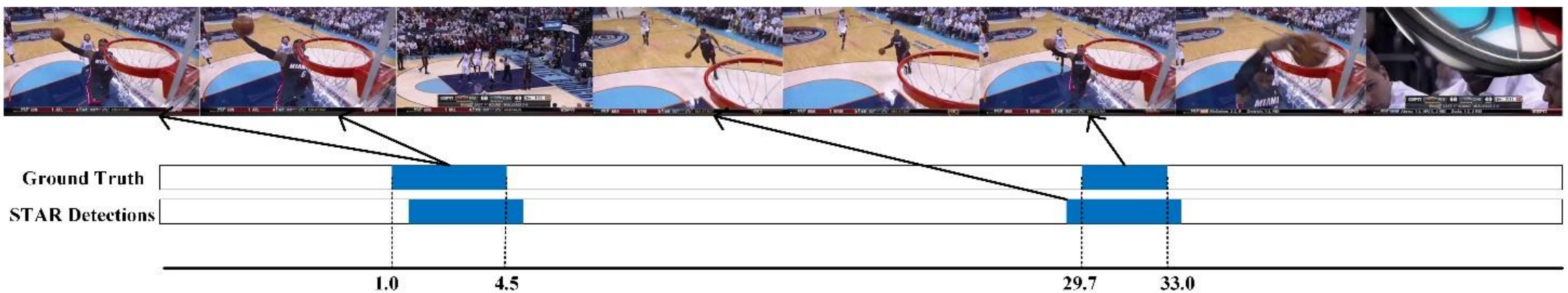

Ground Truth
STAR Detections
1.0
4.5
29.7
33.0

Table 2: Comparison with state-of-the-art on THUMOS'14

Supervision	Method	AP@IoU 0.1	0.2	0.3	0.4	0.5
Fully Supervised	Richard (Richard and Gall 2016)	39.7	35.7	30.0	23.2	15.2
	Shou (Shou, Wang, and Chang 2016)	47.7	43.5	36.3	28.7	19.0
	Yeung (Yeung et al. 2016)	48.9	44.0	36.0	26.4	17.1
	Yuan (Yuan et al. 2016)	51.4	42.6	33.6	26.1	18.8
	Shou (Shou et al. 2017)	–	–	40.1	29.4	23.3
	Yuan (Yuan et al. 2017)	51.0	45.2	36.5	27.8	17.8
	Gao (Gao, Yang, and Nevatia 2017)	54.0	50.9	44.1	34.9	25.6
	Xu (Xu, Das, and Saenko 2017)	54.5	51.5	44.8	35.6	28.9
	Zhao (Zhao et al. 2017)	**66.0**	**59.4**	51.9	41.0	29.8
	Yang (Yang et al. 2018)	–	–	44.1	37.1	28.2
	Chao (Chao et al. 2018)	59.8	57.1	53.2	**48.5**	**42.8**
	Alwasssel (Alwassel et al. 2018)	49.6	44.3	38.1	28.4	19.8
	Lin (Lin et al. 2018)	–	–	**53.5**	45.0	36.9
Weakly Supervised	Wang (Wang et al. 2017)	44.4	37.7	28.2	21.1	13.7
	Singh (Singh and Yong 2017)	36.4	27.8	19.5	12.7	6.8
	Nguyen (Nguyen et al. 2018)	52.0	44.7	35.5	25.8	16.9
	Shou (Shou et al. 2018)	–	–	35.8	29.0	21.2
	Paul (Paul et al. 2018)	55.2	49.6	40.1	31.1	22.8
	Ours	**68.8**	**60.0**	**48.7**	**34.7**	**23.0**

4. Q & A

THANKS !