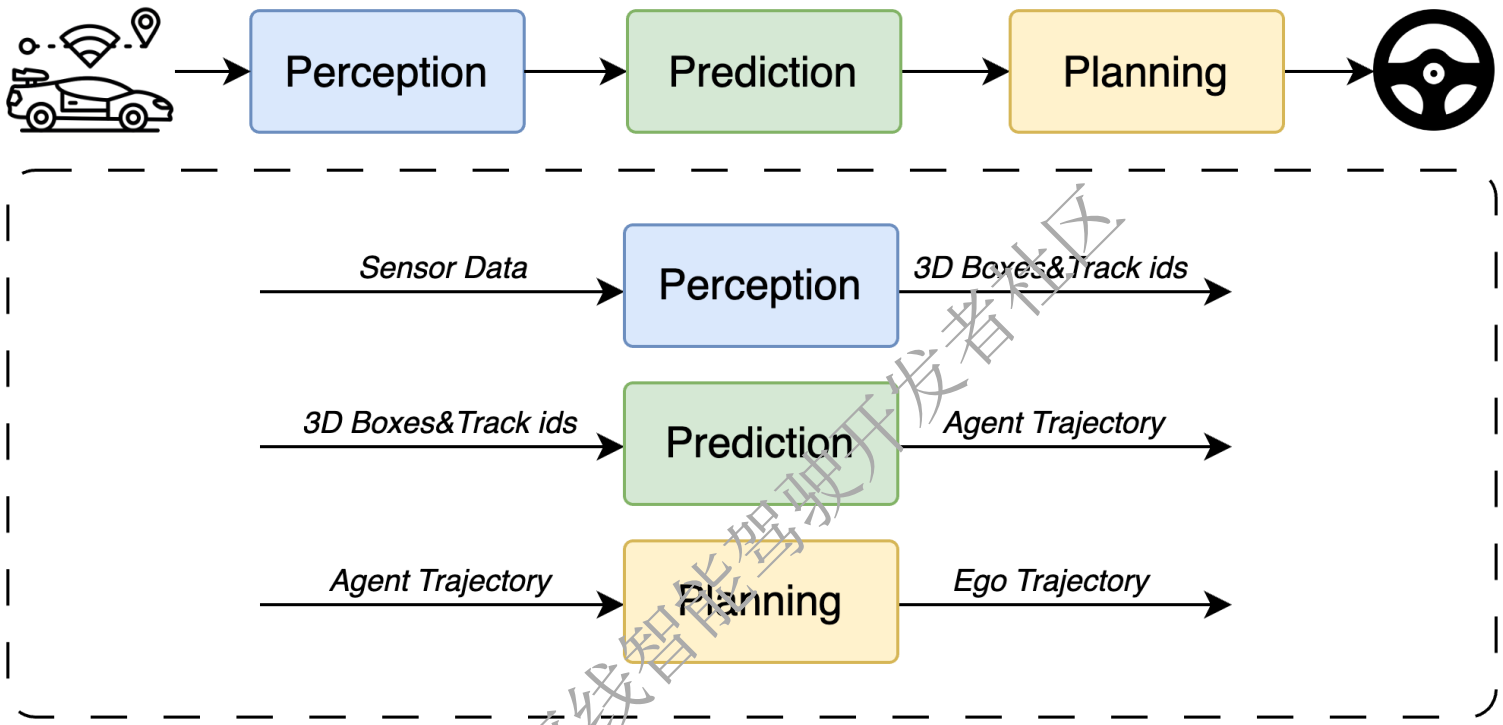
端到端自动驾驶技术的发展及现状

汇报人：李柯宇

目录

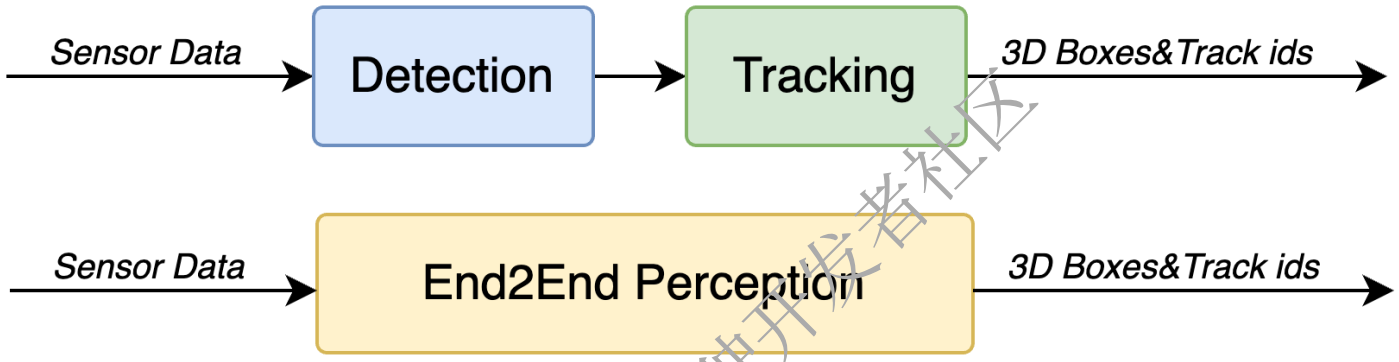
- 端到端定义及分类



- 目前自动驾驶系统的主流方案还是以级联式方案为主；
- 不同模块互不影响，分别开发，耦合度较低；
- 好处是可以并行开发，坏处是误差会累积；
- 不同模块都可以构成独立的端到端系统；

目录

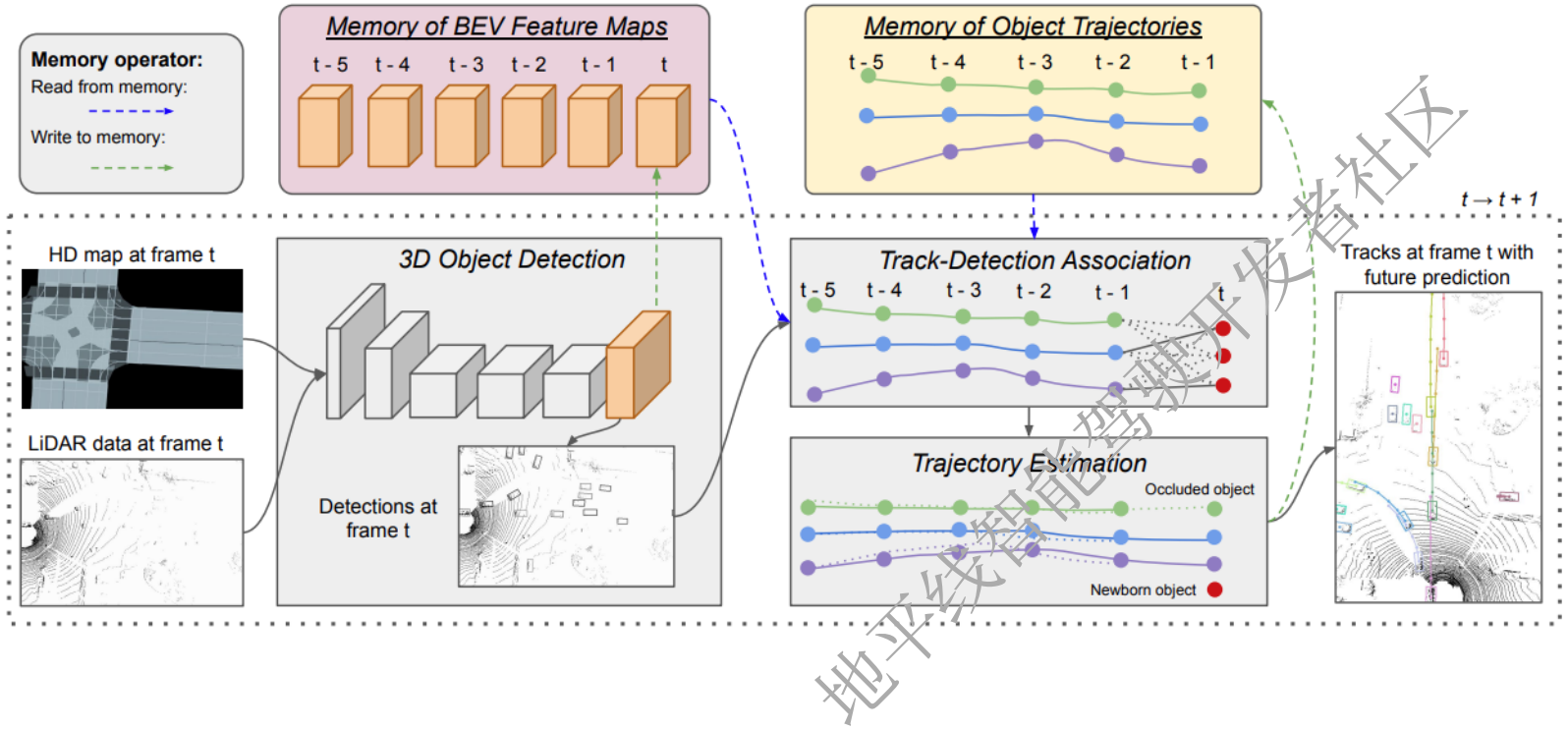
- 感知端到端



- 常规的感知模组可以细分为检测和追踪两部分，分别输出检测框和ID信息；
- 感知端到端模型同时输出检测框和ID信息；
- 常规方法中检测+追踪无法合起来的主要原因在于Tracking模块中的匈牙利匹配操作无法回传梯度；

- 感知端到端

PnPNet



$$C_{i,j} = \begin{cases} \text{MLP}_{\text{pair}}(f(\mathcal{D}_i^t), h(\mathcal{P}_j^{t-1})) & \text{if } 1 \le j \le M_{t-1}, \\ \text{MLP}_{\text{unary}}(f(\mathcal{D}_i^t)) & \text{if } j = M_{t-1} + i, \\ -\text{inf} & \text{otherwise} \end{cases} \tag{7}$$

- 以Learnable的MLP替换Tracking过程中的匈牙利匹配矩阵，避免了匈牙利匹配的梯度无法回传问题，从而实现了端到端的检测+追踪；

Liang M, Yang B, Zeng W, et al. Pnpnet: End-to-end perception and prediction with tracking in the loop[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11553-11562.

- 感知端到端

Sparse4D v3

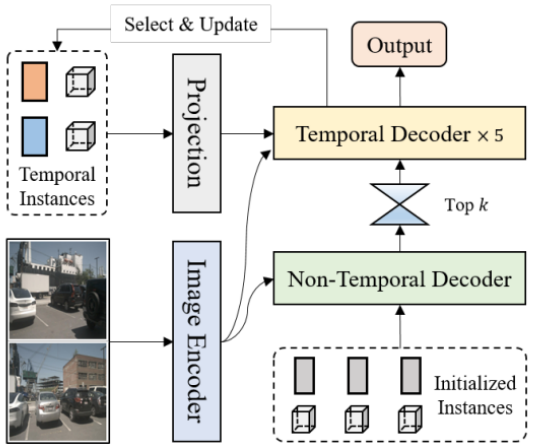


Figure 1: Overview of Sparse4D framework, which input mutli-view video and output the perception results of all frames.

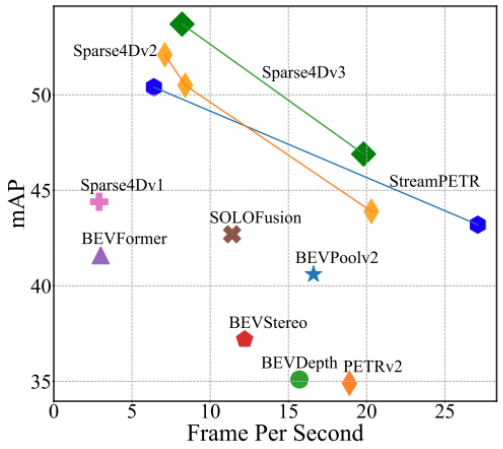
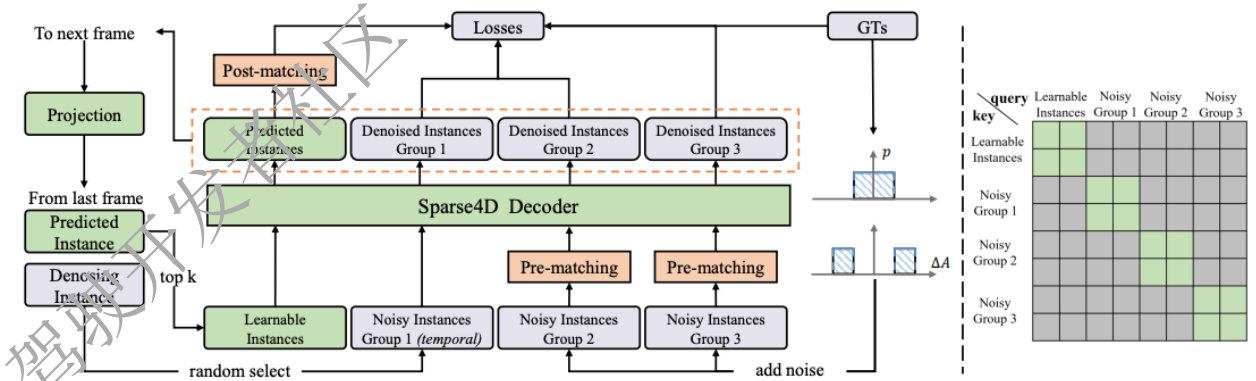


Figure 2: Inference efficiency (FPS) - perception performance (mAP) on nuScenes validation dataset of different algorithms.



(a) Pipeline of training with temporal denosing

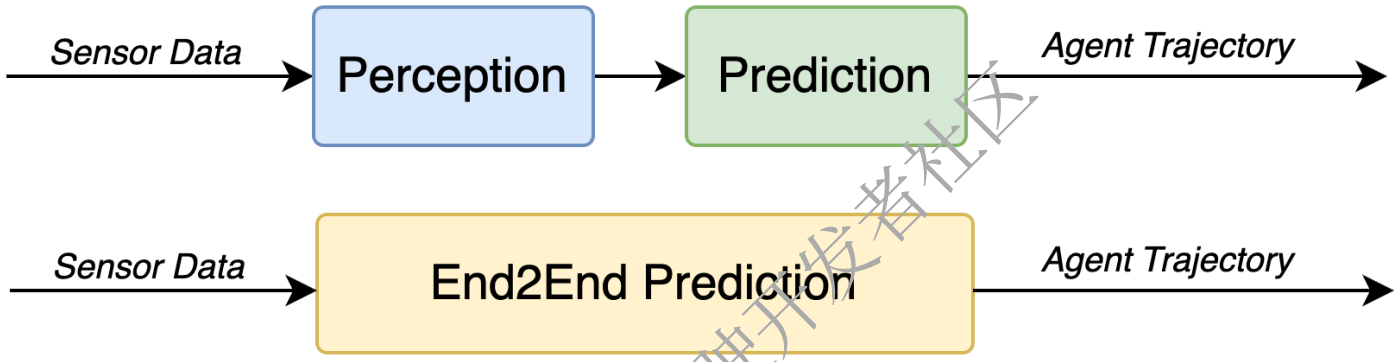
(b) Attention mask of self-attention

- 以DETR范式为基础，以query形式为物体表征的中介，同一个query在不同帧表示同一个物体，从而实现query既输出物体的检测框信息，也输出物体的Track id信息；
- DETR架构先有query(instance)特征，再有检测框信息，而Center-based等方法先有检测框，再有instance信息(RoI取)，所以DETR这种Pipeline天然适合做端到端；

Lin X, Pei Z, Lin T, et al. Sparse4d v3: Advancing end-to-end 3d detection and tracking[J]. arXiv preprint arXiv:2311.11722, 2023.

目录

- 预测端到端



- 常规的预测模组输入是感知的结果，输出是周围交通参与者的未来运动估计；
- 端到端的预测模型输入是传感器数据，输出是周围交通参与者的未来运动估计；

- 预测端到端

FaF

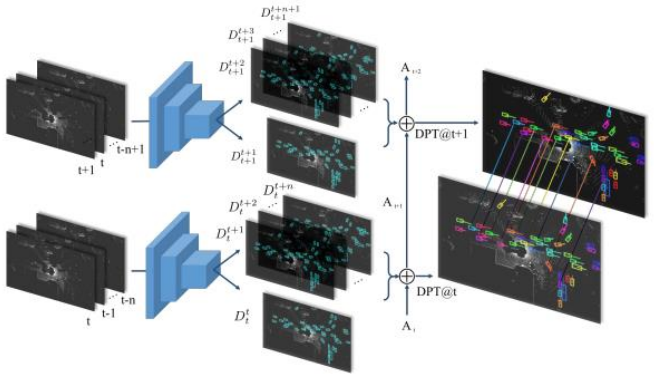
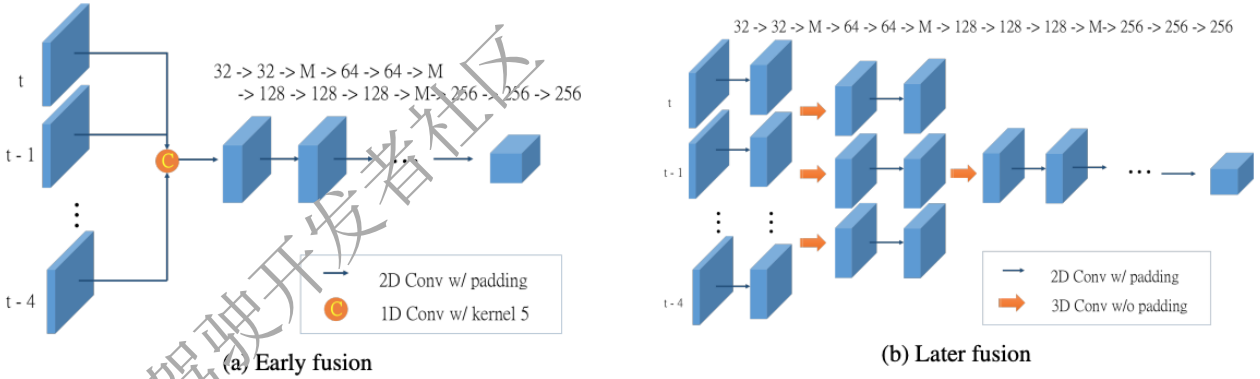


Figure 1: **Overview of our approach:** Our FaF network takes multiple frames as input and performs detection, tracking and motion forecasting.



(a) Early fusion (b) Later fusion

Figure 4: **Modeling temporal information**

- 输入多帧的点云数据，直接输出未来帧的物体的位置；
- 相比目标检测任务，可以看成是未来帧的目标检测任务，有了未来帧的目标检测框，就相当于是运动预测了；

Luo W, Yang B, Urtasun R. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 3569-3577.

- 预测端到端

PnPNet

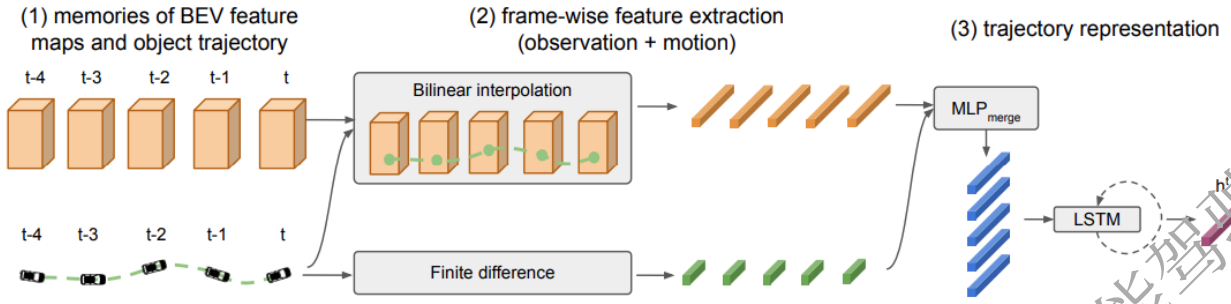


Figure 3. **The proposed trajectory level object representation.** Given an object trajectory, we first extract its sensor observation and motion features at each time step, and then apply an LSTM network to model the temporal dynamics.

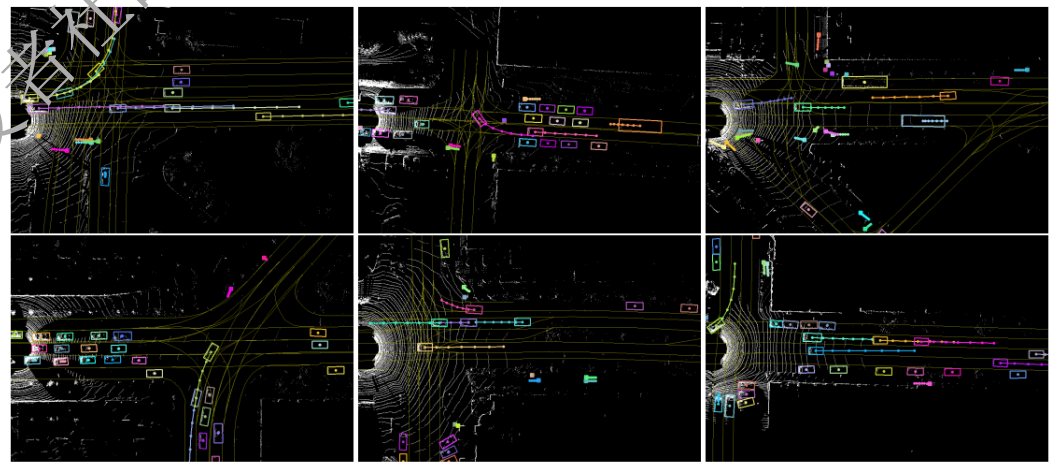
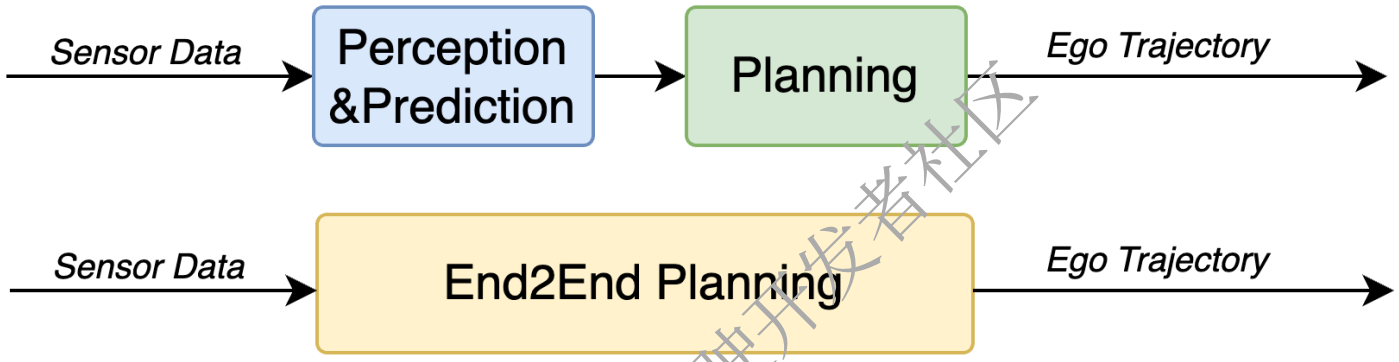


Figure 5. **Qualitative results of PnPNet on ATG4D.** We visualize the perception and prediction results of vehicles and pedestrians up to 100 meters far away, where the ego car is located at the middle left of each frame heading to the right.

- 经过track之后有了物体不同帧的instance级别的特征，然后对同一个物体不同时序上的特征做融合，再根据时序融合之后的特征经过MLP输出物体的未来轨迹；

目录

- 规控端到端



- 常规的规控模块输入是其他交通参与者(和静态要素等)的运动状态以及自车的运动状态，输出是自车未来时刻的运动状态估计，具体形式可以是轨迹点等；
- 规控端到端模型的输入是传感器数据，输出是自车的运动状态估计；

- 规控端到端

UniAD

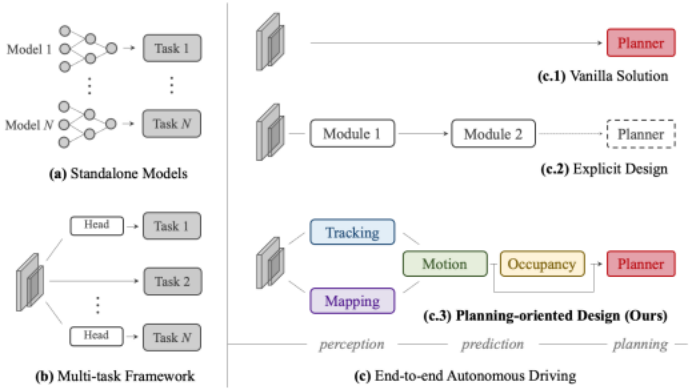


Figure 1. **Comparison on the various designs** of autonomous driving framework. **(a)** Most industrial solutions deploy separate models for different tasks. **(b)** The multi-task learning scheme shares a backbone with divided task heads. **(c)** The end-to-end paradigm unites modules in perception and prediction. Previous attempts either adopt a direct optimization on planning in (c.1) or devise the system with partial components in (c.2). Instead, we argue in (c.3) that a desirable system should be planning-oriented as well as properly organize preceding tasks to facilitate planning.

Design	Approach	Perception			Prediction		Plan
		Det.	Track	Map	Motion	Occ.	
(b)	NMP [101]	✓			✓		✓
	NEAT [19]			✓			✓
	BEVerse [105]	✓		✓		✓	
(c.1)	[14, 16, 78, 97]						✓
(c.2)	PnPNet† [37]	✓	✓		✓		
	ViP3D† [30]	✓	✓		✓		
	P3 [32]					✓	✓
	MP3 [11]			✓		✓	✓
	ST-P3 [38]			✓		✓	✓
	LAV [15]	✓		✓	✓		✓
(c.3)	**UniAD** (ours)	✓	✓	✓	✓	✓	✓

Table 1. **Tasks comparison and taxonomy.** "Design" column is classified as in Fig. 1. "Det." denotes 3D object detection, "Map" stands for online mapping, and "Occ." is occupancy map prediction. †: these works are not proposed directly for planning, yet they still share the spirit of joint perception and prediction. UniAD conducts five essential driving tasks to facilitate planning.

- 以规控为导向，最终目的是自车未来运动的估计；
- 涉及任务多，不同任务既做监督也可做中间结果的输出；
- 感知、预测、规控呈递进式关系，规控能利用到上游所有信息；

Hu Y, Yang J, Chen L, et al. Planning-oriented autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 17853-17862.

- 规控端到端

UniAD

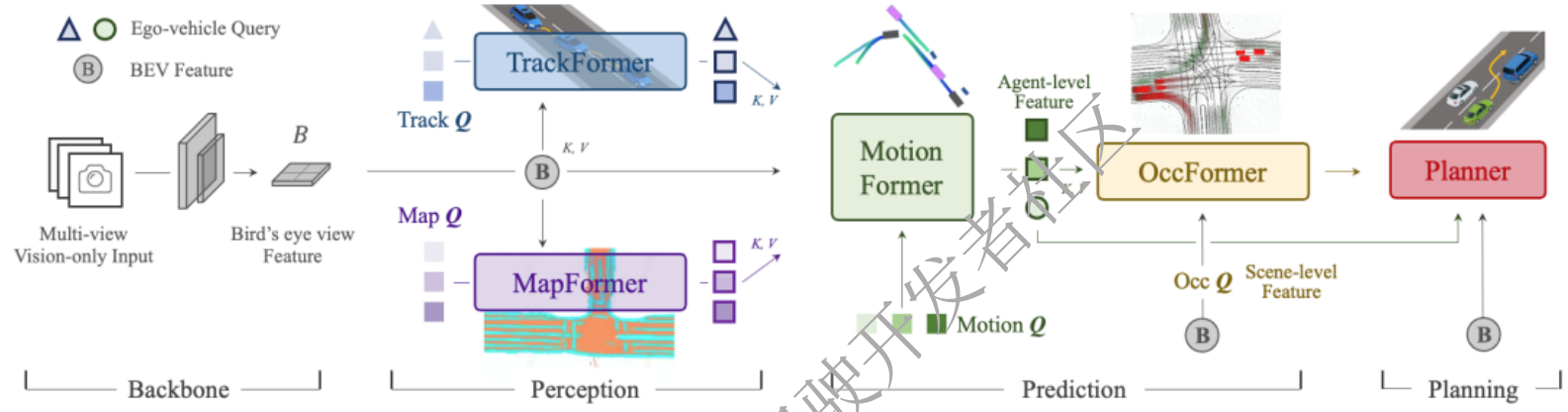


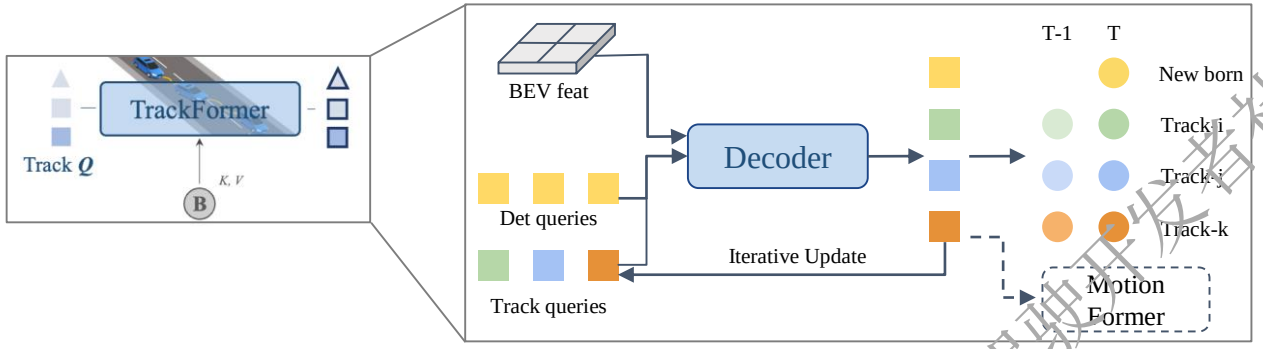
Figure 2. **Pipeline** of Unified Autonomous Driving (UniAD). It is exquisitely devised following planning-oriented philosophy. Instead of a simple stack of tasks, we investigate the effect of each module in perception and prediction, leveraging the benefits of joint optimization from preceding nodes to final planning in the driving scene. All perception and prediction modules are designed in a transformer decoder structure, with task queries as interfaces connecting each node. A simple attention-based planner is in the end to predict future waypoints of the ego-vehicle considering the knowledge extracted from preceding nodes. The map over occupancy is for visual purpose only.

- 以query作为所有任务的中间表示，感知等信息通过query的表征从上游传递到规控端；
- Track query用来检测物体和追踪物体，Map query构建BEV map，Motion query输出agent未来的运动状态估计，Occ query建模occupancy表示，Planner query输出自车未来轨迹规划；
- 整个模型可端到端训练/推理；

- 规控端到端

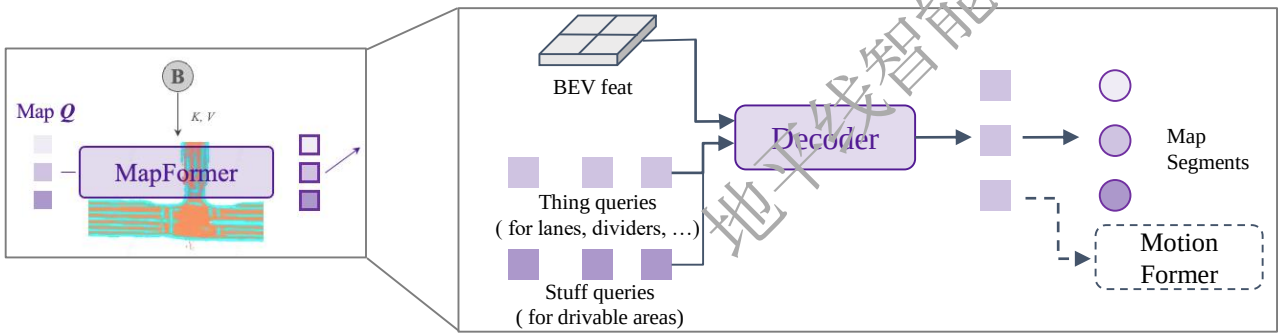
UniAD

TrackFormer



- 端到端检测+追踪，无需后处理操作

MapFormer

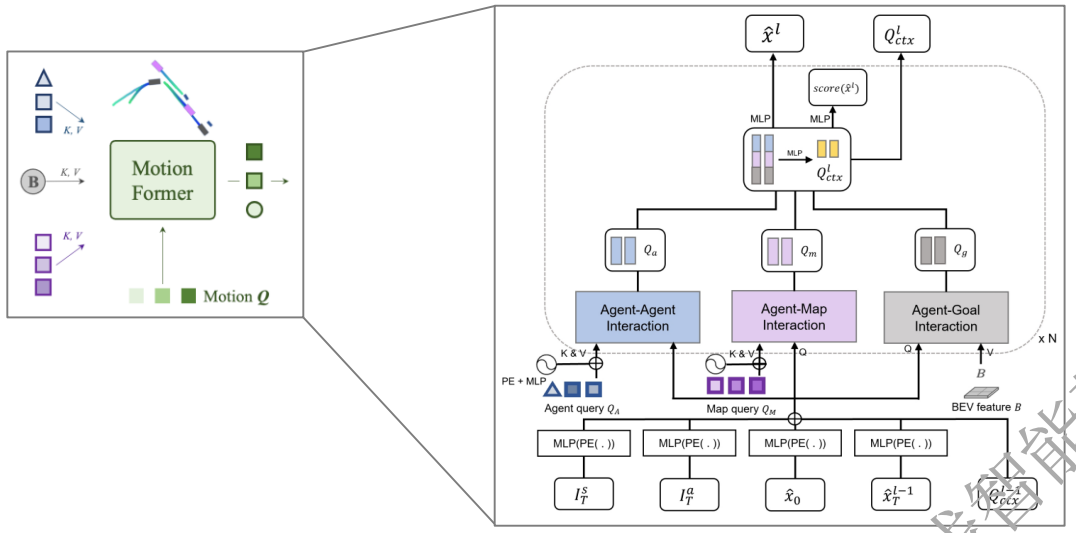


- Map query表征BEV map上的元素

- 规控端到端

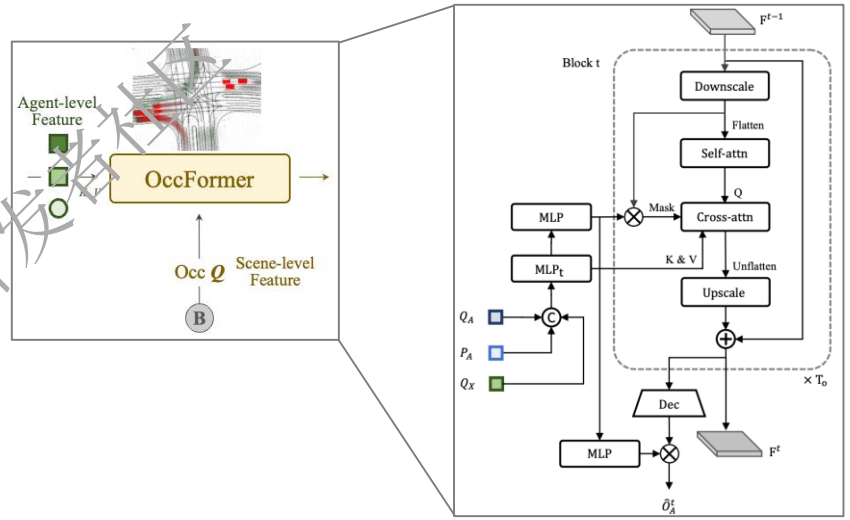
UniAD

MotionFormer



- Agent-agent、agent-map、agent-goal做交互

OccFormer



- 建模agent物体的未来时刻的2D occupancy表示

- 规控端到端

UniAD

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	0.152	-	-	-	-	-	-
6	✓	✓	✓			0.354	1.339	820	0.303	0.672	0.736(-9.7%)	1.066(-12.9%)	0.152	-	-	-	-	-	-
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%)	**1.003**(-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**

Table 2. **Detailed ablations on the effectiveness of each task.** We can conclude that two perception sub-tasks greatly help motion forecasting, and prediction performance also benefits from unifying the two prediction modules. With all prior representations, our goal-planning boosts significantly to ensure safety. UniAD outperforms naive MTL solution by a large margin for prediction and planning tasks, and it also owns the superiority that *no* substantial perceptual performance drop occurs. Only main metrics are shown for brevity. "avg.L2" and "avg.Col" are the average values across the planning horizon. ∗: ID-0 is the MTL scheme with separate heads for each task.

- ID 4-6 不同任务对预测的影响；
- ID 7-9 不同任务对Occ的影响；
- ID 10-12 不同任务对Planning的影响；

- 规控端到端

VAD

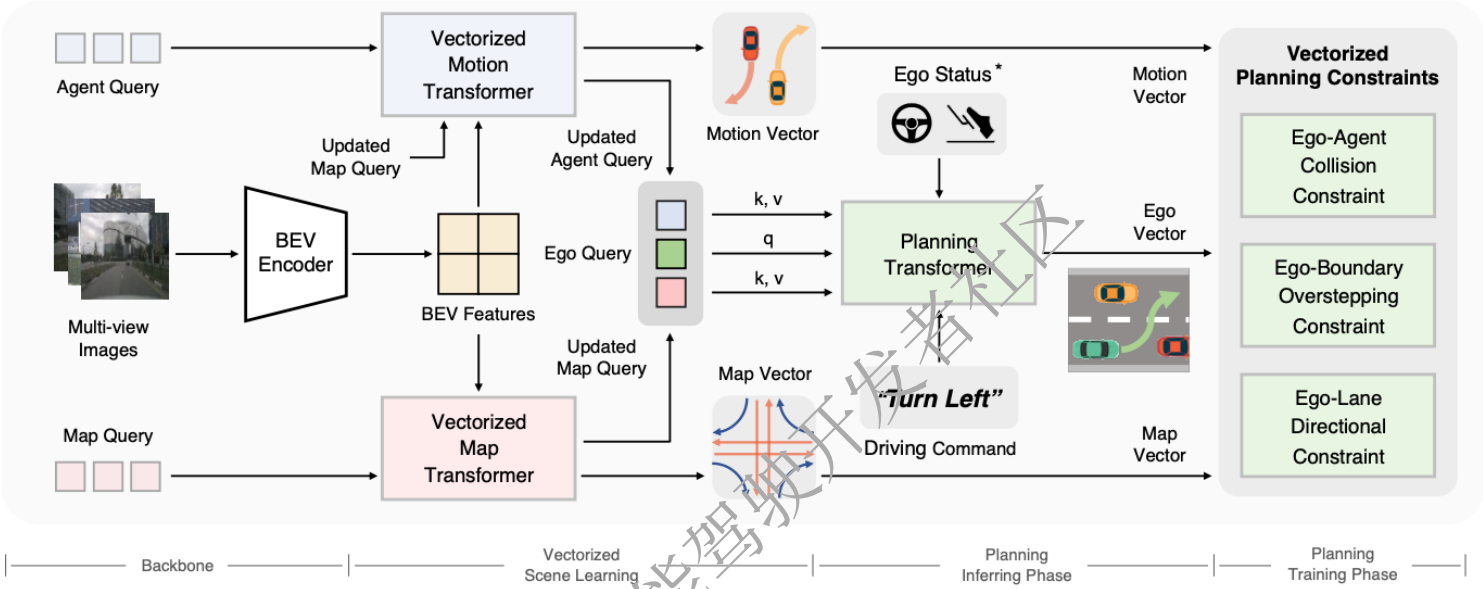


Figure 2. **Overall architecture of VAD.** The full pipeline of VAD is divided into four phases. Backbone includes an image feature extractor and a BEV encoder to project the image features to the BEV features. Vectorized Scene Learning aims to encode the scene information into agent queries and map queries, as well as represent the scene with motion vectors and map vectors. In the inferring phase of planning, VAD utilizes an ego query to extract map and agent information via query interaction and outputs the planning trajectory (represented as ego vector). The proposed vectorized planning constraints regularize the planning trajectory in the training phase. *: optional.

- 不同任务都用向量化思想进行表征；
- 简化了感知任务，同时都用稀疏化的query形式表达；
- 自车ego query跟向量化的信息进行交互，从而获得感知等信息，然后输出自车轨迹预测；

Jiang B, Chen S, Xu Q, et al. Vad: Vectorized scene representation for efficient autonomous driving[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 8340-8350.

- 规控端到端

VAD

Method	L2 (m) ↓				Collision (%) ↓				Latency (ms)	FPS
	1s	2s	3s	Avg.	1s	2s	3s	Avg.		
NMP† [49]	-	-	2.31	-	-	-	1.92	-	-	-
SA-NMP† [49]	-	-	2.05	-	-	-	1.59	-	-	-
FF† [18]	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43	-	-
EO† [24]	0.67	1.36	2.78	1.60	0.04	0.09	0.38	0.33	-	-
ST-P3 [19]	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71	628.3	1.6
UniAD [21]	0.48	0.96	1.65	1.03	**0.05**	0.17	0.71	0.31	555.6	1.8
VAD-Tiny	0.46	0.76	1.12	0.78	0.21	0.35	0.58	0.38	**59.5**	**16.8**
VAD-Base	**0.41**	**0.70**	**1.05**	**0.72**	0.07	**0.17**	**0.41**	**0.22**	224.3	4.5
VAD-Tiny‡	0.20	0.38	0.65	0.41	0.10	0.12	0.27	0.16	59.5	16.8
VAD-Base‡	0.17	0.34	0.60	0.37	0.07	0.10	0.24	0.14	224.3	4.5

Table 1. **Open-loop planning performance.** VAD achieves the best end-to-end planning performance and the fastest inference speed on the nuScenes [1] val dataset. LiDAR-based methods are denoted with †. ‡ denotes taking ego status information as input. In open-loop evaluation, we deactivate ego status information for a fair comparison. FPS of ST-P3 and VAD are measured on one NVIDIA Geforce RTX 3090 GPU. FPS of UniAD is measured on one NVIDIA Tesla A100 GPU.

- 规控端到端

VADv2

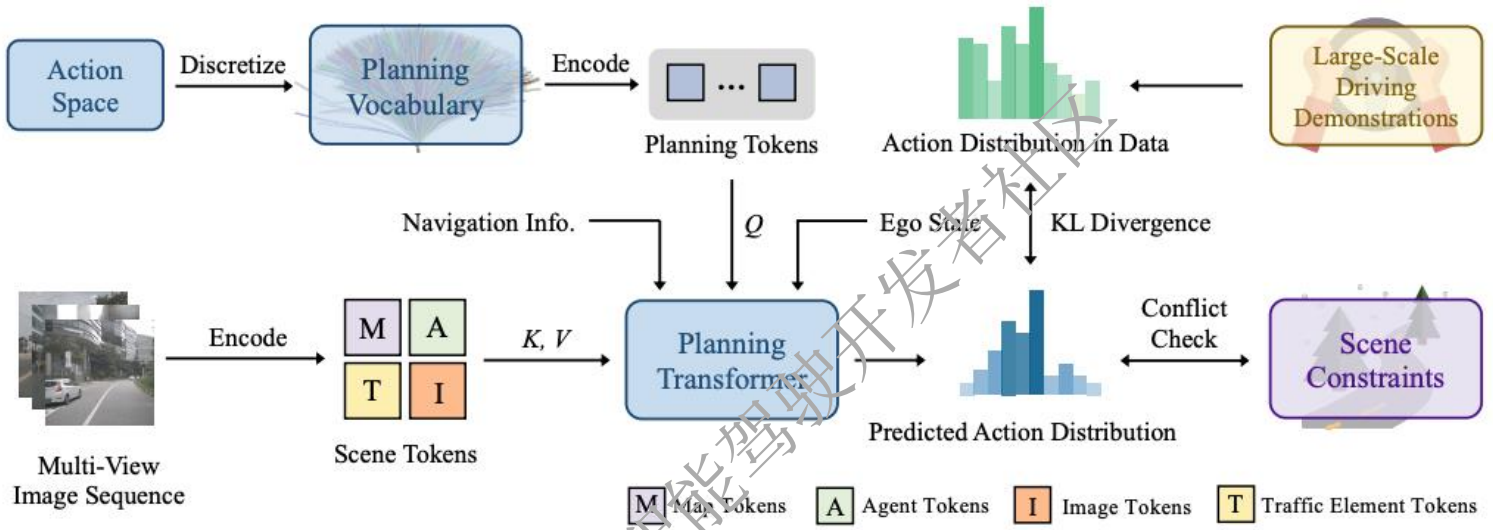


Figure 2. **Overall architecture of VADv2.** VADv2 takes multi-view image sequence as input in a streaming manner, transforms sensor data into environmental token embeddings, outputs the probabilistic distribution of action, and samples one action to control the vehicle. Large-scale driving demonstrations and scene constraints are used to supervise the predicted distribution.

- 输出形式为action的概率分布，相比单纯的轨迹点有了更好的不确定性表达；
- 大幅提高了闭环评测效果；

Chen S, Jiang B, Gao H, et al. VADv2: End-to-End Vectorized Autonomous Driving via Probabilistic Planning[J]. arXiv preprint arXiv:2402.13243, 2024.

- 规控端到端

VADv2

Method	Modality	Reference	Driving Score ↑	Route Completion ↑	Infraction Score ↑
CILRS [9]	C	CVPR 19	7.8	10.3	0.75
LBC [6]	C	CoRL 20	12.3	31.9	0.66
Roach [54]	C	ICCV 21	41.6	96.4	0.43
Transfuser† [40]	C+L	TPAMI 22	31.0	47.5	0.77
ST-P3 [18]	C	ECCV 22	11.5	83.2	-
VAD [23]	C	ICCV 23	30.3	75.2	-
ThinkTwice [21]	C+L	CVPR 23	70.9	95.5	0.75
MILE [16]	C	NeurIPS 22	61.1	97.4	0.63
Interfuser [45]	C	CoRL 22	68.3	95.0	-
DriveAdapter+TCP [20]	C+L	ICCV 23	71.9	97.3	0.74
DriveMLM [49]	C+L	arXiv	76.1	98.1	0.78
VADv2	C	Ours	85.1	98.4	0.87

Table 1. **Closed-loop evaluation on Town05 Long benchmark.**

Method	Modality	Driving Score ↑	Route Completion ↑
CILRS [9]	C	7.5	13.4
LBC [6]	C	31.0	55.0
Transfuser [40]	C+L	54.5	78.4
ST-P3 [18]	C	55.1	86.7
VAD [23]	C	64.3	87.3
VADv2	C	89.7	93.0

Table 2. **Closed-loop evaluation on Town05 Short benchmark.**

- 规控端到端-生成式模型

GAIA-1

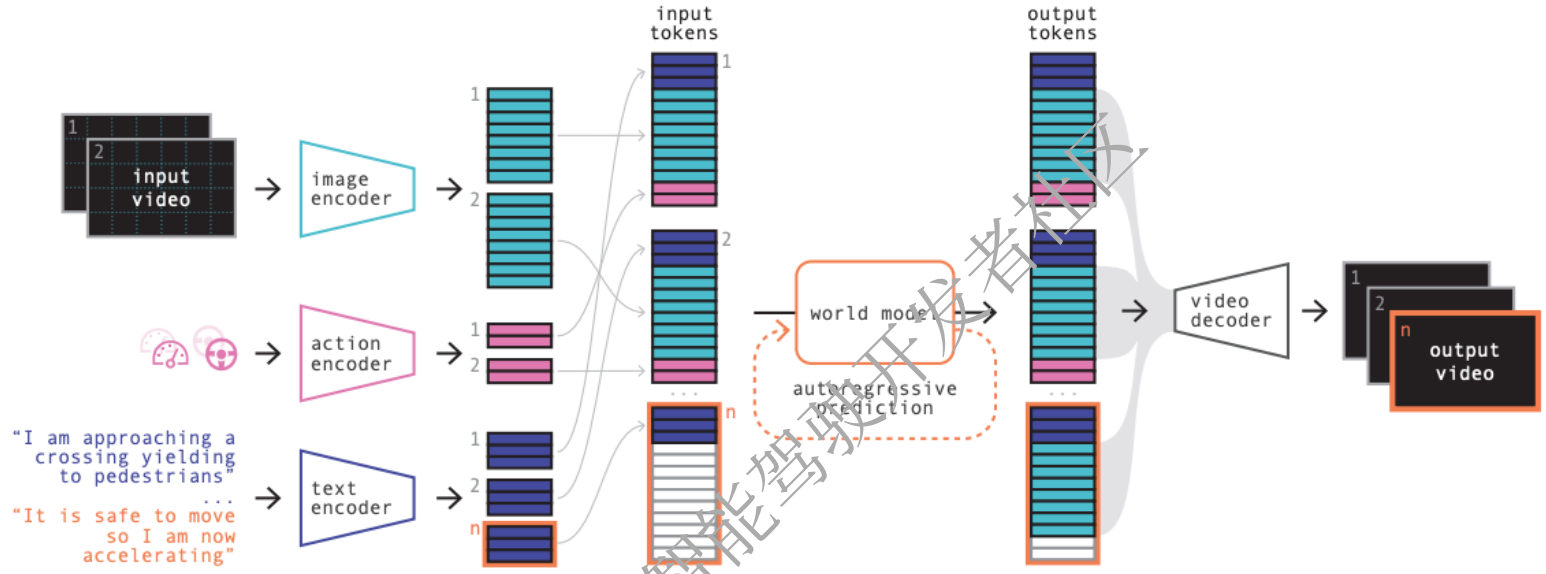


Figure 2: Architecture of GAIA-1. First, we encode information from all input modalities (video, text, action) into a common representation: images, text and actions are encoded as a sequence of tokens. The world model is an autoregressive transformer that predicts the next image token conditioned on past image, text, and action tokens. Finally, the video decoder maps the predicted image tokens back to the pixel space, at a higher temporal resolution.

Hu A, Russell L, Yeo H, et al. Gaia-1: A generative world model for autonomous driving[J]. arXiv preprint arXiv:2309.17080, 2023.

- 规控端到端-生成式模型

ADriver-I

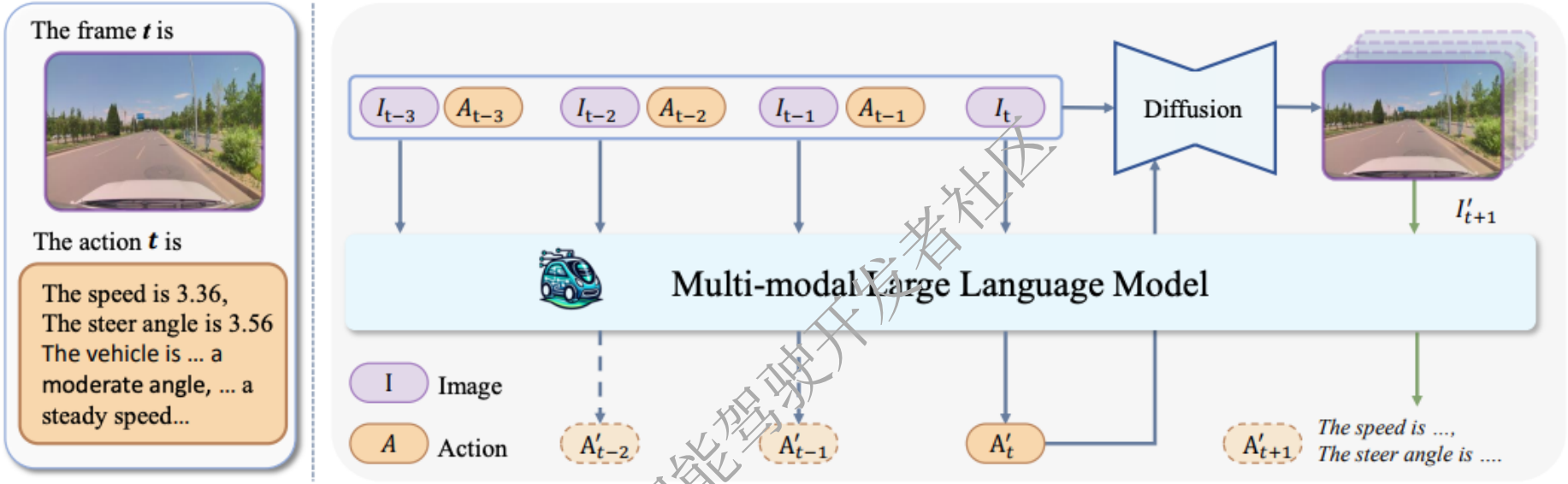


Figure 1. Overview of of our ADriver-I framework. It takes the historical interleaved vision-action pairs $\{I, A\}$ and current visual token as inputs. The multi-modal large language model (MLLM) reasons out the control signal A_t of current frame. The predicted action A_t is further used as the condition prior of video latent diffusion model (VDM) to generate the future four frames. The predicted next frame I'_{t+1} is selected and further input to the MLLM to produce the control signal A'_{t+1}. Such process (green line) can be repeated for infinite times and it achieves the self-learning in the world created by itself. The dashed lines represent the output only in training process.

Jia F, Mao W, Liu Y, et al. Adriver-i: A general world model for autonomous driving[J]. arXiv preprint arXiv:2311.13549, 2023.

目录

目录

• 规控端到端实践问题中困难与挑战

地平线 Horizon Robotics

- **算法方案：**
  - 技术路线不收敛，不同方案存在较大分歧；
  - 暂时没有较成熟的pipeline，包括数据输入形式以及输出形式；
- **数据方面：**
  - 对数据要求较高，包括数据数量以及数据质量；
  - 数据标注需求发生变化，跟过去感知等任务标注需求不同；
  - 评估数据是否有价值的标准发生变化；
- **算力要求：**
  - 训练需要的数据量显著提高，导致对模型算力需求也跟着提高；
- **评测方式：**
  - 开闭环评测存在明显diff，对闭环仿真要求高；


Thank you!
地平线智能驾驶开发者社区

Q&A