OpenDriveLab

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

End-to-end Autonomy

Dr. Hongyang Li

OpenDriveLab at Shanghai AI Lab

Mar 20, 2024

Outline

- **端到端自动驾驶概述**
  - 模块化设计 vs 端到端背景
  - 工业界应用
  - 研究时间线
  - 端到端应用场景
- **主流工作选讲**
  - 第一组：WoR / NEAT / UniAD / DriveAdapter 等
  - 第二组：GenAD / ViDAR / ELM / DriveLM 等
  - 第三组：GAIA-1 / EgoStatus / Panacea
- **当前挑战**
  - 泛化能力、多模态等挑战（8种）
- **未来工作**
  - 与大模型、世界模型等内容结合

OpenDriveLab

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

End-to-end Autonomous Driving

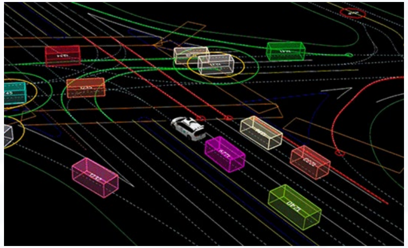
An Introduction

Autonomous Driving (AD) Tasks

Challenge | Various weathers, illuminations, and scenarios

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

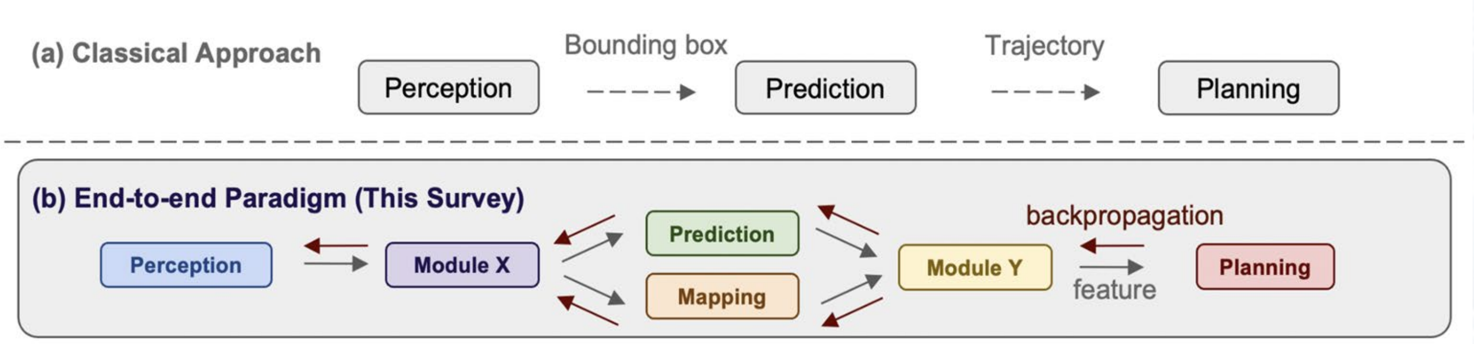
What are around?



How will they go in the future?

Where should I go?

回顾：Why end to end?



https://github.com/OpenDriveLab/End-to-end-Autonomous-Driving

端到端自动驾驶系统：

- 将原始传感器数据作为输入
- 输出轨迹规划，或低级别的控制信号

回顾：Why end to end?

优势

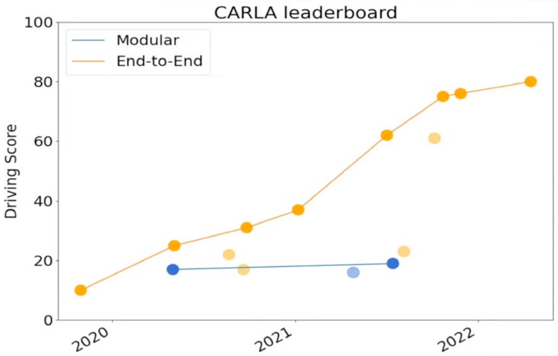
+ 将所有模块合并为一个可**联合训练**的单一模型带来的**便利性**
+ 避免模块化设计带来的级联错误
+ 直接**针对最终任务进行优化（规划/轨迹预测）**
+ 计算效率高 (共享 backbone), 对最终产品友好

回顾：Why end to end?

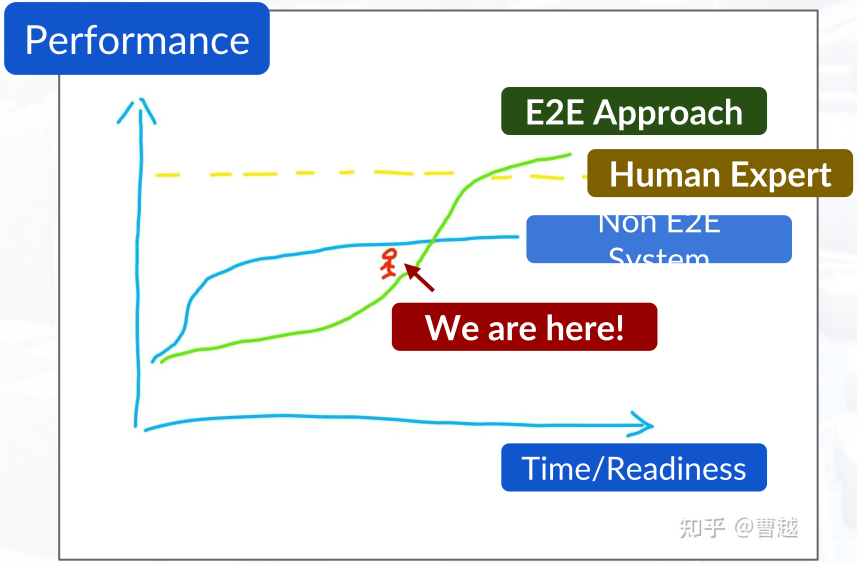
劣势

- 只能在模拟器和机载测试中进行闭环评测(Closed-loop evaluation)
- 缺少真实世界数据
- 可解释性差

E2E vs Non-E2E

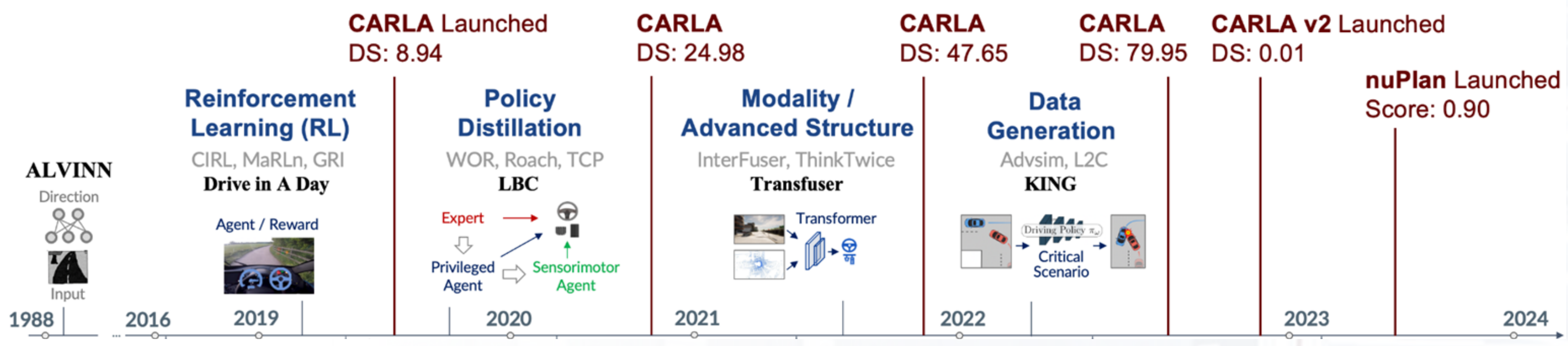


Credit to Andreas Geiger @ CVPR Workshop 2023



Credit to Dr. Yue Cao @ Zhihu

Roadmap | End-to-end Autonomous Driving



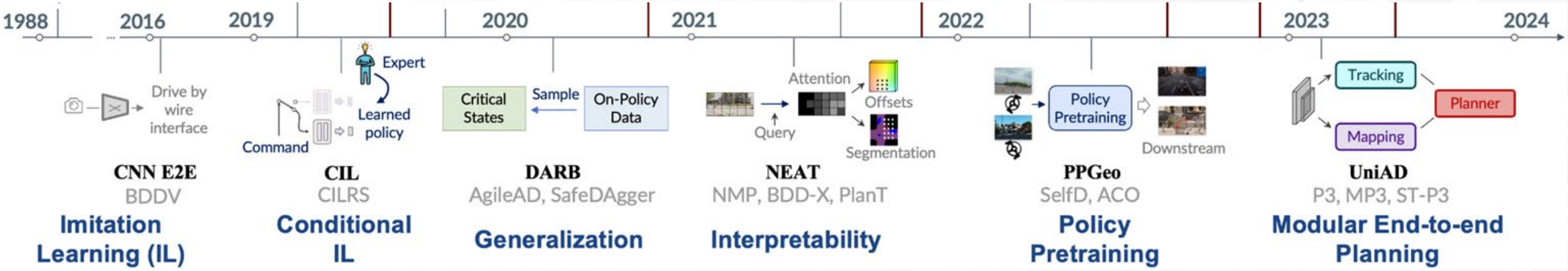
Summary (1/2)

- 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



Trending | End-to-end Autonomous Driving

Industry

TESLA

E2E Vehicle

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

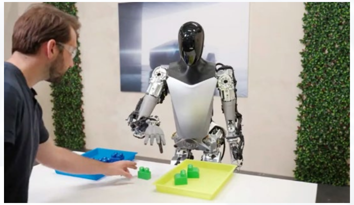
Elon Musk @elonmusk · Aug 26
twitter.com/i/broadcasts/1...

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

E2E Robot



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

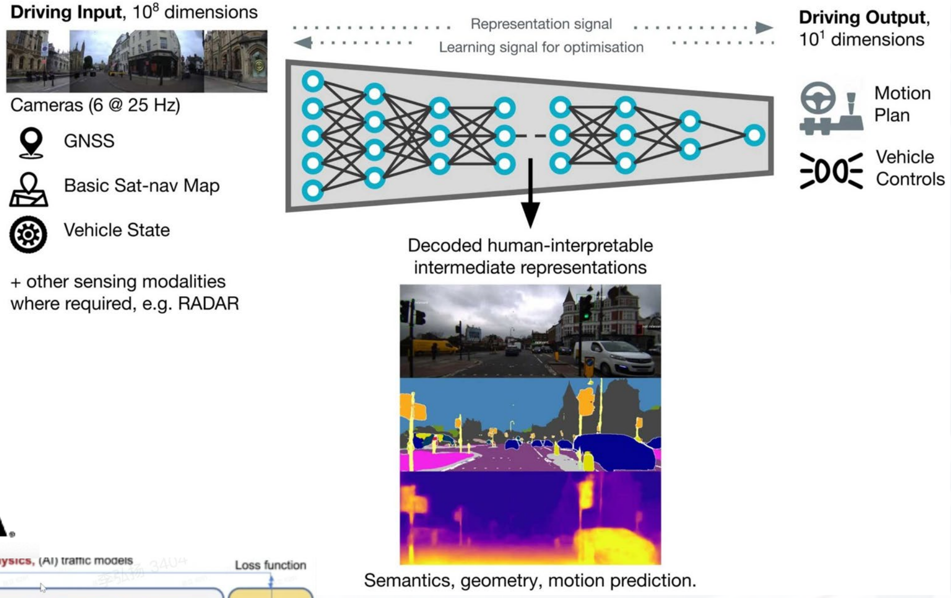
Its neural network is trained fully **end-to-end**: video in, controls out.

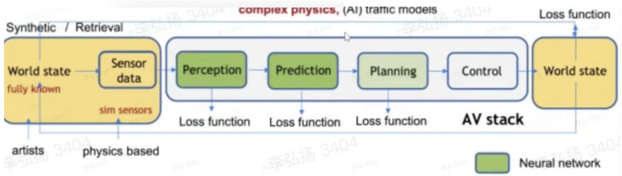
Trending | End-to-end Autonomous Driving

Industry

And many others ...

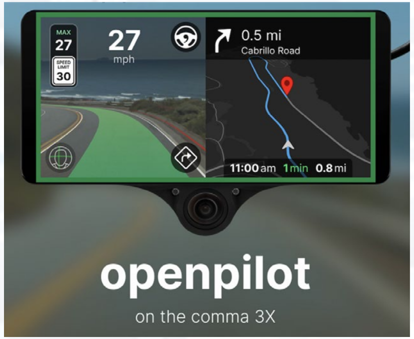
W A Y V E





comma.ai

- *Openpilot is an open source driver assistance system.*
- *Openpilot performs the functions of Automated Lane Centering (ALC) and Adaptive Cruise Control (ACC) for 250+ supported car makes and*



https://arxiv.org/abs/2206.08176

工业界应用

附件4 内容

Public Opinions on Our Survey

- **Paper**
  https://arxiv.org/pdf/2306.16927.pdf
- **Repo (paper collection)**
  https://github.com/OpenDriveLab/End-to-end-Autonomous-Driving

Alex Kendall
@alexgkendall

This is a fantastic, comprehensive and forward-looking survey of academic literature about end-to-end machine learning for autonomous driving. It is a very timely publication as the field is exploding with interest right now.

I'm aligned with the paper's conclusions on open algorithmic challenges. There's loads of insight around opportunities like world modelling, language, foundation models and long-tail robustness. This paper also exposes that academic literature under-appreciates significant industry challenges right now, such as (1) safety, reward modelling and policy alignment against human expectations and risk, or (2) the significance of establishing a synthetic/real-world data engine for training/validation, which are critical to the success of any machine learning system. I'd love to see more work in these areas.

Great to see @AutoVisionGroup @francislee2020, well done!

Awesome Vision Group @AutoVisionGroup · Sep 18

Yann LeCun
@ylecun

A nice survey of end-to-end learning methods for autonomous driving.

Awesome Vision Group @AutoVisionGroup · Sep 18
Why are Tesla @elonmusk and Wayve @alexgkendall @Jamie_Shotton moving towards end-to-end autonomous driving? What is the state-of-the-art in this field? With our friends @francislee2020 we recently wrote an extensive survey paper on this emerging topic: arxiv.org/abs/2306.16927

202339B 端到端自动驾驶

Original 吴双 吴言吴语 2023-10-02 05:42

收录于合集
#自动驾驶 20个 >

这周我们读一篇提交到PAMI的端到端自动驾驶的综述论文:

SUBMITTED TO IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, JUNE 2023

End-to-end Autonomous Driving: Challenges and Frontiers

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger and Hongyang Li

Arxiv链接：https://arxiv.org/abs/2306.16927
可以看到这篇文章在六月份，好像是CVPR会议期间就挂到了arxiv上，当时眼前一亮随手放在了桌面，结果回头就忘了，最近SS兄提醒，就给自己安排了周末作业。由于论文覆盖的内容很多，今天就只聊一聊我个人看到的值得注意或者觉得需要强调的点。

总结：很好的综述，值得看看。

Join Slack Discussions!

https://join.slack.com/t/opendrivelab/shared_invite/zt-244lgu87b-eLonLQzle4wRkg8W8WOUlg

OpenDriveLab

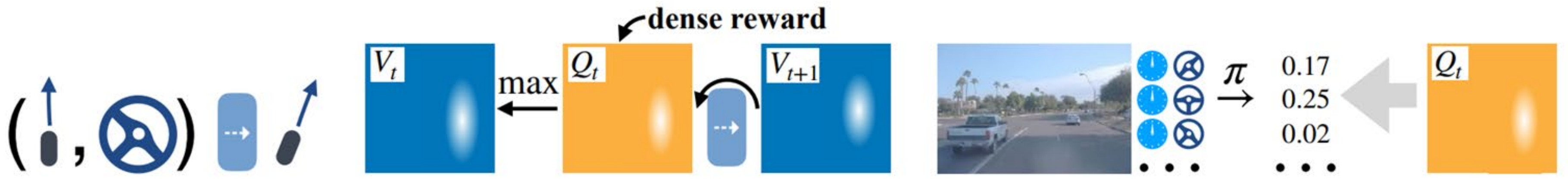
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

主流工作选讲 - Part 1

ST-P3 / PPGeo / NEAT / WoR

主流方法：World on Rails

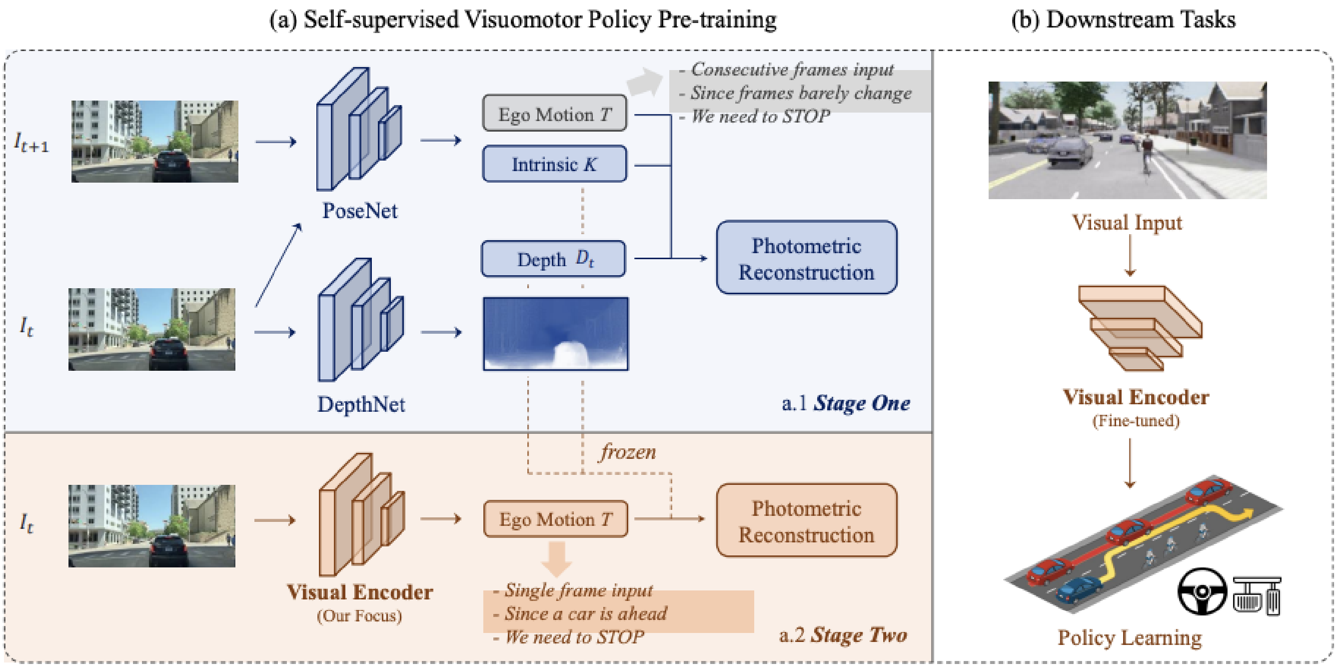
https://arxiv.org/abs/2105.00636



(a) Forward model (b) Bellman update (c) Distillation

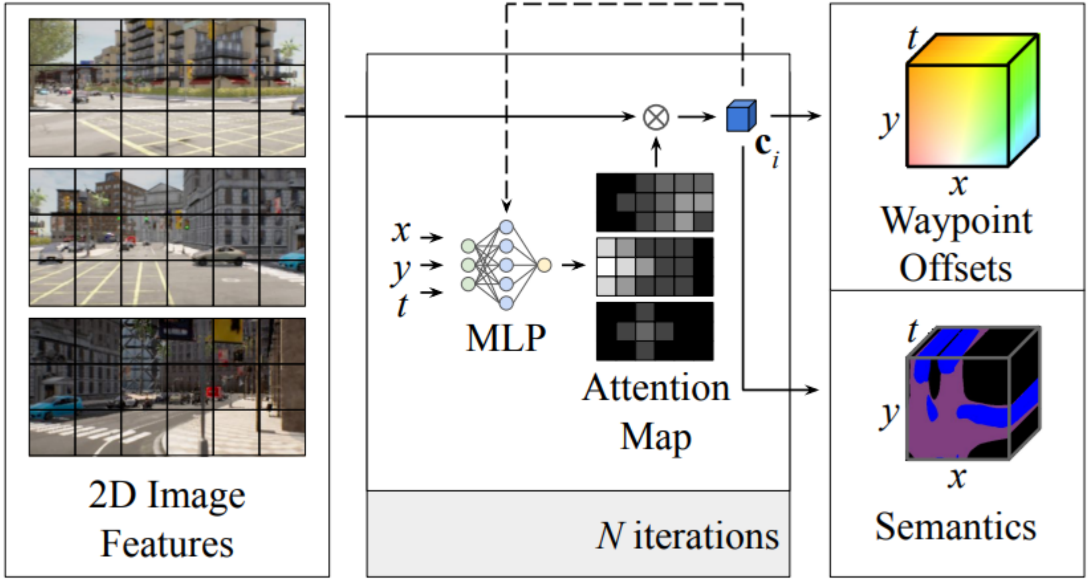
主流方法：PPGeo

https://arxiv.org/pdf/2301.01006.pdf



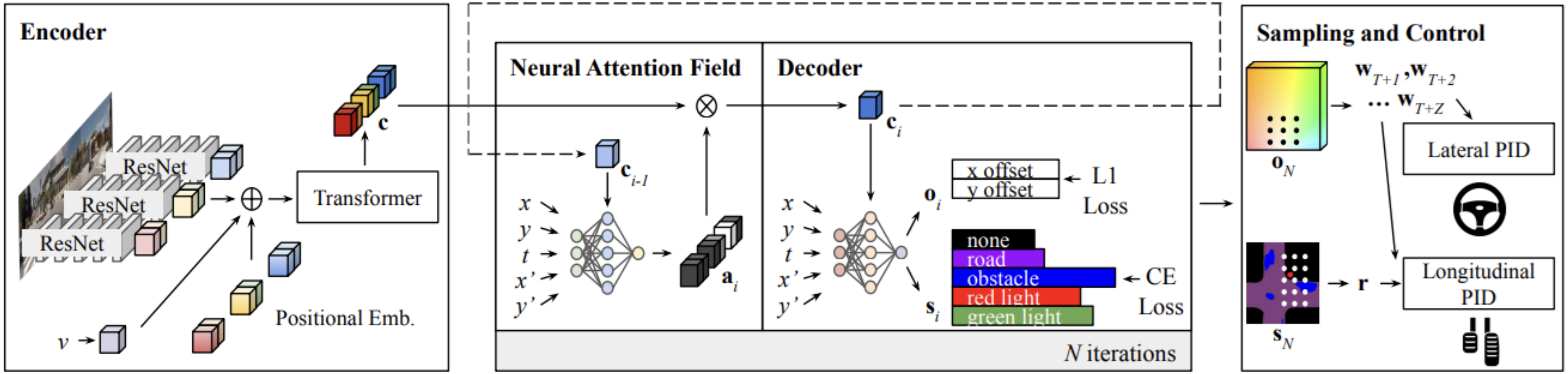
主流方法：NEAT

https://arxiv.org/pdf/2109.04456.pdf



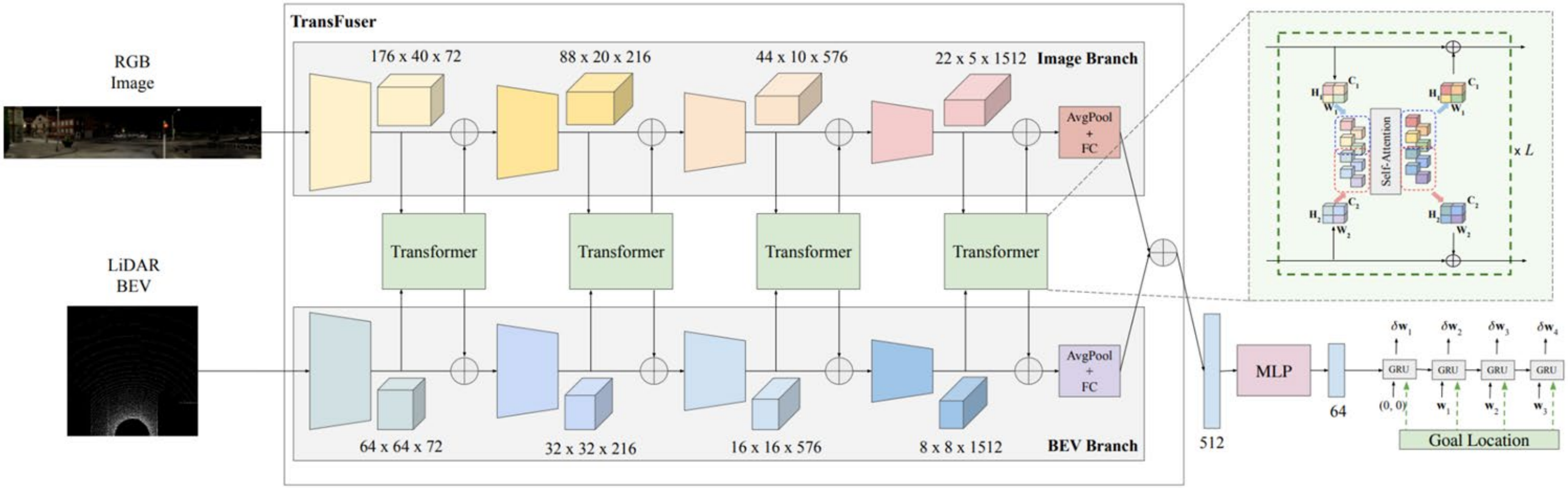
主流方法：NEAT

https://arxiv.org/pdf/2109.04456.pdf



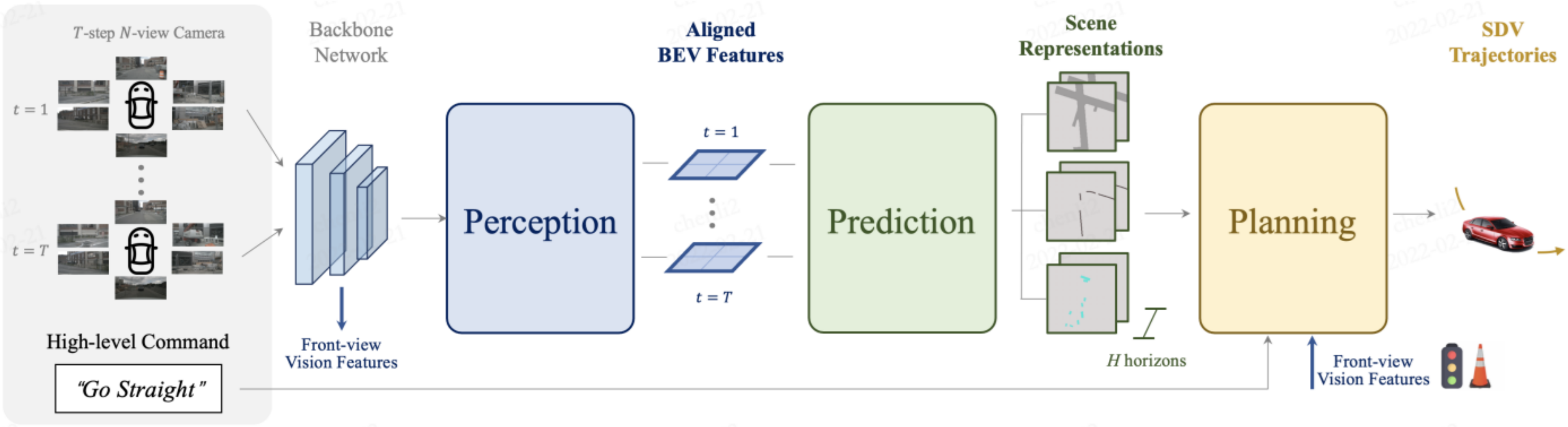
主流方法：TransFuser

https://arxiv.org/abs/2205.15997



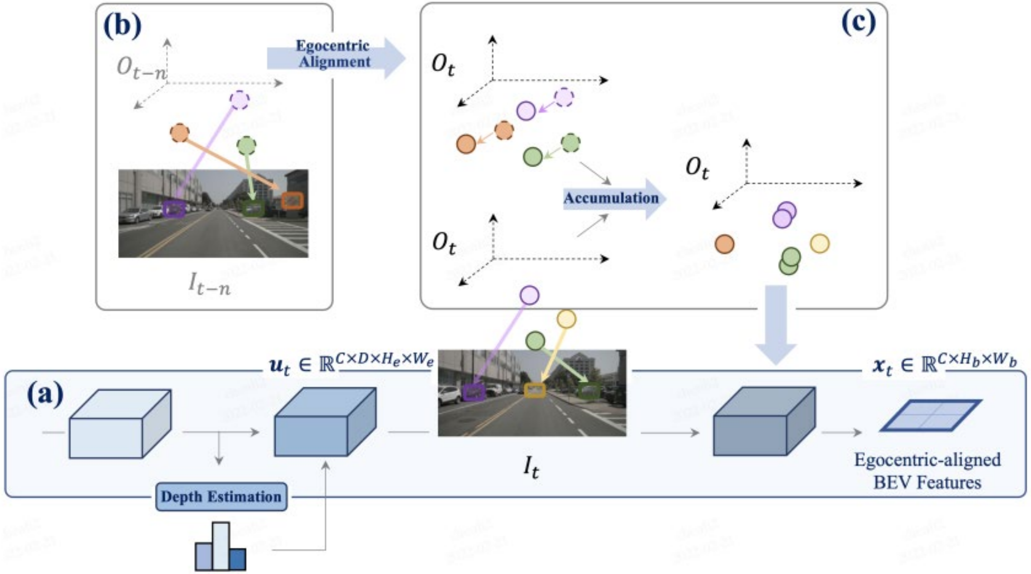
主流方法：ST-P3

https://arxiv.org/abs/2207.07601



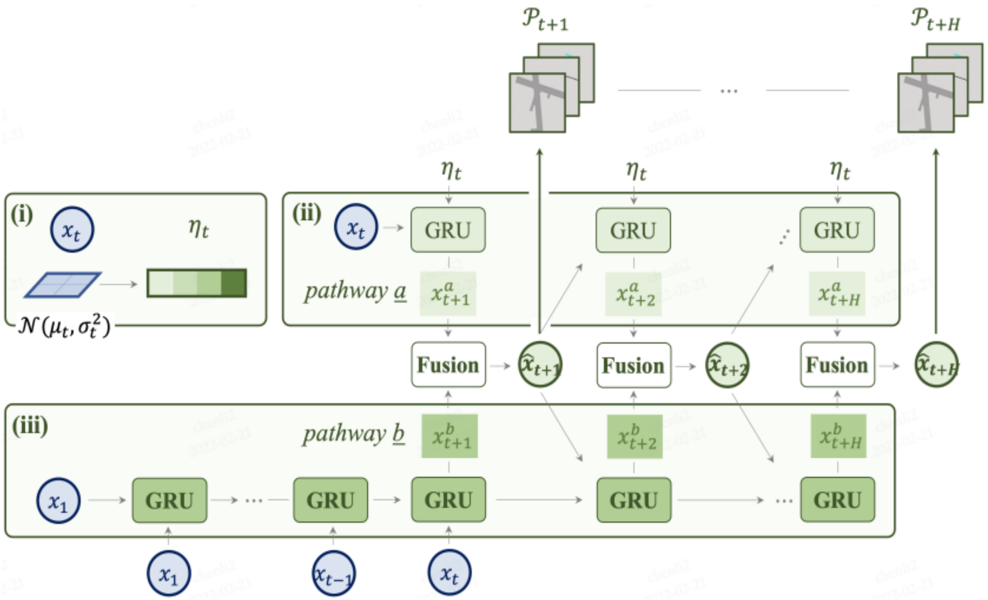
主流方法：ST-P3

https://arxiv.org/abs/2207.07601



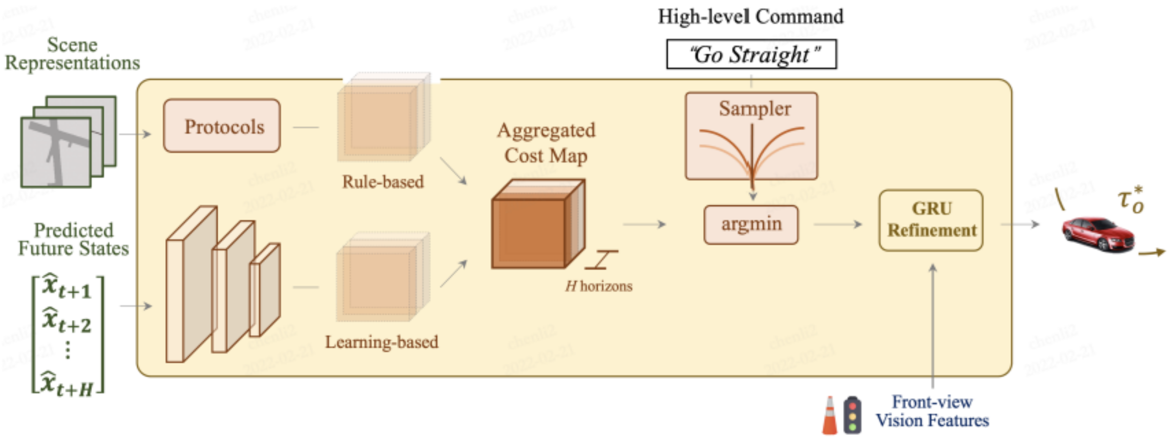
主流方法：ST-P3

https://arxiv.org/abs/2207.07601



主流方法：ST-P3

https://arxiv.org/abs/2207.07601

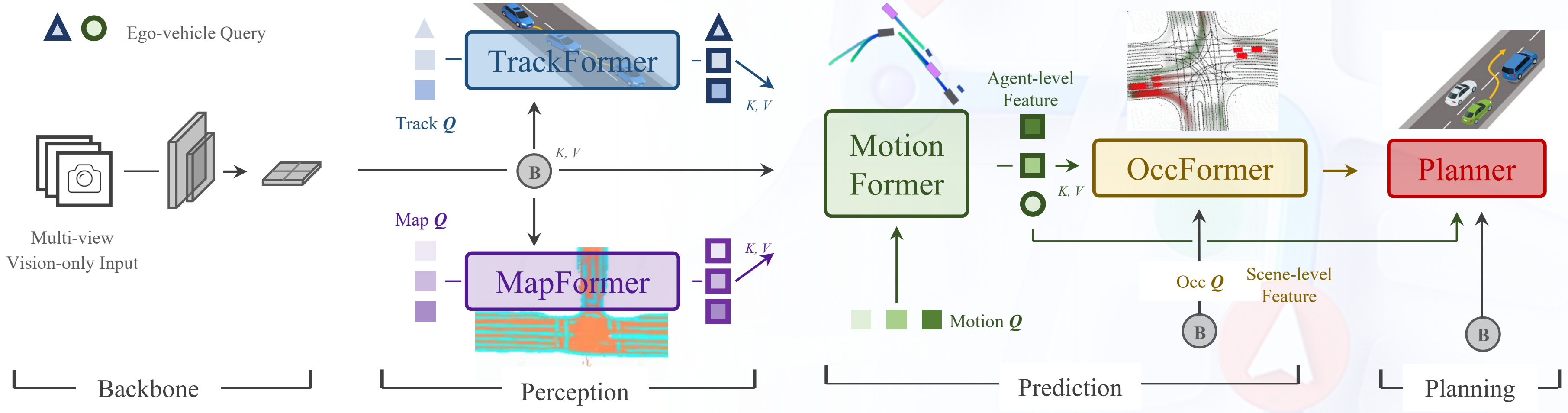


CONTENTS

A quick recap on JUNE 18-22, 2023 CVPR VANCOUVER, CANADA *Best Paper Award*

Planning-oriend Autonomous Driving

UniAD - Pipeline



- Entire pipeline connected by queries
- Tasks coordinated with queries
- Interactions modeled by attention

Unified Query

Transformer-based

First time to unify full-stack AD tasks!

UniAD - Ablation Results

Tasks benefit 🚀 each other and contribute to safe planning

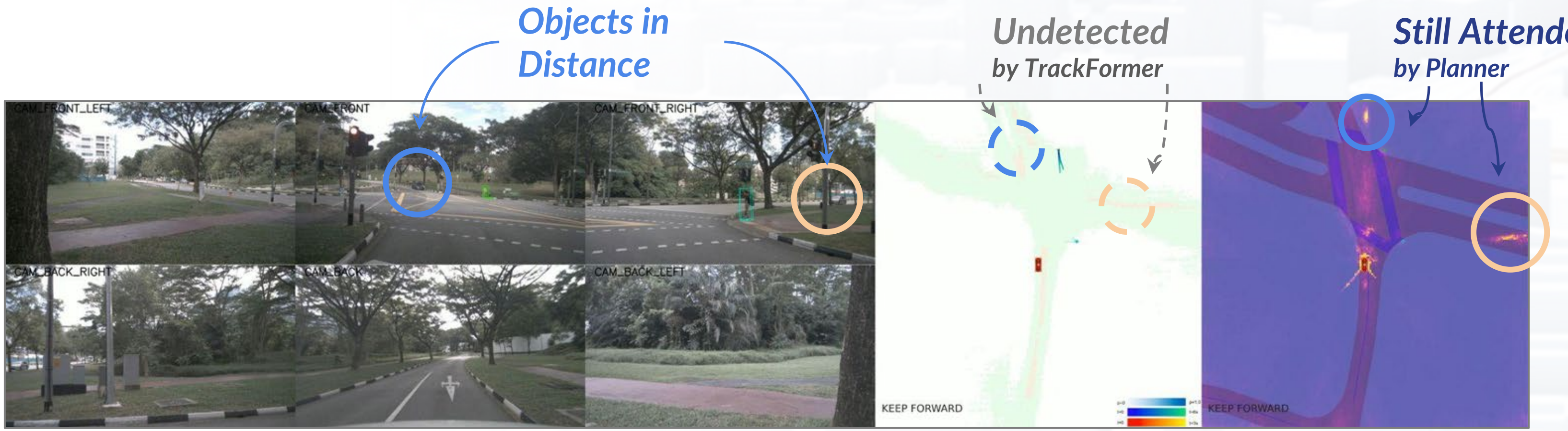
ID	Modules					Tracking			Mapping		Motion Forecasting			Occupancy Prediction				Planning	
	Track	Map	Motion	Occ.	Plan	AMOTA↑	AMOTP↓	IDS↓	IoU-lane↑	IoU-road↑	minADE↓	minFDE↓	MR↓	IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑	avg.L2↓	avg.Col.↓
0*	✓	✓	✓	✓	✓	0.356	1.328	893	0.302	0.675	0.858	1.270	0.186	55.9	34.6	47.8	26.4	1.154	0.941
1	✓					0.348	1.333	791	-	-	-	-	-	-	-	-	-	-	-
2		✓				-	-	-	**0.305**	0.674	-	-	-	-	-	-	-	-	-
3	✓	✓				0.355	1.336	785	0.301	0.671	-	-	-	-	-	-	-	-	-
4			✓			-	-	-	-	-	0.815	1.224	0.182	-	-	-	-	-	-
5	✓		✓			0.360	1.350	919	-	-	0.751	1.109	0.162	-	-	-	-	-	-
6	✓	✓	✓			0.354	1.339	820	0.303	0.672	0.736(-9.7%)	1.066(-12.9%)	0.158	-	-	-	-	-	-
7				✓		-	-	-	-	-	-	-	-	60.5	37.0	52.4	29.8	-	-
8	✓			✓		0.360	**1.322**	809	-	-	-	-	-	62.1	38.4	52.2	32.1	-	-
9	✓	✓	✓	✓		0.359	1.359	1057	0.304	**0.675**	**0.710**(-3.5%)	**1.005**(-5.8%)	**0.146**	**62.3**	39.4	**53.1**	32.2	-	-
10					✓	-	-	-	-	-	-	-	-	-	-	-	-	1.131	0.773
11	✓	✓	✓		✓	**0.366**	1.337	889	0.303	0.672	0.741	1.077	0.157	-	-	-	-	1.014	0.717
12	✓	✓	✓	✓	✓	0.358	1.334	**641**	0.302	0.672	0.728	1.054	0.154	**62.3**	**39.5**	52.8	**32.3**	**1.004**	**0.430**

Conclusion:

- **ID. 4-6:** Track & Map → Motion 🚀
- **ID. 7-9:** Motion 🚀 ↔ Occupancy 🚀
- **ID. 10-12:** Motion & Occupancy → Planning 🚀

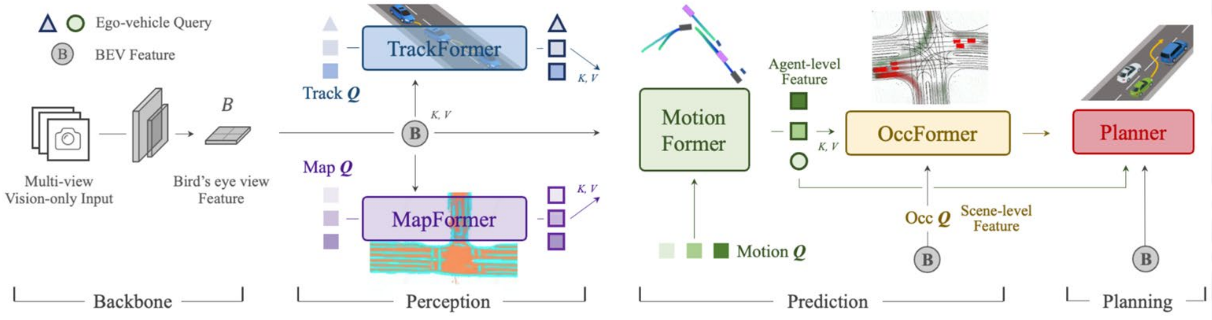
UniAD - Recover from Upstream Errors

Planner could still attend to ‘undetected’



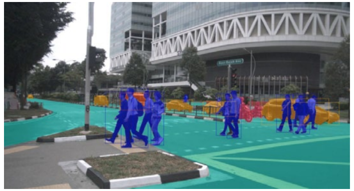
UniAD: One-page Summary

- **Planning-oriented Philosophy**: An end-to-end autonomous driving (AD) framework in pursuit of safe planning, equipped with a wide span of AD tasks.
- **Unified Query** design: *Queries* as interfaces to connect and coordinate all tasks.
- **State-of-the-art (SOTA) Performance** with vision-only input.
- **First Step towards Autonomous Driving Foundation Models**



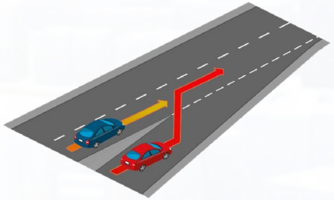
What's next

Tasks, Training Strategies, etc



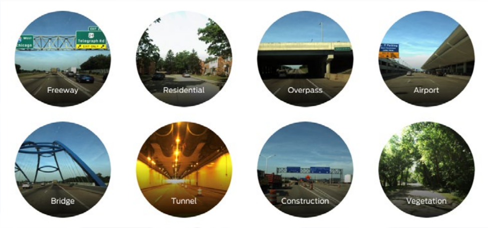
Perception / Visual Abstraction

Closed-loop Evaluation



E2E Challenges

Scale-up?



DriveData / DriveAGI

CONTENTS

From UniAD to DriveAGI ICCV23 PARIS Oral

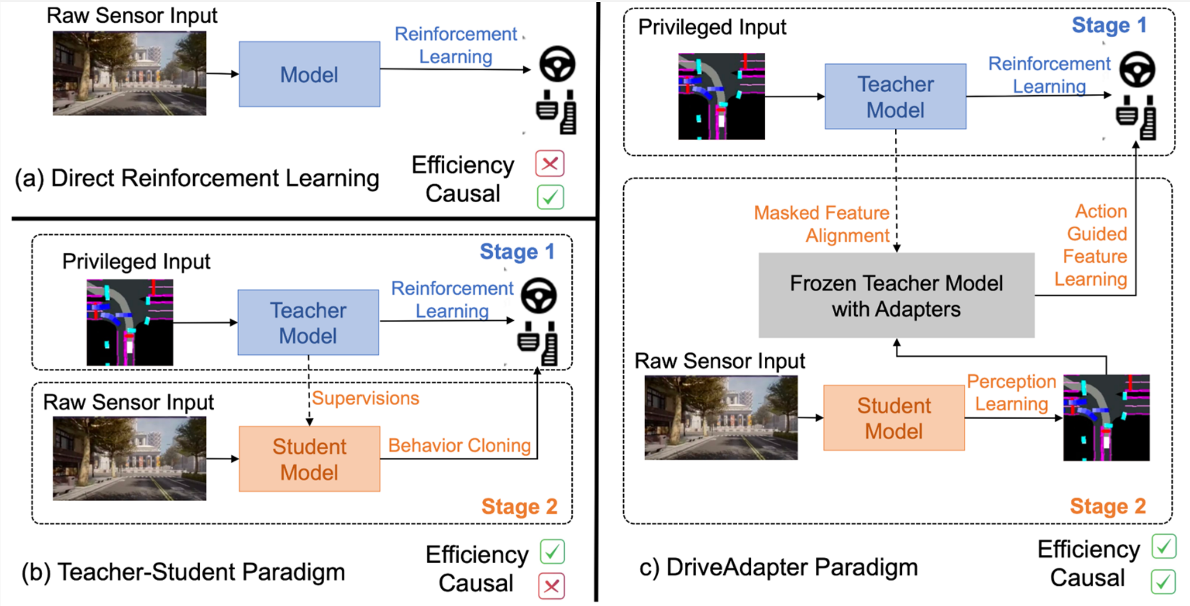
DriveAdapter

Poster: THU-AM-Room “Nord”-155

Github: https://github.com/OpenDriveLab/DriveAdapter

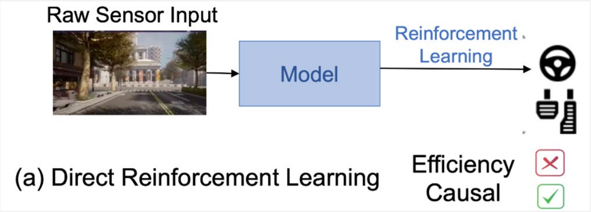
DriveAdapter - Motivation

How to balance the efficiency and causal reasoning ability?



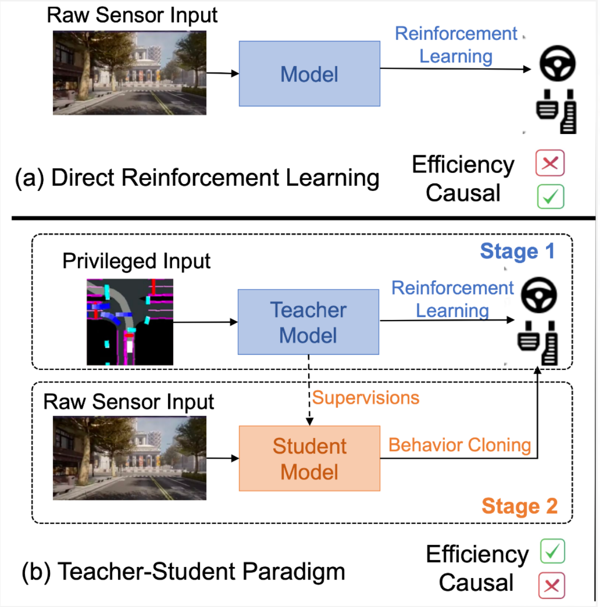
DriveAdapter - Motivation

How to balance the efficiency and causal reasoning ability?



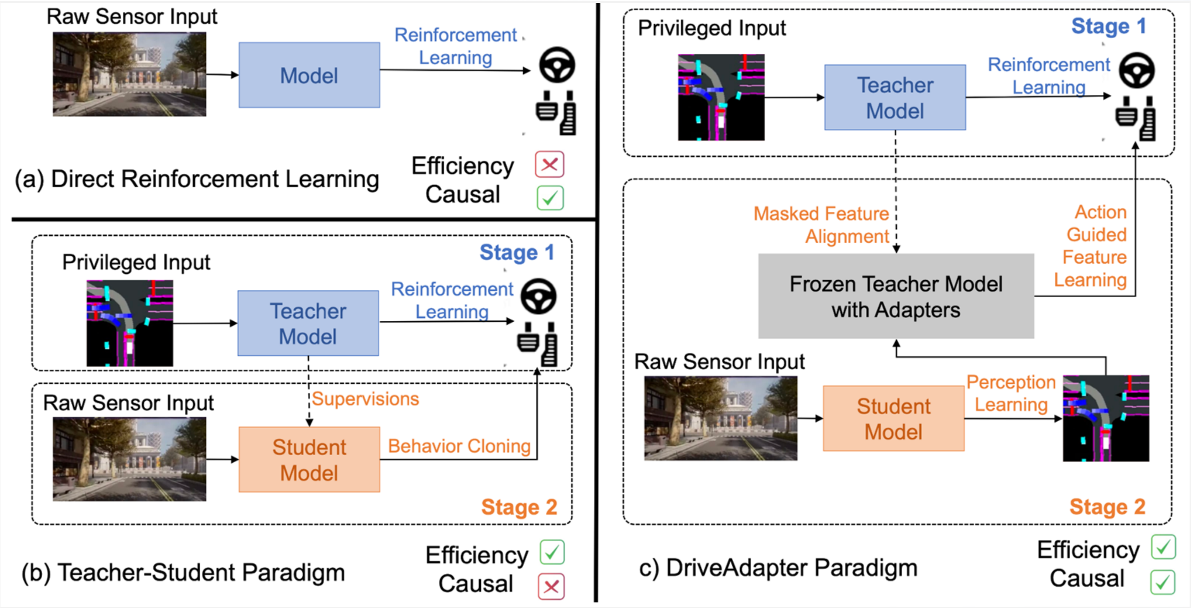
DriveAdapter - Motivation

How to balance the efficiency and causal reasoning ability?



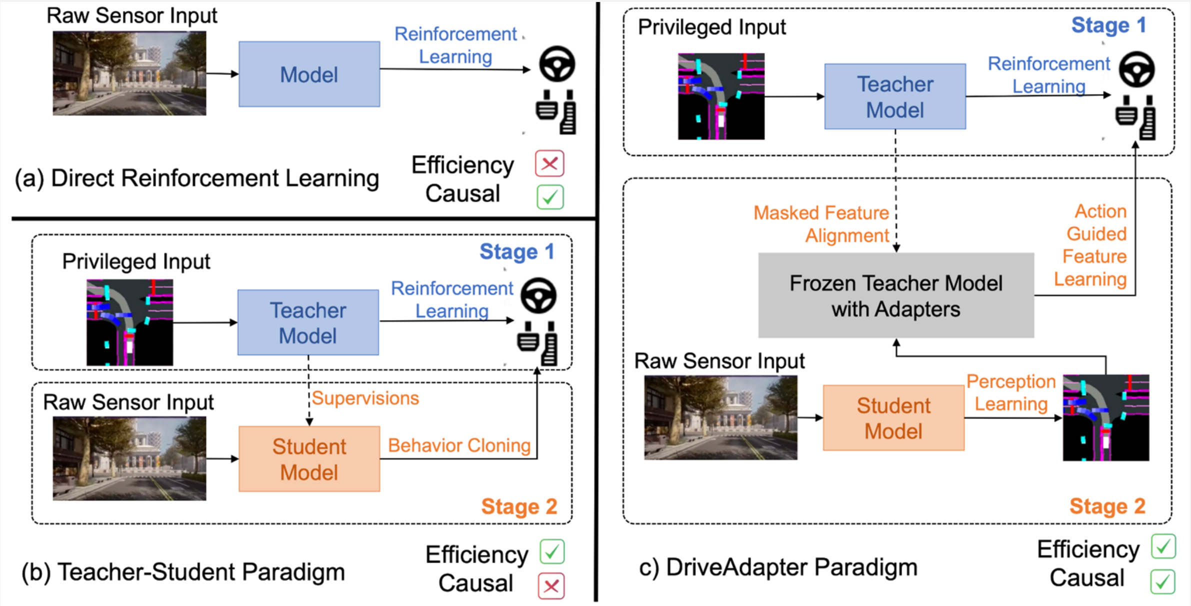
DriveAdapter - Motivation

How to balance the efficiency and causal reasoning ability?



DriveAdapter - Motivation

How to balance the efficiency and causal reasoning ability?



Utilize the strong RL-based privileged teacher model!

- Train a Teacher Model for Planning by RL
- End-to-End Connected by Adapter
- Train a Student Model for Perception

DriveAdapter - Challenge

Challenge 1: Student Model is not perfect

Privileged Input   Perception Result

BEVFusion + Mask2Former
2M training data

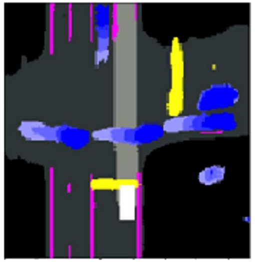
Method	Input	Driving Score ↑
Transfuser [39, 8]	Camera + LiDAR	31.0
LAV [3]	Camera + LiDAR	46.5
Student Model + Frozen Roach	Camera + LiDAR	8.9
Roach [55]	Privileged Info.	74.2
Roach + Rule [50]	Privileged Info.	**87.0**

- Directly feeding the perception results into the teacher model does **NOT** work.

DriveAdapter - Challenge

Challenge 1: Student Model is not perfect

Privileged Input



Perception Result

BEVFusion + Mask2Former
2M training data

Method	Input	Driving Score ↑
Transfuser [39, 8]	Camera + LiDAR	31.0
LAV [3]	Camera + LiDAR	46.5
Student Model + Frozen Roach	Camera + LiDAR	8.9
Roach [55]	Privileged Info.	74.2
Roach + Rule [50]	Privileged Info.	**87.0**

- Directly feeding the perception results into the teacher model does **NOT** work.
- Teacher Model would be the **upper bound** of Student Model's performance

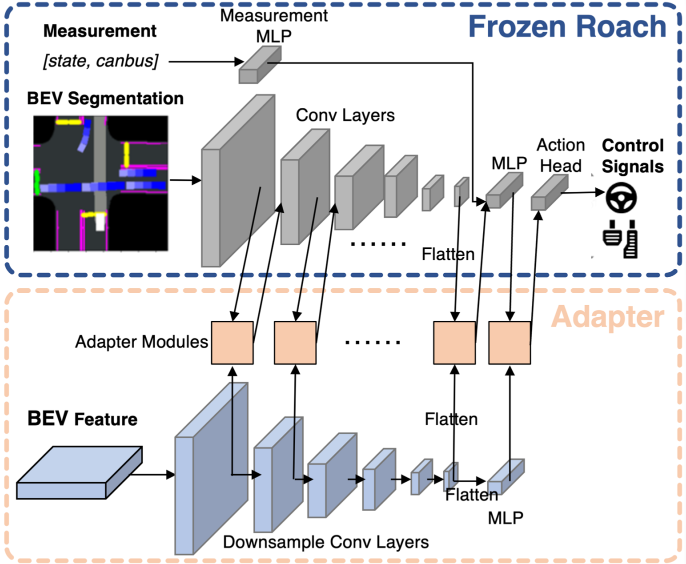
Challenge 2: Teacher Model is not perfect

Example: Emergency brake if there is any obstacle in the front - **require privileged information**

```
## Rules for emergency brake
should_brake = self.collision_detect()
only_ap_brake = True if (control.brake <= 0 and should_brake) else False
if should_brake:
    control.steer = control.steer * 0.5
    control.throttle = 0.0
    control.brake = 1.0
```

DriveAdapter - Method

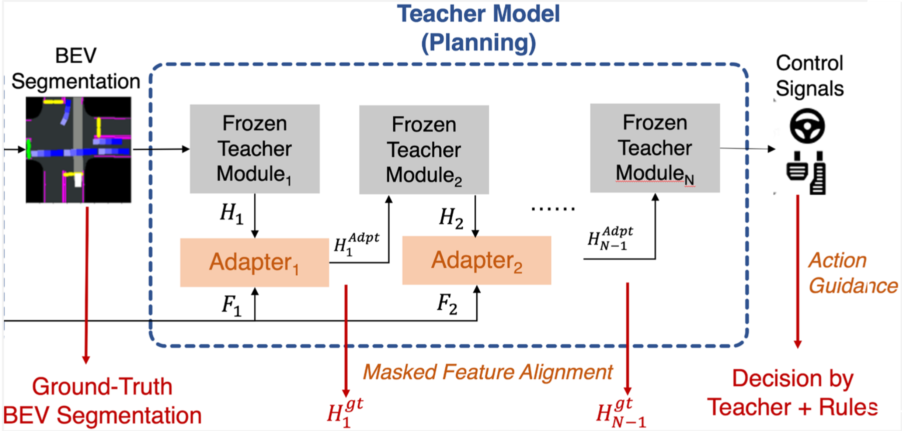
Idea 1: Deal with the distribution shift of between perception GT and prediction



- **Reduce the error in an end-to-end layer-by-layer manner:**
  - Roach (teacher model) = 6 Convs -> flatten -> 4 linears
  - **Adapter module after each layer**
  - Adapter Input: $\boldsymbol{H}_{i-1}^{\text{Adpt}} = \text{Adapter}_{i-1}([\boldsymbol{H}_{i-1}; \boldsymbol{F}_{i-1}])$
  - Adapter Output: $\boldsymbol{H}_i = \text{Teacher}_i(\boldsymbol{H}_{i-1}^{\text{Adpt}})$
  - Adapter Target/Label: GT feature map of teacher

DriveAdapter - Method

Idea 2: Inject the driving knowledge within rules into the model



- **Store the knowledge in the Adapter module:**
  - Target: let the frozen teacher action head output corrected action
  - Mask feature alignment loss for failure cases - not to learn the undesired feature map
  - Directly apply action loss for failure cases - guide the middle feature maps by backpropagation

DriveAdapter - Experiments

🚀 SOTA driving performance on CARLA closed-loop benchmark

Method	Teacher	Student	Reference	DS↑	RC↑	IS↑
CILRS [11]	Rule-Based	Behavior Cloning	CVPR 19	7.8	10.3	0.75
LBC [4]	Imitation Learning	Behavior Cloning + DAgger	CoRL 20	12.3	31.9	0.66
Transfuser [39, 8]	Rule-based	Behavior Cloning	TPAMI 22	31.0	47.5	**0.77**
Roach [55]	Reinforcement Learning	Behavior Cloning + DAgger	ICCV 21	41.6	96.4	0.43
LAV [3]	Imitation Learning	Behavior Cloning	CVPR 22	46.5	69.8	0.73
TCP [50]	Reinforcement Learning	Behavior Cloning	NeurIPS 22	57.2	80.4	0.73
ThinkTwice [26]	Reinforcement Learning	Behavior Cloning	CVPR 23	65.0	95.5	0.69
DriveAdapter	Reinforcement Learning	Frozen Teacher + Adapter	Ours	61.7	92.3	0.69
DriveAdapter + TCP	Reinforcement Learning	Frozen Teacher + Adapter	Ours	**65.9**	94.4	0.72
MILE*† [18]	Reinforcement Learning	Model-Based Imitation Learning	NeurIPS 22	61.1	**97.4**	0.63
Interfuser* [43]	Rule-Based	Behavior Cloning + Rule	CoRL 22	68.3	95.0	-
ThinkTwice* [26]	Reinforcement Learning	Behavior Cloning	CVPR 23	70.9	95.5	0.75
DriveAdapter + TCP*	Reinforcement Learning	Frozen Teacher + Adapter	Ours	**71.9**	97.3	0.74

DriveAdapter - Take-away

- **Breaking the coupling barrier of Perception and Planning:**
  - Driving knowledge from millions of steps of exploration by RL -> ***causal reasoning*** (MDP; reward), ***robustness*** (all kinds of strange cases/scenarios during exploration)
  - ***Efficient*** training for the student model

- **Masked feature distillation:** Combine the knowledge of learning-based teacher and human designed rules

- **Real-world application (potential):** A teacher on large-scale real-world motion dataset , and use DriveAdapter to solve domain adaptation for deployment

- **A Further Step towards Real-world End-to-end Autonomous Driving!**

主流工作选讲 - Part 2

GenAD / ViDAR / ELM / etc

OpenDriveLab

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

星合联盟
STAR LEAGUE

How to scale up the autonomous driving models?

ViDAR

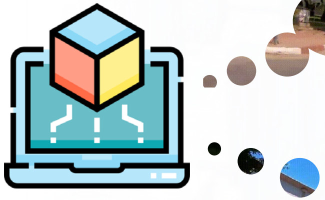
ViDAR - World Model

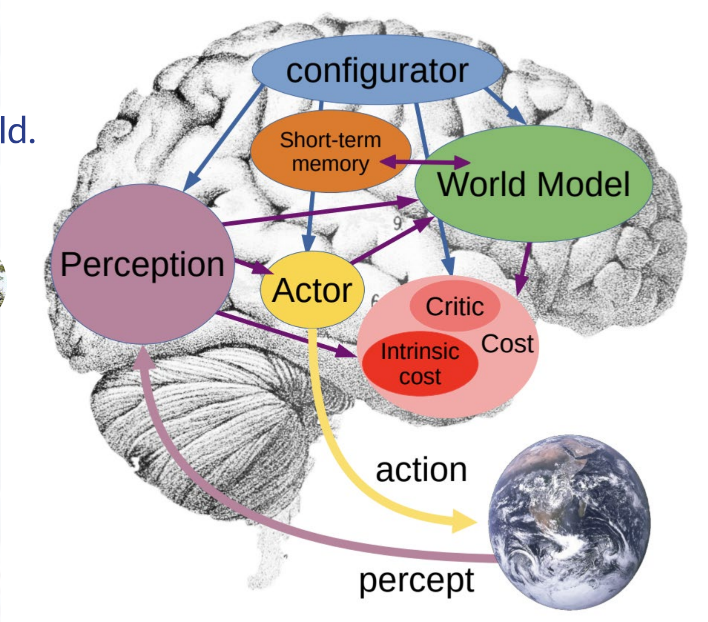
A Path Towards Autonomous Machine Intelligence Version, Yann Lecun

Task / Objective:

- **Represent the world & Learn to predict and re-act**
    - Simulate the world without **REAL** interaction with the world.

What happens if I go straight?





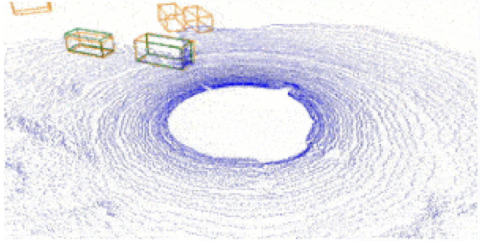
ViDAR - World Model in Driving

+ Action

Point Cloud

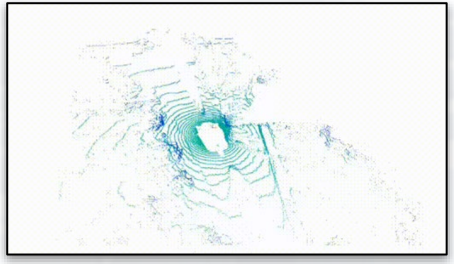
Future Prediction Model

Carnegie Mellon University



S2Net — Point cloud future prediction for planning

World Model



4D-Occ — Ego Future Tracjectory

OpenDriveLab

Point Cloud & Visual Image

ViDAR

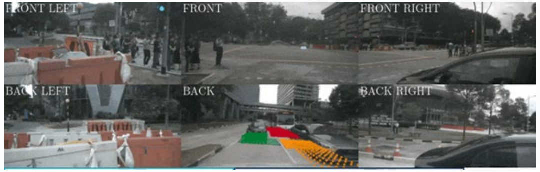
Visual Image

2022

2023

WAYVE

Fiery — visual future prediction for planning.



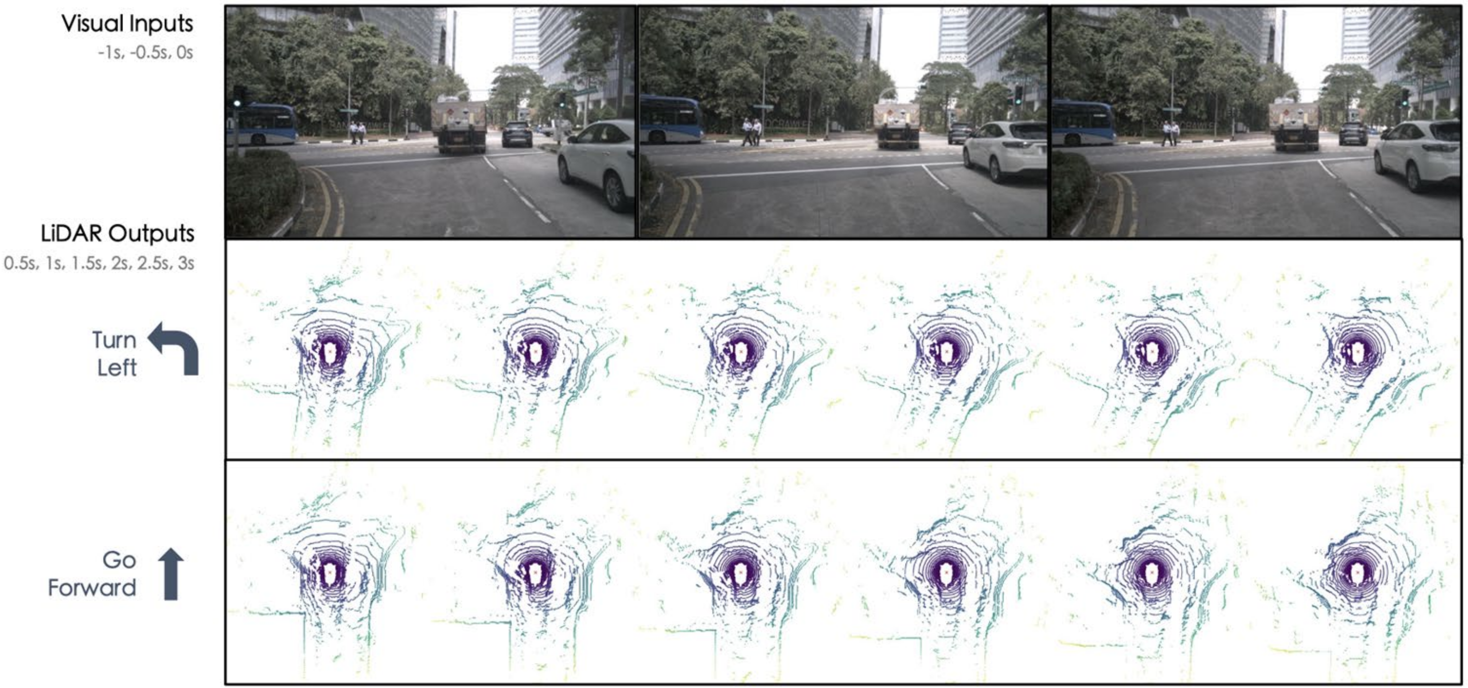
Gaia-I — Text & Steering

DriveDreamer — Box & Image & HDMap

DrivingDiffusion — Layout

ViDAR - World Model in Driving + *Action*

The First Multimodal World Model



Introducing ViDAR,
Visual Point Cloud Forecasting for Scalable Autonomous Driving

Visual Point Cloud Forecasting enables Scalable Autonomous Driving

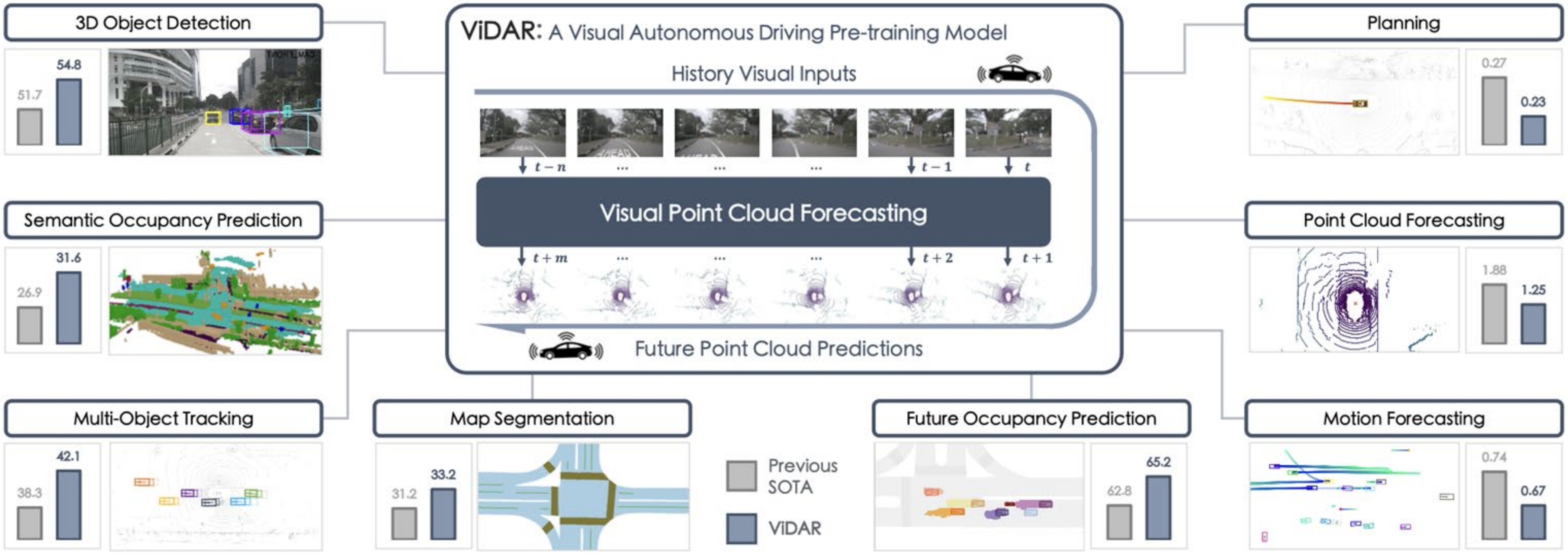
Zetong Yang   Li Chen   Yanan Sun   Hongyang Li

OpenDriveLab and Shanghai AI Lab

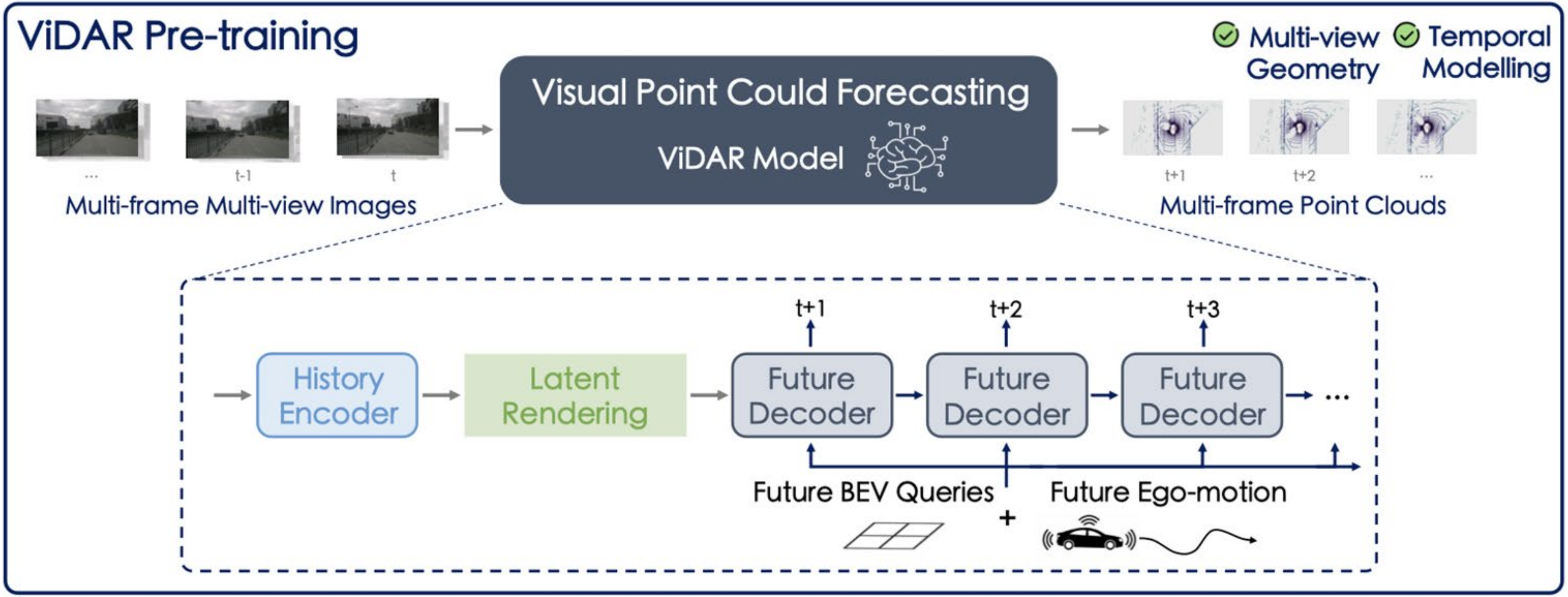
https://github.com/OpenDriveLab/ViDAR

ViDAR | At a Glance

Summary: Training multimodal world model by **Visual Point Cloud Forecasting** and boosting **End-to-End Autonomous Driving**.



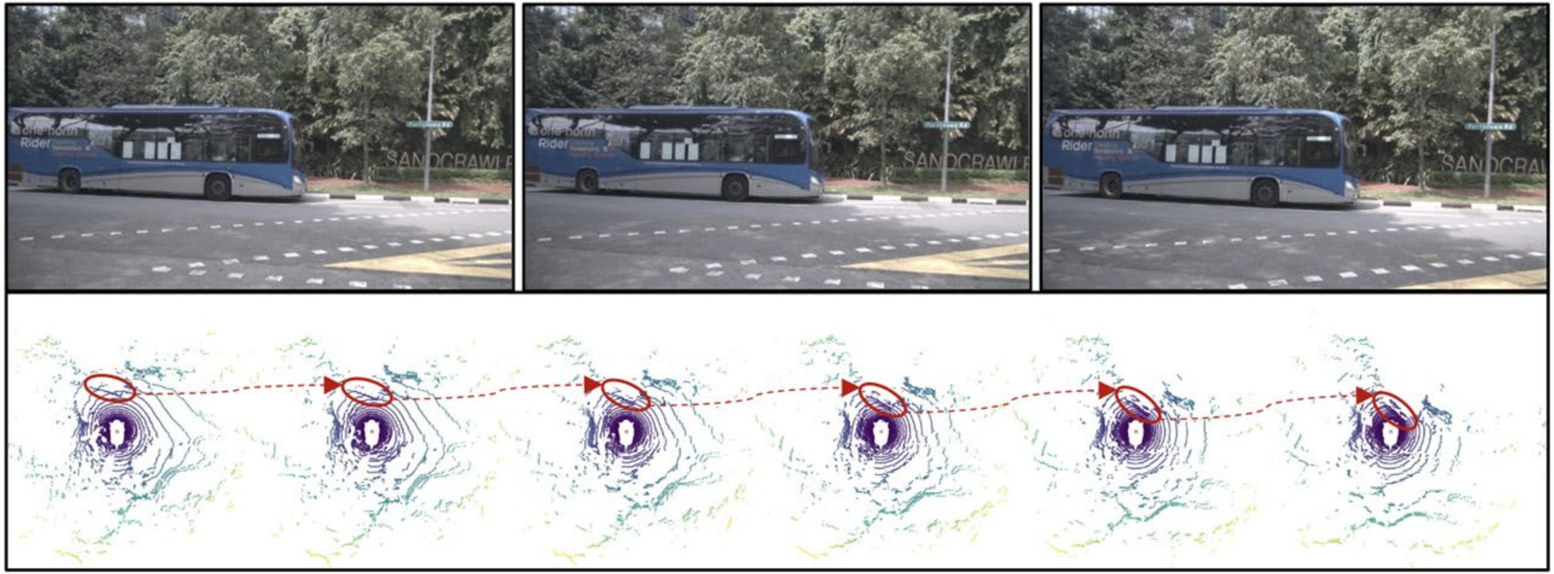
ViDAR | Architecture



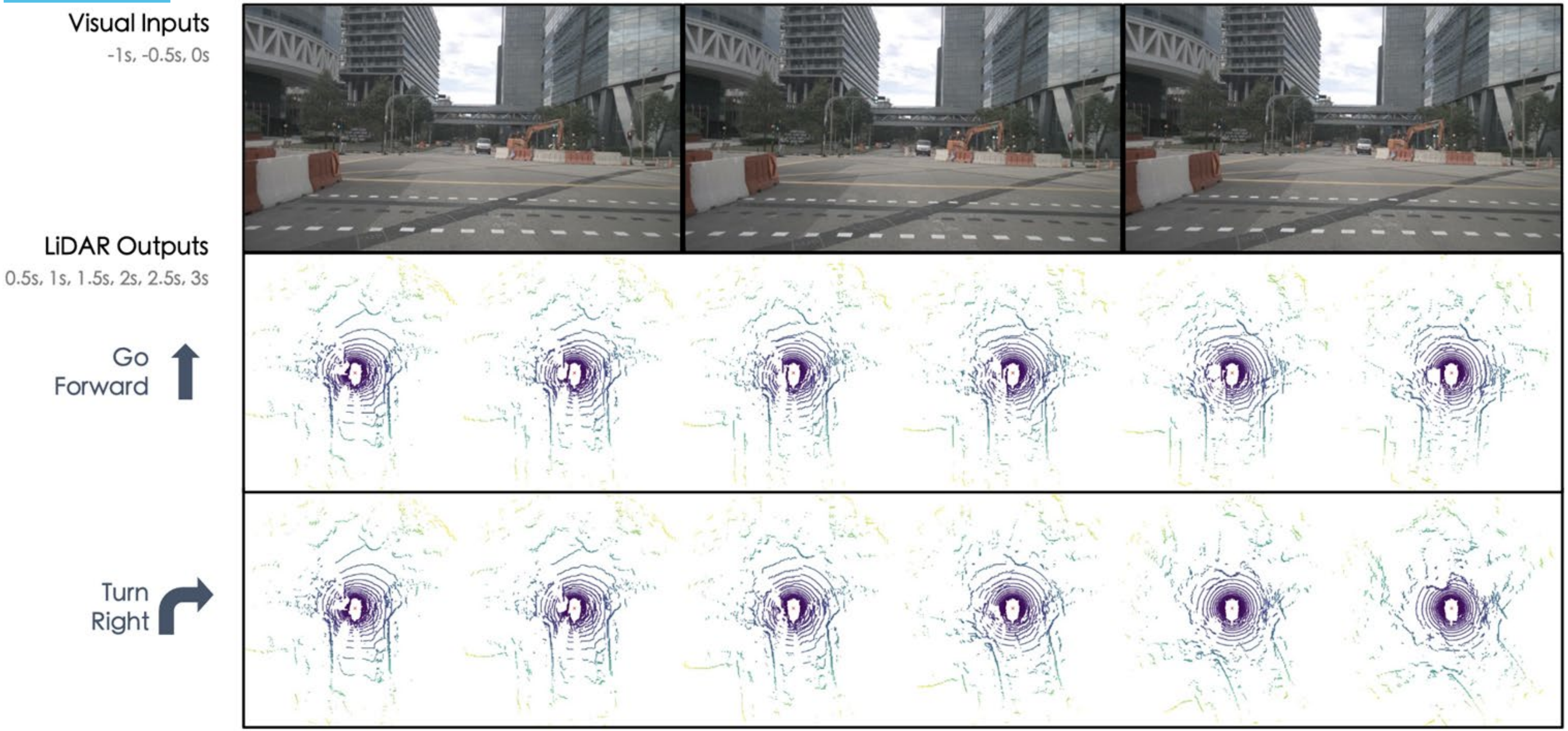
ViDAR | Future Prediction Experiments

Visual Inputs
-1s, -0.5s, 0s

LiDAR Outputs
0.5s, 1s, 1.5s, 2s, 2.5s, 3s



ViDAR | Different Ego Control Experiments



OpenDriveLab

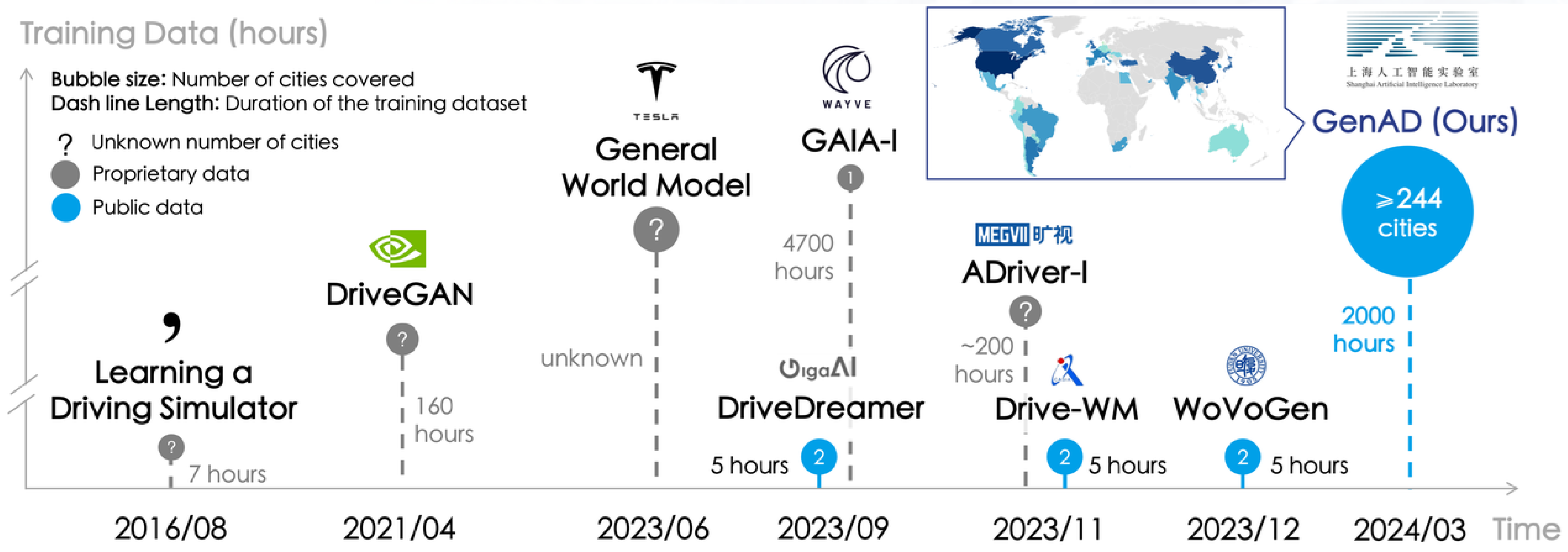
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

星合联盟
STAR LEAGUE

How to scale up the autonomous driving models?

GenAD: Generalized Predictive Model for Autonomous Driving

OpenDV Benchmark



Motivation | What Makes for Generalized AD Model?

Data

+ LLMs pretrained on **trillions of unlabeled text tokens** exhibit great generalization in open-world scenarios.

- However, existing AD models are trained on **limited labeled data**, which hampers its generalization.

LLMs

Unlabeled Text Data

arXiv Common Crawl

- ✅ **Internet scale**: World knowledge.
- ✅ **Free of labeling**: Easy to collect and scale up.

✅ *Good generalization*

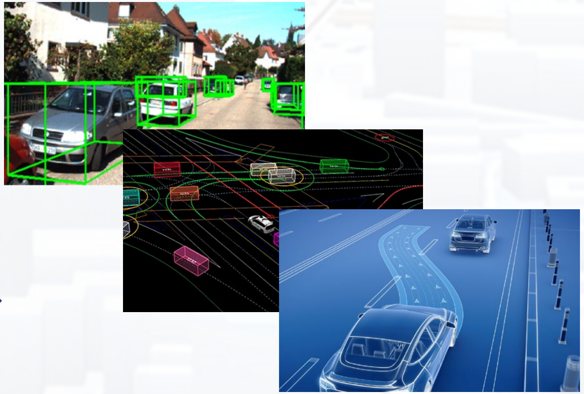
Existing AD Models

Labeled Driving Data

ARGO nuPlan

- ❌ **Small scale**: Limited domain knowledge.
- ❌ **Intricate labeling process**: Unscalable.

❌ *Poor generalization*



- Bbox, map, trajectory, etc.

Motivation | What Makes for Generalized AD Model?

Task / Objective:

- **Supervised Learning**
  - ❌ Hard to scale without sufficient labeled data

No accessible labeled data

UniAD → ❌ *UniAD-XL?*

- **Self-supervised Learning** on Feature Space
  - ✅ Scalable with developed VLMs for supervision. (e.g., DINOv2)
  - ✅ Focused on specific objects (e.g., centered, large ones)
  - ❌ Ignoring details. However, *the devil is in the details*, especially for driving

- Feature map visualization from DINOv2
- Focusing on main objects, while **ignoring fine-grained details**

Motivation | What Makes for Generalized AD Model?

Our finding:

Data: **Massive online driving videos** **+**
Task / Objective: Video Prediction
→ *Scalable and generalized* AD Model

- ✅ **Scalable Data** (easy to collect from the web)
- ✅ "**Self-supervised**" Manner
  - No 3D labeling needed
  - Detail preservation
- ✅ Learning **world knowledge** and **how to drive** inherently

✅ *Good*

Massive YouTube videos, collected worldwide

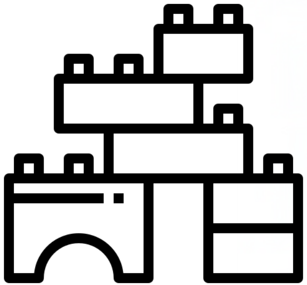
Introducing GenAD,
The First Video Generative Model as World Simulator For Autonomous Driving

Generalized Predictive Model for Autonomous Driving

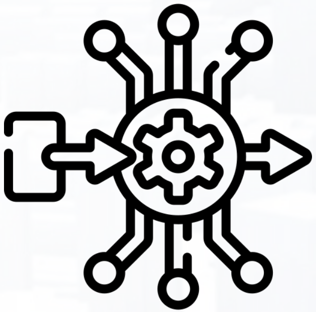
Jiazhi Yang[1*] Shenyuan Gao[2,1*] Yihang Qiu[1*] Li Chen[3,1†] Tianyu Li[1] Bo Dai[1]
Kashyap Chitta[4,5] Penghao Wu[1] Jia Zeng[1] Ping Luo[3] Jun Zhang[2♮]
Andreas Geiger[4,5♮] Yu Qiao[1♮] Hongyang Li[1†]

[1] OpenDriveLab and Shanghai AI Lab [2] Hong Kong University of Science and Technology
[3] University of Hong Kong [4] University of Tübingen [5] Tübingen AI Center

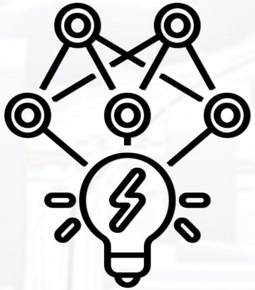
GenAD - Overview



Data



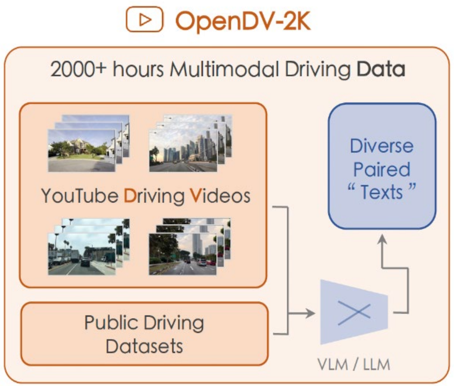
Network Architecture

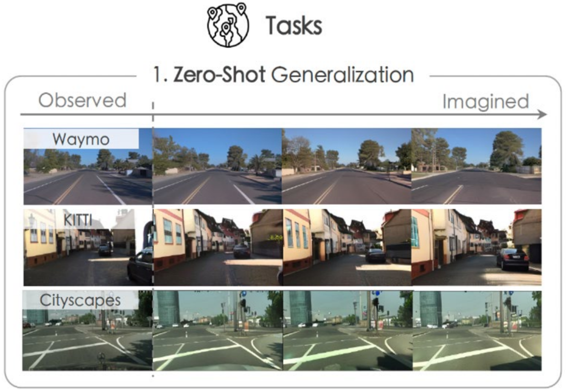


Tasks

GenAD | At a Glance

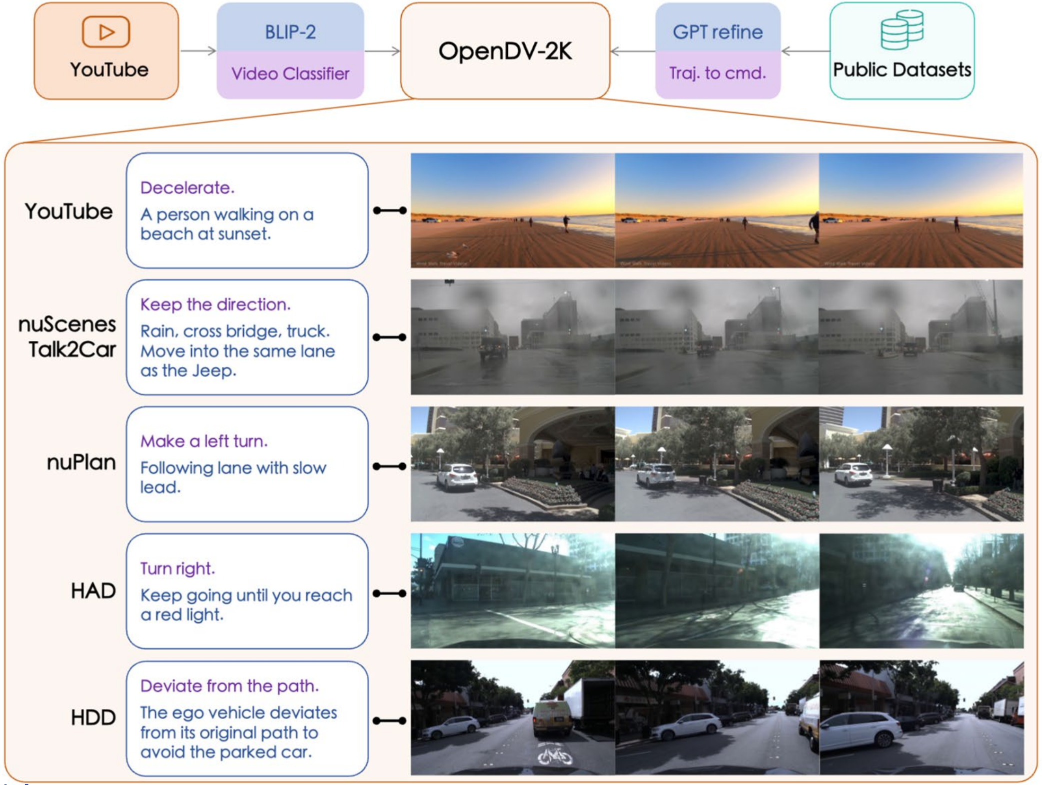
Summary: Training a **billion-scale video prediction model** on **web-scale driving videos**, to enable its **generalization across** a wide spectrum of **domains and tasks**.





Data | OpenDV-2K Dataset

- **Multi-modal** and **Multi-source Dataset**
  - Paired with textual **command** and **context** (annotated by VLMs).
  - Sourced from both **online videos** and **public datasets** for diversity.



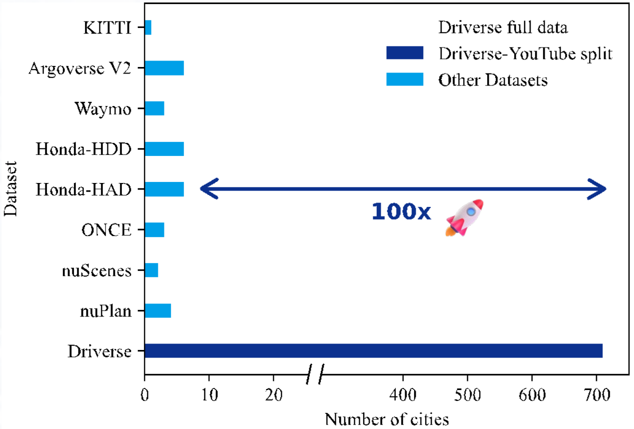
Massive YouTube videos, collected worldwide

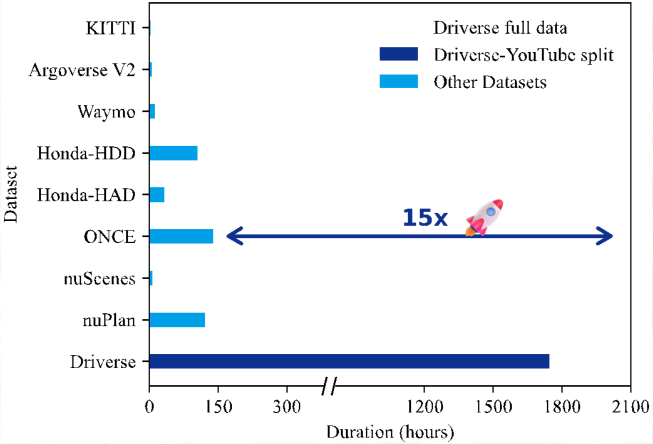
Data | OpenDV-2K Dataset

- *Largest dataset* up-to-date for autonomous driving
- 2059 hours, 709 areas

	Dataset	Duration (hours)	Front-view Frames	Geographic Diversity Countries	Geographic Diversity Cities	Sensor Setup
✗	KITTI [14]	1.4	15k	1	1	fixed
✗	Cityscapes [10]	0.5	25k	3	50	fixed
✗	Waymo Open* [41]	11	390k	1	3	fixed
✗	Argoverse 2* [45]	4.2	300k	1	6	fixed
✓	nuScenes [6]	5.5	241k	2	2	fixed
✓	nuPlan [7]	120	4.0M	2	4	fixed
✓	Talk2Car [12]	4.7	-	2	2	fixed
✓	ONCE [32]	144	7M	1	-	fixed
✓	Honda-HAD [23]	32	1.2M	1	-	fixed
✓	Honda-HDD-Action [38]	104	1.1M	1	-	fixed
✓	Honda-HDD-Cause [38]	32	-	1	-	fixed
✓	OpenDV-YouTube (Ours)	1747	60.2M	$\geq 40^{\dagger}$	$\geq 709^{\dagger}$	uncalibrated
-	**OpenDV-2K** (Ours)	**2059**	**65.1M**	$\geq 40^{\dagger}$	$\geq \mathbf{709}^{\dagger}$	**uncalibrated**

OpenDV-2K (Ours) 🚀

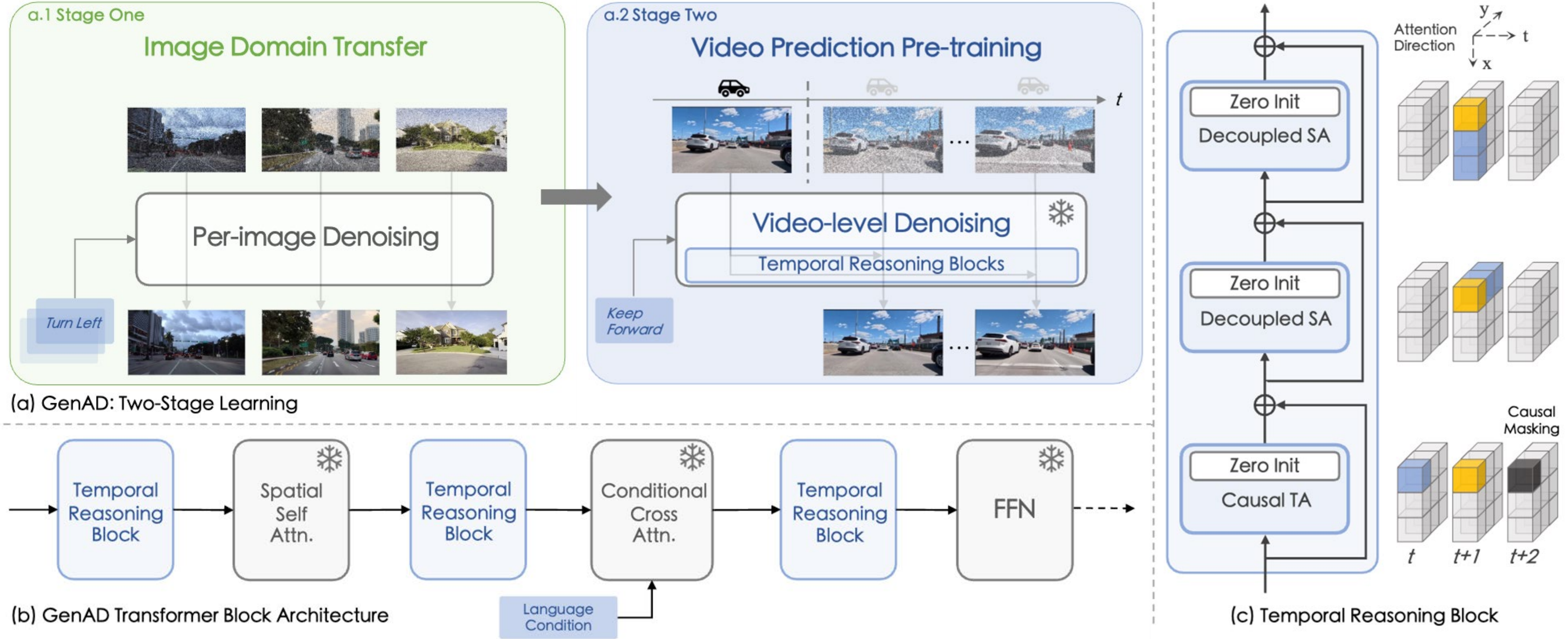




OpenDriveLab

Model | Video Prediction Model for Driving

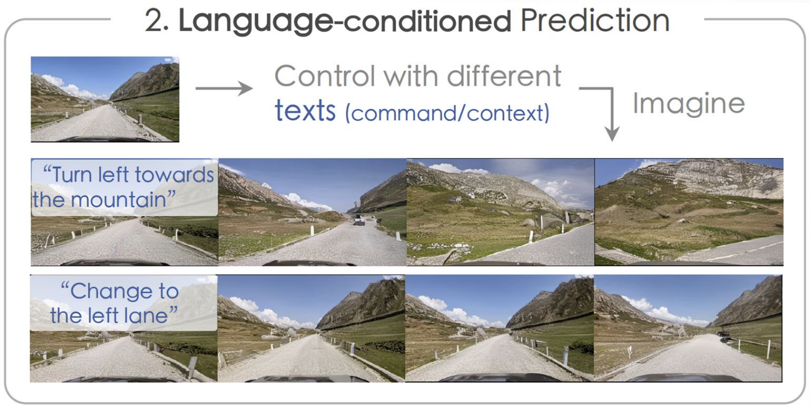
- **GenAD (5.9B)** = SDXL (2.7B) + Temporal Reasoning Blocks (2.5B) + CLIP-Text (0.7B)
- Tuning the **image generation model** (SDXL) into a highly-capable **video prediction model**



Tasks | Zero-shot Generalization (Video Prediction)

Zero-shot video prediction on unseen datasets including Waymo, KITTI and Cityscapes

Tasks | Language-conditioned Prediction

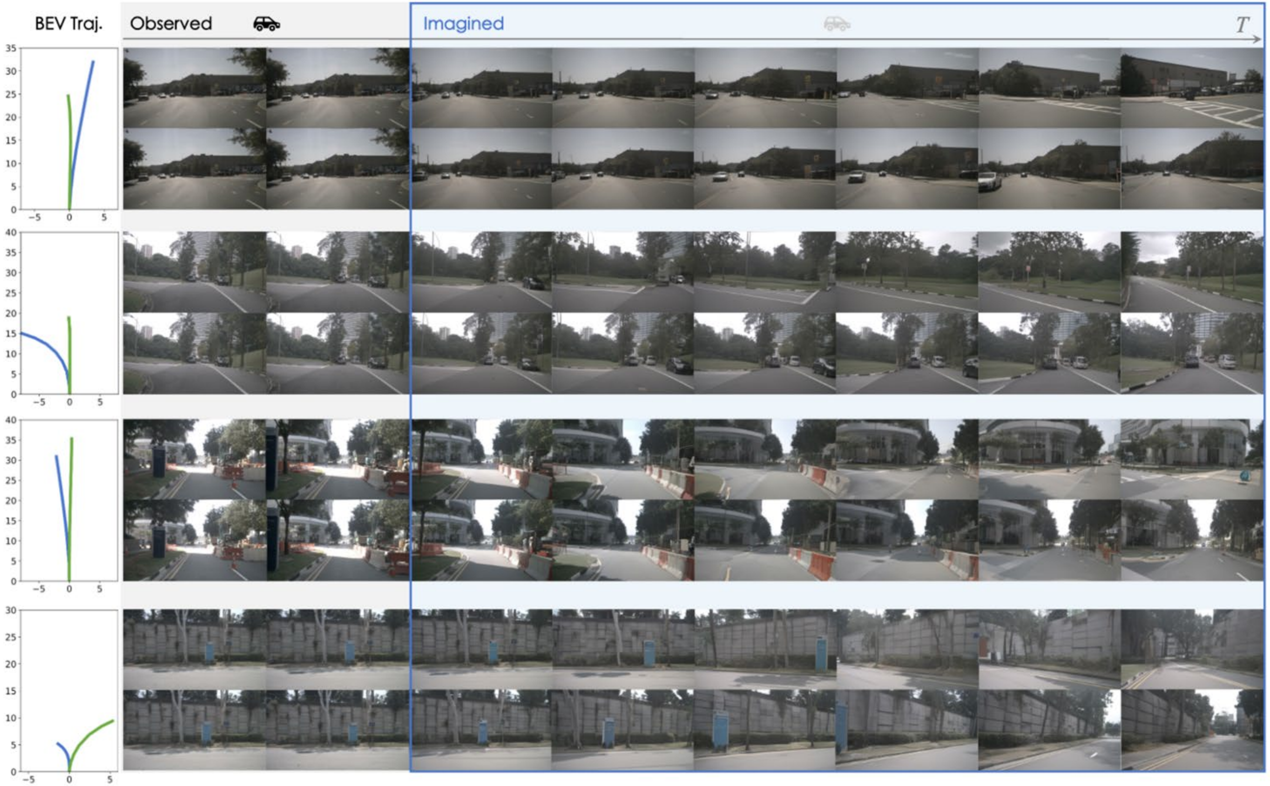


Instruct the future with **free-form texts.**

Tasks | Action-conditioned Prediction (Simulation)

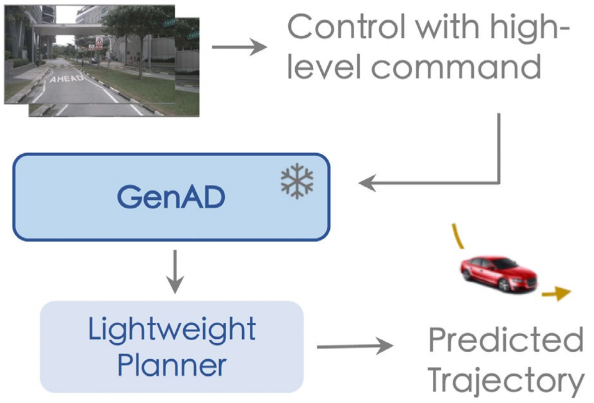
Method	Condition	nuScenes Action Prediction Error (↓)
Ground truth	-	0.9
GenAD	text	2.54
GenAD-act	text + traj.	**2.02**

Table 4. **Task on Action-conditioned prediction**. Compared to GenAD with text conditions only, GenAD-act enables more precise future predictions that follow the action condition.



Simulate the future differently conditioned on **future trajectory.**

Tasks | Planning



Method	# Trainable Params.	nuScenes ADE (↓)	FDE (↓)
ST-P3* [20]	10.9M	2.11	2.90
UniAD* [22]	58.8M	1.03	1.65
GenAD (Ours)	0.8M	1.23	2.31

Table 5. **Task on Planning**. A lightweight MLP with *frozen* GenAD gets competitive planning results with 73× fewer trainable parameters and front-view image alone. *: multi-view inputs.

Training process **speeds up by 3400 times** compared to UniAD (CVPR Best Paper).

More Visualizations on Video Prediction

OpenDriveLab

OpenDriveLab

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

DriveLM: Driving with Graph Visual Question Answering

https://github.com/OpenDriveLab/DriveLM

Trending: Driving + Language

Honda Research Institute USA

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

Explainable Driving Behavior

2019 — Planning

2022 — Prediction

2023 — Full-stack

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

Berkeley DeepDrive

Action description: **Action justification:**

(1) The car is driving ***as*** *there is nothing to impede it.*

Trending: Driving + Language

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**).

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-of-thought.

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.

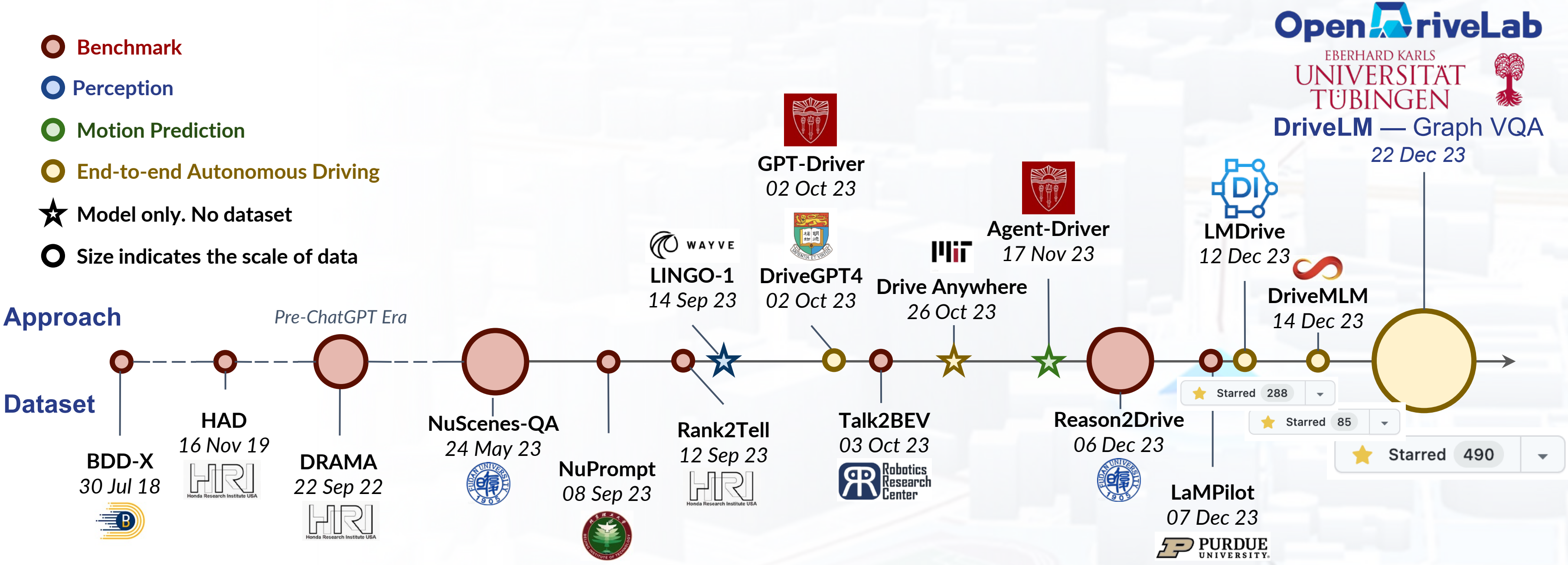
OpenDriveLab

https://arxiv.org/abs/2312.14150

DriveLM: When LLMs meet Driving

In collaboration with 美团

- Largest and high-quality benchmark, up to date.



LLMs in Driving

DriveLM: Driving with Graph Visual Question Answering

Chonghao Sima[4,1*] Katrin Renz[2,3*] Kashyap Chitta[2,3] Li Chen[4,1] Hanxue Zhang[1]
Chengen Xie[1] Ping Luo[4] Andreas Geiger[2,3] Hongyang Li[1]

[1] OpenDriveLab, Shanghai AI Lab [2] University of Tübingen
[3] Tübingen AI Center [4] University of Hong Kong

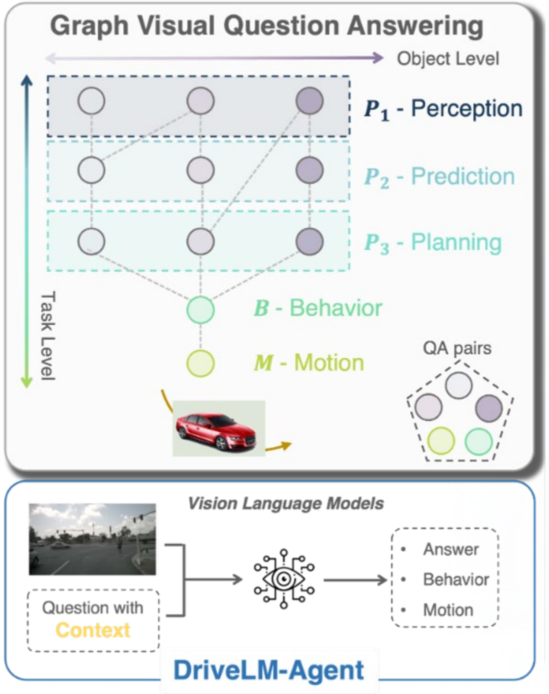
DriveLM - Introduction

- **Generalization** and **Interactivity** in Autonomous Driving.
  - Generalized to **unseen** sensor configuration and objects.
  - Regional / Global (e.g. European) regulations require **explainability** through interaction.

- Recent success in **Vision Language Models**.
  - Good **reasoning** ability, enabled by LLM.
  - **No BEV** representation, since human do not rely on BEV.

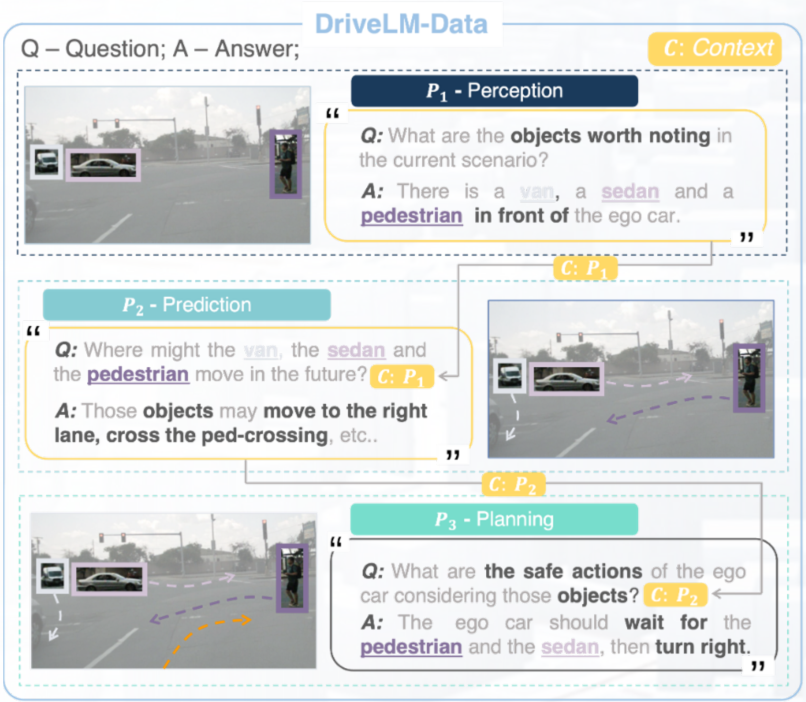
- Why VLM in AD?
  - **Reasoning** ability helps **generalization**.
  - **Language** output provide **interactivity**.

DriveLM - At A Glance



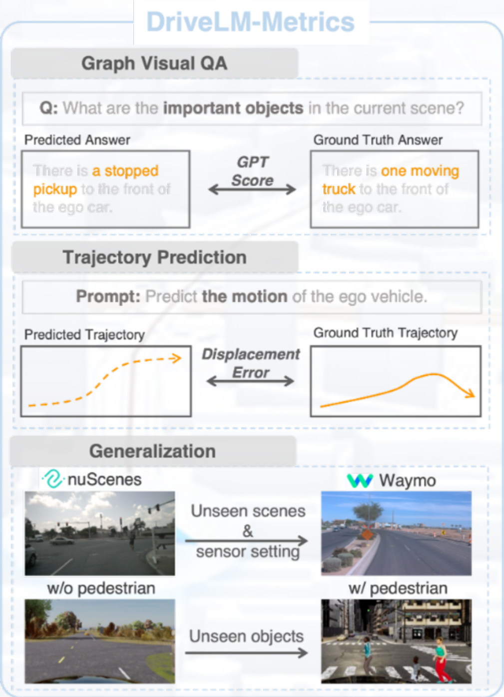
- The critical part is **Graph Visual QA**, upon which we build **data**, model and metrics accordingly.

DriveLM - At A Glance



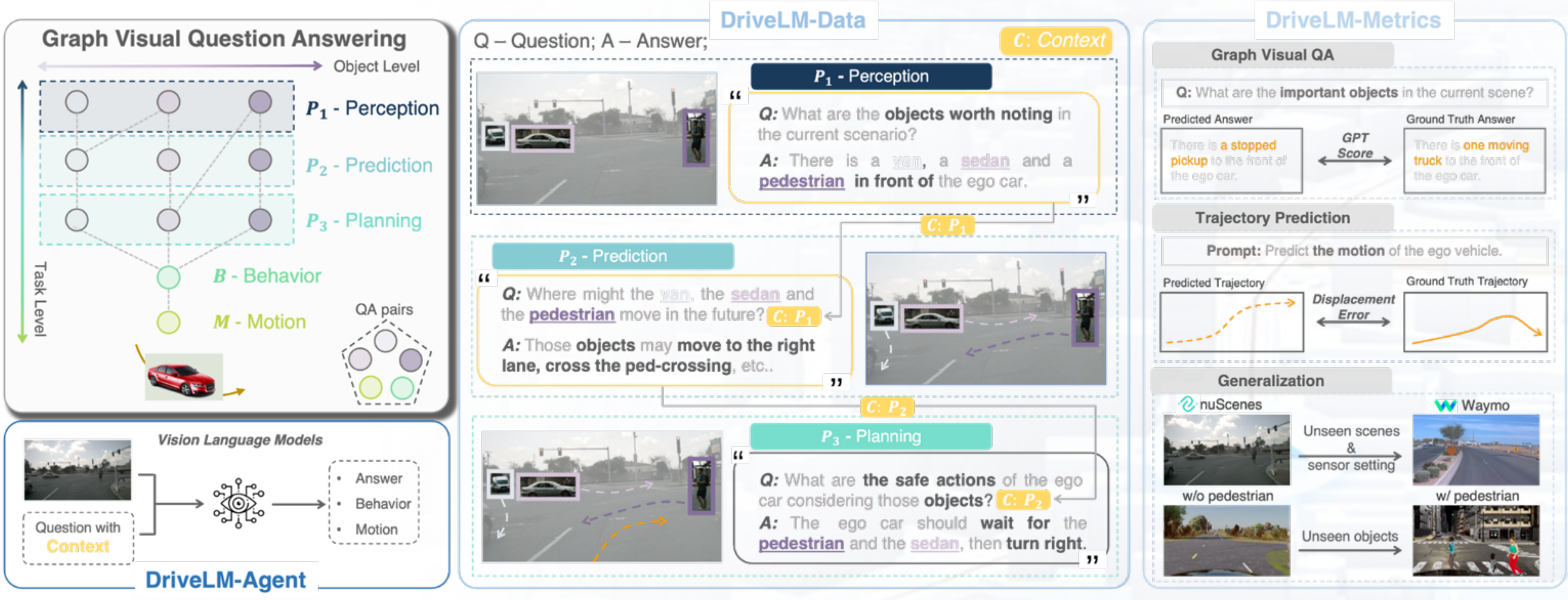
- The critical part is **Graph Visual QA**, upon which we build **data**, model and metrics accordingly.

DriveLM - At A Glance



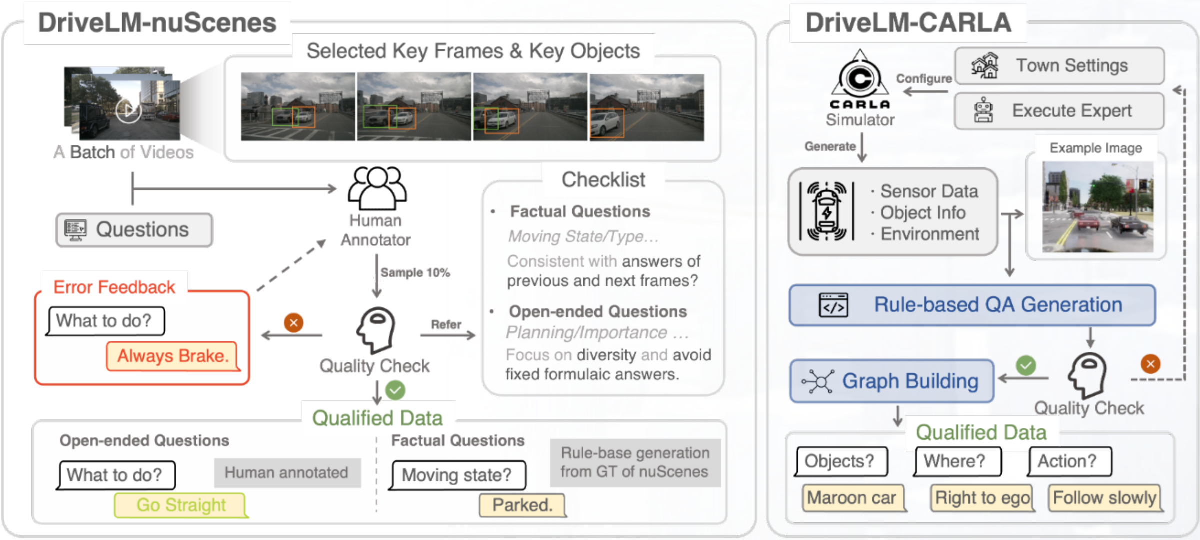
- The critical part is **Graph Visual QA**, upon which we build **data**, model and metrics accordingly.

DriveLM - At A Glance



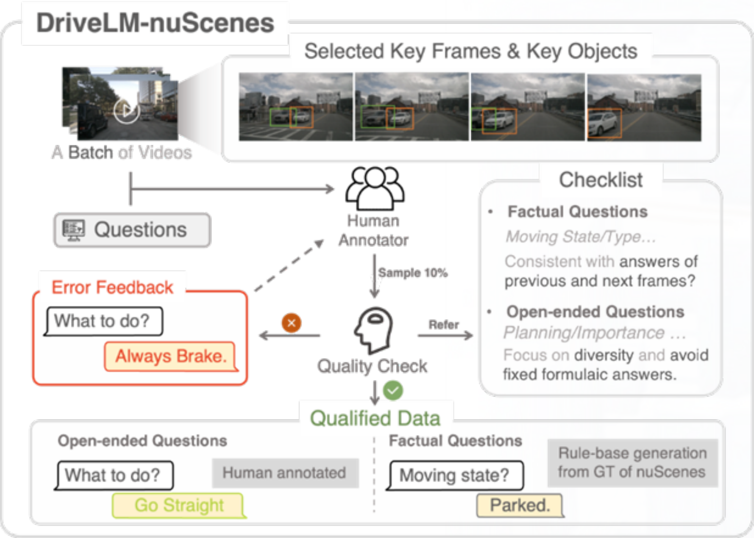
- The critical part is **Graph Visual QA**, upon which we build **data**, model and metrics accordingly.

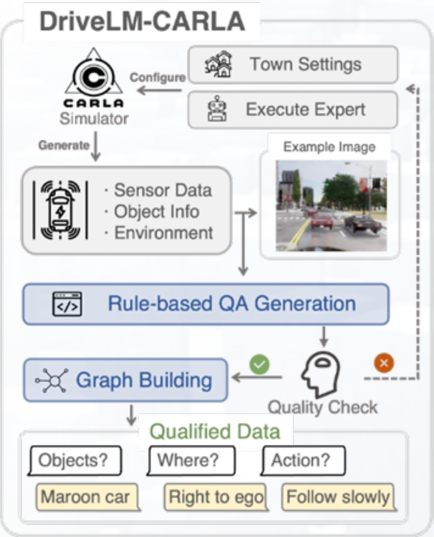
DriveLM - Data

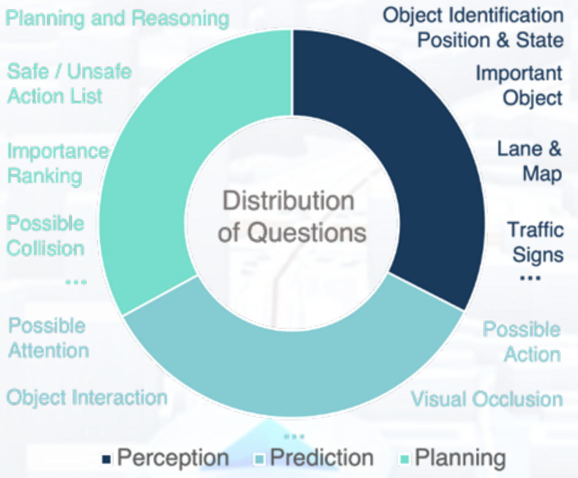


- To ensure the **data quality**, we introduce human annotation with multi-round quality check in nuScenes.
- To **scale-up** annotation, we adopt auto-labelling in CARLA.

DriveLM - Data







- To ensure the **data quality**, we introduce human annotation with multi-round quality check in nuScenes.
- To **scale-up** annotation, we adopt auto-labelling in CARLA.

Diversity matters, spanning from perception to prediction and planning.

DriveLM - Experiments

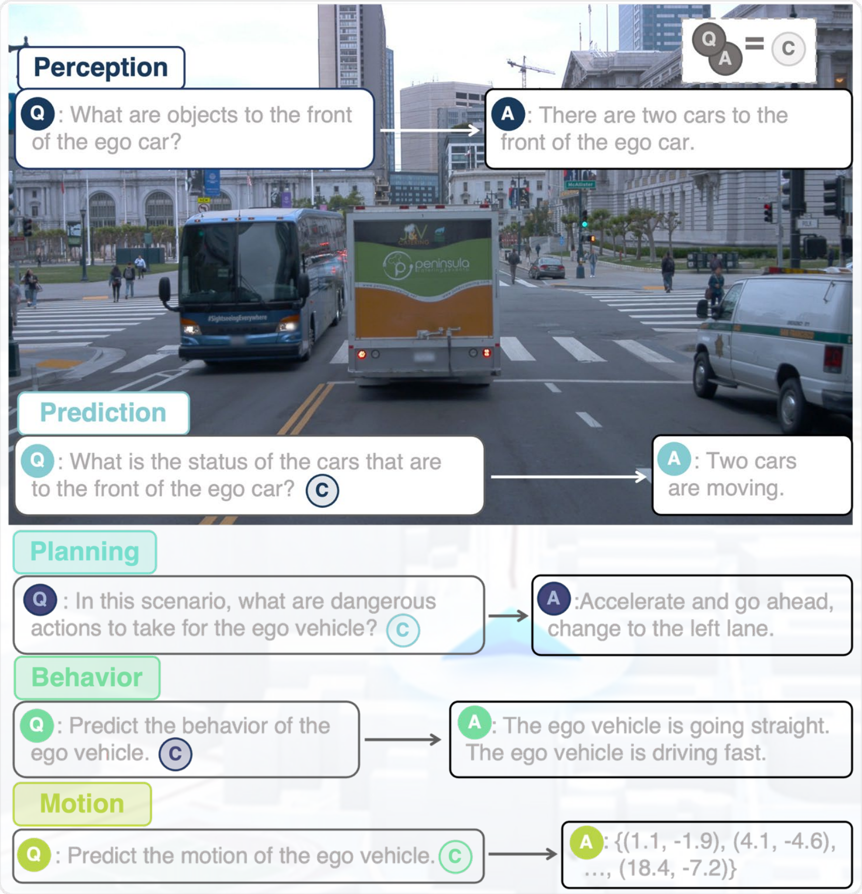
Method	Behavior Context	Motion Context	Behavior (B) Acc. ↑	Speed ↑	Steer ↑	Motion (M) ADE ↓	FDE ↓
Command Mean	-	-	-	-	-	7.98	11.41
UniAD-Single	-	-	-	-	-	4.16	9.31
BLIP-RT-2	-	-	-	-	-	2.78	6.47
DriveLM-Agent	None	B	35.70	43.90	65.20	2.76	6.59
	Chain	B	34.62	41.28	64.55	2.85	6.89
	Graph	B	**39.73**	**54.29**	**70.35**	**2.63**	**6.17**

- Trained on DriveLM-Data (nuScenes-based), DriveLM-Agent (ours) gains **better zero-shot** ability on Waymo scenarios, overpassing other methods by a large margin.

DriveLM - Experiments

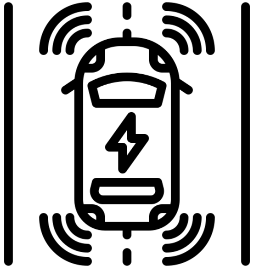
Method	Behavior Context	Motion Context	Behavior (B) Acc. ↑	Behavior (B) Speed ↑	Behavior (B) Steer ↑	Motion (M) ADE ↓	Motion (M) FDE ↓
Command Mean	-	-	-	-	-	7.98	11.41
UniAD-Single	-	-	-	-	-	4.16	9.31
BLIP-RT-2	-	-	-	-	-	2.78	6.47
DriveLM-Agent	None	B	35.70	43.90	65.20	2.76	6.59
	Chain	B	34.62	41.28	64.55	2.85	6.89
	Graph	B	**39.73**	**54.29**	**70.35**	**2.63**	**6.17**

- Trained on DriveLM-Data (nuScenes-based), DriveLM-Agent (ours) gains **better zero-shot** ability on Waymo scenarios, overpassing other methods by a large margin.



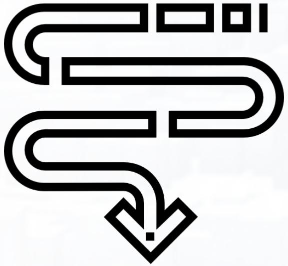
- Qualitative result shows that DriveLM-Agent does **understand the unseen scenarios** in some way.

DriveLM - Limitation



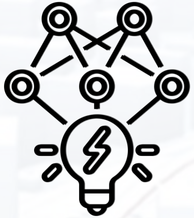
Driving-specific Inputs

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



Closed-loop Planning

DriveLM-Agent 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, DriveLM-Agent suffers from long inference time, which may impact practical implementation.

OpenDriveLab

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

ELM: Embodied Understanding of Driving Scenarios

Embodied Understanding

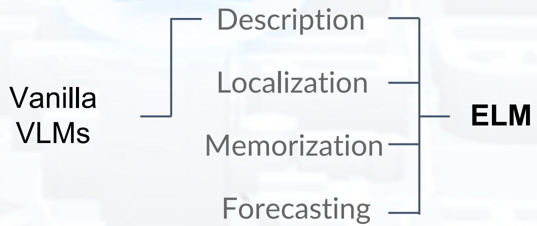
Embodied Understanding of Driving Scenarios

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

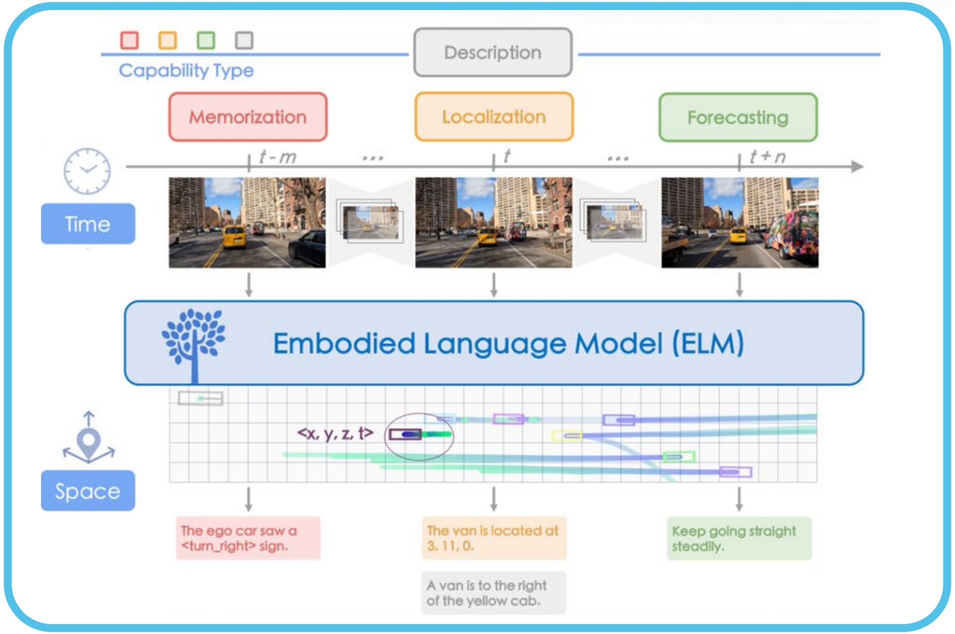
[1] OpenDriveLab, Shanghai AI Lab [2] Shanghai Jiao Tong University
[3] Fudan University [4] University of California, Riverside

ELM - Introduction

- Embodied understanding.
  - **interacting** with environments & **reasoning** via common sense.
- Vision-Language Models.
  - 2D domain: 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.

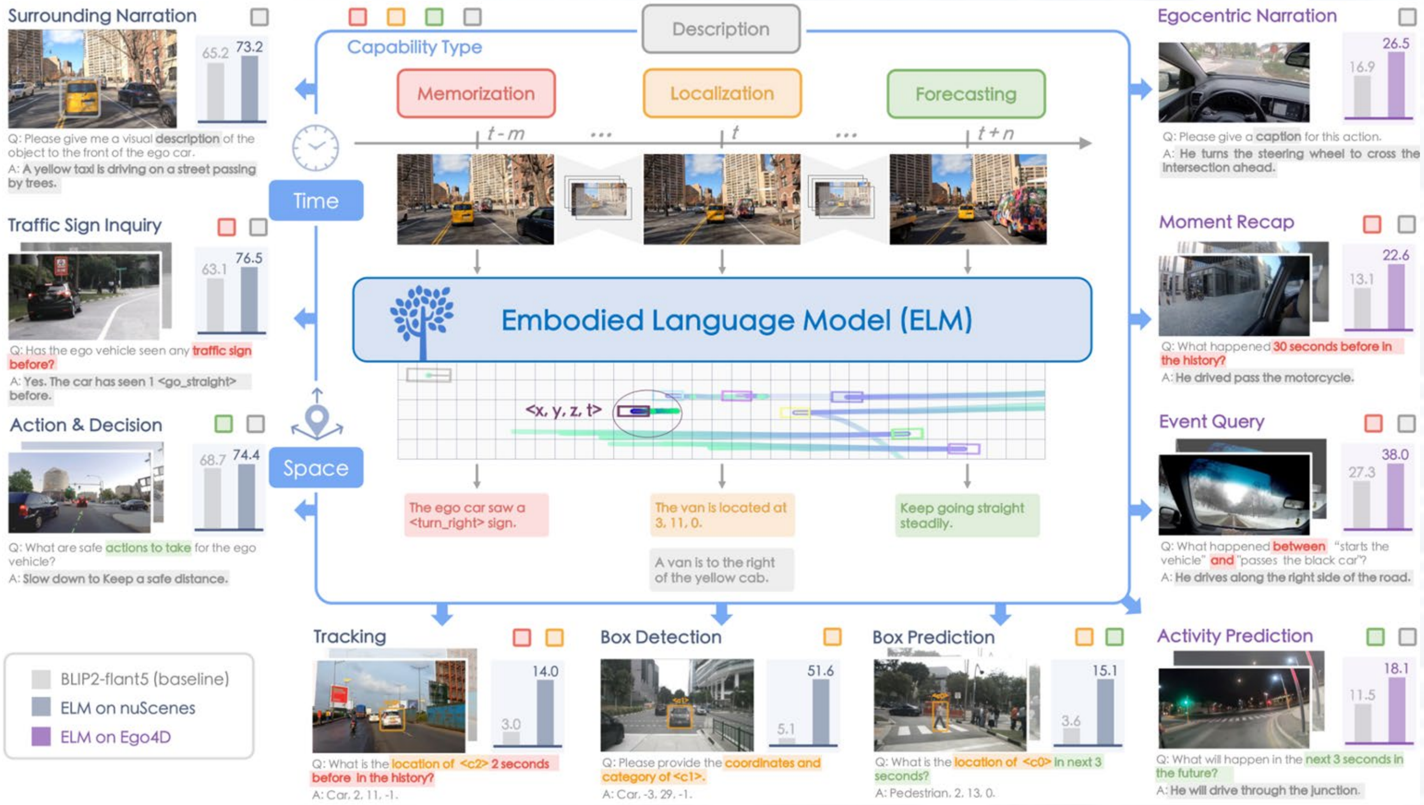


Embodied Understanding At A Glance



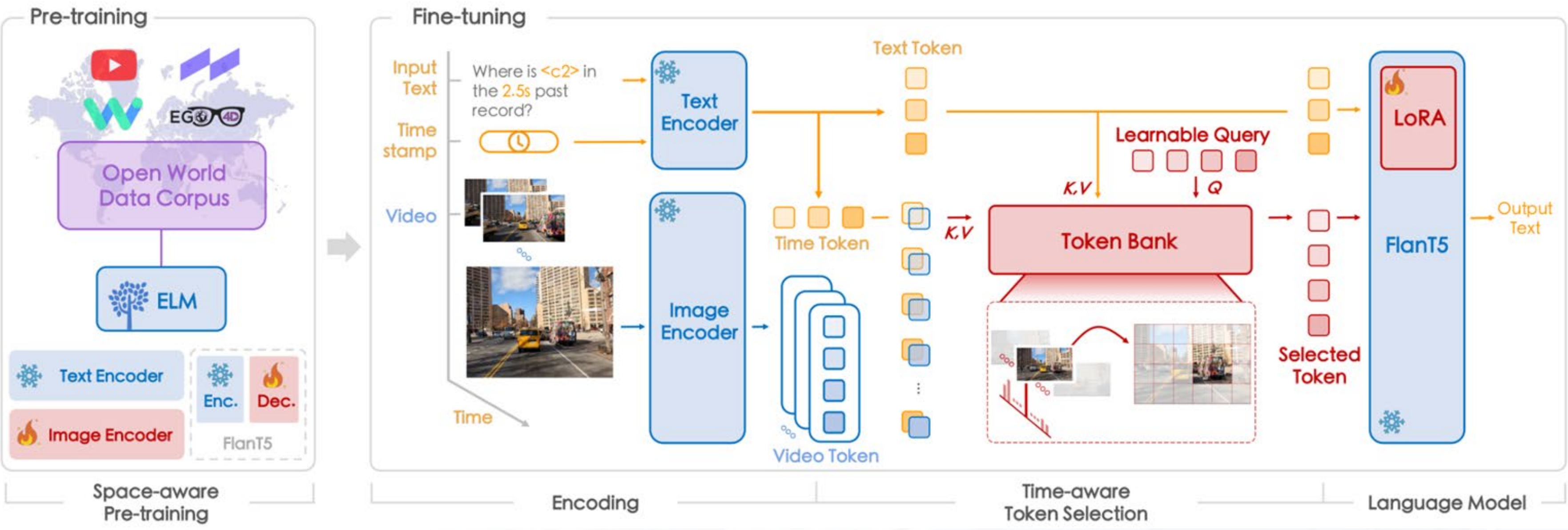
- ELM is an **embodied language model** for understanding the long-horizon driving scenarios in **space** and **time**.
- We expand a wide spectrum of **new tasks** to fully leverage large language models in an embodiment setting.

Embodied Understanding At A Glance



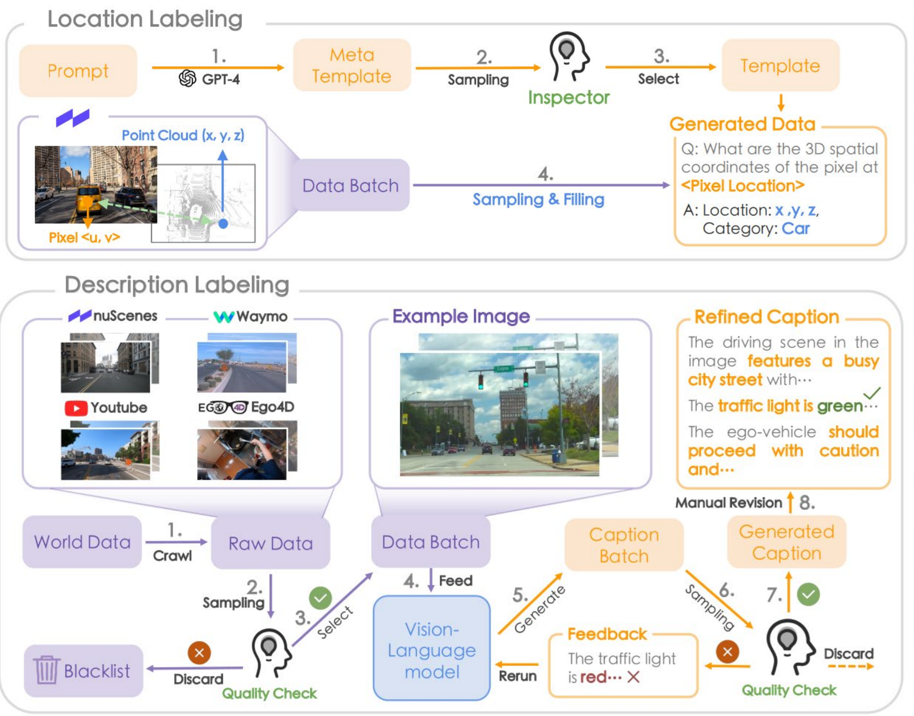
Embodied Understanding

Pipeline



Embodied Understanding

Annotation - Quality check

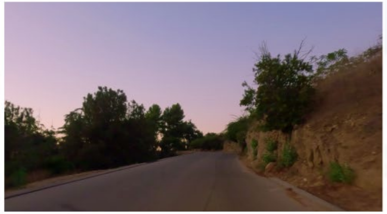


Examples of Description Labels

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.

Q: Please describe the driving scene.



A: The scene is a road with a curvy, winding path, surrounded by trees and hills. The road is empty, with no other vehicles visible. The sun is setting, casting a warm glow over the scene.

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

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.**

Box Prediction

Q: Determine the 3D location in the scene of the 2D pixel at **<c, 865.3, 505.9> 3 seconds later?**

BLIP2-T5: Location: [1.18, 29.96, 1.56], Car
Ours: Location: [0, 38, 0], Pedestrian
GT: Location: [-0.12, 38.45, 0.05], Pedestrian

Traffic Sign Inquiry

Q: Has the ego vehicle seen any **traffic sign** before?

BLIP2-T5: There is no traffic sign in the scene.
Ours: The ego vehicle has seen 1 go_straight, and 1 turn_right before.
GT: The ego vehicle has seen 1 go_straight, and 1 turn_right before.

Memory Recap

Q: "What **happened 12.1 seconds before?**

BLIP2-T5: C stirs the meat in the frying pan
Ours: C moves the meat from the frypan to the plate
GT: C moves the meat from the frypan to the plate

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.
- DriveLM: Driving with Graph Visual Question Answering
  - Use **Graph VQA** as a proxy task to mimic human's driving logic
  - **Some good result under zero-shot setting, but still far from claiming good generalization.**
- 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

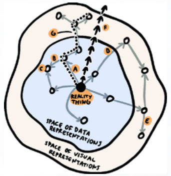
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

End-to-end Autonomous Driving

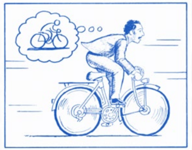
Key Challenges

Challenges in End-to-end Autonomous Driving An Overview

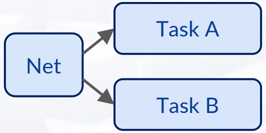
Input Modality



Visual Abstraction

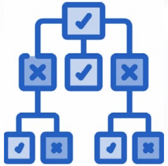


World Model



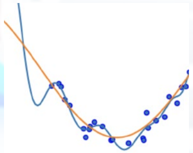
Multi-task Learning

Policy Distillation



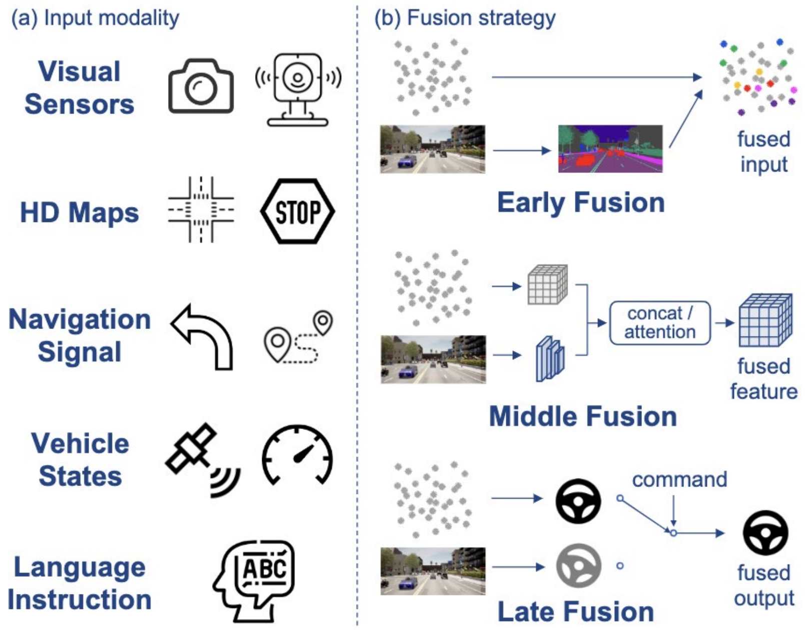
Interpretability

Causal Confusion



Robustness and Generalization

挑战（1/8）- Input Modality



- **Early Fusion:** Combine sensory information before feeding it into the feature extractor
- **Middle Fusion:** Separately encode inputs and then combining them at the feature level
- **Late Fusion:** Combine multiple results from multi-modalities **(Worst Performance)**

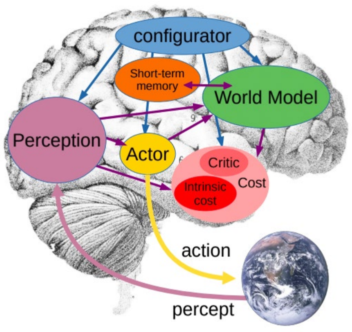
挑战（2/8） - Visual Abstraction

Current methods first pre-train the visual encoder of the network using **proxy pre-training tasks.**

⇩

There inevitably exist possible **information bottlenecks** in the learned representation, and redundant information unrelated to driving decisions may be included.

挑战（3/8） - World Model



In a nutshell:
State of the world at time t: s(t)
Imagined action taken at time t: a(t)
Causal prediction:
s(t+1) = g(s(t),a(t))
where g() is the world model.
Such a *causal* world models enables planning.

	States	Cost / Reward
RL Gyms	- Ego agent - Other objects (**static**) - Background environment	- Success/Fail - Intermediate Reward
Autonomous Driving	- Ego-vehicle - Other vehicles, pedestrians, cyclists, etc (**moving**) - Background environment	- Collision - Comfort - Forward - etc

Complicated!

Hard to define!

A video predictor?

挑战（4/8）- Multi-task Learning

Multi-task learning (MTL)： Jointly perform several related tasks based on a shared representation through separate branches/heads.

Pros

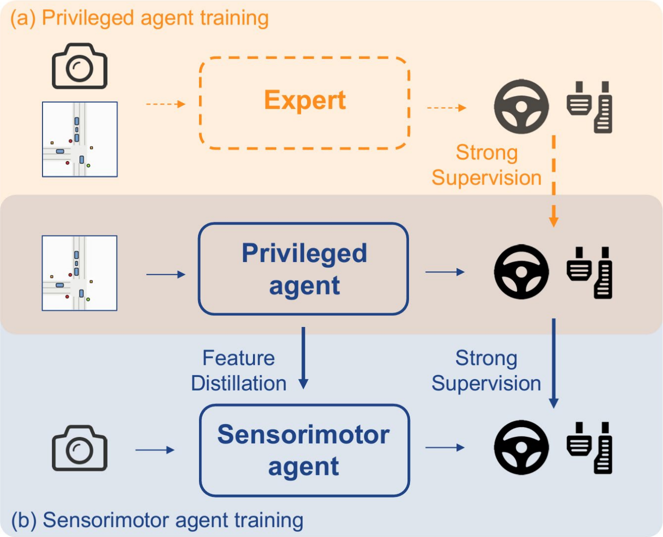
- Significant computational cost reduction
- Related domain knowledge is shared within the shared model

Challenges

- The optimal combination of auxiliary tasks and the appropriate weighting of their losses
- Construct large-scale datasets with multiple types of aligned and high-quality annotations

挑战（5/8）- Policy Distillation

The popular "Teacher-Student" IL Paradigm



- Expert: Ground Truth (GT) to action

Gap

Student: Image to action

- **Expert** (by RL/IL/hand-rule, gt input)
  - Not/Can't perfect, even for a certain benchmark

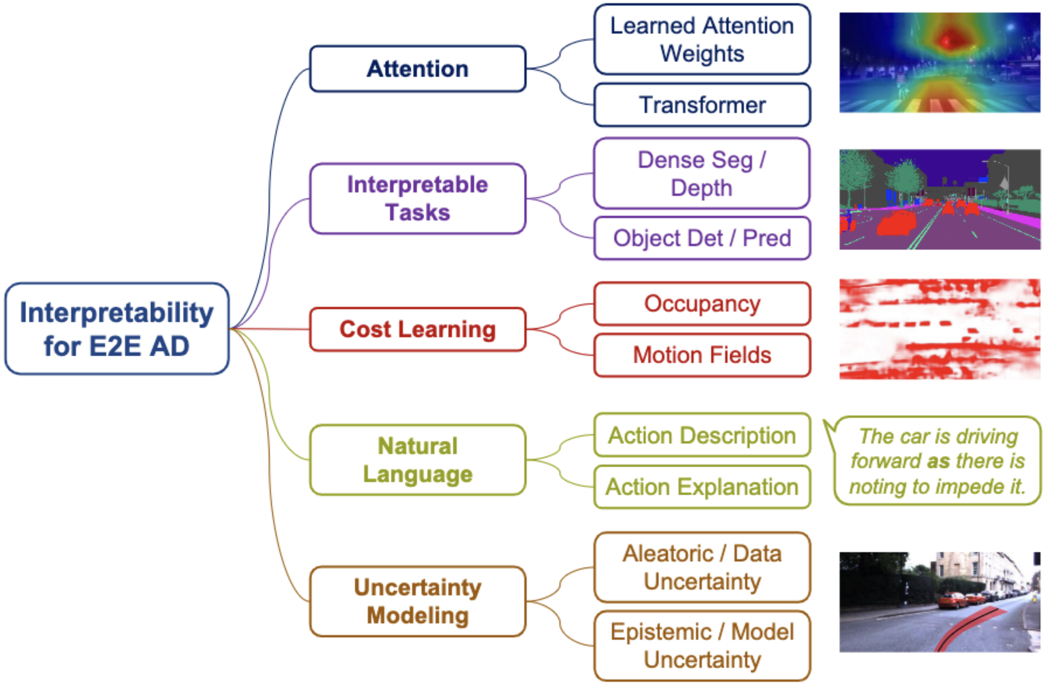
Method	Input	Driving Score ↑
Transfuser [39, 8]	Camera + LiDAR	31.0
LAV [3]	Camera + LiDAR	46.5
Student Model + Frozen Roach	Camera + LiDAR	8.9
Roach [55]	Privileged Info.	74.2
Roach + Rule [50]	Privileged Info.	**87.0**

From DriveAdapter work, ICCV 2023

- What for or How to **Distillation**
  - **Critical** features
  - Input gap - Casual confusion

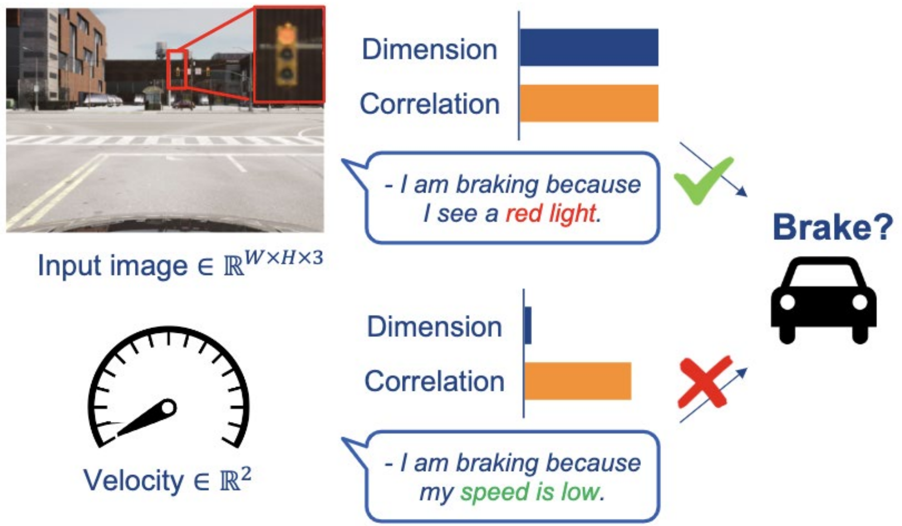
挑战（6/8）- Interpretability

Summary of the different forms of interpretability



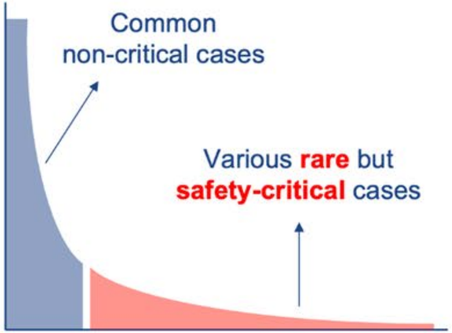
They aid in human comprehension of the **decision-making processes** of end-to-end models, **perception failures**, and the **reliability of the outputs**.

挑战（7/8）- Causal Confusion

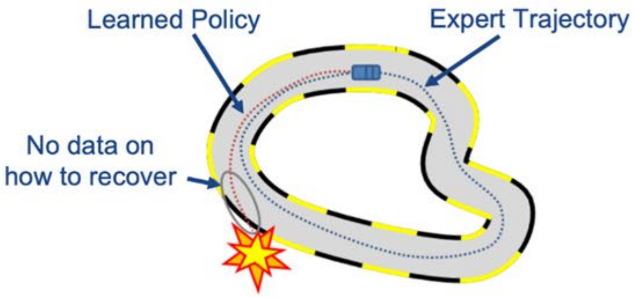


- Driving is a task that exhibits **temporal smoothness**, which makes past motion a reliable predictor of the next action.
- However, methods trained with **multiple frames** can become overly reliant on this shortcut. This is referred to as the **copycat problem** and is a manifestation of **causal confusion.**

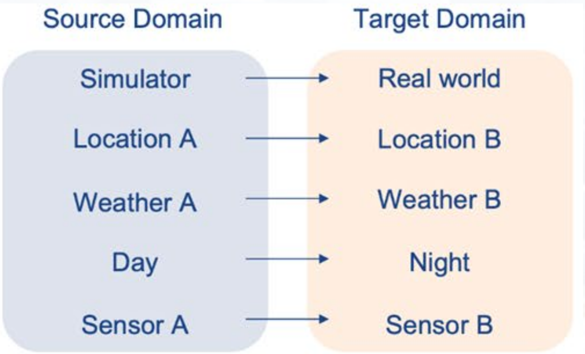
挑战（8/8） - Robustness and Generalization



(a) Long-tailed Distribution



(b) Covariate Shift



(c) Domain Adaptation

OpenDriveLab

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

End-to-end Autonomous Driving

Future Work

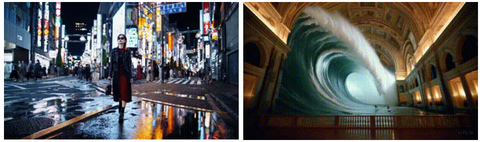
Gap between The Rest and SORA

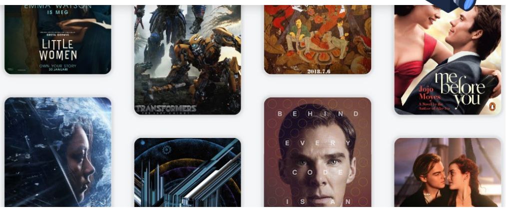
- **High-quality Video Data** Need massive collection, including films
  - Long duration(> 60s) , high resolution, large motion, comprehensive scenarios
  - Existing **public video datasets are inadequate** in both quality and duration. (e.g., webvid 10M, internvid, vimeo 25M)
  - ***Film data*** is a good source. (movies, documentaries, animations, etc.)

- **Spatial-temporal VAE** Need to build from scratch
  - **Videos are highly redundant** in **temporal dimension**,
  - thus should be **compressed** for efficiency.
  - The key ingredient to **long video generation.**
    - SVD (<5s) → SORA (60s)

- **Diffusion Model Architecture** Public, DiT (But need to be extended to video version)
  - Temporal attention alone is not efficient for modeling large motions.
  - **We need (global) spatial-temporal attention**, which requires more compute but yields better results after scaling.

- **Highly-capable Video Captioner** Not available, but have some *weaker* public solutions
  - Annotating **accurate and expressive captions** for each video clip.
  - Public solutions: LLaVA, VideoChat, GPT-4.

	SVD (Previous Sota)	SORA
		
	short duration, small motion, simple scenario	long duration, large motion, complex scenario
Functionality	image → video 1024 x 576 x 4s x 6Hz	text/image/video → video 1920 x 1080 x 60s x 30Hz
		Owe to the compression by spatial-temporal VAE
Model	Spatial VAE + UNet	Spatial-Temporal VAE + DiT (More scalable)
Training data	152M (0.15B) video clips (low quality, short duration, **small motion**, **simple scenarios**), mainly crawled from YouTube	>> 1B video clips *(approx.)* (**high quality**, long duration, **large motion**, **comprehensive scenarios**)



Most of YouTube Videos are noisy, short-period, and in small motion.

We may need film data, which are long-period, highly-dynamic, and highly-aesthetic.
(movies, documentaries, animations, etc.)

Milestone in Computer Vision (1/2)

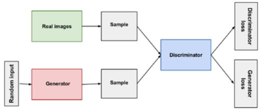
2014.6

Citation: 65k

GAN Université de Montréal

Generative adversarial network

A prominent framework for **generative AI.**

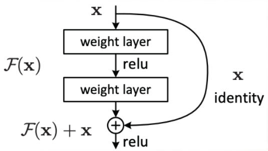


2015.12

Citation: 200k

ResNet

Residual connections enables building deep neural



- Unleashing the power of **deep neural networks.**
- **Stacking more layers ->** better performance.

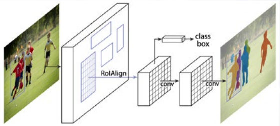
2017.3

Citation: 32k

Mask R-CNN Meta

Segment instances via an **effective mask branch.**

Simple yet effective **object centric learning paradigm.**

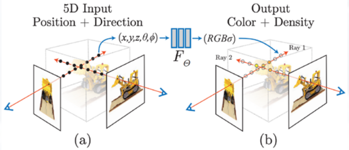


2020.3

Citation: 5k

NeRF Berkeley

Represent **Scenes as Neural Radiance Fields** for View Synthesis



Bridging 2D and 3D representations with multi-view images.

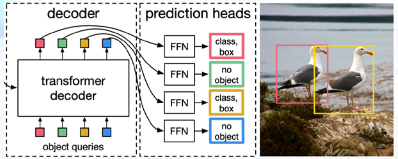
2020.5

Citation: 9k

DETR Meta

Leverage **transformers** for end-to-end **object detection.**

Representing objects as learnable queries dominates vision tasks for its *simplicity* and *flexibility*.



Milestone in Computer Vision (2/2)

Scaling up data and model is the key to success !

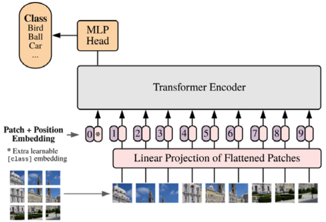
2020.10

Citation: 30k

Vision Transformer

Pure transformer *works in vision regime.*

Unifying the model architecture of vision and language, enabling multi-modal researches.

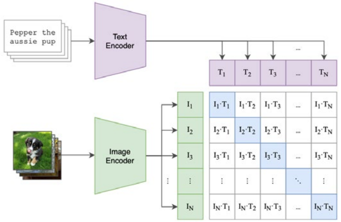


2021.2

Citation:12k

CLIP

Connect texts and images via large-scale contrastive learning.



- A first-time success in leveraging **internet-scale multi-modal corpus.**
- Fueling multi-modal researches like open-vocabulary vision tasks, text-to-image generations, etc.

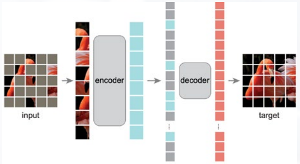
2021.11

Citation: 4k

MAE

Scale up vision models via **masked image modeling.**

A huge success of **self-supervised learning** in vision.

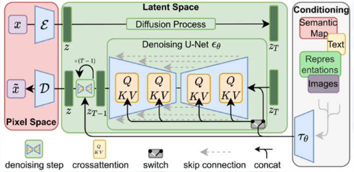


2021.12

Citation: 5k

Latent Diffusion

High-quality image generation via diffusion in the latent space.



- Pre-trained on **billion-scale image-text** data.
- Opening the era of **high-quality content generation**.

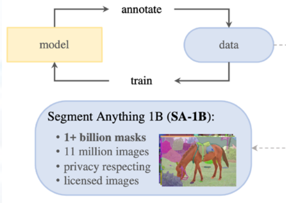
2023.4

Citation: 1k

SAM

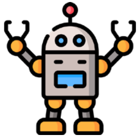
Annotate **1 billion mask with semi-supervised model** in the loop.

A promising way to enlarge vision data via **semi-supervised data engine.**



(c) **Data**: data engine (top) & dataset (bottom)

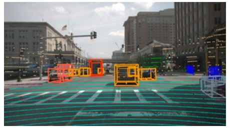
Towards Intelligent, Reliable and Generalizable Autonomy



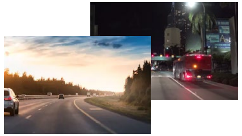

Data-centric Pipeline

Data Collection

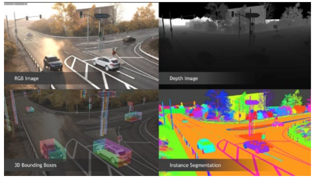
Motional
ARGO AI
lyft



YouTube

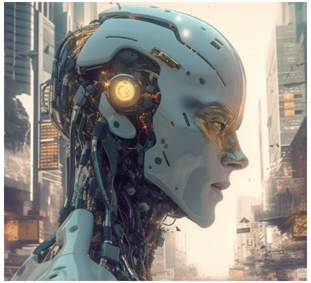


Data Generation



Pre-training DriveCore

Foundation Model

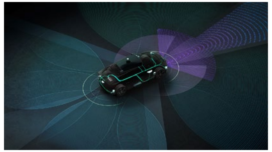



Integrated and General AGI
for autonomous driving

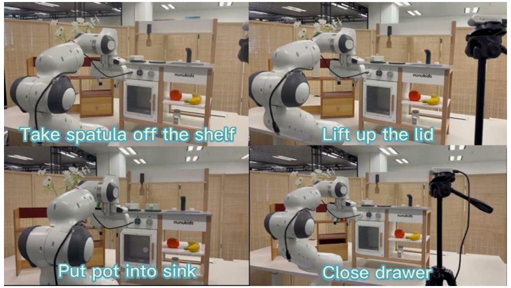
How to formulate?
What's the objective goal?
GenAD (our on-going project)

Applications

Autonomous Driving



Embodied AI

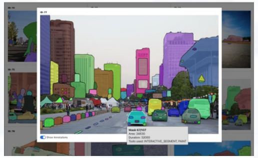


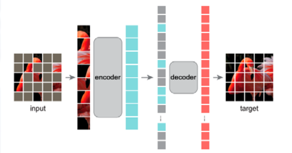
Foundation Models

NLP (LLM)

ChatGPT

General CV





AD System

- Language Interpreter
- Driving Knowledge
- Any more?

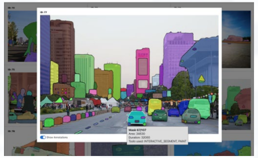
- Vision Abstractor
- Auto-labeling
- Any more?

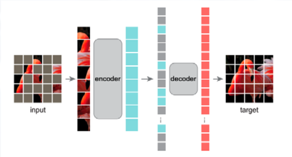
Foundation Models (cont'd)

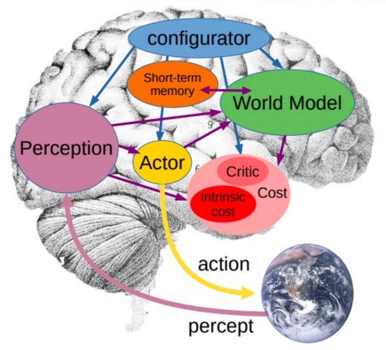
NLP (LLM)

ChatGPT

General CV



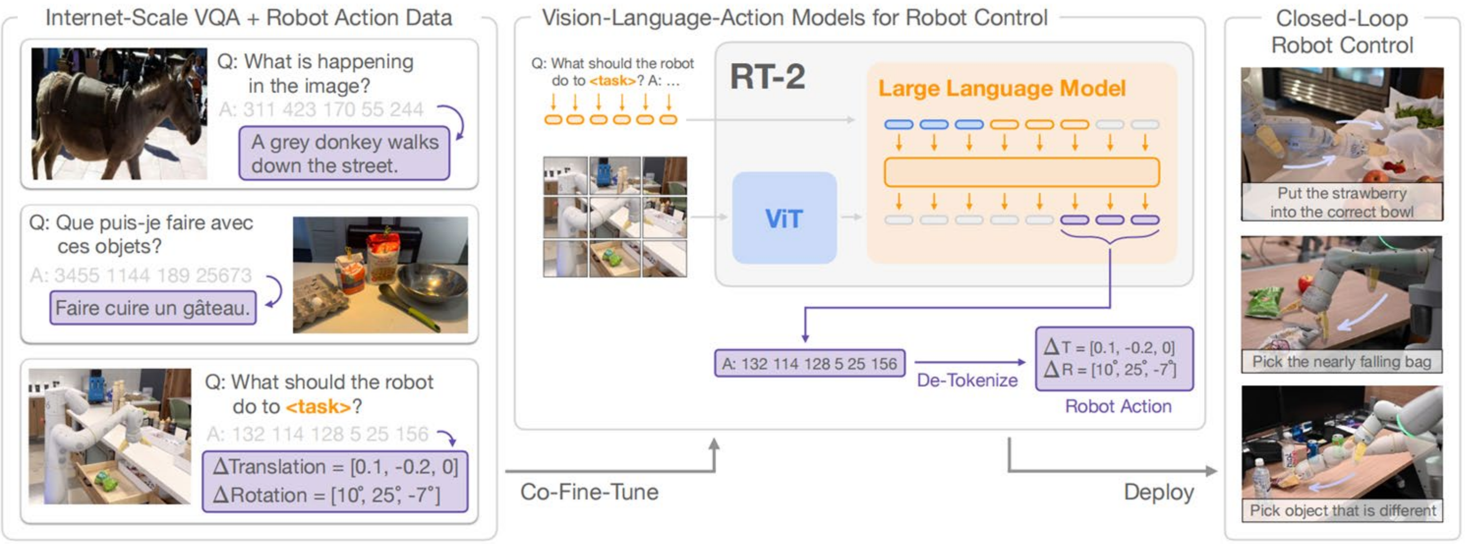




AD System

- Multimodality
- Intelligence
- Generalization

Insight from 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

Analogy to General Domains in CV/NLP/Robotics

General Large Models

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	**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	✗

nuScenes: 4.5h

Industrial Large Models (Application)

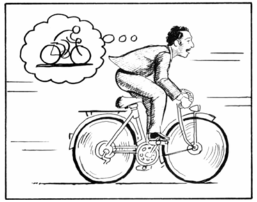
If taken seriously for AD: lots of compute (at least 200 A100s) + **massive amount of data** (at least 10k hours of diverse, high-quality data)

OpenDriveLab

Trending: Recent Work on World Model

From simulated agents to real-world driving systems

RL Agents — 18.3 — 20.3 — **Vision** — 22.6 — 23.6 — **Driving** — 23.6

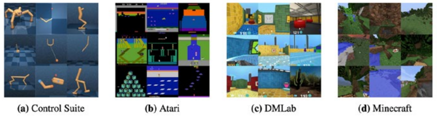


18.3

World Models: Training agents inside their dreams

20.3

Dreamer V1/2/3: Towards general agents with scalable world models



22.6

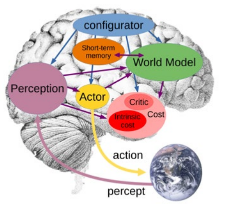
23.6

23.6

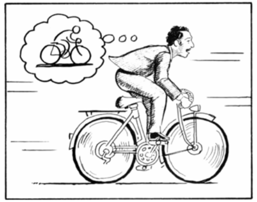
Trending: Recent Work on World Model

From simulated agents to real-world driving systems

Position Paper (by LeCun)
Positioning the developments of world models



RL Agents — 18.3 — 20.3 — Vision — 22.6 — 23.6 — Driving — 23.6

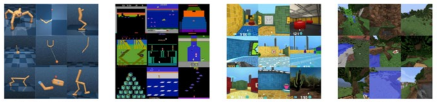


18.3

World Models: Training agents inside their dreams

20.3

Dreamer V1/2/3: Towards general agents with scalable world models



(a) Control Suite (b) Atari (c) DMLab (d) Minecraft

22.6

23.6

I-JEPA: Capturing visual knowledge in self-supervised manner

23.6

Scaling up world models on large corpus of realistic driving videos

General World Model: inhouse data collected around the globe

TESLA

GAIA-1: 4700 hours of driving videos collected in London

World model to generate videos of the driving scenario. **Then what? Is it useful for downstream tasks?** (To be validated)

OpenDriveLab

Personal Take on Foundation Models into Autonomous Driving

End-to-end
Auto Driving

Pros:

1. Scalability
2. Global optimization
3. Easy-to-embed Infra

For:

→ Generalization/Robustness
→ Performance
→ Feasibility for deployment

Personal Take on Foundation Models into Autonomous Driving

End-to-end Auto Driving

Pros:

1. Scalability
2. Global optimization
3. Easy-to-embed Infra

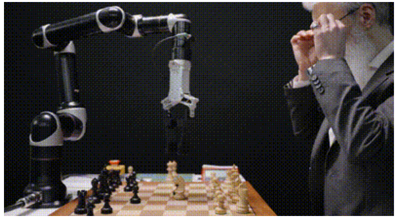
For:

→ Generalization/Robustness
→ Performance
→ Feasibility for deployment

Research OpenAI

Video generation models as world simulators

Mind-blowing Part



Weakness Samples

Some rumors:

- 0.8M GPUs
- 50B video clips from Microsoft (ref: Youtube has 13B videos)
- This a side project from OpenAI

Personal Take on Foundation Models into Autonomous Driving

End-to-end Auto Driving

Pros:

1. Scalability
2. Global optimization
3. Easy-to-embed Infra

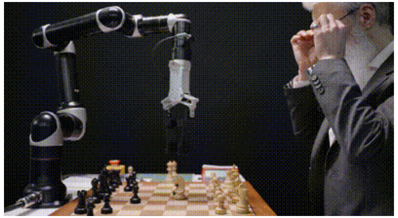
For:

→ Generalization/Robustness
→ Performance
→ Feasibility for deployment

Research OpenAI

Video generation models as world simulators

Mind-blowing Part



Weakness Samples

Some rumors:

- 0.8M GPUs
- 50B video clips from Microsoft (ref: Youtube has 13B videos)
- This a side project from OpenAI

Towards Intelligent, Reliable and Generalizable System

Data-driven | Alg-driven | Metric-driven

- Scaling data in all levels with self-supervised learning
- Simulating the physical world
- Rule of thumbs from foundation models
- Authentic evaluation metric.
- Guarantee reliability and safety.

→ ***Interaction*** between agents and env/physical world

→ Pixel-level ***not*** suffice
Actions require latent abstractions. Depends on task.

End-of-Lecture

OpenDriveLab

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

主流工作选讲 - Part 3

End-to-end Autonomous Driving

OpenDriveLab

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

GAIA-1

End-to-end Autonomous Driving

GAIA-1 | Motivation

https://arxiv.org/abs/2309.17080

Want to solve the problem:

How to predict the various potential outcomes that may emerge in response to the vehicle's actions as the world evolves?

Current limitations:

- **Labeled data:** hard to obtain at scale

- **Simulated data:** low-dimensional representations; hard to capture the complexities of real-world

GAIA-1 can:

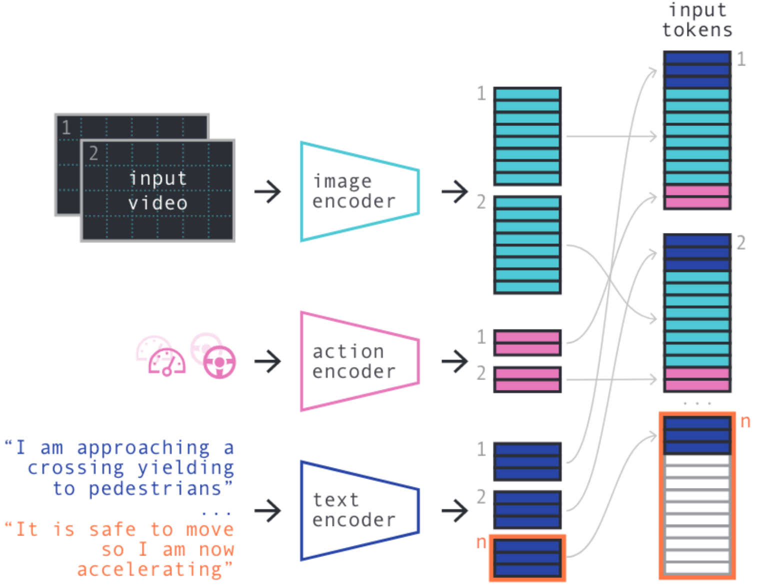
- Combine world models and generative video generation

- Ensure the realism of generative video models and learn meaningful representations

GAIA-1 | Method

https://arxiv.org/abs/2309.17080

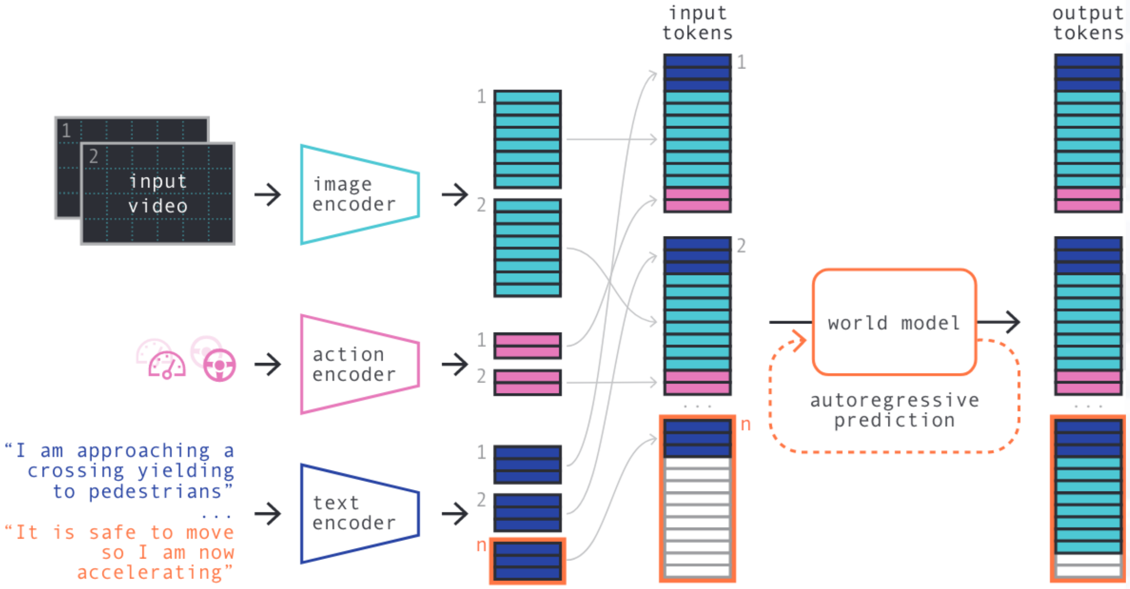
首先，将来自所有输入模态（视频、文本、动作）的信息编码为一个通用的表示，图像、文本和动作被编码为一系列token



GAIA-1 | Method

https://arxiv.org/abs/2309.17080

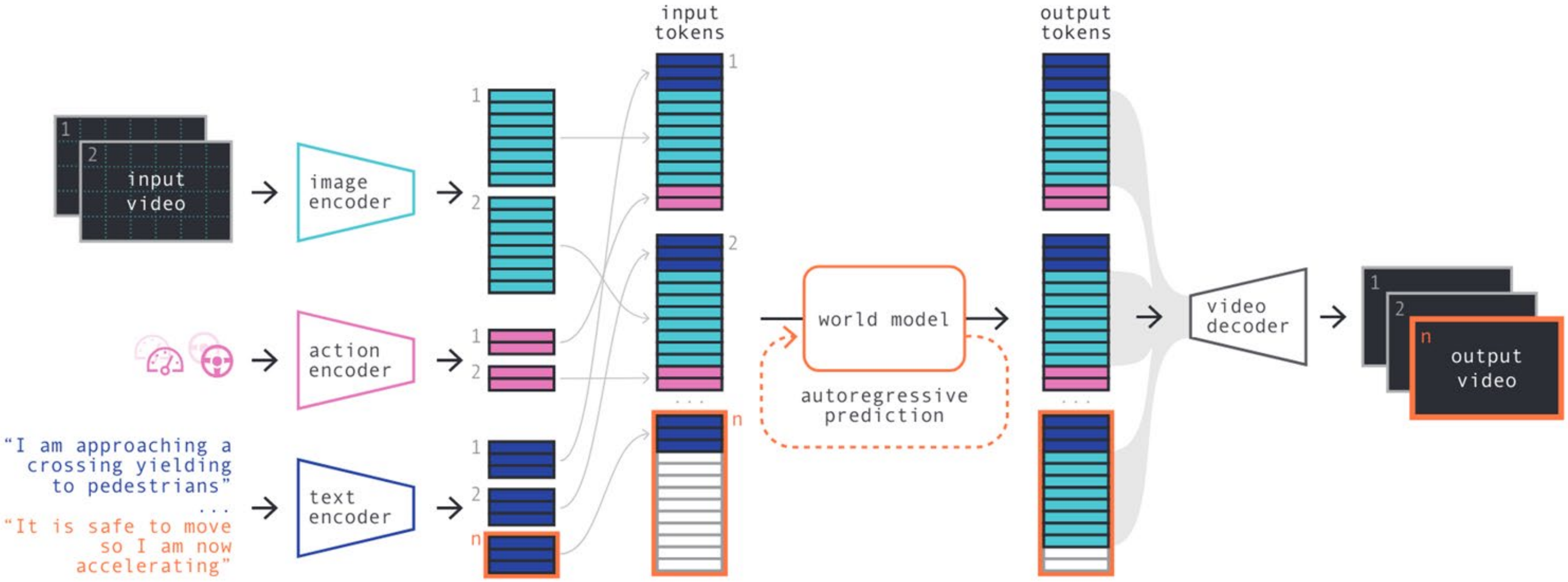
世界模型是一个自回归transformer，它以过去的图像、文本和动作token为条件来预测下一个图像token



GAIA-1 | Method

https://arxiv.org/abs/2309.17080

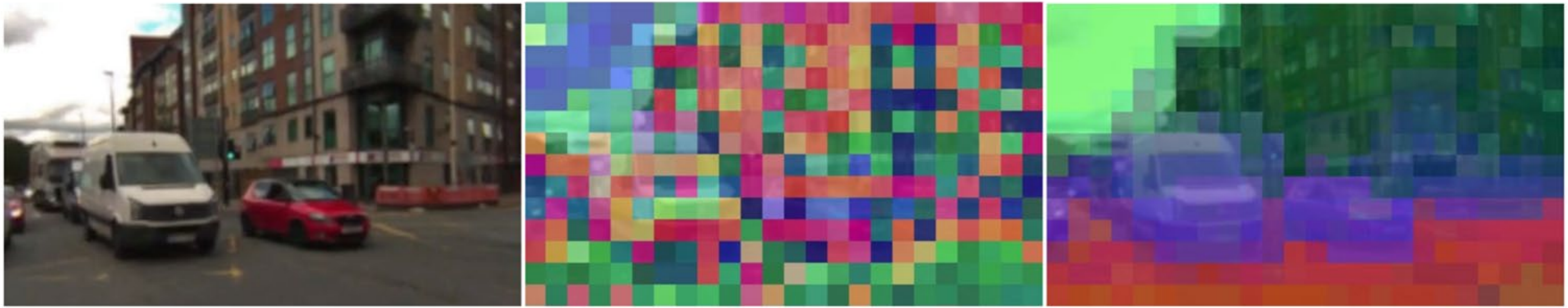
最后，视频解码器以更高的时间分辨率将预测的图像token映射回像素空间



GAIA-1 | Experiment

https://arxiv.org/abs/2309.17080

In **Image Tokenizer,** GAIA-1 guides the compression towards meaningful representations by regressing to the latent features of a pre-trained **DINO** model.

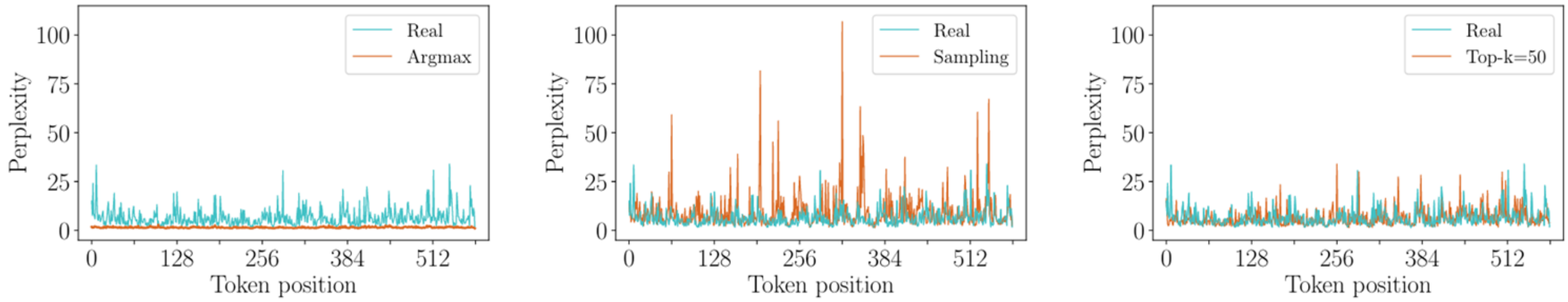


(a) Input image (b) Base VQ-GAN tokens (c) DINO-distilled tokens

GAIA-1 | Experiment

https://arxiv.org/abs/2309.17080

To encourage **diversity** as well as **realism**, GAIA-1 employs **top-k sampling** to sample the next image token from the top-k most likely choices.



(a) Argmax. (b) Sampling. (c) Top-k sampling.

https://arxiv.org/abs/2309.17080

GAIA-1 | Experiment

Images generated by GAIA-1

OpenDriveLab

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

EgoStatus

End-to-end Autonomous Driving

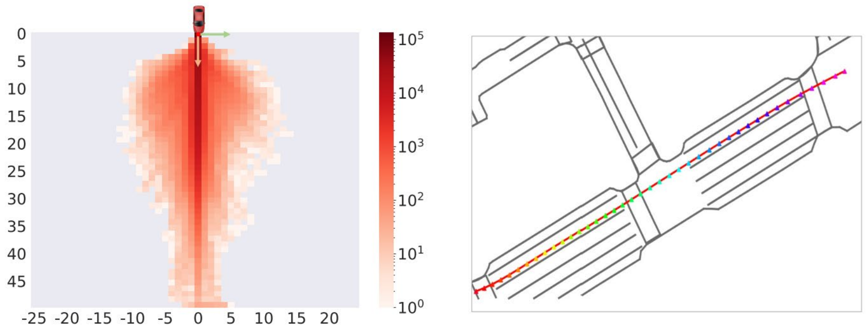
EgoStatus | Motivation

https://arxiv.org/abs/2312.03031

Current prevailing end-to-end autonomous driving methods commonly use **nuScenes** for **open loop evaluation of their planning behavior.**

However:

- **NuScenes** dataset, characterized by **relatively simple driving scenarios**, leads to an underutilization of perception information in end-to-end models.



(a) Trajectory Heatmap

(b) Typical Scene of nuScenes

EgoStatus | Motivation

https://arxiv.org/abs/2312.03031

Current prevailing end-to-end autonomous driving methods commonly use **nuScenes** for **open loop evaluation of their planning behavior.**

However:

- **NuScenes** dataset, characterized by **relatively simple driving scenarios**, leads to an underutilization of perception information in end-to-end models.
- **ADMLP** recently points out that a simple MLP network can also achieve state-of-the-art planning results, **relying solely on the ego status information.**

EgoStatus | Motivation

https://arxiv.org/abs/2312.03031

Current prevailing end-to-end autonomous driving methods commonly use **nuScenes** for **open loop evaluation of their planning behavior.**

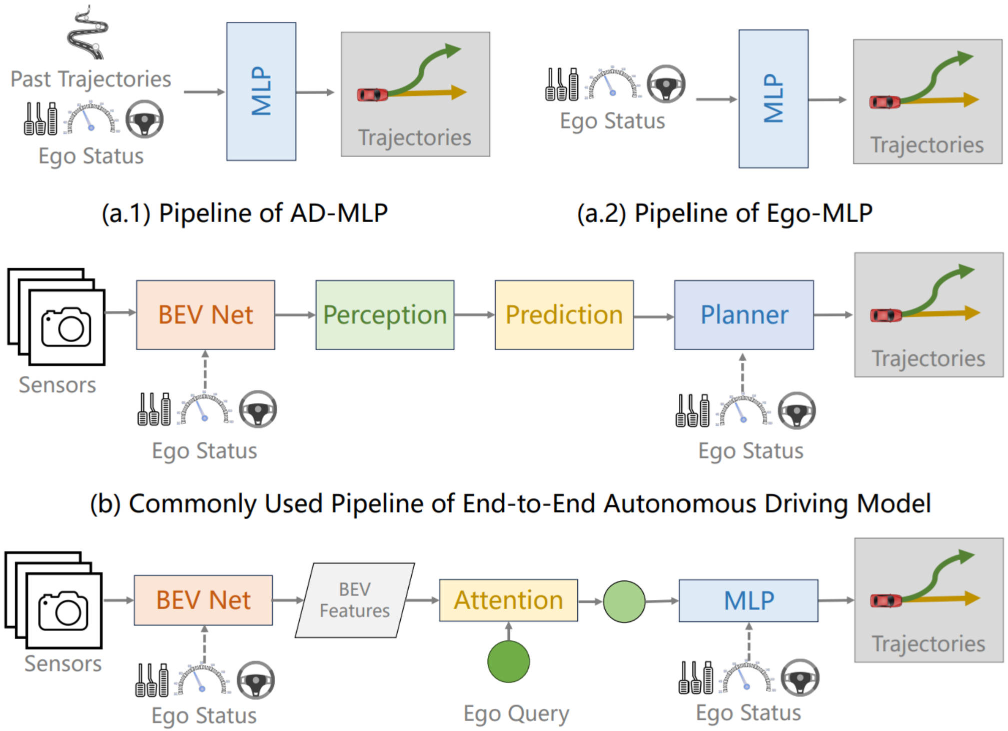
However:

- **NuScenes** dataset, characterized by **relatively simple driving scenarios**, leads to an underutilization of perception information in end-to-end models.
- **ADMLP** recently points out that a simple MLP network can also achieve state-of-the-art planning results, **relying solely on the ego status information.**

Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving?

EgoStatus | Method

https://arxiv.org/abs/2312.03031



(a.1) Pipeline of AD-MLP (a.2) Pipeline of Ego-MLP

(b) Commonly Used Pipeline of End-to-End Autonomous Driving Model

(c) Pipeline of Our BEV-Planner

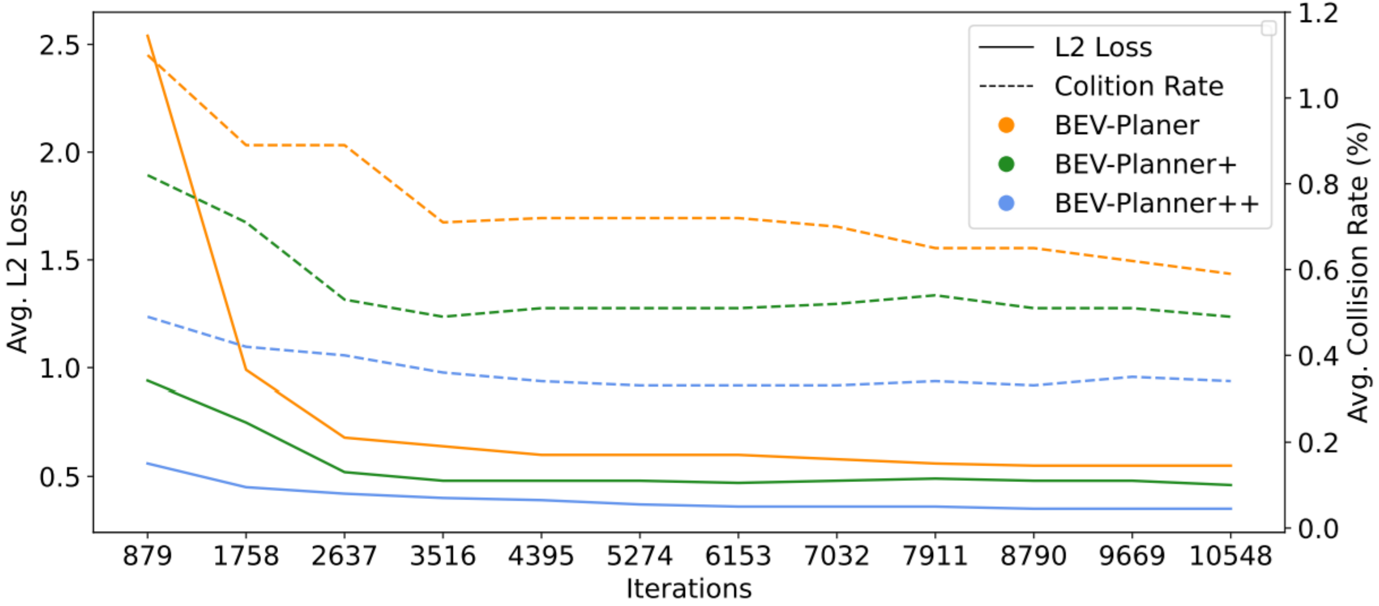
EgoStatus | Experiment

https://arxiv.org/abs/2312.03031

ID	Method	Ego Status in BEV	Ego Status in Planer	L2 (m) ↓ 1s	L2 (m) ↓ 2s	L2 (m) ↓ 3s	L2 (m) ↓ Avg.	Collision (%) ↓ 1s	Collision (%) ↓ 2s	Collision (%) ↓ 3s	Collision (%) ↓ Avg.	Intersection (%) ↓ 1s	Intersection (%) ↓ 2s	Intersection (%) ↓ 3s	Intersection (%) ↓ Avg.	ckpt. source
0	ST-P3	✗	✗	1.59†	2.64†	3.73†	2.65†	0.69†	3.62†	8.39†	4.23†	2.53†	8.17†	14.4†	8.37†	Official
1	UniAD	✗	✗	0.59	1.01	1.48	1.03	0.16	0.51	1.64	0.77	0.35	1.46	3.99	1.93	Reproduce
2	UniAD	✓	✗	0.35	0.63	0.99	0.66	0.16	0.43	1.27	0.62	0.21	**1.32**	3.63	1.72	Official
3	UniAD	✓	✓	0.20	0.42	0.75	0.46	0.02	**0.25**	0.84	0.37	**0.20**	1.33	**3.24**	**1.59**	Reproduce
4	VAD-Base	✗	✗	0.69	1.22	1.83	1.25	0.06	0.68	2.52	1.09	1.02	3.44	7.00	3.82	Reproduce
5	VAD-Base	✓	✗	0.41	0.70	1.06	0.72	0.04	0.43	1.15	0.54	0.60	2.38	5.18	2.72	Official
6	VAD-Base	✓	✓	0.17	0.34	0.60	0.37	0.04	0.27	**0.67**	**0.33**	0.21	2.13	5.06	2.47	Official
7	GoStright	-	✓	0.38	0.79	1.33	0.83	0.15	0.60	2.50	1.08	2.07	8.09	15.7	8.62	-
8	Ego-MLP	-	✓	**0.15**	**0.32**	0.59	**0.35**	**0.00**	0.27	0.85	0.37	0.27	2.52	6.60	2.93	
9	BEV-Planner*	✗	✗	0.27	0.54	0.90	0.57	0.04	0.35	1.80	0.73	0.63	3.38	7.93	3.98	-
10	BEV-Planner	✗	✗	0.30	0.52	0.83	0.55	0.10	0.37	1.30	0.59	0.78	3.79	8.22	4.26	-
11	BEV-Planner+	✓	✗	0.28	0.42	0.68	0.46	0.04	0.37	1.07	0.49	0.70	3.77	8.15	4.21	-
12	BEV-Planner++	✓	✓	0.16	0.32	**0.57**	**0.35**	**0.00**	0.29	0.73	0.34	0.35	2.62	6.51	3.16	-

EgoStatus | Experiment

https://arxiv.org/abs/2312.03031



EgoStatus | Experiment

https://arxiv.org/abs/2312.03031