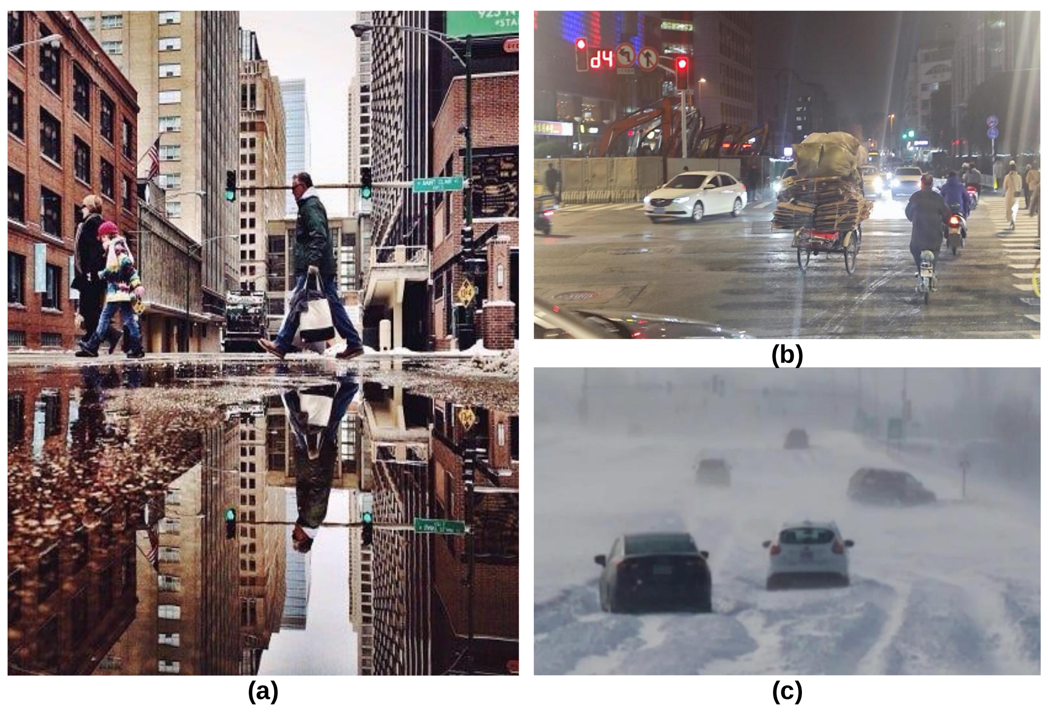
An Advanced Driving Agent with the Multimodal Large Language Model for Autonomous Vehicles

Junzhou Chen, Sidi Lu

Introduction

Limitation of Autonomous Driving System

- Autonomous driving system (ADS): advancing **rapidly**
- **Deep learning-based** ADS heavily rely on training data
- **Corner cases:** ADS often fail to respond appropriately

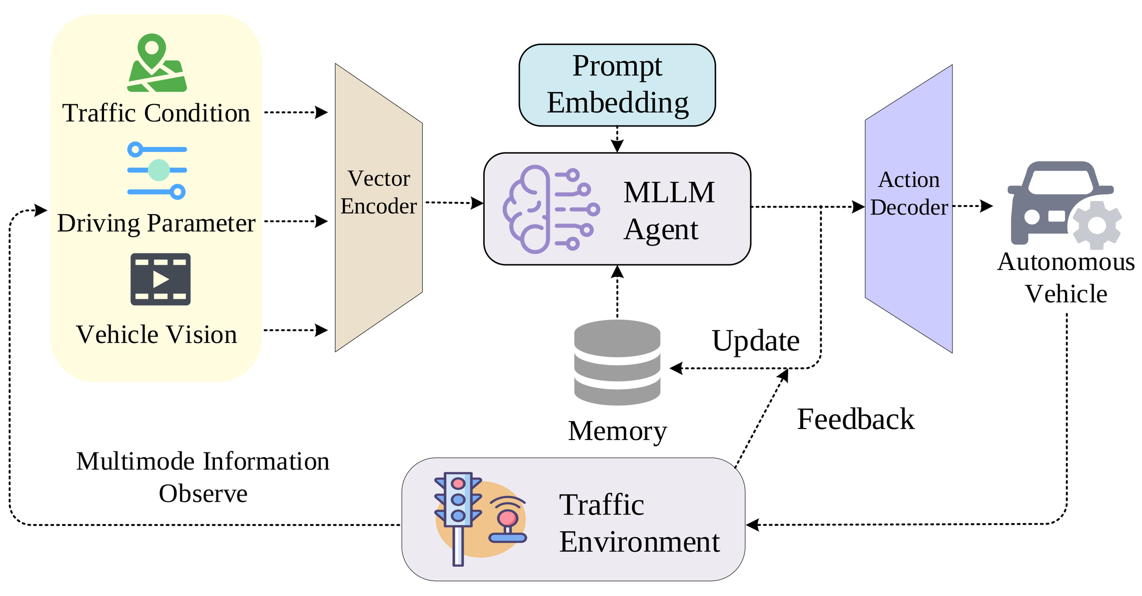


Potential Solutions & Challenges

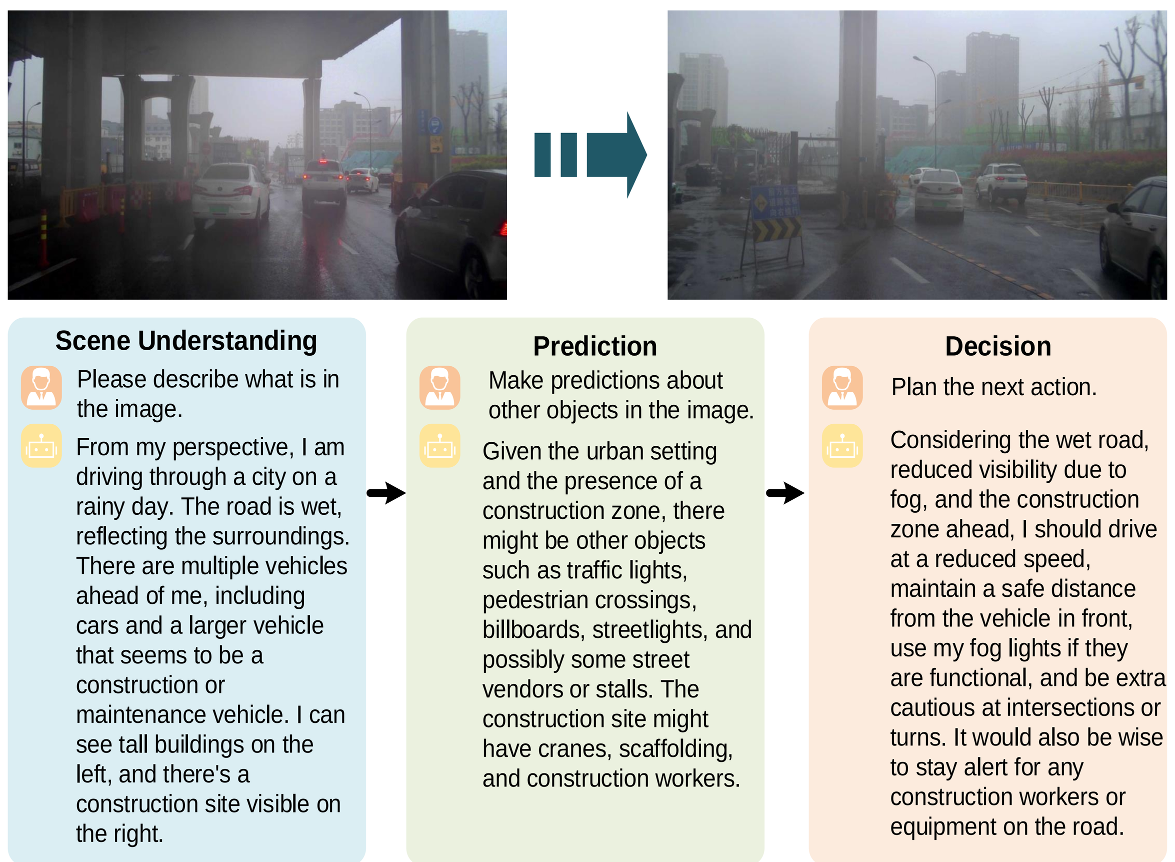
- Addressing the **safety** of ADS and enhancing **reliability**
- Multimodal Large Language Model (MLLMs): enhance **understanding**, **prediction**, and **decision** capacity
- MLLMs: **illusion** problem
- Insufficient multimodal information **fusion**

Experiment Methodology

An end-to-end autonomous driving framework with a Multimodal Large Language Model



Thinking Chain Design



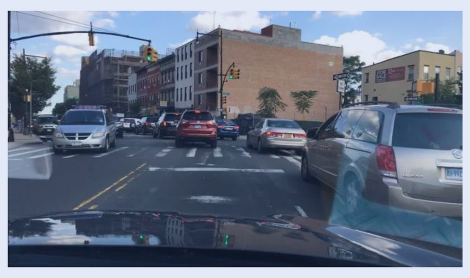
- **Motivation**:
  - **Improved Accuracy**: Enhances decision-making by focusing on interpretability
  - **Reduces Model Illusions**: Stepwise reasoning reduces errors in complex scenarios
- **Rationale and Exemplars:** Embed reasoning steps to improve **logical consistency**
- **Step-by-Step** Process: Break down into three phases
  ➡ **Scene Understanding, Prediction, and Decision**

Visual-Question-Answer Dataset Construction

- Hundreds of images from **BDD100k**
- Label minimal data for GPT-4 **in-context learning**
- ChatGPT-4: automatic annotation & human re-annotation
- Fine-tuning **CogVLM** on VQA dataset ➡minimize model hallucinations

Embed **Chain-of-Thought** into the dataset

 # Vision-Question-Answer dataset case



User prompt: You are now an AI agent driver. While driving on the streets, please answer the following questions from a driver's point of view. Describe what is in the image.

Q1: Scene Understanding Q&A

Assistant response: The road, a four-lane, two-way street with yellow double solid lines and white markings, hosts multiple vehicles, indicating rush hour. A silver sedan, possibly stopped due to a signal or congestion, is closest. Traffic lights display red and green, suggesting varying signals for different lanes. The sky is clear with scattered clouds.

User prompt: Make predictions about the behavior of objects in the image.

Assistant response: Vehicles will resume movement when traffic lights turn green, potentially in waves based on traffic volume. Traffic lights control the flow through different phases. Pedestrians and cyclists, if present, follow their signals. Street scene and weather, clear and sunny, do not impact traffic conditions.

Q2: Prediction Q&A

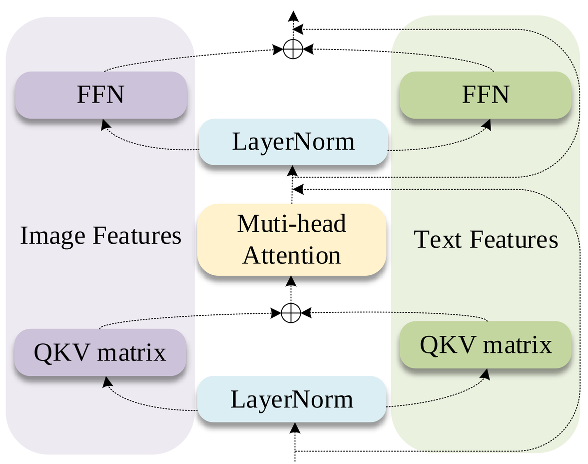
User prompt: Please make decisions about your subsequent driving behavior. Requires you to format the content as *<Example>*:

Assistant response: ACTIONS = {
Speed: 'Maintain speed.',
Direction: 'Continue straight.',
Brake: 'Gently brake if necessary due to traffic or red lights.',
Signaling: 'No signal', ··· }

Q3: Decision Q&A

MLLM Fine-tuning and Deployment

- **CogVLM**: offline MLLM
- Image features have different Query-Key-Value (**QKV**) matrices and Feed-Forward Networks (**FFN**)
  ➡CogVLM: better **visual understanding**



- **Modular framework** for scalability
- **Deployment setting**:
  - 8-bit Quantization: Reduces computation demands
  - Fine-tuned on 4×A100 GPUs
  - Inference: two GeForce RTX 3090 GPUs

Experiment Result

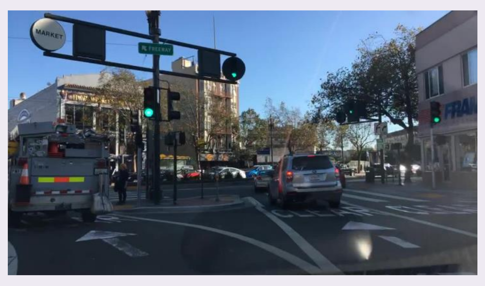
Scene Understanding

Prompt:
This is a picture taken with a car camera. Describe what you see, especially traffic participants, road topology, and traffic lights and traffic conditions (if applicable).

CogVLM:
The image appears to be taken from the perspective of a car's rearview camera, capturing a view of a city street. On the left side of the road, there's a man riding a small truck or cart with goods. There are multiple cars parked on both sides of the road, and a few are moving. Traffic lights can be seen in the distance, and the road seems to be a two-way street with clear markings. The road is flanked by buildings, trees, and some street furniture.

Prediction

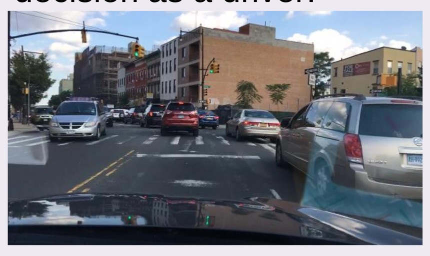
Prompt:
Now that you are a driver, predict the movement of people and vehicles in the picture.



CogVLM:
From the picture, we can see that the traffic lights in the middle of the intersection are currently green, indicating that vehicles can pass through the intersection. In the picture, there are two vehicles waiting at the red light, while several other vehicles are passing through the intersection. On the right side of the picture, there are pedestrians walking on the pedestrian crossing, waiting to cross the road. Therefore, it can be predicted that after the vehicles in the picture pass through the intersection, there will be pedestrians crossing the road, and then vehicles waiting at the red light. After these vehicles pass through, there will be more vehicles waiting to enter the intersection.

Decision

Prompt:
Please give a firm decision as a driver.

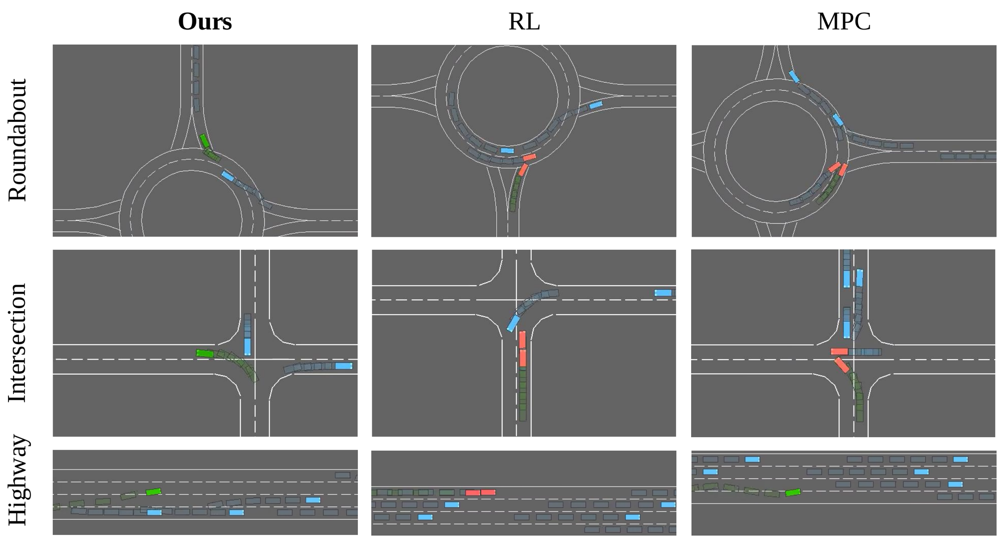


CogVLM uses CoT :
The vehicle is in the oncoming lane and the light is green. Traffic should go straight.

CogVLM without CoT:
The traffic lights are red, and you need to stop.

Simulation Results

- Compared with **Reinforcement Learning-based** (RL) **and Model Predictive Control** (MPC) in the **Highway-env Simulator**



Scenario	Method	Fail	Inefficiency	Average Time
Intersection	RL	10.0%	6.0%	3.9s
	MPC	4.0%	6.0%	4.0s
	Ours	**0.0%**	**2.0%**	**3.8s**
Roundabout	RL	10.0%	8.0%	**4.8s**
	MPC	6.0%	4.0%	5.2s
	Ours	**0.0%**	**2.0%**	5.1s
Highway	RL	12.0%	8.0%	**18.2s**
	MPC	6.0%	4.0%	19.1s
	Ours	**0.0%**	**2.0%**	22.3s

Department of Computer Science

Email: jchen57@wm.edu, sidi@wm.edu

Homepage Lab link Paper link