

Traffic Video Event Retrieval via Text Query using **Vehicle Appearance** and **Motion Attributes**

Tien-Phat Nguyen^{1,2}, Ba-Thinh Tran-Le^{1,2}, Xuan-Dang Thai^{1,2},
Tam V. Nguyen⁴, Minh N. Do⁵, Minh-Triet Tran^{1,2,3}



¹University of Science, Ho Chi Minh City, Vietnam

²Vietnam National University, Ho Chi Minh City, Vietnam

³John von Neumann Institute, Ho Chi Minh City, Vietnam

⁴University of Dayton, U.S.

⁵University of Illinois at Urbana-Champaign, Illinois, U.S.

Outline

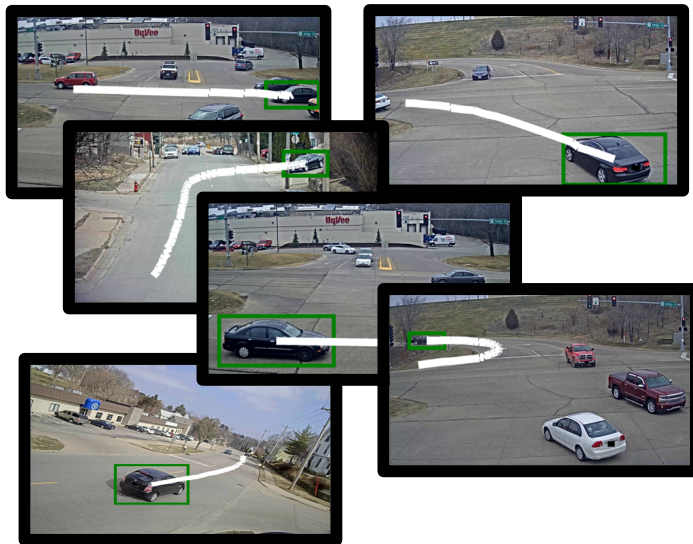
1. Problem statement
2. Proposed Method
 - Textual Attribute Extraction
 - Visual Attribute Extraction
 - Retrieval Model
 - Refinement method
3. Result
4. Conclusion

Problem Statement

Input query

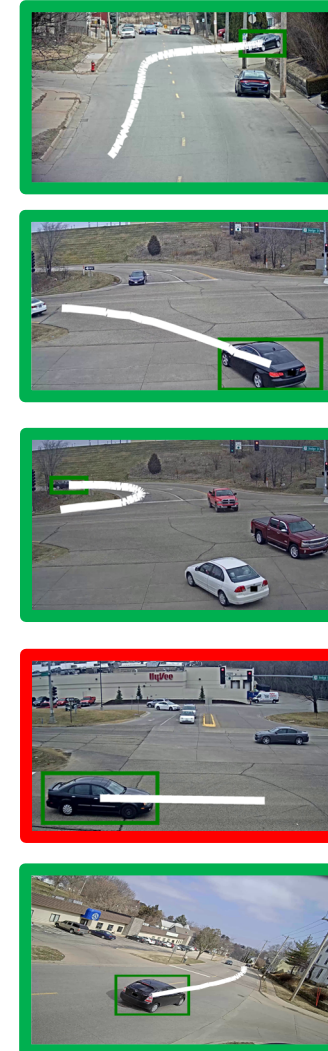
1. A black sedan turns left.
2. A black sedan stops.
3. Cadillac car makes a left turn onto the main street.

Video tracks



Retrieve
System

Rank list

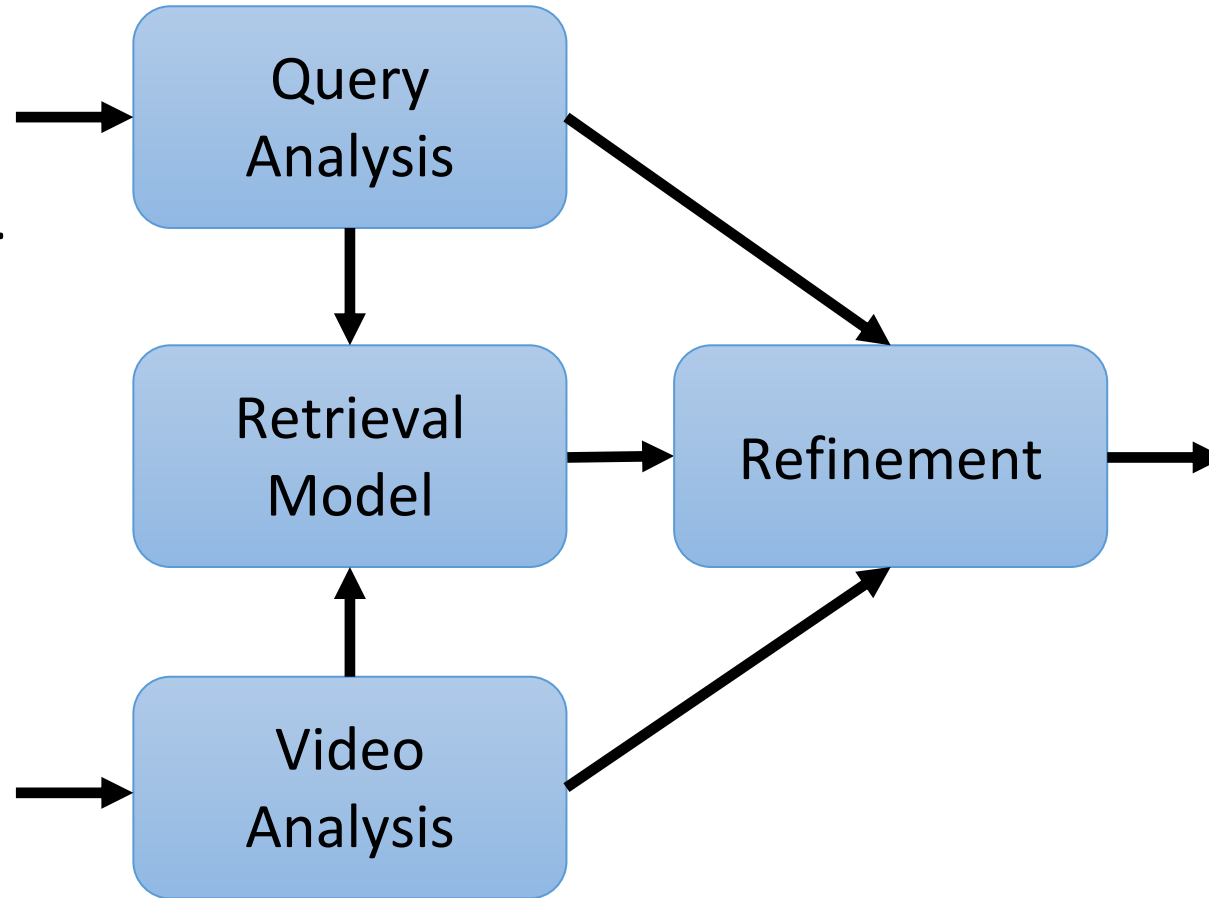
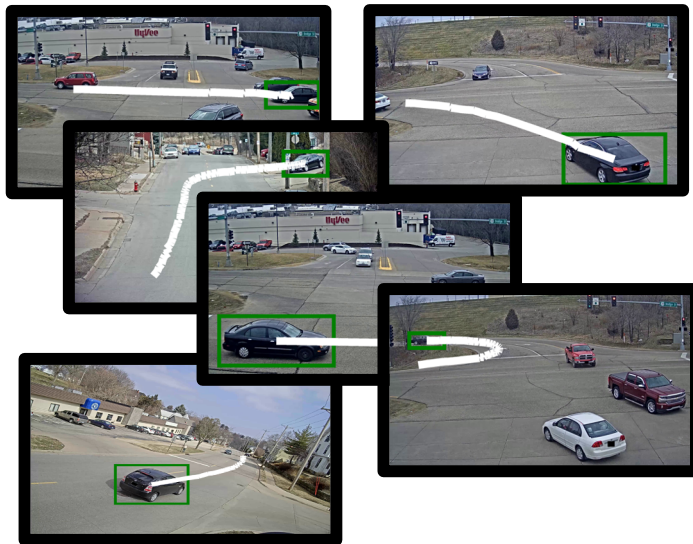


Proposed Method

Input query

1. A black sedan turns left.
2. A black sedan stops.
3. Cadillac car makes a left turn onto the main street.

Video tracks



Rank list

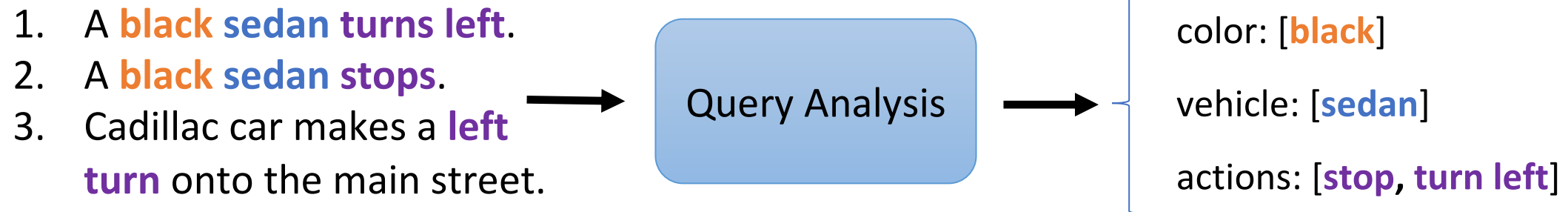


Outline

1. Problem statement
2. Proposed Method
 - Textual Attribute Extraction
 - Visual Attribute Extraction
 - Retrieval Model
 - Refinement method
3. Result
4. Conclusion

Textual Attribute Extraction

- We apply the Semantic Role Labeling (SRL) toolkit [1] to extract three main described characteristics of the target object: *color*, *vehicle type*, *action*.



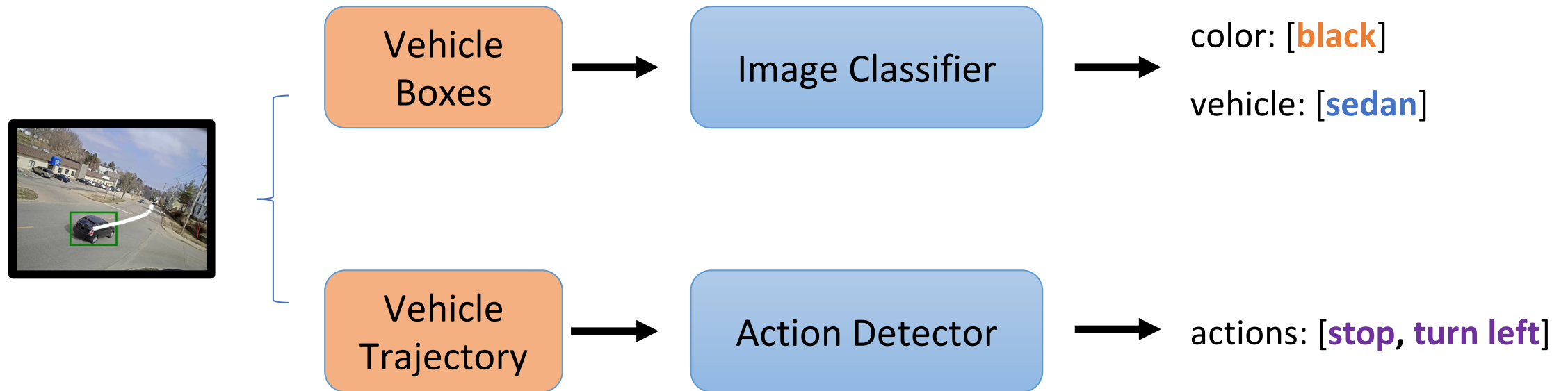
Textual Attribute Extraction - SRL Method

Query	A white semi truck	drives	straight up a hill	behind so many cars
Preprocessed query	<u>a white semi-trailer-truck</u>	<u>drove</u>	<u>straight up a hill</u>	<u>behind so many cars</u>
Extraction result	subject	action	argm_dir	argm_loc
Extraction Subject	subject: semi-trailer-truck color: white			

Outline

1. Problem statement
2. Proposed Method
 - Textual Attribute Extraction
 - Visual Attribute Extraction
 - Retrieval Model
 - Refinement method
3. Result
4. Conclusion

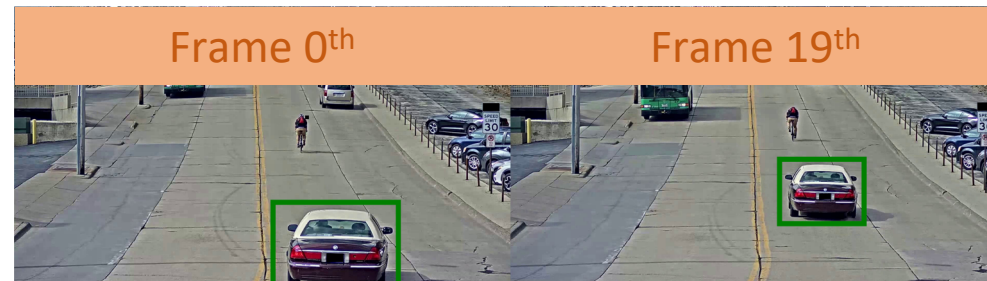
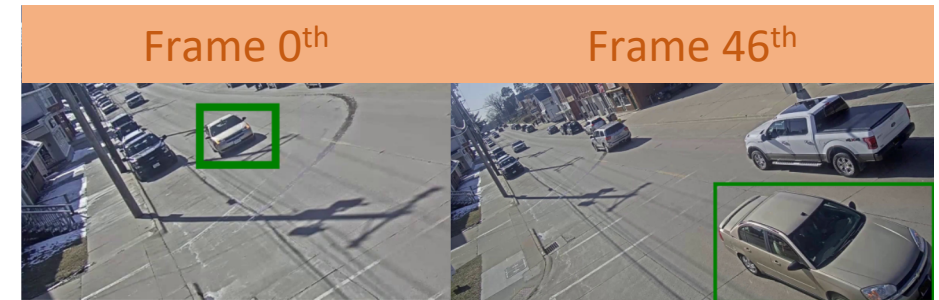
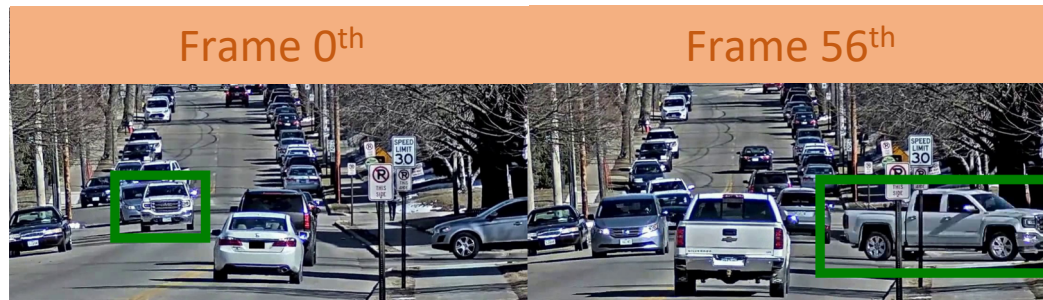
Visual Attribute Extraction



Visual Attribute Extraction

Image Classifier module

- Train with multi-label classification strategy to solve the ambiguous categories problem.
- Backbone: EfficientNet-B5



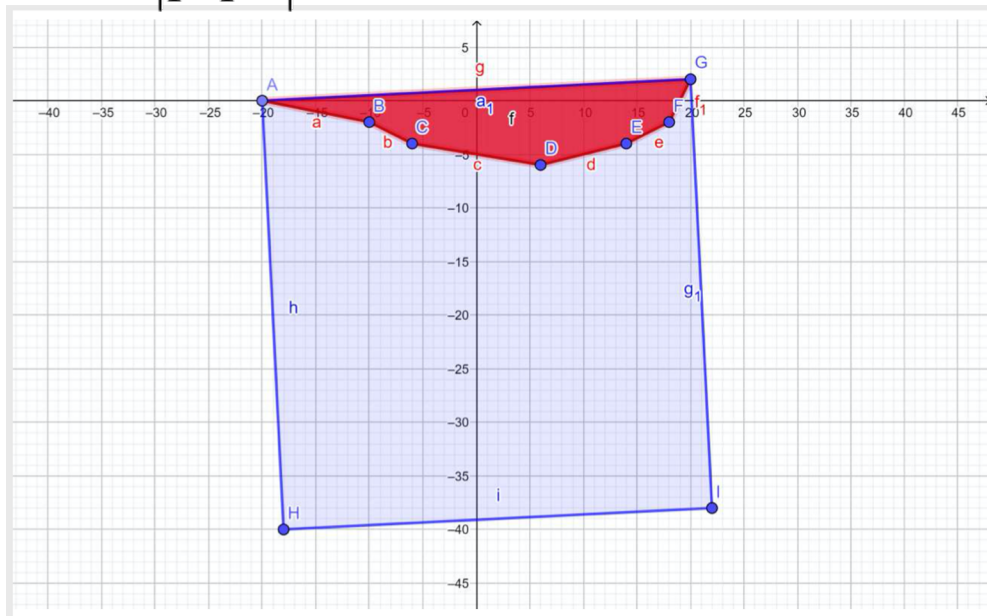
Visual Attribute Extraction

Action Detector module

Turn detection

$$A = \frac{\sum_{i=3}^n \overrightarrow{p_1 p_{i-1}} \times \overrightarrow{p_1 p_i}}{2} \begin{cases} A > 0 & : \text{Left} \\ A < 0 & : \text{Right} \end{cases}$$

$$B = \frac{A}{|p_1 p_n|^2} \longrightarrow \text{Turn or not}$$



Stop detection

$$v_i = \text{dist}(p_{i+k}, p_i)$$

$$v_i < \alpha \times \text{mean}\{v_i\}$$

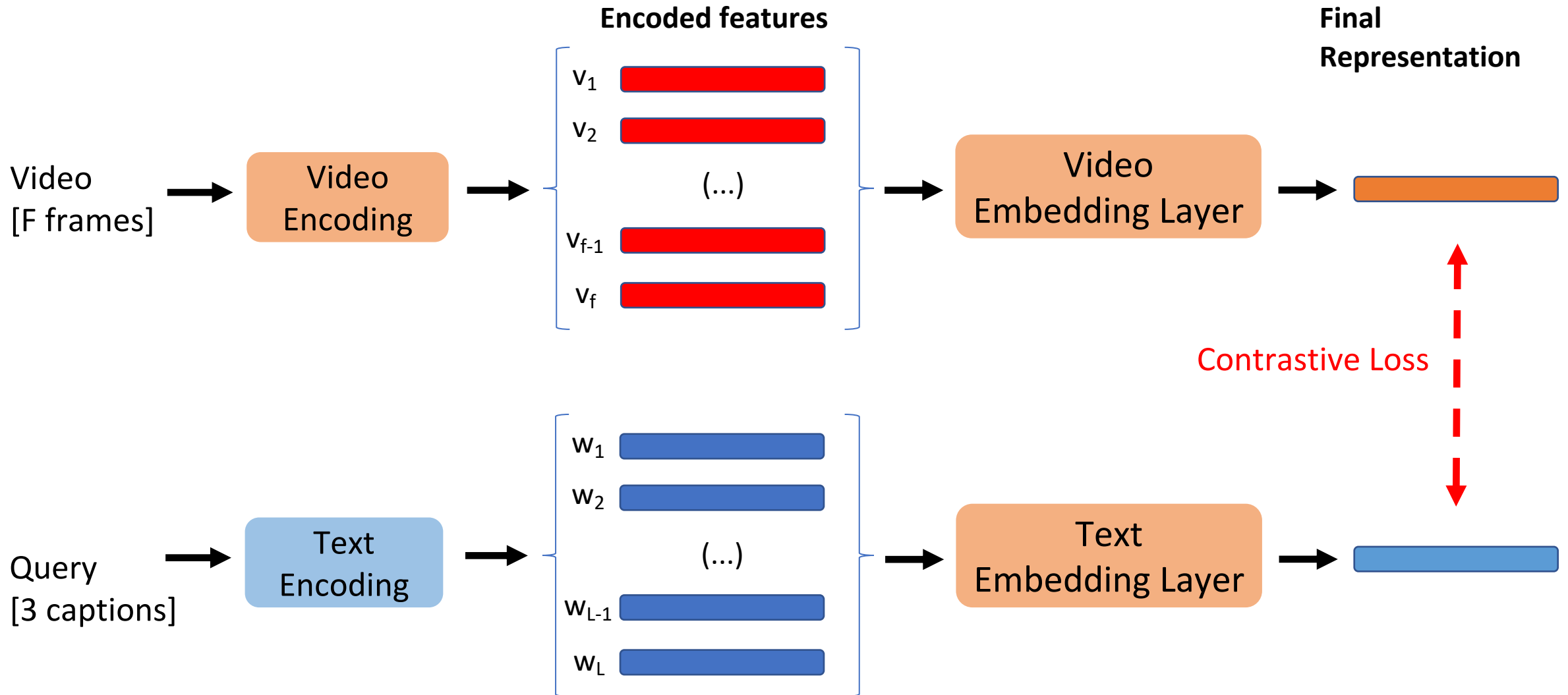
Stop or not

- p_i is the center of bounding box at i^{th} frame
- v_i is the vehicle motion speed at i^{th} frame

Outline

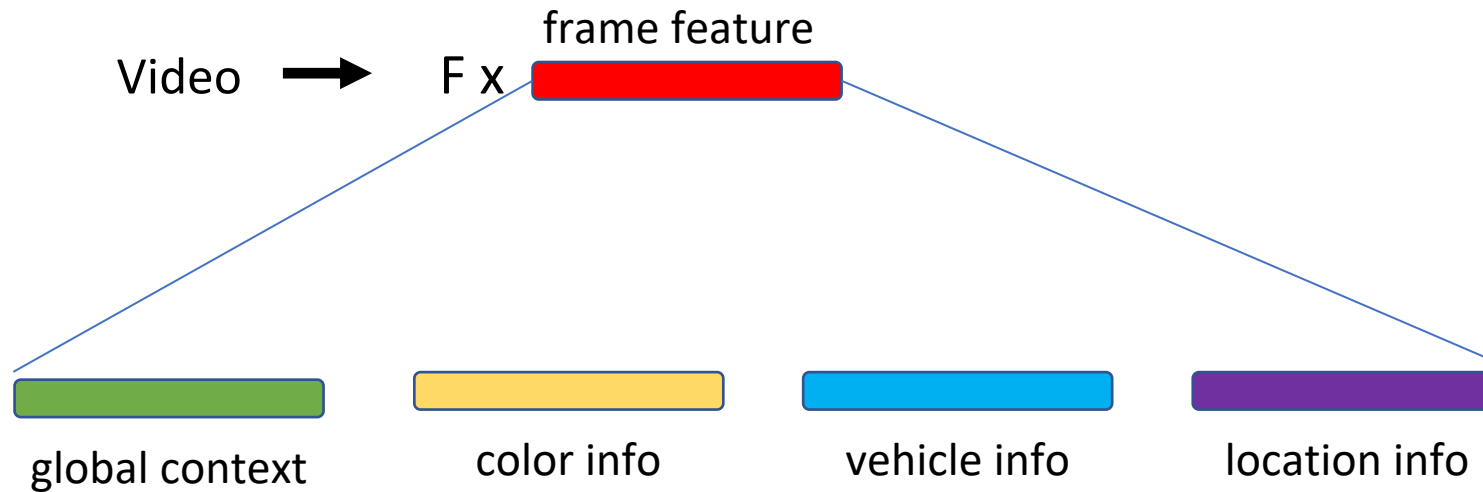
1. Problem statement
2. Proposed Method
 - Textual Attribute Extraction
 - Visual Attribute Extraction
 - Retrieval Model
 - Refinement method
3. Result
4. Conclusion

Retrieval Model



Retrieval Model

- **Text encoding:** pretrained BERT-based model
- **Video encoding:**

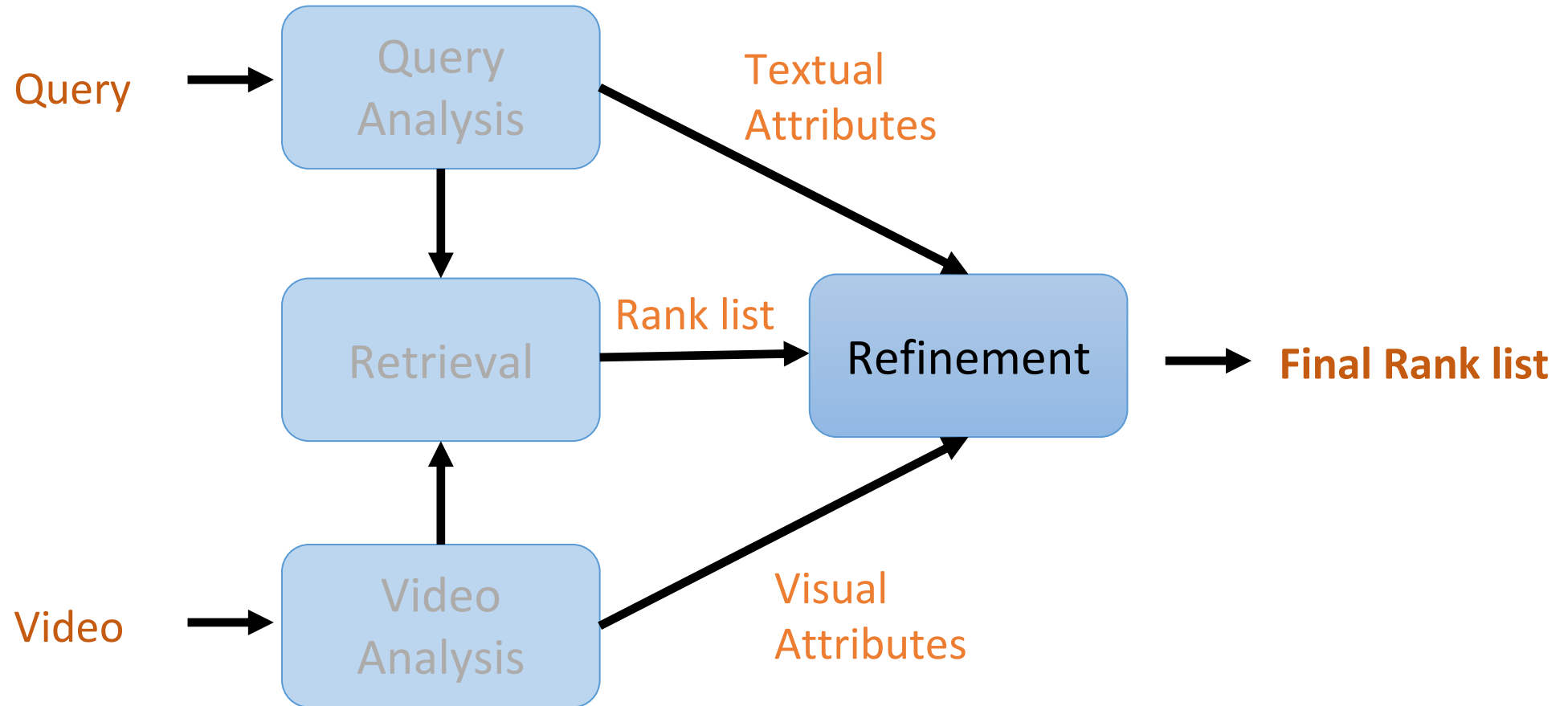


- **Text/video embedding layer:** Transformer-based embedding layer proposed by the COOT framework.

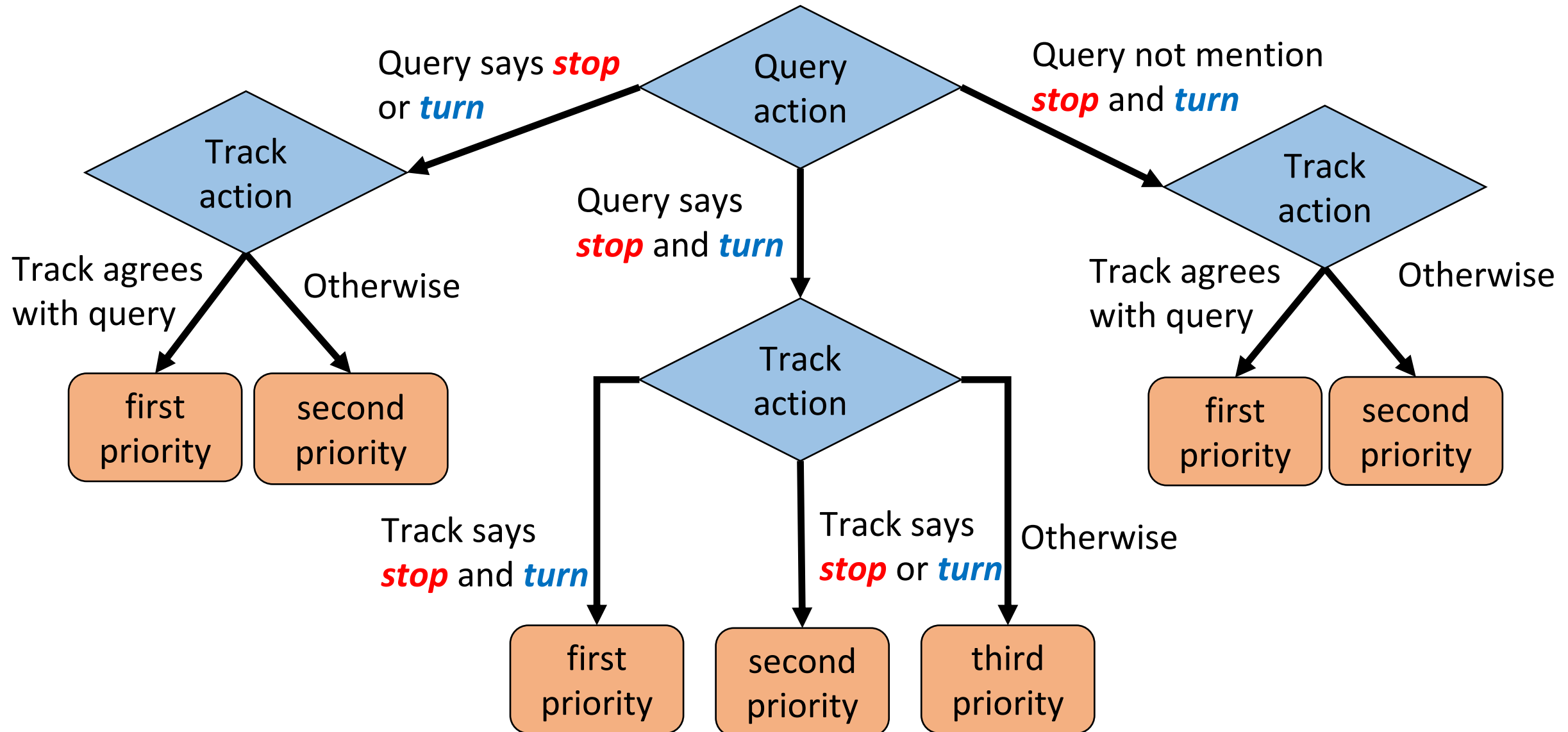
Outline

1. Problem statement
2. Proposed Method
 - Textual Attribute Extraction
 - Visual Attribute Extraction
 - Retrieval Model
 - Refinement method
3. Result
4. Conclusion

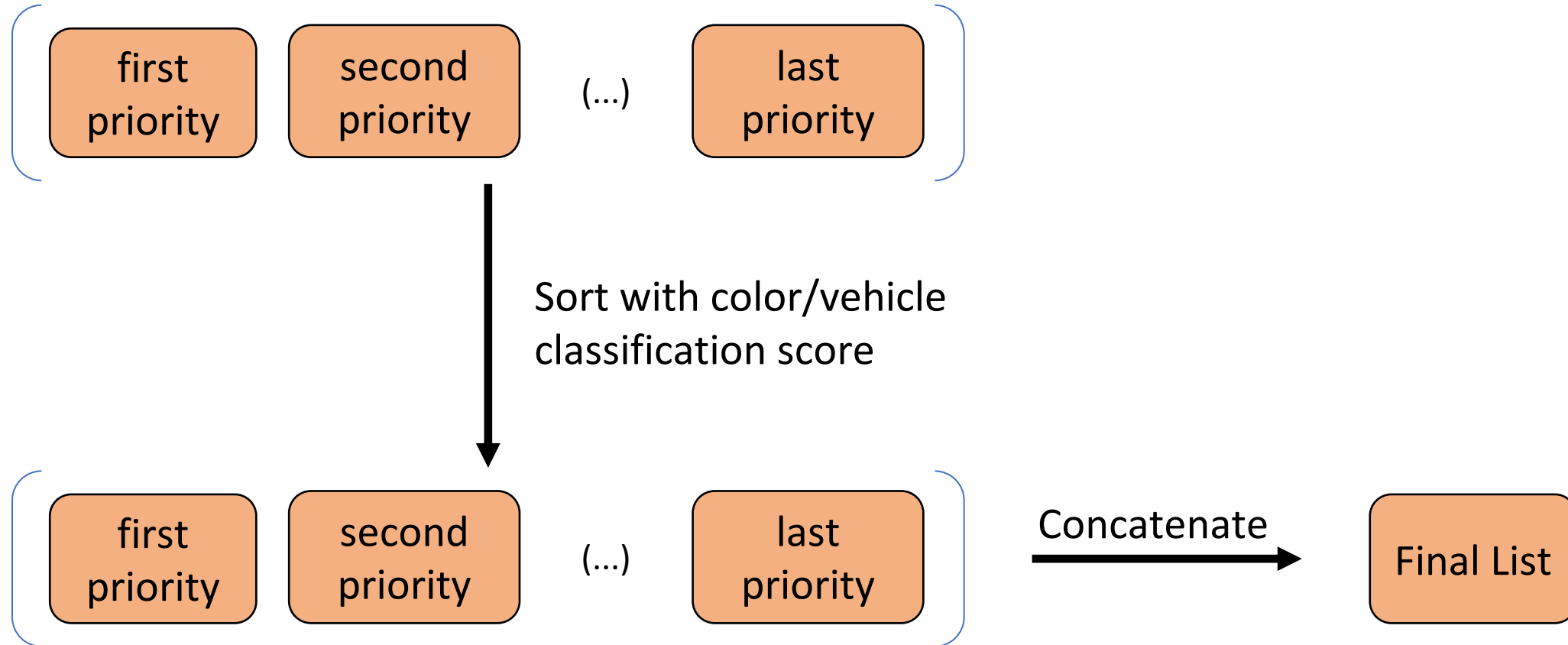
Refinement Method



Refinement Method - Action



Refinement Method – Color/Vehicle Type



Outline

1. Problem statement
2. Proposed Method
 - Textual Attribute Extraction
 - Visual Attribute Extraction
 - Retrieval Model
 - Refinement method
3. Result
4. Conclusion

Result

- Experimental result

Input query	A black sedan turns left. A black sedan stops. Cadillac car makes a left turn onto the main street.		
Top-3 output result			
V1 COOT			
V2 COOT refine col-veh			
V4 COOT refine col-veh-act			

No.	Configuration	MRR	R@5	R@10
V1	COOT	0.1010	0.1453	0.1869
V2	COOT + refine col-veh	0.1433	0.2000	0.3226
V3	COOT + refine col-veh w/o loc	0.1368	0.2113	0.3170
V4	COOT + refine col-veh-act	0.1560	0.2321	0.3302

Result

- Final score on AI City Challenge leaderboard

Rank	Team ID	Team Name	Score
1	132	Alibaba-UTS	0.1869
2	17	TimeLab	0.1613
3	36	SBUK	0.1594
4	20	SNLP	0.1571
5	147	HUST	0.1564
6	13	HCMUS	0.1560
7	53	VCA	0.1548
8	71	aiem2021	0.1364
9	87	Enablers	0.1314
10	6	Modulabs	0.1195
11	51	AIPERT	0.1078
12	11	CE-UIT	0.0852
13	30	Fiberhome-YYDS	0.0850
14	82	CMU INF	0.0184
15	146	UFL-nlp	0.0172

Outline

1. Problem statement
2. Proposed Method
 - Textual Attribute Extraction
 - Visual Attribute Extraction
 - Retrieval Model
 - Refinement method
3. Result
4. Conclusion

Conclusion

Advantages

- Our proposed retrieval system concentrates on vehicle appearance and movement attributes by adding local context modeling and hierarchical attribute refinement.
- We provide an efficient method that produces potential results without ensembling many models.

Further improvements

- Incorporate action modeling to an end-to-end method.
- Model target-objects relations to enhance current performance.

Acknowledgement

- This research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19 and National Science Foundation (NSF) under Grant No. 2025234

Thank you for listening!

Source code is available at:

https://github.com/selab-hcmus/AI_City_2021

