A Language Agent for Autonomous Driving

Authors：Jiageng Mao,Junjie Ye, Yuxi Qian, Marco Pavone，Yue Wang
University of Southern California,
Stanford University,
NVIDIA

汇报人：何杭帅
2024年11月10日

提纲

- **研究背景**
- 系统设计
- 实验评估
- 工作总结

研究背景

Imaging such a picture：

- You are driving your car on a corner road
- **Suddenly**, a ball bounces onto the road in front of you

What would be your first thought?

Human driver Self-driving car

研究背景

Autonomous Driving

Ultimate goal: human-level driving

Conventional approaches :

(a) Conventional Perception-Prediction-Planning Pipeline.

研究背景

Autonomous Driving

Ultimate goal: human-level driving

Conventional approaches :

(a) Conventional Perception-Prediction-Planning Pipeline.

Perception :
Interpret the human perceptual process as object detection or occupancy estimation.

Related work:

- Autonomous Driving: A Comprehensive Survey. IJCV, 2023b.
- Convolutional Occupancy Networks. ECCV, 2020. ‘
- DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries. CoRL, 2022.

研究背景

Autonomous Driving

Ultimate goal: human-level driving

Conventional approaches :

(a) Conventional Perception-Prediction-Planning Pipeline.

Prediction :
Abstract human drivers' foresight of upcoming scenarios as the prediction of future object motions.

Related work:

- IntentNet: Learning to Predict Intention from Raw Sensor Data. CoRL, 2018.
- Perceive, Predict, and Plan: Safe Motion Planning Through Interpretable Semantic Representations. ECCV, 2020.
- Parting with Misconceptions about Learning-based Vehicle Motion Planning. CoRL, 2023.

研究背景

Autonomous Driving

Ultimate goal:
human-level driving

Conventional approaches :

(a) Conventional Perception-Prediction-Planning Pipeline.

Planning :
Emulate the human decision-making process by planning a collision-free trajectory, either using hand-crafted rules.

Related work:

- Congested Traffic States in Empirical Observations and Microscopic Simulations. Physical Review E.
- End-To-End Interpretable Neural Motion Planner. CVPR, 2019.
- Planning-oriented autonomous driving. CVPR, 2023.

研究背景

Perception-prediction-planning framework

· Decompose the driving process into subtasks,efficacy.

· Overly simplifies the human decision-making process and cannot fully model the complexity of driving.

Exists some problem!

Perception:notably redundant, necessitating the detection of all objects in a vast perception range.

研究背景

Perceptionprediction-planning framework

· Decompose the driving process into subtasks,efficacy.

· Overly simplifies the human decision-making process and cannot fully model the complexity of driving

Exists some problem !

Prediction: designed for collision avoidance with detected objects. Nevertheless, they lack deeper reasoning ability inherent to humans.

研究背景

Perceptionprediction-planning framework

· Decompose the driving process into subtasks,efficacy.

· Overly simplifies the human decision-making process and cannot fully model the complexity of driving

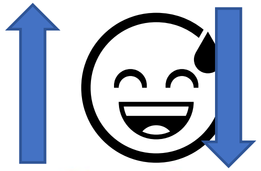
Exists some problem!

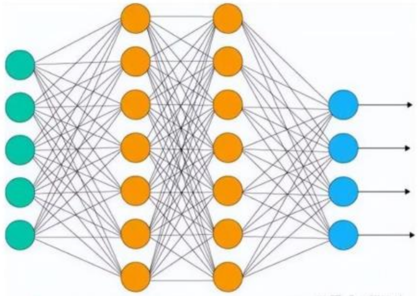
Planning: challenging to incorporate long-term driving experiences and common sense into existing autonomous driving systems.

The major obstacle in development

Form of Knowledge

Expressed in language





Numerical inputs (e.g. perceptual signals, bounding boxes and trajectories)

Differences in **language** and **numerical representation** lead to difficulties in integrating between empirical human knowledge and existing autonomous driving systems.

提纲

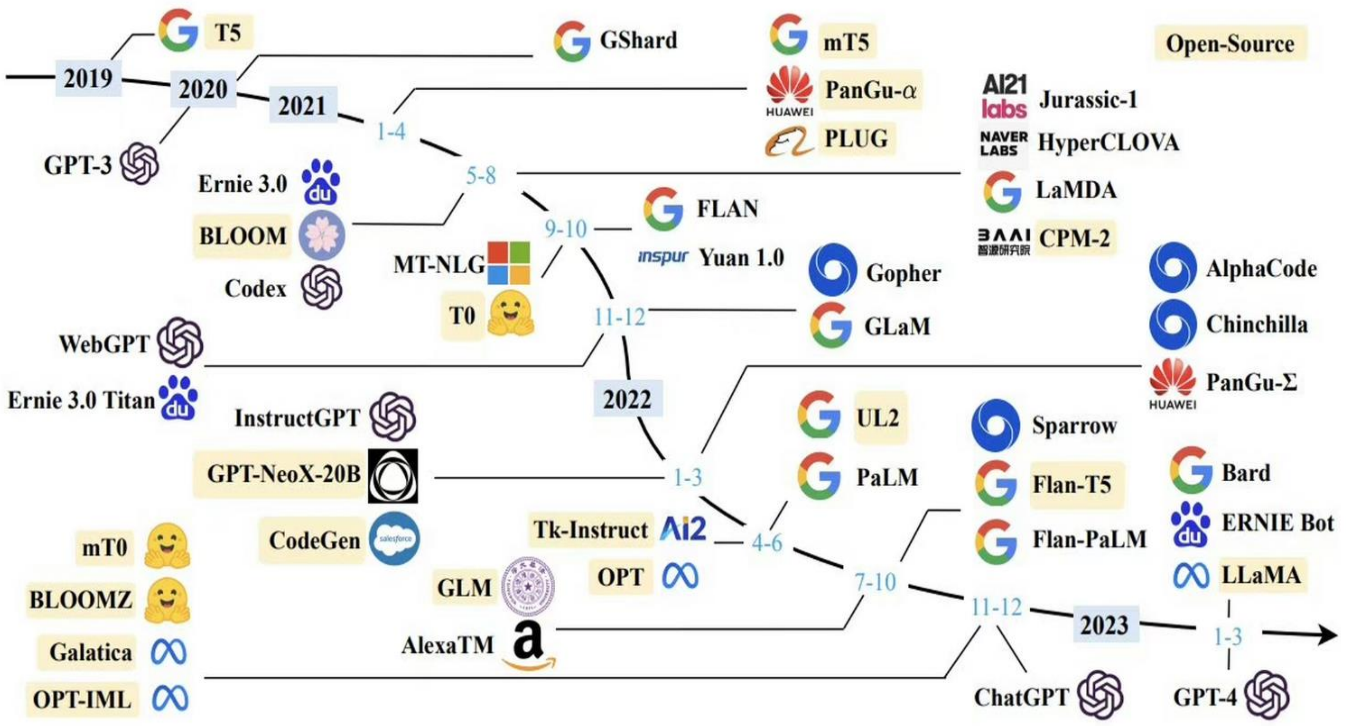
- 研究背景
- **系统设计**
- 实验评估
- 工作总结

系统设计

Agent-Driver

· A cognitive agent empowered by **Large Language Models (LLMs)**

Trained on Internet-scale data, LLMs have demonstrated remarkable capabilities in **commonsense reasoning** and **natural language understanding**

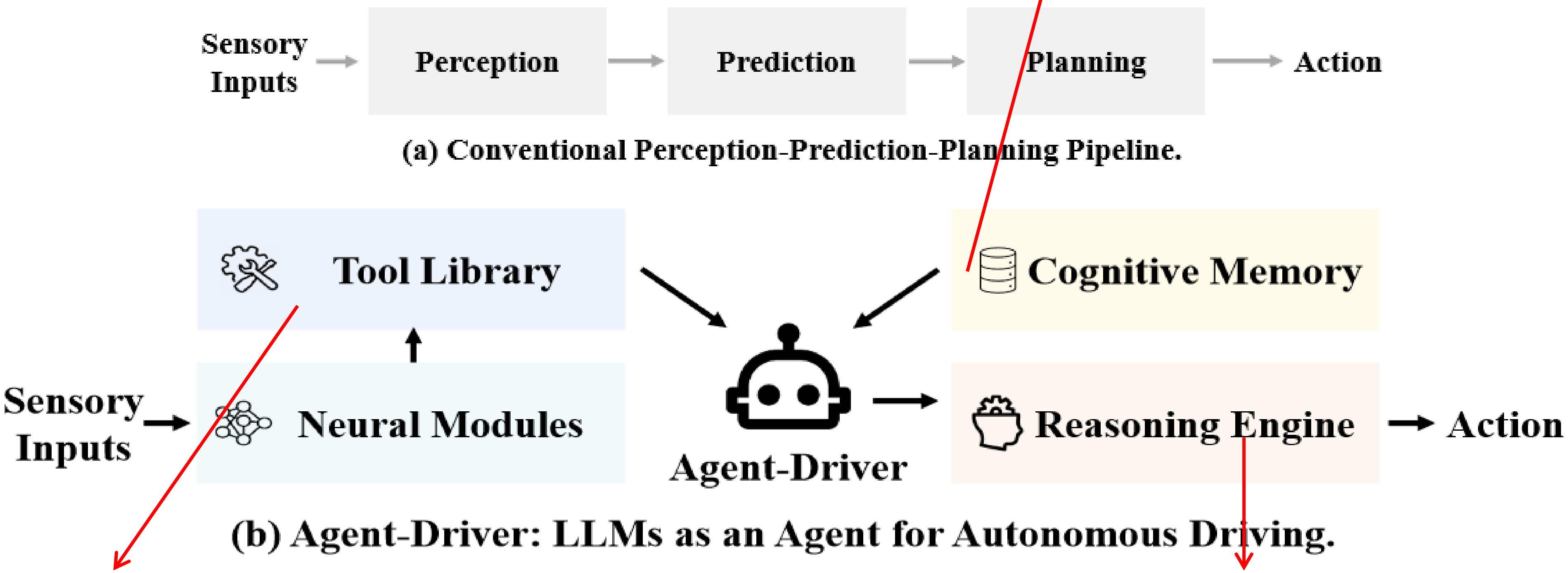


系统设计

Agent-Driver

· Overall Architecture

Explicitly stores common sense and driving experiences, infusing the system with human experiential knowledge



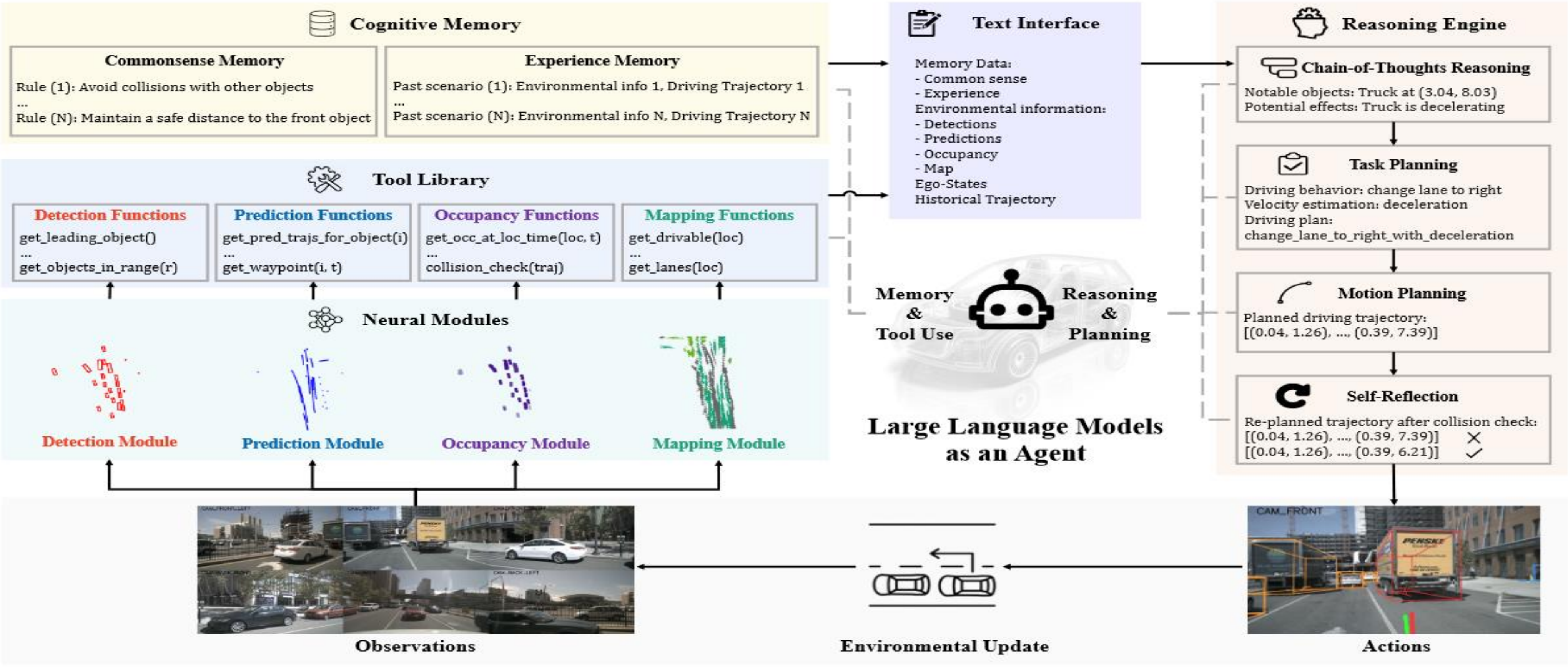
Interfaces with neural modules via dynamic function calls

Processes perception results and memory data to emulate human-like decision-making

系统设计

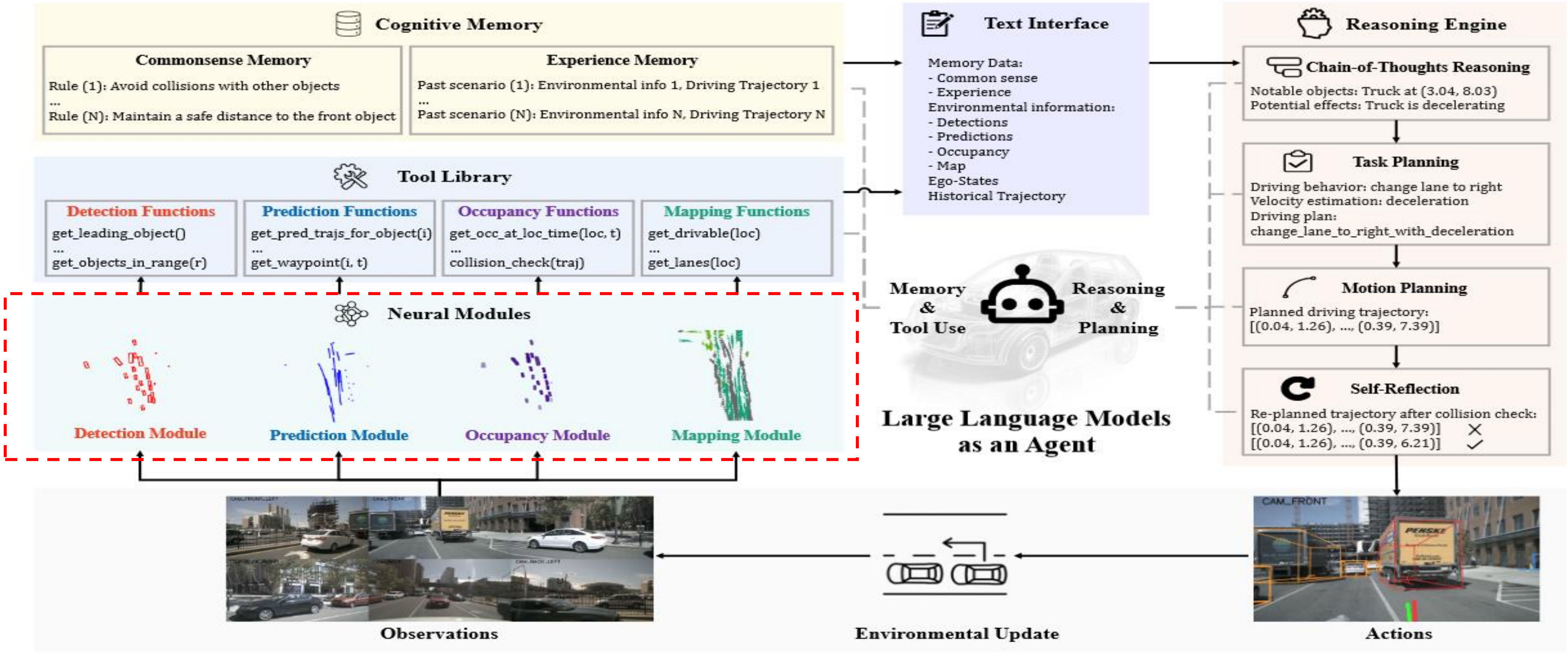
Agent-Driver

· Overall Architecture



Agent-Driver

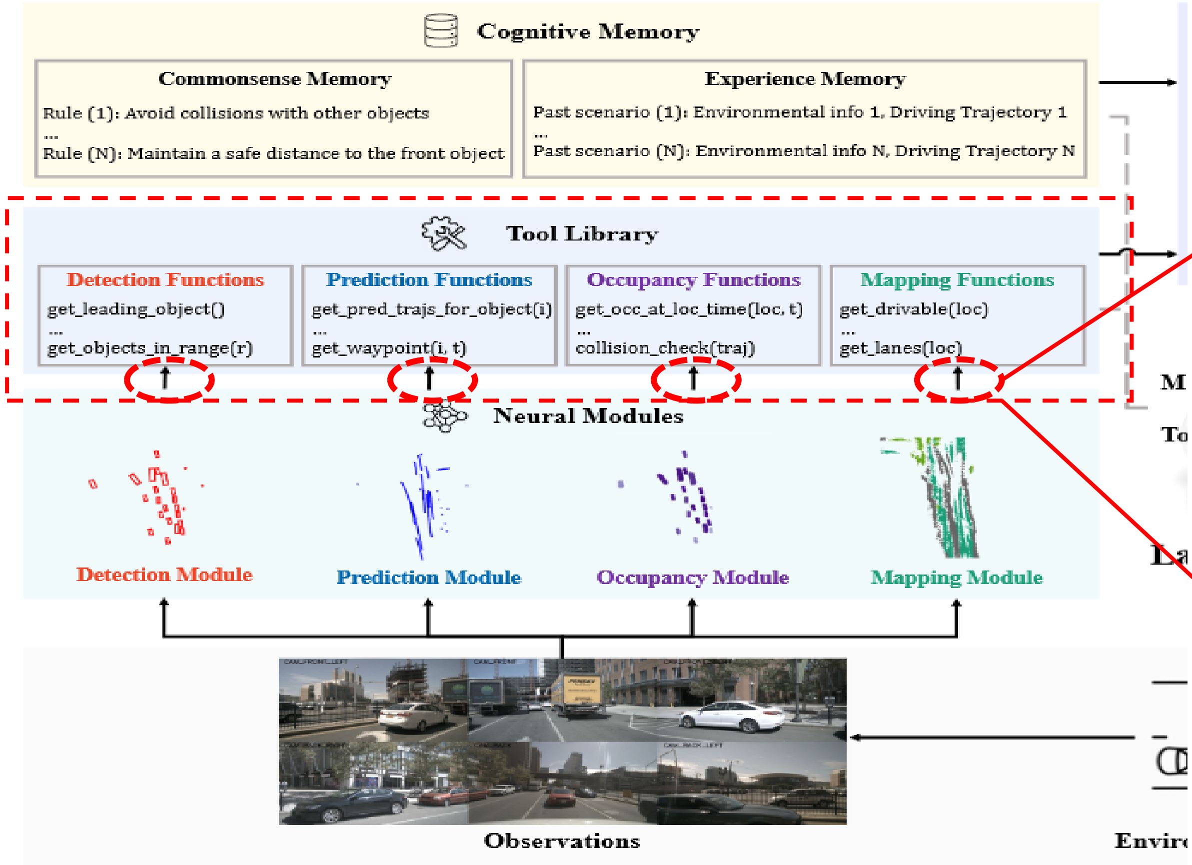
· Neural Modules

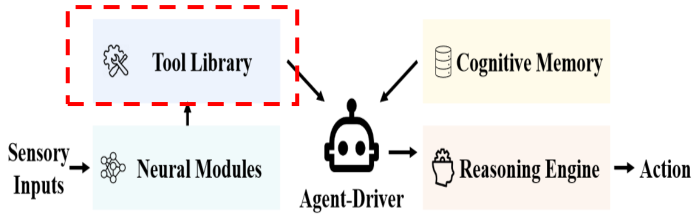


系统设计

Agent-Driver

· Tool Library





(b) Agent-Driver: LLMs as an Agent for Autonomous Driving.

Challenge:
Incorporating human knowledge into neural-network-based driving systems.

Solution:
· Leverage text as a unified interface to connect neural modules.
· dynamically collect text-based environmental information.

系统设计

Agent-Driver

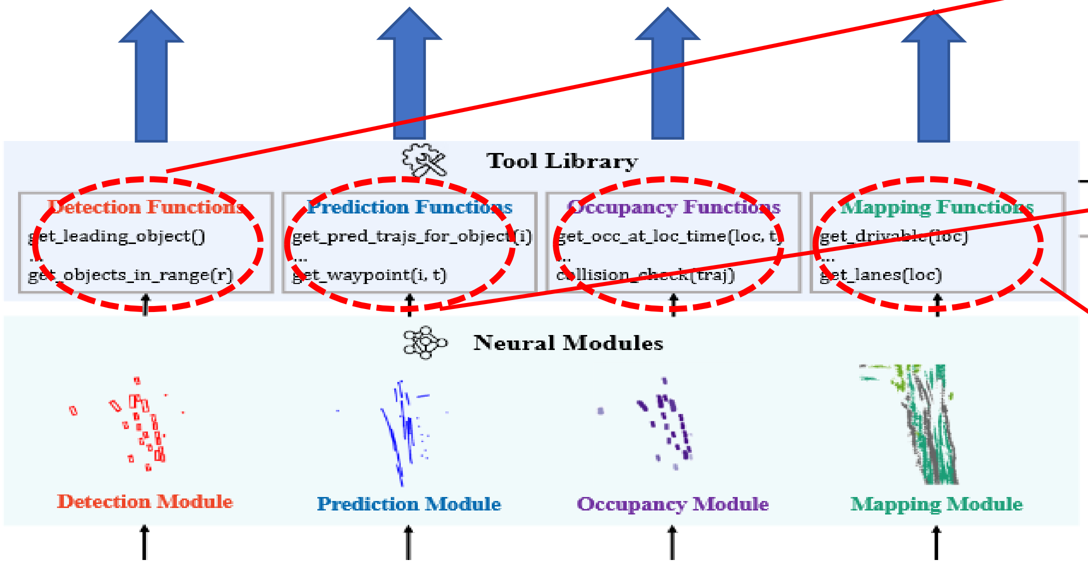
· Tool Library

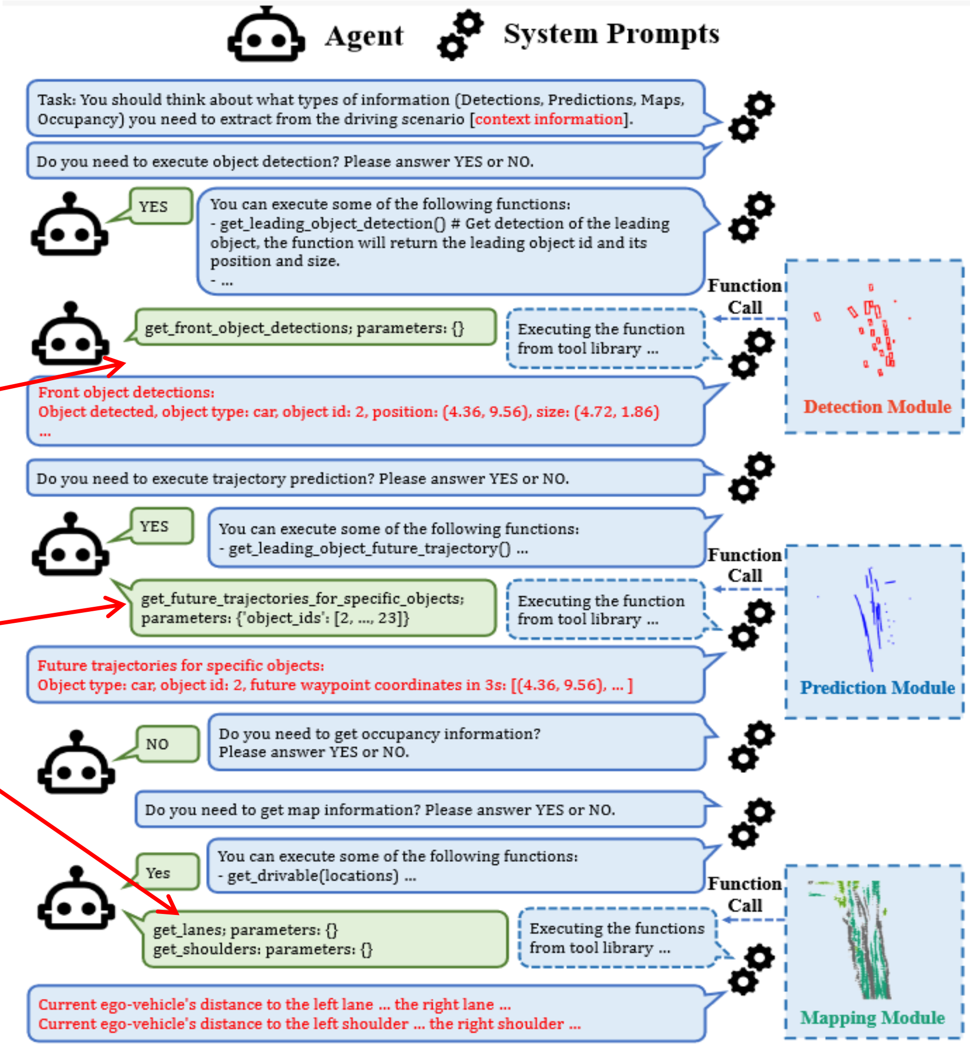
Bounding boxes

Future trajectories

Occupancy grids

Maps

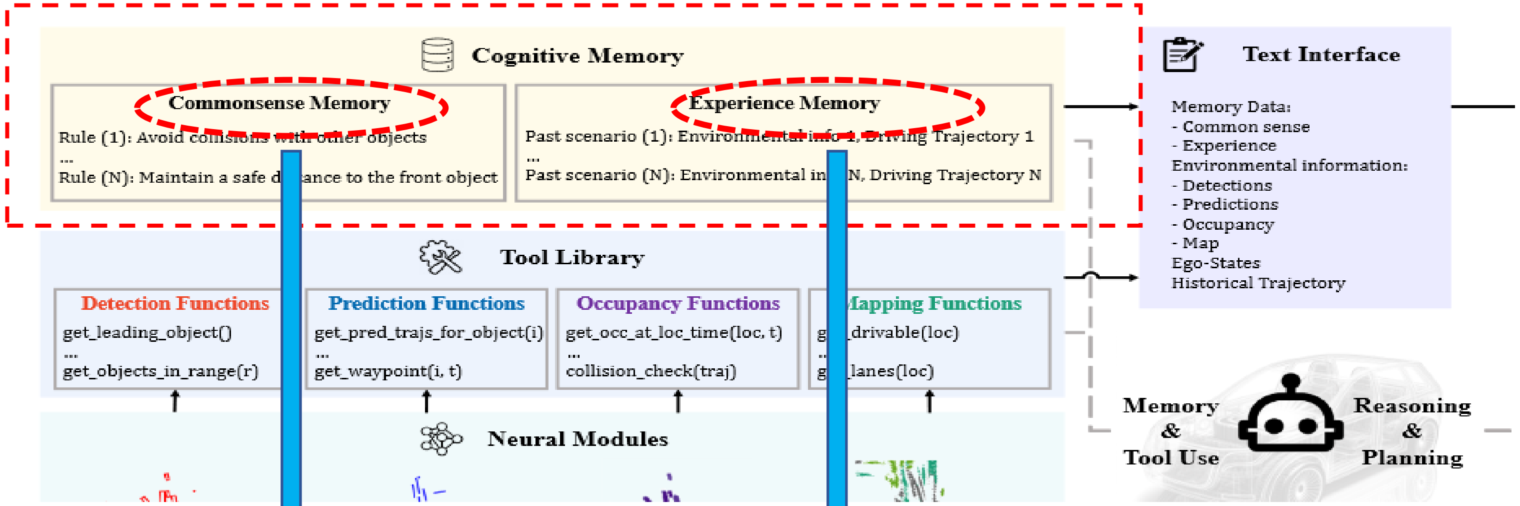


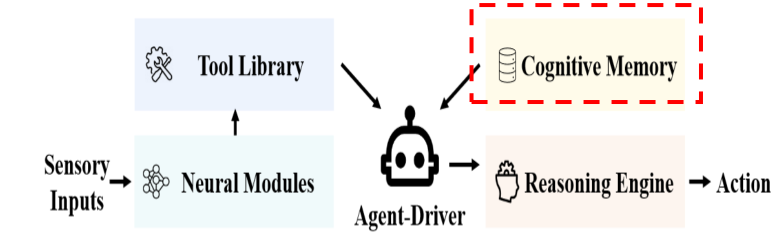


Agent-Driver

· Cognitive Memory

Human ability:
Relying on common sense to navigate, such as obeying local traffic laws and learning from driving experiences in similar situations.





(b) Agent-Driver: LLMs as an Agent for Autonomous Driving.

Interaction with 'Cognitive Memory'

(a) Conventional Perception-Prediction-Planning Pipeline.

Essential knowledge a driver typically needs for driving safely on the road.
(purely text-based and fully configurable)

Series of past driving scenarios, composed of the environmental information,subsequent driving decision at that time

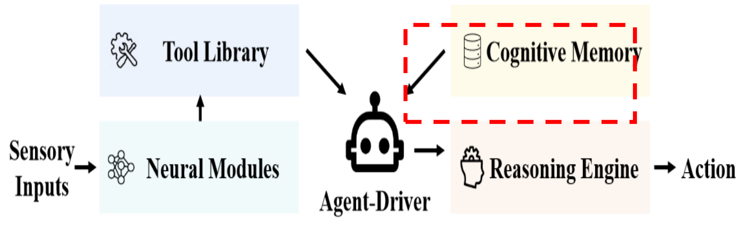
系统设计

Agent-Driver

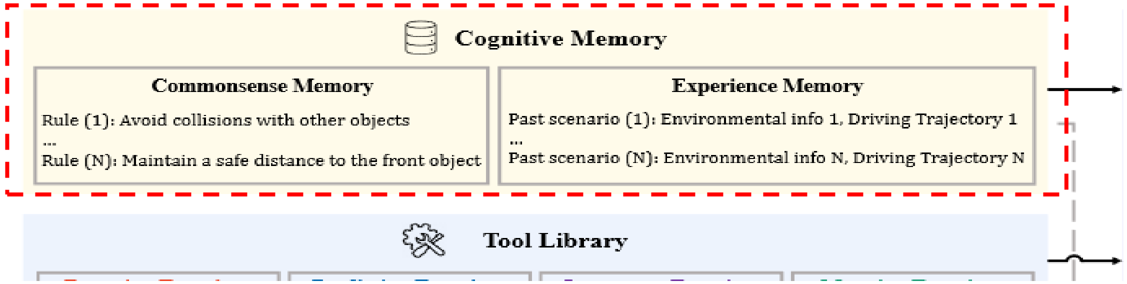
· Cognitive Memory

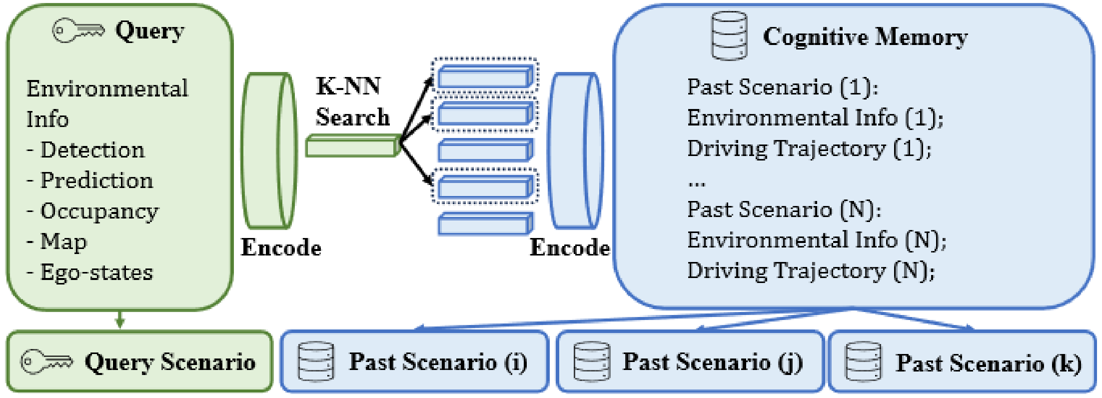
· Two-stage search algorithm

A Purpose-Built Vector Data Management System. SIGMOD 21'



(b) Agent-Driver: LLMs as an Agent for Autonomous Driving.





Stage-1: Vector Search

Stage-1:

· Encode the input query and each record in the memory into embeddings.

· Retrieve the top-K similar records via Knearest neighbors (**K-NN**) search in the embedding space.

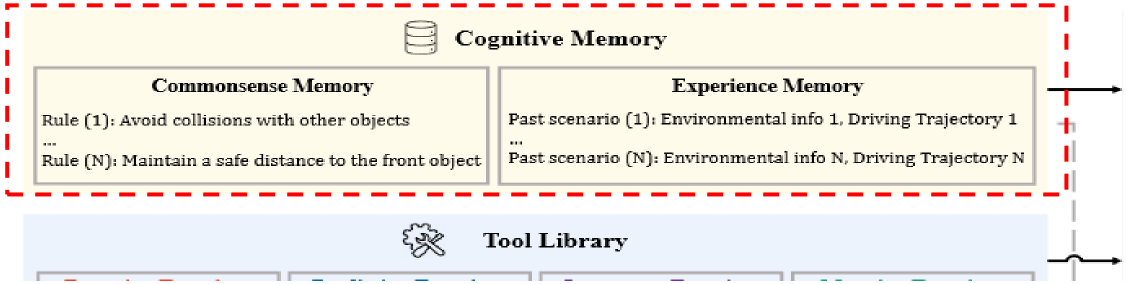
Driving scenarios are quite diverse. Embedding-based search is inherently limited by the encoding methods employed, resulting in insufficient generalization capabilities .

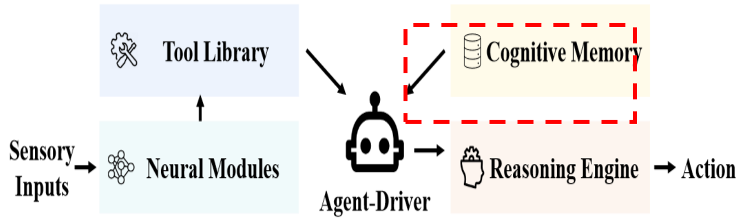
系统设计

Agent-Driver

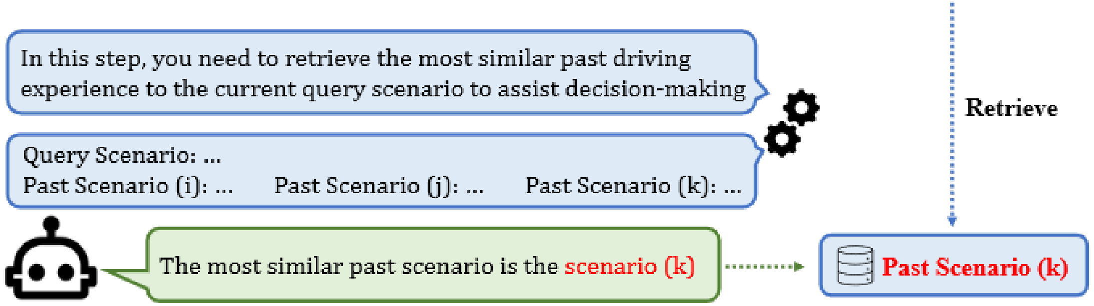
· Cognitive Memory

· Two-stage search algorithm





(b) Agent-Driver: LLMs as an Agent for Autonomous Driving.



Stage-2: LLM-Based Fuzzy Search

Stage-2:

· Incorporates an LLM-based fuzzy search.

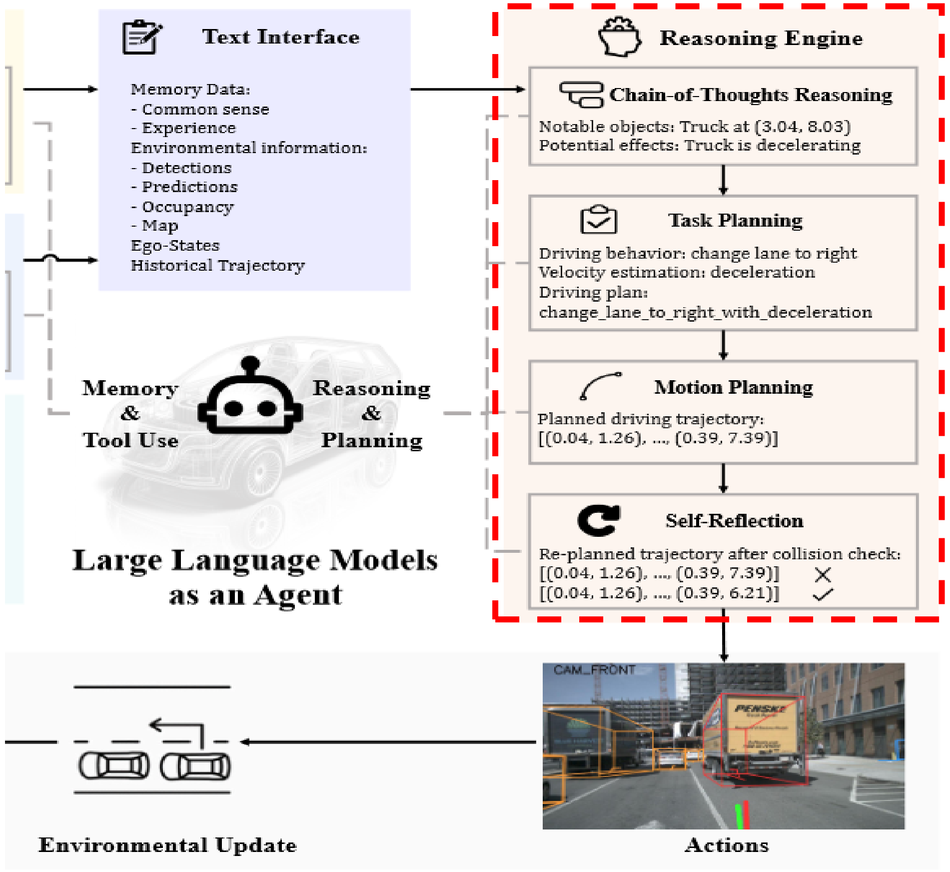
· LLM is tasked to rank these records according to their relevance to the query.

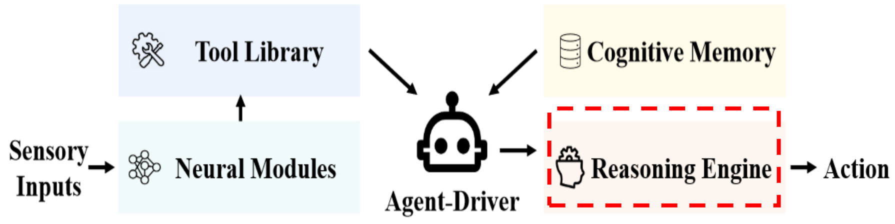
Taking advantage of LLM's capabilities in generalization and inference.

The most similar experiences,common sense,environmental information, form the input to the inference engine.

Agent-Driver

· Reasoning Engine





(b) Agent-Driver: LLMs as an Agent for Autonomous Driving.

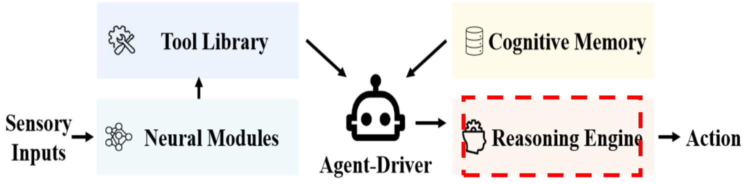
Conventional:
· Directly plan a driving trajectory based on perception and prediction results
· Lacks the reasoning skills inherent in human drivers and is ill-equipped to handle complex driving scenarios.

Agent-Driver:
· Incorporates reasoning ability into the driving decision-making process.
· Consists of four core components: **Chain-of-Thoughts Reasoning**, **Task Planning**, **Motion Planning**, and **Self-Reflection**.

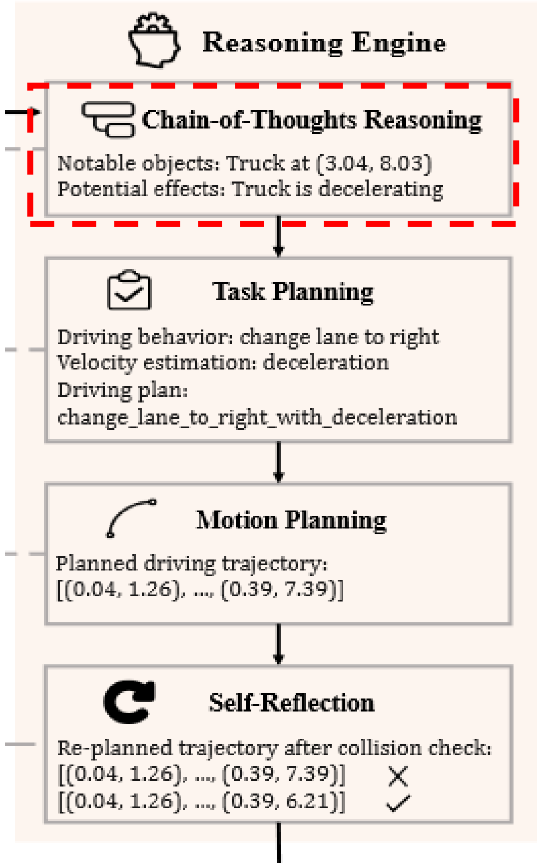
系统设计

Agent-Driver

- Reasoning Engine
  - Chain-of-Thought Reasoning



(b) Agent-Driver: LLMs as an Agent for Autonomous Driving.



In-context learning

Human-annotated example

A list of key objects and their potential effects in text.

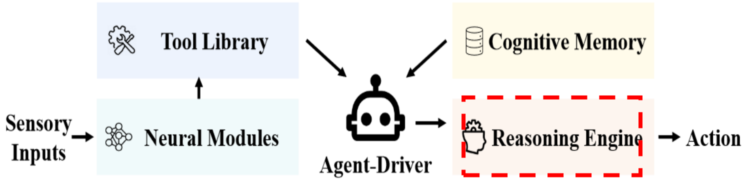
Successfully aligns the reasoning power of the LLM with the context of autonomous driving, leading to improved reasoning accuracy.

系统设计

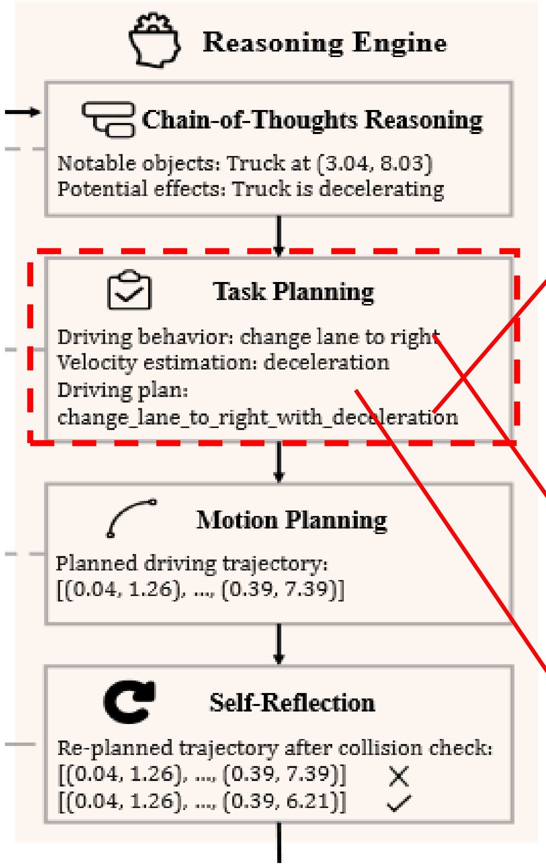
Agent-Driver

· Reasoning Engine

  · Task Planning



(b) Agent-Driver: LLMs as an Agent for Autonomous Driving.



High-level driving plans

Low-level motion planning

The traditional approach is just motion planning

defined as

Combination of
- **discrete driving behaviors**
- **velocity estimations**

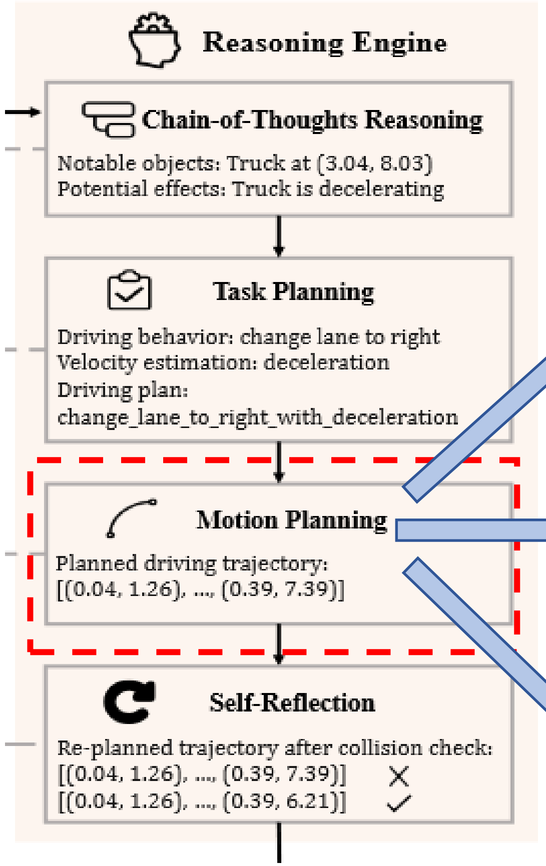
Instruction LLM through contextual learning to develop high-level driving plans based on environmental information, memory data, and chained-thinking reasoning results

系统设计

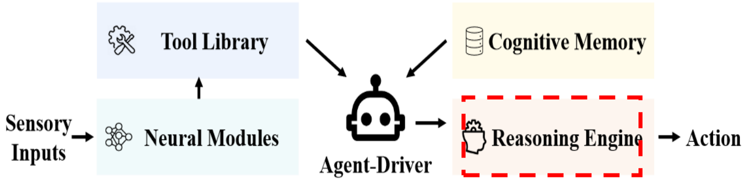
Agent-Driver

· Reasoning Engine

· Motion Planning



- Aims to devise a safe and comfortable trajectory for driving.
- Each trajectory is represented as a sequence of waypoints.



(b) Agent-Driver: LLMs as an Agent for Autonomous Driving.

Re-formulate motion planning as a **language modeling problem**.

Input： environmental information, memory data, reasoning results, and high-level driving plans.

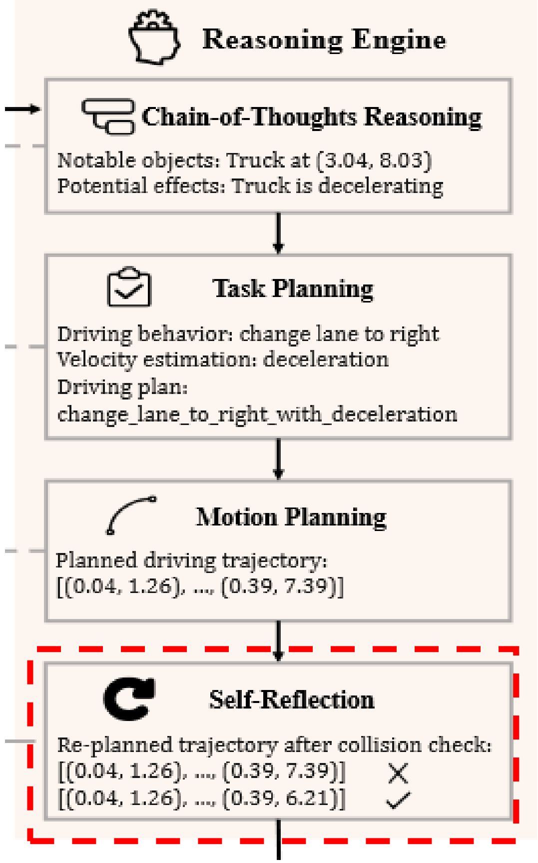
Onput : text-based driving trajectories through reasoning.

系统设计

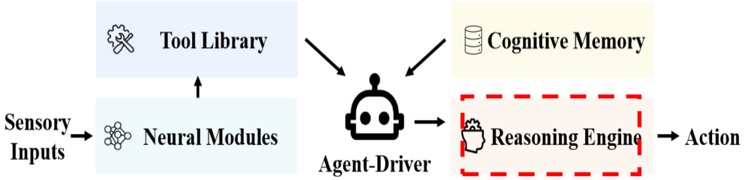
Agent-Driver

· Reasoning Engine

  · Self-Reflection



- A crucial ability in humans' decision-making process, aiming to re-assess the former decisions and adjust them accordingly.



(b) Agent-Driver: LLMs as an Agent for Autonomous Driving.

$$\tau^* = \min_{\tau} \mathcal{C}(\tau, \hat{\tau}) = \min_{\tau} \lambda_1 ||\tau - \hat{\tau}||_2 + \lambda_2 \mathcal{F}_{col}(\tau).$$

For a planned trajectory $\hat{\tau}$ from the motion planning module, the collision check function in the tool library is first invoked to check its collision.

If collision detected, we refine the trajectory $\hat{\tau}$ into a new trajectory τ^* by optimizing the cost function C.

提纲

- 研究背景
- 系统设计
- 实验评估
- 工作总结

实验评估-Benchmarks

Open-loop autonomous driving:

· **Dataset**: nuScenes , containing 1000 driving scenarios and ~34,000 keyframes covering a wide range of locations and weather conditions.

· **Evaluation metrics:** Referring to previous work, **L2 error** and **collision rate** are used to evaluate the planning performance.

ST-P3: End-to-End Vision-Based Autonomous Driving via Spatial-Temporal Feature Learning. ECCV22'

VAD: Vectorized Scene Representation for Efficient Autonomous Driving. ICCV23'

Close-loop autonomous driving:

· **Benchmarking**: adopt the **Town05-Short** benchmark powered by the **CARLA simulator**,Includes 10 challenging driving routes, each with 3 intersections and a high density of dynamic agents.

· **Evaluation metrics :** route completion and driving score, which takes into account comfort and safety.

Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR21'

实验评估-Realization details

Base model: gpt-3.5-turbo-0613

Motion planning: refer to Mao et al. (2023a), using human driving trajectories from the nuScenes training set to fine-tune the LLM.

Neural modules: modules from Hu et al. (2023) were used; perceptual modules from LAV (Chen & Krähn ¨ uhl, 2022) were used, and the rest of the system was kept consistent.

Training and evaluation protocols: the training setup and evaluation protocols from Chen & Krähn ¨ uhl (2022) were followed to ensure fair comparisons.

Learning from all vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Planning-oriented autonomous driving. CVPR23’

实验评估-Function Interaction

- **Comparison with State-of-the-art Methods**
  - **Open-Loop Results**

	Method	L2 (m) ↓ 1s	2s	3s	Avg.	Collision (%) ↓ 1s	2s	3s	Avg.
ST-P3 metrics	ST-P3 (Hu et al., 2022) ECCV	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
	VAD (Jiang et al., 2023) ICCV	0.17	0.34	**0.60**	0.37	0.07	0.10	0.24	0.14
	GPT-Driver (Mao et al., 2023a)	0.20	0.40	0.70	0.44	0.04	0.12	0.36	0.17
	Agent-Driver (ours)	**0.16**	**0.34**	0.61	**0.37**	**0.02**	**0.07**	**0.18**	**0.09**
UniAD metrics	NMP (Zeng et al., 2019)	-	-	2.31	-	-	-	1.92	-
	SA-NMP (Zeng et al., 2019)	-	-	2.05	-	-	-	1.59	-
	FF (Hu et al., 2021)	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
	EO (Khurana et al., 2022)	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
	UniAD (Hu et al., 2023)	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
	GPT-Driver (Mao et al., 2023a)	0.27	0.74	1.52	0.84	0.07	0.15	1.10	0.44
	Agent-Driver (ours)	**0.22**	**0.65**	**1.34**	**0.74**	**0.02**	**0.13**	**0.48**	**0.21**

Table 2: **Open-loop planning performance compared to the state-of-the-arts.** Agent-Driver significantly outperforms prior works in terms of L2 and collision rate. Our approach attains more than 30% performance gains in collisions compared to the state-of-the-art methods.

实验评估-Function Interaction

- **Comparison with State-of-the-art Methods**
  - **Closed-Loop Results**

Methods	Driving Score ↑	Route Completion ↑
CILRS (Codevilla et al., 2019)	7.47	13.40
LBC (Cui et al., 2021)	30.97	55.01
Transfuser (Prakash et al., 2021)	54.52	78.41
ST-P3 (Hu et al., 2022)	55.14	86.74
VAD (Jiang et al., 2023)	**64.29**	<u>87.26</u>
Agent-Driver (Ours)	<u>57.33</u>	**91.37**

Table 1: **Closed-loop planning performance compared to the state-of-the-arts.** Agent-Driver yields the best route completion and an on-par driving score compared to prior arts.

实验评估-Function Interaction

- **Few-shot Learning**

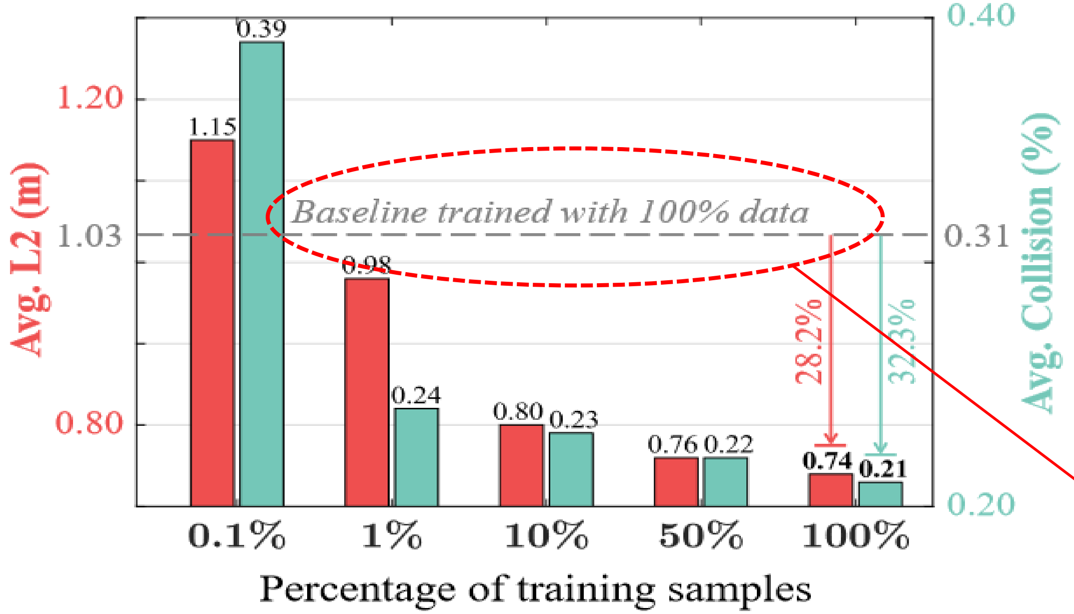


Figure 6: **Few-shot learning.** The motion planner in Agent-Driver fine-tuned with 1% data exceeds the state-of-the-art (Hu et al., 2023) trained on full data, verifying its few-shot learning ability.

Planning-oriented autonomous driving. CVPR23'

实验评估-Function Interaction

- **Interpretability**

*****Environmental information:*****
Front object detections:
Front object detected, object type: pedestrian, object id: 0, position: (-4.32, 13.85), size: (0.76, 0.91)
...
Front object detected, object type: pedestrian, object id: 9, position: (2.23, 19.39), size: (0.57, 0.69)
Future trajectories for specific objects:
Object type: pedestrian, object id: 0, future waypoint coordinates in 3s: [(-4.44, 14.50), ...]
...
Object type: pedestrian, object id: 9, future waypoint coordinates in 3s: [(1.74, 19.76), ...]
Map information (road shoulders):
Current ego-vehicle's distance to left shoulder is 0.5m and right shoulder is 4.5m

*****Common sense:*****
- Maintain a safe distance from the objects in front of you...
*****Past driving experience for reference:*****
Most similar driving experience from memory with similarity score: 0.77:
Scenario information: ...
The planned trajectory in this scenario for your reference:[(0.04,2.49), ... ,(0.30,12.26)]

*****Chain-of-thoughts reasoning:*****
- Notable objects: pedestrian at (0.80,18.81), moving to (-2.53,20.89) at 3.0 second
- Potential effects: may collide if continue driving at this speed.
...
*****Task planning:*****
Behavior: forward; Speed: deceleration; Driving plan: move forward with a deceleration
*****Motion planning:*****
Trajectory: [(-0.03, 2.47), (-0.10, 4.84), (-0.19, 7.10), (-0.29, 9.25), (-0.39, 11.29), (-0.49, 13.22)]
*****Self-reflection:*****
No collision. No change to the motion planning result.

*****Environmental information:*****
Front object detections:
Front object detected, object type: car, object id: 4, position: (-2.11, 14.95), size: (1.96, 4.76)
...
Front object detected, object type: car, object id: 6, position: (5.31, 32.79), size: (1.90, 4.48)
Future trajectories for specific objects:
Object type: pedestrian, object id: 4, future waypoint coordinates in 3s: [(-2.39, 14.80), ...]
...
Object type: pedestrian, object id: 6, future waypoint coordinates in 3s: [(5.32, 32.78), ...]
Map information (lanes):
Current ego-vehicle's distance to left lane is 1.5m and right lane is unknown

*****Common sense:*****
- Avoid collision with other objects...
*****Past driving experience for reference:*****
Most similar driving experience from memory with similarity score: 0.45:
Scenario information: ...
The planned trajectory in this scenario for your reference:[(-0.14,0.98), ... ,(-5.10,8.27)]

*****Chain-of-thoughts reasoning:*****
- Notable objects: car at (-2.11,14.95), moving to (-2.84,14.53) at 1.5 second
- Potential effects: inside the safety zone of the ego-vehicle at 1.5 second.
...
*****Task planning:*****
Behavior: turn left; Speed: deceleration; Driving plan: turn left with a deceleration
*****Motion planning:*****
Trajectory: [(-0.11,0.94), (-0.31,1.81), (-0.62,2.75), (-1.16,3.88), (-1.84,4.93), (-2.95,6.29)]
*****Self-reflection:*****
No collision. No change to the motion planning result.

Figure 7: **Interpretability of Agent-Driver.** In the referenced images, planned trajectories of our system and human driving trajectories are in red and green respectively. Agent-Driver extracts meaningful objects (in yellow) from all detected objects (in blue) via the tool library. The reasoning engine further identifies notable objects (in red). Messages from the tool library, cognitive memory, and reasoning engine are recorded in colored text boxes. Every message is documented and our system is conducted in an interpretable and traceable way.

实验评估-Function Interaction

- **Compatibility with Different LLMs**

Method	L2 (m) ↓ 1s	2s	3s	Avg.	Collision (%) ↓ 1s	2s	3s	Avg.
Llama-2-7B	0.25	0.69	1.47	0.80	0.02	0.27	0.78	0.35
gpt-3.5-turbo-1106	0.24	0.71	1.47	0.80	0.03	0.08	0.63	0.25
gpt-3.5-turbo-0613	0.22	0.65	1.34	0.74	0.02	0.13	0.48	0.21

Table 3: **Compatibility to different LLMs.** Agent-Driver realizes satisfactory motion planning performance utilizing different types of LLMs as agents.

实验评估-Function Interaction

- **Stability**

  · LLMs typically suffer from arbitrary predictions—they might produce invalid outputs. (e.g., hallucination or invalid formats)

Percentage of training samples	0.10%	1%	10%	50%	100%
Number of invalid outputs	2	0	0	0	0

Table 4: **Stability of Agent-Driver exposed to different amounts of training samples.** With only 1% training samples ($\sim$ 230 samples), Agent-Driver produces *zero* invalid output.

实验评估-Function Interaction

- **In-Context Learning vs. Fine-Tuning**

Modules: CoT Reason.+Task Plan.	Modules: Motion Plan.	Avg. L2 (m)	Avg. Col. (%)
Fine-tuning	In-context learning	1.81	0.79
In-context learning	In-context learning	1.90	0.79
Fine-tuning	Fine-tuning	0.72	0.22
In-context learning	Fine-tuning	0.74	0.21

Table 5: **In-context learning vs. fine-tuning.** In-context learning performs slightly better in reasoning and task planning. Fine-tuning is indispensable for motion planning.

提纲

- 研究背景
- 系统设计
- 实验评估
- 工作总结

总结

Agent-Driver

总体上：

Introduces Agent-Driver, a novel human-like paradigm，leverage LLMs as an agent to schedule different modules in autonomous driving.

设计上：

Propose a tool library, a cognitive memory, and a reasoning engine to bring human-like intelligence into driving systems.

实验上：

Extensive experiments on real-world driving datasets , confirm the effectiveness, small amount of learning capability, and interpretability of agent driving.

The experiments are all deeply significant with comparisons to top conference papers.

These findings reveal the potential of LLM as an agent in human-level intelligent driving systems.

总结

个人idea

- For this paper, LLM might also be able to combine multimodal information such as **vision** and **speech** to fuse processing.
- The adaptability of LLMs to new situations and emergencies is still limited. Consideration could be given to incorporating an “e-learning” module into the system so that LLMs can update their knowledge base and memory at any time, thus realizing real-time adaptation and self-adjustment .

Q&A