OpenDriveLab

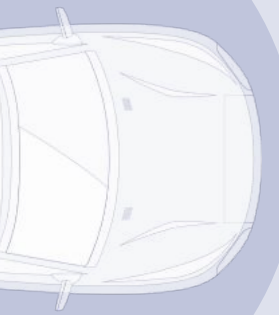
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

端到端自动驾驶：前沿与挑战
End-to-end Autonomous Driving: Challenges and Frontiers

Li Chen
OpenDriveLab at Shanghai AI Lab
June 10, 2024

Outline

- 课程目标
  - 掌握端到端自 动驾驶系统的特点与基 础方法
  - 了解端到端自 动驾驶算法的典型 问题与挑战
  - 了解端到端自 动驾驶前沿算法



Outline

- **端到端自动驾驶系统概述** / Introduction
  - **背景与动机** / Motivation
  - **发展历程** / Roadmap
- **基础方法** / Method
- **验证与评测** / Benchmarking
- **关键研究内容与挑战** / Frontiers and Challenges
- **未来可能与讨论** / Future Trends and Discussions
- Q & A

OpenDriveLab

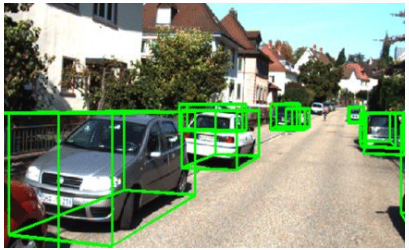
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

端到端自动驾驶系统概述 / Introduction

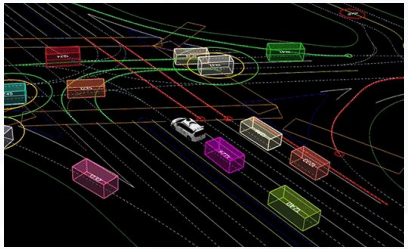
Autonomous Driving (AD) Tasks

Challenge | Various weathers, illuminations, and scenarios

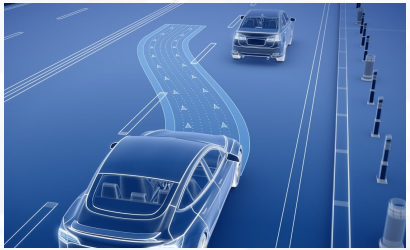
Perception → (Bounding boxes) → Prediction → (Waypoints) → Planning → (Trajectory)



What are around?



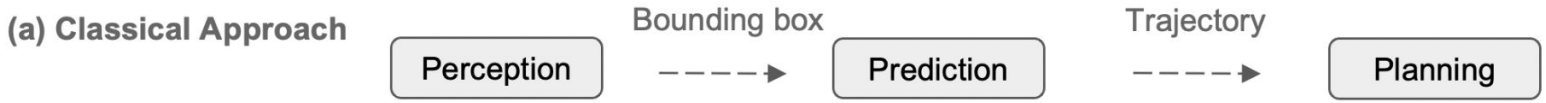
How will they go in the future?



Where should I go?

Conventional Autonomous Driving (AD) Systems

Modular-based



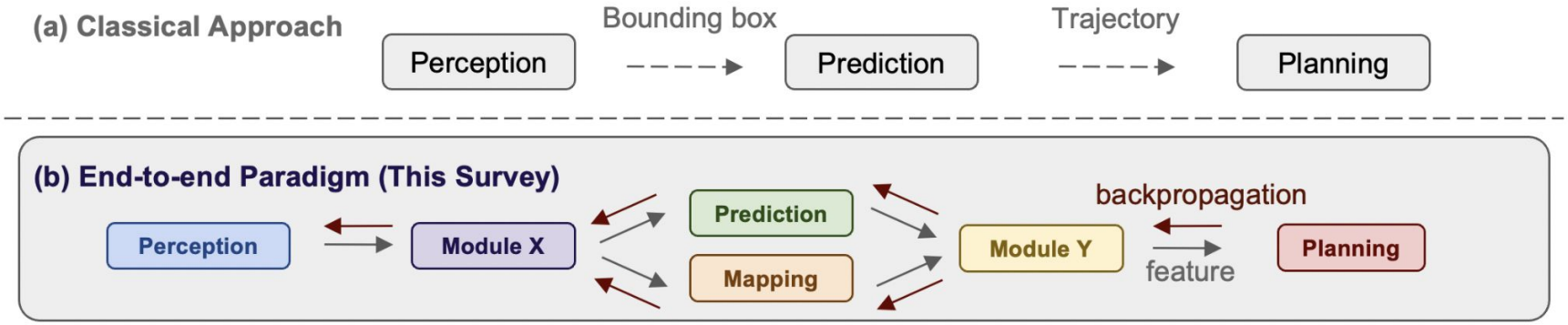
Pros

- **Independent** teams for module development
  - Dataset friendly
  - Quantitatively evaluation for intermediate tasks
  - Great interpretability
- Parallel onboard deployment

Cons

- Error accumulation, information loss. **Results, instead of features** are traversed across modules.
  - Openset problems
  - Long-tail problems

Motivation | Why End-to-end Autonomous Driving?



End-to-end autonomous driving system - A suite of fully differentiable programs that:

- take raw sensor data as input
- produce a plan and/or low-level control actions as output

Motivation | Why End-to-end Autonomous Driving?

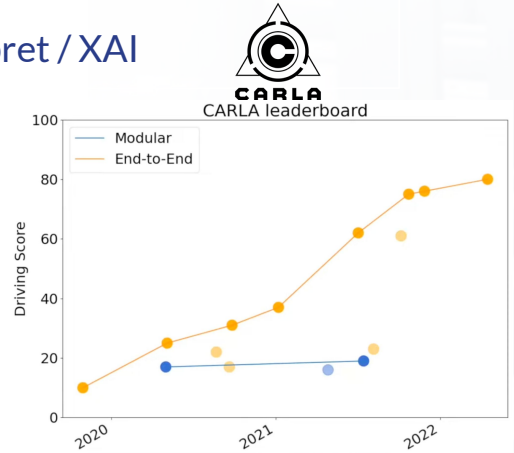
Advantages

+ **Simplicity** in combining all modules into a single model that can be **joint trained**
+ Preventing cascading errors in modular design
+ Directly optimized **toward the ultimate task**, planning / trajectory prediction
+ Computational efficiency (all shared backbone), production-level friendly

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

Disadvantages

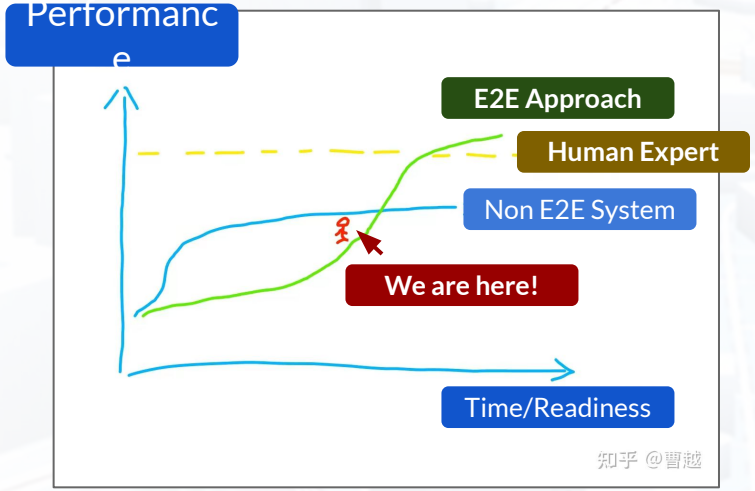
- Closed-loop evaluation only available in simulator / onboard test
- Lack of real-world data
- Hard to interpret / XAI



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 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.

E2E Robot

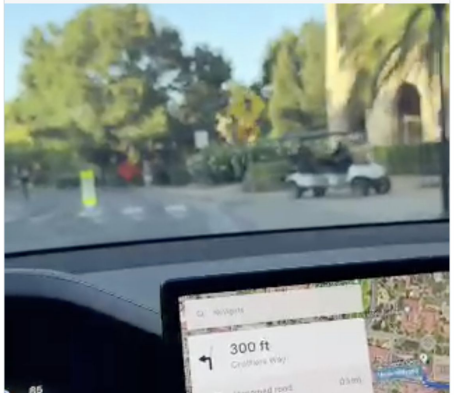
Tesla Optimus @Tesla_Optimus · Sep 24
Optimus can now sort objects autonomously

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!

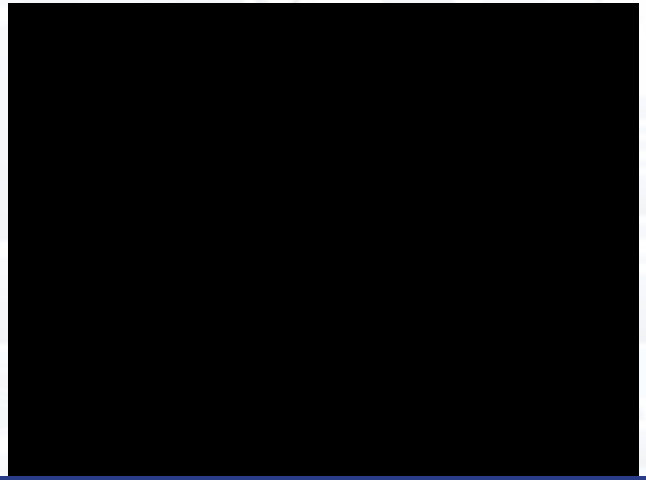
Elon Musk @elonmusk · Aug 26
twitter.com/i/broadcasts/1...



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



CNBC Report | Discussion Thread on Zhihu | Live Stream

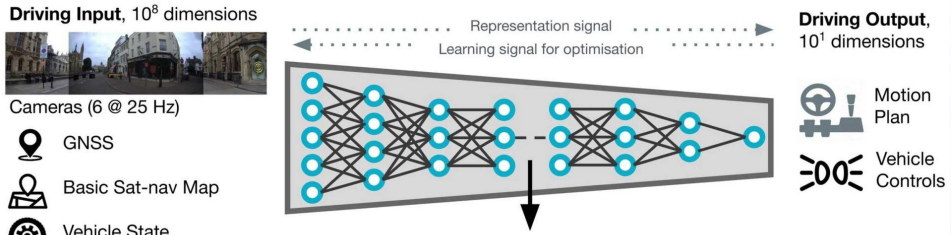
OpenDriveLab

Trending | End-to-end Autonomous Driving

Industry

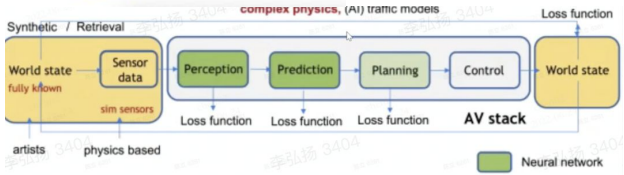
And many others …

WAYVE



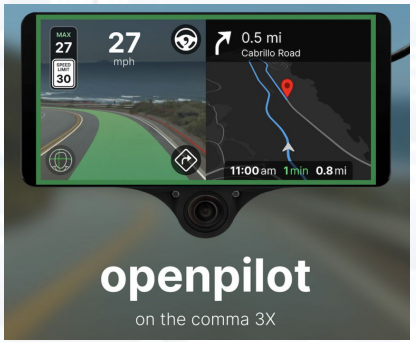
Semantics, geometry, motion prediction.

NVIDIA



comma.ai

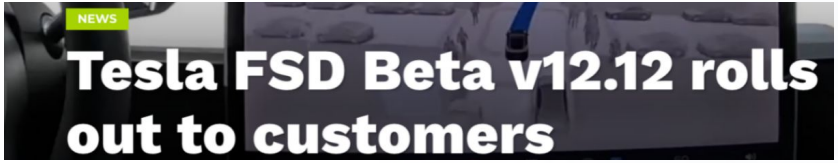
- *Openpilot is an open source driver assistance system.*
- *Openpilot performs the functions of Automated Lane Centering (ALC) and Adaptive Cruise Control (ACC) for 250+ supported car makes and models.*



https://arxiv.org/abs/2206.08176

Application | End-to-end Autonomous Driving

Industry

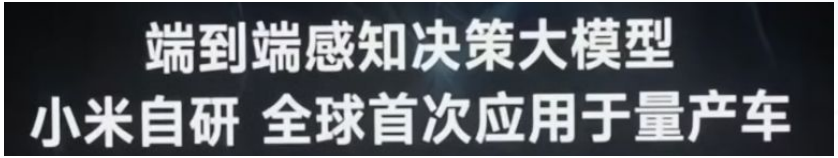


- *Tesla website*

- *Mobileye at CES 2024*

- *Wayve website*

- *XPENG (小鹏) website*



- *Mi (小米) Automobile Technological Event 2024*

Trending | End-to-end Autonomous Driving

Academia / Hybrid

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Transfuser, KING, Misconceptions, DA-RB, etc.

TEXAS The University of Texas at Austin

LBC, WoR, LAV, etc.

UNIVERSITY OF TORONTO

Uber

Waabi

NMP, P3, MP3, UniSim (sort of)

BOSTON UNIVERSITY

LbW, SelfD, CaT, AnyD, etc.

ETH zürich

Roach, etc

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

MMFN, Carl-lead, PMP, etc

Valeo

GRI

Horizon Robotics

VAD, VAD v2

UCLA

MetaDrive, ACO, CAT, etc

USC

GPT-Driver, Agent-Driver, etc

Berkeley UNIVERSITY OF CALIFORNIA

GenAD, etc

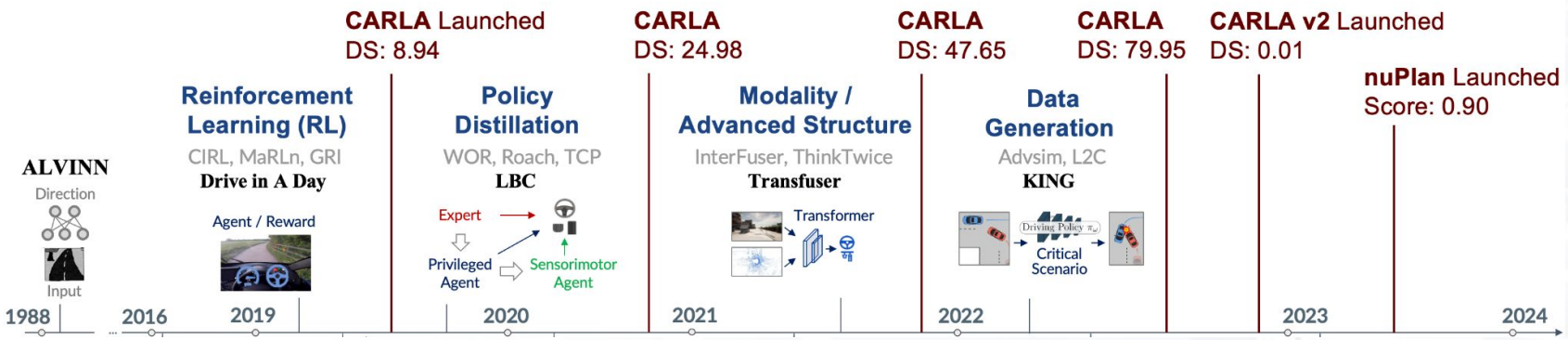
清華大學 Tsinghua University

SparseDrive, ADAPT, etc

OpenDriveLab

Work from our Team introduced later

Roadmap | End-to-end Autonomous Driving



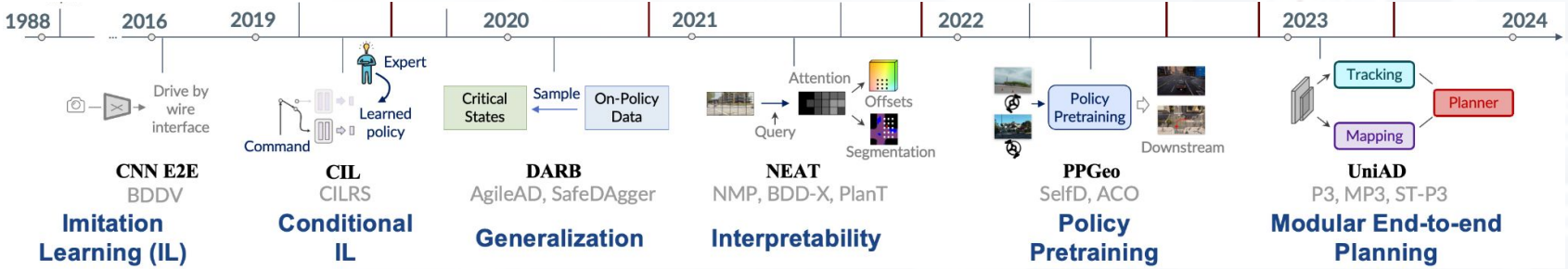
Summary (1/2)

- CARLA leaderboard gets much improved over the years. With new mapping / scenarios (leaderboard 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

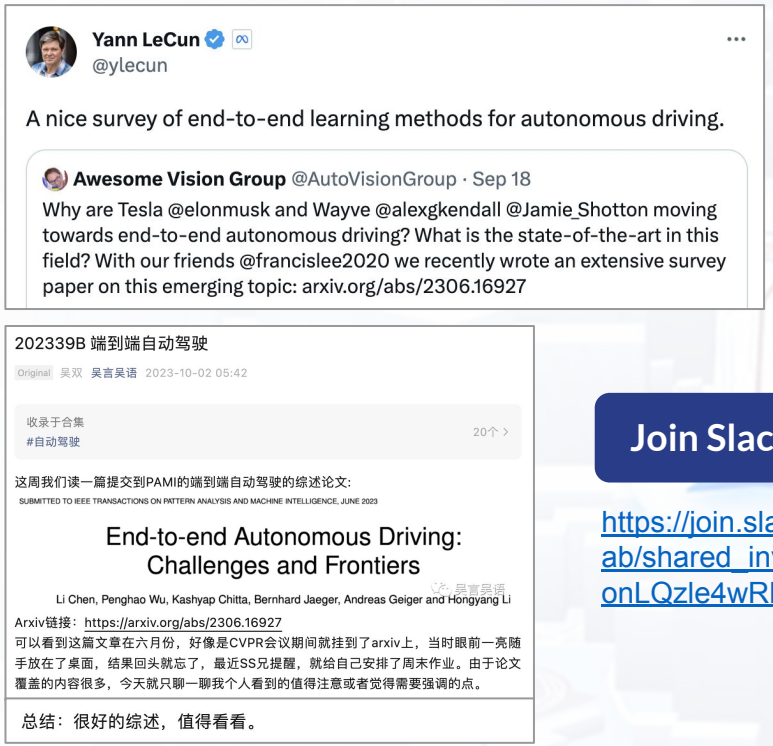
Summary (2/2)

- 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



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



Join Slack Discussions!

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

[1] Chen et al. End-to-end Autonomous Driving: Challenges and Frontiers. arXiv, 2023.

OpenDriveLab

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

基础方法 / Method

Types of Learning

Supervised Learning:

- Dataset: $\{(x_i, y_i)\}$ (x_i = data, y_i = label) Goal: Learn mapping $x \mapsto y$
- Examples: Classification, regression, imitation learning, affordance learning, etc.

Unsupervised Learning:

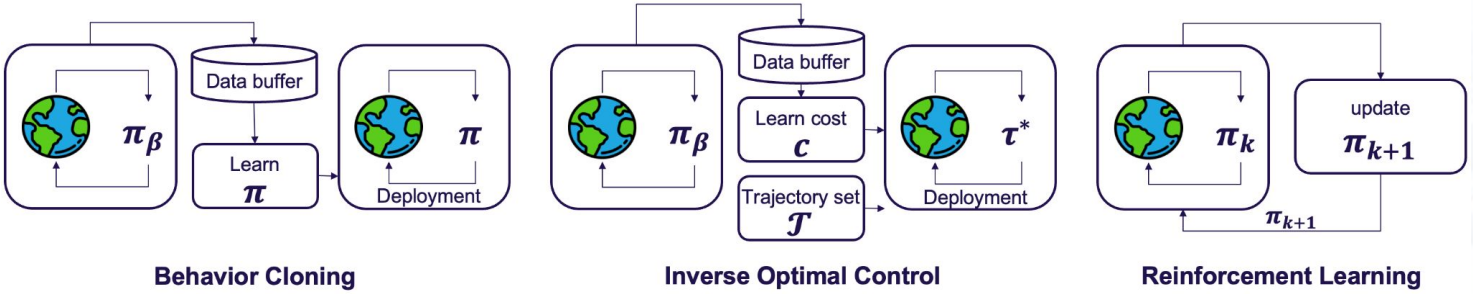
- Dataset: $\{(x_i)\}$ (x_i = data) Goal: Discover structure underlying data
- Examples: Clustering, dimensionality reduction, feature learning, etc.

Reinforcement Learning:

- Agent interacts with environment which provides numeric reward design
- Goal: Learn how to take actions in order to maximize reward
- Examples: Learning of manipulation or control tasks (everything that interacts)

Credit to Andreas Geiger, Lecture: Self-Driving Cars

Types of Learning for E2E AD



Imitation Learning: Learn a policy from **expert demonstrations**

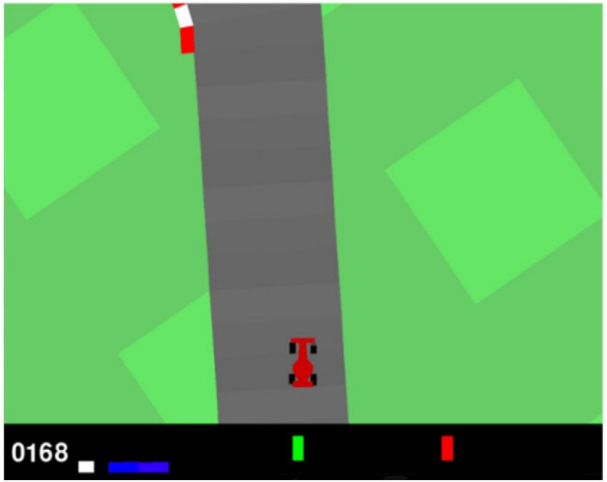
- Expert demonstrations are provided
- Supervised learning problem
- Behavior Cloning, Inverse Optimal Control (Inverse Reinforcement Learning)

Reinforcement Learning: Learn a policy through **trial-and-error**

- No expert demonstrations given
- Agent discovers itself which actions **maximize the expected future reward**
  - The agent interacts with the environment and obtains reward
  - The agent discovers good actions and improves its policy π

Imitation Learning

Trainer
(Human Driver)

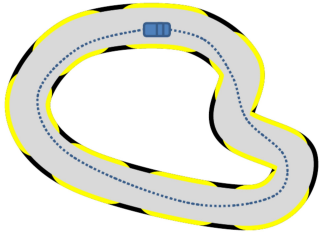


Trainee
(Neural Network)

Credit to Andreas Geiger, Lecture: Self-Driving Cars

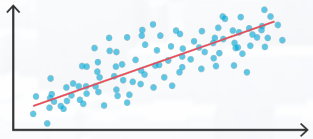
Imitation Learning in a Nutshell

Expert Trajectories

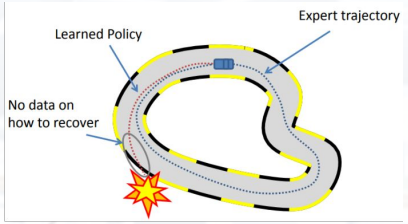


Dataset

Supervised Learning



Test Execution



Hard coding policies is often difficult → Rather use a data-driven approach!

- **Given**: Demonstrations or demonstrator
- **Goal**: Train a policy to mimic decision
- **Variants**: Behavior Cloning, Inverse Optimal Control (Inverse Reinforcement Learning)

Credit to Andreas Geiger, Lecture: Self-Driving Cars

Imitation Learning: Trajectory or Control?

- **Trajectory + PID**
  LBC/Transfuser/LAV/…
- **Control**
  CIL/CILRS/WoR/…

PROS

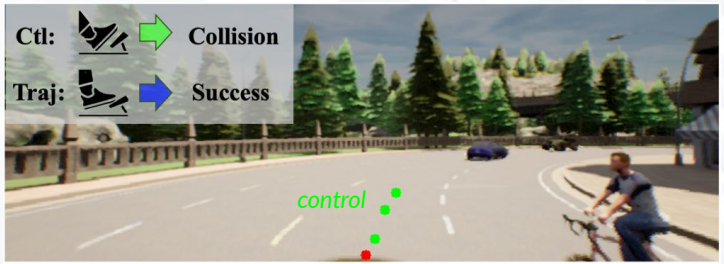
- **long-term strategy**
- Combined with other modules to enhance obstacle avoidance ability

- End-to-end optimization; elegant structure
- Utilization of advanced RL methods

CONS

- Imperfect trajectory following
- Simple controller; **Inertial problem**

- **Focus on current timestamp**; incontinuous output
- Lack of intention-related information
- Coupling with vehicle dynamics, challenging to generalize



- However, no clear performance gap in existing research papers we think.
- While, trajectory may be more appropriate for practical application.

[1] Wu et al. Trajectory-guided Control Prediction for End-to-end Autonomous Driving: A Simple yet Strong Baseline. NeurIPS, 2022.

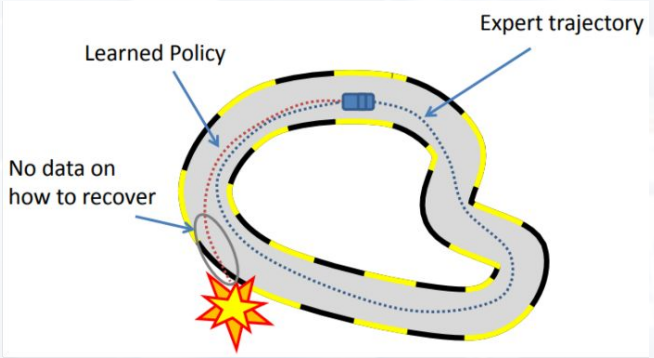
Imitation Learning - Behavior Cloning

Behavior cloning makes I.I.D. assumption

- Next state is sampled from states observed during expert demonstration
- Thus, next state is sampled independently from action predicted by current policy

What if π_θ makes a mistake?

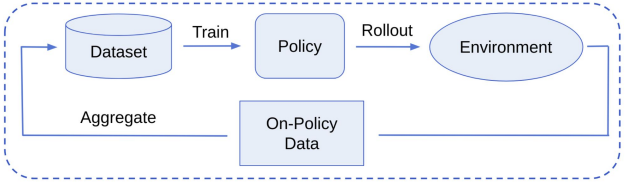
- Enters new states that haven't been observed before
- New states not sampled from same (expert distribution anymore)
- Cannot recover, catastrophic failure in the worst case



What can we do to overcome this train/test distribution mismatch?

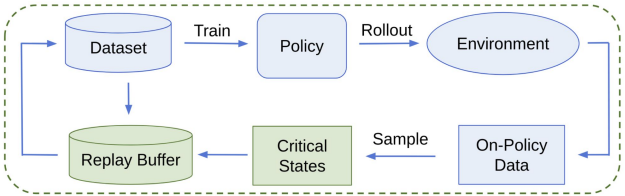
Credit to Andreas Geiger, Lecture: Self-Driving Cars

Imitation Learning - Behavior Cloning



What can we do to overcome this train/test distribution mismatch?
- Data Aggregation (**DAgger**)

- Iteratively build a set of inputs that the final policy is likely to encounter based on previous experience. Query expert for aggregate dataset
- But can easily overfit to main mode of demonstrations
- High training variance (random initialization, order of data)

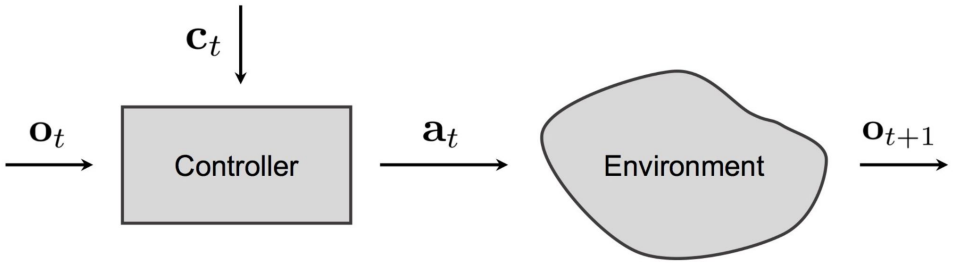


DAgger with Critical States and Replay Buffer

- Sample critical states from the collected on-policy data based on the utility they provide to the learned policy in terms of driving behavior
- Incorporate a replay buffer which progressively focuses on the high uncertainty regions of the policy's state distribution

[1] Ross, et al. A Reduction of Imitation Learning and Structured Prediction on No-Regret Online Learning. AISTATS, 2011.
[2] Prakash, et al. Exploring Data Aggregation in Policy Learning for Vision-based Urban Autonomous Driving. CVPR, 2020.
Credit to Andreas Geiger, Lecture: Self-Driving Cars

Imitation Learning - Conditional Behavior Cloning

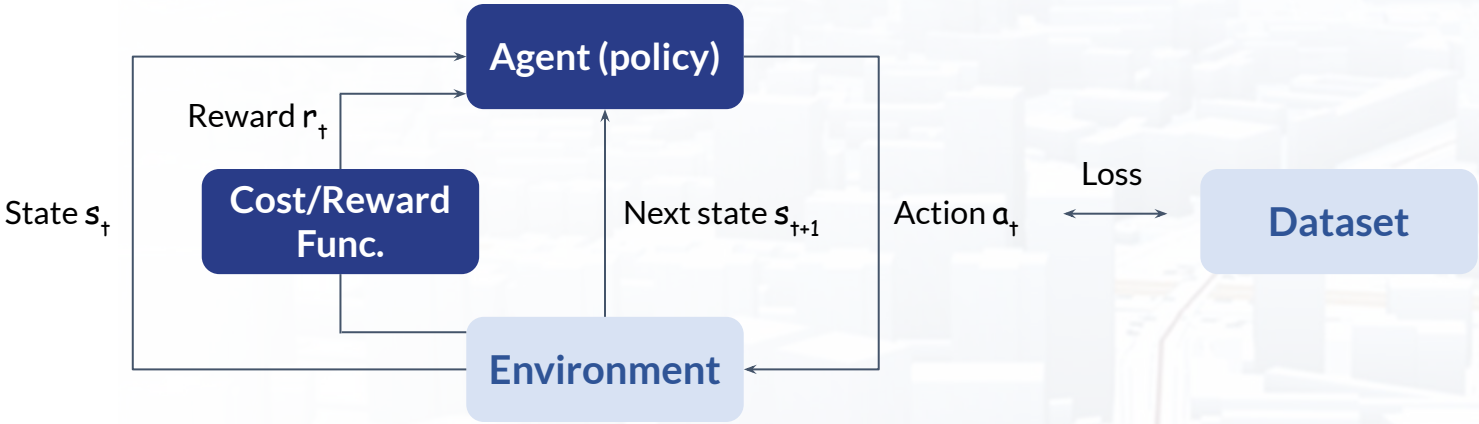


Idea:

- Condition controller on **navigation command** c ∈ {left, right, straight}
- High-level navigation command can be provided by consumer GPS, i.e., telling the vehicle to **turn left/right** or **go straight** at the next intersection
- This removes the task ambiguity induced by the environment
- State s_t : current image Action a_t : steering angle & acceleration

[1] Codevilla et al. End-to-end Driving via Conditional Imitation Learning. ICRA, 2018.

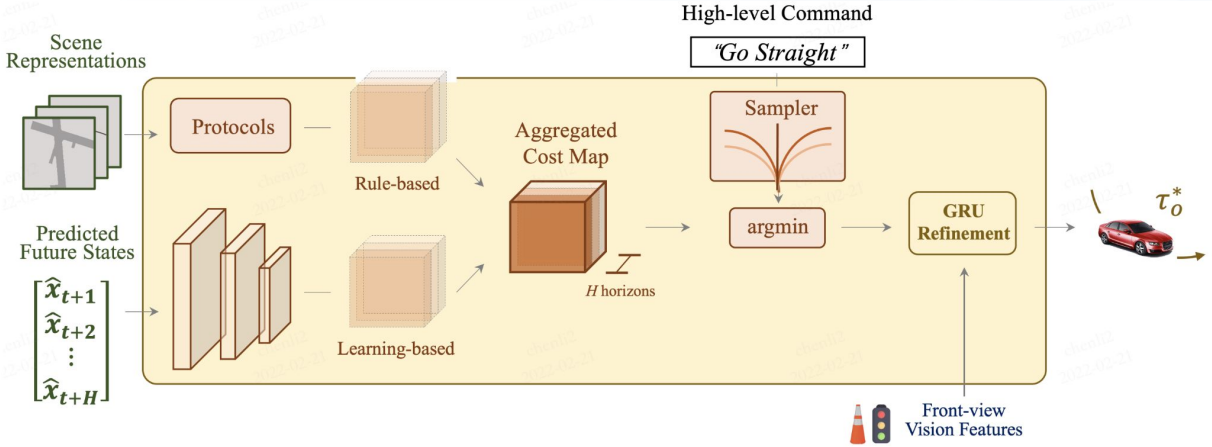
Imitation Learning - Inverse Optimal Control



Inverse Optimal Control (Inverse Reinforcement Learning):

- Agent observes environment state s_t at time t
- Agent sends action a_t at time t, based on the cost/reward function, to the environment
- Environment returns the new state s_{t+1} to the agent

Imitation Learning - Inverse Optimal Control



Objective cost function:

- **Safety Cost**: not collide with other detected objects within future periods; not overlap with road boundaries; maintain a safe distance at high velocity
- **Comfort and Progress**: penalize large lateral acc., jerk, or curvature; reward forwarding to designated directions
- **Learned Cost** Volume: unspecified terms → emergent ability?

[1] Hu et al. ST-P3: End-to-end Vision-based Autonomous Driving via Spatial-Temporal Feature Learning. ECCV, 2022.

Imitation Learning - Summary

Advantages of Imitation Learning, esp. Behavior Cloning

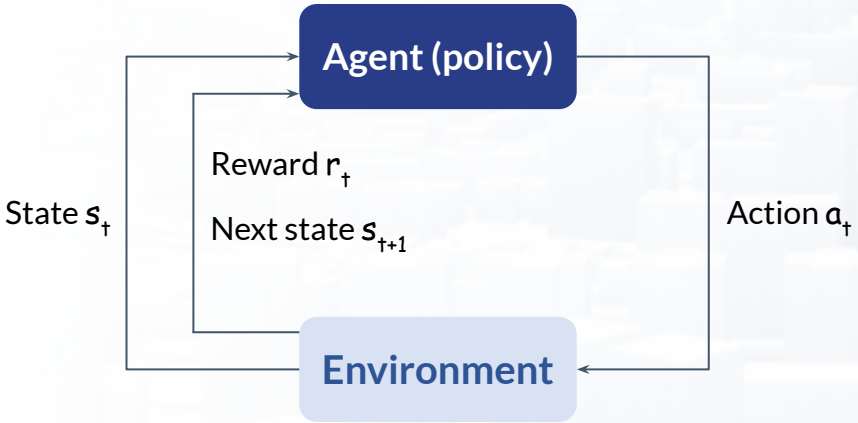
- **Easy** to implement
- **Cheap annotations** (just driving while recording visual sensor and actions)
- Entire model trained **end-to-end**
- Conditioning removes ambiguity at intersections

Challenges of Imitation Learning, esp. Behavior Cloning

- Behavior cloning uses **I.I.D. assumption** which is violated in practice
- No memory (can't remember speed signs, etc.)
- Mapping is difficult to interpret ("**black box**"), despite visualization techniques
- More discussion later.

Credit to Andreas Geiger, Lecture: Self-Driving Cars

Reinforcement Learning

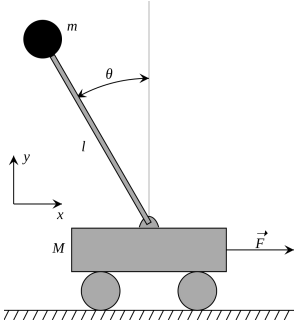


Reinforcement Learning:

- Agent observes environment state s_t at time t
- Agent sends action a_t at time t to the environment
- Environment returns the reward r_t and its new state s_{t+1} to the agent

Credit to Andreas Geiger, Lecture: Self-Driving Cars

Reinforcement Learning

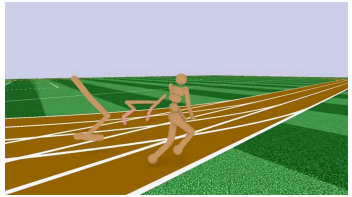


Cart Pole Balancing

- **Objective**: Balance pole on moving cart
- **State**: Angle, angular vel., position, vel.
- **Action**: Horizontal force applied to cart
- **Reward**: 1 if pole is upright at time t

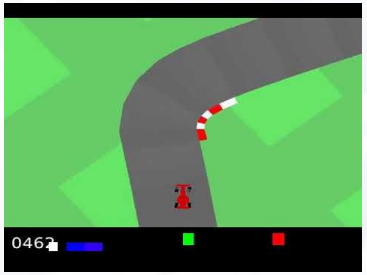
Atari Games

- **Objective**: Maximize game score
- **State**: Raw pixels of screen (210x160)
- **Action**: Left, right, up, down
- **Reward**: Score increase/decrease at t



Robot locomotion (in mujoco)

- **Objective**: Make robot move forward
- **State**: Position and angle of joints
- **Action**: Torques applied on joints
- **Reward**: 1 if upright & forward moving



Self-driving (in gym/CarRacing)

- **Objective**: Lane following
- **State**: Image (96x96)
- **Action**: Acceleration, Steering
- **Reward**: - per frame, + per tile

Credit to Andreas Geiger, Lecture: Self-Driving Cars

Reinforcement Learning

Advantages of Reinforcement Learning

- **Straightforward** idea, early attempts
- Easy be trained with privileged simulator information
- Exploration & Exploitation **relieve causal confusion**

Challenges of Reinforcement Learning

- Large search/state-action space (π), especially in outdoor driving scenarios; **long training** times
- Deep Q-Learning
  - Uniform sampling from replay buffer → all transitions equally important
  - Simplistic exploration strategy
  - Action space is limited to a **discrete set of actions** (otherwise, expensive test-time optimization required)
- Trained in simulation mainly. **Sim2Real gap**.
- More discussion later.

Credit to Andreas Geiger, Lecture: Self-Driving Cars

OpenDriveLab

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

验证与评测 / Benchmarking

Benchmarking

Real-world

DARPA Grand Challenge

Mcity, UMich

On-field tests (c.f. Waymo)

Autonomous Racing

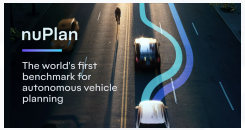
Online/Closed-loop

CARLA

NVIDIA

grand theft auto V

VISTA



Metric:

- Driving Score =
  Route Completion * ∏ Infraction Penalty

Offline/Open-loop

NUSCENES

WAYMO

ARGO AI

Lyft

HUAWEI

Baidu 百度

Metric:

- L2 = | GT - Pred |
- Collision rate

Benchmarking - Closed-loop Simulation

Parameter Init.

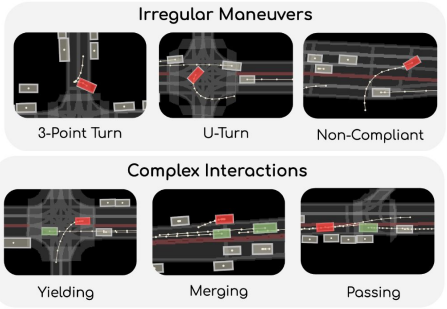
Procedural Generation

- Sample probabilistic distribution of simulation properties with algorithms
- Hand-tuned rules & params.

Data-Driven

- Sample from logs
- Generate by models

Traffic Sim.



TrafficSim, CVPR'21

Rule-Based

- Intelligent Driver Model (IDM) (proves to be effective)

Data-Driven

- Generate by models
- Waymo Sim Agent

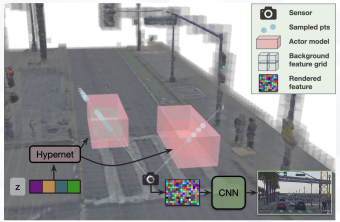
Sensor Sim.

Graphic-Based

- 3D models with assets, plus physical rendering process and sensors modeling
- Physical occlusion, shadows, reflections, etc.

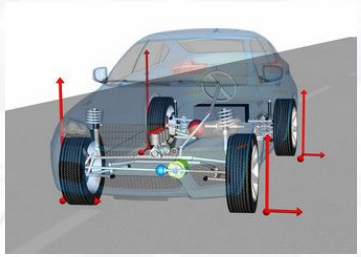
Data-Driven

- NeRF & 3D Gaussian Splatting
- GAN & Diffusion



UniSim, CVPR'23

Vehicle Dyn. Sim.



Simplified Vehicle Model

- Unicycle / Bicycle model
- Multi-body system

Data-Driven

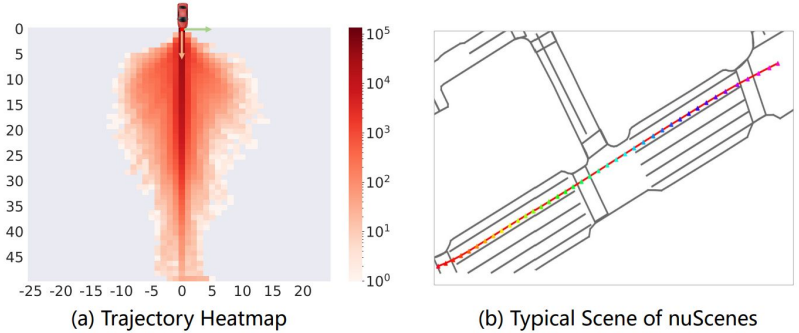
- NN-based learned from data

Benchmarking - Open-loop Evaluation

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

[1] Li et al. Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving? CVPR, 2024.

EgoStatus | Motivation

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.
- **AD-MLP** paper recently points out that a simple MLP network can also achieve state-of-the-art planning results, **relying solely on the ego status information.**

[1] Li et al. Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving? CVPR, 2024.

EgoStatus | Motivation

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.
- **AD-MLP** paper 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?

[1] Li et al. Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving? CVPR, 2024.

EgoStatus | Conclusion

- The **planning performance** of existing **open-loop** autonomous driving models based on **nuScenes** is highly affected by **ego status**
- Existing planning **metrics** fall short of fully capturing the true performance of models.

The development of more appropriate ***datasets*** and ***metrics*** represents a more critical and urgent challenge to tackle

[1] Li et al. Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving? CVPR, 2024.

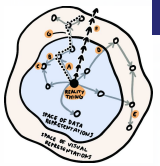
OpenDriveLab

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

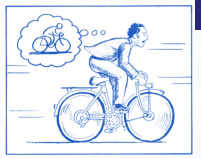
关键研究内容与挑战 / Frontiers and Challenges

Key Challenges

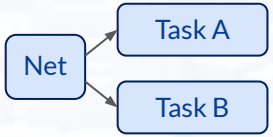
Sensing and Input Modalities



Visual Abstraction

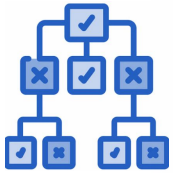


World Modeling for MBRL



Multi-task Learning

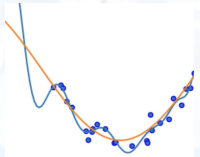
Policy Distillation



Interpretability

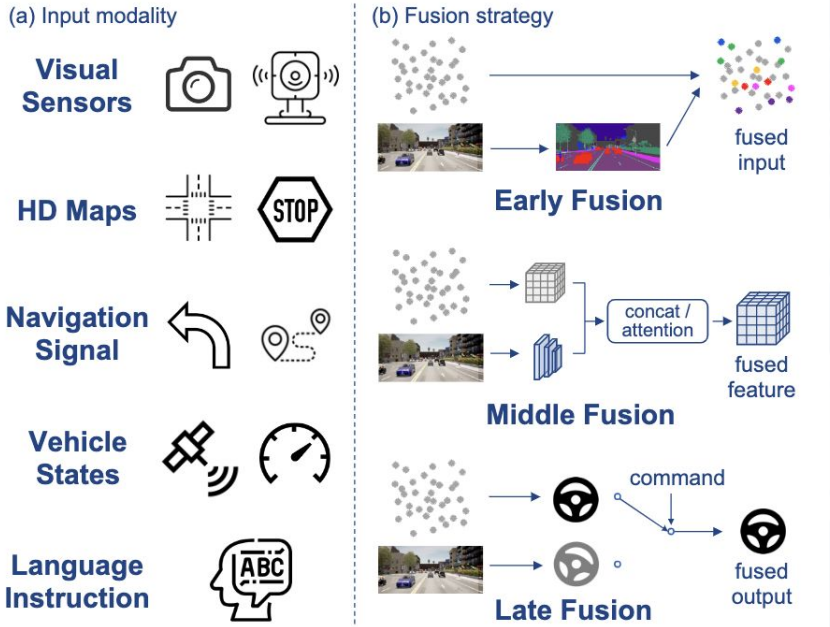
Safety Guarantee

Causal Confusion



Robustness and Generalization

Dilemma over Sensing and Input Modalities



Early Fusion

- Combine sensory information before feeding it into the feature extractor
- Concatenation (imgs, or point-painting)

Middle Fusion

- Separately encode inputs and then combining them at the feature level
- Transformer attention
- Fusion under BEV space

Late Fusion

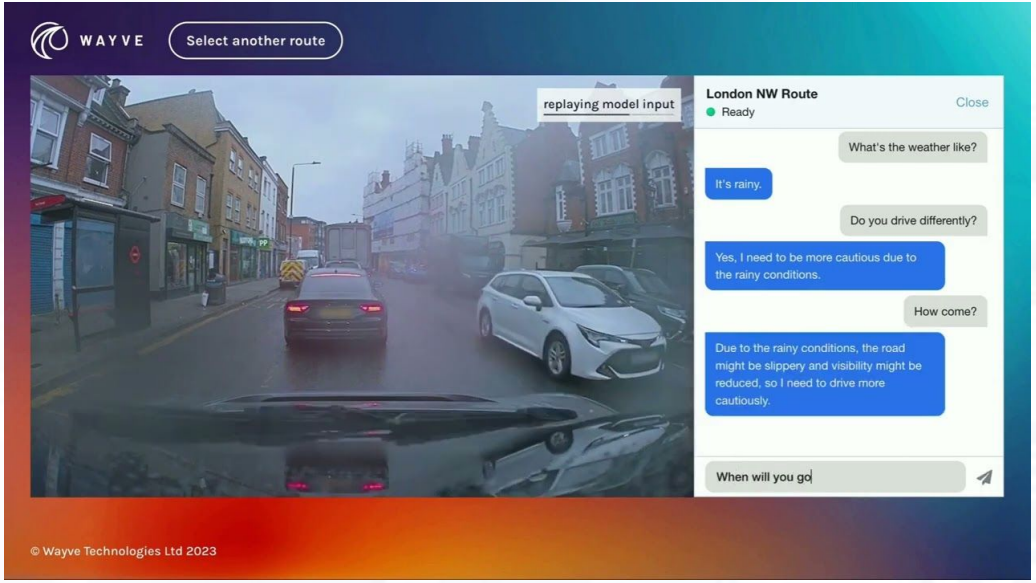
- Combine multiple results from multi-modalities

Challenge: Various sensors possess distinct perspectives, data distributions (similar issues in perception tasks). Some unique inputs such as vehicle states and navigation signals exist for E2E driving.

Dilemma over Sensing and Input Modalities - Language

Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

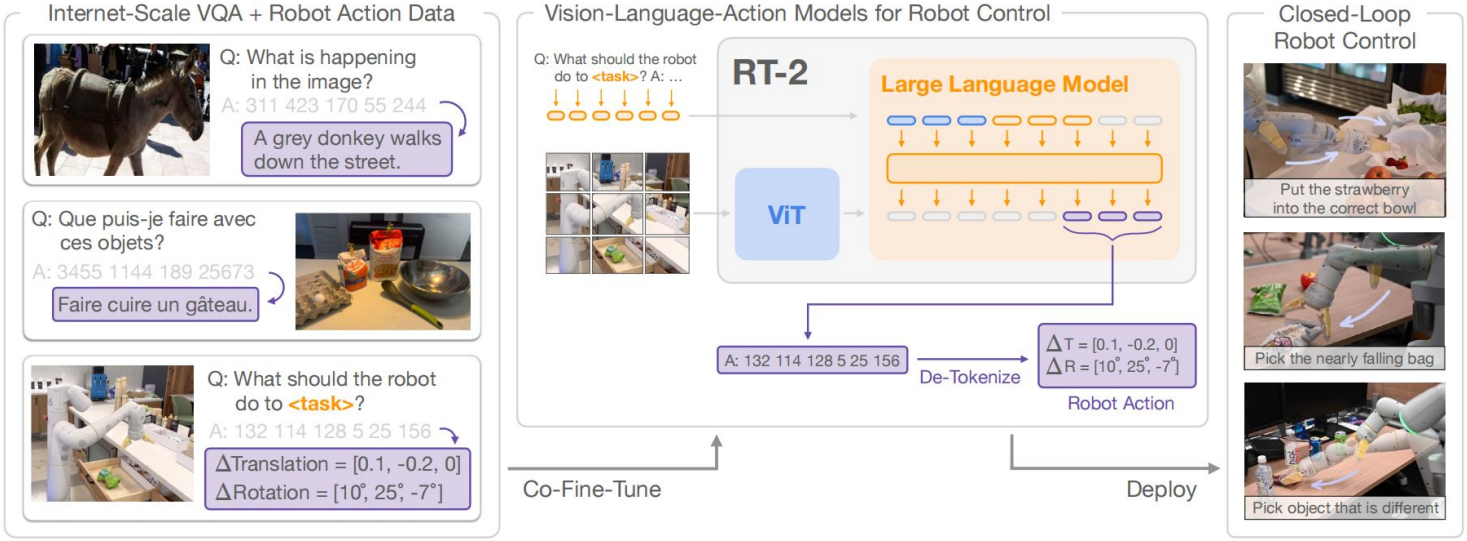
Vision Language Navigation, CVPR'18



LINGO-2, arXiv'24

Insight from Robotics / Embodied AI

RT-2, arXiv'23



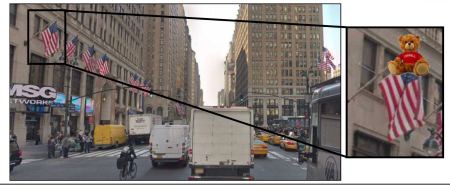
- How vision-language models trained on Internet-scale data can be incorporated directly into **end-to-end robotic control**
- Goal: to **boost generalization** and enable emergent semantic reasoning

- 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

Dilemma over Sensing and Input Modalities - Language

Outdoor Navigation

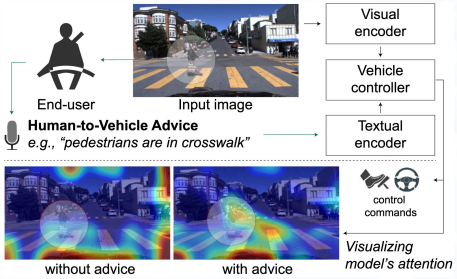


Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light, As you come to the third traffic light you will see a white building on your left with many American flags on it. Touchdown is sitting in the stars of the first flag.

TOUCHDOWN, CVPR'19

- Discrete action space (Forward/Left/Right/Stop)
- Crowdsourcing routes or collected from Google Maps

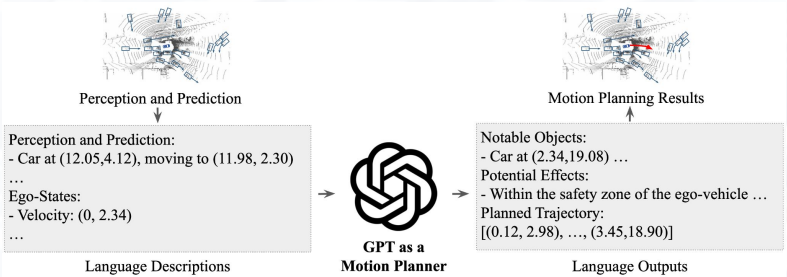
Linguistic Instruction



HAD, CVPR'19

- **Grounding** objects/areas with human advice, and predict control signals with attention
- Human annotated videos

LLM-assisted E2E Driving



GPT-Driver, NeurIPS Workshop'23

- Encode perception/prediction/etc results into texts (or tokens) and **prompt with LLMs** for planning and behavior explanation
- OR: construct **visual question answering** dataset to train VLMs

Existed Opportunity: Language can provide certain **high-level instructions**, enhancing human-machine-interaction.

Current Challenge: LLMs (or VLMs) require **long inference** time, lack of **quantitative accuracy** and stability of outputs.

Dependence on Visual Abstraction

State space / Sensor input

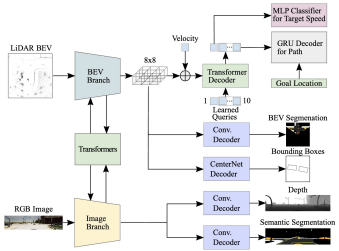
Key function:

- Compact intermediate representations
- Pre-trained visual encoders accelerate downstream training, especially for RL methods (sample efficiency)

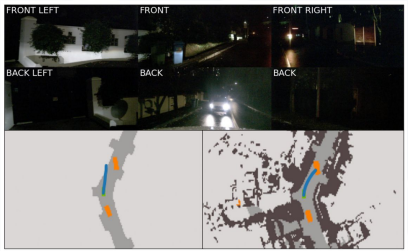
Dependence on Visual Abstraction - Representation Design

Representation forms of visual inputs

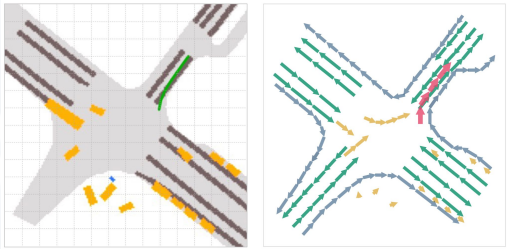
- CNN vs Transformer:
  - CNN still dominates in E2E driving than ViT. Why?
- Bird's-eye-view (BEV) & Occupancy grid:
  - BEV is good for sensor/temporal fusion and facilitates downstream tasks
  - 3D occupancy can capture irregular objects and used for collision avoidance
- Maps:
  - BEV segmentation, vectorized lanelines, centerlines and topology, lane segments, etc.



Transfuser++, ICCV'23



Scene as Occupancy, ICCV'23



VAD, ICCV'23

However, if explicit representations are necessary for performance improvement is unclear. This is a debate about modular E2E vs. scaling up "black box" models.

Dependence on Visual Abstraction - Representation Learning

State space / Sensor input

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.

Dependence on Visual Abstraction - Representation Learning

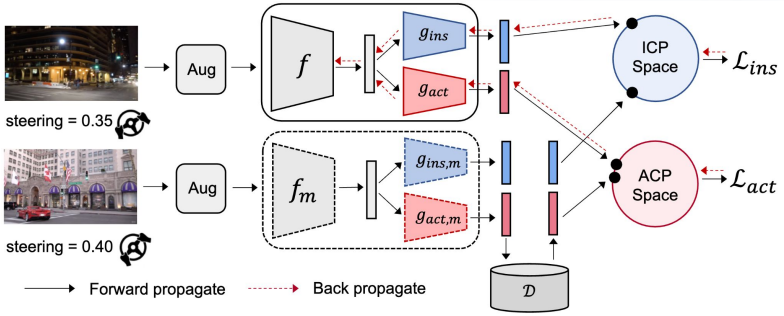
Traditional RL methods' pretraining: abstracted information, also helpful for Sim2Real

- Semantic segmentation
- Depth map
- VAE (Variational autoencoder)

Pre-training from large-scale vision data

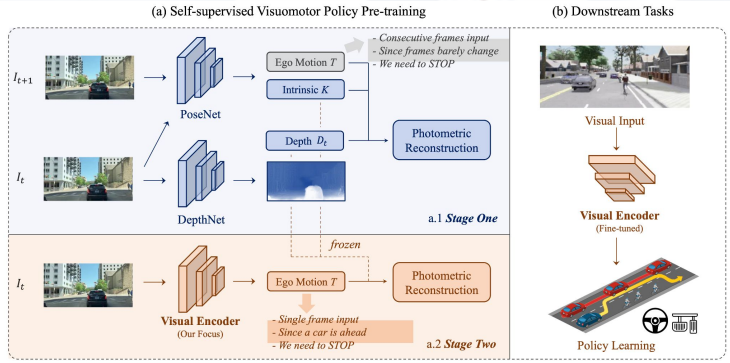
YouTube ⇨ **Large-scale Policy Pre-training**

Contrastive Learning



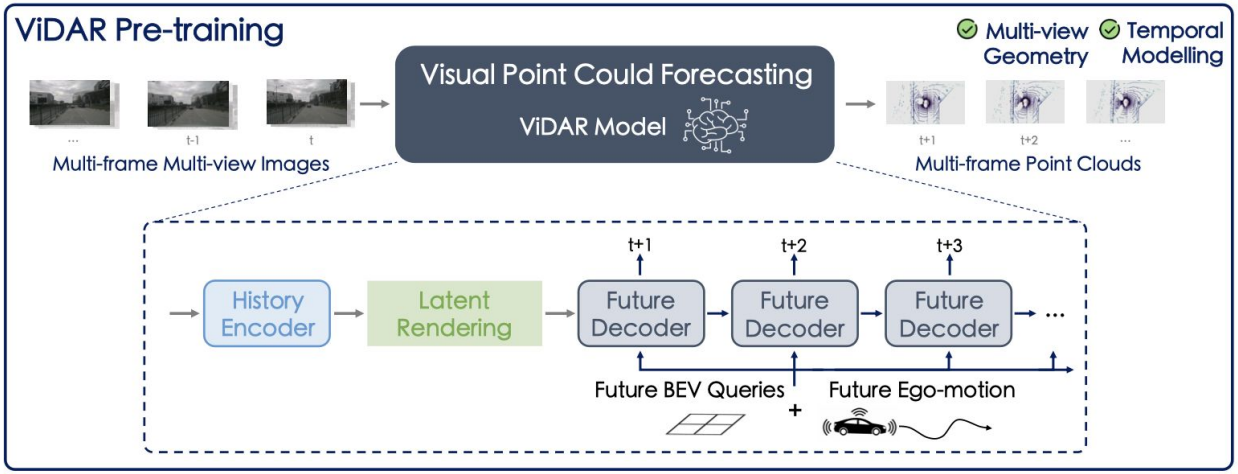
ACO, ECCV'22

Self-supervised Geometric Modeling



PPGeo, ICLR'23

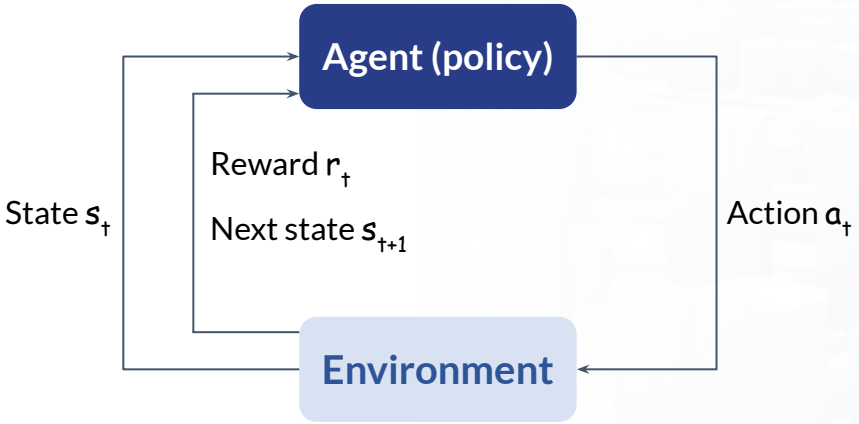
Representation Learning - ViDAR

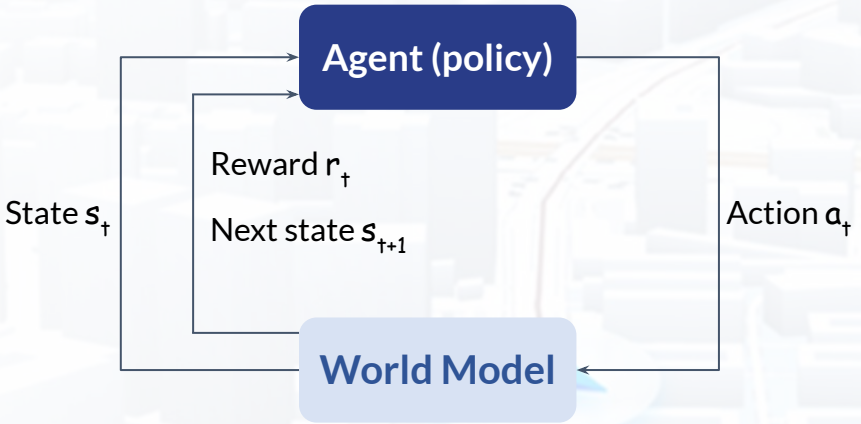


Method	Detection		Tracking			Mapping		Motion Forecasting			Future Occupancy Prediction				Planning	
	NDS ↑	mAP ↑	AMOTA↑	AMOTP↓	IDS↓	IoU-lane↑	IoU-road↑	minADE↓	minFDE↓	MR↓	IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑	avg.L2↓	avg.Col.↓
UniAD	49.36	37.96	38.3	1.32	1054	31.3	69.1	0.75	1.08	0.158	62.8	40.1	54.6	33.9	1.12	0.27
ViDAR	**52.57**	**42.33**	**42.0**	**1.25**	**991**	**33.2**	**71.4**	**0.67**	**0.99**	**0.149**	**65.4**	**42.1**	**57.3**	**36.4**	**0.91**	**0.23**

By fine-tuning on UniAD, ViDAR boosts perception, prediction, and planning at the same time by a large margin, providing a new solution for using large-scale unlabeled data.

Complexity of World Modeling





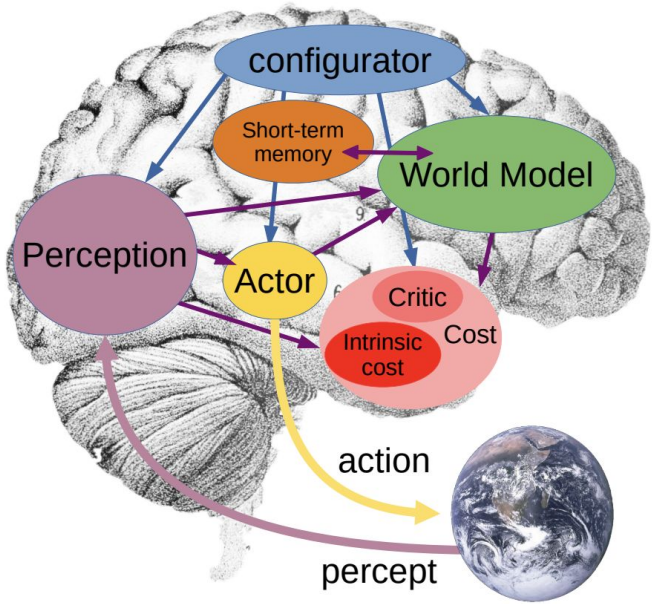
Complexity of World Modeling

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.

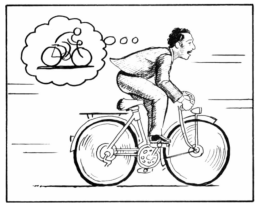
What happens if I go straight?



Trending: Recent Work on World Model

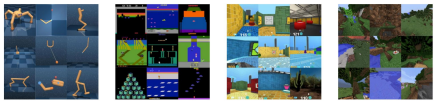
From simulated agents to real-world driving systems

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



World Models:
Training agents inside their dreams

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

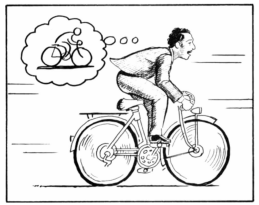


(a) Control Suite (b) Atari (c) DMLab (d) Minecraft

Trending: Recent Work on World Model

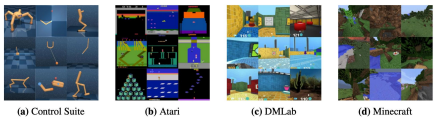
From simulated agents to real-world driving systems

RL Agents — **Vision** — **Driving**

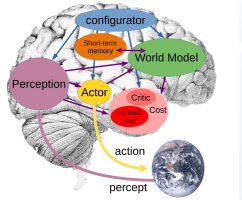


18.3 **World Models:** Training agents inside their dreams

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



22.6 **Position Paper** (by LeCun) Elaborating the developments of world models



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

23.6 **Scaling up world models on large corpus of realistic driving videos**

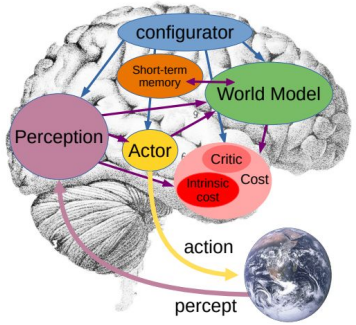
General World Model: inhouse data collected around the globe

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

Complexity of World Model



	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?

Complexity of World Model

Forms of world model

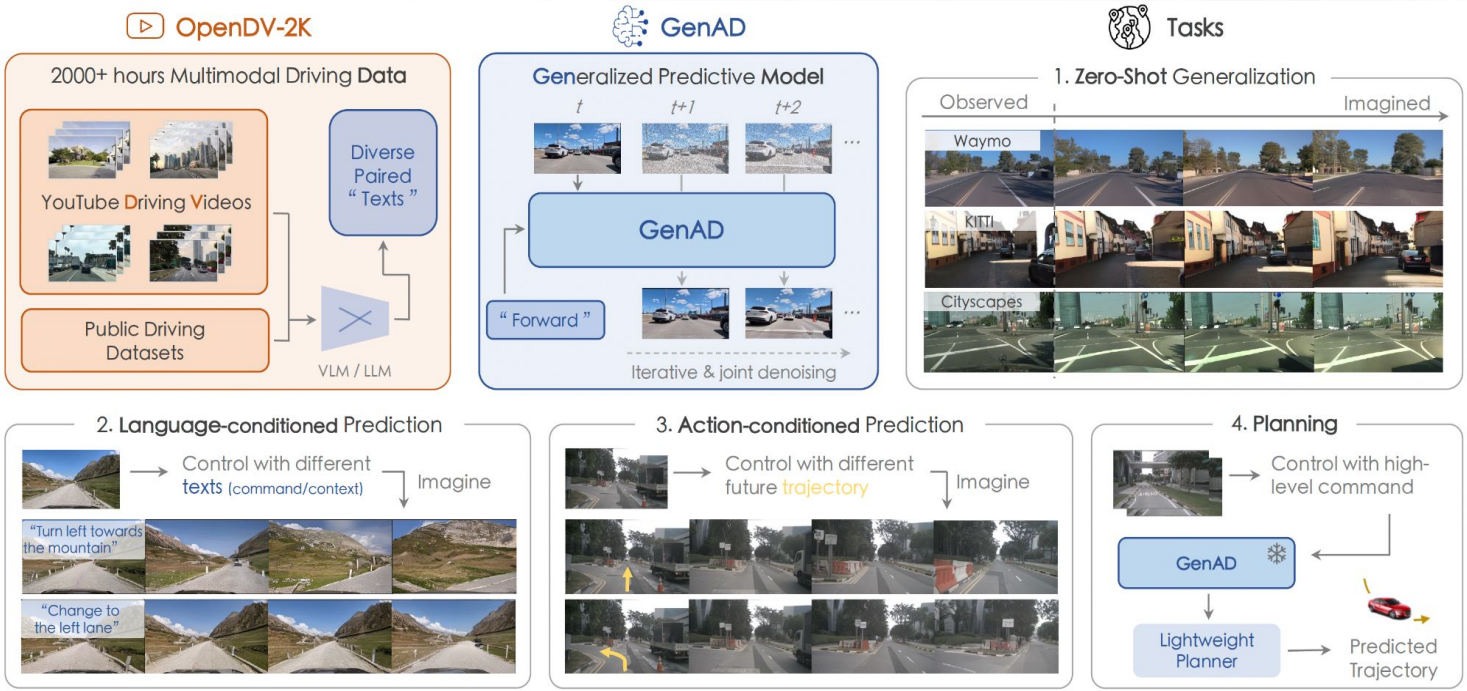
- Images
- Point cloud (LiDAR) and Occupancy
- BEV maps
- Integrated Motion Prediction and Planning
- Latent model (LSTM/GRU-based, etc)

Application of world model

- Reward for RL/Sampling/etc
- Decode planning with Inverse Dynamics Model
- Serve as pre-training (fine-tuned for planning)

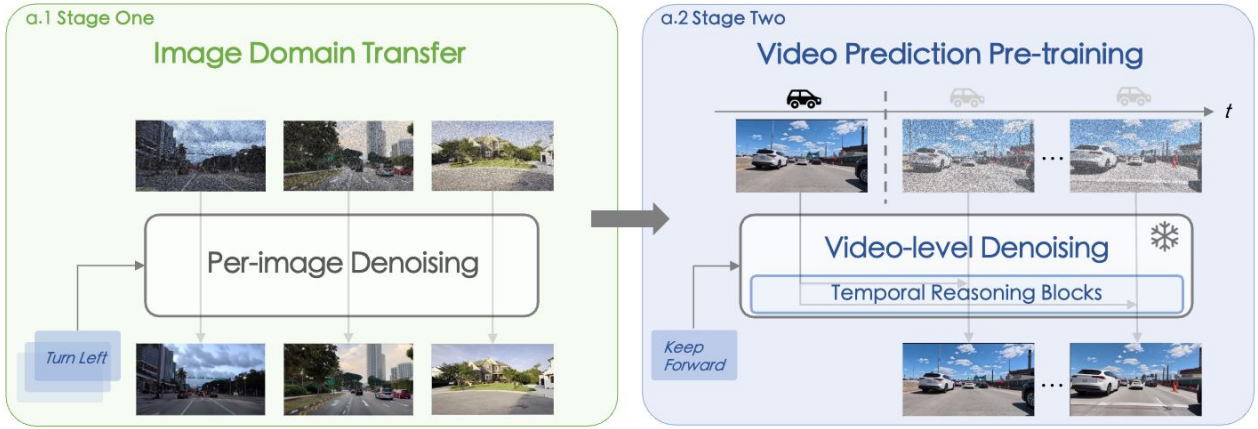
Complexity of World Model | GenAD

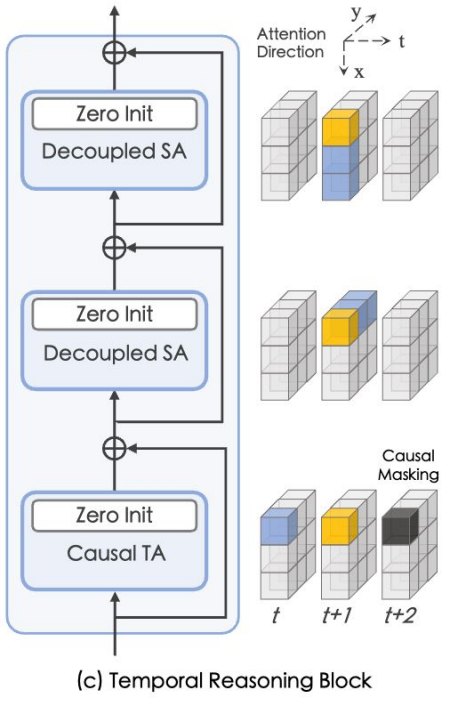
Summary: Training a **billion-scale video prediction model** on **web-scale driving videos**, to enable its **generalization across** a wide spectrum of **domains and tasks**.



Complexity of World Model | GenAD

- **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**

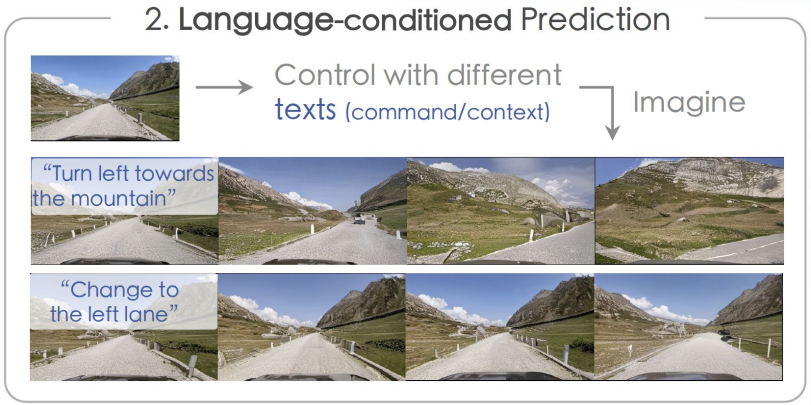




Tasks | Zero-shot Generalization (Video Prediction)

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

Tasks | Language-conditioned Prediction



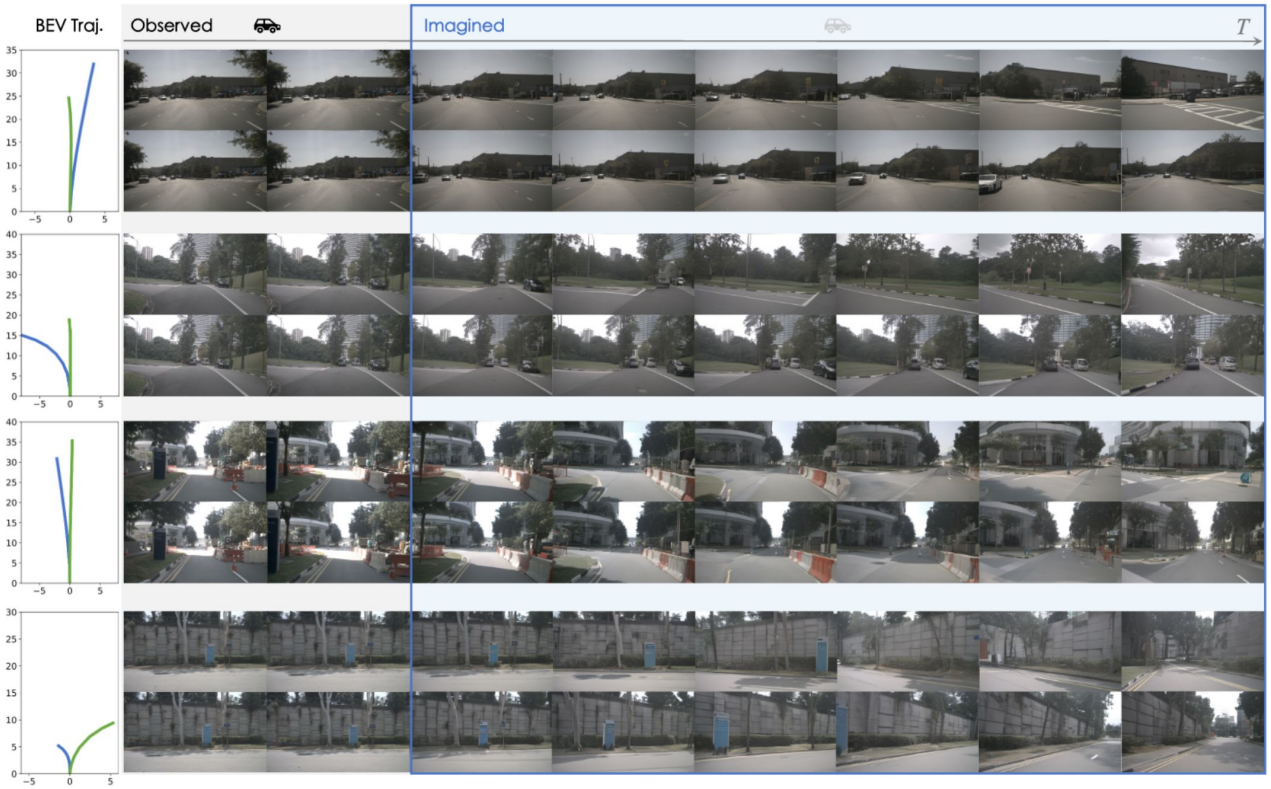
Instruct the future with **free-form texts.**

Tasks | Action-conditioned Prediction (Simulation)

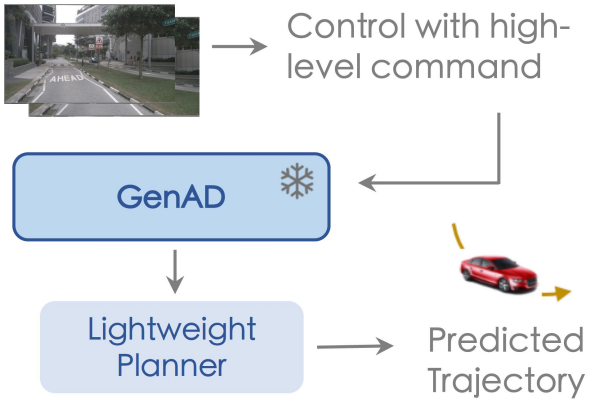
Method	Condition	nuScenes Action Prediction Error (↓)
Ground truth	-	0.9
GenAD	text	2.54
GenAD-act	text + traj.	**2.02**

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

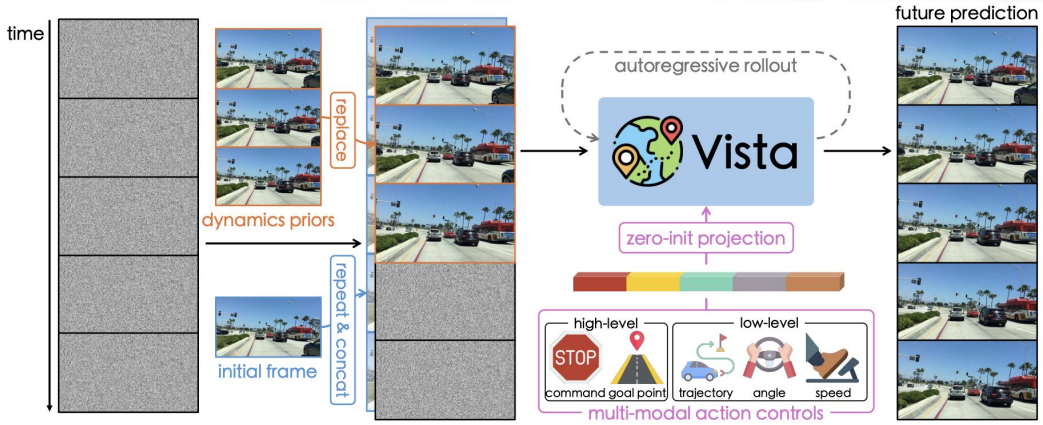


Method	# Trainable Params.	nuScenes ADE (↓)	nuScenes FDE (↓)
ST-P3* [20]	10.9M	2.11	2.90
UniAD* [22]	58.8M	1.03	1.65
GenAD (Ours)	0.8M	1.23	2.31

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).

Vista: Generalized action conditions



Control with traj & angles

(translated to commands for vis)

Forward | Left | Right | Stop

[1] Gao et al. Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability. arXiv, 2024.

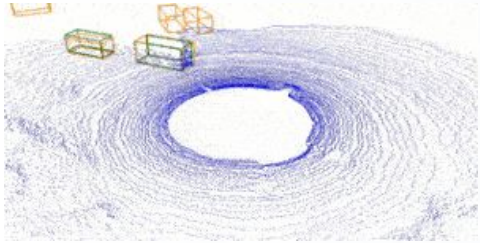
ViDAR - World Model in Driving

+ Action

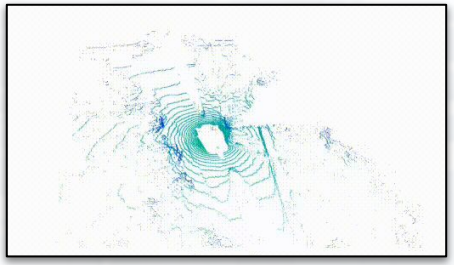
Future Prediction Model → World Model

Point Cloud

Carnegie Mellon University



S2Net — Point cloud future prediction for planning



4D-Occ — Ego Future Tracjectory

OpenDriveLab

Point Cloud & Visual Image

ViDAR

Visual Image

2022

2023

WAYVE

FIERY — visual future prediction for planning.

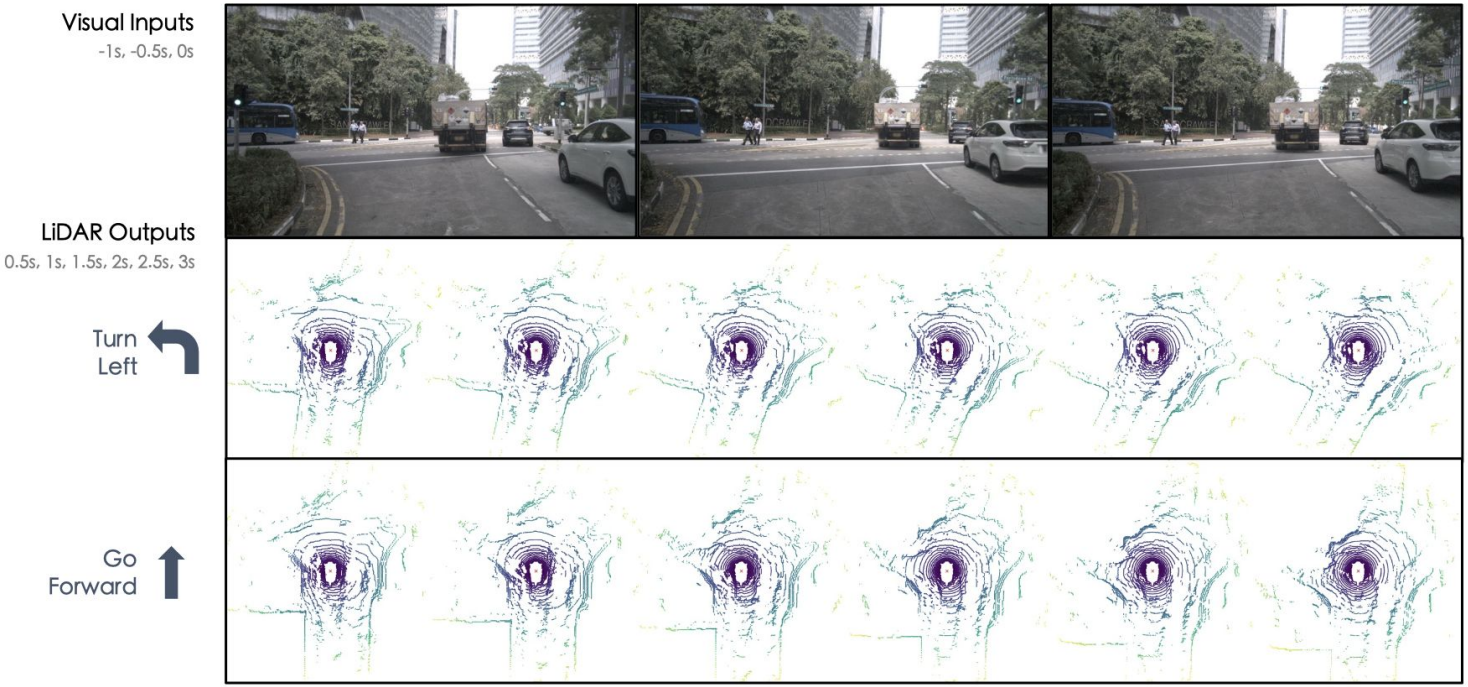
GAIA-1 — Text & Steering

DriveDreamer — Box & Image & HDMap

DrivingDiffusion — Layout

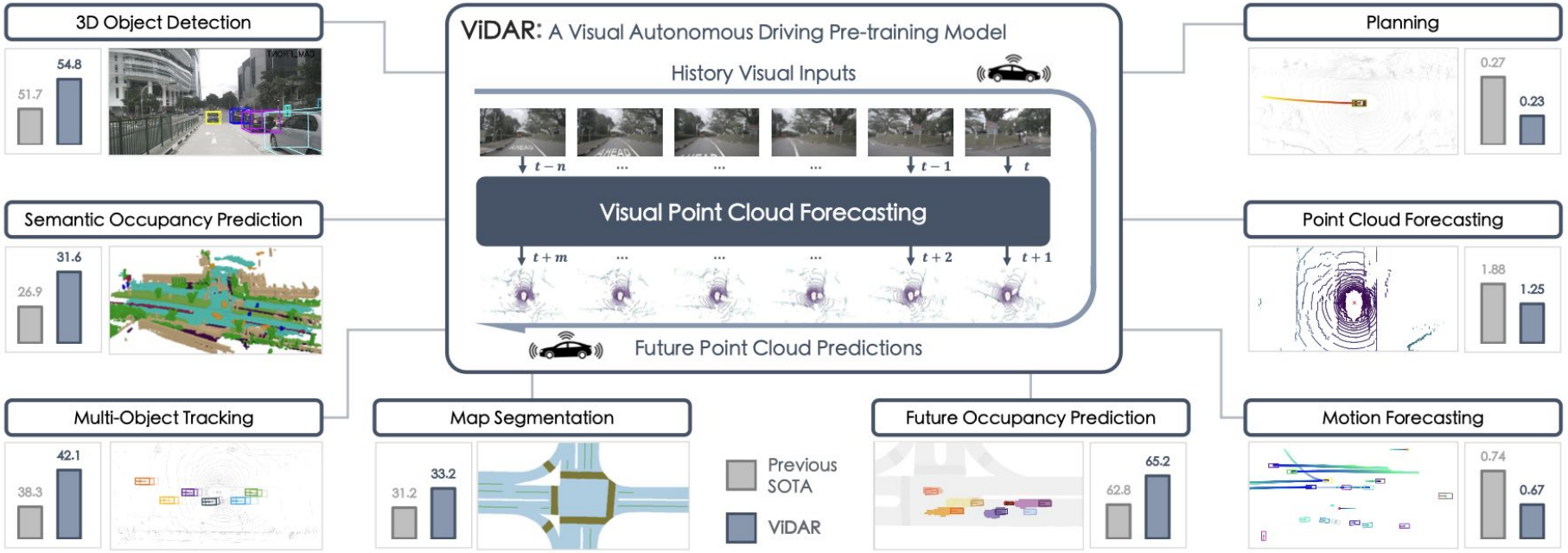
ViDAR - World Model in Driving

The First Multimodal World Model

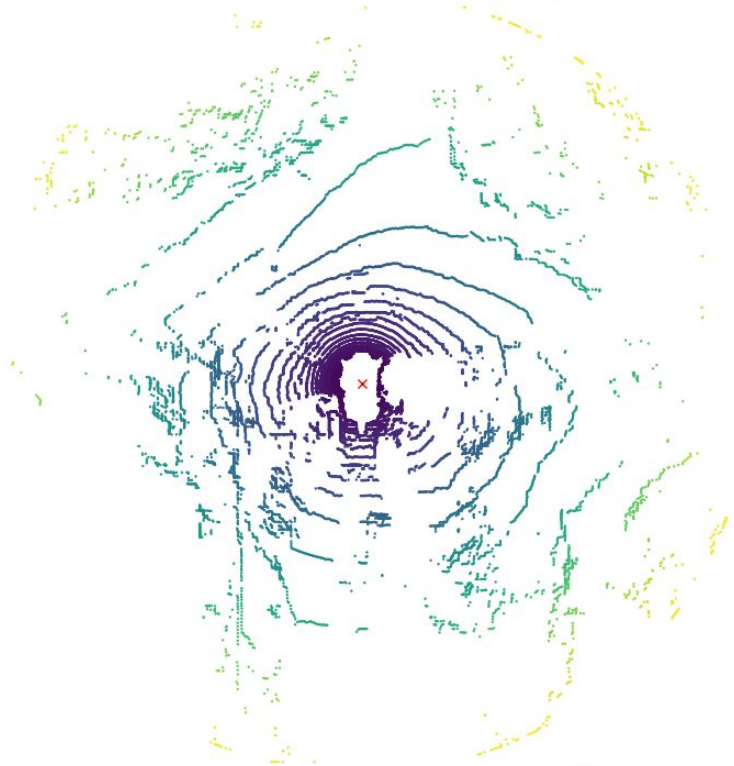


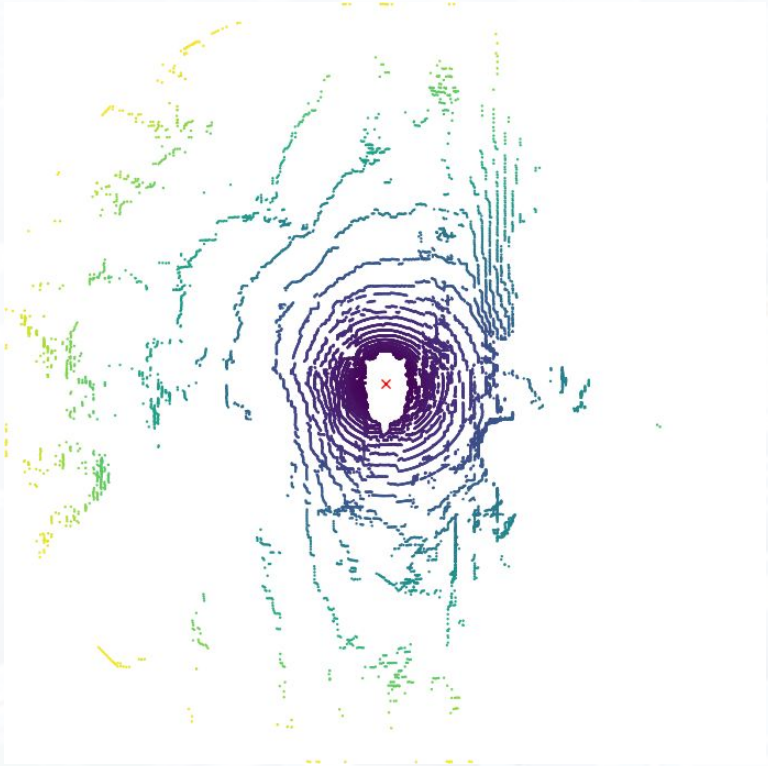
ViDAR - World Model in Driving

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



ViDAR | Future Prediction Experiments





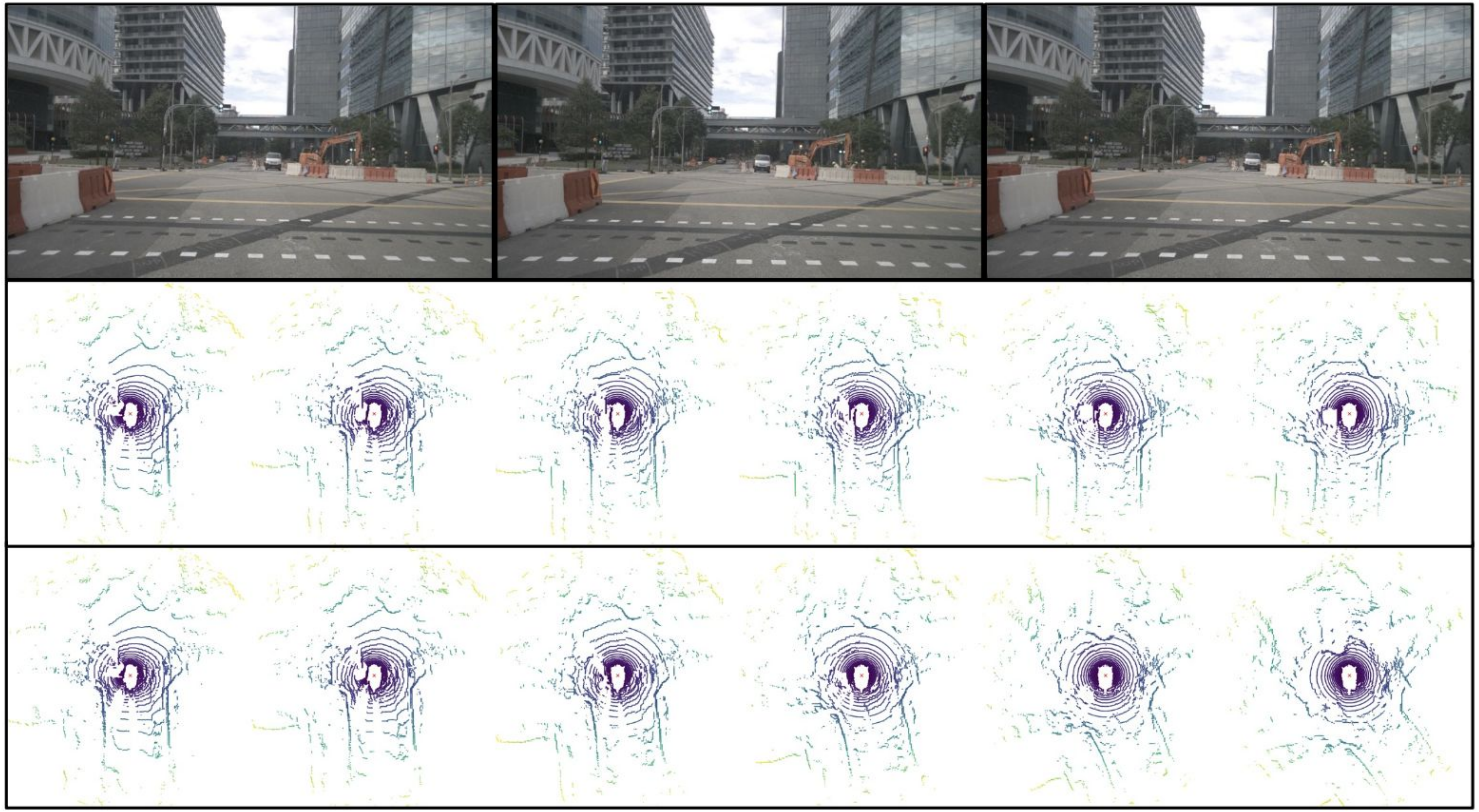
ViDAR | Different Ego Control Experiments

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

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

Go Forward

Turn Right



ViDAR | Downstream Experiments

Method	Detection		Tracking			Mapping		Motion Forecasting			Future Occupancy Prediction				Planning	
	NDS ↑	mAP ↑	AMOTA↑	AMOTP↓	IDS↓	IoU-lane↑	IoU-road↑	minADE↓	minFDE↓	MR↓	IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑	avg.L2↓	avg.Col.↓
UniAD	49.36	37.96	38.3	1.32	1054	31.3	69.1	0.75	1.08	0.158	62.8	40.1	54.6	33.9	1.12	0.27
ViDAR	**52.57**	**42.33**	**42.0**	**1.25**	**991**	**33.2**	**71.4**	**0.67**	**0.99**	**0.149**	**65.4**	**42.1**	**57.3**	**36.4**	**0.91**	**0.23**

By fine-tuning on UniAD,

ViDAR boosts perception, prediction, and planning at the same time by a large margin, providing a new solution for using large-scale unlabeled data.

Reliance on 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

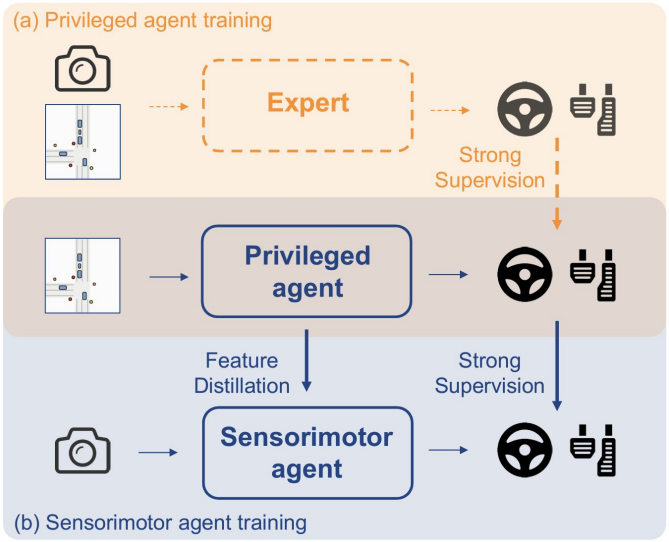
- The **optimal combination** of auxiliary tasks and the appropriate weighting of their losses
- Construct **large-scale datasets** with multiple types of aligned and high-quality annotations

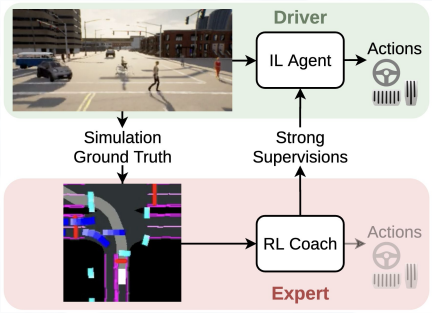
Typical multi-tasks in E2E driving:

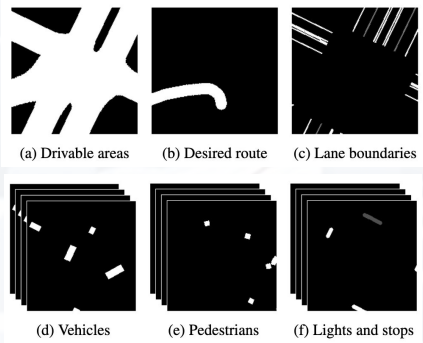
- Early works:
  - **Semantic segmentation**: high-level understanding
  - **Depth estimation**: 3D geometry
- Recent works:
  - **3D object detection**
  - **BEV segmentation**
  - **Visual affordance**, including traffic light states, distance to opposite lanes, etc

Inefficient Experts and Policy Distillation

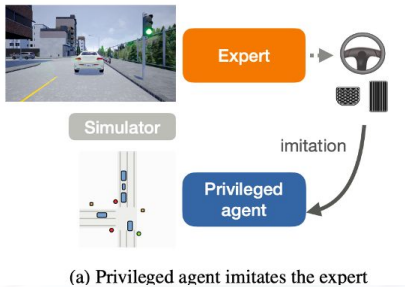
The popular "Teacher-Student" IL Paradigm



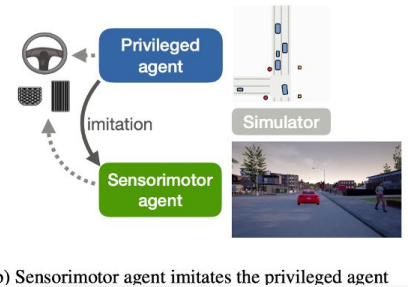




Roach, ICCV'21



(a) Privileged agent imitates the expert



(b) Sensorimotor agent imitates the privileged agent

Learning by Cheating, CoRL'21

- Ease training and better generalization (with privileged information)
- Can be queried with any states, instead of logged states only

Inefficient Experts and Policy Distillation

- Expert: Ground Truth (GT) to action

Gap

- Student: Img to action

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

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**

Student (IL etc, sensor input):

- Not/Can't perfect, even with large-scale data and have visually great representations

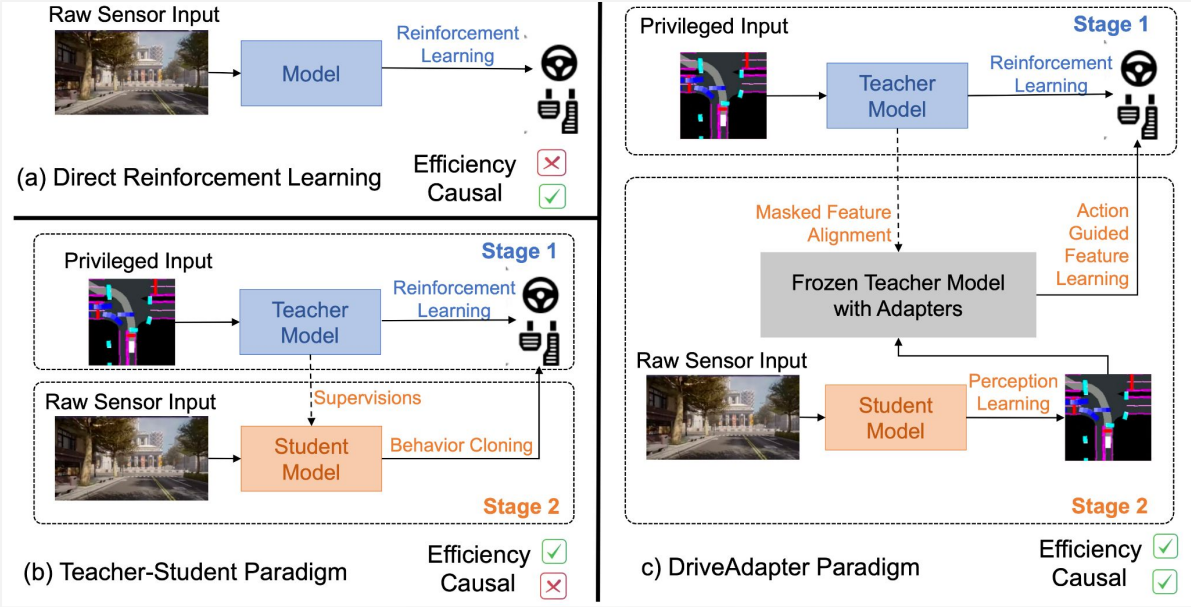
Privileged Input

Perception Result

BEVFusion + Mask2Former
2M training data

DriveAdapter

How to balance the efficiency and causal reasoning ability?



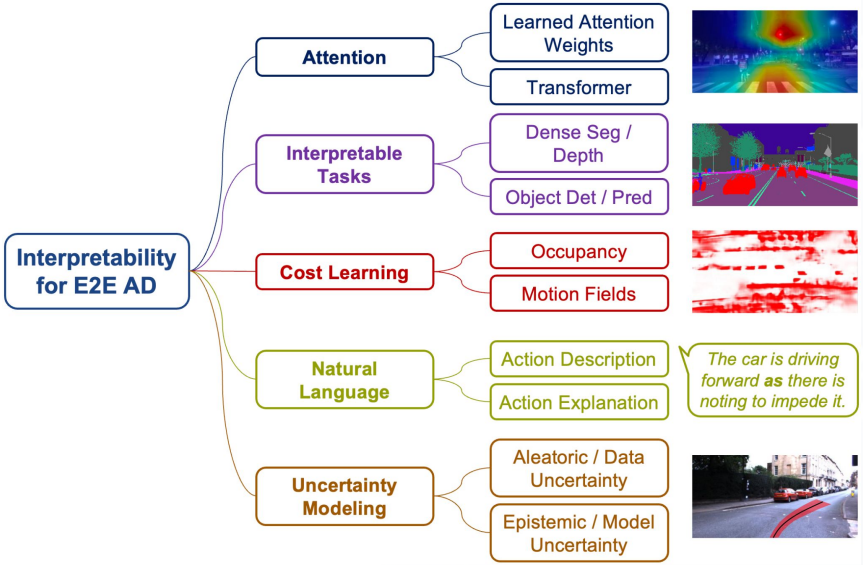
Utilize the strong RL-based privileged teacher model!

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

[1] Jia et al. DriveAdapter: New Paradigm for End-to-End Autonomous Driving to Alleviate Causal Confusion. ICCV, 2023.

Lack of Interpretability

Summary of the different forms of interpretability



They aid in human comprehension on the:

- Decision-making processes of end-to-end models
- Perception failures
- Reliability of the outputs

Lack of Safety Guarantees

Safety of the intended functionality (SOTIF)

- Applicable design, verification, and validation measures

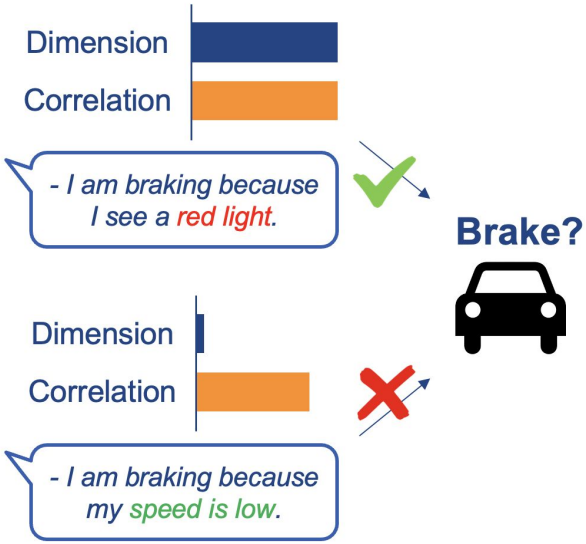
Modular driving stacks

- Safety-related constraints or optimizations, within motion planning or speed prediction modules
- Integrated into E2E models as post-process steps or safety checks
- Detection and motion prediction results can be used in post-processing procedures

Causal Confusion

Input image $\in \mathbb{R}^{W \times H \times 3}$

Velocity $\in \mathbb{R}^2$



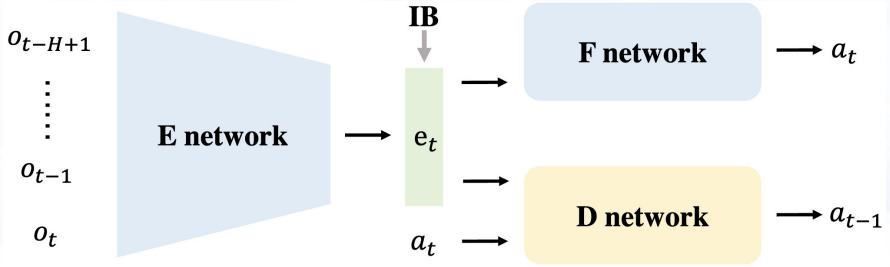
- Driving is a task that exhibits **temporal smoothness**, which makes past motion a reliable predictor of the next action.
- However, methods trained with **multiple frames** can become overly reliant on this shortcut. This is referred to as the **copycat problem** and is a manifestation of **causal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

Causal Confusion

Current Solutions

- Adversarial model predicts ego's past action → min-max optimization trains the model to eliminate its past from intermediate layers
- Random dropout
- Upweighting keyframes in the training loss, where a decision change occurs
- Predict action residuals instead of actions
- Use stacked LiDAR points

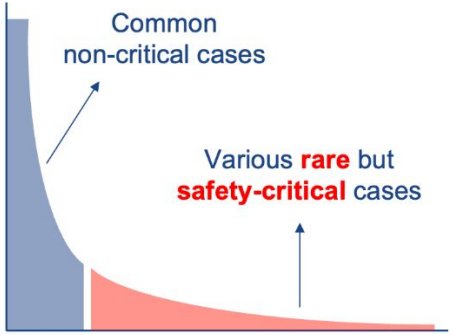
Still challenging



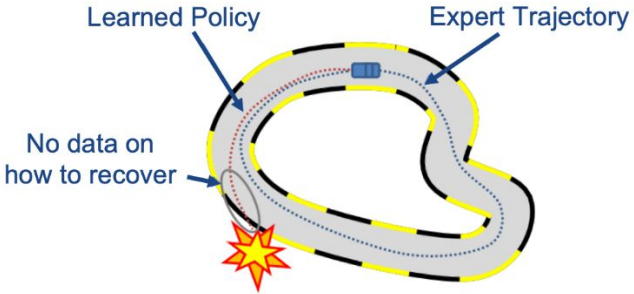
Fighting copycat agents in behavior cloning from observation histories, NeurIPS' 20

- **Problem:** F learns to rely heavily on the information about a_{t-1} in e_t to predict a_t
- **Target**: Remove information about a_{t-1} in e_t
- **Method**:
  - Adversarial (to remove information) network D predicts a_{t-1} from e_t
  - E maximizes the conditional entropy $H(a_{t-1}|e_t)$
- **Issue:** removing all maybe counterproductive; the copycat problem arises only when a_t and a_{t-1} are highly correlated

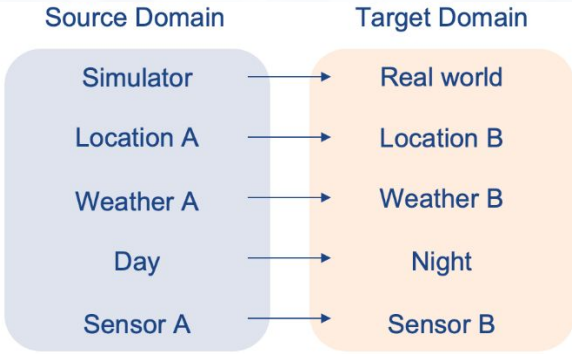
Lack of Robustness



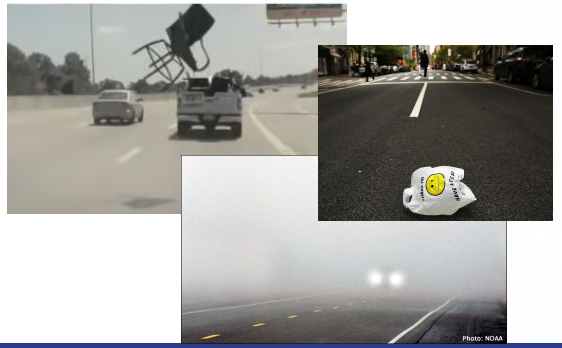
(a) Long-tailed Distribution



(b) Covariate Shift



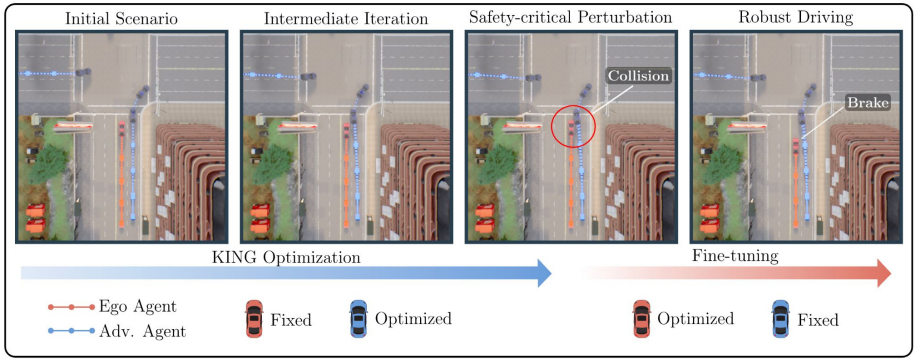
(c) Domain Adaptation



Lack of Robustness - Long-tail Distribution

Current Solutions

- **Hand-crafted scenarios** for more diverse data in simulation, especially for OOD perception
- Non-ego agents' prediction to promote data diversity (philosophy of **integrated prediction and planning**)
- **Importance-sampling** to accelerate evaluation of rare-event probabilities
- Adversarial attacks
  - Bayesian Optimization
  - Policy gradient for generation
  - Hand-crafted modification on agents' trajectories



KING, ECCV'22

OpenDriveLab

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

未来可能与 讨论 / Future Trends and Discussions

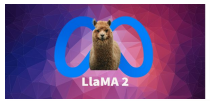
Foundation Models

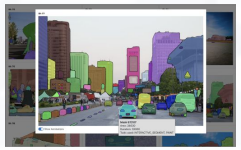
NLP (LLM)

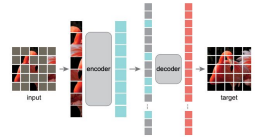
Multimodal (VLM, etc)

General CV

ChatGPT







AD System

- Language Interpreter
- Driving Knowledge
- any more?

- Vision Abstractor
- Auto-labeling
- any more?

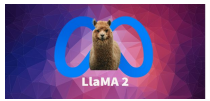
Foundation Models (cont'd)

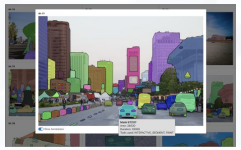
NLP (LLM)

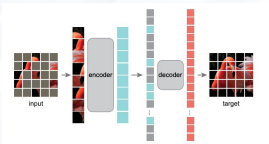
Multimodal (VLM, etc)

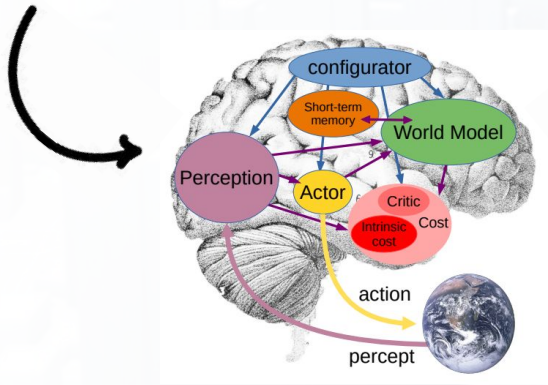
General CV

ChatGPT







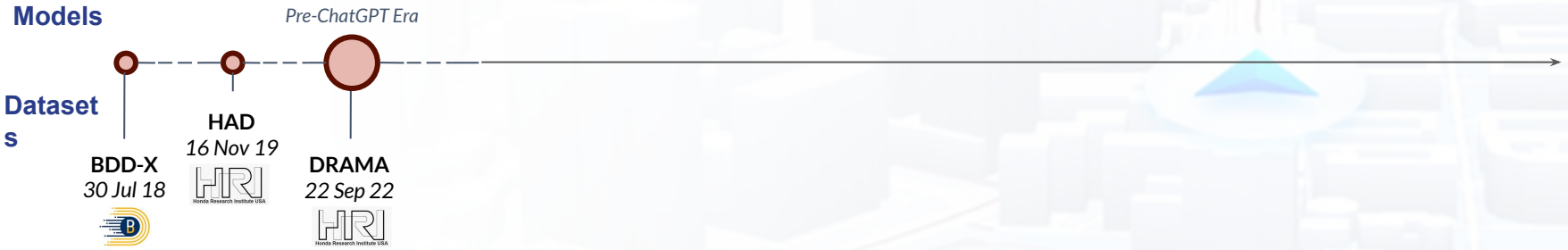


AD System

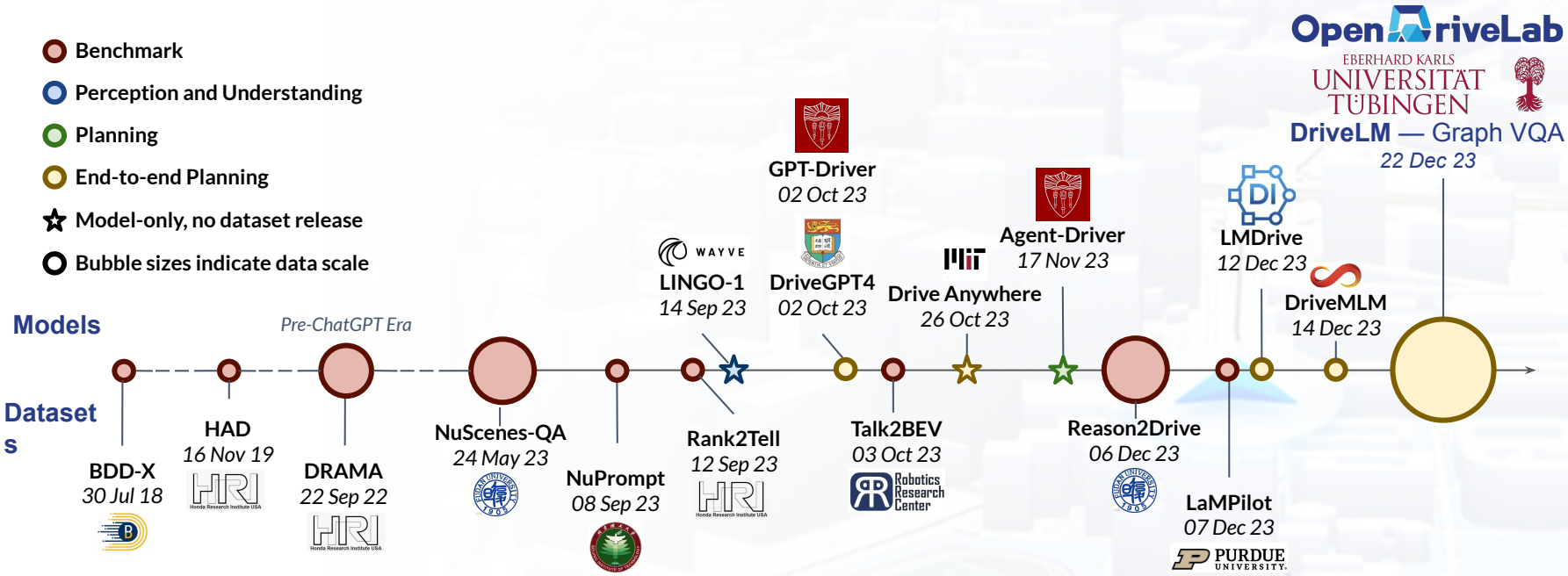
- Multi-modal
- Intelligence
- Generalization

Trending in E2EAD | Driving + Language

- Benchmark
- Perception and Understanding
- Planning
- End-to-end Planning
- Model-only, no dataset release
- Bubble sizes indicate data scale



Trending in E2EAD | Driving + Language



- https://github.com/OpenDriveLab/DriveLM

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**

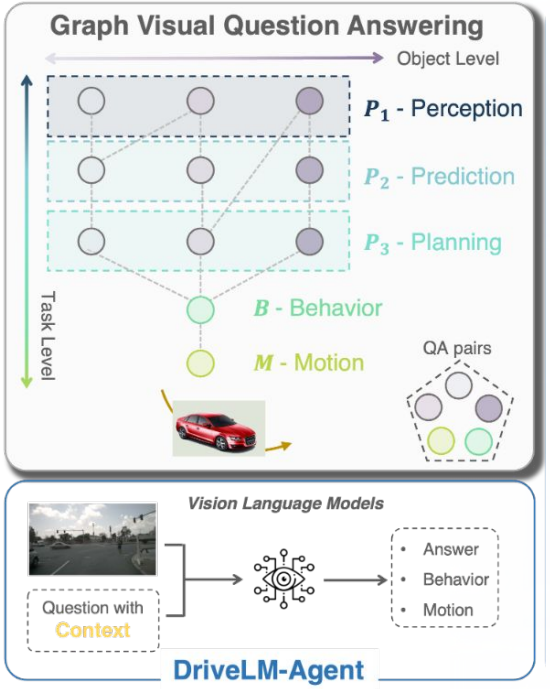
- Good **reasoning** ability, enabled by LLM
- No BEV representation, since human do not rely on BEV

Why VLM in AD?

- **Reasoning** ability helps **generalization**
- **Language** output provides **interactivity**

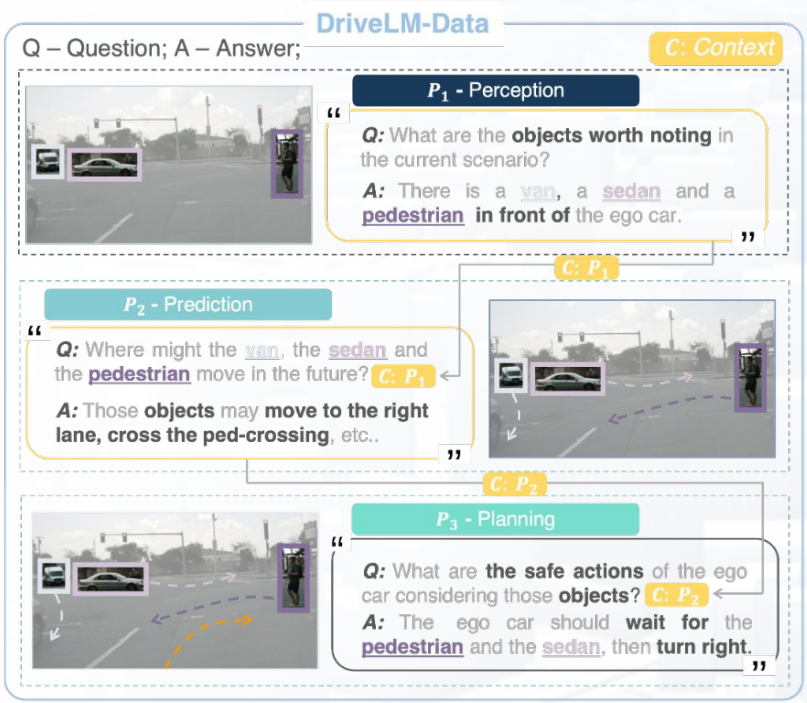
[1] Sima et al. DriveLM: Driving with Graph Visual Question Answering. arXiv, 2023.

DriveLM | At a Glance



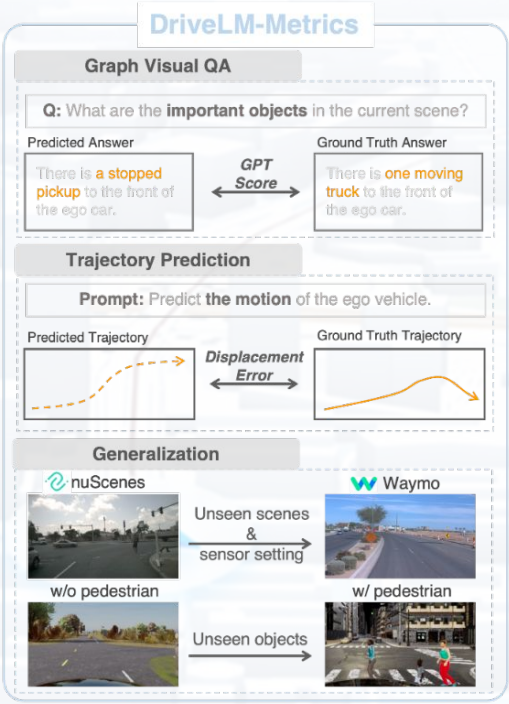
The critical part is **Graph Visual QA**, upon which we build **data**, model and metrics accordingly

DriveLM | At a Glance



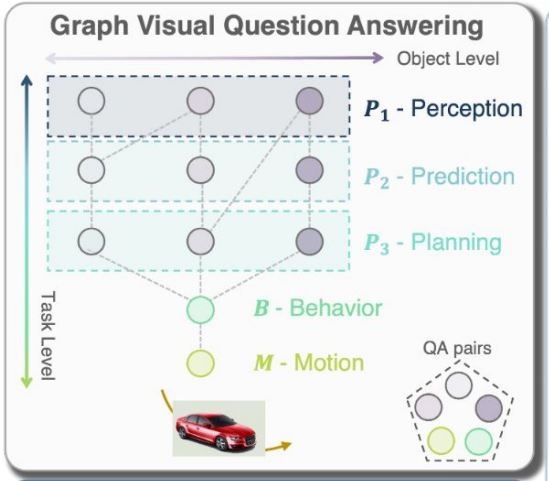
The critical part is **Graph Visual QA**, upon which we build **data**, model and metrics accordingly

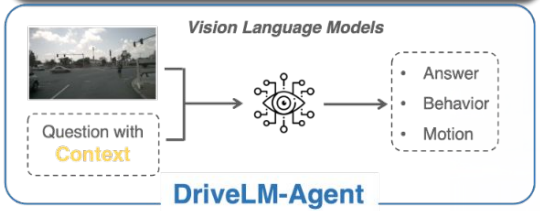
DriveLM | At a Glance

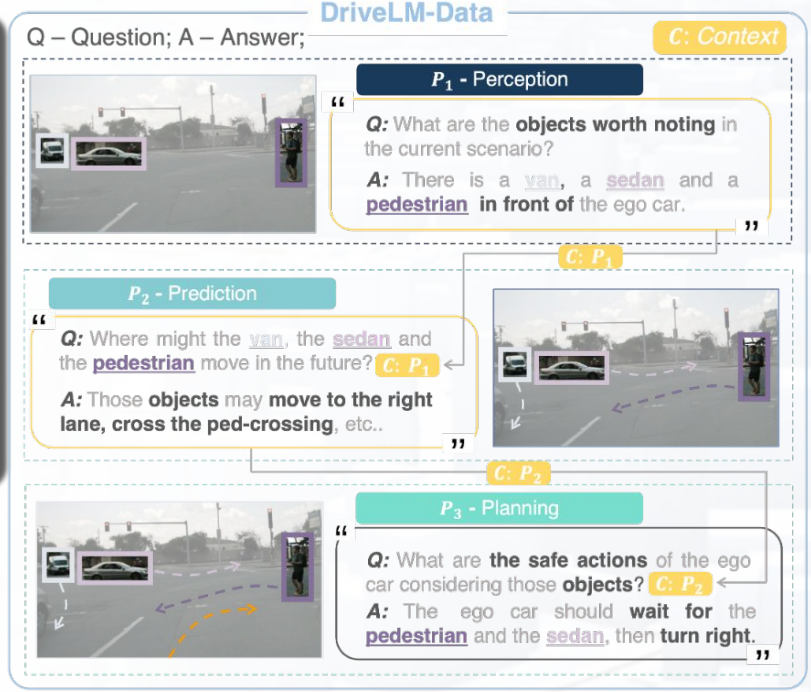


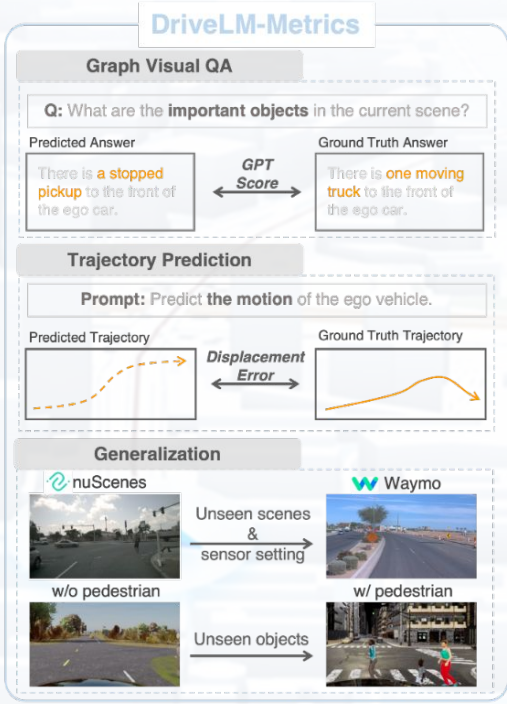
The critical part is **Graph Visual QA**, upon which we build **data**, model and metrics accordingly

DriveLM | At a Glance



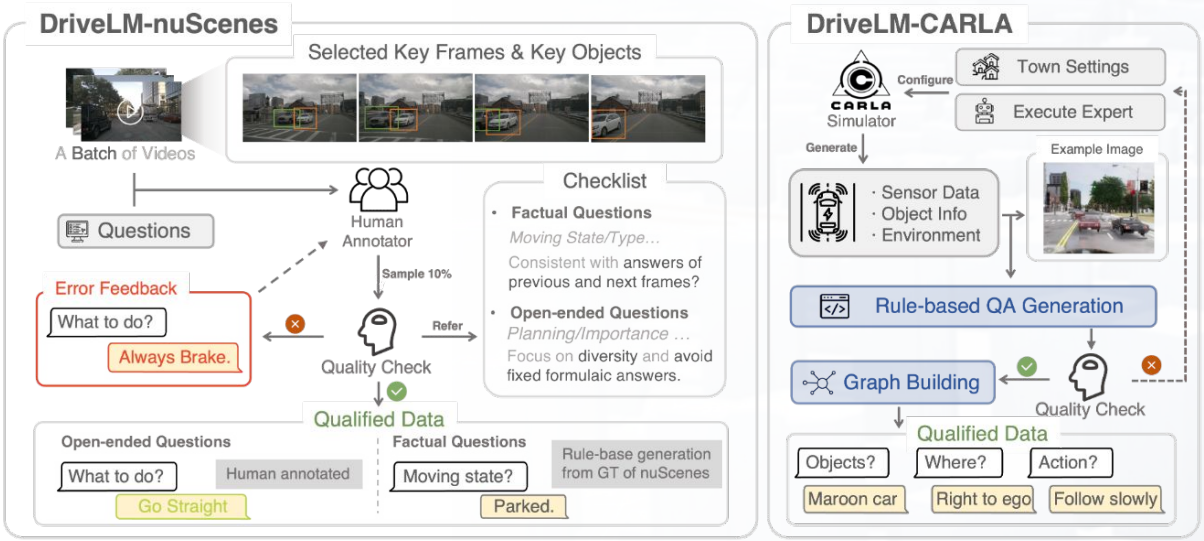






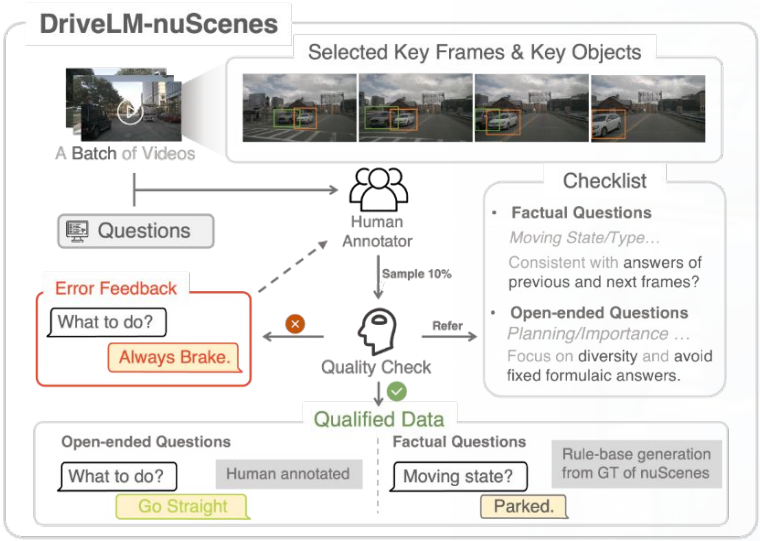
The critical part is **Graph Visual QA**, upon which we build **data**, model and metrics accordingly

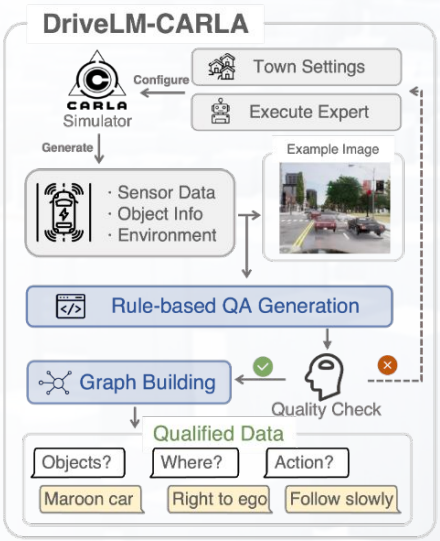
DriveLM | Data

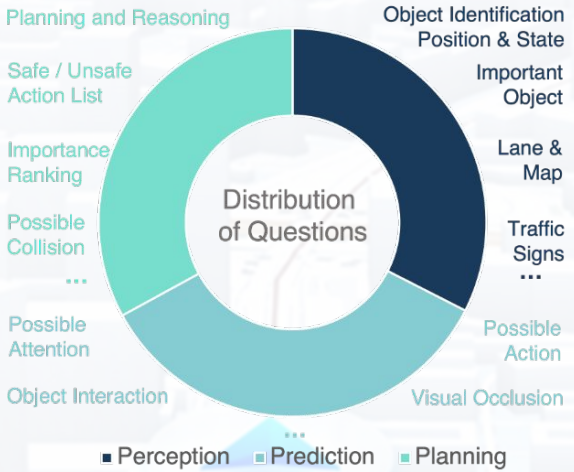


- To ensure the **data quality**, we introduce human annotation with multi-round quality check in nuScenes
- To **scale up** annotation, we adopt auto-labelling in CARLA

DriveLM | Data



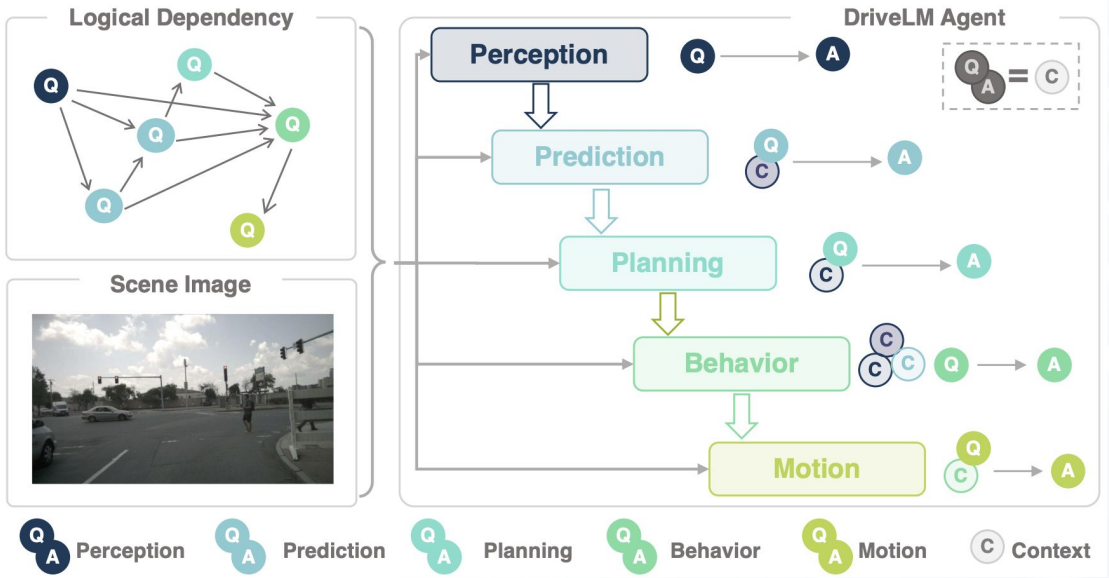




- To ensure the **data quality**, we introduce human annotation with multi-round quality check in nuScenes
- To **scale up** annotation, we adopt auto-labelling in CARLA

Diversity matters, spanning from perception to prediction and planning

DriveLM | Agent



- 🧩 General and scalable VLM architecture
- 🌍 Web-scale pre-training
- 🛠️ Fine-tuned end-to-end for planning
- 💡 Interpretable and interactive

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**

Conclusion:

- 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

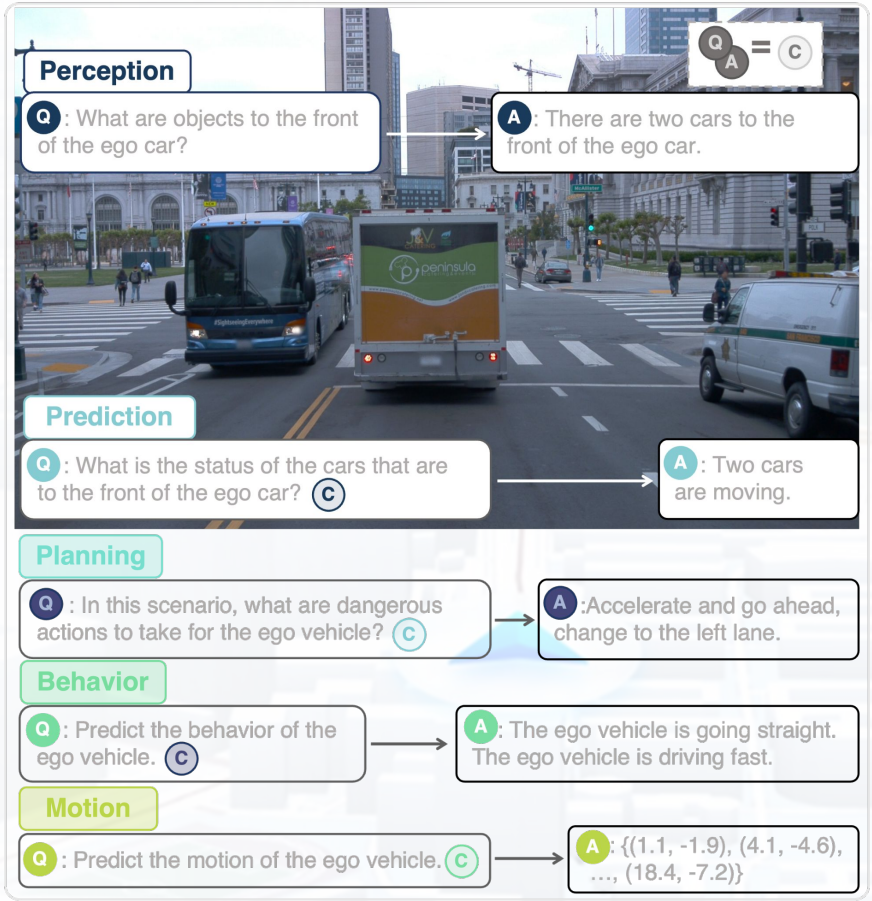
Method	Behavior Context	Motion Context	Behavior (B) Acc. ↑	Speed ↑	Steer ↑	Motion (M) ADE ↓	FDE ↓
Command Mean	-	-	-	-	-	7.98	11.41
UniAD-Single	-	-	-	-	-	4.16	9.31
BLIP-RT-2	-	-	-	-	-	2.78	6.47
DriveLM-Agent	None	B	35.70	43.90	65.20	2.76	6.59
	Chain	B	34.62	41.28	64.55	2.85	6.89
	Graph	B	**39.73**	**54.29**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

Conclusion:

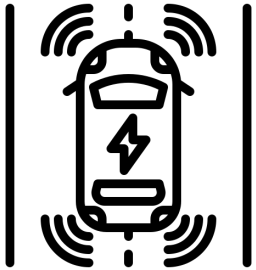
- Trained on DriveLM-Data (nuScenes-based), DriveLM-Agent (ours) gains **better zero-shot ability** on Waymo scenarios, overpassing other methods by a large margin.

Conclusion:

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

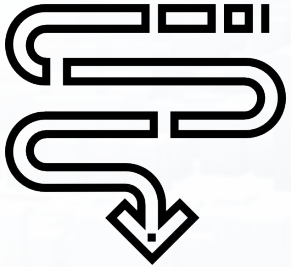


DriveLM - Limitation



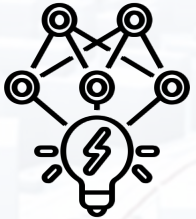
Driving-specific Inputs

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.

One-page Takeaway

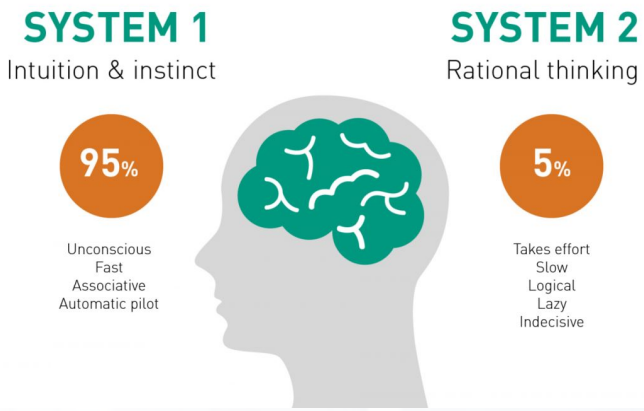
End-to-end Autonomous Driving

- Challenge: **Generalization & Explainability**
- Recent trend: use vision language model to **embed "world knowledge"** to solve 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**

System 1&2



System 1: fast, automatic, frequent, emotional, stereotypic, unconscious

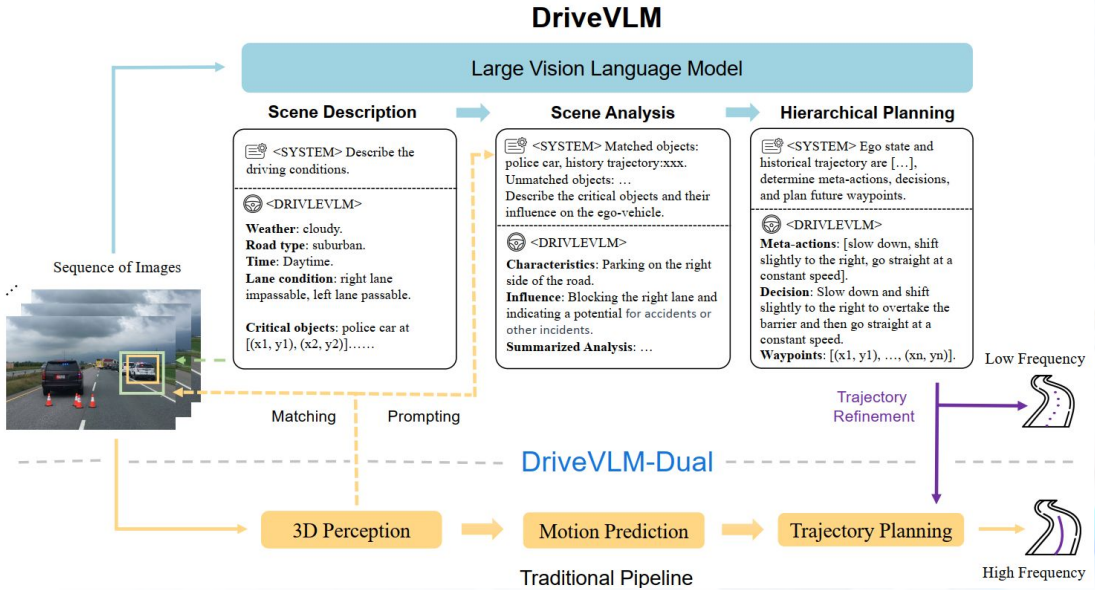
- Determine that an object is at a greater distance than another
- Drive a car on an empty road

System 2: Slow, effortful, infrequent, logical, calculating, conscious

- Look for the woman with the grey hair
- Park into a tight parking space

[1] Daniel Kahneman. Thinking, Fast and Slow. 2011.

System 1&2



Idea

- Incorporate results from the traditional pipeline to LVLM, as prompts
- Chain-of-thought VLM
- Take the low frequency VLM planning output as a reference to refine traditional trajectory planning (selectively attend to the additional information)

[1] Tian et al. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. arXiv, 2024.

OpenDriveLab

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

Q&A

End