

# Vision-Language Models and Autonomous Driving

## 视觉语言大模型与自动驾驶

Hang Zhao  
赵行

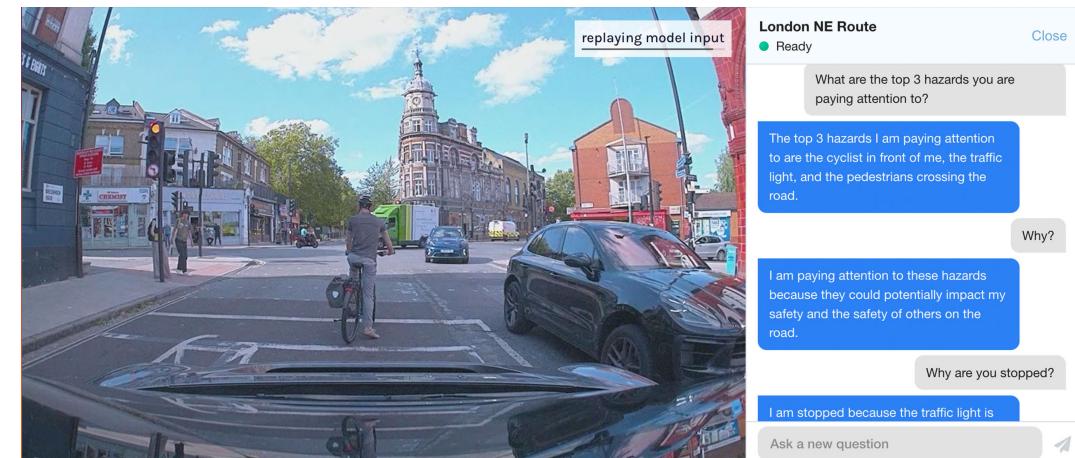
清华大学交叉信息院  
助理教授

2024.06.08

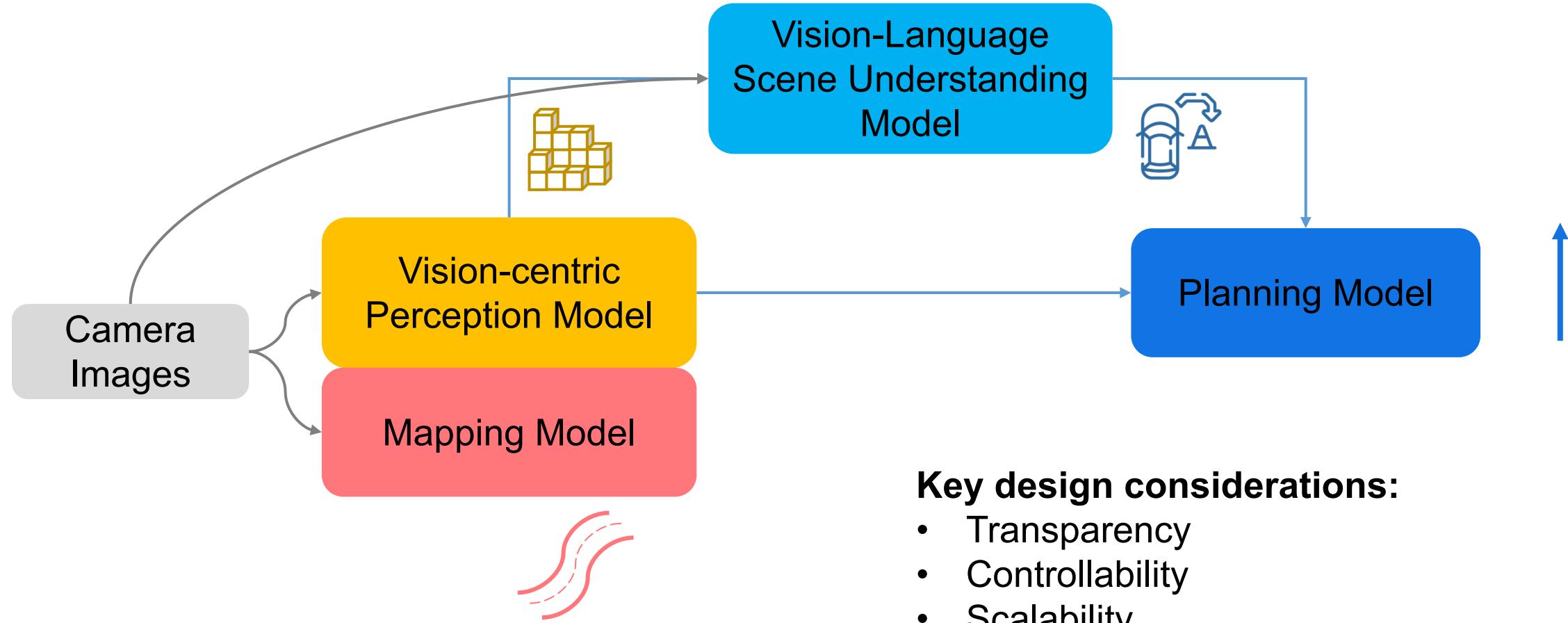


# The future is Vision(-Language)-Centric

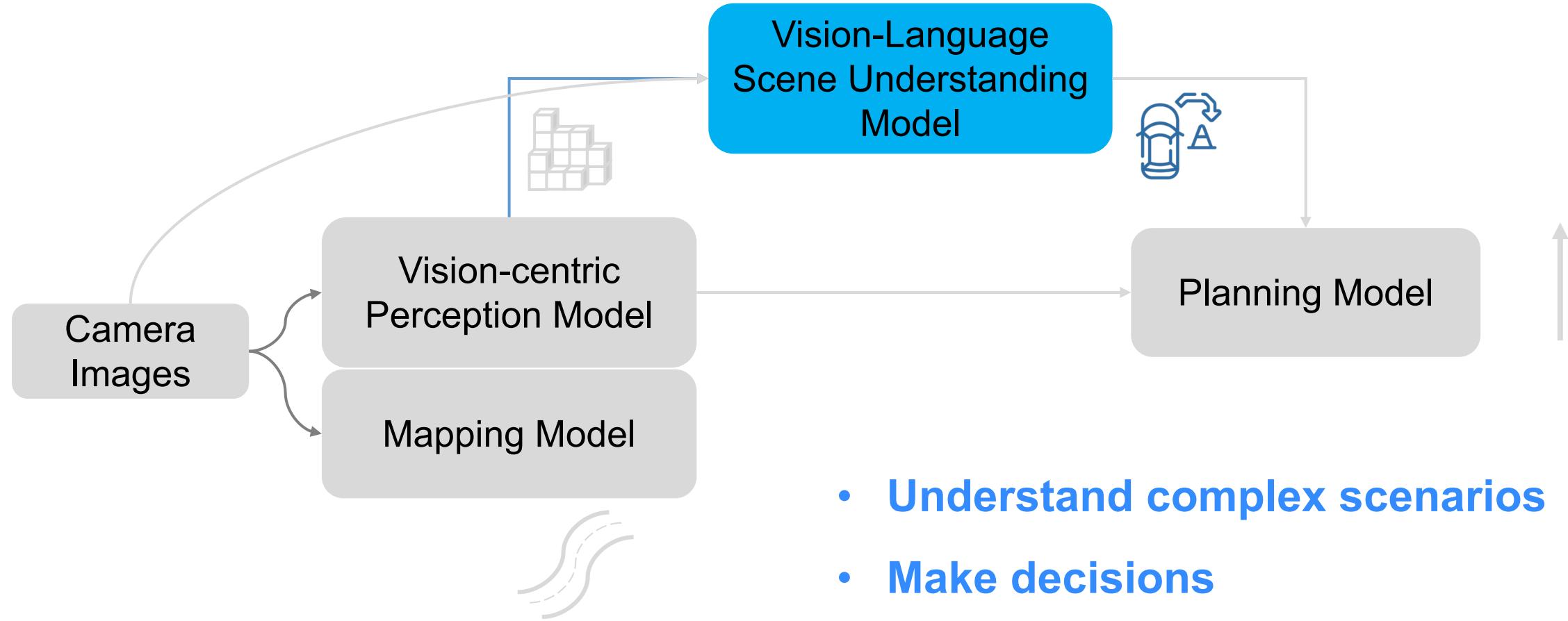
- Vision offers **dense observations**, with **rich attributes**
- Language enables **holistic scene understanding and reasoning**
- \* LiDAR is still required for the sake of safety



# Future Self-driving Tech Stack – My Guess



# Future Self-driving Tech Stack – My Guess



# AGI should come before self-driving?



# Vision-Language Models

What is unusual about this image?

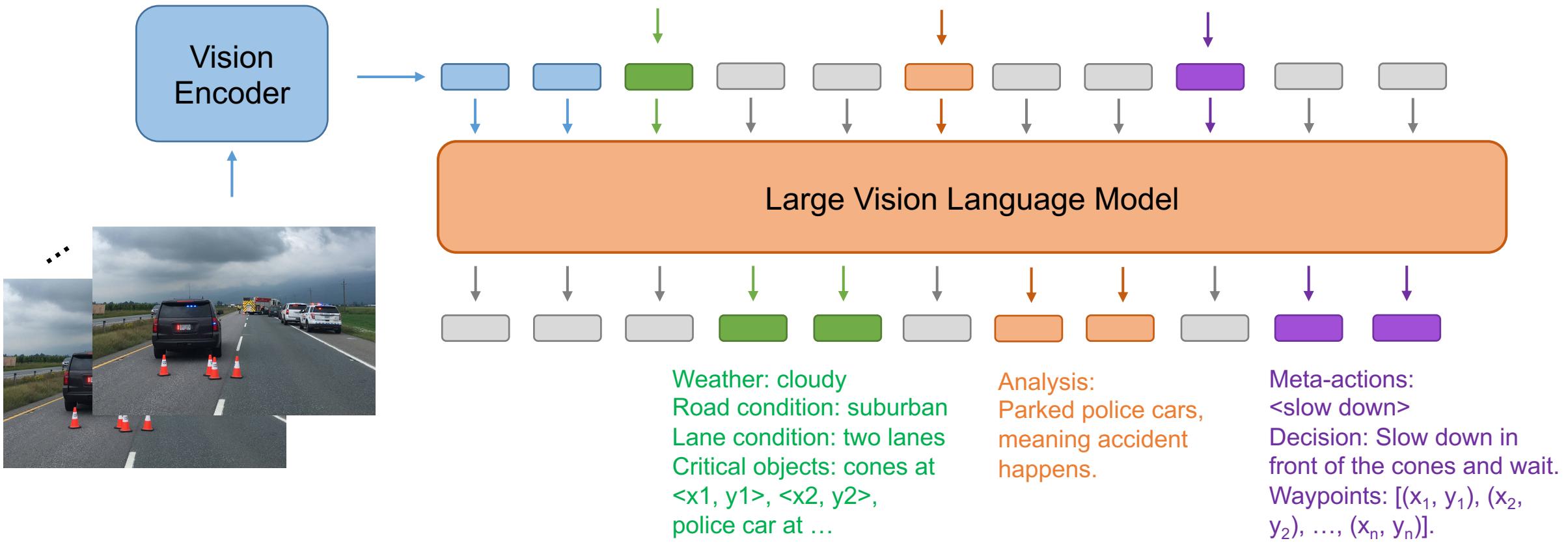


Source: [Barnorama](#)

The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

GPT4-V Technical Report

# DriveVLM



DriveVLM, Xiaoyu Tian, Junru Gu, et al. arXiv:2402.12289

# DriveVLM - Limitations

- VLMs have poor 3D spatial grounding ability.
  - Perception: false positives/negatives, inaccurate distance
  - Planning: jaggy trajectories, prone to collision
- Hard to deploy on the car for real-time ( $>10\text{Hz}$ ) inference.

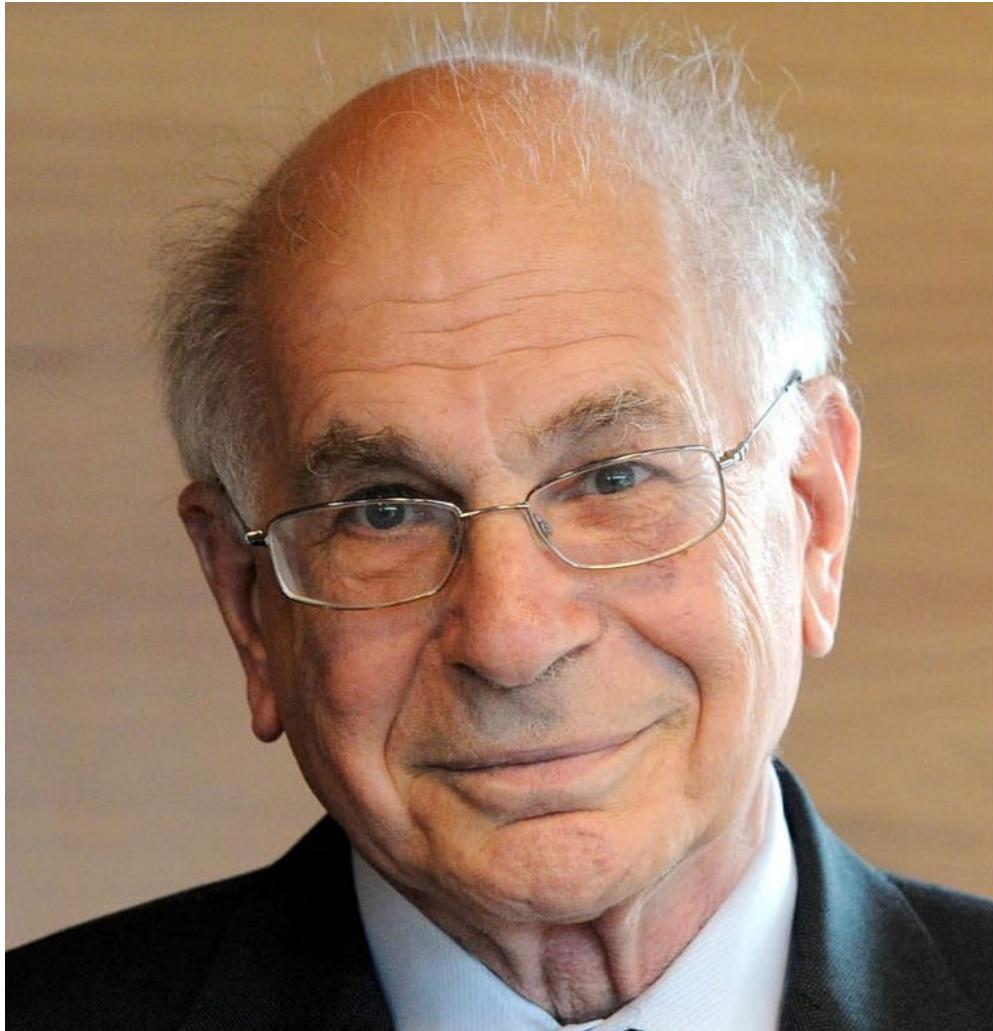
# Dual System for Autonomous Driving

- 95+% of time we are driving without thinking (muscle memory)
- But reasoning is required for long-tail scenarios



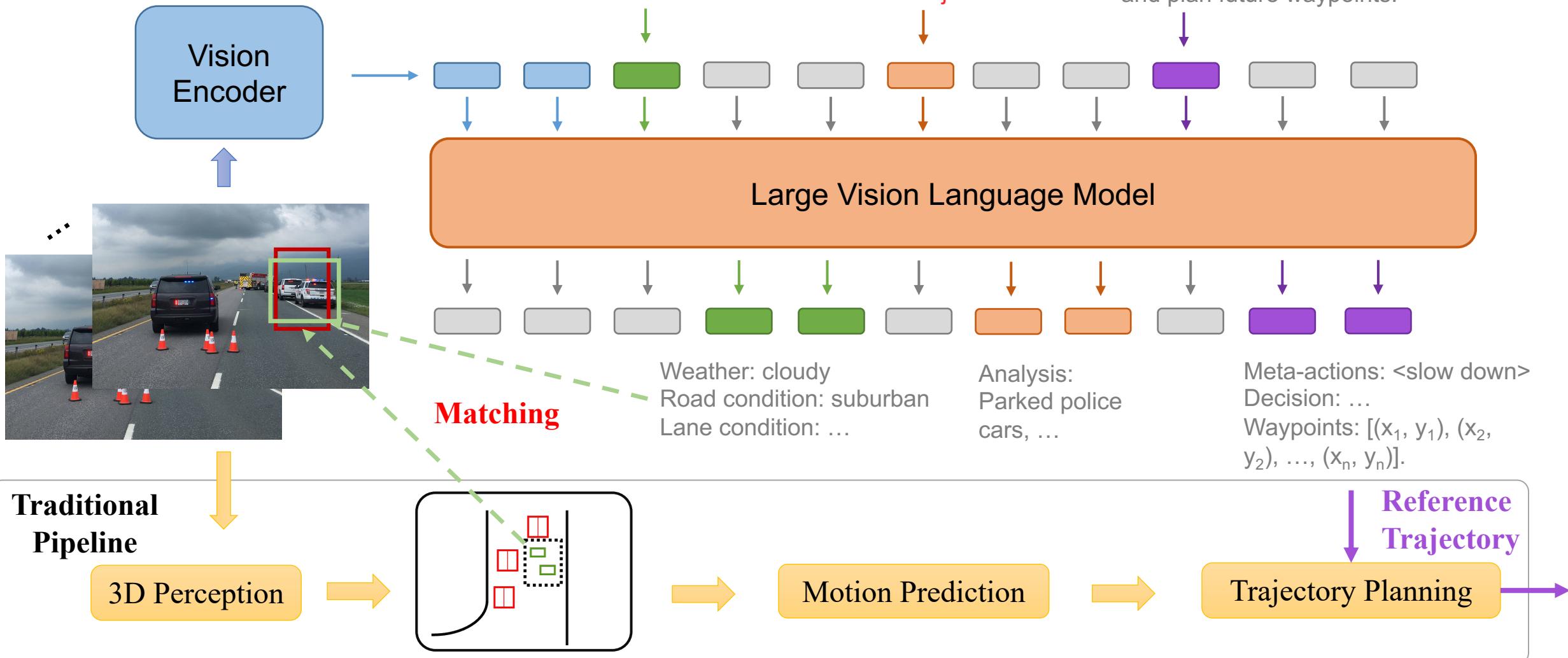
DriveVLM, Xiaoyu Tian, Junru Gu, et al. arXiv:2402.12289

# Dual System - Thinking Slow and Fast



THINKING,  
FAST AND SLOW  
DANIEL  
KAHNEMAN

# DriveVLM-Dual



# SUP-AD Dataset

## Complex and Long-tail Driving Scenarios



# SUP-AD Dataset Annotations

Scene Summary:

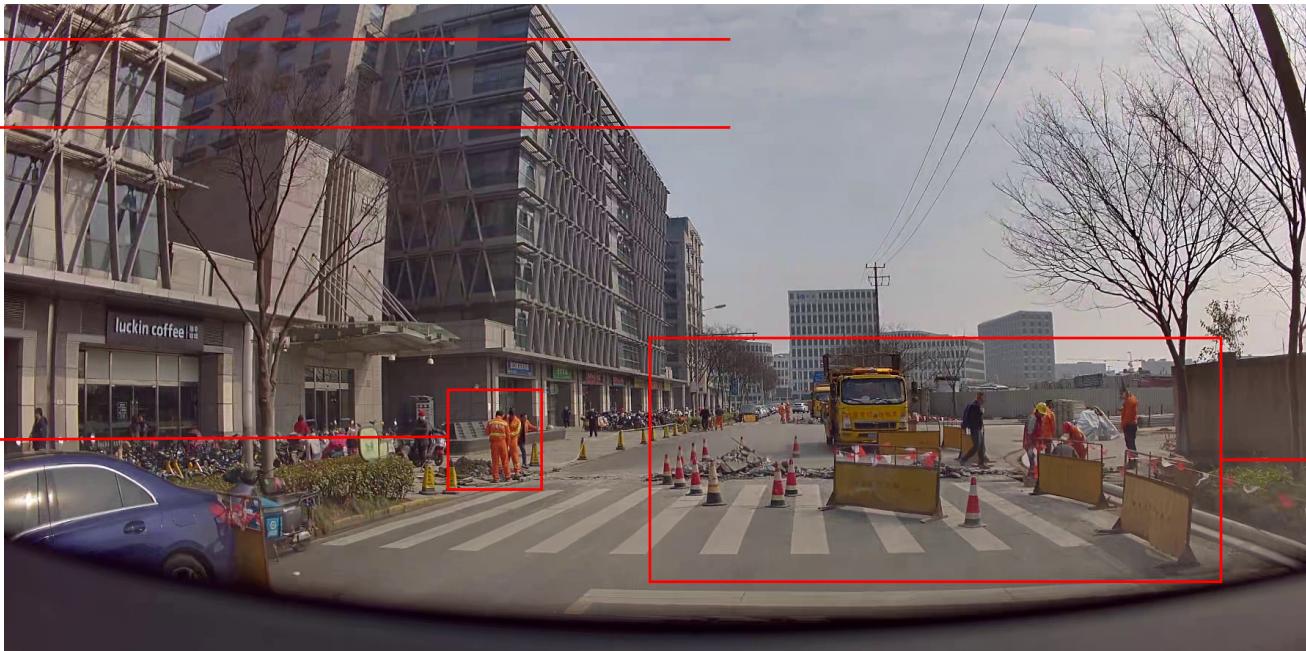
The ego vehicle is moving at a constant speed along the current lane, with ongoing road construction work ahead; there are three construction workers working on the left side of the lane at the roadside.

Weather: Sunny

Time: Daytime

Critical Object:

Class:	Three Construction Workers
Characteristics:	Construction work on the side of the lane to the left of the host vehicle
Influence:	Affects the normal speed of the host vehicle



Road Environment: Construction

Critical Object:

Class:	Construction Zone
Characteristics:	Road repair in front of the host vehicle lane
Influence:	Affects the host vehicle to drive straight normally

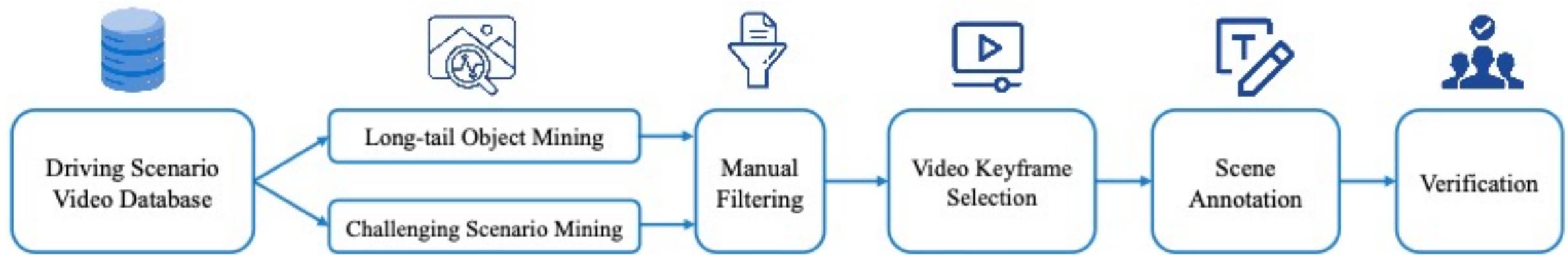
Meta Action:

["Slow down", "Change lane to the left", "Go straight slowly"]

Decision Description:

Decelerate and change lanes to the left, keeping a safe distance from the construction workers on the left front side.

# Dataset Construction Pipeline



# SUP-AD Dataset

## Complex and Long-tail Driving Scenarios



# SUP-AD Dataset Annotations

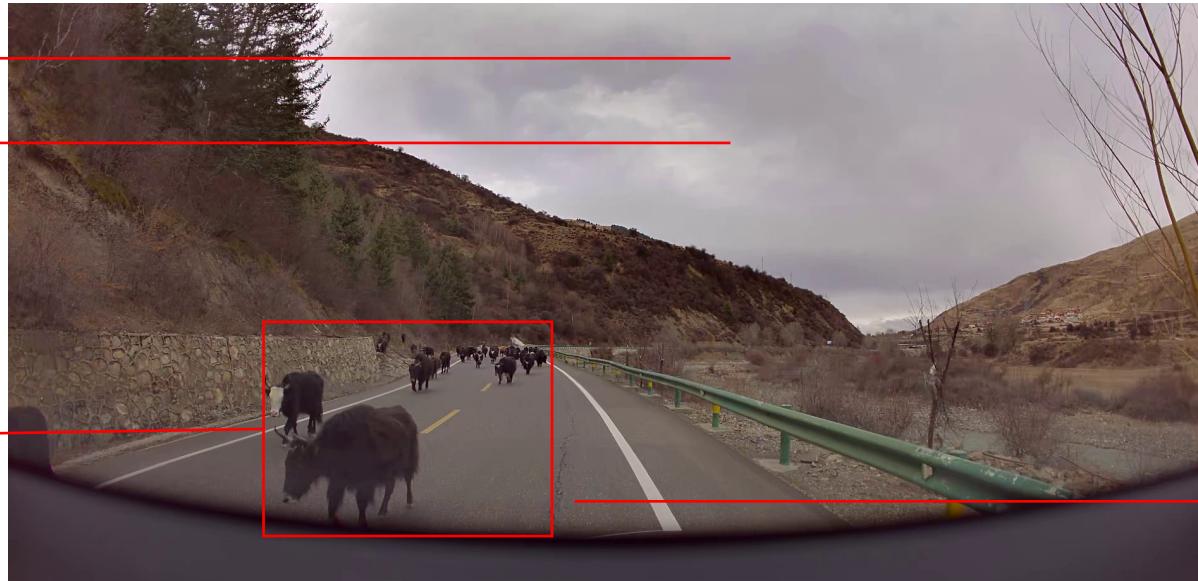
Scene Summary: The ego vehicle is traveling straight in the current lane, and there is a group of slow-moving cows ahead in the lane.

Weather: Cloudy

Time: Daytime

Critical Object:

Class:	A herd of cows
Characteristics:	Slowly moving ahead in the ego lane
Influence:	Influencing the driving speed of the ego vehicle



Lane Options: Own Lane

Road Environment: Mountain

Meta Action: ["Slow down", "Go straight slowly"]

Decision Description: Reduce speed and proceed slowly while maintaining a safe distance from the group of cows.

# SUP-AD Dataset

## Complex and Long-tail Driving Scenarios

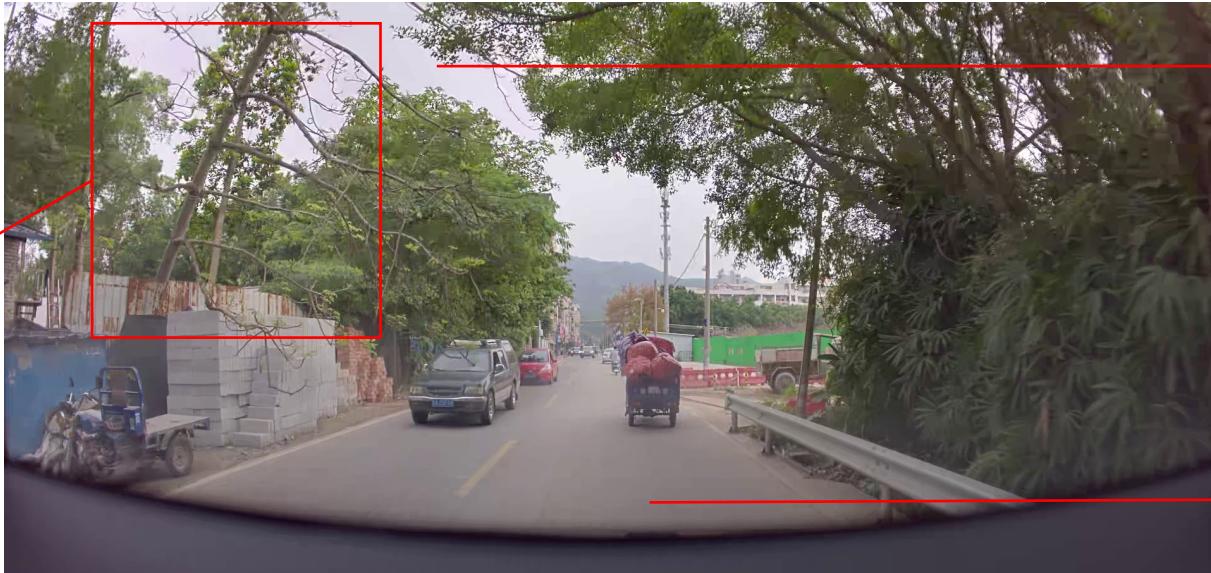


# SUP-AD Dataset Annotations

Scene Summary: The ego vehicle is moving forward on the current road, and a tree suddenly falls towards the ego vehicle from the left front side.

Critical Object:

Class:	Tree
Characteristics:	Leaning towards our vehicle on the left front
Influence:	Blocking our vehicle from moving forward



Weather: Cloudy

Time: Daytime

Lane Options: Own Lane

Road Environment: National Road

Meta Action: [“Slow down rapidly”, “Stop”, “Wait”]

Decision Description: Immediately decelerate and come to a stop, wait for the fallen tree to be cleared before resuming driving.



These are a few frames taken at regular intervals from a driving video:. Suppose you are driving in this scenario, please describe the scene condition.



Scene Condition:

Weather:Sunny.

Time:Day.

Road Environment:Urban.

Lane Options:Own Lane, Left Lane.

Ego Lane Position:Rightmost Lane.



Please list the critical objects in the front view that affect ego vehicle's next driving decisions, and box them step by step.



There are 1 critical object in the front view. The objects include ['Fallen tree'].

Object 1:<ref>Fallen tree</ref><box>(353,413),(483,566)</box>.

Unmatching 2D Objects: - Fallen tree at img region [353, 413, 483, 566].

Describe the characteristics of the object and indicate its influence based on the matching information





Characteristics: In the lane on the left front side of the ego vehicle. | Influence: Blocking the future path of the ego vehicle.



Ego-States:

Historical Trajectory (last 2 seconds):[(-6.21,-0.05),(-12.39,-0.20),(-18.43,-0.43),(-24.44,-0.67)]

Navigation: MAINTAIN.

Please make a summary of the current scene based on the information provided and your previous analysis.



The ego vehicle is driving straight ahead at a constant speed in the current lane, with a fallen tree blocking the left lane ahead



Based on the above analysis and the current scenario, please determine the driving meta actions and decisions you will take next.



Meta Actions: ['Slow down', 'Shift slightly to the right', 'Go straight slowly'].

Decisions: Slow down, shift to the right to avoid the fallen tree, and continue driving.



Based on the above decisions and analysis, please plan the 3s future trajectory of the ego vehicle.



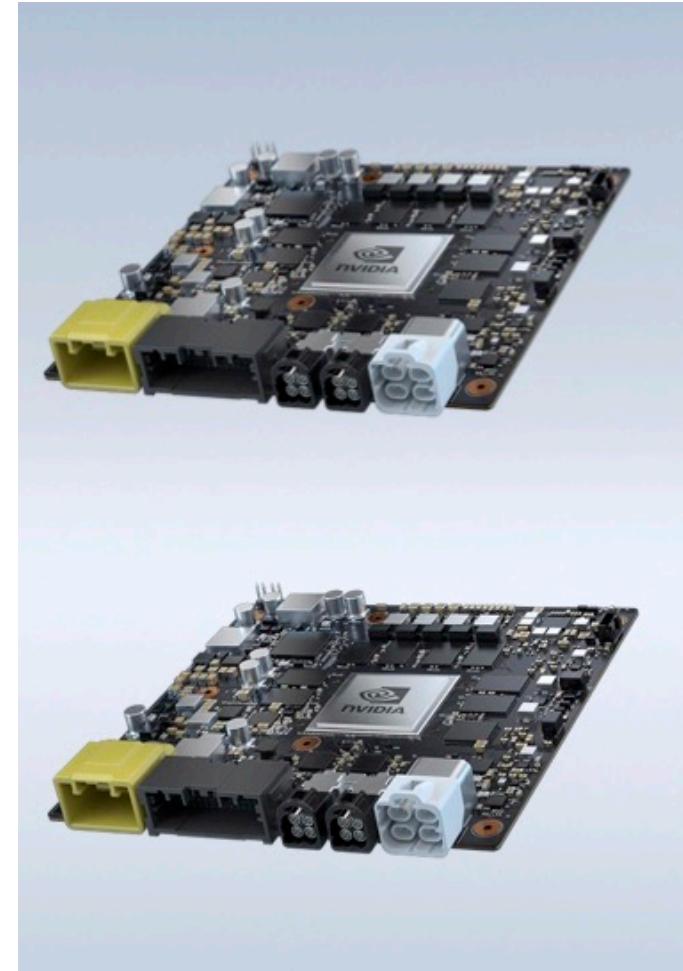
[(6.1,-0.0),(12.1,-0.1),(18.1,-0.2),(24.1,-0.3),(30.1,-0.4),(36.1,-0.5)]



# DriveVLM-Dual Onboard Deployment

- Two Orin X's
- Orin X-1: Fast System
- Orin X-2: Slow System DriveVLM
- Speculative sampling speed 1-2 Hz

Credit to Li Auto team





# Acknowledgement



Xiaoyu Tian  
Tsinghua University



Junru Gu  
Tsinghua University

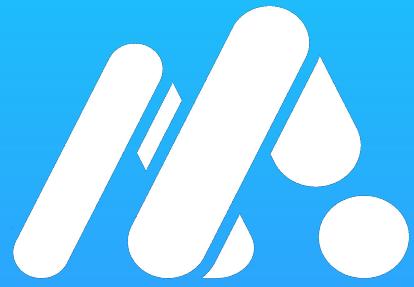


Yicheng Liu  
Tsinghua University



清华大学  
Tsinghua University





MARS Lab  
THE END THANKS