DBUS: Human Driving Behavior Understanding System

Guangyu Li*, Bo Jiang*, Zhengping Che, Xuefeng Shi, Mengyao Liu, Yiping Meng, Jieping Ye, and Yan Liu

DiDi

USC University of Southern California

ICCV 2019
Seoul, Korea

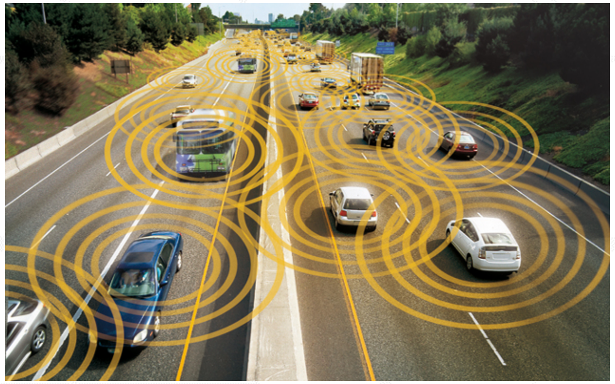
Outline

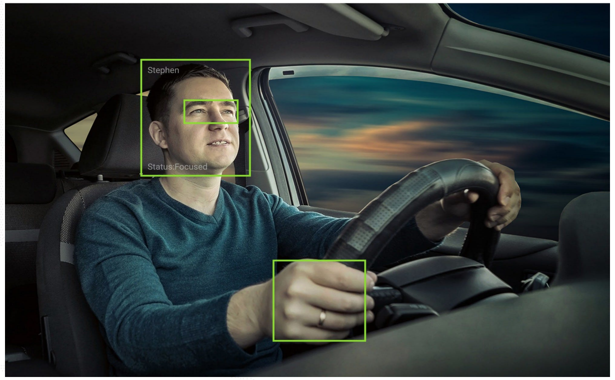
- Introduction
- Motivation
- Overview
- System Architecture
- Driving Behavior Analysis
- Experiment & Case Study
- Summary

Introduction

- What is human driving behavior understanding?
  - Understand how humans drive and interact with environments.

- Driving behavior understanding in intelligent transportation systems
  - Autonomous driving (AD)
  - Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I)
  - Driving safety monitoring systems
    - *Top priority for ride-sharing platform & fleet management*





Introduction

- Why is human driving behavior understanding essential?
  - AD:
    - Proposing human-like P&C strategies, building the realistic simulator, etc.
  - V2V/V2I:
    - Providing insight for an efficient and convenient design of systems, etc.
  - Driving safety monitoring system:
    - Real-time driving safety monitoring and comprehensive driving safety profiling, etc.
- Key *challenges* come from
  - Sophisticated real-world traffic scenarios -- *Environment*
  - Huge diversities of different driving styles -- *Drivers*

Motivation

- Collection of human driving data on ride-sharing platform
  - *Heterogeneous* -- including videos, GPS/IMU signals, etc.
  - *Large-scale* -- offering potential for in-depth analysis

- Major challenges of human driving understanding
  - Handling different complex tasks
    - Perception, reasoning, attention, planning, etc.
  - Multiple-level analyzation
    - Low-level *driving action recognition* -> high-level driver's *attention & cause inference*
  - Heterogeneous data source
    - Each type of data has its own properties and works with different methodologies
  - System efficiency
    - Which plays a crucial role for practical deployment

Motivation

- Existing works for human driving behavior understanding
  - Only on relatively small dataset
  - Only analysis it at single level or on subset of tasks
  - Only leverage single type of data

- Our proposed solution
  - **DBUS: D**riving **B**ehavior **U**nderstanding **S**ystem

Overview

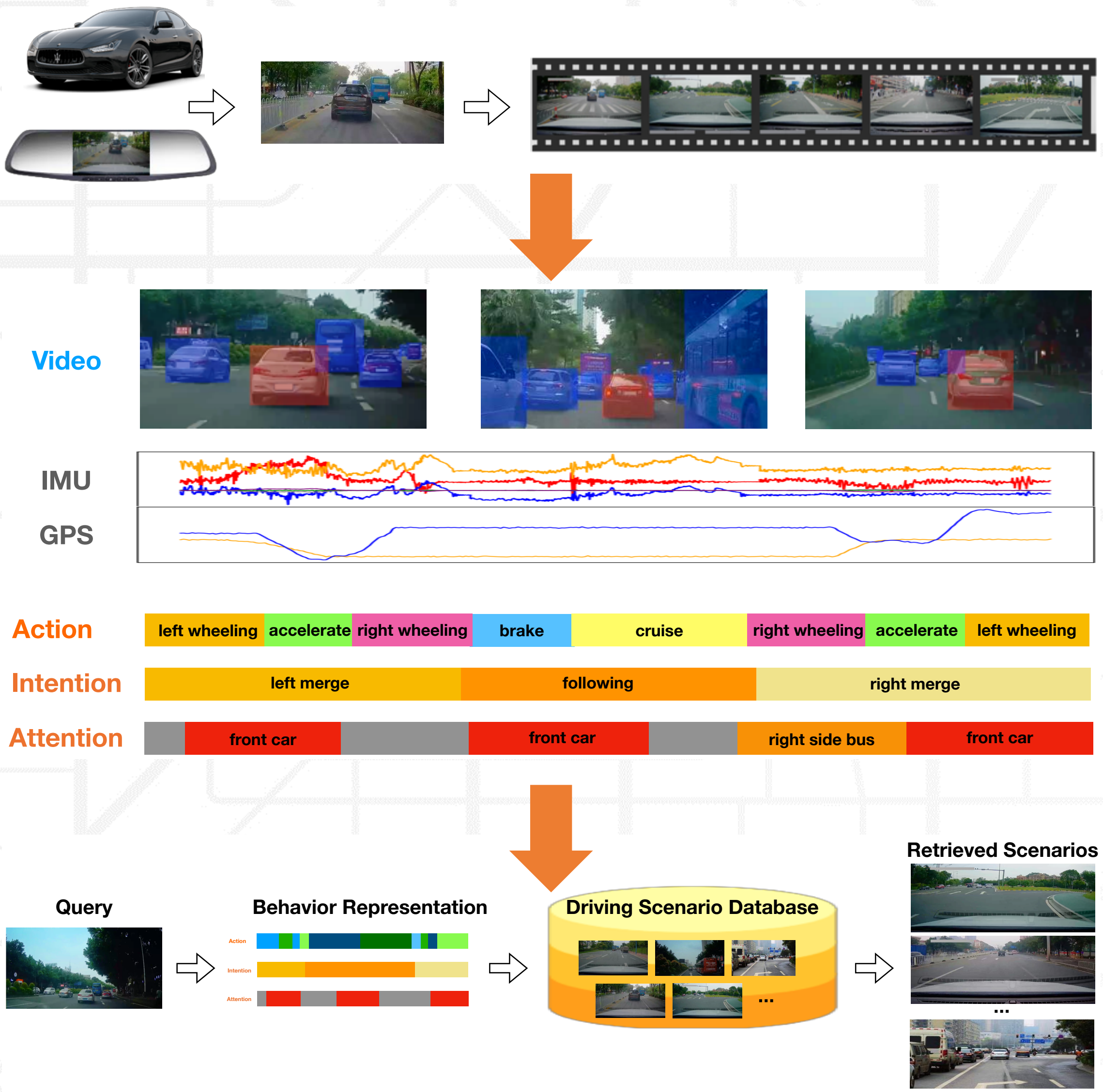
- Problem formulation
  - Input – Human driving scenario $\mathcal{D} = (\mathbf{V}, \mathbf{S})$ with time horizon T
    - $\mathbf{V} = \{\mathrm{v}_t\}_{t=1}^{T}$ refers frames from front-view camera
    - $\mathbf{S} = \{s_t\}_{t=1}^{T}$ denotes GPS/IMU signals
  - Output – Three-level structured representation of driving behavior $\mathcal{B} = (\mathbf{M}, \mathbf{W}, \mathbf{A})$
    - $\mathbf{M} = \{\mathrm{m}_t\}_{t=1}^{T}$ refers pre-defined basic driving actions
    - $\mathbf{W} = \{\mathrm{w}_t\}_{t=1}^{T}$ refers pre-defined driver's intentions
    - $\mathbf{A} = \{\mathrm{a}_{mask}^{t}, \mathrm{a}_{obj}^{t}\}_{t=1}^{T}$ refers driver's attention object categories

Overview

- We consider
  - 9 basic driving actions
  - 8 driver's intentions
  - 8 driver's attention object categories

Behavior	# of Categories	Categories
Driving Action ($\mathbf{m}$)	9	{ *left_accelerate*, *left_cruise*, *left_brake*, *straight_accelerate*, *straight_cruise*, *straight_brake*, *right_accelerate*, *right_cruise*, *right_brake* }
Driving Intention ($\mathbf{w}$)	8	{ *following*, *left_turn*, *right_turn*, *left_lane_change*, *right_lane_change*, *left_merge*, *right_merge*, *U_turn* }
Driving Attention ($\mathbf{a}_{obj}$)	8	{ car, bus, truck, person, bicycle, motorcycle, tricycle, traffic light }

Overview



Real-World Human Driving Scenarios

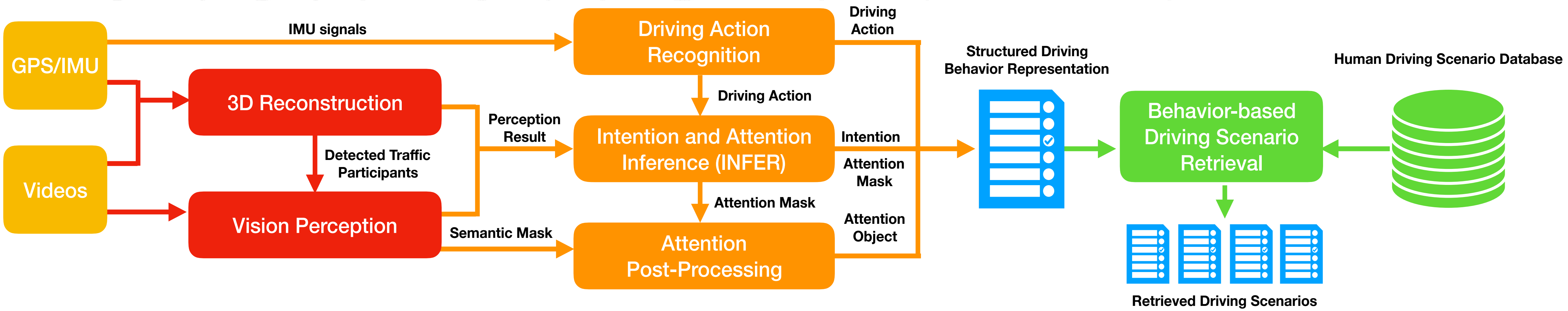
Structured Driving Behavior Representation

Behavior-Based Driving Scenarios Retrieval

Overview

- Major tasks and applications
  - Driving behavior analysis
    - Joint inference of intention $\mathbf{W}$ and attention $\mathbf{A}$
  - Driving scenario search & retrieval
    - Given $\mathcal{B} = (\mathbf{M}, \mathbf{W}, \mathbf{A})$, retrieve top-$K$ relevant $\{\mathcal{D}_k = (\mathbf{V}_k, \mathbf{S}_k)\}_{k=1}^{K}$ from a *massive* database

System Architecture



- Major modules of **DBUS**
  - **Perception**
    - Vision perception and 3D reconstruction of driving scenarios, etc.
  - **Driving Behavior Analysis** – ***core module***
    - Generates 3-level *structured representation* with perception & GPS/IMU signals
  - **Driving Scenario Retrieval**
    - Efficient behavior-based retrieval of relevant driving scenarios

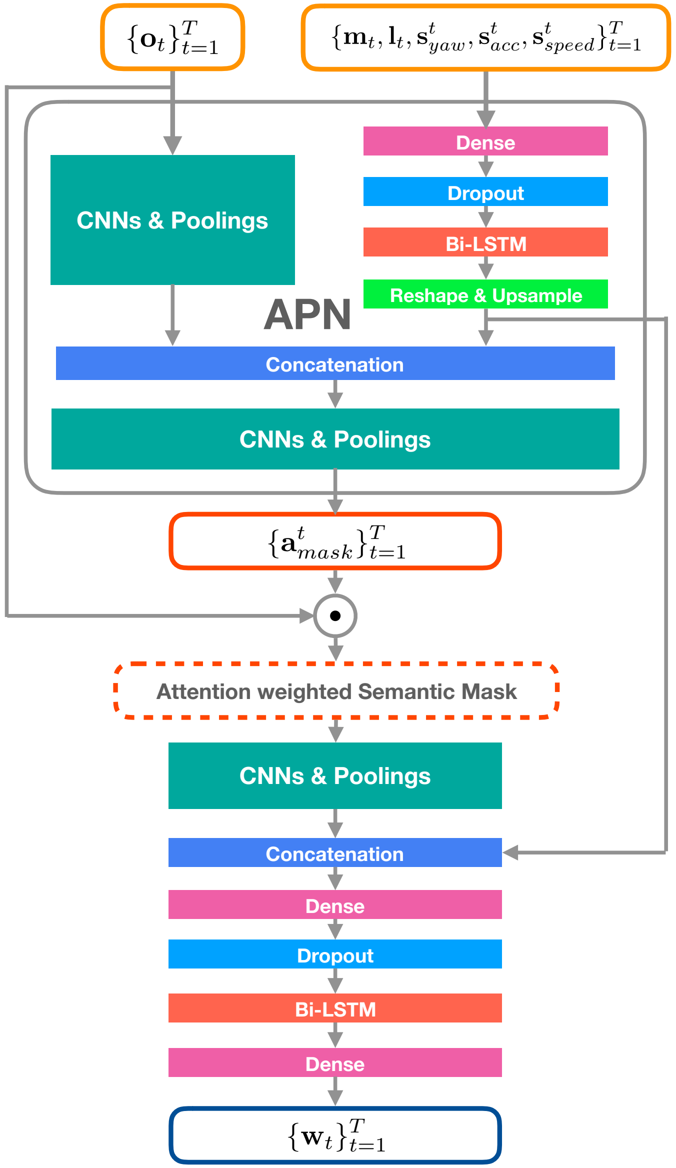
System Architecture

- Workflow of DBUS
  1. Feed $\mathcal{D} = (\mathbf{V}, \mathbf{S})$ -> **Perception** module -> $\mathcal{P} = (\mathbf{O}, \mathbf{D}, \mathbf{L})$
     - $\mathbf{O}$: Semantic mask of detections; $\mathbf{D}$: Following distance; $\mathbf{L}$: Relative location
  2. Feed $\mathcal{P} = (\mathbf{O}, \mathbf{D}, \mathbf{L})$ + GPS/IMU -> **Driving Behavior Analysis** -> $\mathcal{B} = (\mathbf{M}, \mathbf{W}, \mathbf{A})$
  3. The **Driving Scenario Retrieval** takes $\mathcal{B}$ and returns top-K relevant $\mathcal{D}$

Notation			Type	Definition
Raw Data ($\mathcal{D}$)	Video ($\mathbf{V}$)	$\mathbf{v}$	image	front-view video frames
	GPS/IMU ($\mathbf{S}$)	$vs._{speed}$	$\mathbb{R}$	vehicle speed in GPS signals
		$vs._{acc}$	$\mathbb{R}$	forward accelerate in IMU signals
		$vs._{yaw}$	$\mathbb{R}$	yaw angular velocity in IMU signals
Perception Result ($\mathcal{P}$)	Objects ($\mathbf{O}$)	$\mathbf{o}$	mask	semantic mask of detected traffic participants and traffic lights
	Distance ($\mathbf{D}$)	$\mathbf{d}$	$\mathbb{R}$	distance between ego-vehicle and nearest front traffic participants
	Locations ($\mathbf{L}$)	$\mathbf{l}$	category	vehicle's location on the road based on the lane perception results
Behavior Representation ($\mathcal{B}$)	Action ($\mathbf{M}$)	$\mathbf{m}$	category	basic driving actions
	Intention ($\mathbf{W}$)	$\mathbf{w}$	category	driving intention
	Attention ($\mathbf{A}$)	$\mathbf{a}_{mask}$	mask	driving attention mask
		$\mathbf{a}_{obj}$	category	object category of driving attention

Driving Behavior Analysis

- Basic driving action inference
  - Based on GPS/IMU signals $\mathbf{S}$ with a rule-based manner
- Intention and attention inference
  - Introduced a deep model **INFER**
    - Attention proposal network (APN)
    - Intention inference network
  - Input – Set of features $\{o_t, d_t, l_t, vs._{yaw}^{t}, vs._{acc}^{t}, vs._{speed}^{t}\}_{t=1}^{T}$
  - Output – Driving intention $\{w_t\}_{t=1}^{T}$ & attention mask $\{a_{mask}^{t}\}_{t=1}^{T}$
    - Note: we use $\{a_{mask}^{t}\}_{t=1}^{T}$ + $\{o_t\}_{t=1}^{T}$ to find $\{a_{obj}^{t}\}_{t=1}^{T}$



Experimental Settings

- Dataset

# of Samples	Time Horizon *T*	Sampling Rate	Video Resolution
2759	25	5Hz	1920 × 1080

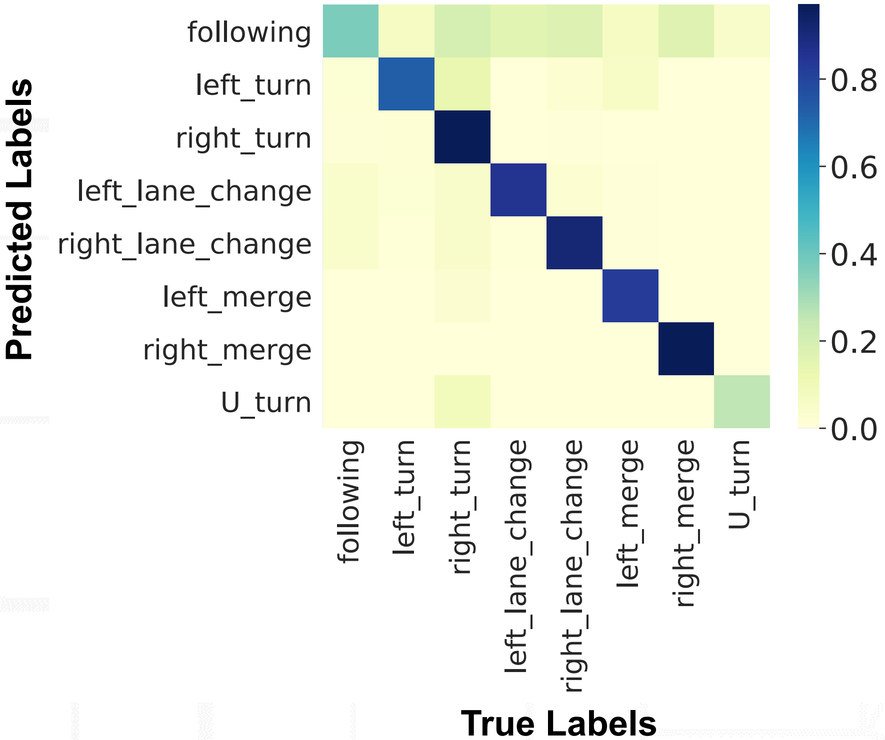
- Baselines
  - SVM – predicts driving intentions $\{\mathrm{w}_t\}_{t=1}^{T}$ only
  - XGBoost – predicts $\{\mathrm{w}_t\}_{t=1}^{T}$ only
  - INFER-NO-SM – without using $\{\mathrm{o}_t\}_{t=1}^{T}$, predicts $\{\mathrm{w}_t\}_{t=1}^{T}$ only
  - INFER-ONLY-SM – only using $\{\mathrm{o}_t\}_{t=1}^{T}$, predicts both $\{\mathrm{w}_t\}_{t=1}^{T}$ & $\{\mathrm{a}_{mask}^{t}\}_{t=1}^{T}$
  - INFER-NO-ATTN – without inferring $\{\mathrm{a}_{mask}^{t}\}_{t=1}^{T}$, outputs $\{\mathrm{w}_t\}_{t=1}^{T}$ only

Quantitative Results

- Results of attention & intention prediction

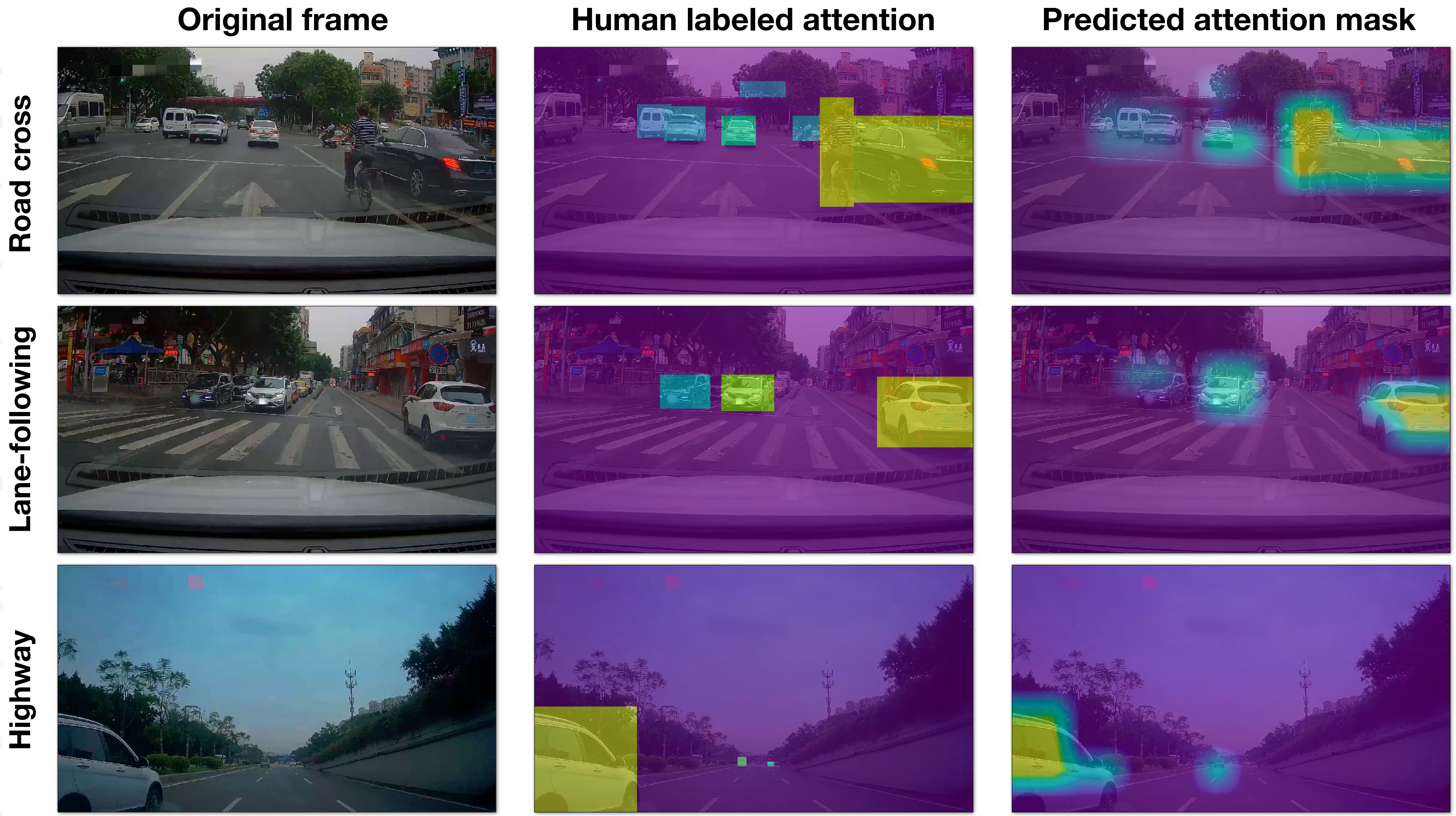
	MSE (Attention masks)	ACC (Intentions)
SVM	-	0.193
XGBoost	-	0.258
INFER-NO-SM	-	0.276
INFER-ONLY-SM	0.032	0.693
INFER-NO-ATTN	-	0.628
INFER	**0.025**	**0.772**

- Confusion matrix



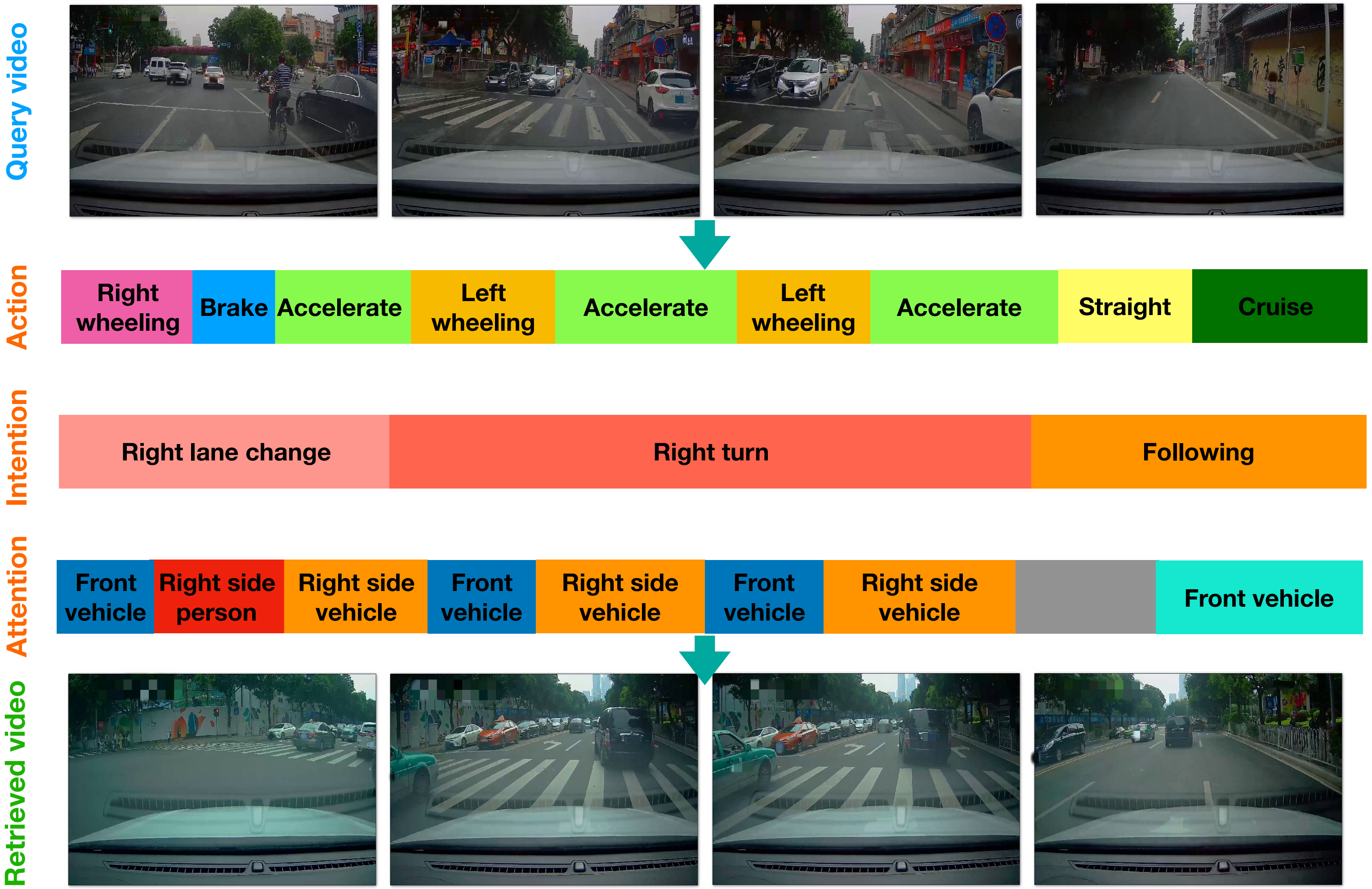
Case Study

- Attention inferred by DBUS



Case Study

- Retrieved driving scenario with query & its structured representation



Summary

- **DBUS**: Human **D**riving **B**ehavior **U**nderstanding **S**ystem
  - Integrated with **Perception**, **Driving Behavior Analysis** and **Driving Scenario Retrieval**
  - A structured representation of driving behavior is designed with multi-level understanding
  - Developed INFER, jointly infers driver's intention & attention
  - Provides a total solution for mining *large-scale* human driving behavior data

- Future works
  - Design semi-supervised learning method
    - To alleviate the demand for data annotation
  - Optimize the system with pruning & compression of deep neural networks
    - To deploy DBUS into on-vehicle embedded platforms

Thank you!

Our Poster is @ Room 401 #155

DiDi AI Labs http://ai.didiglobal.com/

USC Melady Lab https://melady.usc.edu/