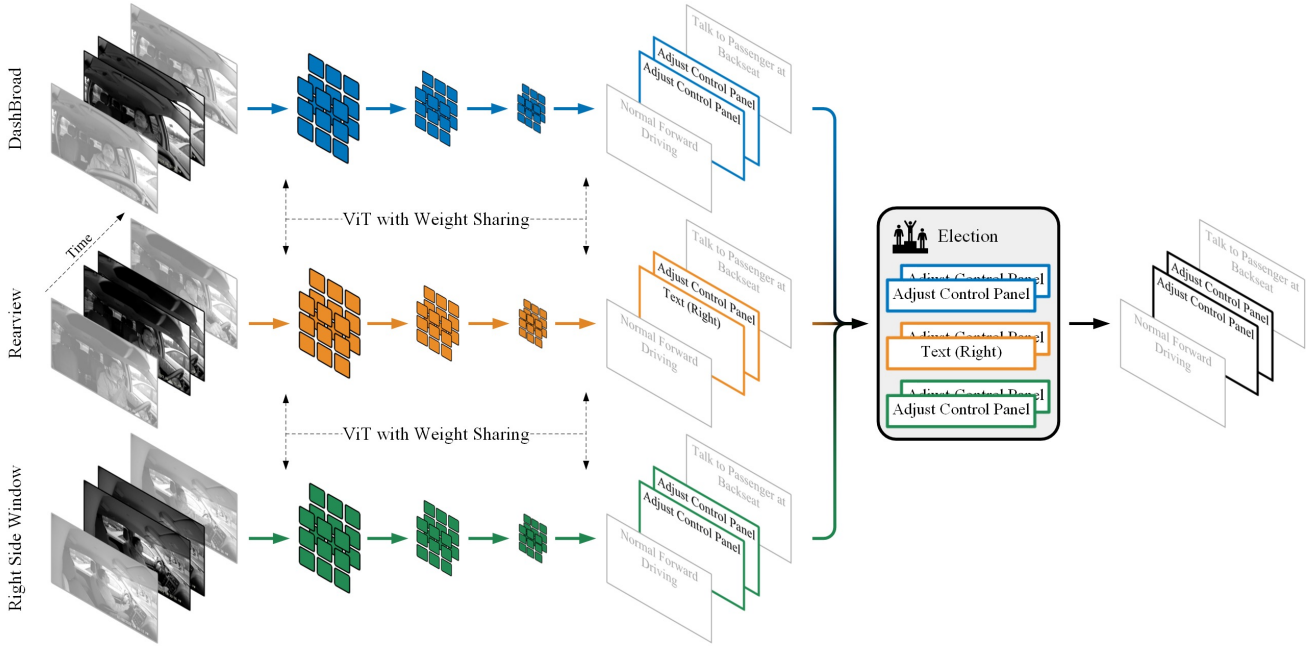
M2DAR: Multi-View Multi-Scale Driver Action Recognition with Vision Transformer

Yunsheng Ma, Liangqi Yuan, Amr Abdelraouf, Kyungtae Han, Rohit Gupta, Zihao Li, Ziran Wang

Purdue University, College of Engineering

Toyota Motor North America, InfoTech Labs

@The 7th AI City Challenge

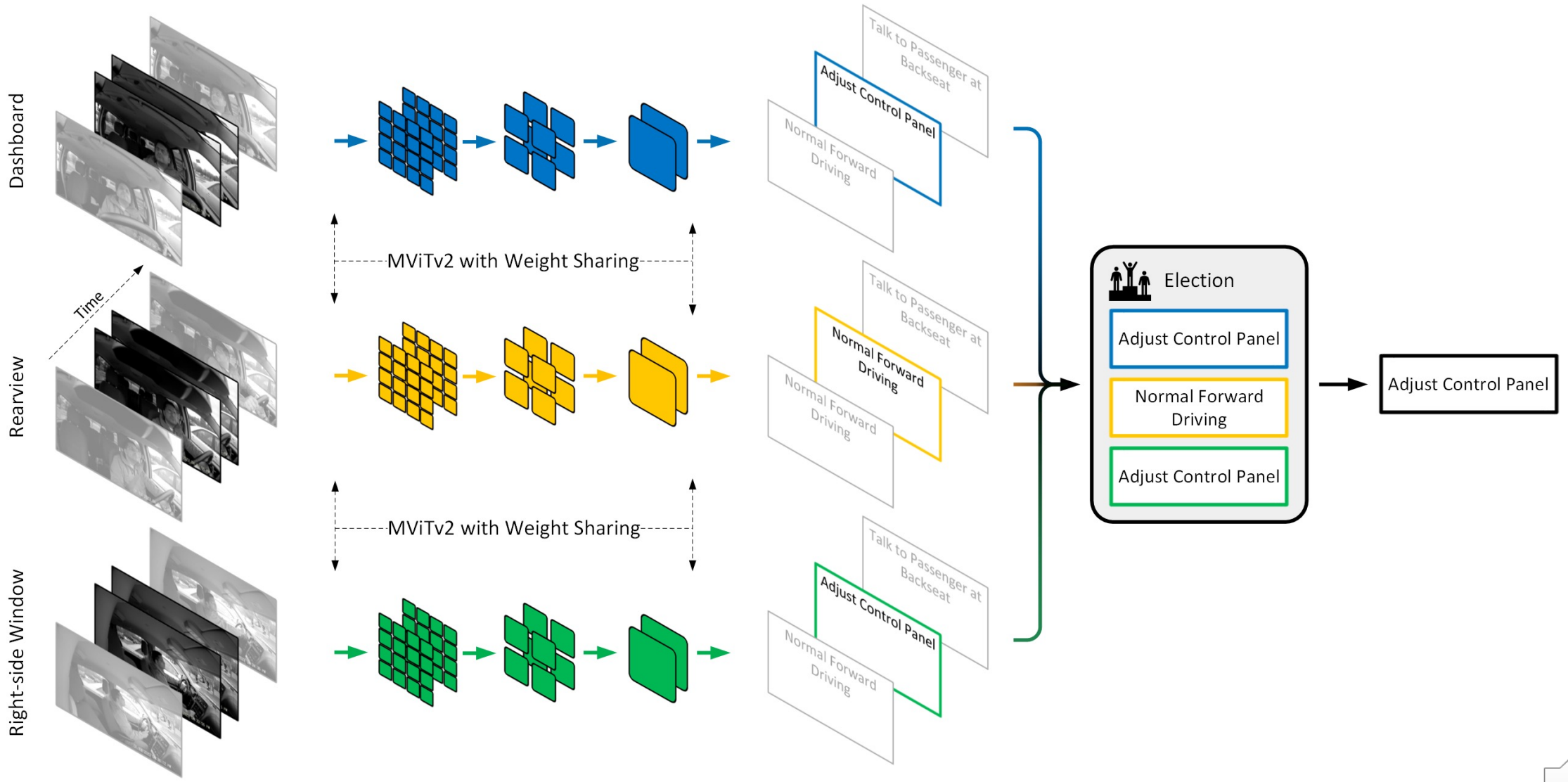


Background

- Distracted driving is a major cause of traffic accidents
- AI City Challenge 2023 has released a comprehensive dataset and organized a competition on naturalistic driving action recognition
- Our goal is to accurately determine the start and end times and identify the specific actions performed by a driver in each video, using input from multiple camera views.

Dashboard **Rear-View** **Right-Side-Window**

Methodology - Overview



Li, Yanghao, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. "Mvitv2: Improved multiscale vision transformers for classification and detection." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4804-4814. 2022.

Election Stage - Overview

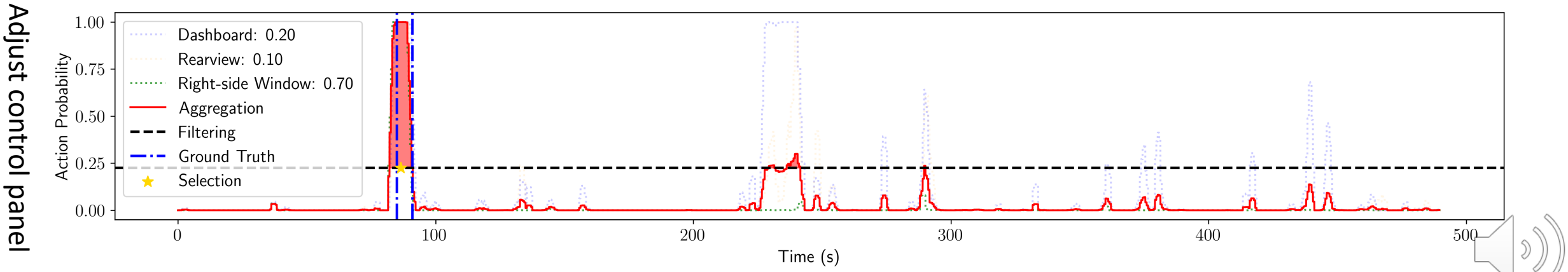
- Input: $\mathbf{p} \in \mathbb{R}^{T \times |\mathcal{C}| \times M}$
- Four steps:
  - Aggregation (AGG)
  - Filtering (FLTR)
  - Merging (MRG)
  - Selection (SEL)

Election Stage – Aggregation (AGG)

- Fuse information from various camera views

- Applies convolution operation to the input probability matrix using convolution kernels

- Convolution kernels weight information from each camera view differently for each action category

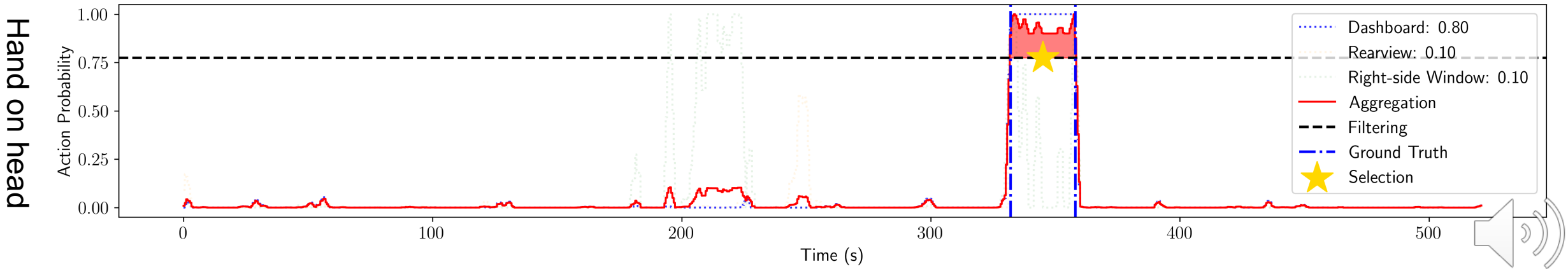


Election Stage – Filtering (FLTR)

- Identifies initial action candidates

- Extracts continuous frames with probability scores that exceed a predefined threshold for each action category

- Ensures a balance between recall and precision

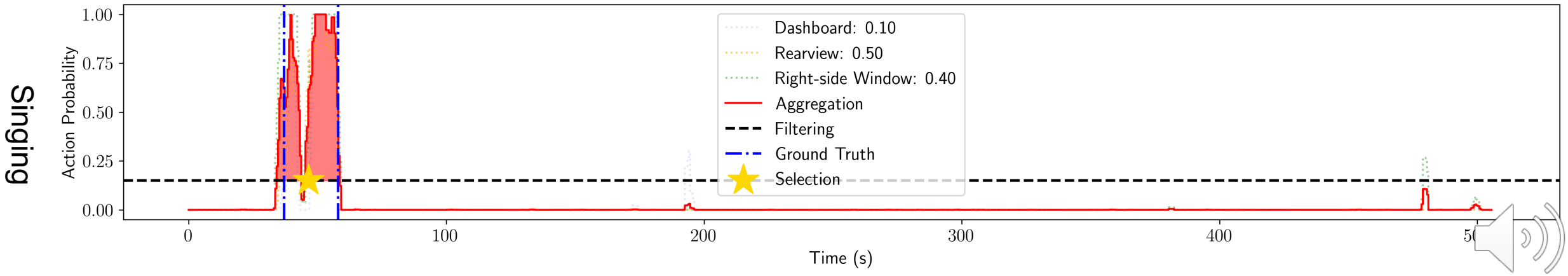


Election Stage – Merging (MRG)

- Merges clips that have a small temporal gap between them

- Iteratively compares the temporal distance between each pair of adjacent action candidate clips

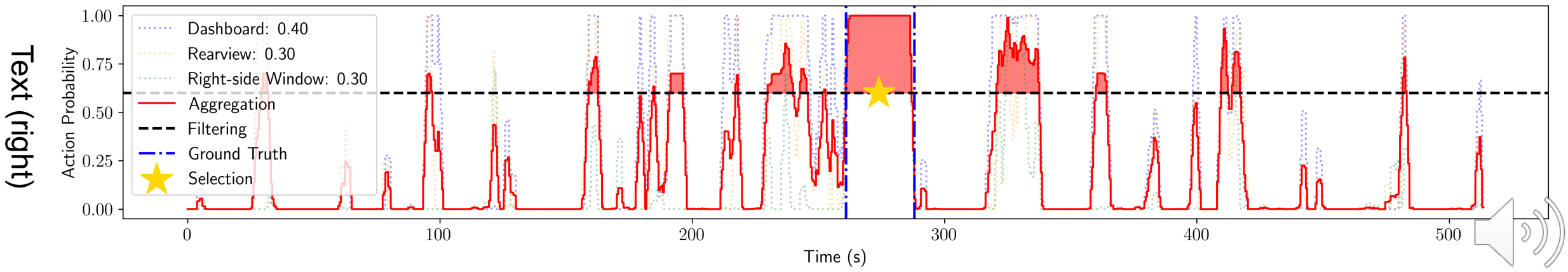
- Merges them if the distance is less than the predefined gap threshold



Election Stage – Selection (SEL)

- Computes the average score of all merged candidates for each action category

- Chooses the one with the highest average score as the final action candidates



Conclusion

- This work presents a multi-view multi-scale framework for detecting distracted driving behaviors in untrimmed videos
- The framework achieved **an overlap score of 0.5921** on the A2 test set of the AI City Challenge 2023 Track 3
- The proposed framework has the potential to aid in the development of more effective driver monitoring systems and ultimately improve road safety.