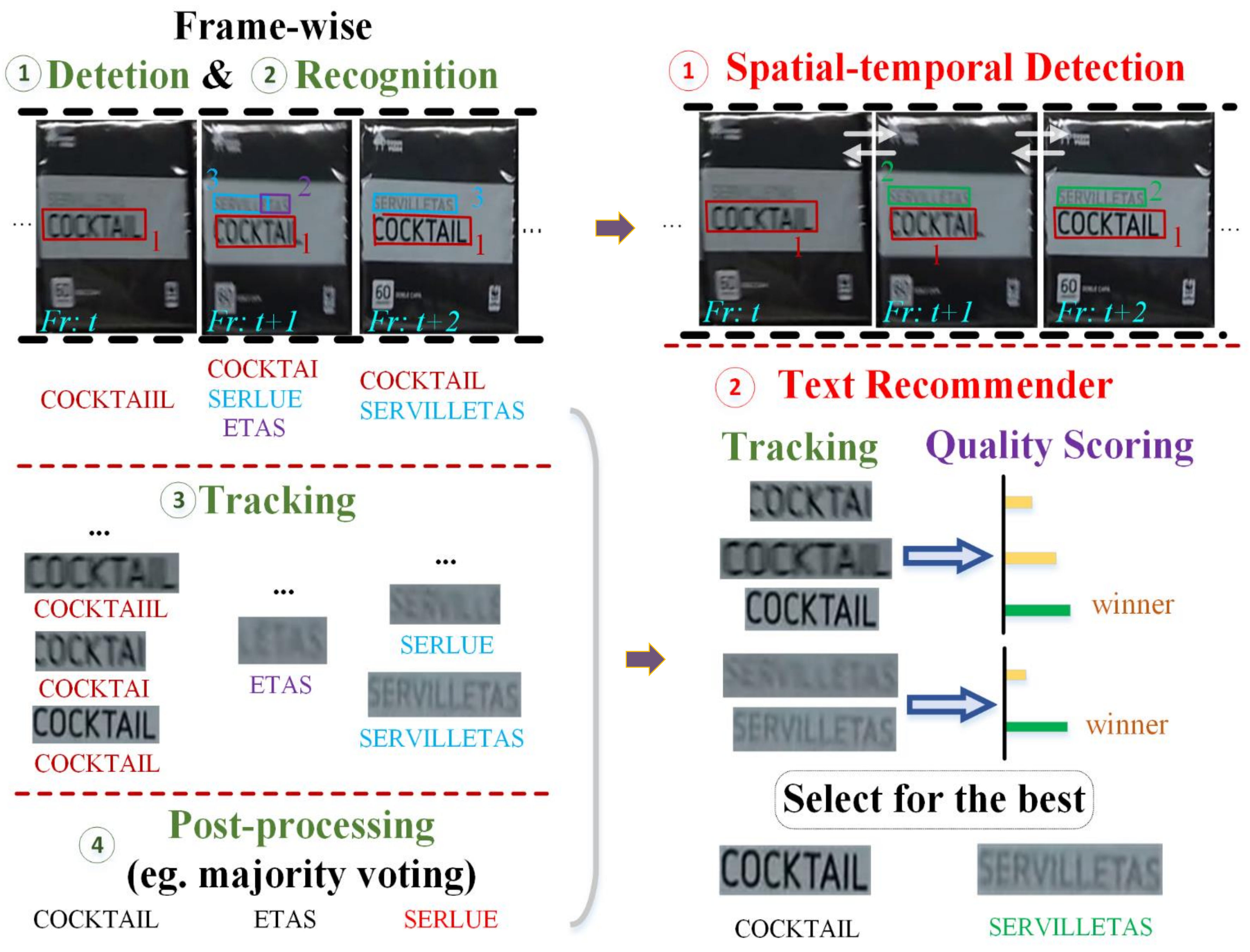
21 - 25 October 2019, Nice, France

You Only Recognize Once: Towards Fast Video Text Spoting

Zhanzhan Cheng[1,2]; Jing Lu; Yi Niu; Shiliang Pu   Fei Wu   Shuigeng Zhou
Hikvision Research Institute[1]   Zhejiang University[2]   Fudan University

➢ Motivation of This Work

- ***existing multi-stage pipeline**: localize and recognize in each frames, track for text streams, then post-precess. Two problems:*
  - ***excessive computation cost from repetitive recognition***
  - ***unstable recognition results due to low-quality text***



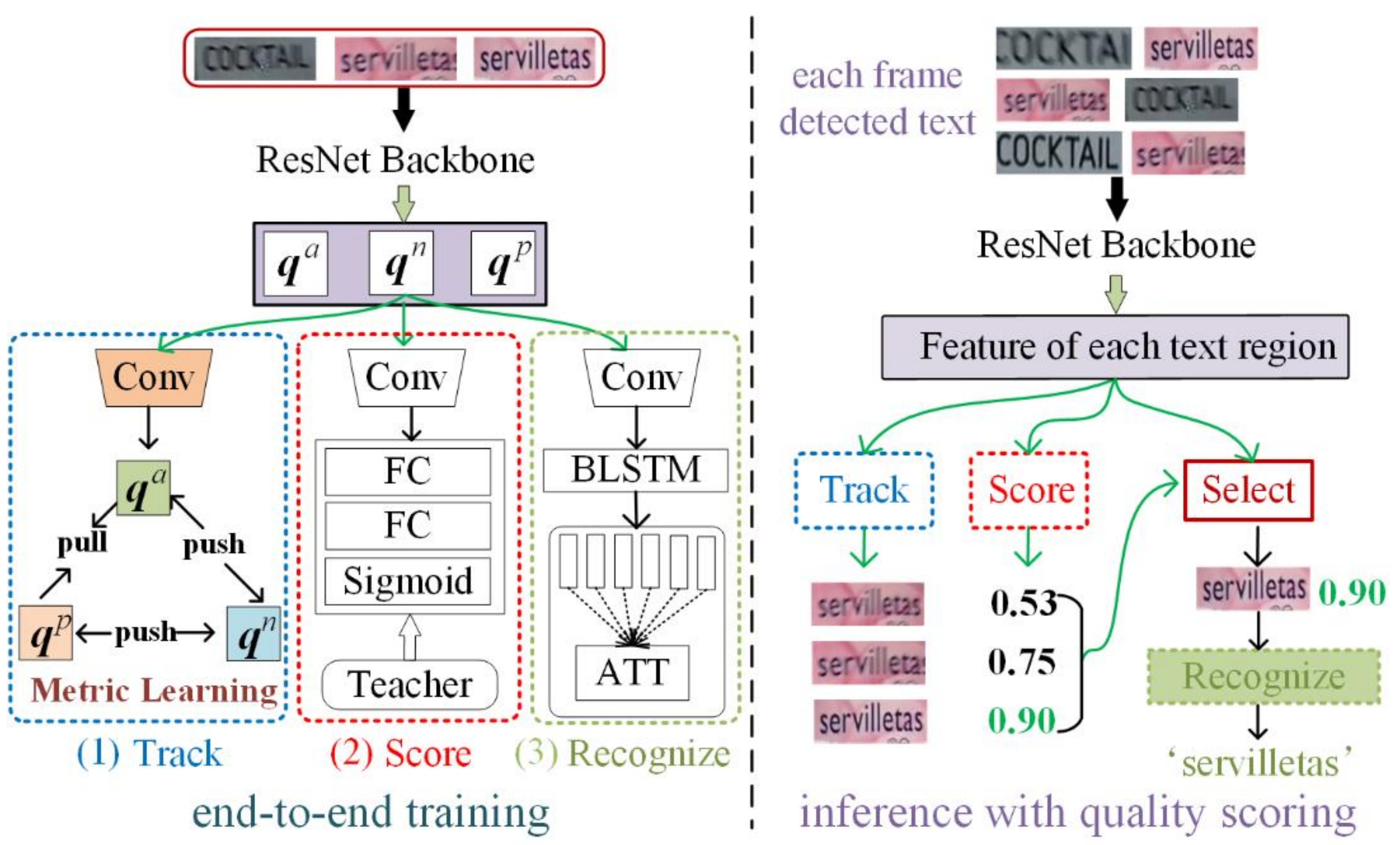
✓ existing framwork   ✓ Our framwork

- ***key idea**: select the highest quality text region from each text stream to be recognized once instead of one-by-one, which:*
  - ***speeds up text spotting by avoiding repetitive recognition***
  - ***leads to more robust recognition results by filtering out low-quality text***

➢ Main Contributions

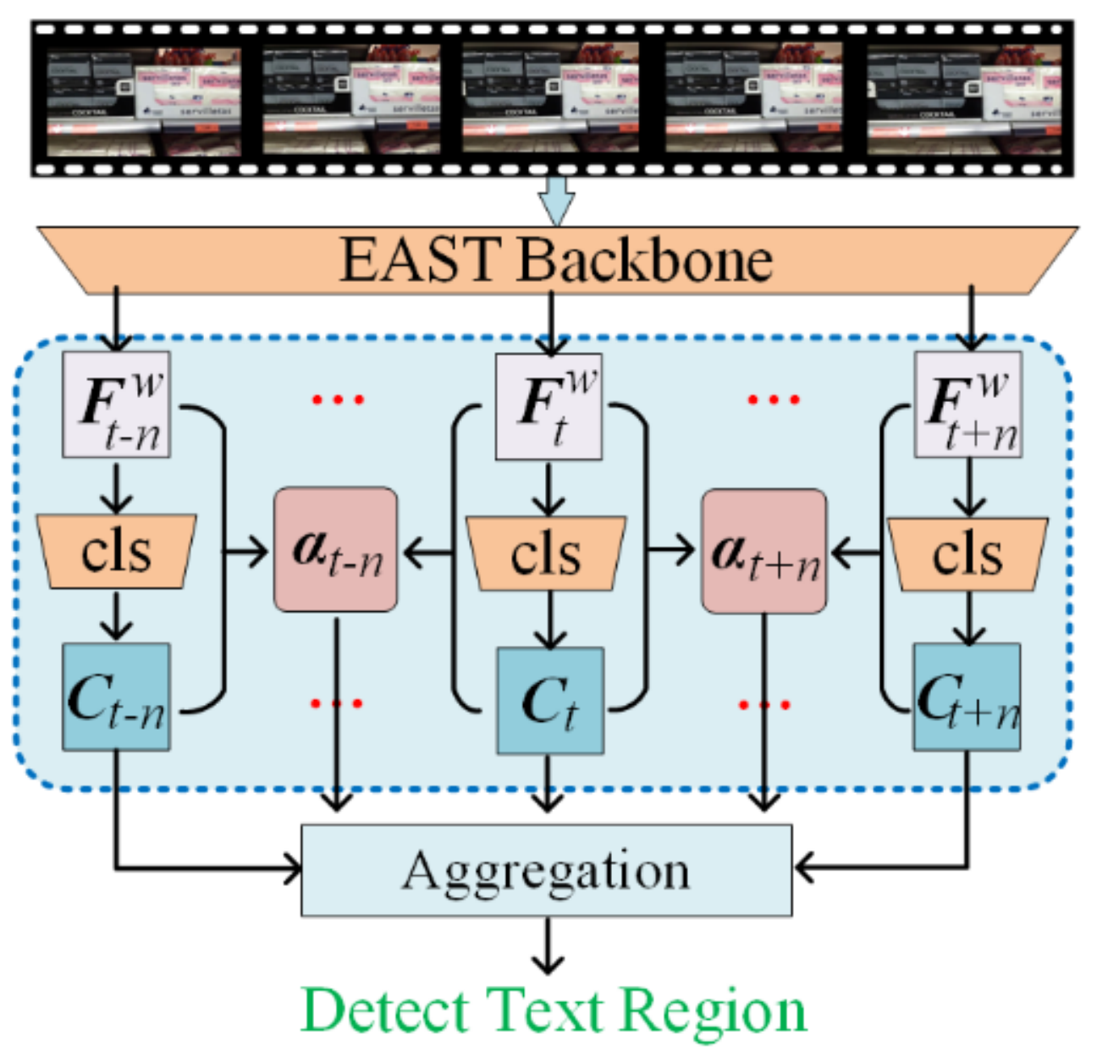
- *an unifed **two-stage** framework YORO consisting of a spatial-temporal detector and a text recommender for fast video text spoting.*
- *a novel text recommender for selecting the highest-quality text from text streams, then only recognizing the selected text regions once.*

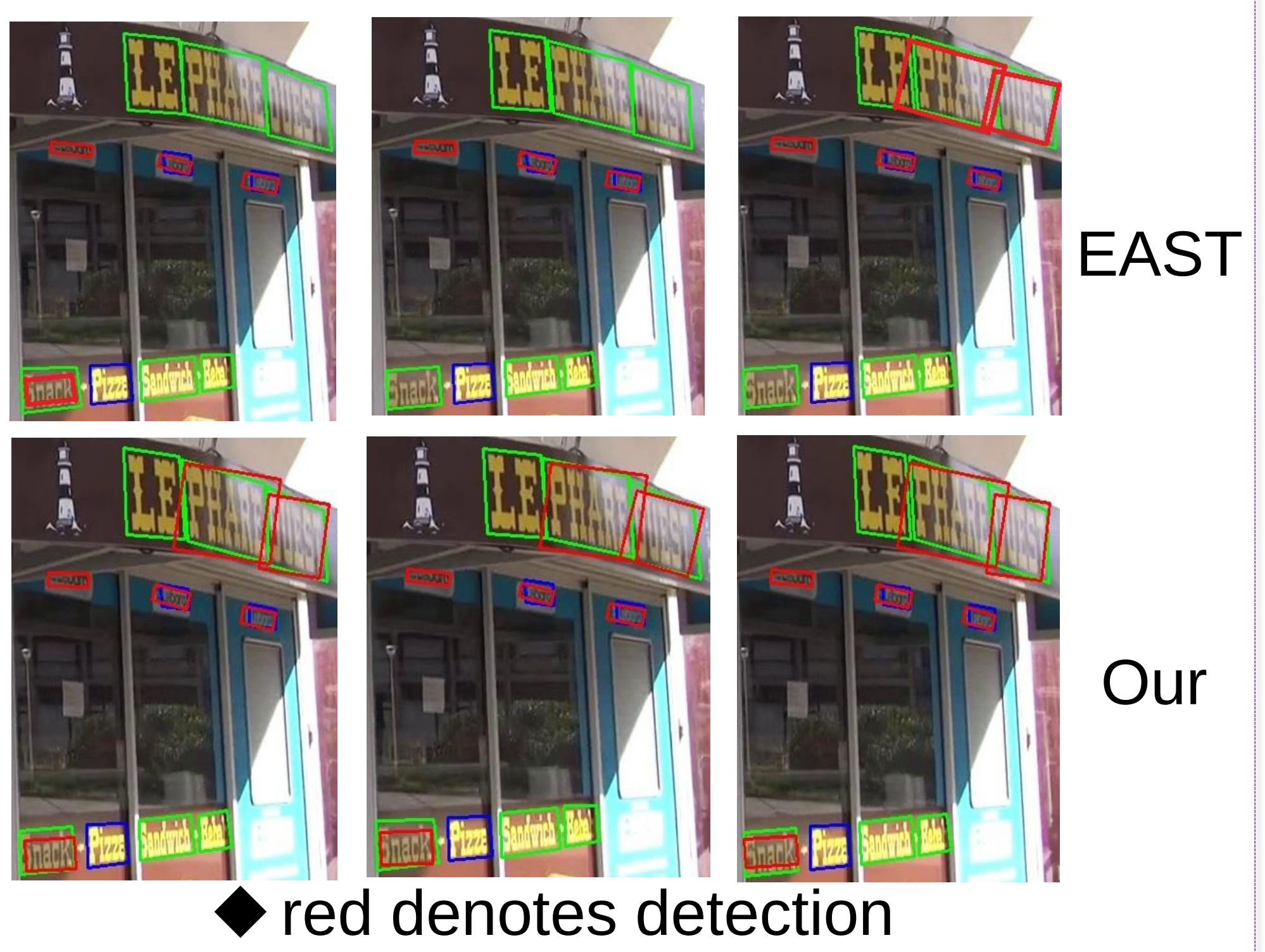
Text Recommender



- *a spatial-temporal detector for robustly recall more text by referring to temporal relationship among different frames.*

✓ self-attention based aggregation



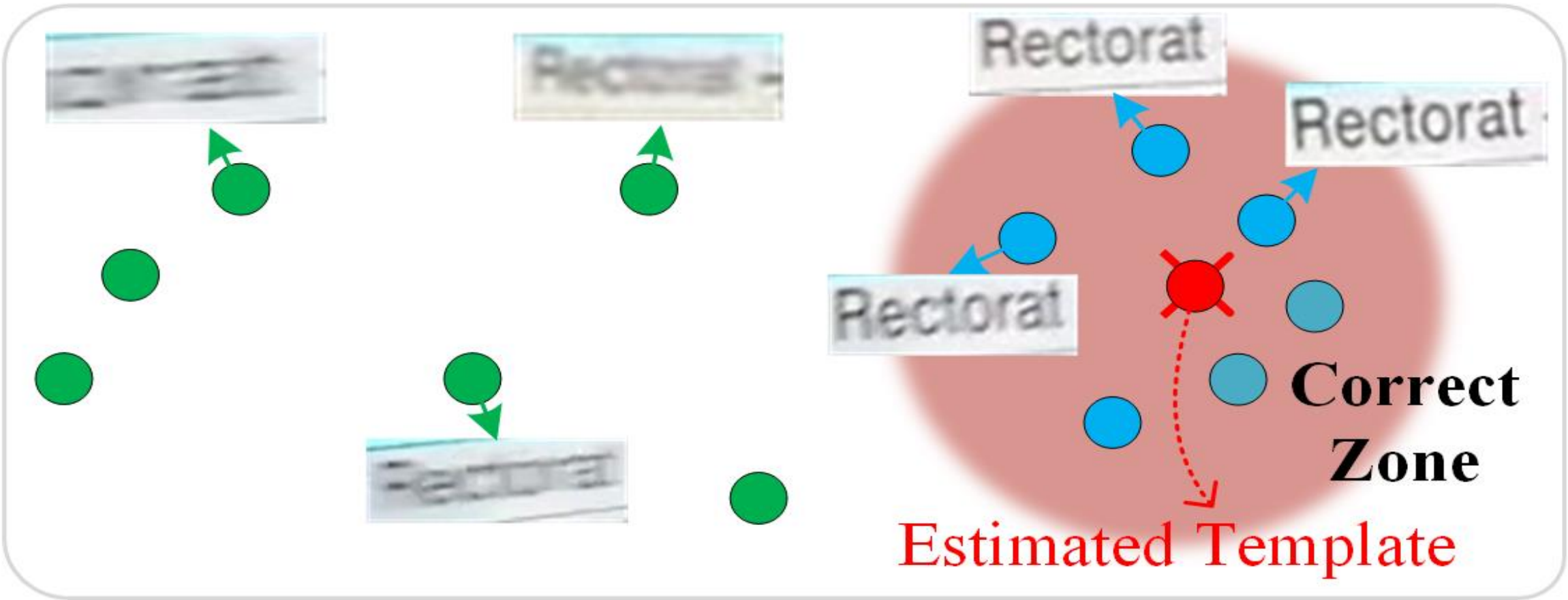


◆ red denotes detection

◆ **Program ID: P2C-11**   **Email:lujing6@hikvison.com**

➢ Key Component

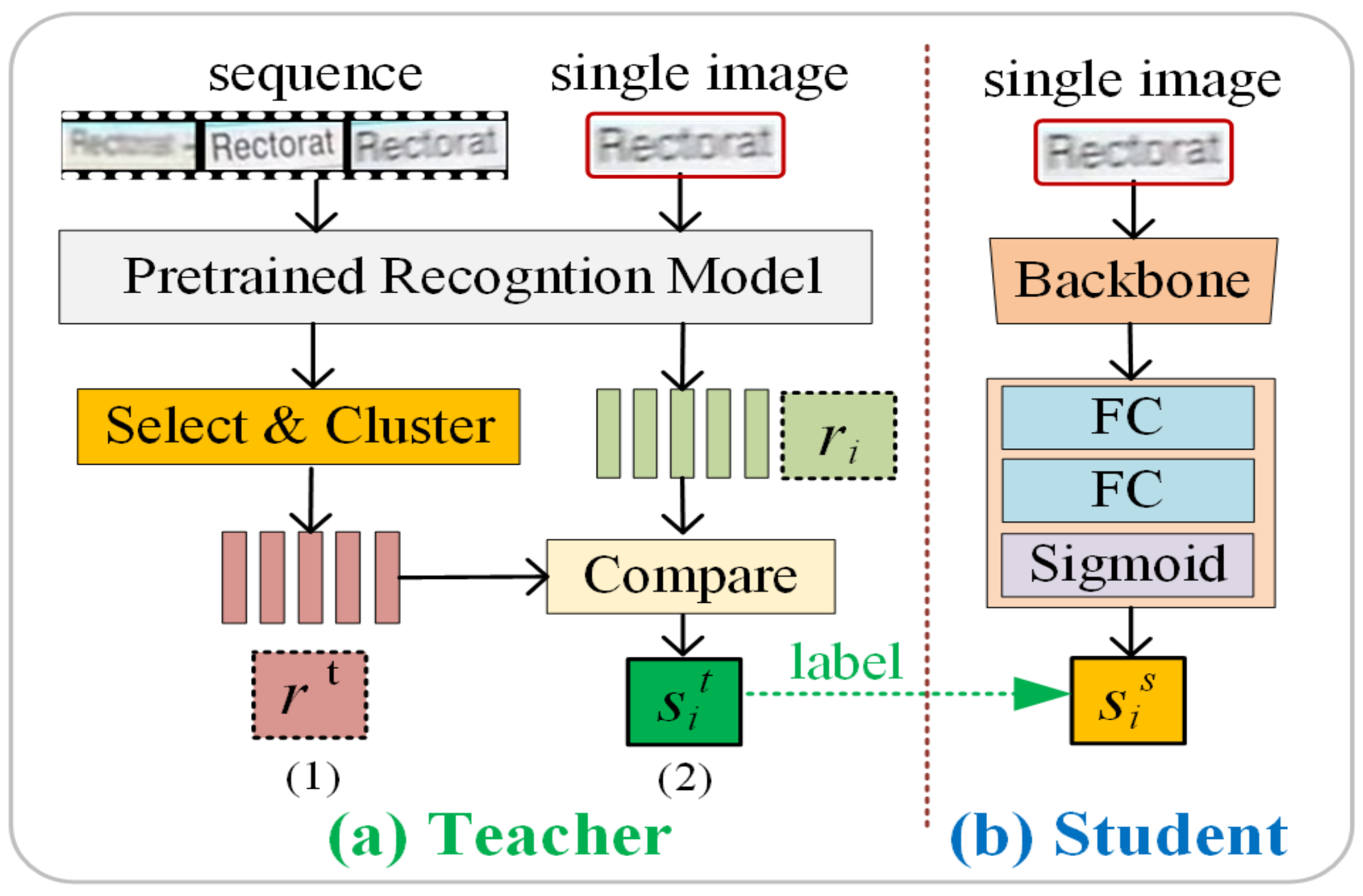
● *mechanism of quality scoring network*

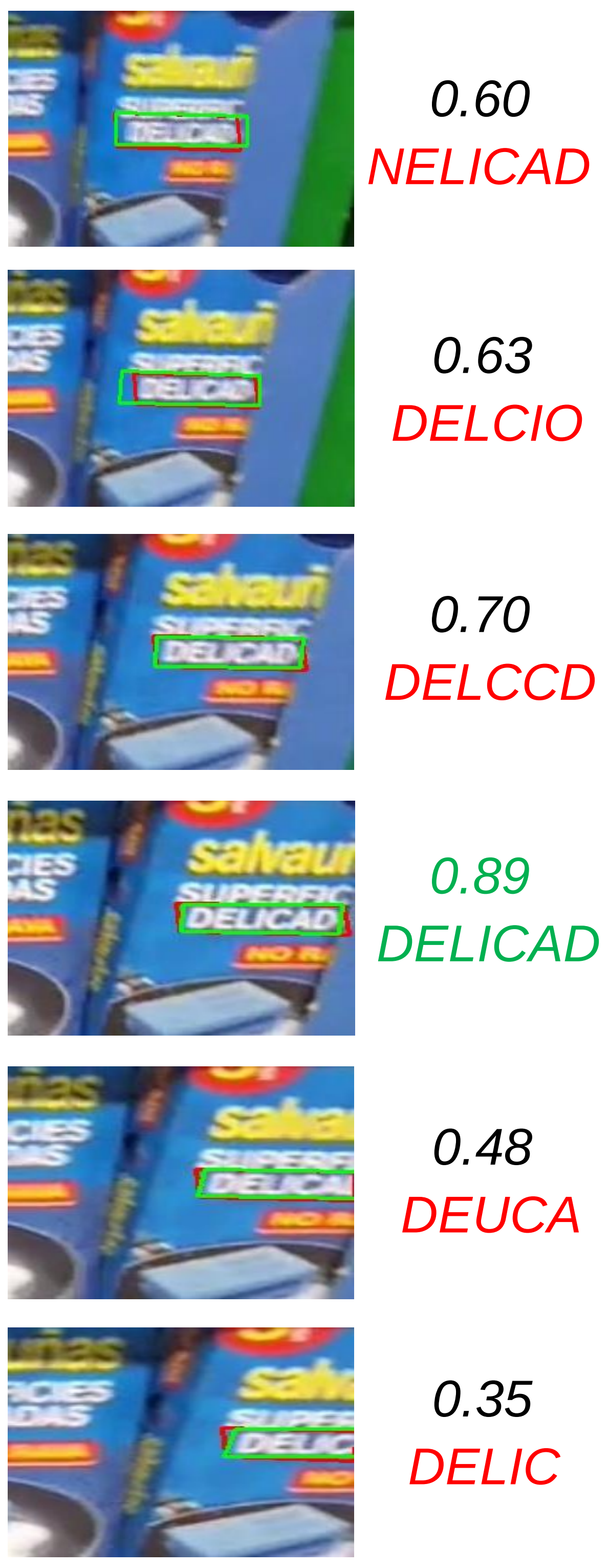


$$① \quad r^t = kmeans(r_1^{cor}, r_1^{cor}, \ldots, r_k^{cor})$$

$$② \quad s_i^t = \frac{r^t \odot r_i}{||r^t|| * ||r_i||}$$

● *teacher- student architecture*





➢ Experiments & Ablation

- ***ablation**:*

✓ performance and speed comparison with other frame selection methods

Methods	*QSHR* (IC13/IC15)	*RCR* (IC13/IC15)	*FPS*
PCW	74.55/75.83	66.06/66.32	4.52
HFP	75.32/76.34	68.30/68.56	
TR ($\mathcal{L}_S$)	77.89/79.69	68.89/69.41	324.58
TR ($\mathcal{L}_S$+$\mathcal{L}_T$)	78.64/80.36	69.12/69.82	
TR ($\mathcal{L}_S$+$\mathcal{L}_R$)	81.23/83.03	69.92/70.69	
TR ($\mathcal{L}$)	**81.74/83.29**	**70.18/70.95**	

✓ effectiveness of each module

D-BASE	✓	✓		
D-ST			✓	✓
TR ($\mathcal{L}_S$)	✓		✓	
TR		✓		✓
PRE$_s$	69.91	72.84	64.88	68.28
REC$_s$	54.34	61.73	61.54	67.21
F-score	61.15	66.83	63.17	**67.74**

PCW: select with recognition confidence   RCR: rate of correctly recognizing selected text regions
HFP: select by majority voting   D-BASE: single frame detection by east
QSHR: quality selection hitting rate   TR($\mathcal{L}_{(*)}$): text recommender trained only with tracking, scoring or recognition loss

- ***comparison with state-of-art**:*

Methods	*REC*	*PRE*	*F-measure*
Khare et al. [20]	41.40	47.60	44.30
Zhao et al. [58]	47.02	46.30	46.65
Shivakumara [42]	53.71	51.15	50.67
Yin et al. [55]	54.73	48.62	51.56
Wang et al. [52]	51.74	58.34	54.45
D-BASE	56.21	**85.76**	67.91
D-ST	**60.23**	81.45	**69.25**

✓ detection on IC13

Method	$MOTP_R$	$MOTA_R$	ATA_R
Stradvision [18]	0.69	0.57	0.29
Deep2Text [18]	0.62	0.35	0.19
Wang et al. [53]	0.70	0.69	0.60
Ours	**0.76**	**0.69**	**0.63**

✓ end-to-end on IC15

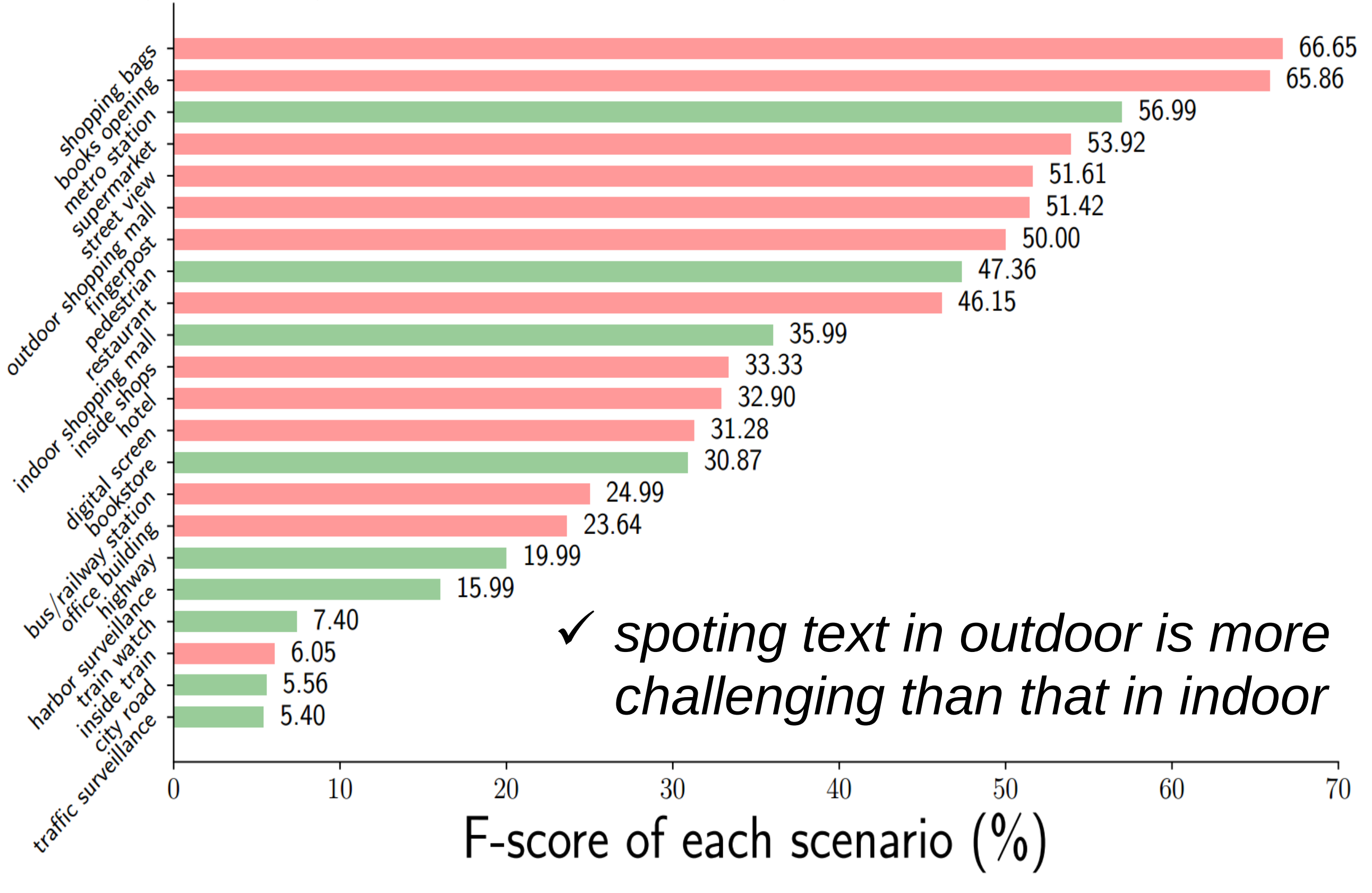
➢ Proposed Dataset (LSVTD)

- *existing video scene text datasets: limited scale and scenes, which may restrain research of video scene text spoting.*
- *our collected dataset:*

Datasets	#scenarios	#videos	#frames	#instances	quality?
Merino [28]	4	–	–	–	
Minetto [30]	–	5	3599	8706	
IC13 [19]	7	28	15277	93934	✓
YVT [33]	–	30	13500	–	
IC15 [18]	7	49	27824	–	✓
LSVTD	**22**	**100**	**66700**	**569300**	✓

✓ *22 indoor/outdoor real-world scenarios (100 videos)*   ✓ *multilingual*

- *end-to-end evalations on our dataset.*



✓ *spoting text in outdoor is more challenging than that in indoor*