SAE INTERNATIONAL® Mobility, Advanced™



SAE INTERNATIONAL



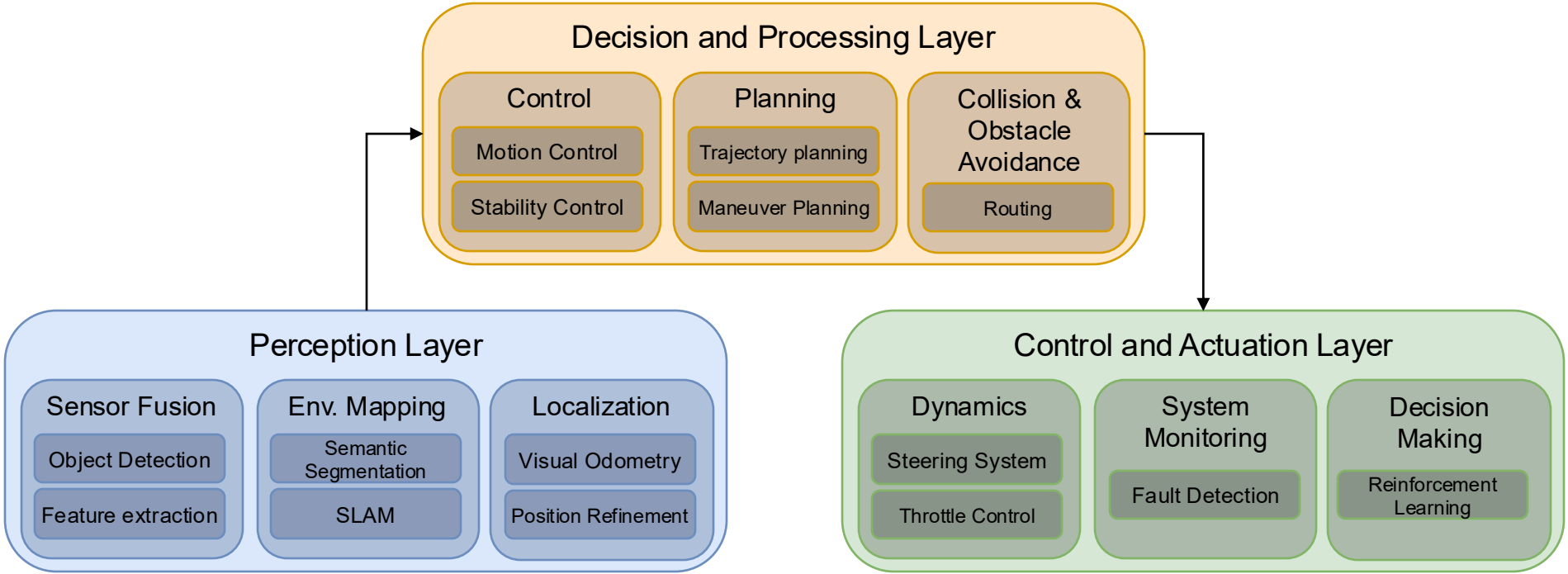
Avoiding the CRASH

A Vision-Language Model Evaluation of Critical Traffic Scenarios

CLEMSON UNIVERSITY

TIGERSEC

DNN in ADS


Decision and Processing Layer
Control
Motion Control
Stability Control
Planning
Trajectory planning
Maneuver Planning
Collision & Obstacle Avoidance
Routing
Perception Layer
Sensor Fusion
Object Detection
Feature extraction
Env. Mapping
Semantic Segmentation
SLAM
Localization
Visual Odometry
Position Refinement
Control and Actuation Layer
Dynamics
Steering System
Throttle Control
System Monitoring
Fault Detection
Decision Making
Reinforcement Learning

Challenges

DNN Challenges

- Imperfect Generalization
- Black Box Problem
- Model Drift
- Continuous Retraining
  - Resource Demands

Human Drivers Challenges

- In emergency situations, drivers experience stress
  - Cloud judgement
  - Impulsive reaction
  - Failure to act

LLMs and VLMs

- Large Language Models (LLMs)
  - Pretrained on massive datasets
  - Learning capabilities
  - Semantic understanding

- Vision Language Models (VLMs)
  - LLMs + capable of understanding image data
    - LLaVA-7B [1]
    - MoE-LLaVA [2]

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," University of Wisconsin-Madison, Microsoft Research, Columbia University. [Online]. Available: https://llava-vl.github.io

[2] B. Lin et al., "MoE-LLaVA: Mixture of Experts for Large Vision-Language Models," arXiv:2401.15947 [cs.CV], Jan. 2024. [Online]. Available: https://arxiv.org/abs/2401.15947

Research Questions

Goal: Evaluate VLMs ability to handle challenging driving scenarios

1. Can VLMs overcome Human Driver Limitations?
2. Can VLMs overcome DNN-Based AV Systems Limitations?
3. How can we improve conventional metrics to evaluate crash scenarios?

Methodology

- Crash videos to evaluate VLMs' ability to handle challenging driving scenarios
  - **AVC**: cases where AVs failed to detect obstacles
  - **HDC**: cases where human drivers made poor decisions under pressure

Autonomous Vehicle Crashes (**AVC**)



Human-driven Crashes (**HDC**)



Crash Scenarios

Category	Scenario	Description	Speed (m/s)	Duration (s)	Frame Count	Distance (m)
AVC	Fail to Stop at A[illegible]ent (AVC-FTSA) 	Secondary collision after failing to stop.	25.43	6	35	152.63
	Off-Road S[illegible] (AVC-ORS) 	Vehicle halts off the road unexpectedly.	24.42	6	34	146.48
	Lane Wea[illegible] (AVC-LW) 	Vehicle drifts out of its lane.	22.76	2	9	45.74
	Fail to Stop at Int[illegible]ection (AVC-FTSI) 	Misses stopping at a busy intersection.	1.51	4	25	6.09
	Fail to Stop Sudden Brake (AVC[illegible]SSB) 	Rear-end collision after sudden brake.	20	2	12	40
HDC	Fail to Stop on [illegible]ge (HDC-FTSM) 	Collision due to improper merging.	21.91	10	53	219.78
	Fail to Yield Cross-Traffic (HDC-FYCT) 	Crash from not yielding to cross-traffic.	23.06	11	55	254.32
	Fail to Brake [illegible]me (HDC-FBT) 	Delayed braking leads to impact.	22.35	10	51	223.47
	Red Light Vio[illegible]on (HDC-RLV) 	Runs red light, causing a collision.	3.26	9	44	6.05
	Unsafe U-T[illegible] (HDC-UUT) 	Unsafe U-turn results in crash.	8.05	11	52	16.76

Table 1: Collected dataset of crash scenarios

1. Can VLMs overcome Human Driver ?

MoE-LLaVA description

Crash: … *you should slow down*



PNR: … *you should slow down as it is crucial to maintain a safe distance from the car in front of you, which is a white car.*

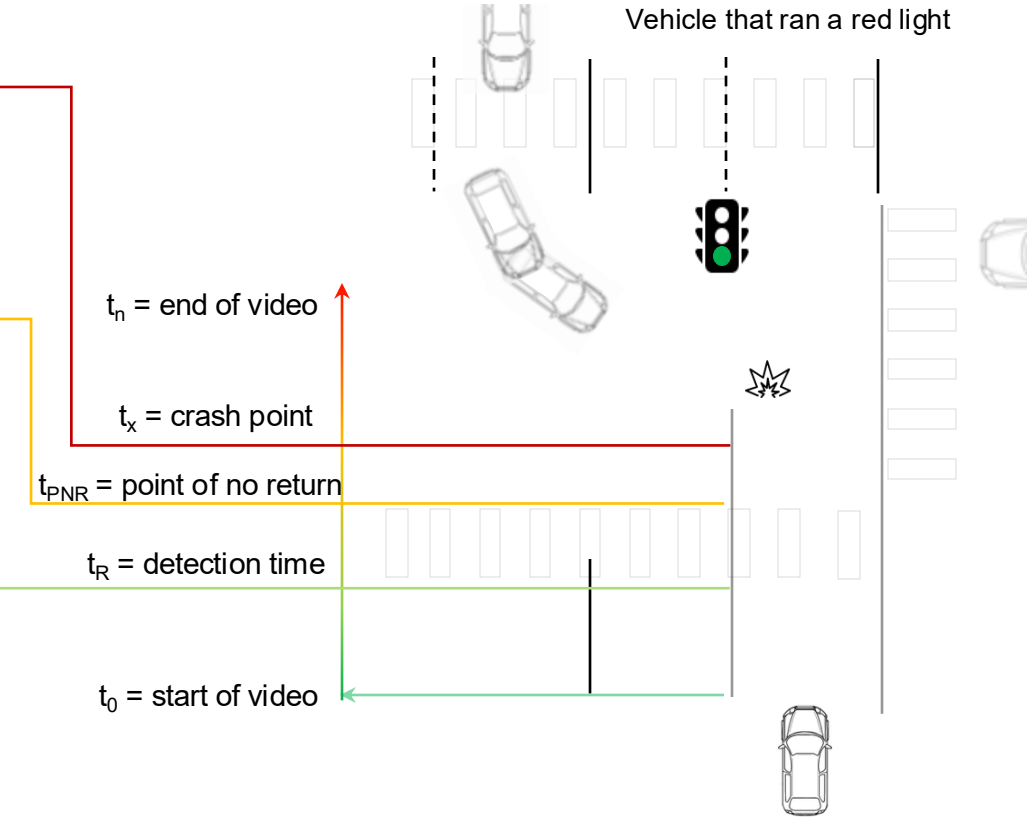


Detection time: … ***a car*** *is approaching a crosswalk …*



Start of video: …*you are driving at 8 mph and approaching a four-way intersection…*



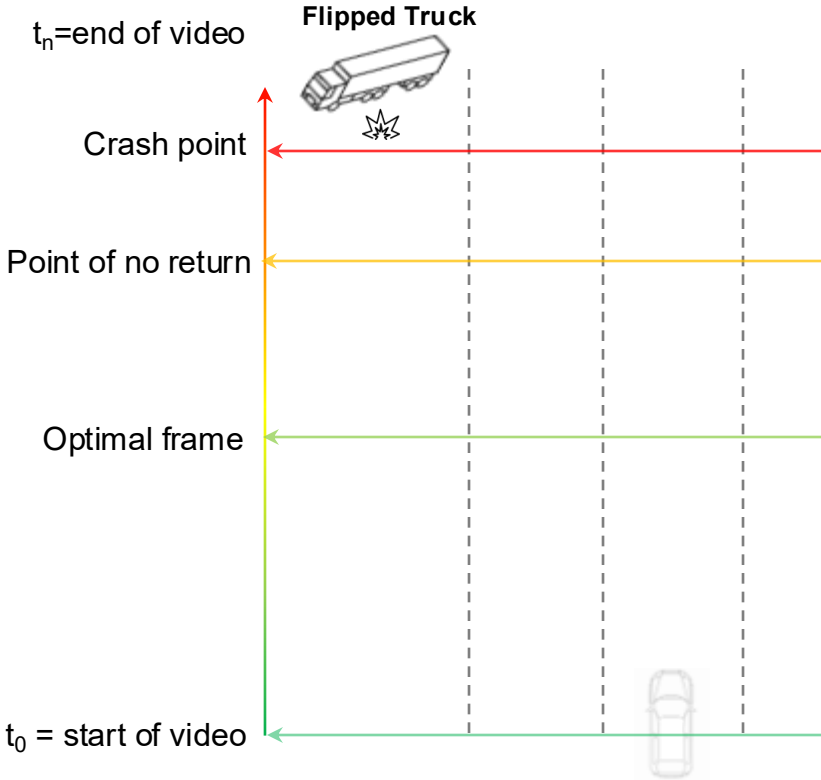


1. Can VLMs overcome Human Driver ?

Scenario	ΔPNR Time (Human - LLaVA-7B) [s]	ΔPNR Time (Human - MoE-LLaVA) [s]	ΔDetection Time (Human - LLaVA-7B) [s]	ΔDetection Time (Human - MoE-LLaVA) [s]
HDC-FTSM	1.55	1.22	1.4	2.57
HDC-FYCT	0.9	0.9	1.68	1.68
HDC-FBT	1.43	1.43	0.3	-0.2
HDC-RLV	0.44	0.44	1.2	1.37
HDC-UUT	0.08	0.08	1.08	1.25
Average	**0.88**	**0.81**	**1.13**	**1.33**

Table 2: PNR detection for HDC scenarios using LLaVA-7B and MoE-LLaVA

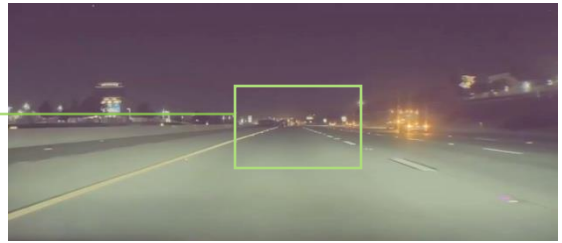
2. Can VLMs overcome DNN-Based AV Systems?



MoE-LLaVA description



Point of No Return: *…you should slow down and maintain a safe speed to avoid any potential accidents or collisions. In the image, there is a truck on the road, and it appears to be in a precarious position*



Detection*: … driving at 57 mph on a dark highway at night, it is essential to maintain a safe distance from the vehicle ahead and be aware of potential obstacles or traffic conditions…*



Start of video*: As you are driving at 57 mph on a dark highway at night, it is essential to maintain a safe distance from the vehicle ahead and be aware of potential obstacles or traffic conditions…*

2. Can VLMs overcome DNN-Based AV Systems?

Scenario	Total Frames	Optimal PNR	PNR LLaVA-7B	LLaVA-7B Description	PNR MoE-LLaVA	MoE-LLaVA Description
AVC-FTSA	35	25	22	there is a truck in front of ...you should slow down	21	...there is a truck on the road in a precarious position...
AVC-ORS	34	13	12	...you should come to a complete stop	12	...you need to come to a complete stop
AVC-LV	9	5	5	...there is a car in front of you	5	car ahead of you, and you should react
AVC-FTSI	25	10	6	...you should come to a complete stop... slow down and exercise caution...	4	...slow down and come to a complete stop at the crosswalk
AVC-FTSSB	12	6	7	slow down and exercise caution…	9	...maintain a safe distance from the truck ahead

Table 4: PNR detection for AVC scenarios using LLaVA-7B and MoE-LLaVA.

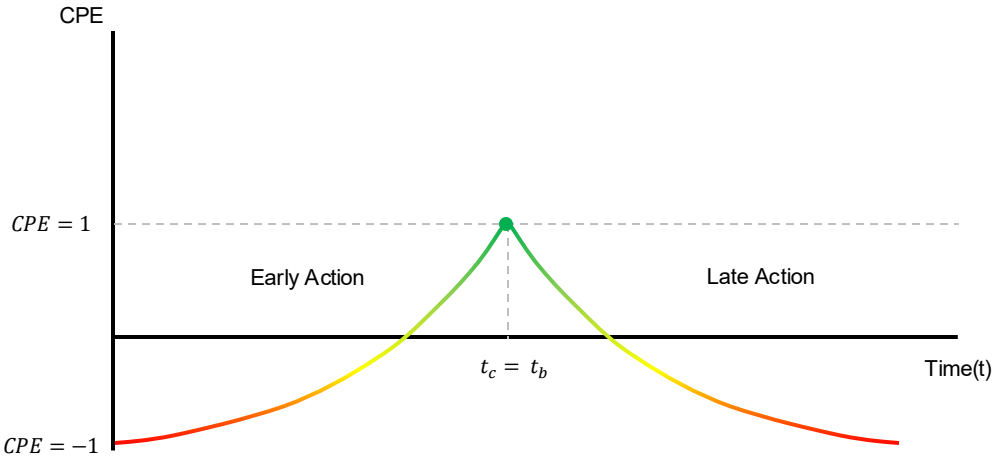
3. Crash Prevention Efficiency (CPE)

- Measures how well a VLM can respond to a potential crash
  - Assess both timing of the detection
  - Proximity to a predefined PNR.

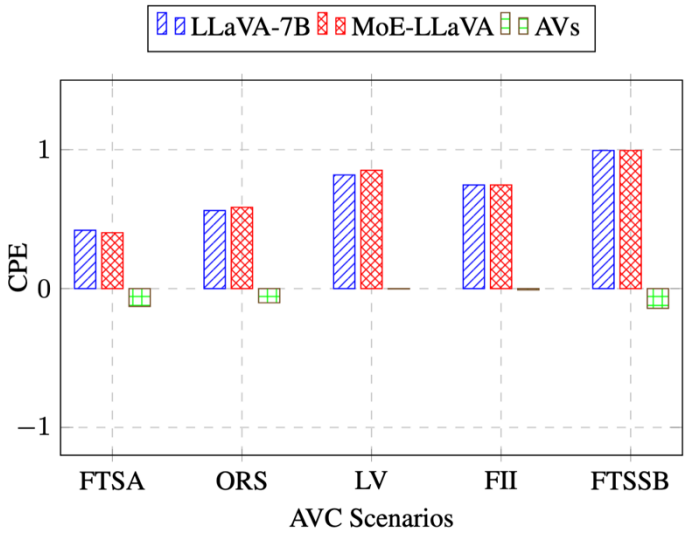
- PNR: latest moment when action must be taken to avoid crash
- $t_b = \frac{d}{v}$ maximum allowable time to take action before vehicle reaches crash point
- $t_d =$ time at which VLM detects crash
- $t_r =$ time required to react after detection
- $t_c = t_d + t_r =$ time for detection and reaction, called critical time

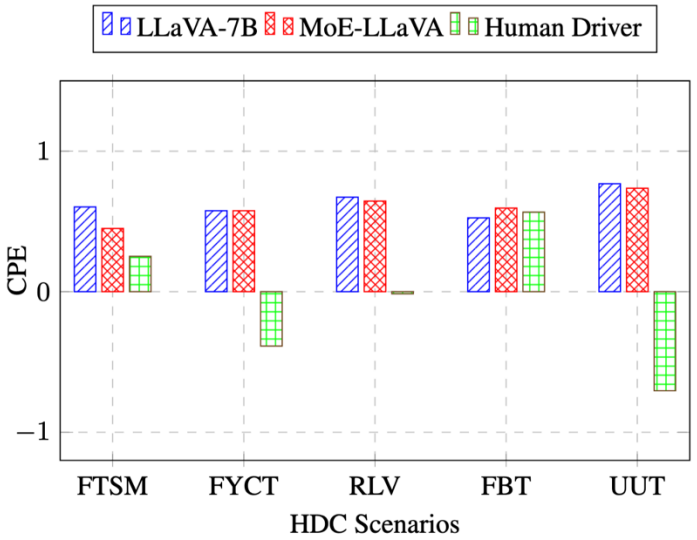
3. Crash Prevention Efficiency (CPE)

$$CPE = \begin{cases} e^{-\alpha(t_b - t_c)}, & \text{if } t_c < t_b \text{ (Early Action)} \\ 1, & \text{if } t_c = t_b \text{ (On-Time Action)} \\ e^{-\beta(t_c - t_b)}, & \text{if } t_c > t_b \text{ (Late Action)} \end{cases}$$



CPE Results


LLaVA-7B
MoE-LLaVA
AVs
CPE
1
0
−1
FTSA
ORS
LV
FII
FTSSB
AVC Scenarios


LLaVA-7B
MoE-LLaVA
Human Driver
CPE
1
0
−1
FTSM
FYCT
RLV
FBT
UUT
HDC Scenarios

Limitations

- Latency
- Computational demands of VLMs present a limitation

Conclusions

- VLMs show better performance by making more timely decisions that could have prevented accidents.
- Results indicate that VLMs have potential to surpass existing ADS in accident prevention.
- Our CPE metric effectively measures how AV systems detect crash scenarios

Contact Info

- Thank you

- David Fernandez*
  dferna3@clemson.edu
- Pedram MohajerAnsari*
  pmohaje@clemson.edu
- Amir Salarpour
  asalarp@clemson.edu
- Mert D. Pesé
  mpese@clemson.edu