OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

ELM: Embodied Understanding of Driving Scenes

(A 60 min Talk)

Online

August 2024

Yunsong Zhou, Shanghai Jiao Tong University / Shanghai AI Lab

Outline

- **Introduction to End-to-end Autonomous Driving**
  - Modular vs End-to-end
  - Industry Applications
  - Research Roadmap
- **Challenges**
- **Mainstream Lineup of Work**
  - Embodied Understanding of Driving Scenarios

Introduction

End-to-end Autonomous Driving

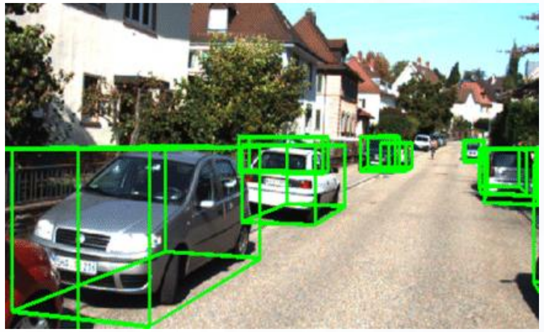
OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

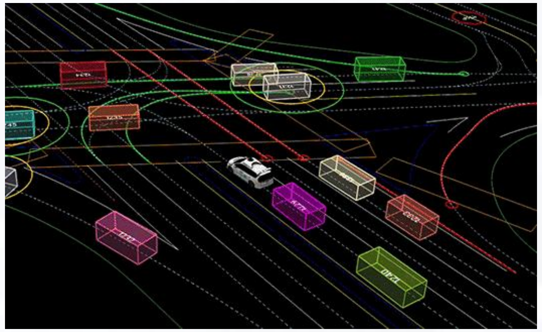
Problem setting | Autonomous Driving (AD) Tasks

Challenge | Various weathers, illuminations, and scenarios

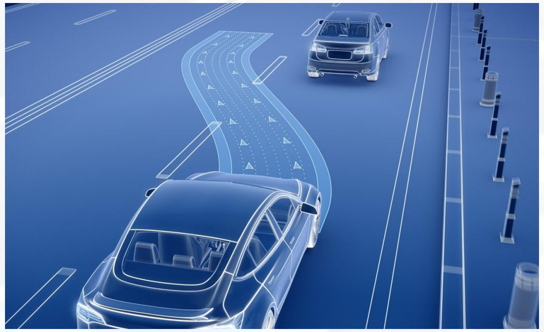
Perception → Bounding boxes → Prediction → Waypoints → Planning → Trajectory



What are around?



How will they go in the future?

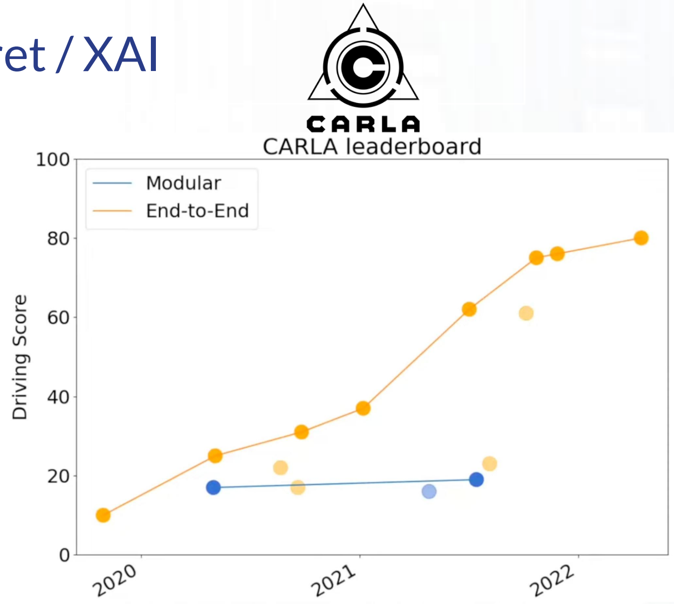


Where should I go?

Motivation | Why End-to-end (E2E) Autonomous Driving?

Disadvantages

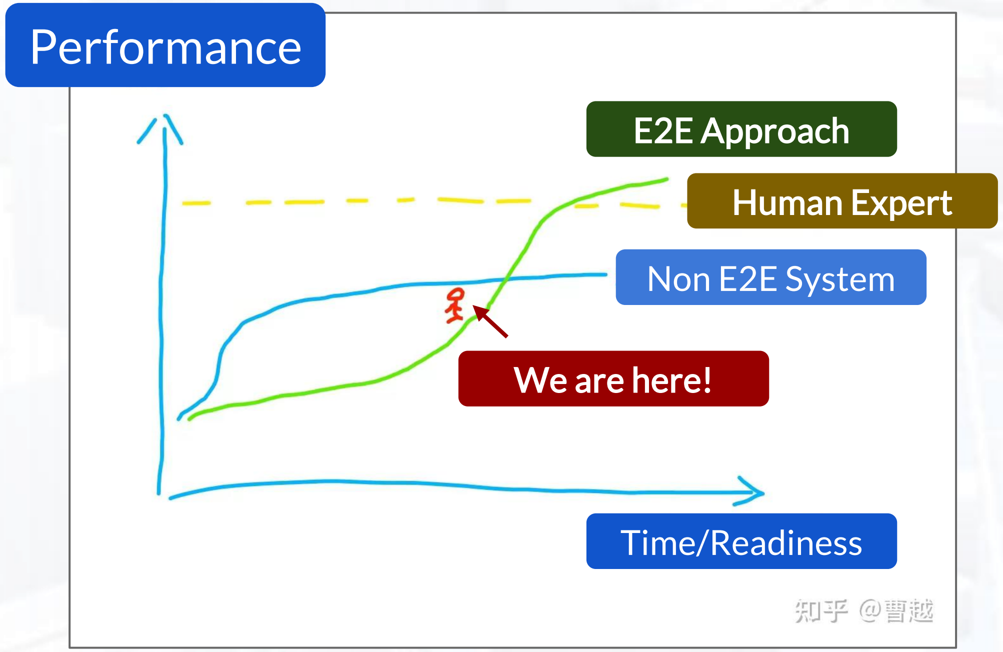
- Closed-loop evaluation only available in simulator / onboard test
- Lack of real-world data
- Hard to interpret / XAI



https://leaderboard.carla.org/

Credit to Andreas Geiger @ CVPR Workshop 2023

E2E vs Non-E2E



Credit to Dr. Yue Cao @ Zhihu

Trending | End-to-end AD (E2EAD)

Industry

TESLA

E2E Vehicle

...
v12 is reserved for when FSD is **end-to-end AI,** from images in to steering, brakes & acceleration out.

E2E Robot

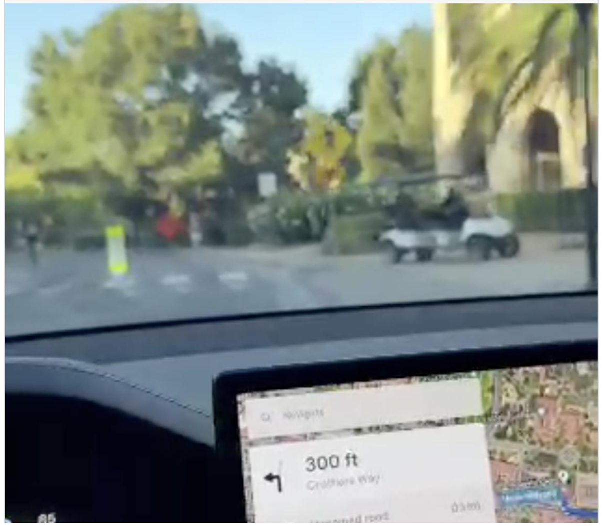
Tesla Optimus @Tesla_Optimus · Sep 24
Optimus can now sort objects autonomously 🤖

Its neural network is trained fully **end-to-end**: video in, controls out.

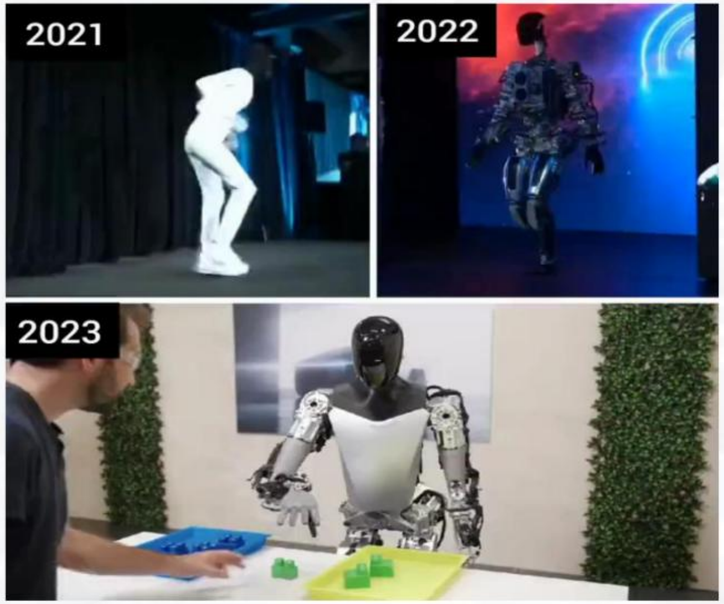
Ashok Elluswamy @aelluswamy

This end to end neural network approach will result in the safest, the most competent, the most comfortable, the most efficient, and overall, the best self-driving system ever produced. It's going to be very hard to beat it with anything else!

Elon Musk @elonmusk · Aug 26
twitter.com/i/broadcasts/1...



No hard-code.
Completely learning on its own.
End-to-end, video to neural network to controls.
Don't need map data at all, only coordinates!
No cellular connection needed.

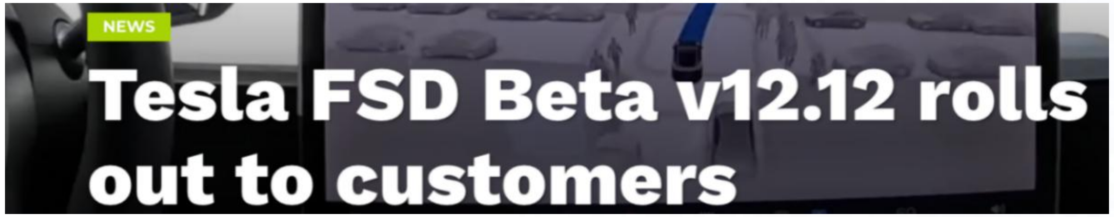


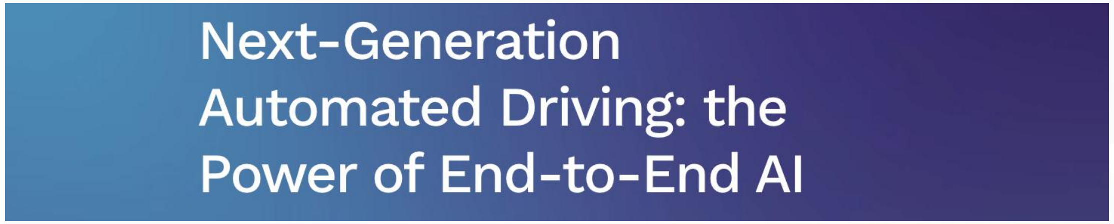
My Opinion

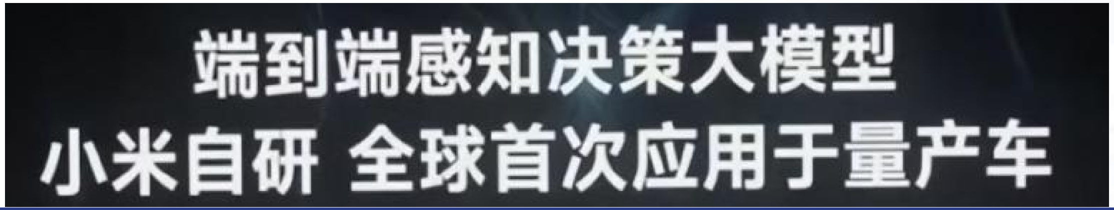
- Probably e2e as a backup module
- Massive high-quality data prevail
- Mapless is promising and feasible

Application | End-to-end Autonomous Driving

Industry

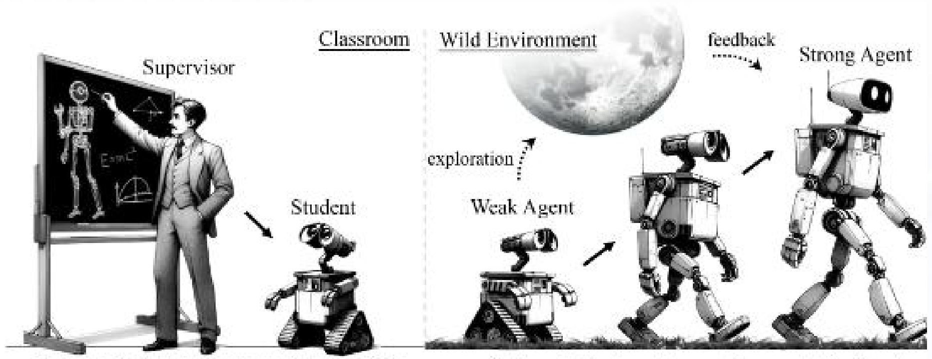






- *Tesla website*
- *Mobileye at CES 2024*
- *Wayve website*
- *DeepRoute (元戎启行) at GTC 2024*
- *Mi (小米) Automobile Technological Event 2024*

研究背景与现状 - 端到端



AGI（L5）

Drive Anywhere/Anytime

↓

泛化、智能、安全

80% 自动驾驶问题
已被解决（L2）

剩余10%
Corner case in Daily Life
靠规则永远写不完（L3/L4）

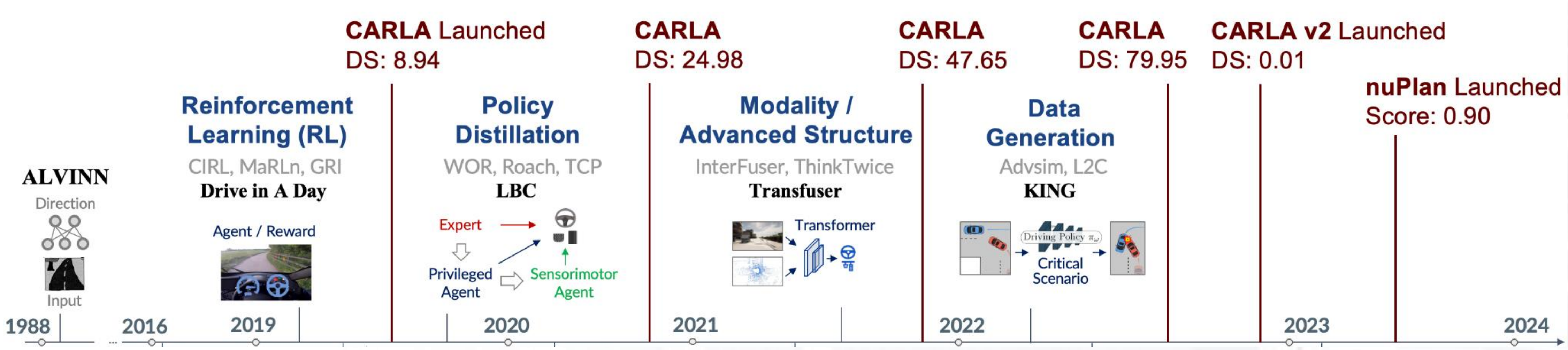
→ **加数据?**
如何采? ←
如何训?

↓

通用的端到端具身智能体进化方法

Bad Cases in Tesla FSD v12.3 Test drive before CVPR'24

Roadmap | End-to-end Autonomous Driving



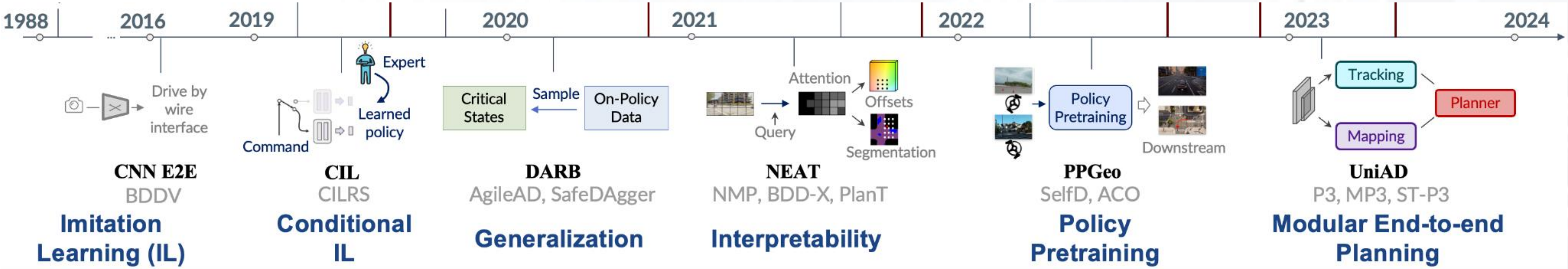
Summary (1/2)

- Carla leaderboard gets much improved over the years. With new mapping / routes (Carla v2) and nuPlan benchmark, this field got so much to do.
- RL method is prevalent in the beginning (since it's natural)
- Input modality and more advanced structure boosts the performance

Roadmap | End-to-end Autonomous Driving

Summary (2/2)

- The First Neural Net based method dates back to 2016 using Imitation Learning
- Learned policy from Experts (IL), with data augmentation, could prevail in performance
- Interpretability, with explicit design in the network stands out recently
- End-to-end design comes to obsess many merits in previous attempt



Analogy to General Domains in CV/NLP/Robotics

General Large Models

Industrial Large Models (Application)

Domain	Method Abbreviation	Institute / Time	Data Scale	Public?
NLP (LLM)	GPT-4	OpenAI / 2023.3	13T tokens	✗
	LLaMA 2	Meta / 2023.7	2T tokens	☑
Vision	ViT-22B	Google / 2023.2	4B images	✗
Vision Language (LLM backend)	BLIP-2	Salesforce / 2023.1	129M images-text pairs	☑
Autonomous Driving nuScenes: 4.5h	**DriveAGI (GenAD)**	OpenDriveLab / 2023.11	**2000 h videos (public)**	☑
	GAIA-1	Wayve / 2023.6	4700 h videos	✗
	World Model Demo	Tesla / 2023.6	Unknown (Large-scale)	✗
Robotics (LLM backend)	PaLM-E	Google / 2023.3	Unknown (Large-scale)	✗
	RT-2	DeepMind / 2023.7	1B img-text pairs / 13 robots / 17 months	✗

If taken seriously for AD: lots of compute (at least 200 A100s) + **massive amount of data** (at least 10k hours of diverse, high-quality data)

OpenDriveLab

Trending in E2EAD | Driving + Language

Honda Research Institute USA

Go straight at an intersection then turn left.
There are **construction cones** on the road.

HAD — human-to-vehicle driving advice, highlighting key objects.

Explainable Driving Behavior

BDD-X — one-sentence explanation of driving behavior.

Berkeley DeepDrive

Action description: *(1) The car is driving*

Action justification: ***as** there is nothing to impede it.*

2019 — Planning

2022 — Prediction

2023 — Full-stack

Trending in E2EAD | Driving + Language

Explainable Driving Behavior

BDD-X — one-sentence explanation of driving behavior.

Berkeley DeepDrive

2019

HAD — human-to-vehicle driving advice, highlighting key objects.

Planning

2022

Rank2Tell — reasoning for the rank of objects' importance level.

Talk2Car — a description of how to reach the goal point from current position.

DRAMA — caption about important objects and future decision.

Prediction

2023

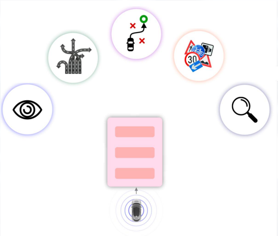
DriveLM — perception-prediction-planning driving description with graph VQA.

Dataset	Source Dataset	# Frames	Avg. captions / QA per annotated frame	Total captions / QA in Perception	Total captions / QA in Prediction	Total captions / QA in Planning	Logic among captions/QA pairs
nuScenes-QA [47]	nuScenes	34,149	13.5	460k**	✗	✗	None
nuPrompt [66]	nuScenes	34,149	1.0	35k*	✗	✗	None
HAD [31]	HDD	25,549	1.8	25k	✗	20k	None
BDD-X [30]	BDD	26,228	1	26k	✗	✗	None
DRAMA [40]	DRAMA	17,785	5.8	85k	✗	17k	Chain
Rank2Tell [51]	Rank2Tell	5,800	-	-	✗	-	Chain
DriveLM-nuScenes	nuScenes	4,871	**91.4**	144k*	153k	146k	**Graph**
DriveLM-CARLA	CARLA	**183,373**	20.5	**2.46M****	**578k****	**714k****	**Graph**

Table 1. **Comparison of DriveLM-nuScenes & -CARLA with Existing Datasets.** * indicates semi-rule-based labeling (w/ human annotators), ** indicates fully-rule-based (no human annotators), and - means publicly unavailable. DriveLM-Data significant advances annotation quantity, comprehensiveness (covering **perception, prediction and planning**), and logic (chain to **graph**).

Full-stack

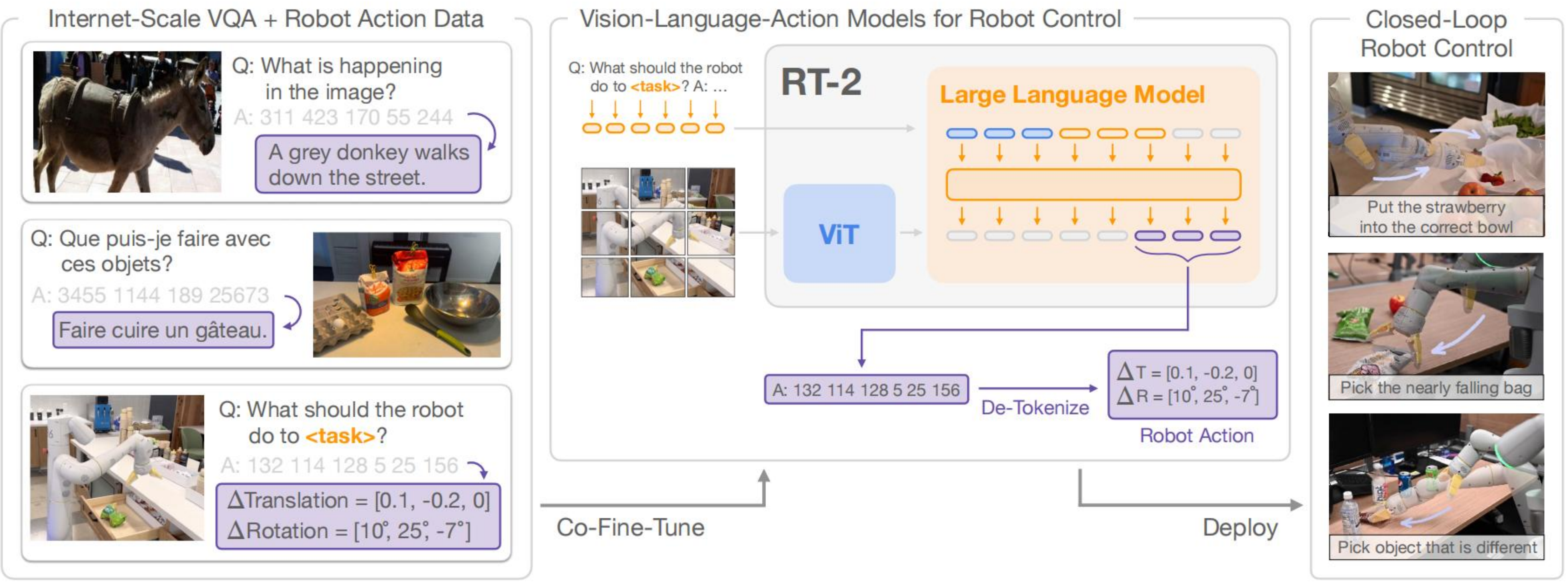
LINGO-1 — commentary for explaining driving behaviours.



For now, language into driving is marginal (trivial).
Serves only as a "commentator". We haven't verified (or seen) the effectiveness.

OpenDriveLab

Insight | VLM in Robotics / Embodied AI



- How vision-language models trained on Internet-scale data can be incorporated directly into **end-to-end robotic control**
- Goal: to **boost generalization** and enable emergent semantic reasoning

- Robotic tasks naturally fits into language at dissecting tasks step by step using language (prompt).
- Is it the **right way** to open the language tool box as does in Robotics for Autonomous Driving?

Key ingredient(s): huge amount of data (not public) + language prompt to dissect tasks

OpenDriveLab

OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Grand Challenge 2024

CVPR 2023 AD Challenge - Recap

https://opendrivelab.com/AD23Challenge.html

Overview

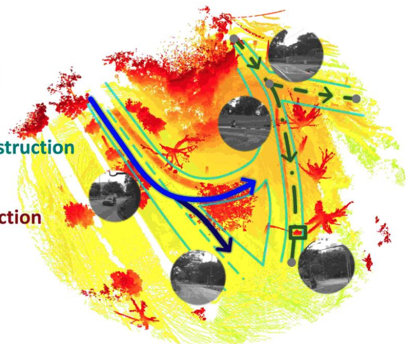
Tracks

Track 1
OpenLane Topology

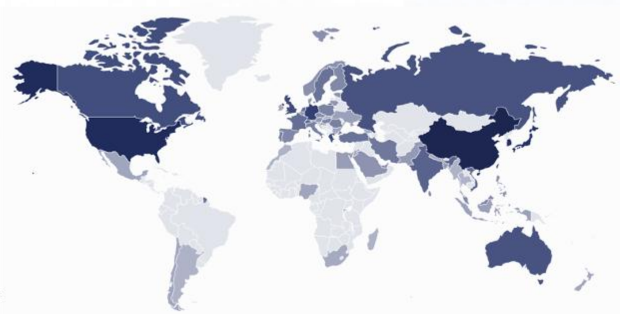
Track 2
Online HD Map Construction

Track 3
3D Occupancy Prediction

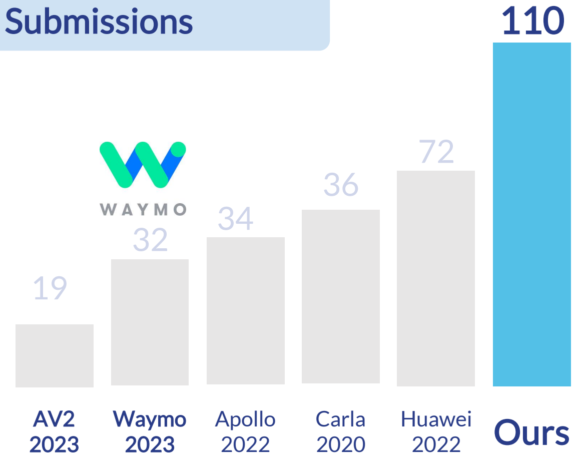
Track 4
nuPlan Planning



Internationalization



Submissions



Teams

> 2.3k
Submissions

> 15
Contries / Regions

> 68k
Website / Social Media Views

Competing Teams

Attention from All Sides

NVIDIA Research ... Award at CVPR

Bosch Center for Artificial Intelligence (BCAI)

#CVPR2023 #IndustrialAI #automateddriving #research #competition #nuplan

旷视科技

近日，为期三个月的 CVPR 2023 自动驾驶国际挑战赛比赛结果揭晓。旷视研究院在OpenLane Topology 赛道中击败 30 余支国内外队伍，夺得冠军。

Team 42dot Wins 2nd Place in the Autonomous Driving Challenge at CVPR 2023

International Teams Advance AV Resear... 2023 nuPlan Challe...

Challenges of Mass Production Autonomous Driving in China

And the Recent Progress from Xpeng Motors in 2023

ころんびあ @colum2131

実はTuringでCVPR Challenge 2023参加していましたー！

自分はTrack 1のOpenLane Topologyという走行レーンや交通信号機・標識の検出や、それら間にトポロジー(意味的な接続性)があるか分類するタスクにチームで参加し...

什么样的方案，夺得了CVPR自动驾驶挑战赛冠军？

雷峰网 机器之心 2023-06-19 19:21 发表于北京

自动驾驶圈正被"去高精地图"风暴席卷。

今年6月中旬，一年一度的AI顶会 CVPR 2023 在加拿大举行，同期还有场自动驾驶国际挑战赛，这场挑战赛吸引了来自15个国家的270多支队伍参赛，有2300多件作品投稿，赛况激烈。

36Kr 旷视刷榜了什么自动驾驶比赛？

旷视研究院参加的这个比赛，是CVPR 2023专门面向自动驾驶感知决策系统设立的挑战赛。

CVPR 2024 Autonomous Grand Challenge

120,000 USD Bonus!

https://opendrivelab.com/challenge2024/

Timeline

2023.12

Contest tracks open for comments

2024.02

Track Announcement

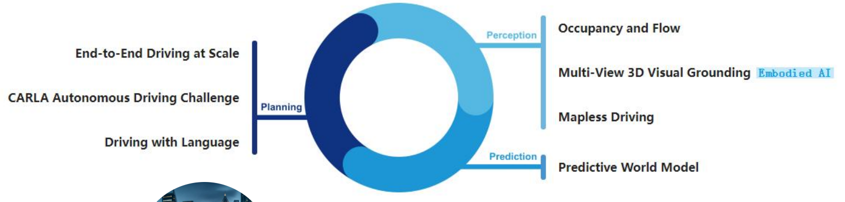
2024.03

Challenge Starts
Dataset / Evaluation Metrics Public

2024.06
Workshop @ Seattle

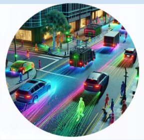
Competition ends
Announcement of winning teams

Seven Brand-new Tracks



End-to-End Driving at Scale

Predictive World Model

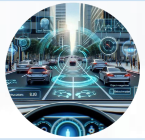


Occupancy and Flow

Multi-View 3D Visual Grounding

CARLA AD Challenge

Driving with Language



Mapless Driving

Cooperation from All Sides

OpenCVLab
OpenRobotLab
清華大學 Tsinghua University
Hugging Face
Motional
NVIDIA
CARLA
美团 Meituan Autonomous Delivery
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
WAYVE
机器之心 Synced

OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

ELM: Embodied Understanding of Driving Scenarios

Embodied Understanding

Embodied Understanding of Driving Scenarios

Yunsong Zhou[1,2*] Linyan Huang[1*] Qingwen Bu[1,2*] Jia Zeng[1] Tianyu Li[1,3]
Hang Qiu[4] Hongzi Zhu[2†] Minyi Guo[2] Yu Qiao[1] Hongyang Li[1†]

[1] OpenDriveLab, Shanghai AI Lab [2] Shanghai Jiao Tong University
[3] Fudan University [4] University of California, Riverside

Credits:
https://github.com/OpenDriveLab/ELM, https://arxiv.org/abs/2403.04593

OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

星合联盟
STAR LEAGUE

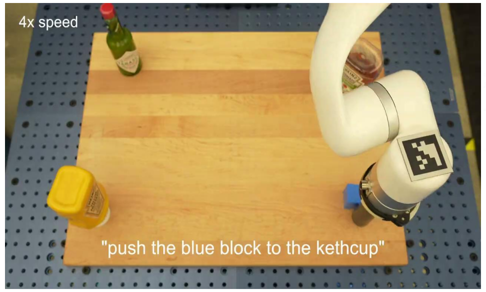
Introduction

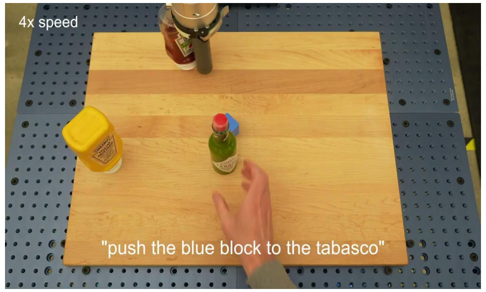
ELM - Introduction

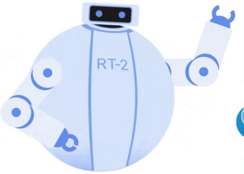
Embodied Understanding

Interacting with environments & **reasoning** via common sense

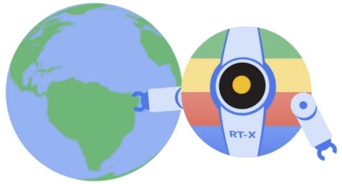
> This makes transfer of internet scale knowledge to robots more direct, and may provide a more scalable class of approaches in the future.







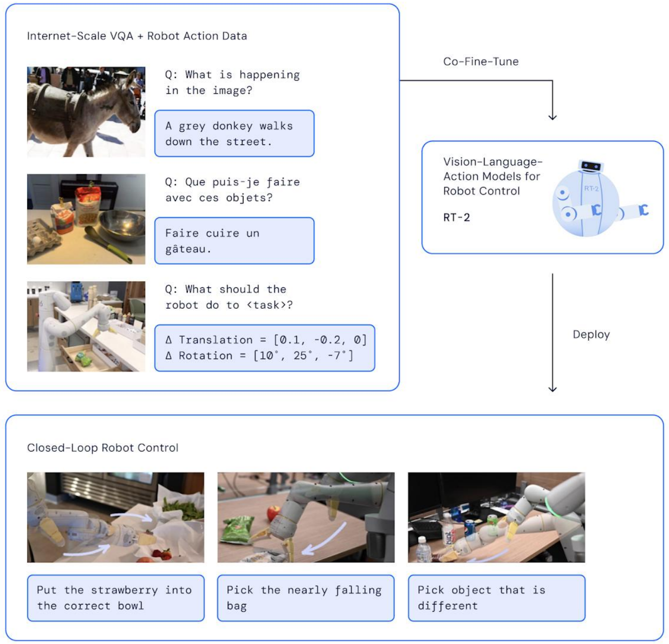
Google DeepMind



PaLM-E 562B

BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

TOYOTA RESEARCH INSTITUTE



Credits: Google DeepMind, RT-2, RT-X, PaLM-E

OpenDriveLab

ELM - Introduction

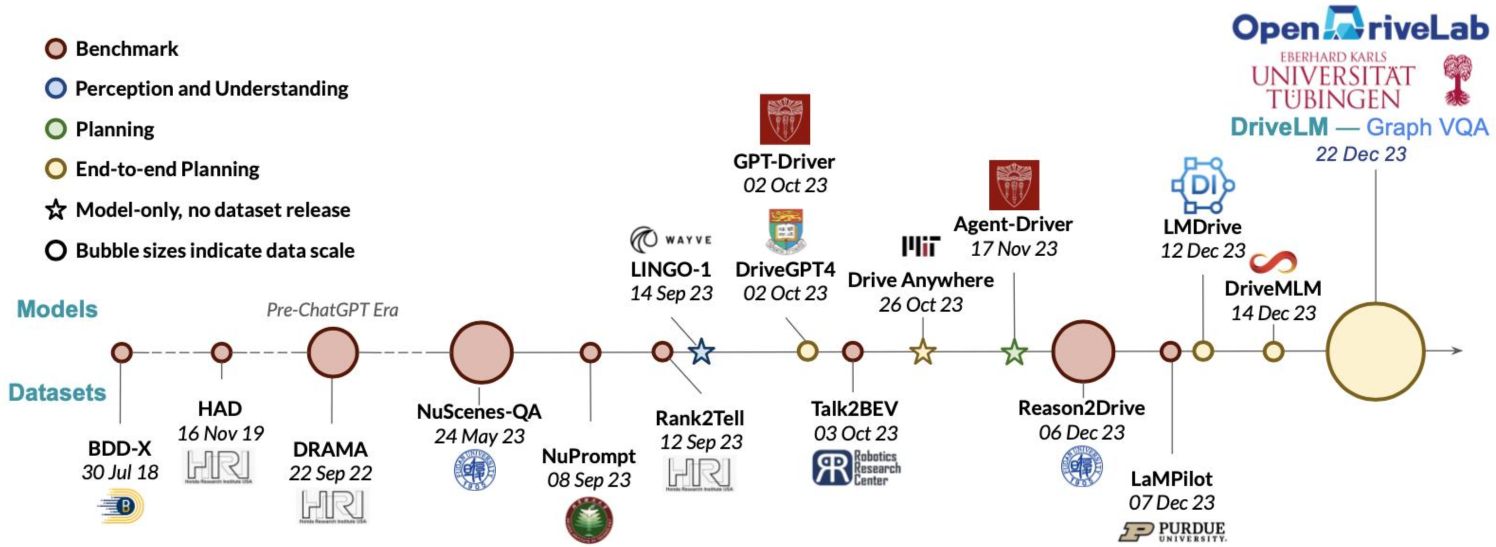
Embodied Understanding

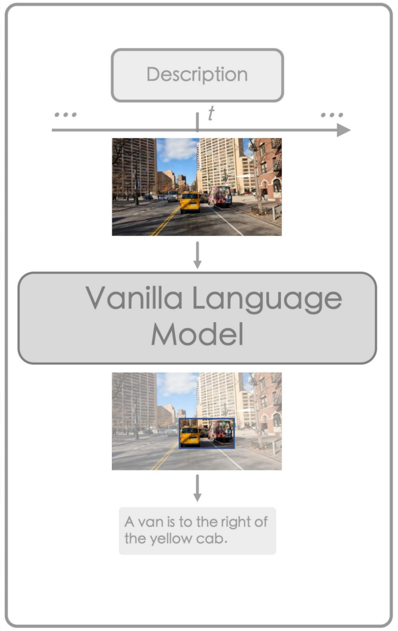
Interacting with environments & **reasoning** via common sense

Vision Language Models

Only focus on 2D domain, *i.e.*, **description**

Credits: DriveLM; Open-sourced Data Ecosystem in Autonomous Driving

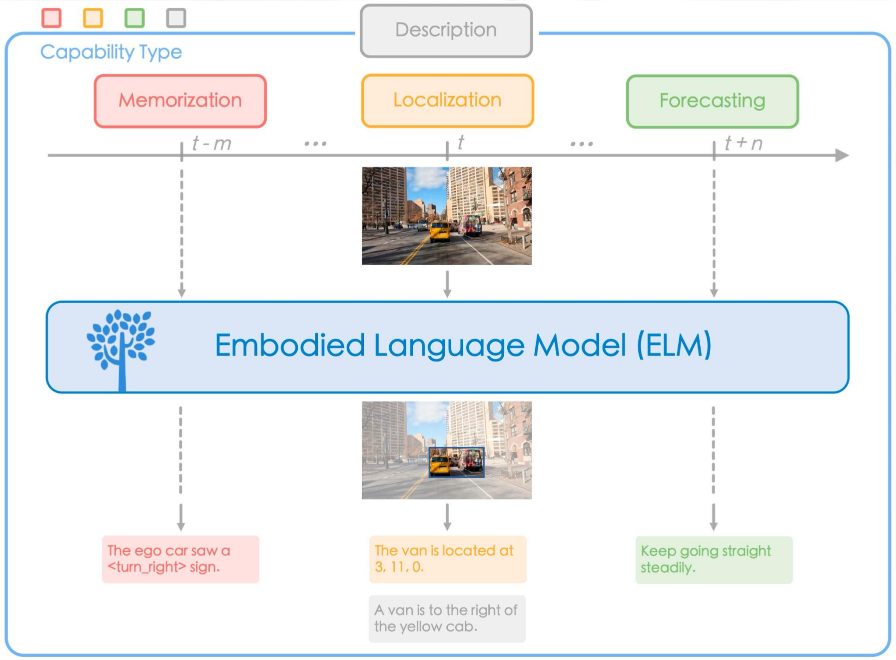




ELM - Introduction

Embodied Understanding	Vision Language Models	Embodied Language Model
Interacting with environments & **reasoning** via common sense	Only focus on 2D domain, *i.e.*, **description**	Expanding vanilla VLMs to driving scenes

- Task: embodied understanding of driving scenarios.
- Capabilities: description, **localization**, **memorization**, **forecasting**.
- Model: **ELM** with long-horizon **space** and **time**.
- Benchmark: A spectrum of tasks in an embodiment setting.



ELM - Introduction

Embodied Understanding

Interacting with environments & **reasoning** via common sense

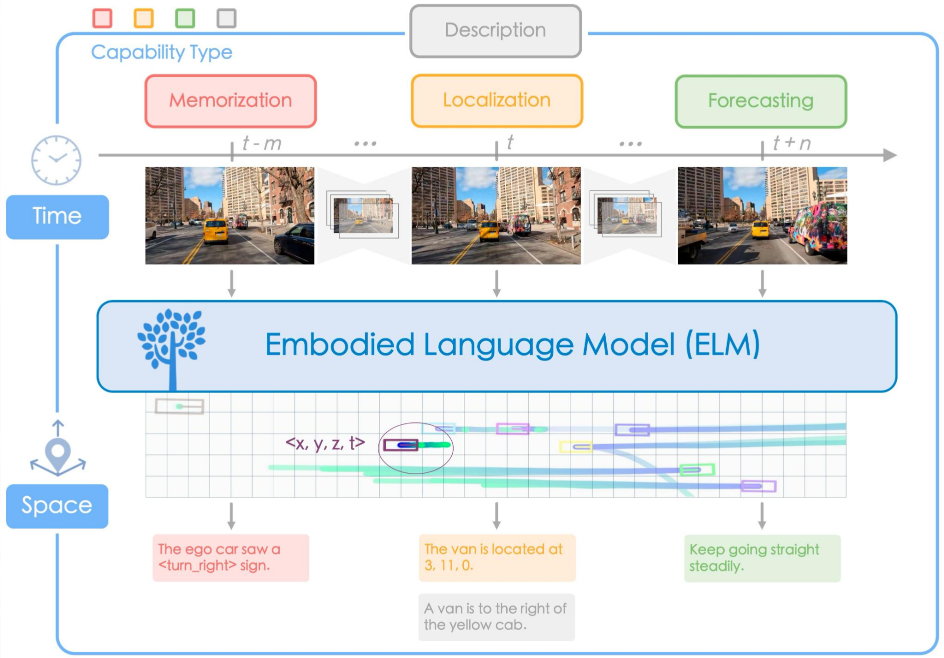
Vision Language Models

Only focus on 2D domain, *i.e.*, **description**

Embodied Language Model

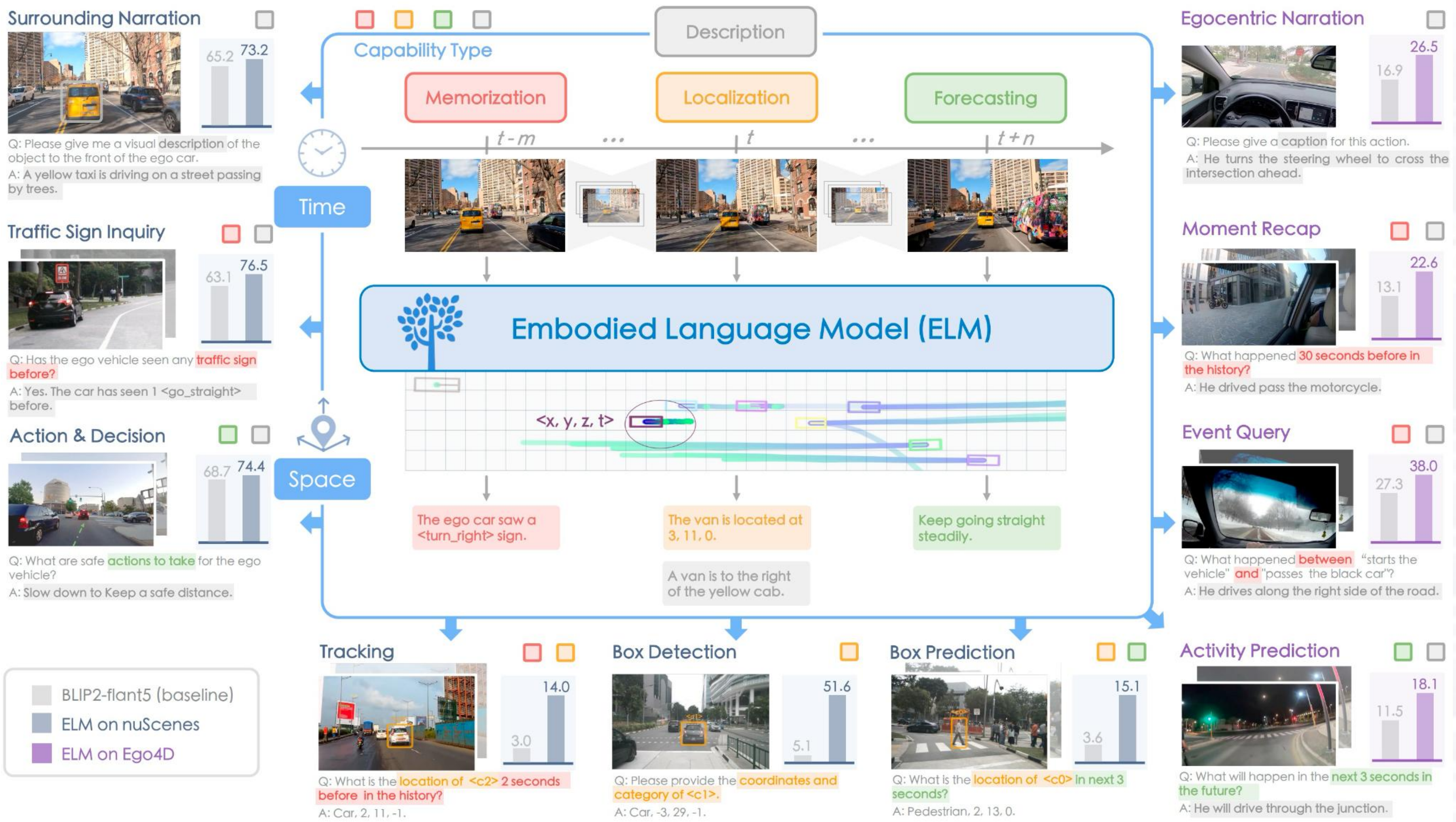
Expanding Vanilla VLMs to Driving Scenes

- Task: embodied understanding of driving scenarios.
- Capabilities: description, **localization**, **memorization**, **forecasting**.
- Model: **ELM** with long-horizon **space** and **time**.
- Benchmark: A spectrum of tasks in an embodiment setting.



Embodied Understanding

At A Glance



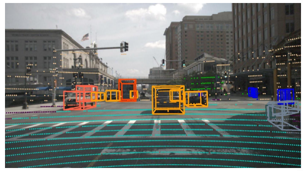
We expand a wide spectrum of **new tasks** to fully leverage large language models in an embodiment setting

ELM - The Big Picture

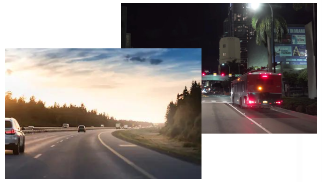
Data-centric Pipeline

Data Collection

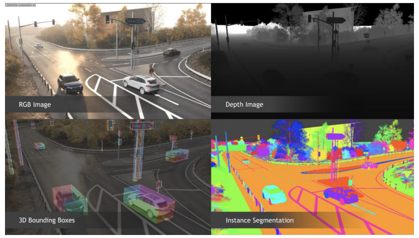
Motional
ARGO AI
lyft



YouTube

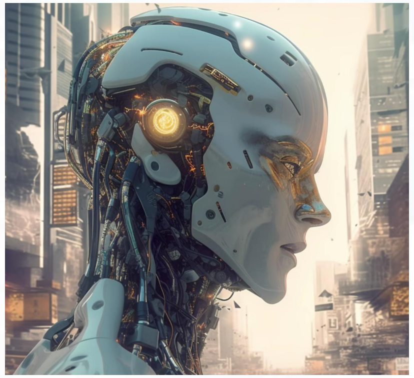


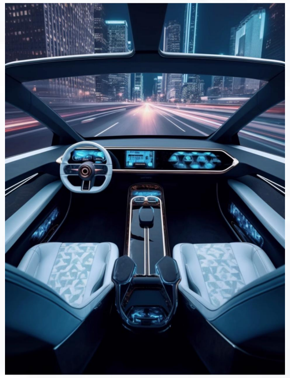
Data Generation



Partial photo by courtesy of online resources.

Pre-training DriveCore





Universal Foundation Model
for autonomous driving

How to formulate?
What's the objective goal?
Space and **time**?

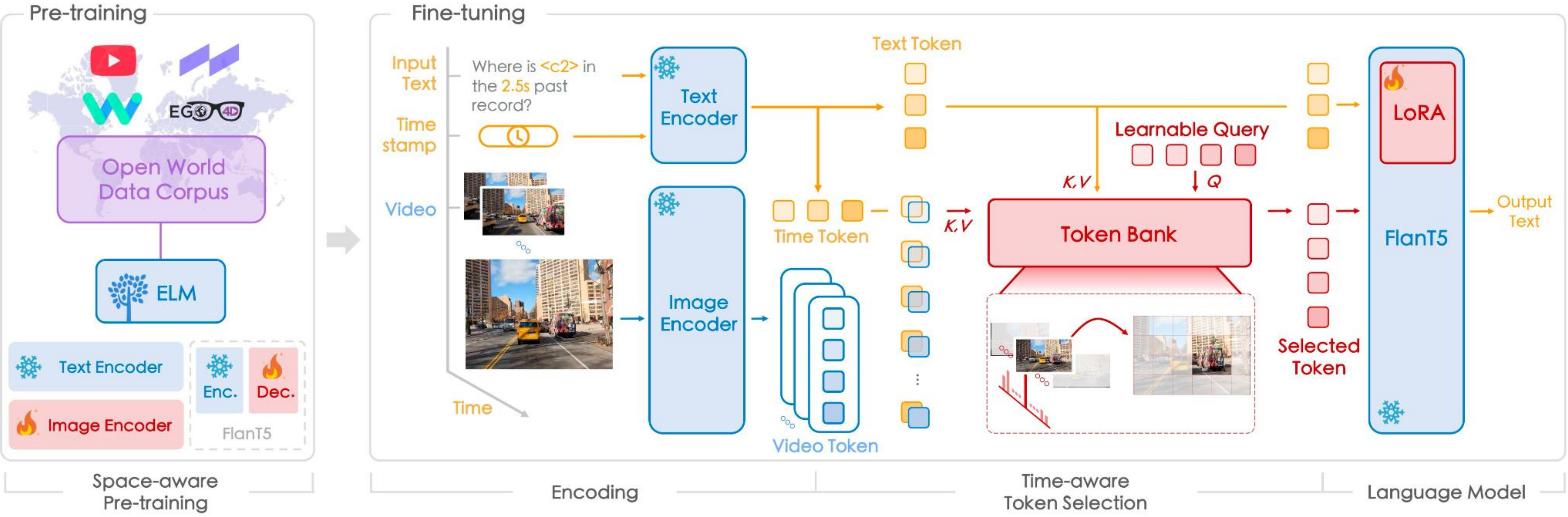
Benchmark

Autonomous Driving (nuScenes)

Robotics (Ego4D)

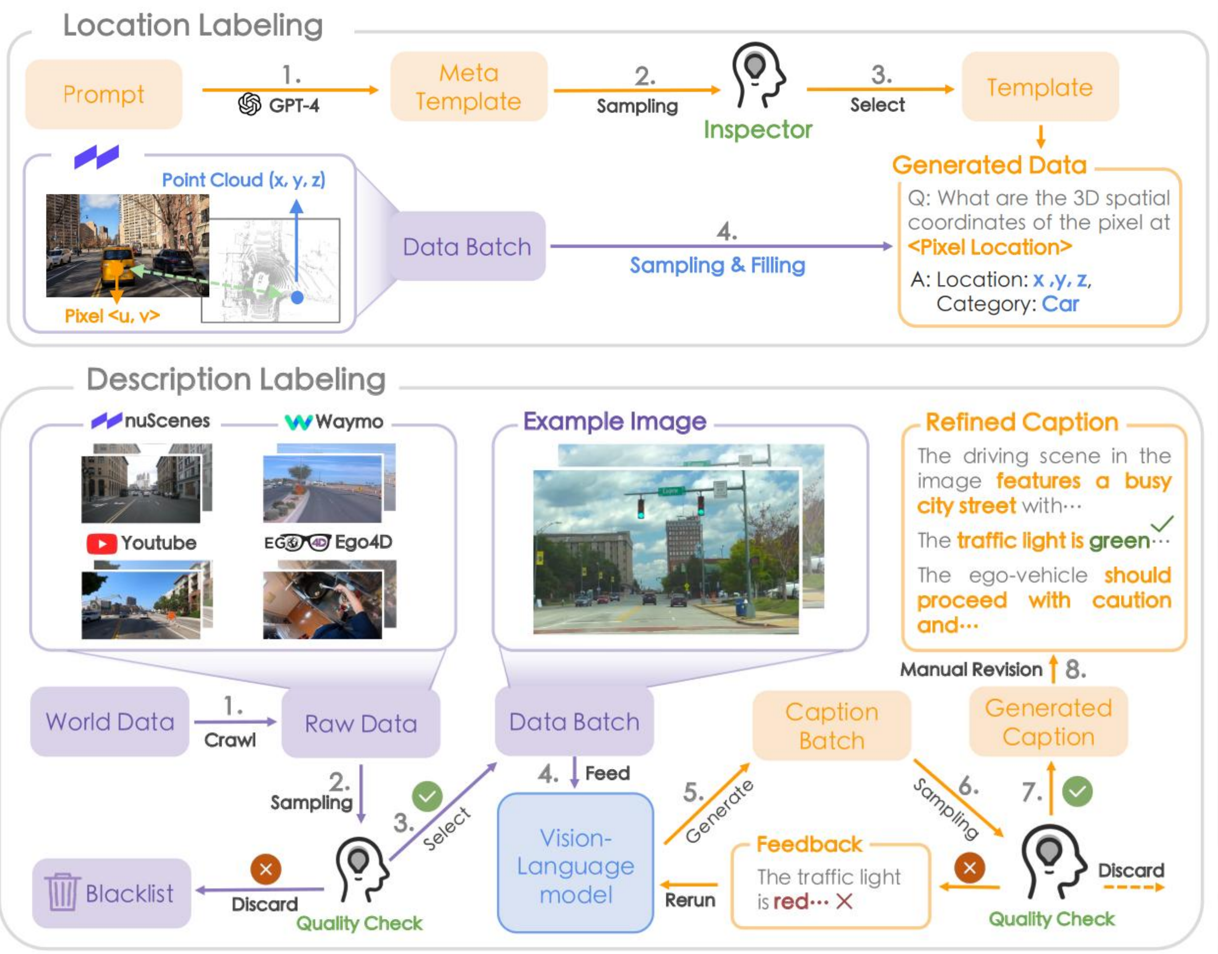
Embodied Understanding

Pipeline



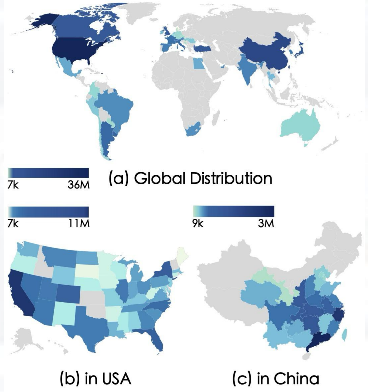
Embodied Understanding

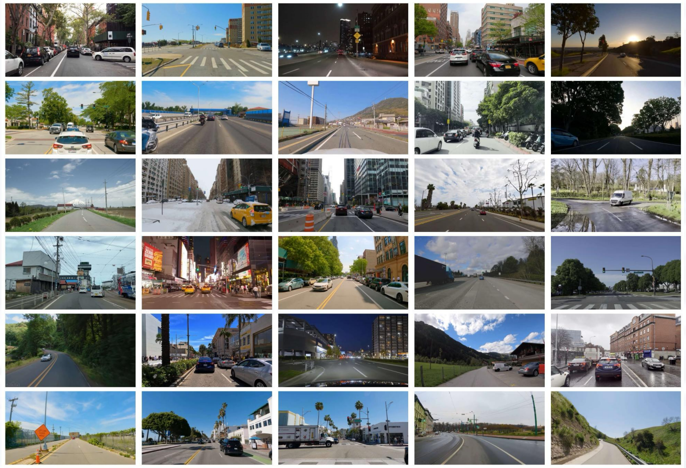
Data Collecting



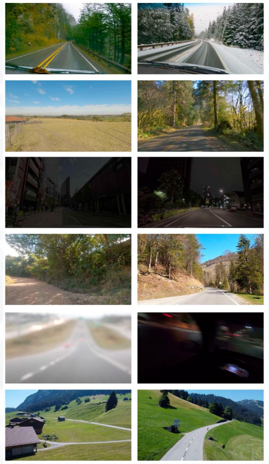
Data Status

Method	Pre-train Data	#	Time	City	Anno
LLaVA [48]	COCO [47]	150K	-	-	Des
VideoChat [44]	Self-Collected	18K	-	-	Des
Vid-ChatGPT [56]	ActivityNet-200 [6]	100K	-	-	Des
nuScences-QA [61]	nuScenes [7]	460K	5.5	2	Des
DriveGPT4 [82]	BDD-X [38]	28K	77	4	Des
LLM-driver [11]	Self-Collected	160K	-	-	Des
ELM (Ours)	nuScenes [7]	7.4M	5.5	2	Des, Loc
	Waymo [70]	450K	6.4	3	Des
	YouTube	1.1M	1474	700	Des
	Ego4D [25]	300K	1638	74	Des



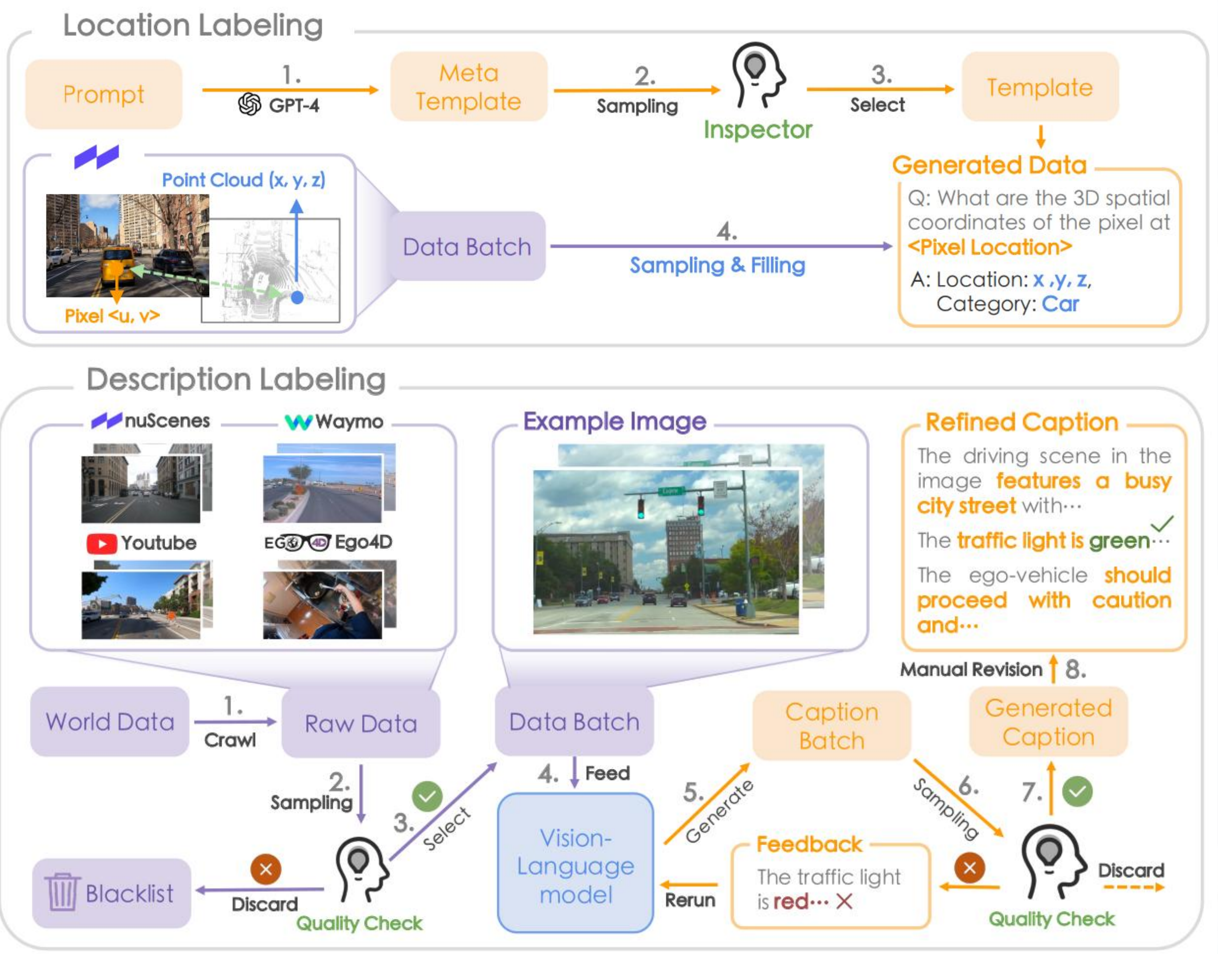


Discarded

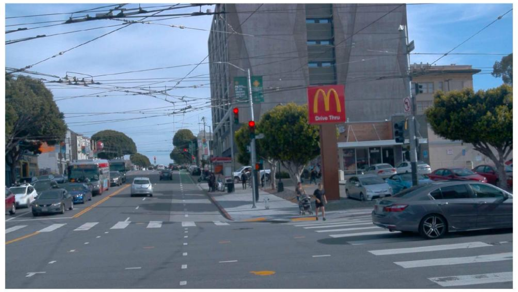


Embodied Understanding

Annotation - Quality check



Examples of Description Labels



Raw: The scene is a busy city street with a red traffic light, a McDonald's restaurant, and a crosswalk. The traffic elements include traffic lights, cars, and pedestrians. The ego-vehicle is traveling in the left lane, and the traffic light is red···

Refined: The scene is a busy city street with a red traffic light, a McDonald's restaurant, and a crosswalk. The traffic elements include traffic lights, cars, and pedestrians. The ego-vehicle is traveling in the right lane, and the traffic light is red···

Raw: The driving scene in the image features a car driving under a bridge, with a truck carrying a load of wood in the background. The car is traveling in the left lane, and it appears to be moving in the same direction as the truck.

Refined: The driving scene in the image features a car driving under a bridge, with a truck carrying a load of wood in the background. The car is traveling in the middle lane, and it appears to be moving in the same direction as the truck.

Q: How should the ego-vehicle drive at this point?

A: The ego-vehicle should follow the traffic light's instructions and **wait for the light to turn green before proceeding.**

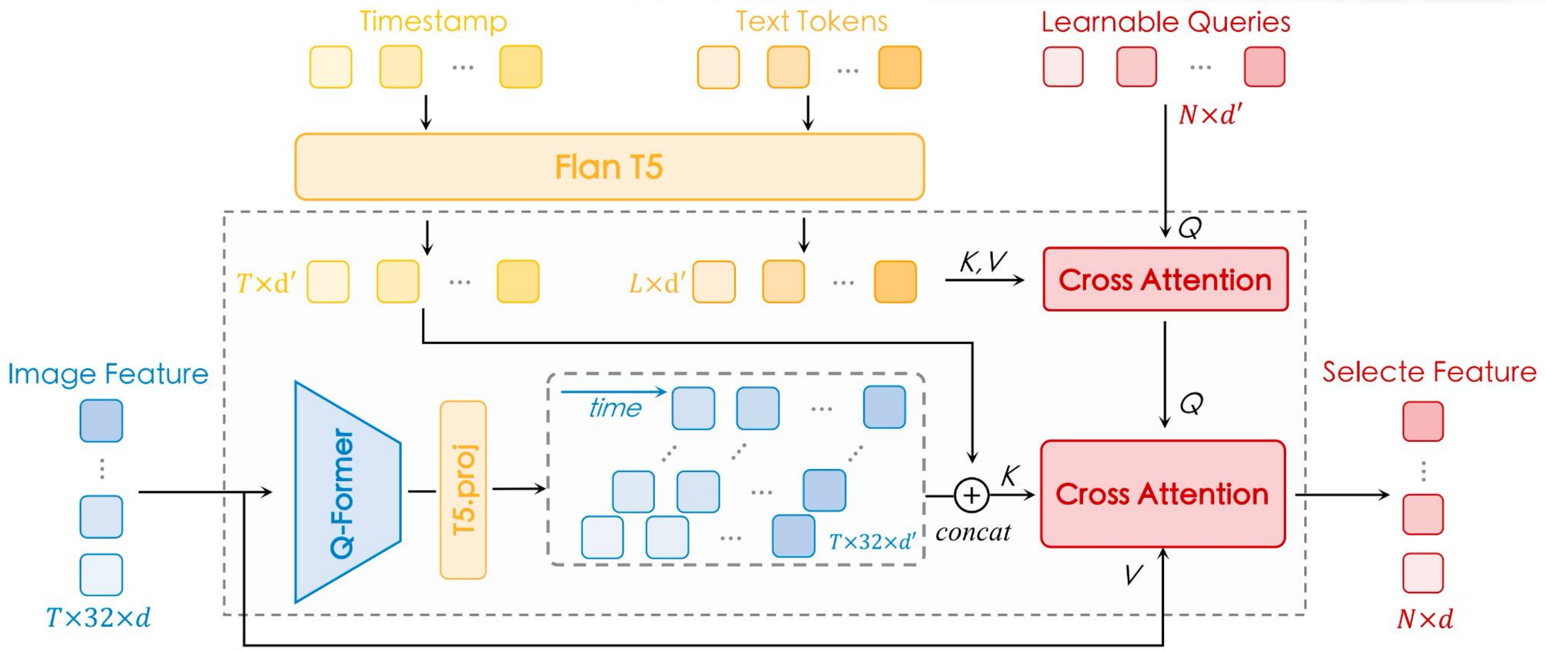
Q: Are there any potential dangers in the scenario?

A: The ego-vehicle should continue **driving through the intersection**, following the traffic light's instructions. It is important to **be cautious and attentive to the pedestrians to ensure safety for all involved.**

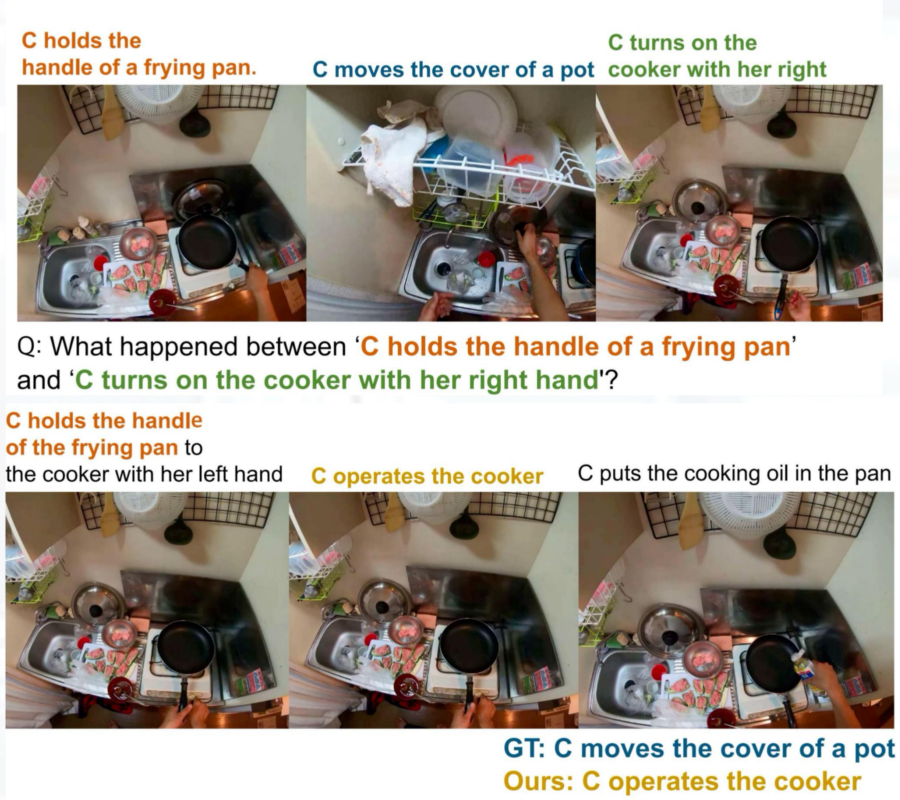
Embodied Understanding

Time-Aware Token Selection

- Utilize text prompts as guidance to select tokens wisely.
- Alignment among text, timestamps, and videos.



Examples of Selection



OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

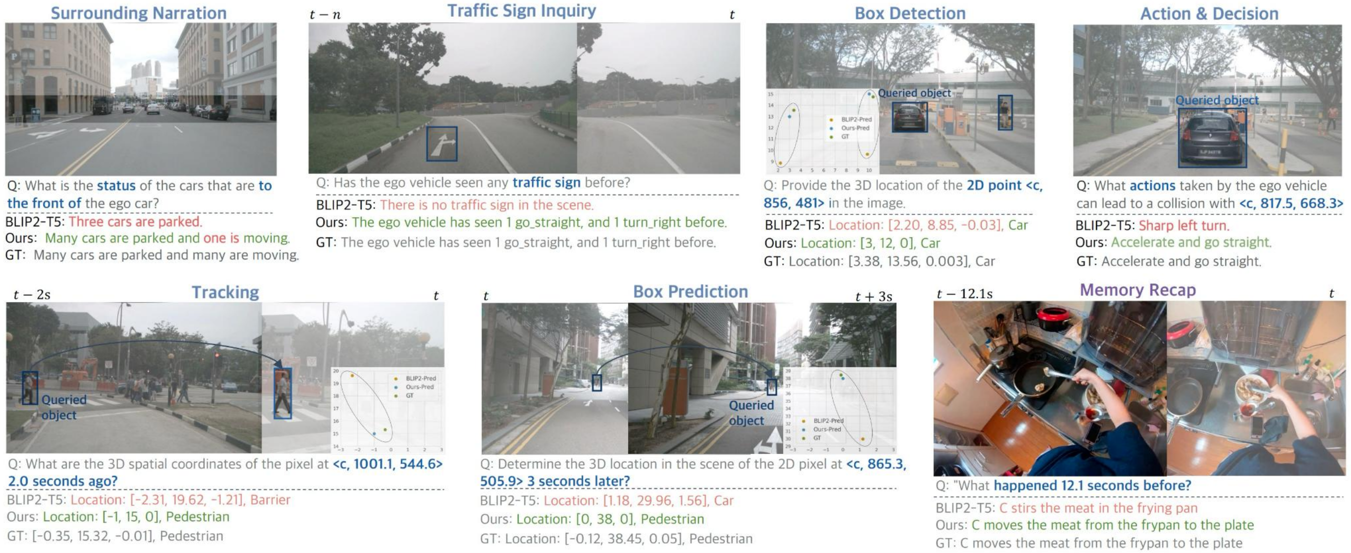
Experiments

Embodied Understanding

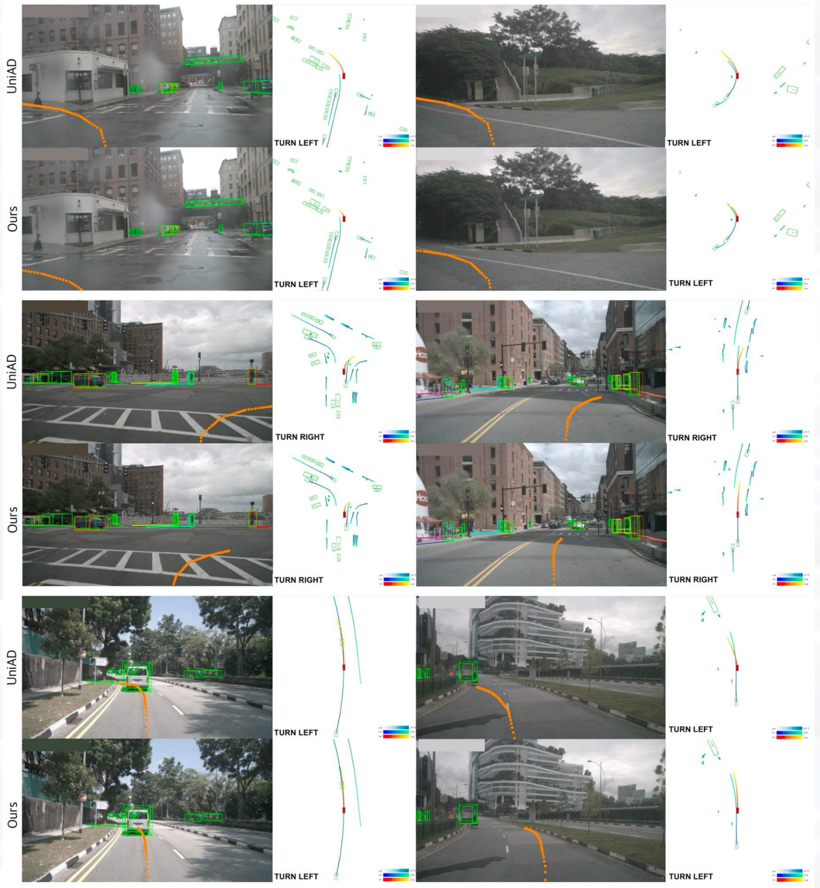
Benchmark

- Performing tasks for embodied understanding .

Tasks	Fine-tune Dataset	Capability				Statistics		
		Description	Localization	Memorization	Forecasting	S(m) / R(m)	T (s) / F	#
Surrounding Narration	nuScenes [7]	✓				30 / 5	0.5 / 1	142K
Traffic Sign Inquiry		✓		✓		30 / 1	3.5 / 7	20K
Action & Decision		✓			✓	30 / 5	3.5 / 7	301K
Box Detection			✓			50 / 1	0.5 / 1	232K
Tracking			✓	✓		50 / 1	3.5 / 7	131K
Box Prediction			✓		✓	50 / 1	3.5 / 7	133K
Egocentric Narration	Ego4D [25]	✓				20 / 3	3 / 1	357K
Moment Recap		✓		✓		20 / 3	60 / 20	70K
Event Query		✓		✓		20 / 3	60 / 20	70K
Activity Prediction		✓			✓	20 / 3	60 / 20	69K



End-to-end Planning

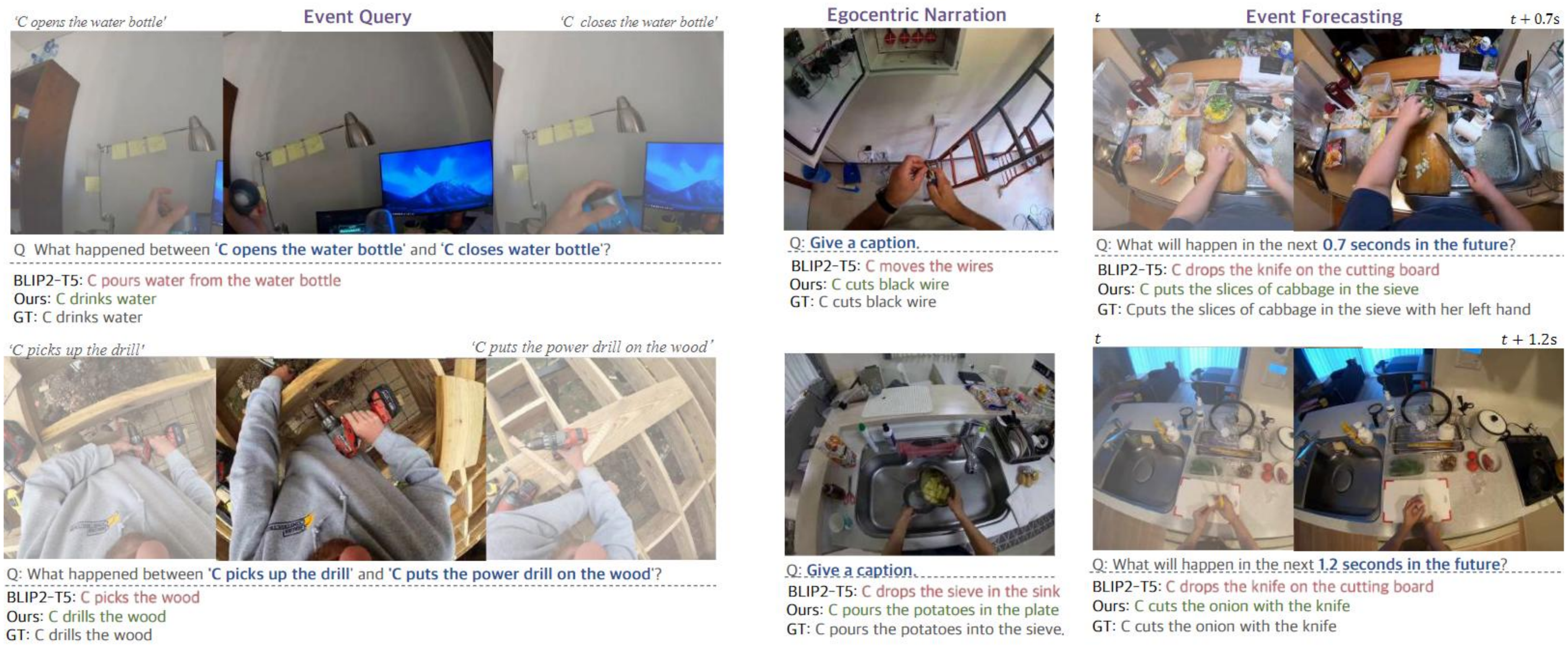


Embodied Understanding

Benchmark

Embodied Understanding

Benchmark



Embodied Understanding

Experiments

Methods	Tracking		Box Detection		Box Prediction		Traffic Sign Inquiry			Surrounding Narration			Action & Decision		
	Pr@1	Pr@2	Pr@1	Pr@2	Pr@1	Pr@2	C	R	B	C	R	B	C	R	B
BLIP2-opt [27]	0.1	0.1	0.1	0.2	0.2	0.5	23.0	26.9	20.5	8.1	19.7	21.2	8.4	11.5	11.1
BLIP2-flant5 [27]	3.0	6.0	5.1	10.5	3.6	6.3	63.1	39.4	31.4	65.2	64.9	27.9	68.7	71.4	43.1
LLaMA-Ada. [17]	6.1	10.5	8.3	14.9	7.5	12.5	68.3	66.6	61.6	67.0	77.5	**60.1**	72.3	76.8	**64.7**
LLaVA [32]	5.5	9.3	28.5	31.2	6.1	10.2	51.1	58.5	50.8	64.9	64.6	41.2	64.4	62.4	57.9
Otter [26]	10.0	17.2	41.8	46.9	8.9	15.8	62.8	41.1	32.4	60.0	64.2	13.3	69.2	73.2	53.0
VideoChat [28]	0.4	0.9	0.1	0.3	0.1	0.2	25.3	21.9	11.7	21.7	29.2	12.2	29.6	33.2	13.1
Vid-ChatGPT [36]	0.1	0.6	0.1	1.0	0.3	1.2	49.6	57.1	48.6	61.0	69.6	37.2	53.6	58.5	43.5
ELM (Ours)	**14.0**	**23.3**	**51.6**	**56.9**	**15.1**	**24.4**	**76.5**	**71.2**	**63.9**	**73.2**	**78.7**	29.8	**74.4**	**83.3**	41.2

(a) **nuScenes.** We outperform the best previous methods on most metrics across the six tasks on nuScenes which validates the generality of our model.

Methods	Moment Recap			Event Query			Egocentric Narration			Activity Prediction		
	C	R	B	C	R	B	C	R	B	C	R	B
BLIP2-opt [27]	1.2	8.9	6.8	7.8	28.4	14.7	5.2	19.8	10.7	2.7	18.7	9.6
BLIP2-flant5 [27]	13.1	31.9	12.5	27.3	33.0	16.6	16.9	33.5	15.4	11.5	31.2	11.3
LLaMA-Ada. [17]	11.2	30.2	12.3	37.5	**47.2**	28.1	18.4	34.2	15.3	13.1	31.2	12.8
LLaVA [32]	9.6	28.3	12.1	**39.8**	44.6	**29.9**	6.5	28.2	11.6	8.4	28.0	13.0
Otter [26]	11.4	29.6	10.5	27.1	38.3	19.1	14.1	31.4	13.9	11.1	29.4	10.3
VideoChat [28]	13.2	32.5	13.8	34.5	42.2	26.4	20.7	35.0	**17.6**	12.1	32.4	14.1
Vid-ChatGPT [36]	10.0	31.1	13.3	27.9	36.5	20.9	10.2	21.7	10.4	9.4	30.5	12.6
ELM (Ours)	**22.6**	**36.7**	**19.4**	38.0	43.1	27.6	**26.5**	**37.7**	16.9	**18.1**	**34.1**	**17.0**

(b) **Ego4D.** We extended the model to Ego4D dataset and verified the generality of our token bank module on four tasks.

Methods	# param
BLIP2-opt	2.7B
BLIP2-flant5	2.7B
LLaMA-Ada.	7B
LLaVA	7B
Otter	7B
VideoChat	7B
Vid-ChatGPT	7B
ELM (Ours)	2.7B

(c) **Adopted LLM params.**

OOD Evaluation

Method	ADE↓	FDE↓	Time(s)↓
Command Mean	7.98	11.41	-
UniAD-single [34]	4.16	9.31	0.56
Flamingo [3]	2.78	5.31	1.47
ELM	**2.28**	**4.27**	1.61

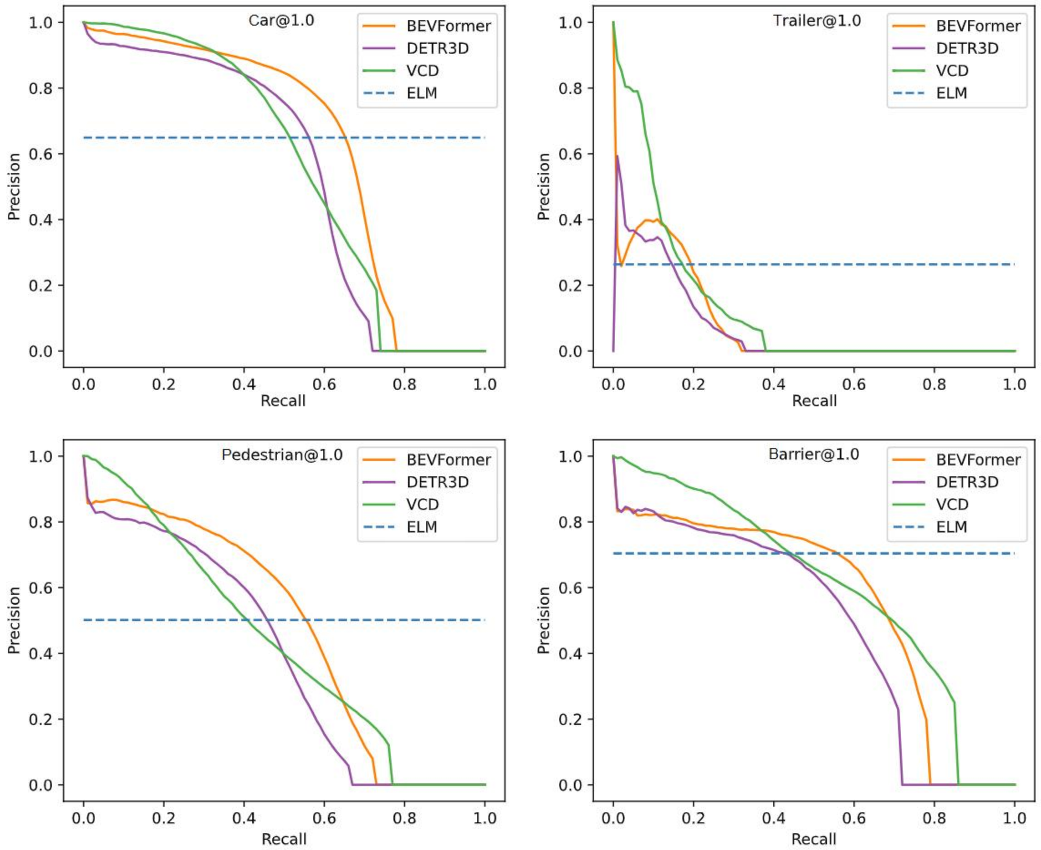
Q: What is the unusual about the driving scene?

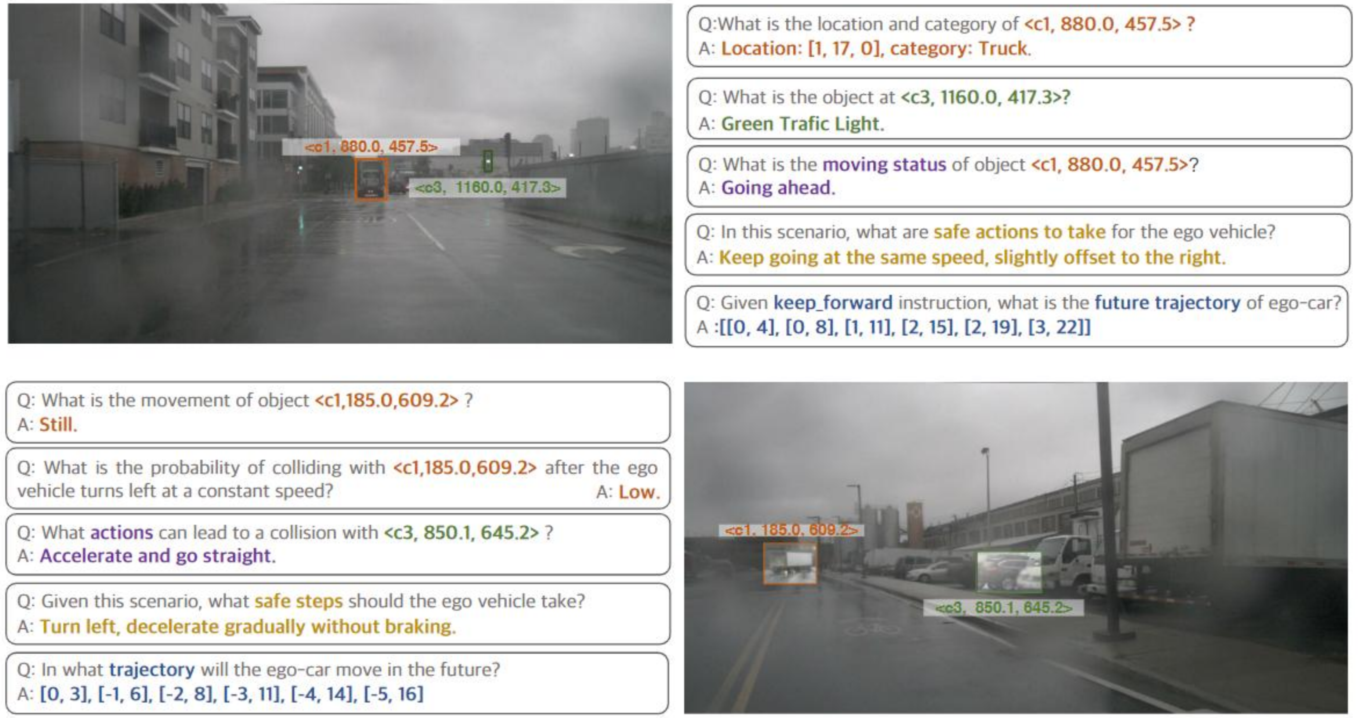
A: The scene shows a busy city street with a construction site on the side of the road. There is a traffic light visible in the background. The unusual element in this scene is the presence of a construction site, which may cause traffic disruptions or delays.

OpenDriveLab

Embodied Understanding

Experiments





Embodied Understanding

Limitations

Driving-specific Inputs

ELM cannot handle common setting such as LiDAR or multi-view images as input, limiting its information source.

Closed-loop Planning

ELM is evaluated under an open-loop scheme, while closed-loop planning is necessary to see if it can handle corner cases.

Efficiency Constraints

Inheriting the drawbacks of LLMs, ELM suffers from long inference time, which may impact practical implementation.

One-page Takeaway

- End-to-end Autonomous Driving
  - Challenge: **Generalization & Explainability**
  - Recent trend: use vision language model to **embed "world knowledge"** to solve the challenges.

- ELM: Embodied Understanding of Driving Scenarios
  - Revive driving scene understanding by delving into **embodied** settings, along with capacities, tasks, and rubrics.
  - Expand vanilla VLMs to process long horizon **space** and **time** (open-world data & module design).

OpenDriveLab

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

END