NEURAL INFORMATION PROCESSING SYSTEMS

PointTAD: Multi-Label Temporal Action Detection with Learnable Query Points

Jing Tan[1*], Xiaotong Zhao[2], Xintian Shi[2], Bin Kang[2], Limin Wang[1,3†]

[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]Platform and Content Group (PCG), Tencent [3]Shanghai AI Lab

南京大學 NANJING UNIVERSITY

MCG MULTIMEDIA COMPUTING GROUP 媒体计算研究组

PCG 平台与内容事业群 PLATFORM AND CONTENT GROUP

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

Problem
Temporal Action Detection (TAD)

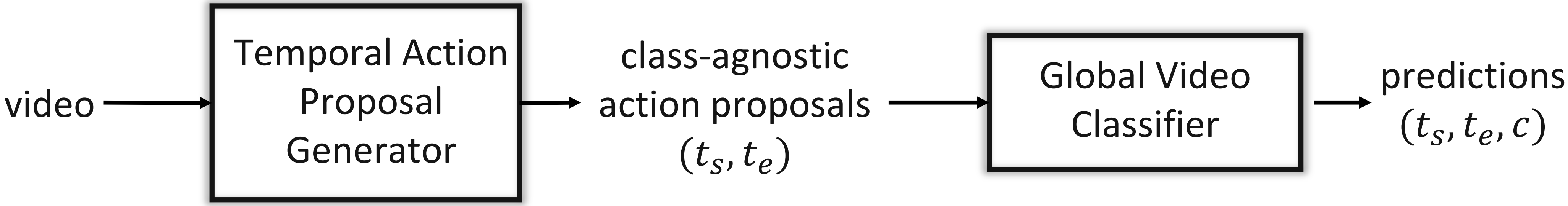
- Detecting the temporal span and class label of actions in untrimmed videos.

Problem
Temporal Action Detection (TAD)

- Non-overlapping instances
- Single-label annotations

Problem
Temporal Action Detection (TAD)

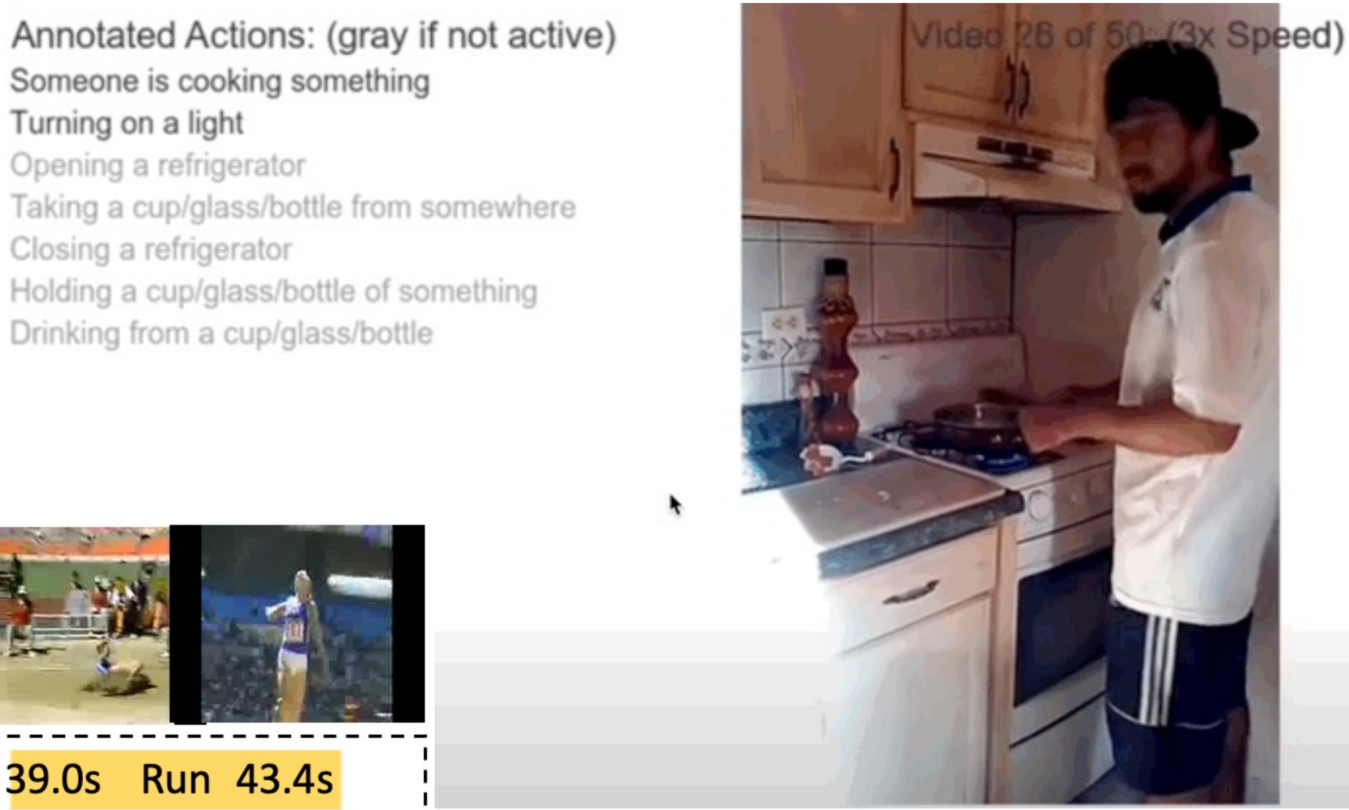
Mainstream approaches:



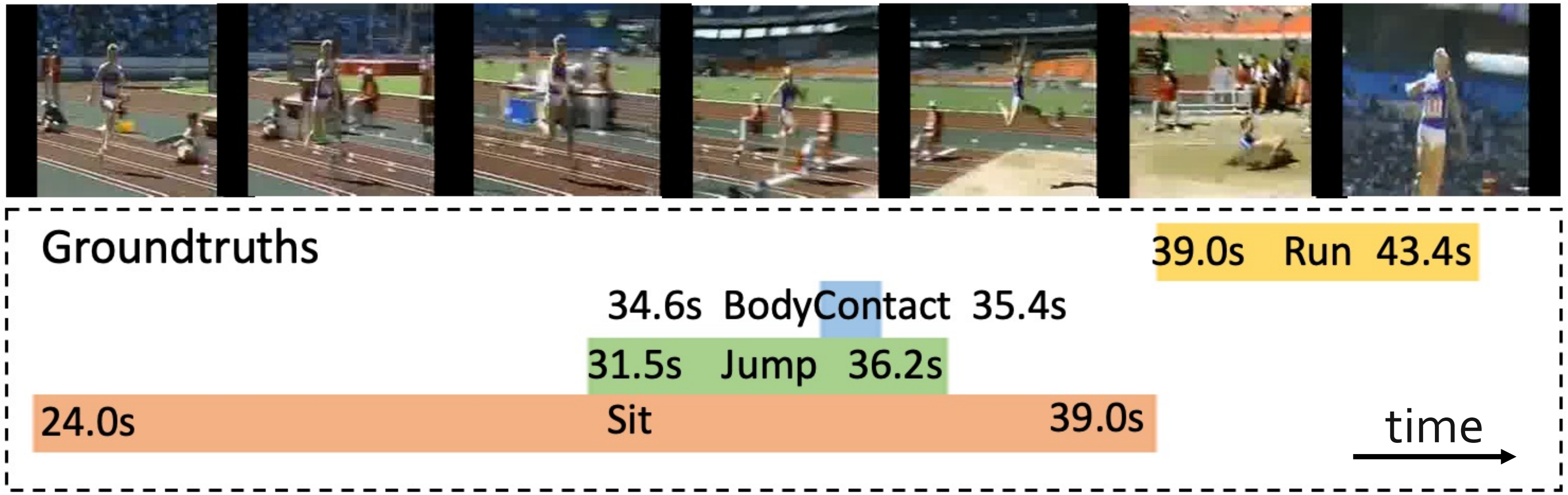
Real-world scenario: different classes of actions often co-occur in videos!

Problem
Multi-label Temporal Action Detection

- A more challenging TAD setup
  - Concurrent instances
  - Complex action relations

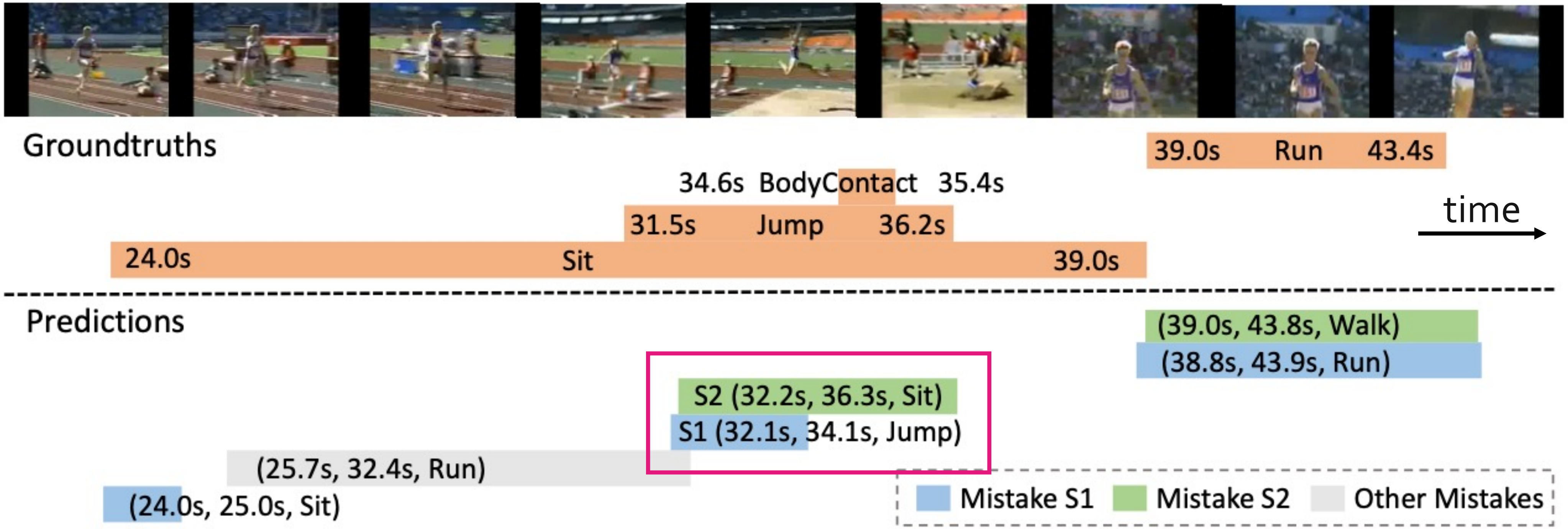


source: https://prior.allenai.org/projects/charades



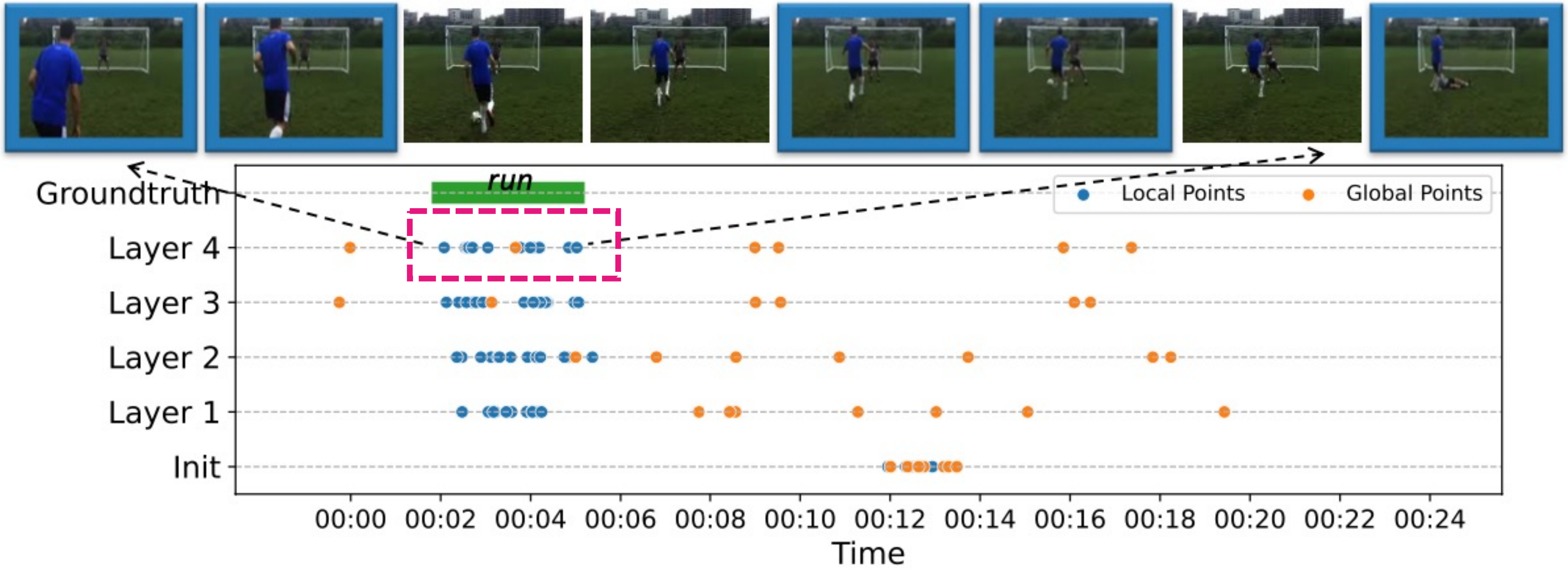
Related work
Segment-based TAD methods

- **Segment-based TAD methods** either capture incomplete action segments or get misclassified over good localization.



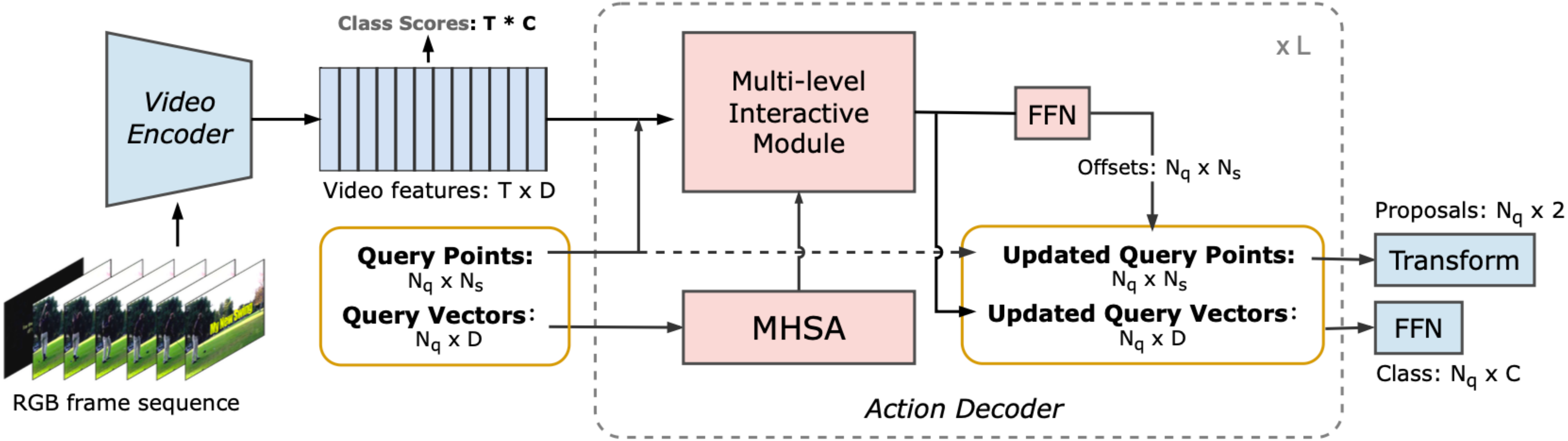
Our approach

- We introduce **Learnable Points** to handle both boundary frames and semantic keyframes of actions.



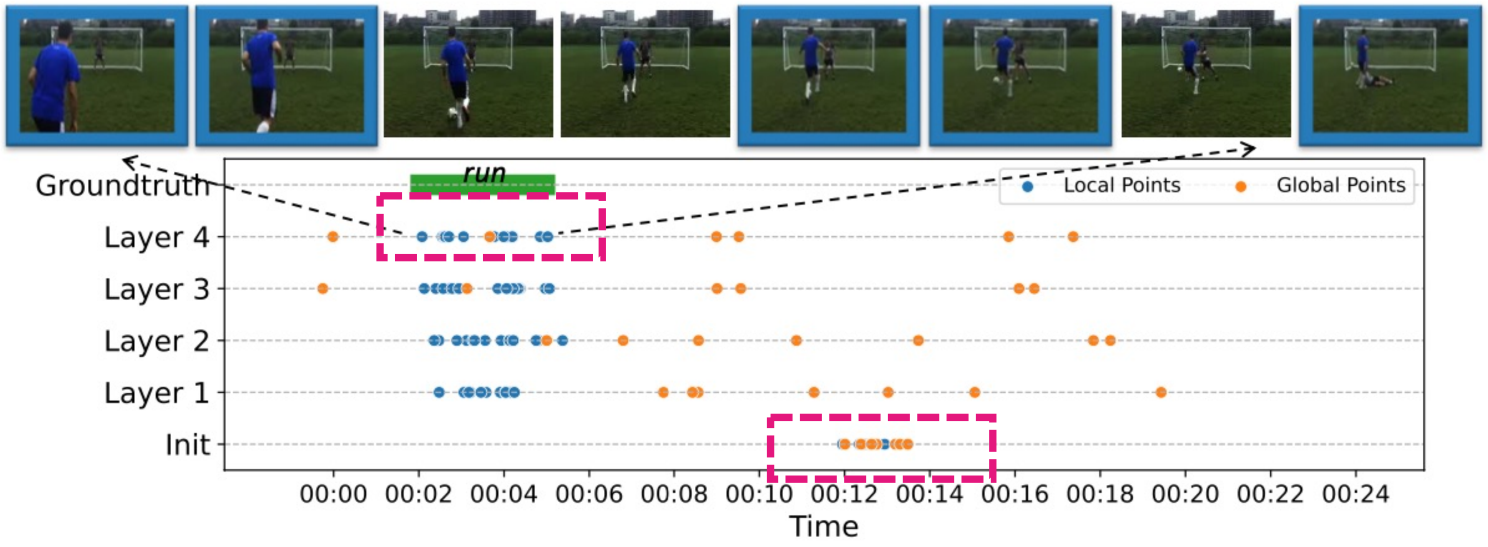
Method Overview

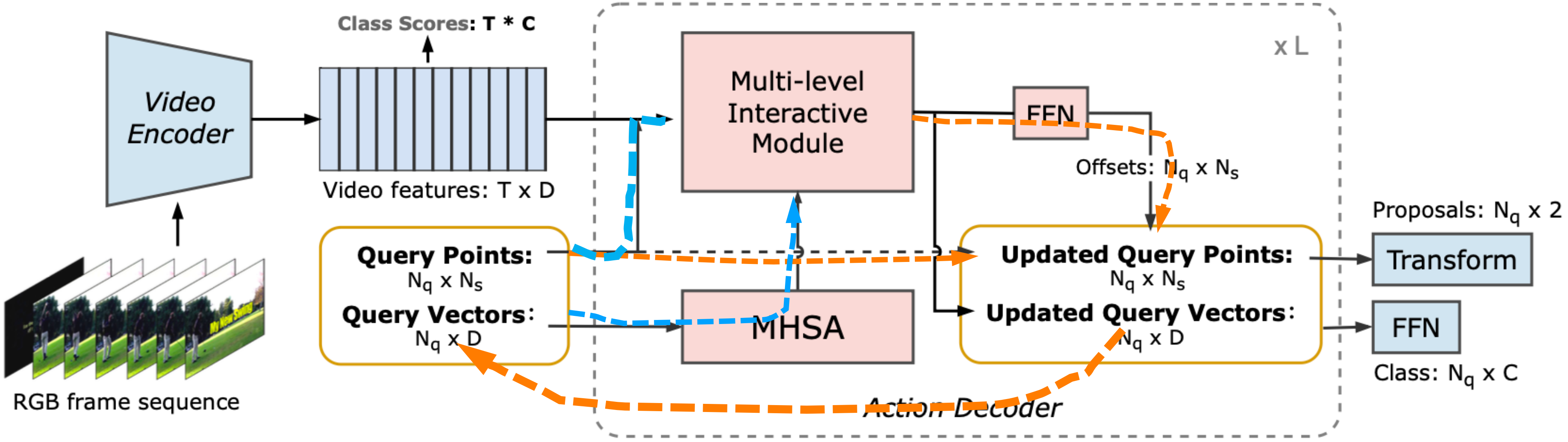
NEURAL INFORMATION PROCESSING SYSTEMS



Method
Learnable Query Points

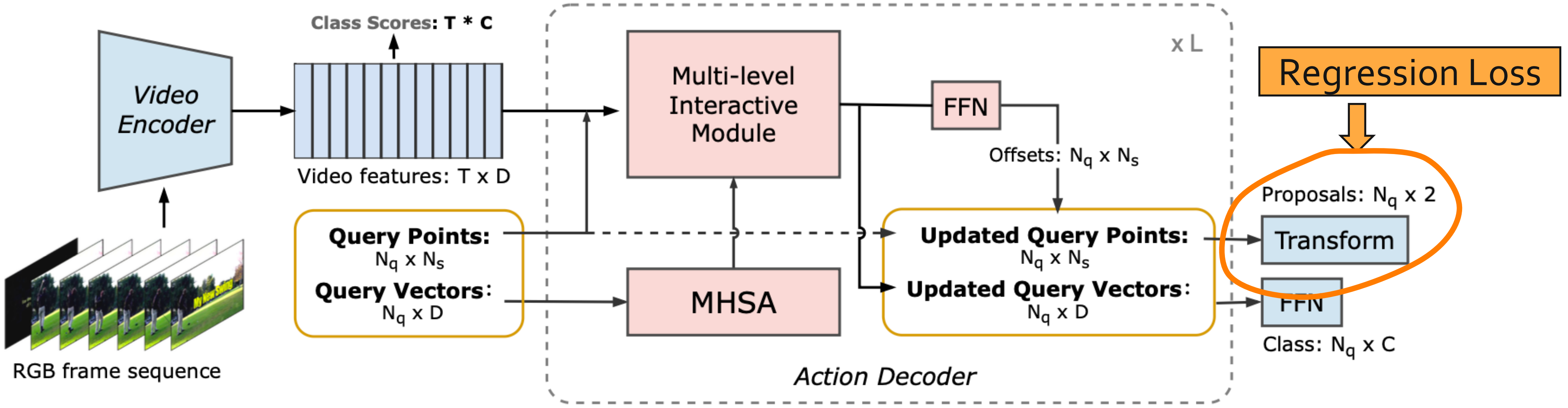
- Iterative Point Refinement.





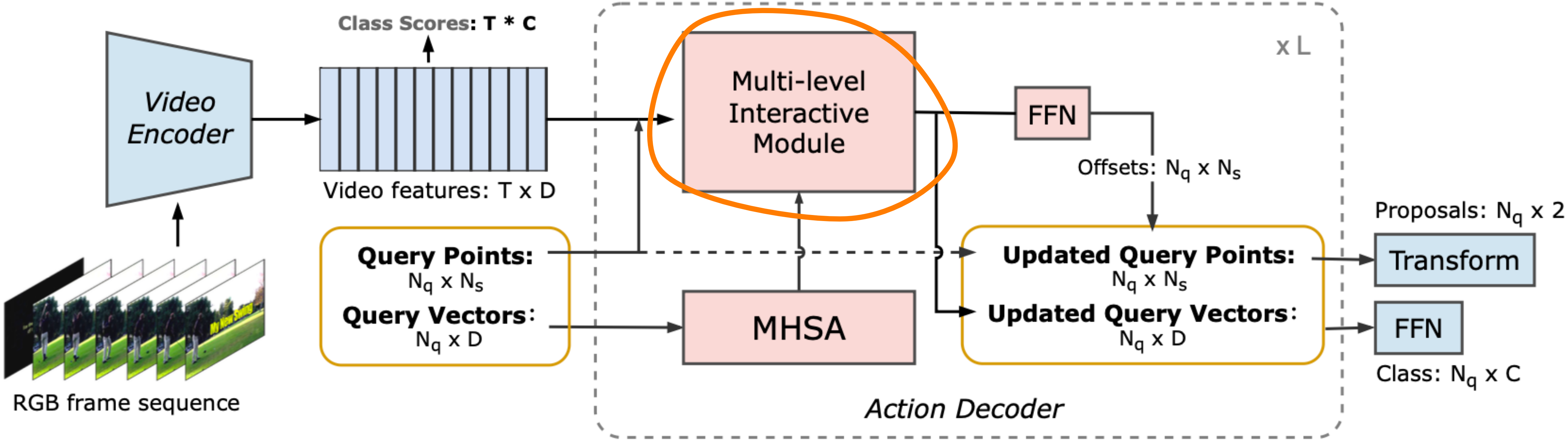
Method
Learnable Query Points

- Learning Query Points with pseudo segments
  - Point to segment transformation $\mathcal{T}: \mathcal{P} = \{t_j\}_{j=1}^{N_s} \rightarrow \mathcal{S} = (t^s, t^e)$



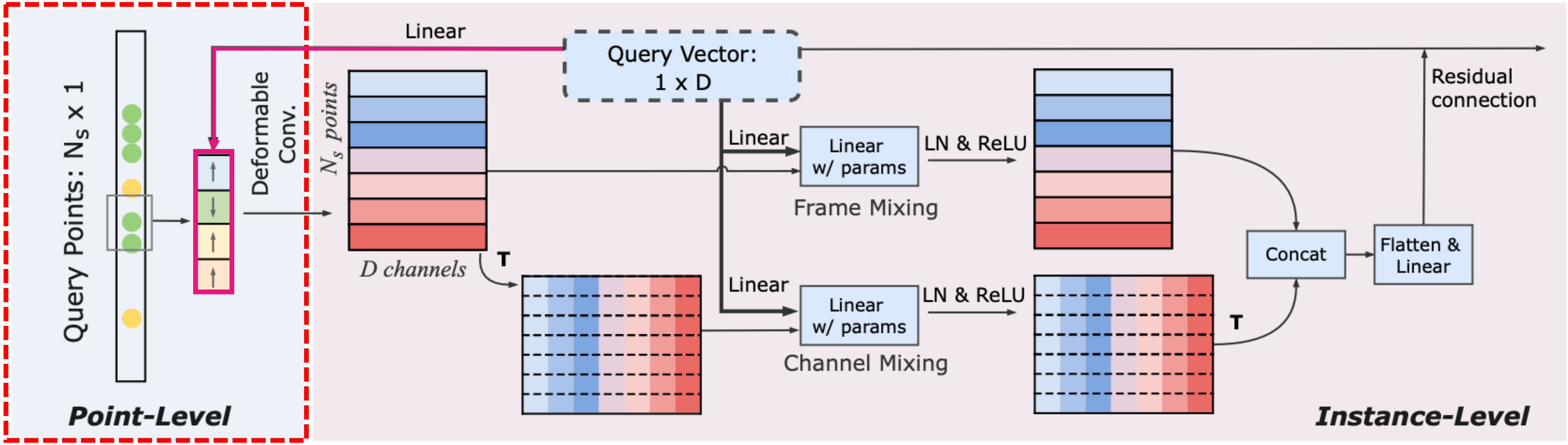
Method
Multi-level Interactive Module

- Point-level local deformation
- Instance-level adaptive mixing



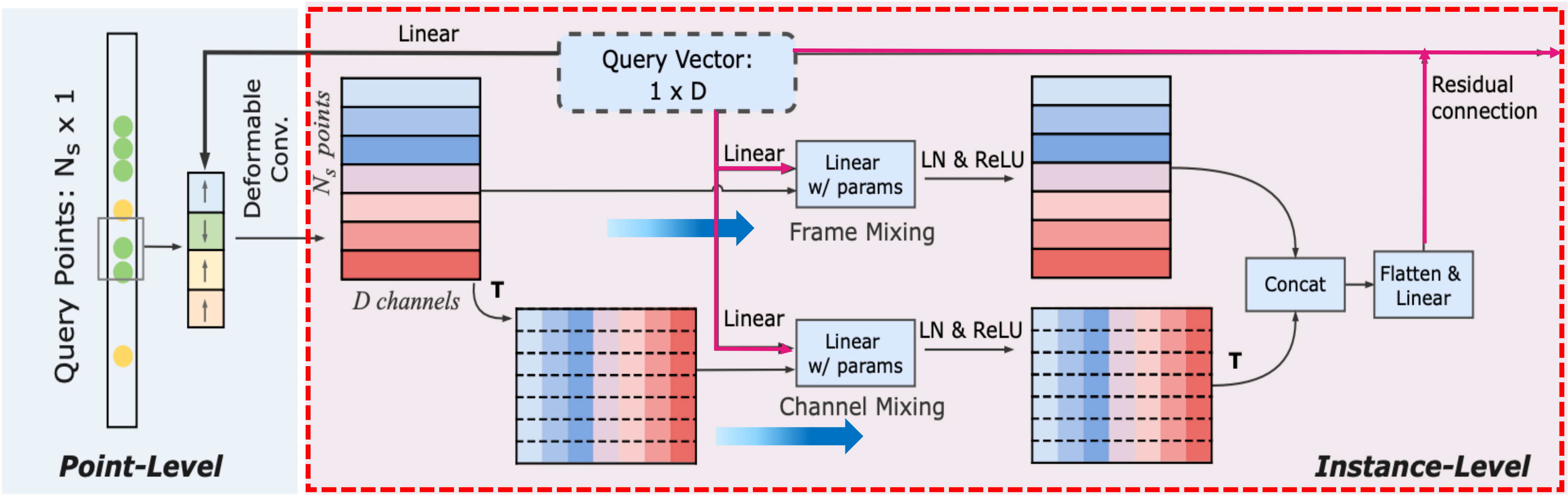
Method
Multi-level Interactive Module

- **Point-level local deformation**: capture local temporal cue.



Method
Multi-level Interactive Module

- **Instance-level adaptive mixing**: explore frame relations and channel dynamics
  - **Channel-wise** and **frame-wise** dynamic convolution in **parallel**



Results

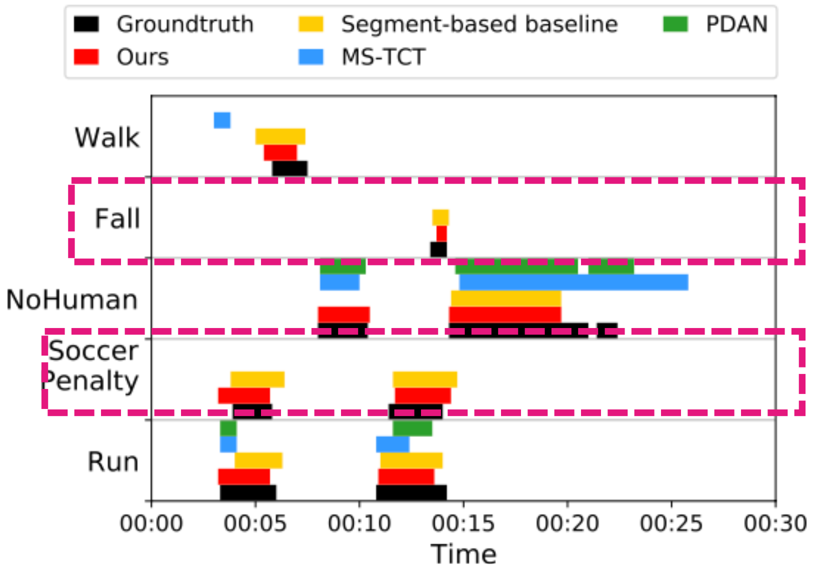
- Quantitative comparison
  - Introduced detection-mAP from classic TAD for instance-wise detection evaluation.

Table 1: **Comparison with the state of the art** on the MultiTHUMOS test set and Charades test set, under detection-mAP (%) and segmentation-mAP(%).

Methods	Modality	MultiTHUMOS		Charades	
		Det-mAP	Seg-mAP	Det-mAP	Seg-mAP
R-C3D [44]	RGB	-	-	-	17.6
Super-event [29]	RGB	-	36.4	-	18.6
TGM [30]	RGB	-	37.2	-	20.6
TGM [30]	RGB+OF	-	44.3	-	21.5
PDAN [9]	RGB	17.3/17.1*	40.2	8.5	23.7
Coarse-Fine [16]	RGB	-	-	6.1	25.1
MLAD [40]	RGB	14.2/13.9*	42.2	-	18.4
MLAD [40]	RGB+OF	-	51.5	-	23.7
CTRN [7]	RGB	-	**44.0**	-	25.3
CTRN [7]	RGB+OF	-	51.2	-	27.8
AGT [27]	RGB+OF	-	-	-	28.6
MS-TCT [8]	RGB	16.2/16.0*	43.1	7.9	**25.4**
Ours	RGB	**21.5/21.4***	39.8	**11.1**	21.0

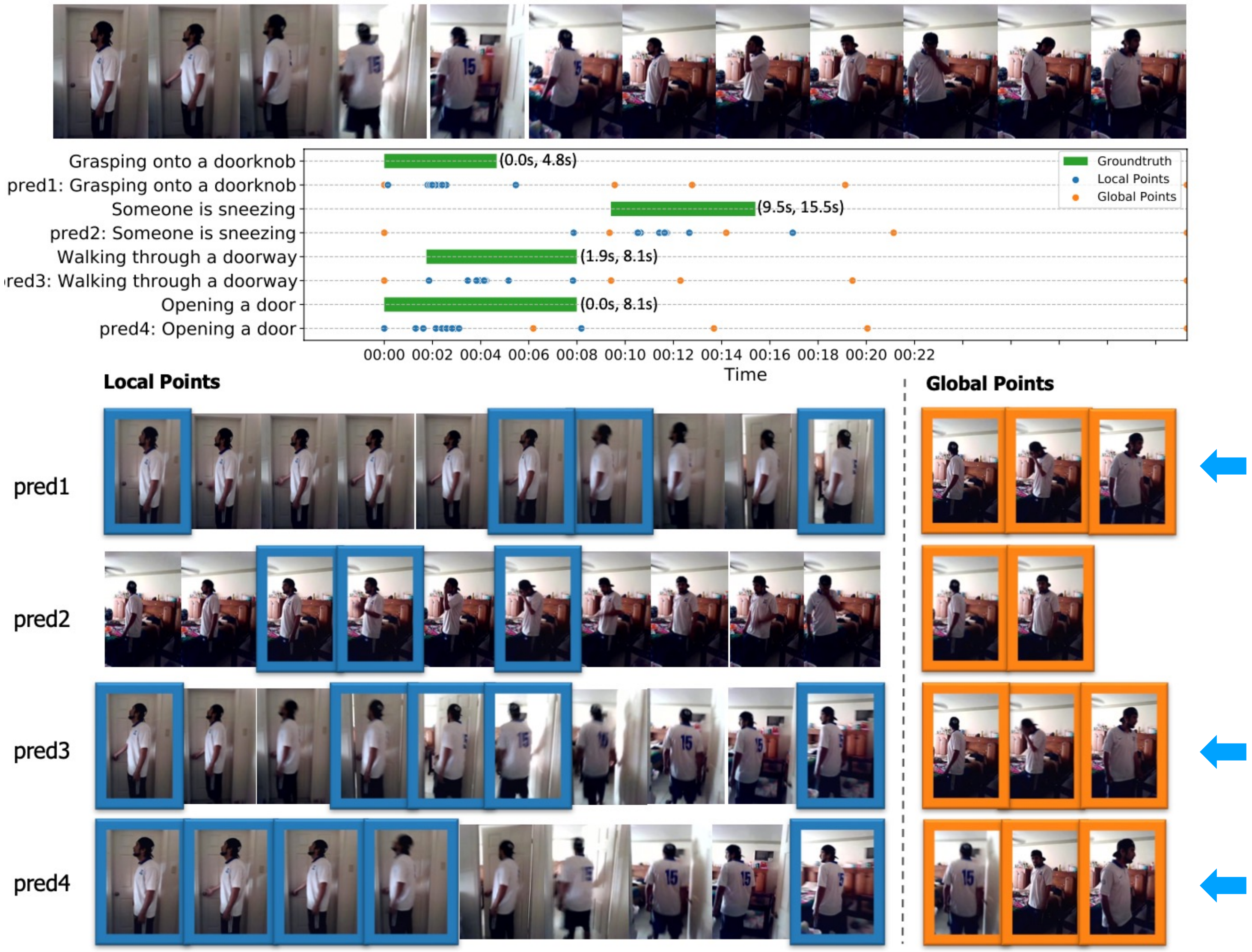
* *indicates detection results excluding NoHuman class.*

- Qualitative comparison



- More precise temporal boundaries
- More instances detected at harder categories

Visualization



- For highly overlapping groundtruths, local query points capture different frames for different actions.

NEURAL INFORMATION PROCESSING SYSTEMS

Thanks for your attention!

Code is available at https://github.com/MCG-NJU/PointTAD

南京大學 NANJING UNIVERSITY

MCG MULTIMEDIA COMPUTING GROUP 媒体计算研究组

PCG 平台与内容事业群 PLATFORM AND CONTENT GROUP

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory