合肥工業大學

端到端自动驾驶技术发展与未来展望

贾伟

2025.10.24

合肥工业大学计算机与信息学院（人工智能学院）

报告提纲

背景

端到端自动驾驶

端到端自动驾驶（End-to-End Autonomous Driving）是主流的自动驾驶范式，旨在通过一个统一的模型或流程，直接从原始传感器输入（如摄像头、激光雷达等）生成最终的驾驶行为（如控制指令、行驶轨迹等），尽量减少或消除传统模块化架构中人为设定的任务边界（如感知、预测、规划、控制等）。

优点

- 架构简单、减少模块依赖
- 全局优化能力更强
- 可从海量数据中学习经验
- 无需复杂人工特征设计
- 适应复杂、非结构化道路能力强

不再依赖“感知→预测→规划→控制”多个模块，系统更紧凑，减少误差传递和逻辑冲突。

模型从原始输入直接学习驾驶行为，能够学到最优策略，而不是被人为规则限制。

通过真实驾驶数据模仿人类驾驶行为，有利于持续优化和快速迭代（如特斯拉FSD）。

自动提取视觉、环境特征，避免手工标注车道线、目标检测等步骤。

比传统规则系统更擅长处理无标线道路、乡村路、杂乱交通等复杂环境。

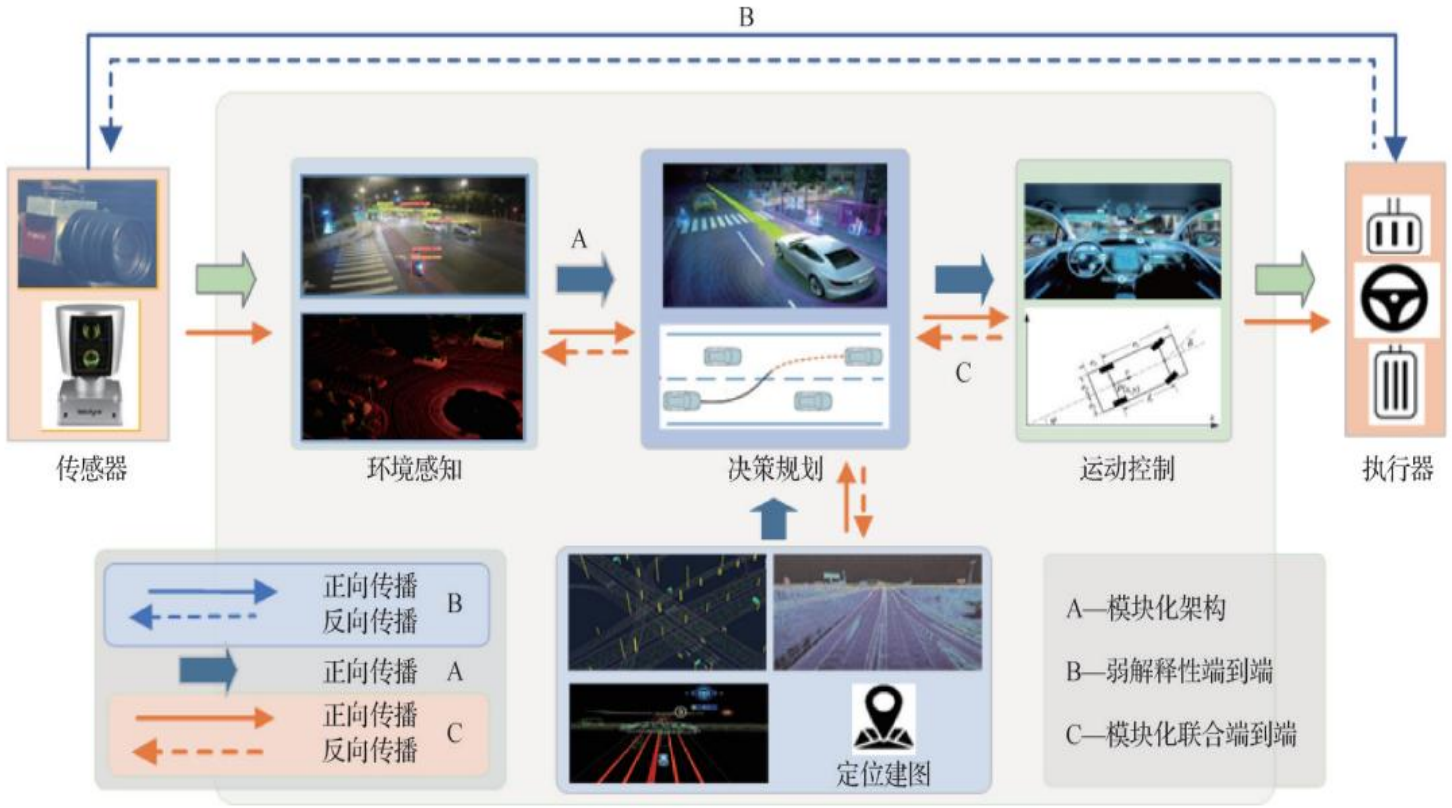
端到端自动驾驶

端到端自动驾驶（End-to-End Autonomous Driving）是主流的自动驾驶范式，旨在通过一个统一的模型或流程，直接从原始传感器输入（如摄像头、激光雷达等）生成最终的驾驶行为（如控制指令、行驶轨迹等），尽量减少或消除传统模块化架构中人为设定的任务边界（如感知、预测、规划、控制等）。

缺点

■ 可解释性差（黑盒问题）	很难理解模型做某个决策的原因，事故追责困难，缺乏监管透明度。
■ 对极端和罕见场景（长尾问题）处理不足	对少见危险情况容易失效。
■ 数据和算力需求巨大	要覆盖各种道路、天气、文化驾驶习惯，需要海量真实或仿真数据和高算力训练。
■ 难调试、难验证安全性	出问题时难定位在哪个环节出错，也难以通过功能安全认证（ISO 26262等）。
■ 交通规则和伦理难显式嵌入	很难实现“明确遵守红灯”“必须停车让人”等硬约束规则。
■ 鲁棒性限制	传感器误差、遮挡、恶劣天气下模型容易输出错误控制指令。

从传统模块化到端到端



- **传统模块化：**感知、定位与建图模块、预测、规划、控制等功能独立开发，优化目标不一致，易累积误差。

- **一段式端到端：**用一个学习模型把传感器输入直接映射到车辆控制指令。感知至控制一体化学习，特征共享充分，但可解释性弱、调试困难。

- **模块化端到端：**传统模块化与端到端思想的折中与融合，保留一定的模块化结构，但在模块内部或模块之间引入数据驱动的模型，并尽可能实现端到端的联合优化。模块间可微，端到端协同优化，兼顾整体性能与可解释性。

从传统模块化到端到端

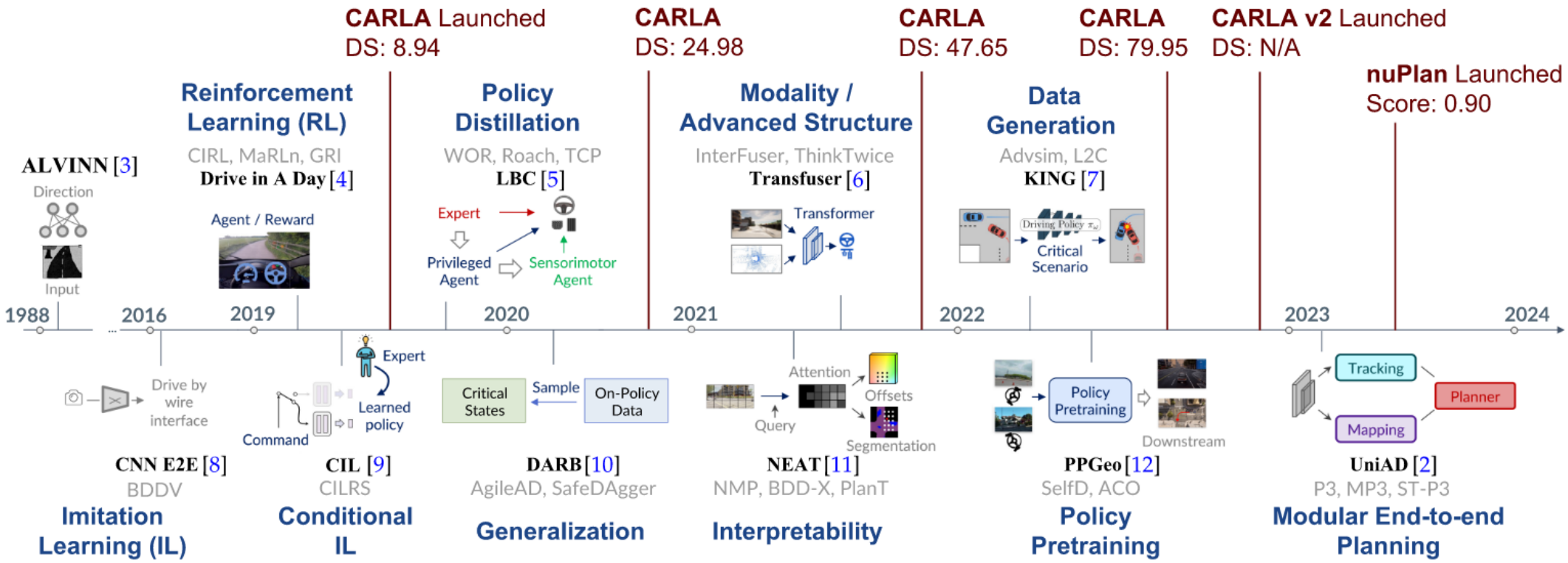
对比维度	一段式端到端（Single-Step E2E）	模块化端到端（Modular E2E）
流程结构	图像 → 直接输出方向盘/油门	图像 → 中间表示（如轨迹 / BEV / 世界模型）→ 控制
可解释 / 可调试性	很差，内部过程黑箱	较好，可分阶段优化、监控和安全验证
工程落地难度	高，难以满足安全和法规要求	低，更接近现有自动驾驶感知—预测—规划架构
代表系统	NVIDIA PilotNet、Tesla 早期方案	Tesla FSD 12（规划端到端，但保留世界模型）、Waymo ChauffeurNet、Apollo E2E Planner
主流企业使用程度	研究型、实验性	工业界主流选择

模块化端到端为什么成为主流?

1. **兼顾端到端的高效率 + 传统架构的可控性**
   - 将感知、预测、规划集成统一，但不是“一步到控制信号”，而是“输出轨迹 / 世界状态”。
2. **更容易满足安全性与法律监管要求**
   - 可以监控中间结果（如障碍物检测、轨迹规划），便于验证和冗余设计。
3. **与L2~L4级自动驾驶部署兼容性强**
   - 车厂现有系统都是模块化架构（感知-预测-规划），E2E模块能无缝嵌入。

端到端自动驾驶发展路线

2024 TPAMI综述论文：End-to-End Autonomous Driving: Challenges and Frontiers



DAVE-2 模仿学习 → 强化学习 Learning to drive a day → Transformer 与多模态TransFuser → UniAD 模块化端到端 → 生成式人工智能 与端到端LMDriver

端到端自动驾驶发展路线

2024 TPAMI综述论文：End-to-End Autonomous Driving: Challenges and Frontiers

一、思想萌芽与可行性验证（1980s - 2000s）

✓ 标志性工作：1989年的 ALVINN (Autonomous Land Vehicle In a Neural Network)

二、深度学习的复兴与初步探索（2010s中期）

✓ 标志性工作：2016年NVIDIA发表的论文《End to End Learning for Self-Driving Cars》

三、性能提升与可解释性探索（2010s后期）

✓ 模仿学习主导：行为克隆（BC）成为主流方法

✓ 强化学习探索：开始在仿真环境中尝试RL方法

✓ 多模态融合：开始整合相机、激光雷达等多种传感器

四、深化期：架构优化与性能提升（2020-2023）

✓ Transformer应用：引入注意力机制处理多模态数据

✓ BEV表示：鸟瞰图成为主流中间表示方式

✓ 世界模型：开始构建预测环境动态的内部模型

五、融合期：大模型与基础模型整合（2023至今）

✓ 大语言模型融合：LLM提供高级语义理解和指令解析

✓ 多模态基础模型：视觉-语言联合建模能力突破

✓ 分层智能架构：快慢系统协同决策

✓ **世界模型和VLA成为主流范式**

端到端自动驾驶发展路线

2024 TPAMI综述论文：End-to-End Autonomous Driving: Challenges and Frontiers

1. **演进规律**：体现了**从分离到集成、从规则到学习、从单一到融合**、从**不智能到智能**的技术演进规律
2. **核心驱动力**：
   - **人工智能技术的飞跃**
   - **传感技术、网络的进步**
   - **数据规模扩大**：从少量数据到海量级样本
   - **计算能力提升**：从有限算力到GPU加速
   - **算法创新**：从简单网络到Transformer架构
   - **评估体系完善**：从开环到闭环测试
3. **当前状态**：正处于**大模型与传统方法融合**的关键时期，注重**安全性、可解释性和人性化体验**
4. **未来方向**：向**更智能、更安全、更人性化**的自动驾驶系统发展，最终实现真正可信赖的自动驾驶

端到端自动驾驶技术已经从最初的概念验证发展到现在的系统集成阶段，未来将通过与大语言模型和基础模型的深度融合，实现更加智能和人性化的自动驾驶体验。

经典端到端算法(E2E)

端到端自动驾驶范式

2024 TPAMI综述论文：End-to-End Autonomous Driving: Challenges and Frontiers

范式	核心思想	优势	主要挑战	典型应用场景
■ 行为克隆(BC)	监督学习，模仿专家动作	简单高效，无需奖励函数	复合错误，因果混淆	大多数早期和基础端到端模型
■ 逆强化学习(IRL)	从专家行为反推奖励函数	泛化能力强，具解释性潜力	计算复杂，奖励模糊	需要更强泛化能力的系统
■ 强化学习(RL)	试错交互，最大化累积奖励	能发现超专家策略，考虑长期回报	样本效率低，奖励设计难，安全性差	主要在仿真中微调或研究
■ 策略蒸馏	教师(特权)-学生(传感器)学习	性能强大，缓解因果混淆	需要特权信息，蒸馏效率	比较先进的系统

- 端到端自动驾驶的技术范式经历了从纯模仿到模仿与反推奖励结合，再到利用特权信息蒸馏的演进。
- 强化学习由于其在现实世界中的挑战，目前主要扮演辅助优化的角色。
- **未来的趋势将是融合世界模型、基础模型、VLA和更大规模数据的混合范式，以构建更通用、更安全、更智能的自动驾驶系统。**

端到端自动驾驶范式

输出	输入	输出	是否有显式模块	核心方法	优点	缺点
1、一段式端到端	原始传感器数据（图像/点云）	控制信号 或 轨迹	无	深度学习（CNN/Transformer）	简单、数据驱动、潜在泛化强	黑箱、难解释、难落地
2、模块化端到端	原始数据 或 中间特征	控制 / 轨迹	有（但模块更智能）	深度学习 + 联合优化	工程友好、可解释性较强、可优化	仍有模块边界，设计复杂
3、模仿学习型	传感器数据 + 人类驾驶行为	驾驶行为（控制/轨迹）	可有可无	行为克隆 / 条件模仿学习	行为自然、数据直观	依赖数据质量、泛化性有限
4、强化学习型	传感器数据 + 环境状态	驾驶策略 / 行为	可有可无	RL算法（PPO/SAC等）	可优化长期目标、自探索策略	训练不稳定、安全性难保障

5、世界模型+ 端到端：先让模型学习预测环境将如何响应自己的行为（即构建世界模型），再基于此进行规划与决策

6、视觉-语言-行动模型：引入语言理解能力，使自动驾驶系统能理解导航指令、交通标志语义、与人类交互等

7、世界模型+基础模型+VLA+更大规模数据的混合范式

8、新的突破?

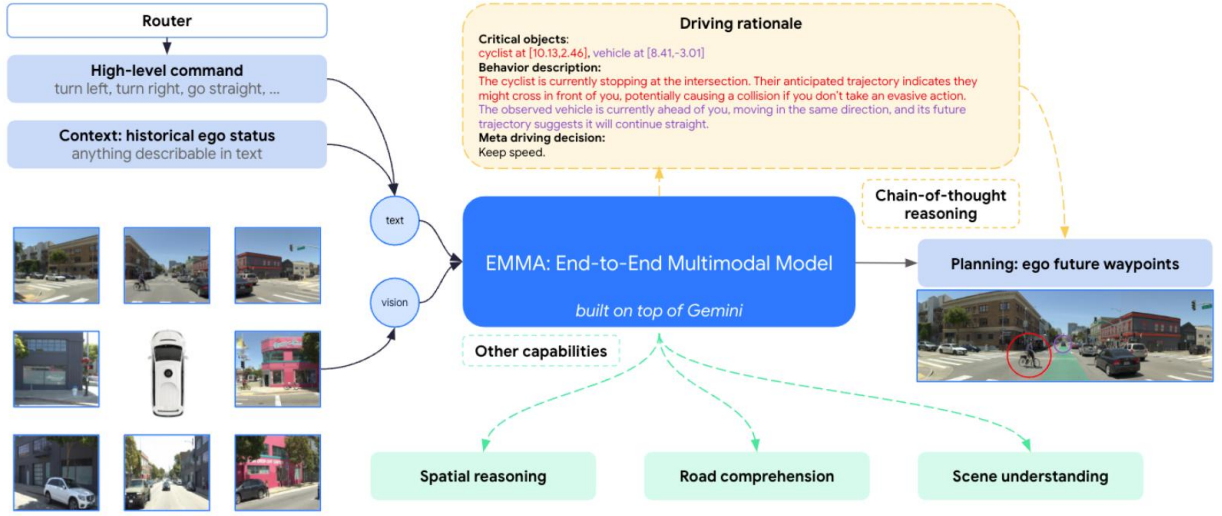
经典端到端算法-分类

类型	代表方法	输入	输出	特点
经典CNN端到端	PilotNet (NVIDIA, 2016)	单目图像	方向盘转角	最经典，简单直接，仅控制转向
条件模仿学习	Codevilla et al. (2018)	图像 + 驾驶指令	控制信号	引入高级语义指令，增强可控性
准端到端（轨迹规划）	ChauffeurNet (Waymo, 2018)	抽象环境表示	轨迹/控制	不是原始像素输入，但结构统一
多模态融合端到端	TransFuser (2021)	多摄像头 + Transformer	控制/轨迹	强调多视角融合与Transformer应用
全栈统一端到端	UniAD (Hu et al., 2023)	多传感器	规划/控制	感知-预测-规划一体化，工业级

- **纯图像到控制的端到端模型（如PilotNet）**由于缺乏可解释性和应对复杂场景的鲁棒性，在实际量产中较少单独使用
- 当前工业界更倾向于采用**“模块化+部分端到端”或“统一规划架构”**（如UniAD），在保证性能的同时提升可解释性与安全性
- **多模态融合（视觉+雷达+LiDAR）**与**Transformer架构**正在成为端到端自动驾驶模型的研究热点
- **仿真与真实数据结合、数据增强、安全校验机制**也是端到端模型落地的关键

一段式端到端

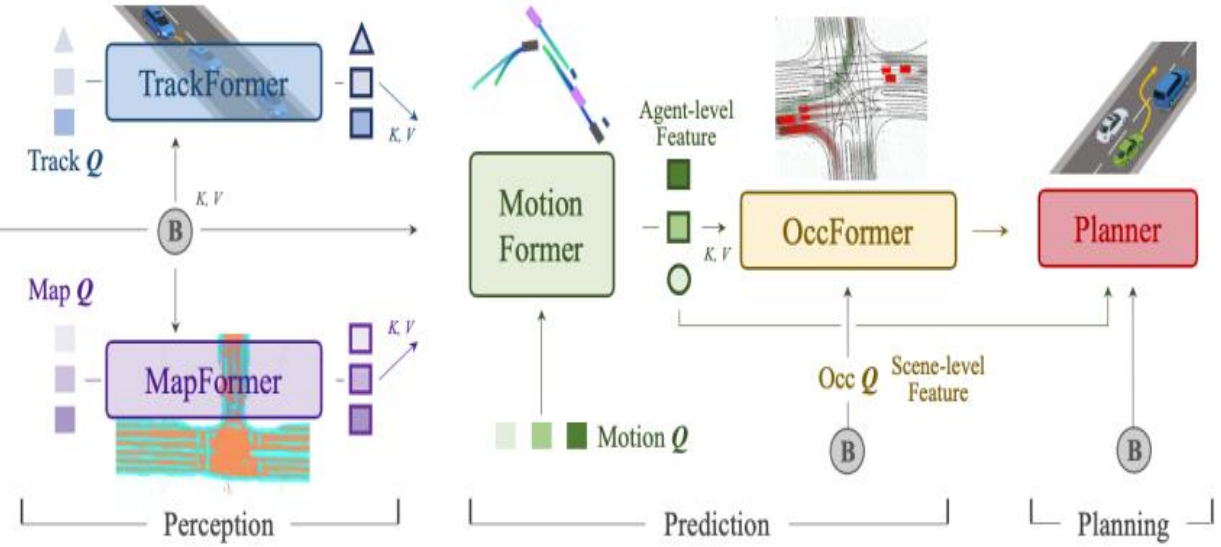
EMMA是以MLLM为核心的一段式架构



2025 EMMA: End-to-end multimodal model for autonomous driving

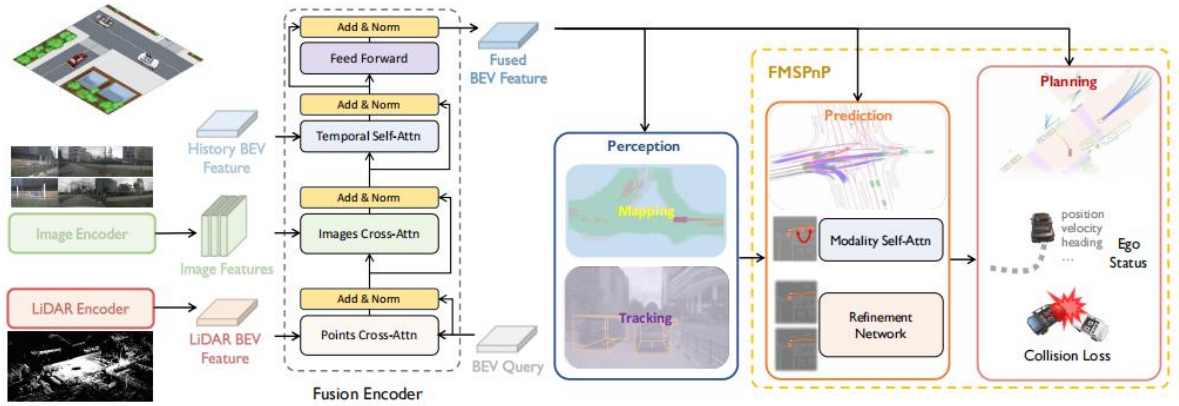
- **核心模型**EMMA：基于Gemini构建的端到端多模态模型，负责整合文本和视觉输入，进行多模态理解与推理。
- **推理与决策：**通过链式思考进行推理与决策，最终生成自动驾驶车辆的未来路径点，指导车辆行驶。
- EMMA 的核心理念是：将多模态大语言模型视为自动驾驶系统的核心，充分利用其两大优势：海量世界知识和强大推理能力
- 架构创新：首次系统性地将 MLLM 作为核心基础模型应用于端到端自动驾驶，提出了统一的视觉-语言框架。
- 通用性验证：证明了通用模型通过多任务协同训练可以超越单任务模型。
- 可解释性增强：通过思维链推理，使模型的决策过程更透明、更可信。

模块化端到端



CVPR 2023-Planning-oriented autonomous driving

- UniAD是以规划为导向的理念设计，利用了从驾驶场景中的前置节点到最终规划的联合优化优势。
- 所有感知和预测模块均采用Transformer解码器结构设计，以任务查询作为连接每个节点的接口。
- 该论文首次提出感知决策一体化的自动驾驶通用大模型UniAD，开创了以全局任务为目标的自动驾驶大模型架构先河，标志着自动驾驶技术的重要突破。

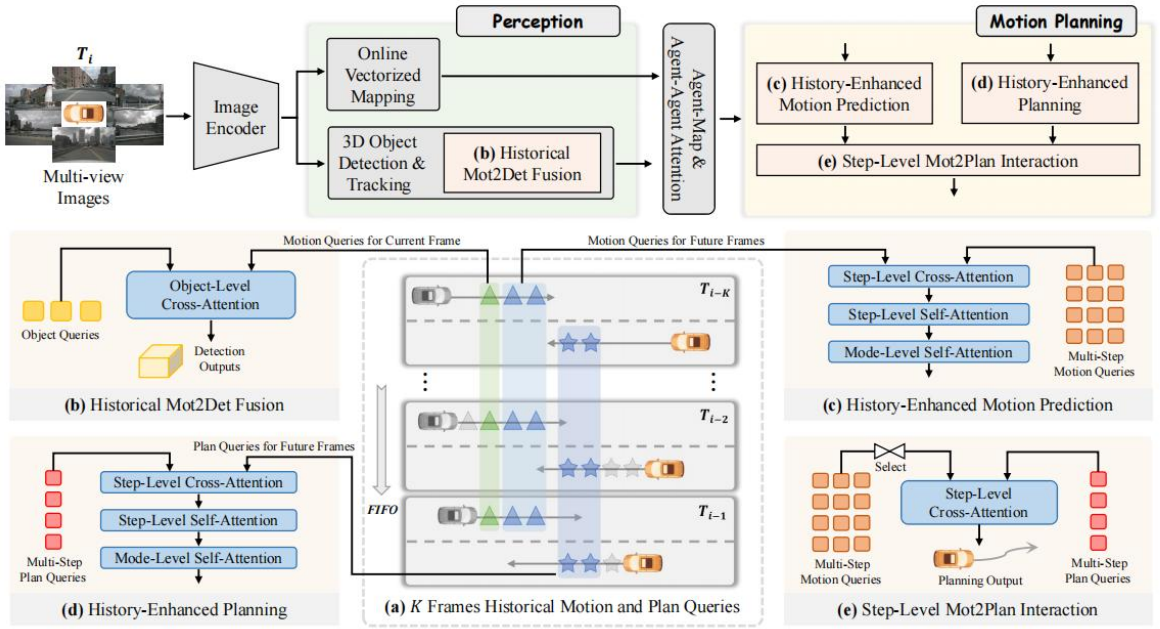


2023-Multi-modality fusion for prediction and planning tasks of autonomous driving

- FusionAD 的核心在于 **“融合”** 与 **“统一”** 。它通过一个精心设计的**多模态BEV融合编码器**，将摄像头和LiDAR的互补优势深度融合，生成强大的统一环境表征。并在此基础上，通过 **FMSPnP模块** 对预测和规划任务进行针对性优化，最终实现了各项任务的性能飞跃。
- 性能突破：它证明了将多模态融合从感知扩展至预测和规划，能带来巨大的性能增益，
- 为如何构建一个统一、高效的多模态多任务自动驾驶框架提供了具体的设计范例和实现细节

模块化端到端

BridgeAD 是一个典型的、先进的模块化端到端自动驾驶框架



Bridging past and future: End-to-end autonomous driving with historical prediction and planning

- **感知模块：**通过在线矢量化建图，从图像中构建环境地图。同时，模块进行3D目标检测与跟踪，并创新性地融合目标的历史运动数据，提升了检测的准确度与鲁棒性。
- **运动规划模块：**历史增强运动预测、历史增强规划和步级与运动规划交互，提升运动预测的一致性、准确性并优化规划结果。
- BridgeAD 的核心思想是：**通过设计多步查询，并对历史信息进行“分时复用”（当前帧用于感知，未来帧用于规划），从而在端到端自动驾驶框架中高效地桥接了过去与未来。**
- 指明方向：它揭示了更充分地利用历史信息是提升端到端驾驶系统连贯性和安全性的关键。
- 提供范式：其“多步查询”和“分模块融合”的设计为后续研究提供了一个新颖且有效的范式。

经典端到端算法

End-to-End Autonomous Driving: Challenges and Frontiers

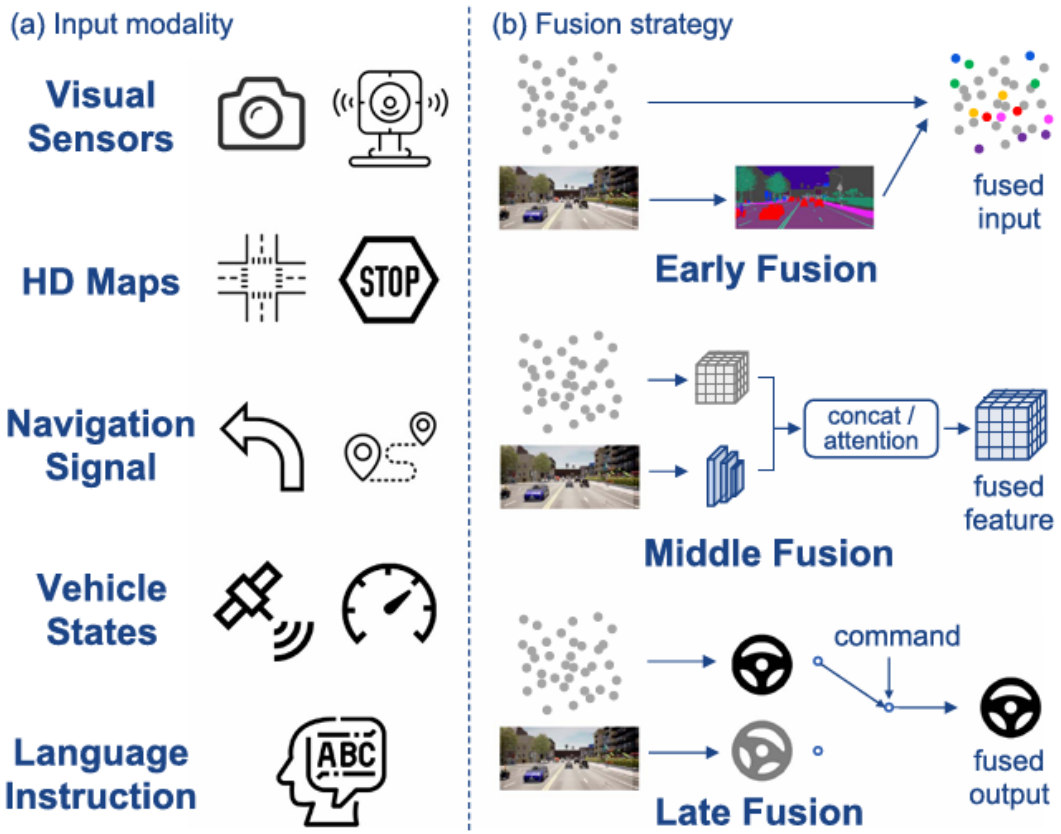
端到端自动驾驶方法的实现主要基于两种学习方法：模仿学习和强化学习。

端到端方法数据输入：视觉信息、高清地图、导航信号、自车状态、语言指令等。

现阶段输入多为多种数据融合输入。

端到端方法所遇到的挑战：关于感知和输入模态的困境、对视觉抽象的依赖、基于模型的强化学习世界建模的复杂性、对多任务学习的依赖、低效的专家和策略蒸馏、缺乏可解释性、缺乏安全保障、因果混淆、缺乏鲁棒性

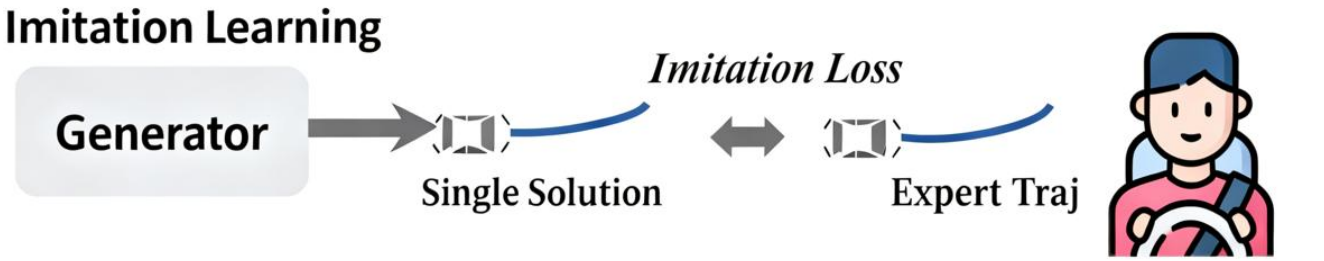
未来发展前景：零样本和小样本学习、模块化端到端规划、数据引擎（自动化处理数据，提高训练数据有效性）、基础模型（核心是 “一次预训练，多任务复用”）



端到端输入数据的不同融合时机

经典端到端算法--基于模仿学习

模仿学习（Imitation Learning）：
模仿学习是一种简单强大的方法，使用高质量的人类驾驶数据产生类似人类的行为。

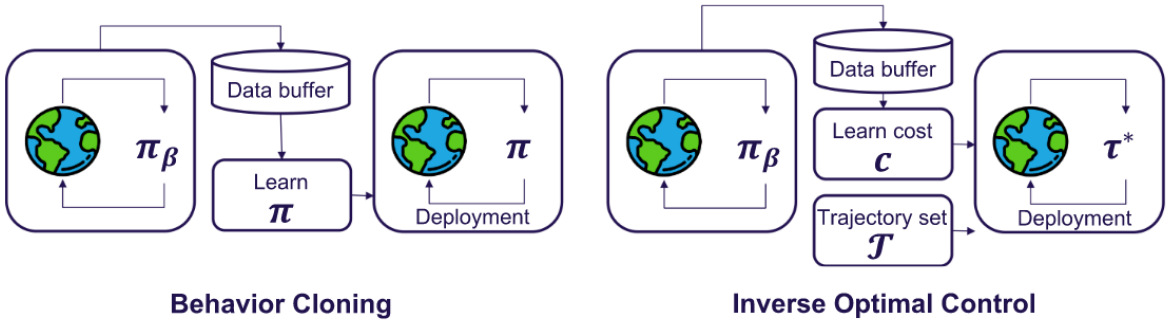


EvaDrive: Evolutionary Adversarial Policy Optimization for End-to-End Autonomous Driving

优点

- **安全可靠**：汲取人类经验，减少决策失误。
- **体验舒适**：行为自然拟人，提升驾乘感受。
- **泛化优势**：适应未知场景，优于规则系统。
- **成本低廉**：利用数据闭环，降低开发成本。

模仿学习

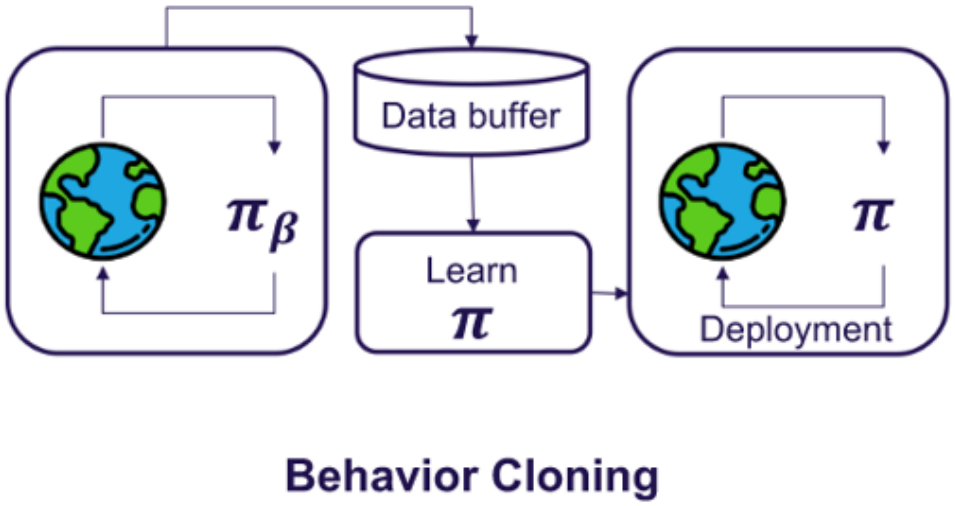


模仿学习主要分为两方面：行为克隆，逆最优控制（也叫作逆强化学习）。

行为克隆

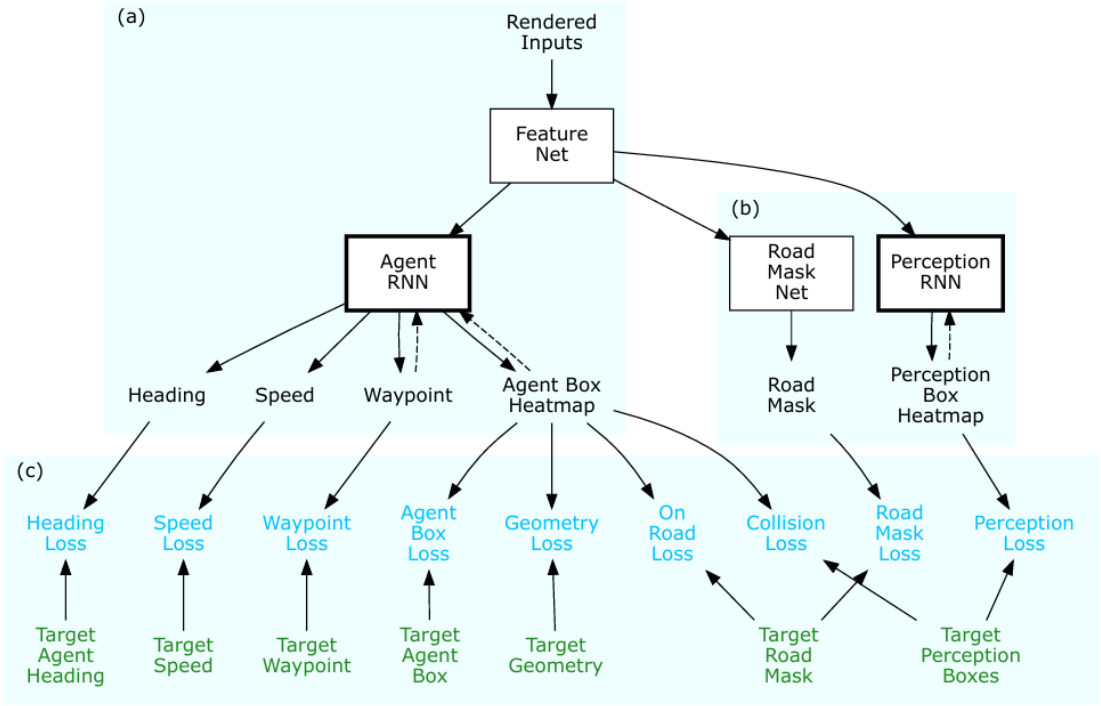
行为克隆是一种直接的学习方案。假设我们有许多专家的示例数据，它们以这样的形式出现：<s1,a1>，<s2,a2>，...，<sn,an>，

其中，si代表当前的环境，ai代表当前环境下专家采取的动作。



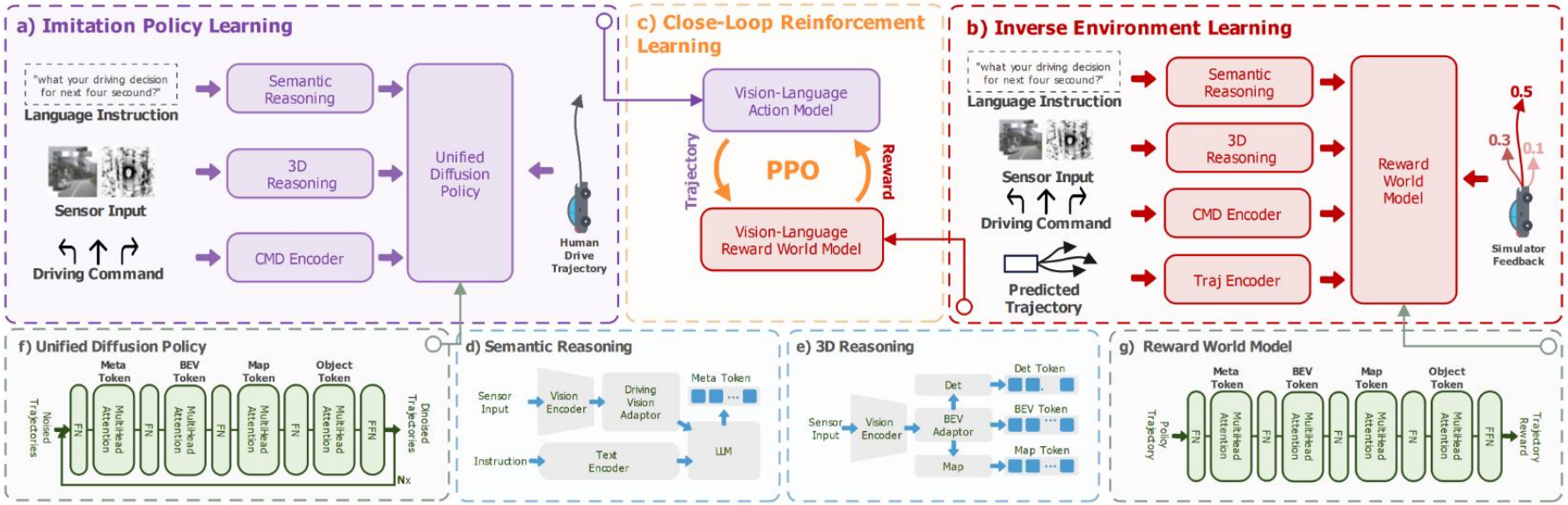
Behavior Cloning

提出了一个创新的自动驾驶模仿学习框架，其核心思想可以概括为：**"通过模仿专家驾驶行为的同时，主动合成最差驾驶场景（如碰撞、偏离道路等），并引入专门的环境损失函数来惩罚不良行为，从而显著提升模仿学习在自动驾驶中的鲁棒性和安全性。"**



2018-Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst

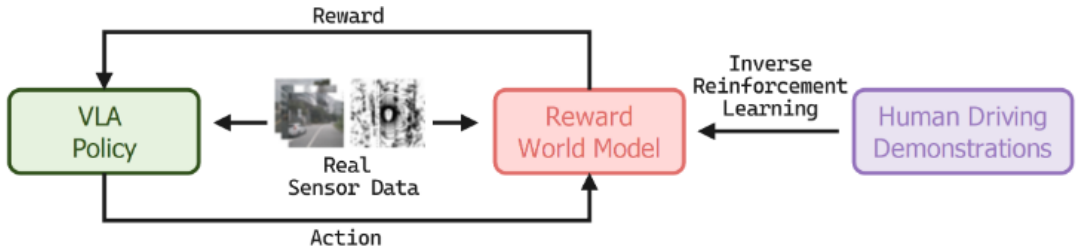
逆强化学习



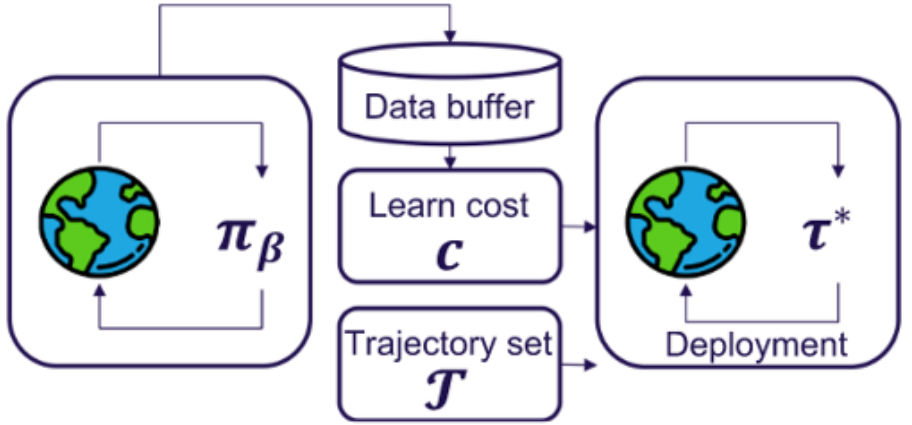
逆强化学习解决了行为克隆的核心痛点——它不直接模仿“动作”，而是先**学习人类驾驶的“潜在目标”**。

本质是“从示范中反推奖励函数，再用强化学习优化动作”。

IRL-VLA: Training an Vision-Language-Action Policy via Reward World Model



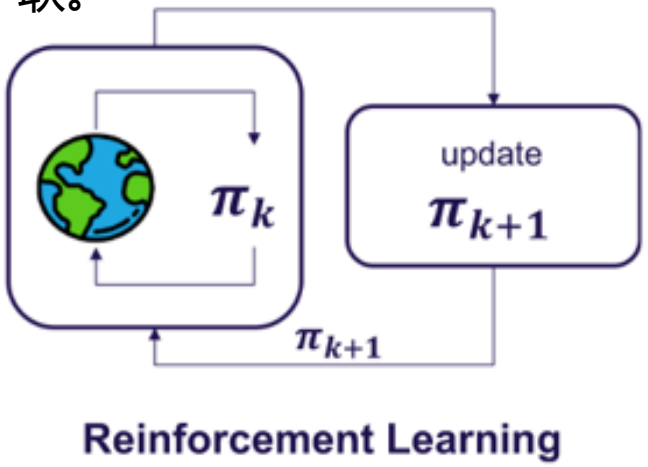
利用逆强化学习学习出专家行为的奖励函数，再利用强化学习进行调整策略。



Inverse Optimal Control

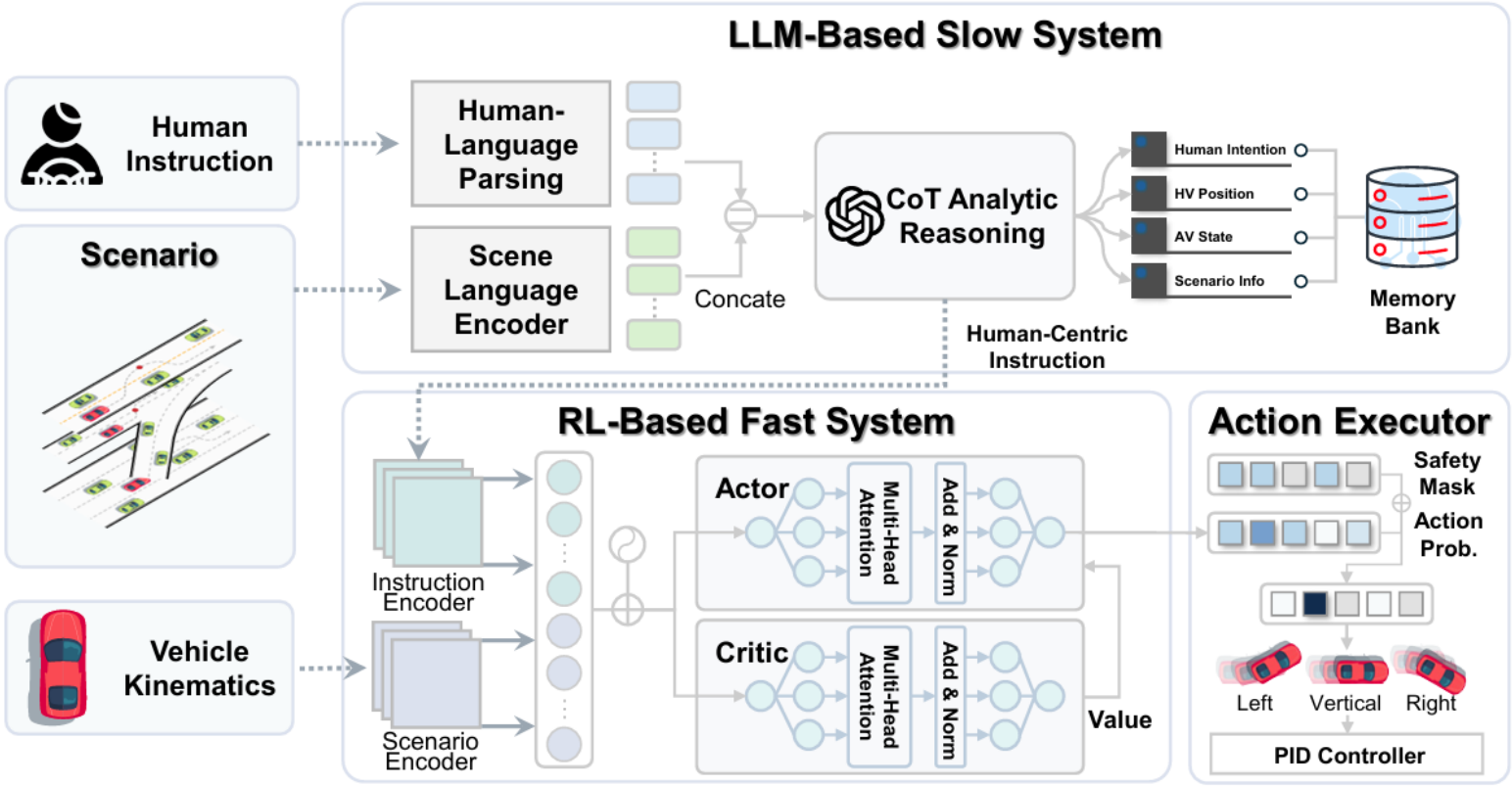
经典端到端算法--基于强化学习

模仿学习的短板在于缺乏环境奖励信号与机制，易受多重交互与因果倒置影响，难以梳理人类驾驶行为的因果逻辑与决策思路，不利于自动驾驶训练；
强化学习可通过设计多目标奖励函数，促使模型掌握动作与结果的因果联系，还能明确构建状态与动作的因果结构，剔除虚假关联。



Reinforcement Learning

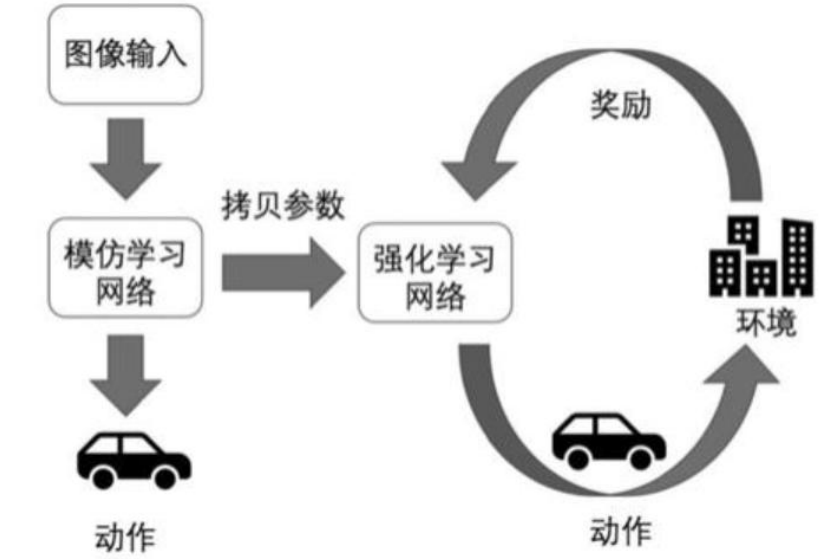
提出了一个创新的快慢双系统架构。**LLM作为“慢”系统**，负责将模糊的用户指令转化为结构化的指导信号。**而RL代理作为“快”系统**，负责在实时环境中做出快速且安全的决策。这种架构通过解耦高级决策与低级控制，实现了用户意图的精准理解与复杂交通环境中的灵活应对。



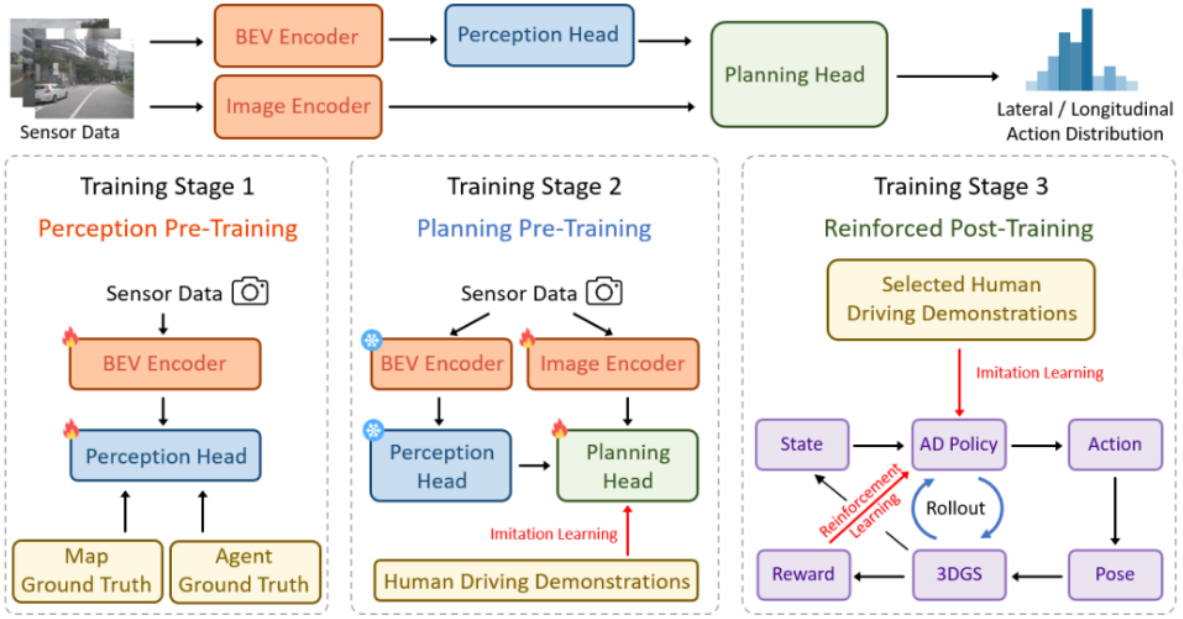
2025-Towards Human-Centric Autonomous Driving: A Fast-Slow Architecture Integrating Large Language Model Guidance with Reinforcement Learning

强化学习与模仿学习结合

结合模仿学习和强化学习，自动驾驶系统可以先通过模仿学习快速掌握基本驾驶技能，再通过强化学习进一步优化驾驶策略。



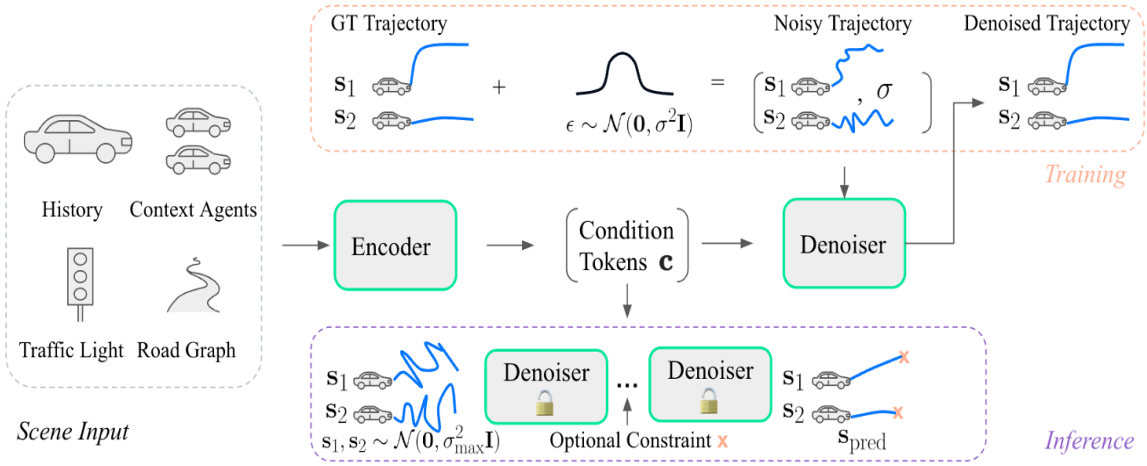
RAD采用三阶段训练范式：**感知预训练阶段**，以监督学习训练模型识别驾驶场景关键元素，构建环境认知；**规划预训练阶段**，依托大规模真实驾驶示范数据，借模仿学习初始化动作概率分布，规避强化学习冷启动问题；**强化后训练阶段**，通过强化学习与模仿学习协同，完成策略微调。



RAD: Training an end-to-end driving policy via large-scale 3dgs-based reinforcement learning

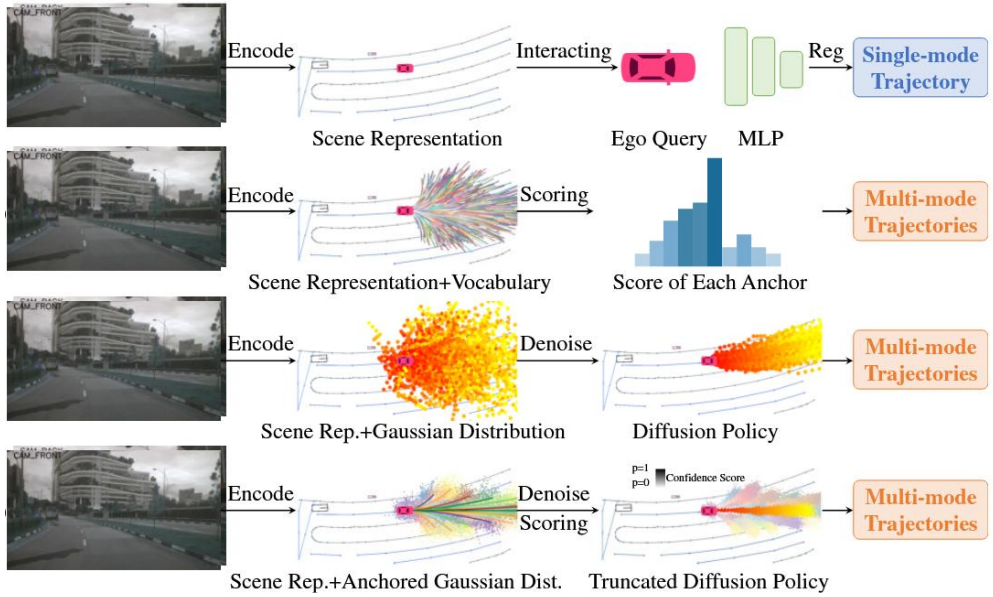
经典端到端算法--基于扩散模型

- 驾驶行为本质上是多模态的，而扩散模型具备强大的多模态建模能力，这使得扩散模型得以应用到自动驾驶领域。
- 在自动驾驶领域，扩散模型主要用于轨迹生成、训练数据合成、运动预测、不确定性估计与安全增强。



Motiondiffuser: Controllable multi-agent motion prediction using diffusion

轨迹生成



DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving

扩散模型主要分为两步实现，前向加噪和反向去噪：

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$$

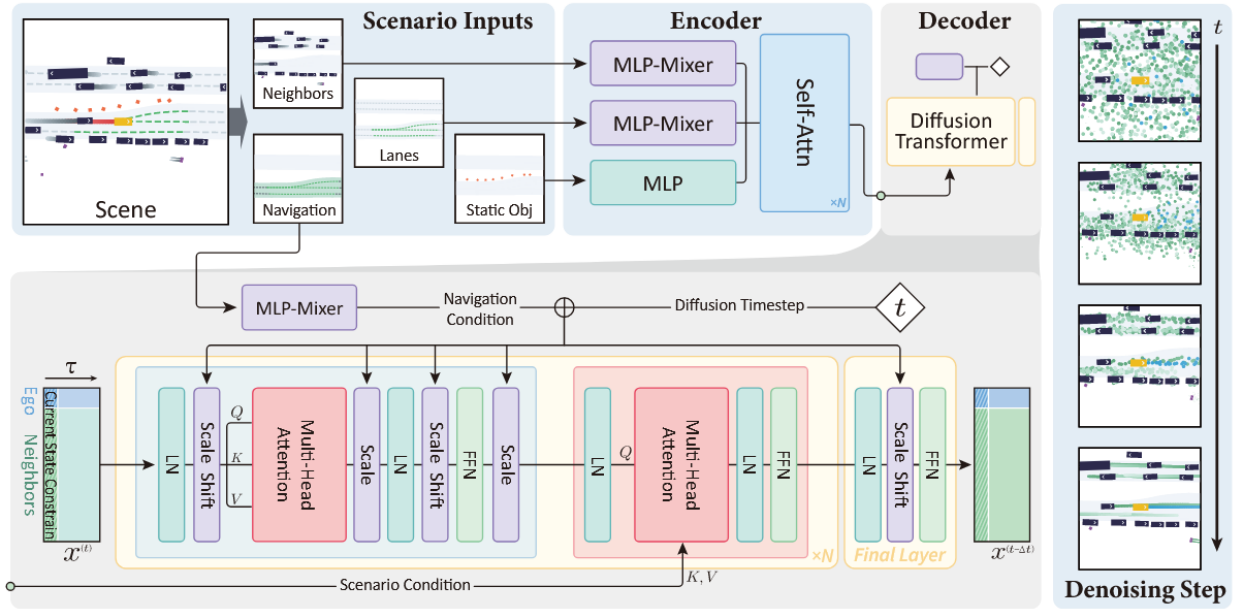
前向加噪

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

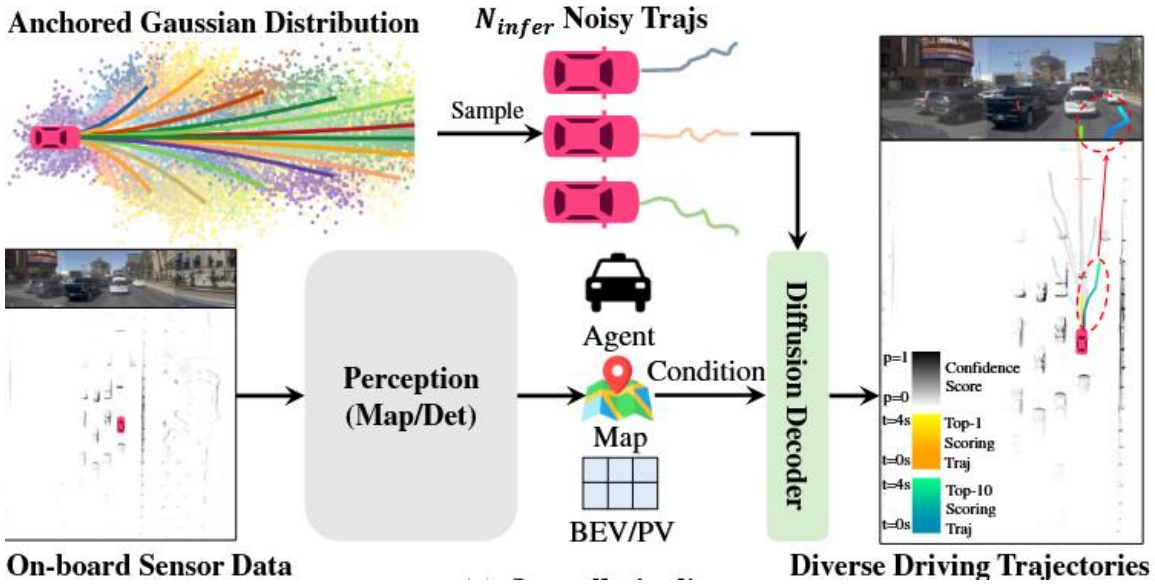
反向去噪

经典端到端算法--基于扩散模型

Diffusion-Planner充分利用扩散模型的表达能力和灵活的引导机制来进行高质量的自主规划。引入基于Transformer的架构，通过扩散目标对运动预测和规划任务中的多模态数据分布进行联合建模。分类器指导用于将规划行为与安全或用户偏好的驾驶风格保持一致。



diffusion-based planning for autonomous driving with flexible guidance



DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving

DiffusionDrive核心思想是提出一种“截断扩散策略”，将生成式扩散模型高效且有效地应用于端到端自动驾驶的实时规划中，解决了传统扩散模型在此领域面临的计算开销大和模式坍塌问题。

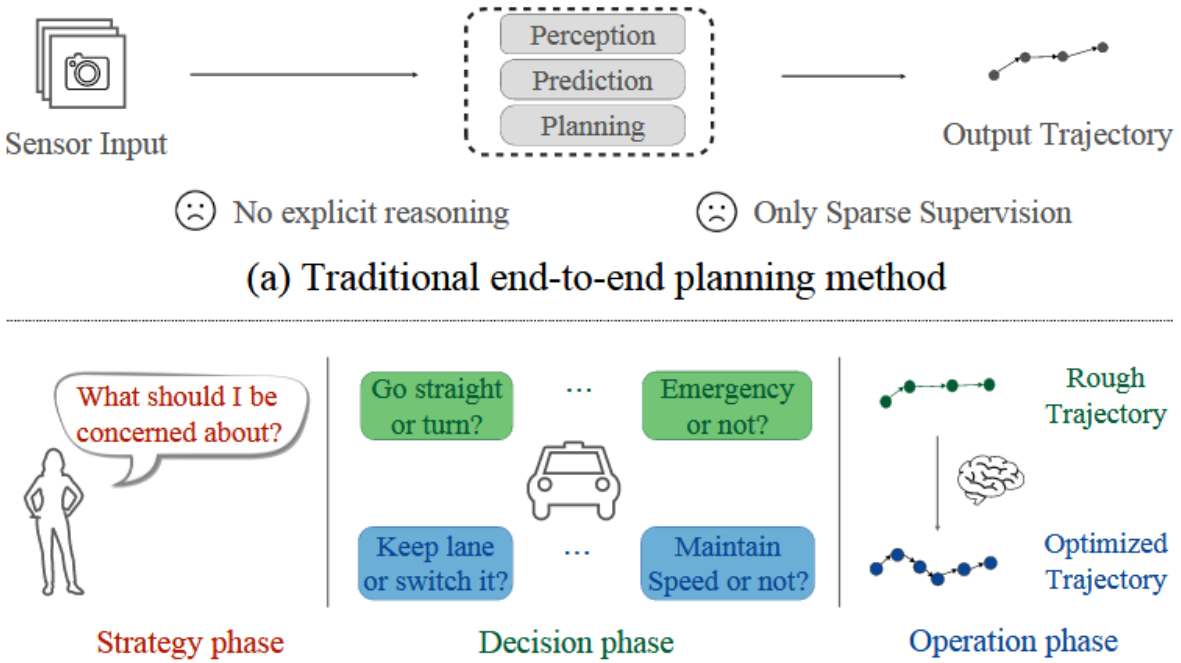
端到端--最新相关工作

ReAL-AD: Towards Human-Like Reasoning in End-to-End Autonomous Driving（ICCV2025）

现有的端到端方法往往依赖于固定和稀疏的轨迹监督，限制了它们捕获人类驾驶员自然使用的分层推理过程的能力。

主要贡献：

- 提出了一种新的推理增强的端到端自动驾驶框架ReAL-AD，该框架明确地纳入了分层决策，并将轨迹规划与人类的认知过程联系起来；
- 引入了用于VLM驱动决策集成的战略推理注入器、用于结构化控制的战术推理集成器和用于分层轨迹细化的分层轨迹解码器，以确保从推理到执行的一致性。

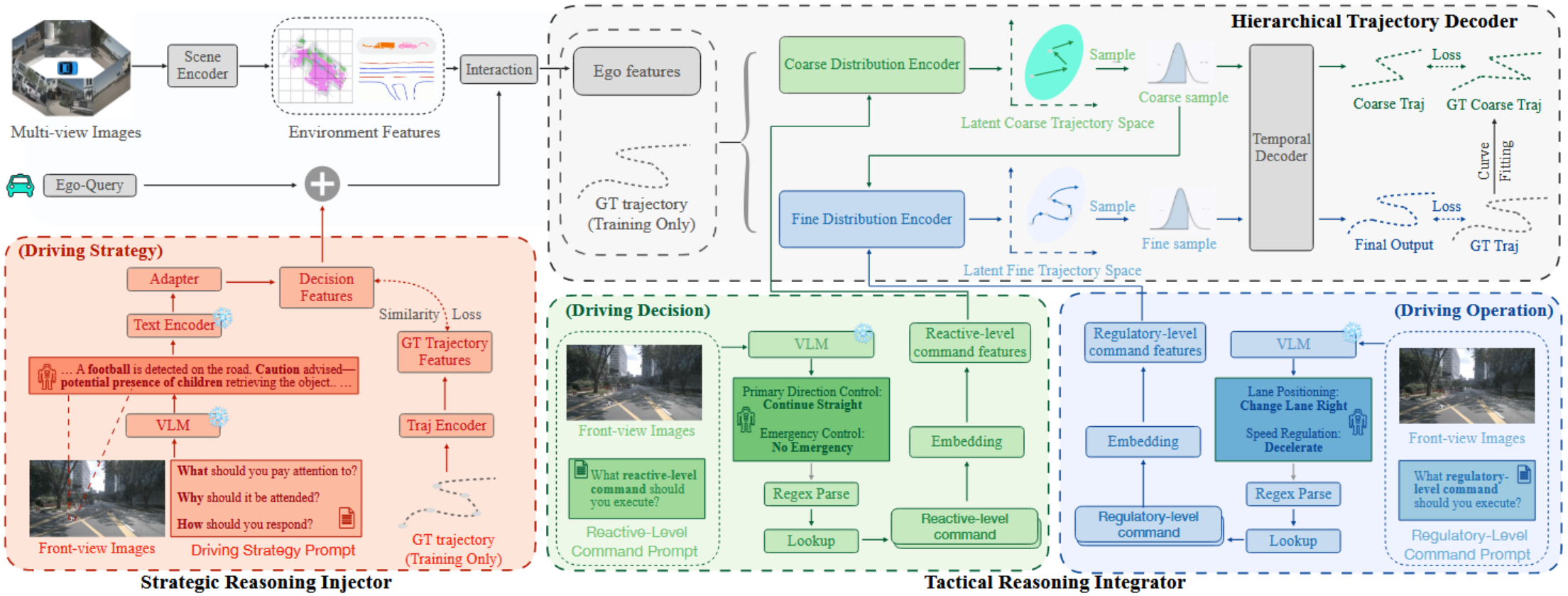


端到端网络与人类驾驶逻辑的比较

（a）描述了端到端自动驾驶系统的工作流程和局限性，（b）描述了人类驾驶员的结构化决策过程。

端到端--最新相关工作

ReAL-AD: Towards Human-Like Reasoning in End-to-End Autonomous Driving（ICCV2025）



这是一种推理增强学习框架，基于三层人类认知模型（驾驶策略、驾驶决策和驾驶操作）来构建自动驾驶中的决策过程，并引入视觉-语言模型（VLMs）以增强环境感知和结构化推理能力。

端到端--最新相关工作

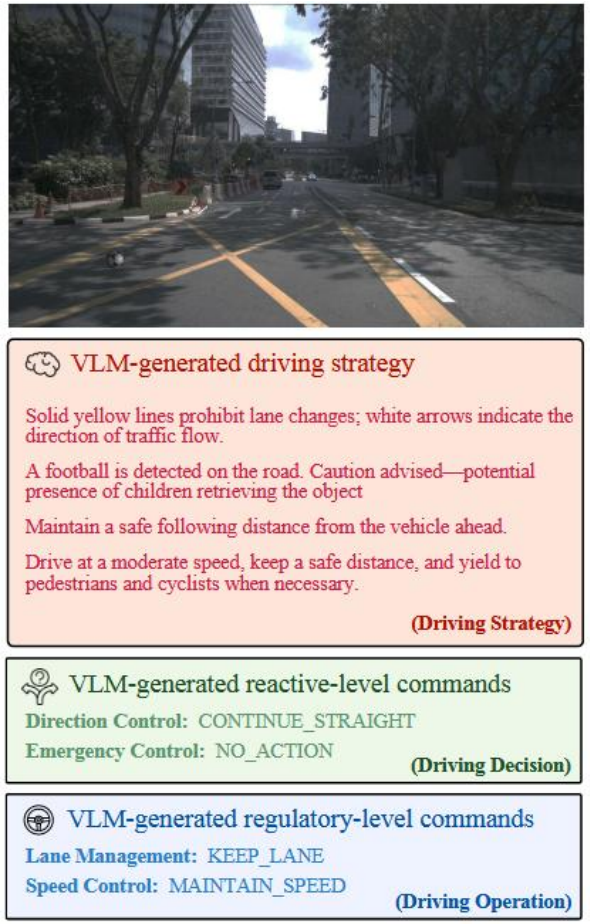
ReAL-AD: Towards Human-Like Reasoning in End-to-End Autonomous Driving（ICCV2025）

该框架主要分为三个模块分别对应（驾驶策略、驾驶决策和驾驶操作）：

Strategic Reasoning Injector：利用 VLM 生成的洞察解析复杂交通情境，形成高层驾驶策略；

Tactical Reasoning Integrator：将策略意图进一步细化为可解释的战术选择；该模块强制输出四类指令（方向/紧急/车道/速度），其中方向和紧急属于反应级别驾驶决策，车道和速度属于经过思考的驾驶操作命令，类似于人的直觉与理性。

Hierarchical Trajectory Decoder：模拟人类“直觉—细化”的决策过程，先建立粗略运动模式，再逐步精修为详细轨迹。

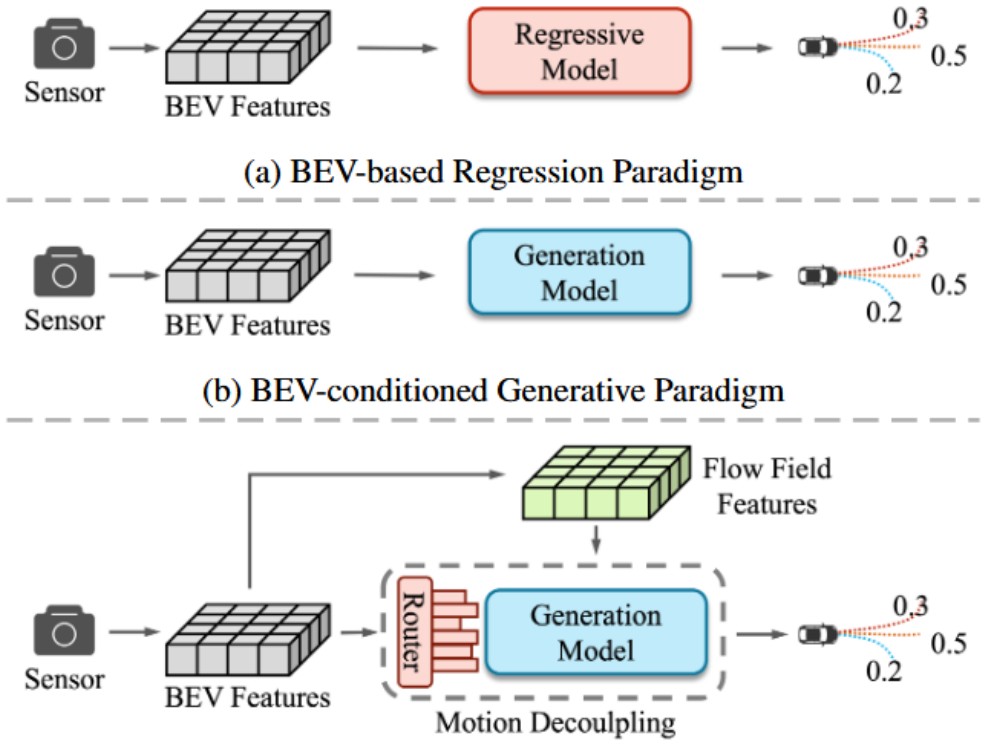


实例VLM输出

端到端--最新相关工作

FlowDrive: Energy Flow Field for End-to-End Autonomous Driving（arxiv2025）

- 自动驾驶需同时应对障碍物硬约束与规则类软语义，但现有端到端框架依赖隐式 BEV 特征，缺乏风险与指导先验的显式建模，难以保障规划安全可解释，提出 FlowDrive 框架。
- FlowDrive 通过引入**物理可解释的能量流场**，为端到端自动驾驶规划提供了显式的安全与语义先验。其**“流场学习 → 锚点优化 → 解耦生成”**的框架设计，巧妙地解决了现有方法中隐含学习不透明和任务耦合冲突的两大痛点。
- **可解释性**：能量场使模型的决策过程变得可视、可理解，有助于调试和建立对自动驾驶系统的信任。
- **性能提升**：显式的先验指导与解耦的架构设计共同作用，带来了规划性能的显著提升。
- **新范式**：它为端到端自动驾驶领域提供了一种新的思路，即如何将传统的、具有明确物理意义的控制理论（如人工势场法）与现代的、数据驱动的深度学习方法进行融合创新。

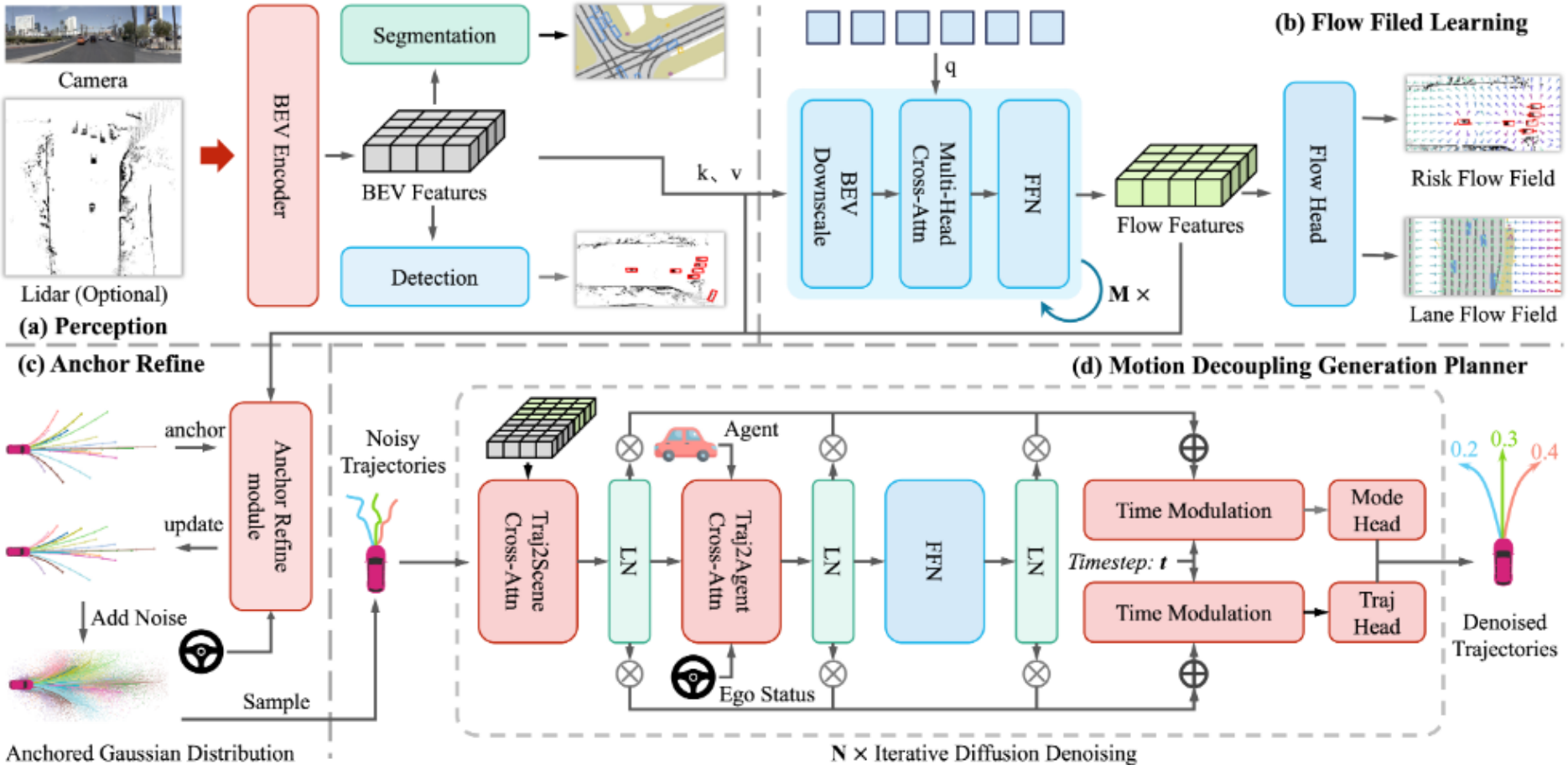


不同端到端范式的比较

（a）直接从BEV表示中预测轨迹的基于回归的范式。（b）基于BEV特征采样轨迹的生成范式。（c）提出了基于BEV特征和流场特征的解耦生成范式。

端到端--最新相关工作

FlowDrive: Energy Flow Field for End-to-End Autonomous Driving（arxiv2025）

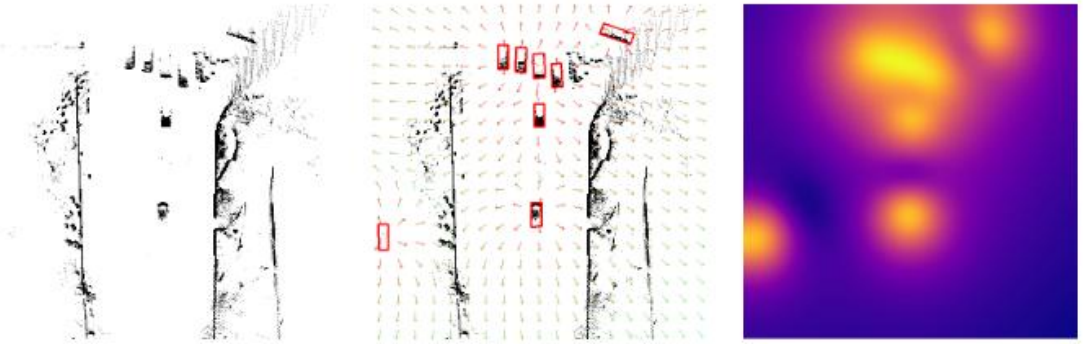


可视化危险区域和重要区域（车道线等），使模型更具有解释性

端到端--最新相关工作

FlowDrive: Energy Flow Field for End-to-End Autonomous Driving（arxiv2025）

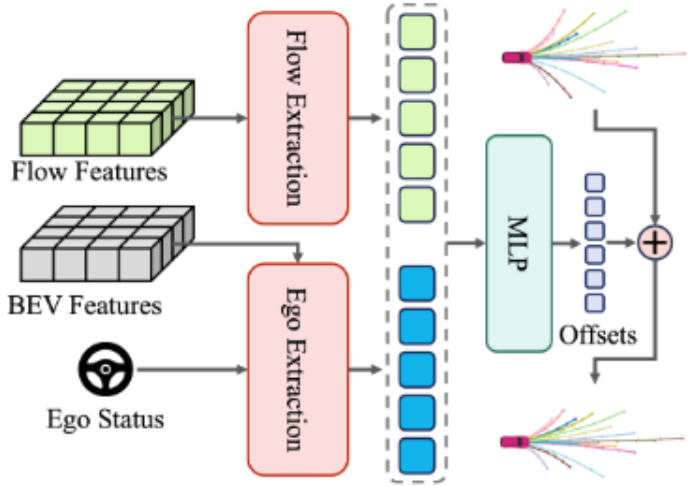
流场学习：引入风险势场和车道吸引力场，编码安全和语义先验。通过对 BEV 特征下采样，与可学习查询交互提取流特征，解码得到能量图。



(a) Lidar BEV　(b) Flow Field　(c) Flow Field Feature

流感知锚定轨迹优化：为提升规划轨迹与底层驾驶场景的空间对齐度，提出流感知锚定轨迹优化模块，以场景自适应的方式调整初始锚定点。

运动解耦生成规划器：FlowDrive 采用条件扩散框架对多模态轨迹分布进行建模。为缓解模式预测和轨迹生成之间的梯度干扰和表示纠缠问题，引入特征级解耦策略，将两者的学习目标分离，针对不同任务进行有针对性的特征提取。



流感知锚点优化模块

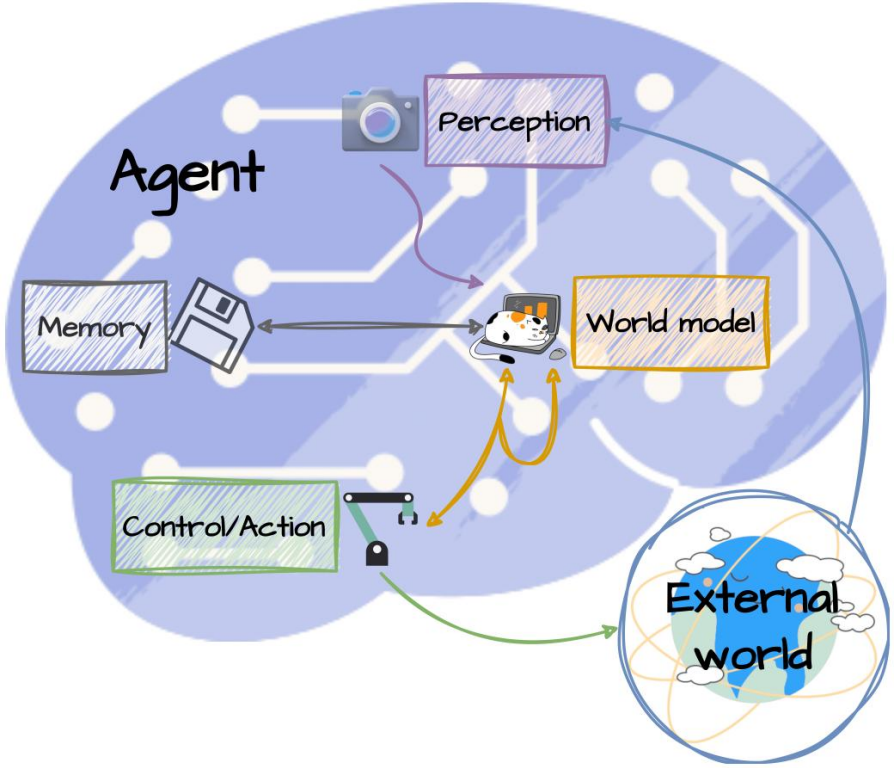
世界模型(WM)

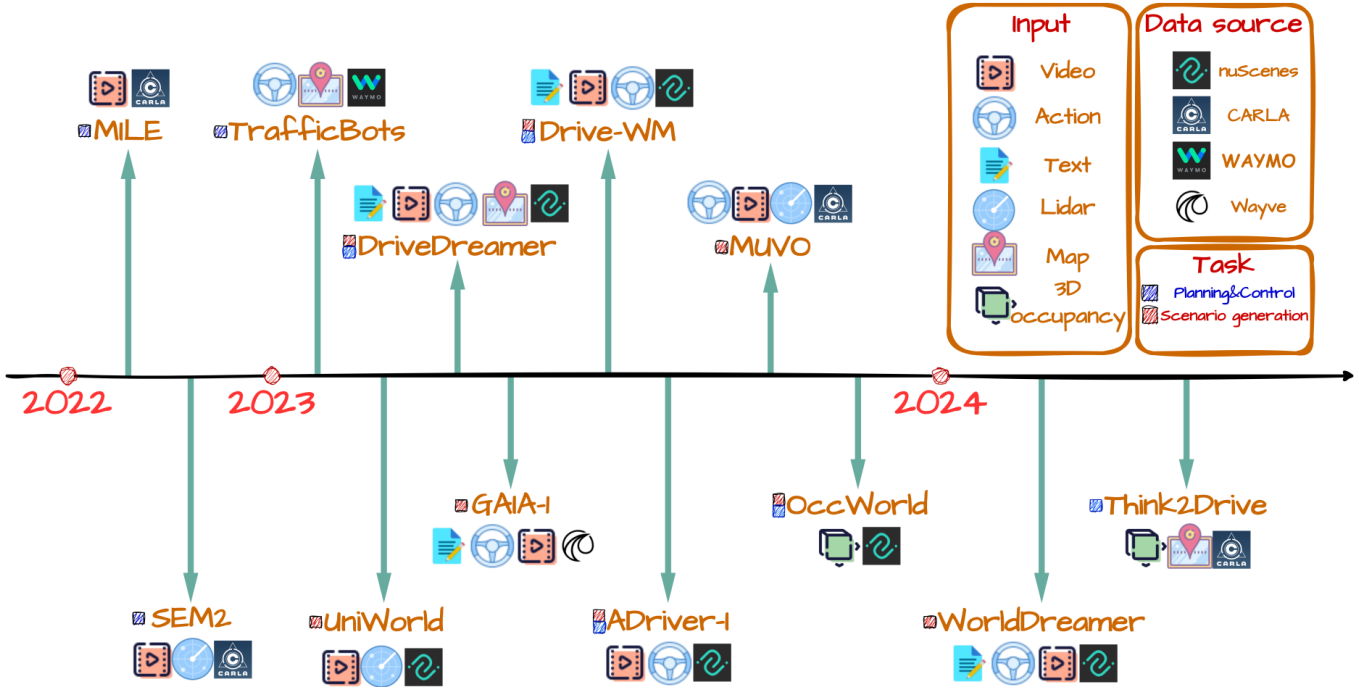
世界模型 (WM)

世界模型旨在让自动驾驶系统学会像一个“物理学家”和“心理学家”一样去理解和推演世界。它不仅仅是感知“现在是什么”，更是要构建一个对环境如何运作的内部模拟器，从而能够预测未来可能发生什么。

自动驾驶要实现 “安全落地”，**核心是让车辆像人一样“理解环境、预判未来、规划动作”**—— 而世界模型正是这一能力的核心载体。

世界模型即利用历史场景观测信息加上预设条件预测未来智能驾驶场景变化和自车响应的模型。世界模型的核心任务有两个，一个是预测未来场景，二是动作规划。





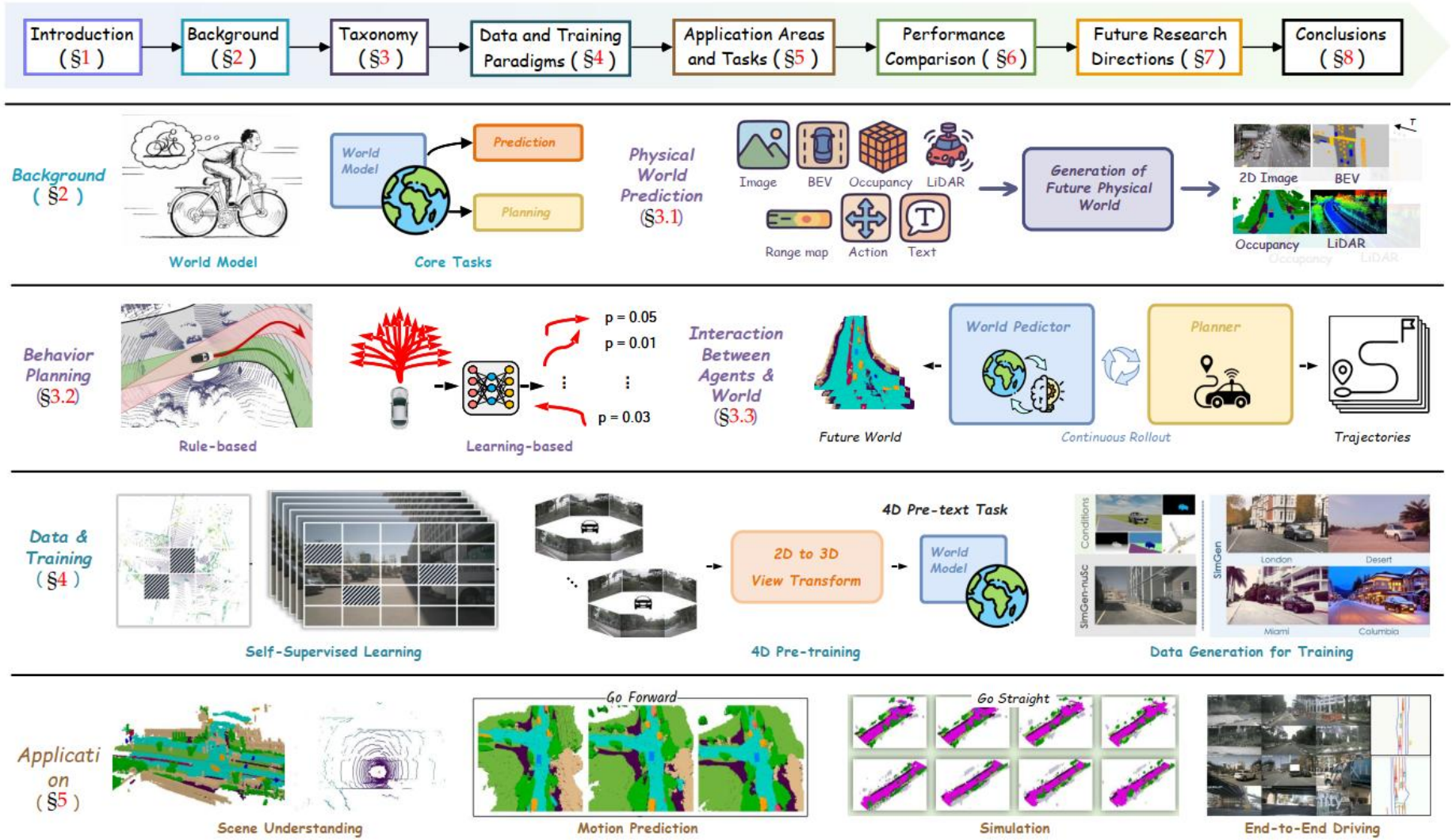
世界模型 (WM)

作用	具体描述
1. 环境建模（Understanding）	从传感器数据中学习世界的结构与状态，比如车道、车辆、行人、信号灯等，并转换为可用的内部表征。
2. 未来预测（Prediction）	预测其他交通参与者的未来轨迹、行为和环境变化，如前车刹车、行人横穿马路。
3. 可模拟驾驶（Simulation）	允许模型在内部“脑中模拟”多种驾驶结果，类似人类的“心理预演”。比如尝试变道是否安全。
4. 支撑规划与决策（Planning）	规划模块可使用世界模型提供的未来场景进行路径选择、速度优化、避障等决策。
5. 数据高效学习（Sample Efficiency）	强化学习或模仿学习能在世界模型中离线训练策略，减少真实道路测试需求，降低成本与风险。

世界模型 (WM)

近年来，自动驾驶世界模型技术百花齐放，但也面临 “方法碎片化、任务无共识、落地难衔接” 的困境：有的模型擅长静态场景重建，却搞不定动态车辆预测；有的聚焦单一传感器，多模态融合时漏洞百出。

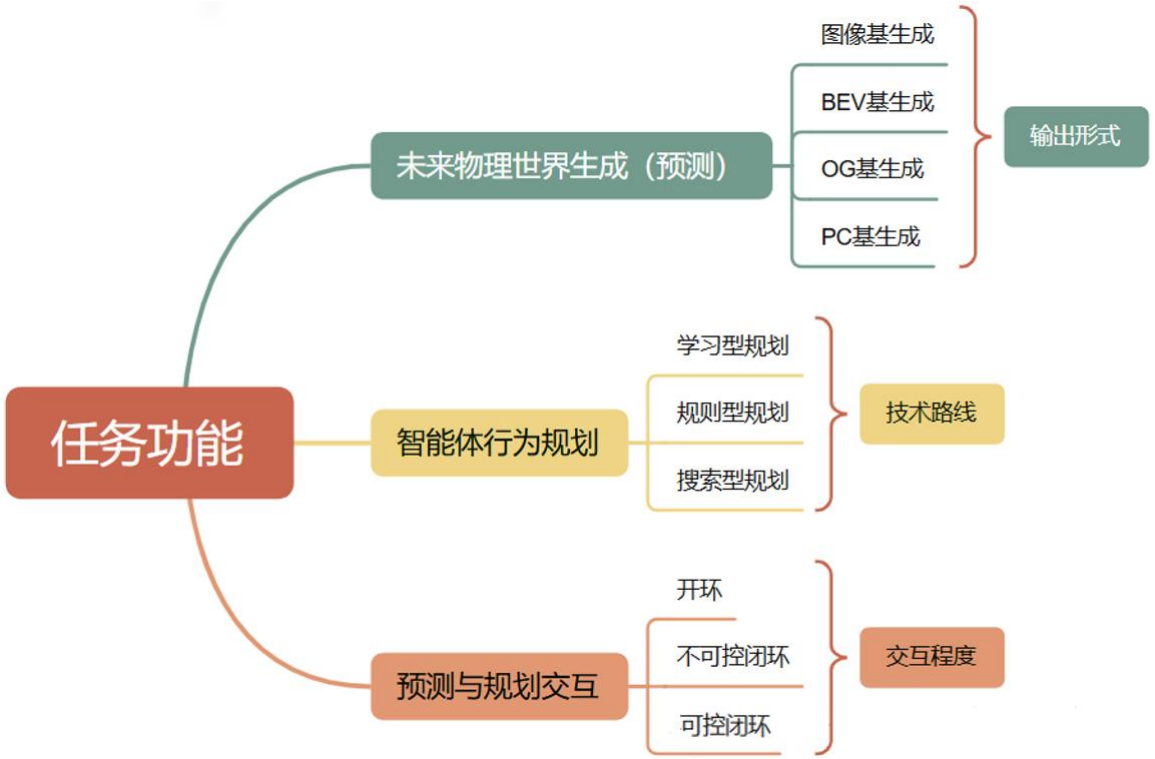
浙江大学 CCAI 团队的综述论文《A Survey of World Models for Autonomous Driving》首次提出 “三维分类框架”，系统梳理 190 + 篇前沿文献，覆盖数据训练、模型设计、落地应用全链条，提供了一份 “从理论到落地” 的完整指南。



综述介绍了世界模型的关键组成部分——未来物理世界的生成、智能代理的行为规划以及它们之间的迭代；强调了在自动驾驶中训练模型的各种方法，包括自监督学习范式、预训练策略和数据生成的创新方法；展示了自动驾驶中世界模型应用的四个领域：场景理解、运动预测、仿真和端到端驾驶。

世界模型 (WM)

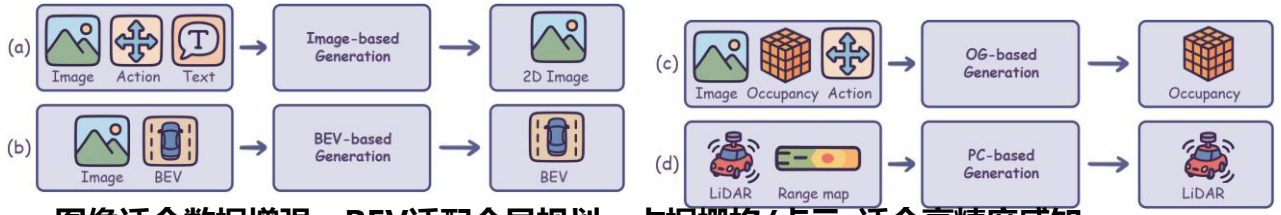
“三维分类框架” -- （维度一）任务功能维度



世界模型三大任务功能

让车辆“看见未来”

预测环境未来状态（动态物体运动、静态场景变化），是世界模型的“感知基础”。



图像适合数据增强，BEV适配全局规划，占据栅格/点云 适合高精度感知。

让车辆“选对动作”

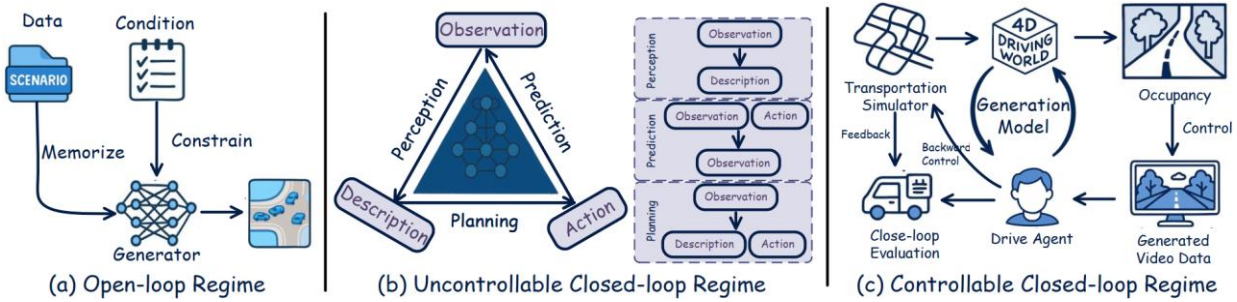
为车辆生成安全、高效的轨迹，是世界模型的 “决策核心”。

- 学习型规划：用RL/Transformer学决策逻辑，如通过“视觉思维链（CoT）” 生成轨迹；融合世界模型预训练；
- 规则型规划：基于物理规则/手工成本函数，如IDM模型适合跟车场景，RRT*擅长避障路径搜索，优点是“安全可解释”，缺点是应对复杂路况灵活度低；
- 搜索型规划：将状态空间离散为图，用A*/Dijkstra算法找最优路径，如Hybrid-State A*考虑车辆动力学约束，适合“高动态场景”（如路口转弯）

学习型适合复杂城区，规则型适合高速简单场景，实际落地常“学习+规则”融合。

让车辆“越用越聪明”

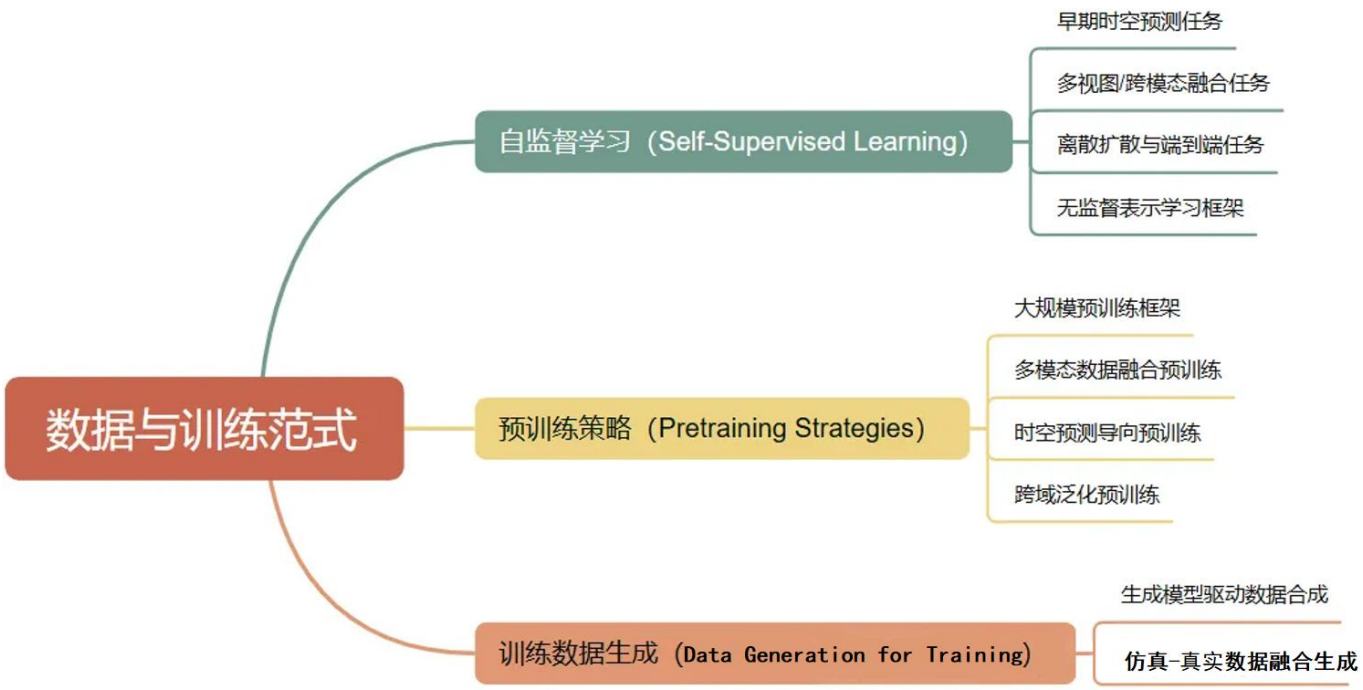
打破 “预测和规划脱节” 的瓶颈，形成 “预测指导规划、规划反哺预测” 的闭环，是世界模型落地的关键。



当前正从“不可控闭环”向“可控闭环”升级，这是解决“极端场景泛化”的核心方向。

世界模型 (WM)

“三维分类框架”-- （维度二）训练范式维度



世界模型三大训练范式

1.自监督学习

- 早期时空预测任务：如UnO/EO，早期基础方法，通过4D占用或LiDAR对比学习环境动态，降低标注依赖；
- 多视图/跨模态融合任务：如RenderWorld/ViDAR，融合2D-3D数据，用自监督信号提升多模态场景理解；
- 离散扩散与端到端任务：如COPILOT4D/SSR，针对无标注数据或端到端任务，用离散扩散、稀疏表示简化训练；
- 高效无监督框架，专注BEV嵌入预测，无生成/对比机制，预训练速度快5倍，兼顾速度与环境细节捕捉。

2.大规模预训练

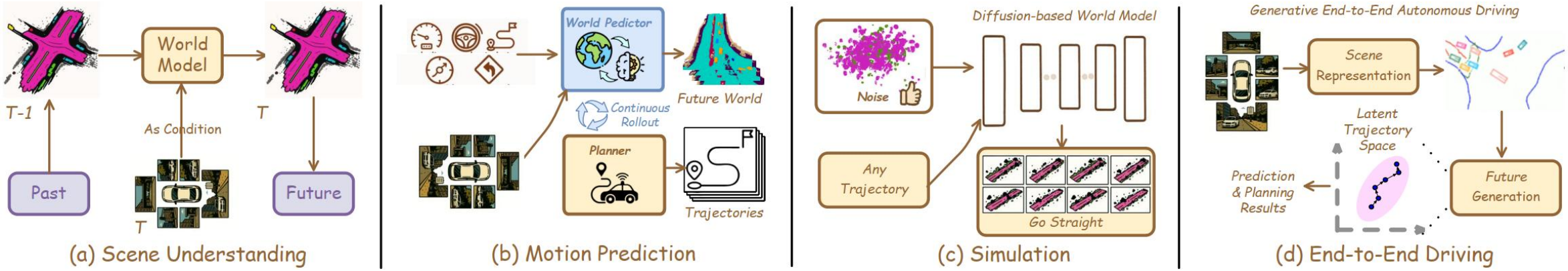
- 多模态融合（ViDAR++/UniPAD）：打破传感器模态壁垒，提升多源数据适配性；
- 时空预测（UniWorld/DriveWorld）：以“预测未来”为目标，强化动态场景建模能力；
- 跨域泛化（BEVWorld）：统一latent空间，解决“仿真训练、真实部署”的域适应问题。

3.生成式数据增强

- 生成模型驱动（OccSora/InfinityDrive 等）：生成高保真、长时序或可控场景，覆盖长尾场景（如极端天气、复杂交互）；
- 仿真-真实融合（SimGen/DrivingDojo）：弥补纯仿真“不真实”缺陷，生成贴近真实路况的数据，提升模型鲁棒性。

世界模型 (WM)

“三维分类框架” -- （第三维度）应用场景维度



1. 场景理解：环境感知的基础

提供更丰富、更准确的3D环境表征

•核心需求：重建3D环境、识别静态/动态目标。

•适配方法：OG/PC生成（如OccWorld提升占用预测精度）+自监督学习（如ViDAR对齐多模态数据）；

•代表案例：ViDAR融合相机与LiDAR，实现复杂道路的语义分割。

2. 运动预测：动态交互的关键

预测周围交通参与者的未来状态

•核心需求：预测其他智能体（车辆、行人）的未来轨迹。

•适配方法：BEV生成（如FIERY预测多模态实例轨迹）+可控闭环交互（如 TrafficBots模拟多智能体行为）；

•代表案例：OccWorld用3D占用表示预测动态演化，无需大量标注。

3. 仿真：安全测试的核心

构建高保真的虚拟环境，用于算法测试和验证

•核心需求：生成高保真场景，替代部分真实路测。

•适配方法：图像/OG生成（如OccSora生成 16秒时序场景）+训练数据生成（如 SimGen缩小仿真-真实差距）；

代表案例：OccSora的4D占用仿真，可测试 “车辆遮挡+行人横穿”的极端场景。

4. 端到端驾驶：决策闭环的终极形态

作为核心组件，连接感知、预测和规划，实现一体化决策

•核心需求：直接映射传感器输入到驾驶动作。

•适配方法：学习型规划（如LAW预测latent特征优化动作）+自监督学习（如SSR用稀疏表示减少监督需求）；

•代表案例：LAW的端到端框架，L2误差平均0.61m，碰撞率0.30%

世界模型 (WM)

未来展望：四大路径，突破落地瓶颈

要实现世界模型在复杂城市环境的落地，需聚焦四大前沿方向：

1.自监督世界模型：降低标注依赖 现有自监督方法（如SelfOcc）仍比全监督模型mIoU低10%，未来需结合“跨模态重建 +物理损失”（如图像生成LiDAR，LiDAR验证几何），同时探索无标签4D占用预测，进一步降低标注成本。

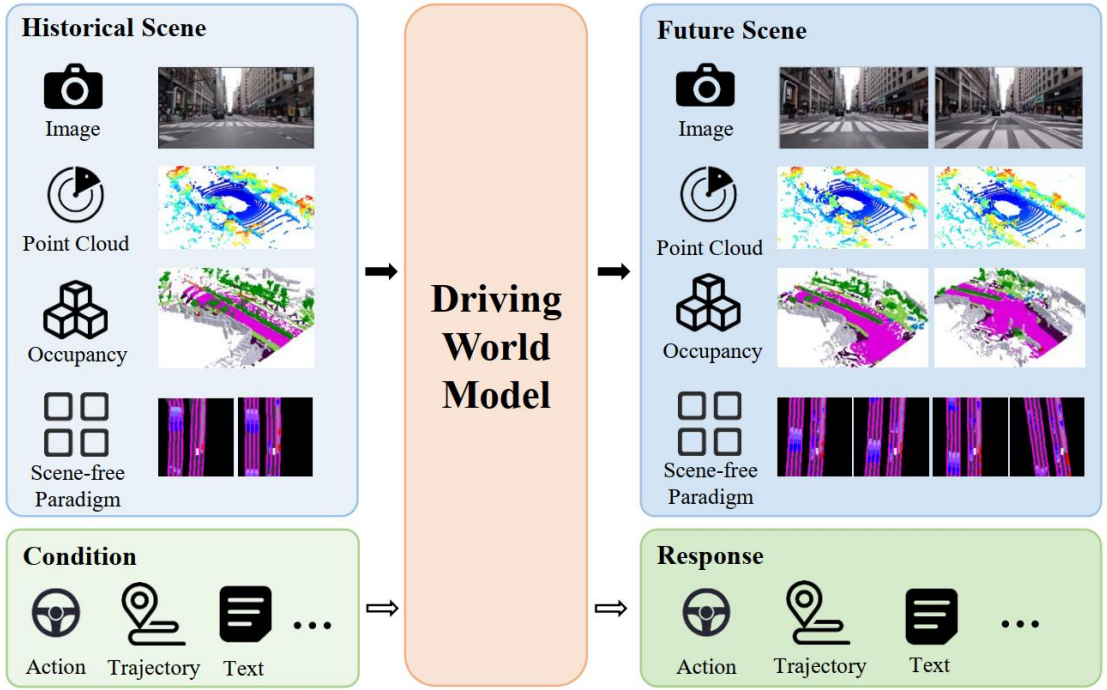
2.多模态融合：统一传感器输入 当前多模态方法需人工设计传感器适配器（如相机→BEV、LiDAR→BEV），未来需构建 “通用模态嵌入”，自动对齐异步传感器（如雷达延迟、相机曝光差异），同时融入语言指令（如“避让行人”），提升可控性。

3.先进仿真：提升物理真实性 现有仿真（如CARLA）在“物理细节”（如轮胎摩擦、雨天反光）上与真实场景差距大，未来需融合“扩散生成+可微物理求解器”（如扩散模型生成4D占用，物理引擎保证牛顿定律），同时支持文本注入罕见场景（如“突发行人”）。

4.高效架构：适配车载硬件 当前世界模型（如GAIA-1）推理耗时达数百毫秒，无法满足AD实时性（需<100ms），未来需通过“稀疏 tokenization”（如仅关注动态物体）、“边缘量化”（车载GPU适配）、“联邦学习”（fleet数据更新），平衡性能与效率。

世界模型 (WM)

华中科大与百度联合撰写的综述《The Role of World Models in Shaping Autonomous Driving: A Comprehensive Survey》按 “预测模态”也提出了一种“三维分类体系”：2D 场景演化（GAIA-1/DriveDreamer等）、3D 场景演化（OccWorld/Copilot4D等）、无场景范式（Think2Drive/TrafficBots等）。



范式类型	核心目标	代表工作	解决的自动驾驶痛点
2D 场景演化	生成 photoreal 2D 视频，保证时空一致性	GAIA-1（扩散 decoder）、DriveDreamer（多模态控制）、Vista（高保真仿真）	传统 2D 数据少、场景单一，补充雨天 / 夜间等长尾视频
3D 场景演化	预测 3D 几何结构（占用 / 点云），保留空间关系	OccWorld（3D 占用 token 化）、Copilot4D（点云离散扩散）、GaussianWorld（高斯建模）	纯 2D 无法感知深度，解决 “近处障碍物漏检” 问题
无场景范式	不预测细节场景，聚焦 latent 状态 / 多智能体行为	Think2Drive（latent 空间 RL）、TrafficBots（智能体人格建模）	实时决策需快速响应，避免 3D 计算带来的 latency

世界模型 (WM)

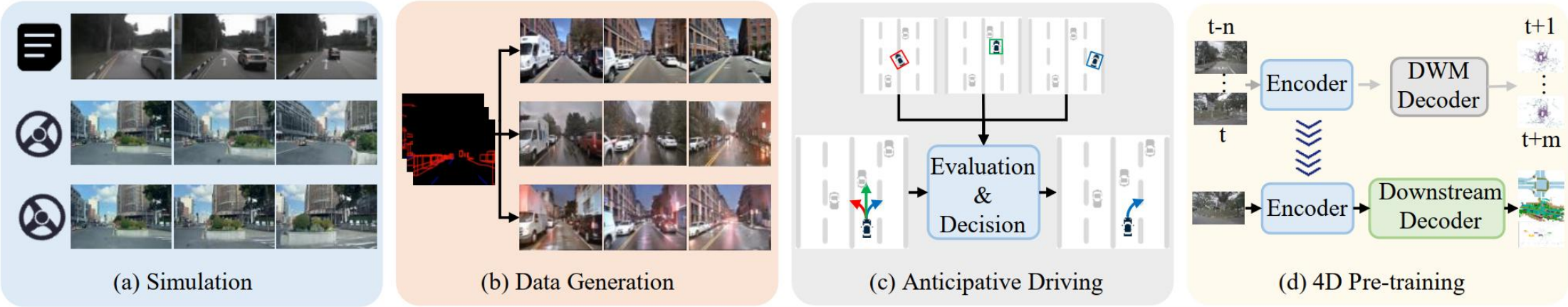
梳理出四大高价值应用：

1. 仿真：解决传统模拟器“泛化差” 传统模拟器（如CARLA）场景固定、与真实世界差距大，WM则能“按指令生成多样场景”——比如「Vista」可输入“雨天+早晚高峰”指令，生成1000段不同加塞场景的视频，用于测试规划算法的避险能力，比真实路测成本降低90%；

2. 数据生成：填补“长尾场景”缺口 自动驾驶最缺“事故 / 极端天气”等罕见数据，WM可定向合成——「DrivePhysica」通过 3D框约束生成“道路落物 + 急刹”场景，用这些数据微调3D检测器，召回率提升18%；「LidarDM」生成高保真LiDAR 点云，解决 LiDAR 数据采集贵的问题；

3. 预见性驾驶：让规划“更安全” 通过预测未来场景优化轨迹，比如「Drive-OccWorld」将WM与规划器结合，先预测未来3秒的3D 占用，再据此生成“无碰撞轨迹”，在nuScenes测试中，L2轨迹误差从0.96m 降至0.67m，碰撞率从0.71% 降至0.29%；

4. 4D 预训练：降低标注依赖 用大量无标注数据做4D预训练（预测时空演化），提升下游任务性能——「ViDAR」通过视觉点云预测预训练，让BEV检测器在少标注数据下（仅10%标注），NDS指标仍达 72.3，比直接训练高 15%。



世界模型 (WM)

指出WM落地“六大卡脖子问题”：

1.稀缺数据：用WM“自己造数据反哺自己” 现状：真实3D数据少、标注贵；解决思路：用已有少量数据训WM，再用WM生成大量合成 3D数据（如占用、点云），反哺WM自身训练，形成“数据 - 模型 - 更多数据”闭环；

2.效率低：解耦场景降低计算成本 现状：3D场景预测（如4D占用）需大量显存，推理慢；解决思路：参考「DFIT-OccWorld」，只预测动态物体的流动，静态场景通过姿态变换获取，计算量降低60%；

3.仿真不可靠：加入物理约束+多模态验证 现状：WM生成场景常出现“物理错误”（如车辆突然升空）；解决思路：「DrivePhysica」的做法 —— 加入3D流、碰撞约束，同时用相机/LiDAR 跨模态验证，确保场景符合物理规律；

4.统一任务：融合语言与预测 - 规划 现状：WM只做预测，与规划、语言指令脱节；解决思路：像「OccLLaMA」那样，让WM理解文本指令（如“避让施工车”），同时统一预测与规划，实现“指令→预测→轨迹”端到端；

5.多传感器建模：利用非对齐数据降低成本 现状：多传感器（相机/LiDAR）数据需精准对齐，采集成本高；解决思路：探索“非对齐/非配对” 数据融合，比如用单独的相机视频和LiDAR点云训WM，通过BEV空间统一模态；

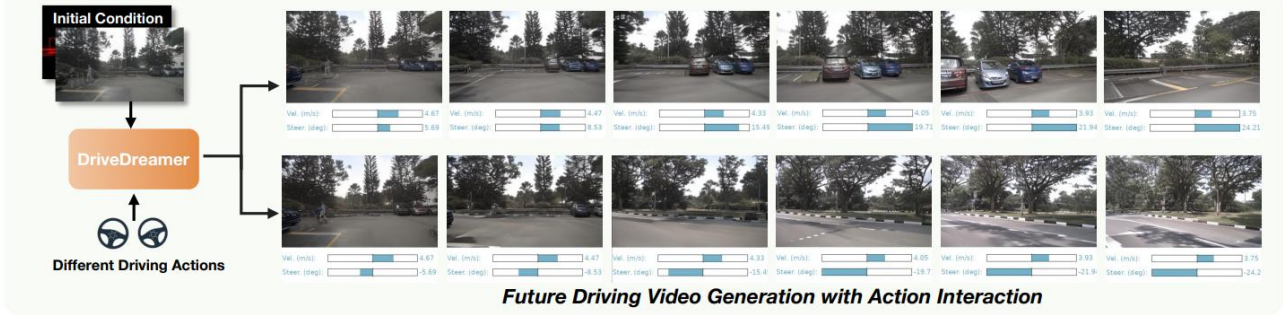
6.攻防缺失：警惕 adversarial攻击 现状：无研究关注WM的抗攻击能力；解决思路：开发针对WM的adversarial样本（如路面微小贴纸导致预测错误），同时设计防御策略（如多模态交叉验证），避免攻击引发事故。

世界模型 (WM)--数据生成

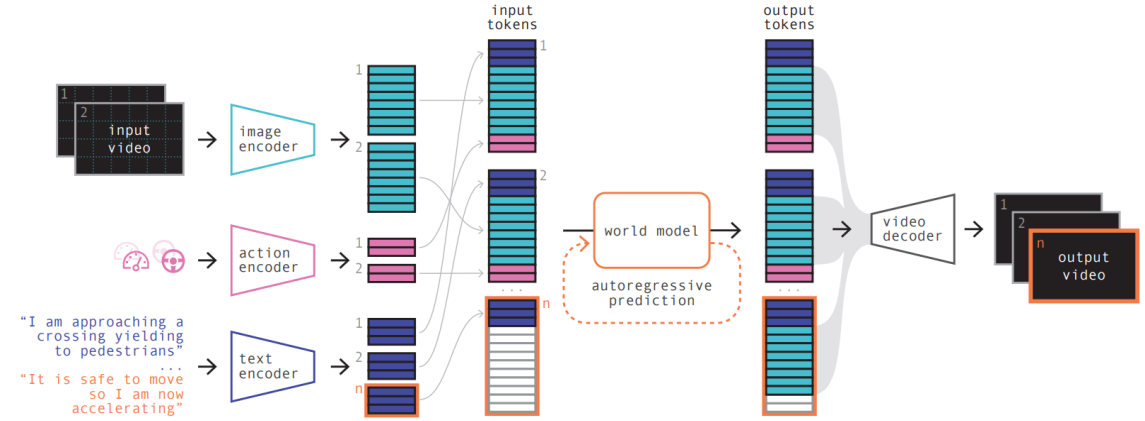
预测环境未来状态（动态物体运动、静态场景变化），是世界模型的“感知基础”。世界模型用于数据生成，降低落地成本，突破“标注依赖与长尾数据” 瓶颈。

GAIA-1代表了一种能使用视频、文本和动作输入创建逼真驾驶视频的世界模型。DriveDreamer输入HD地图和3D框等元素，允许更精确的控制和更深入的理解，从而提高视频生成质量。

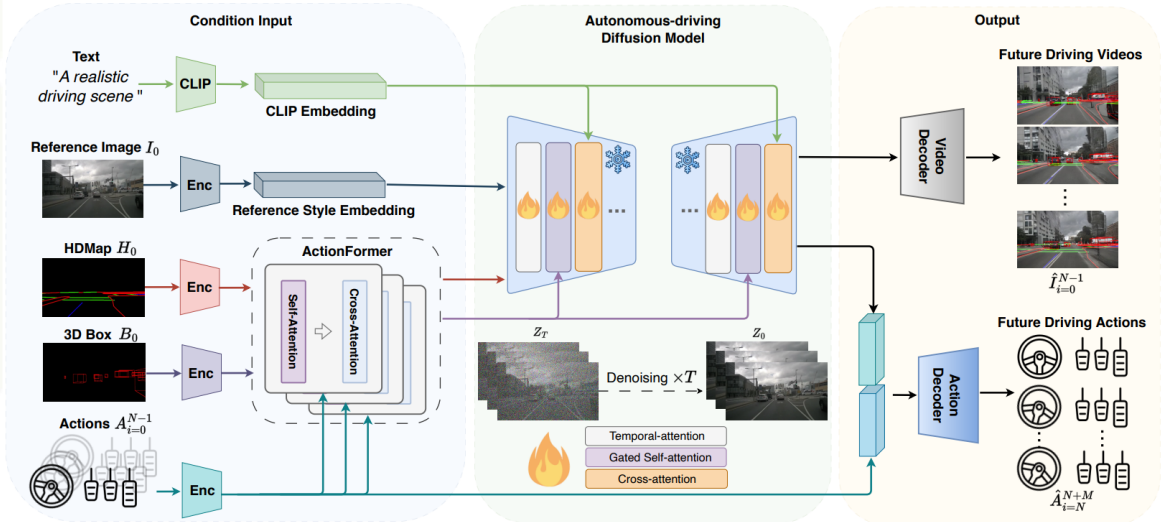
Driving Video Generation with Traffic Condition and Different Text Prompts



Future Driving Video Generation with Action Interaction



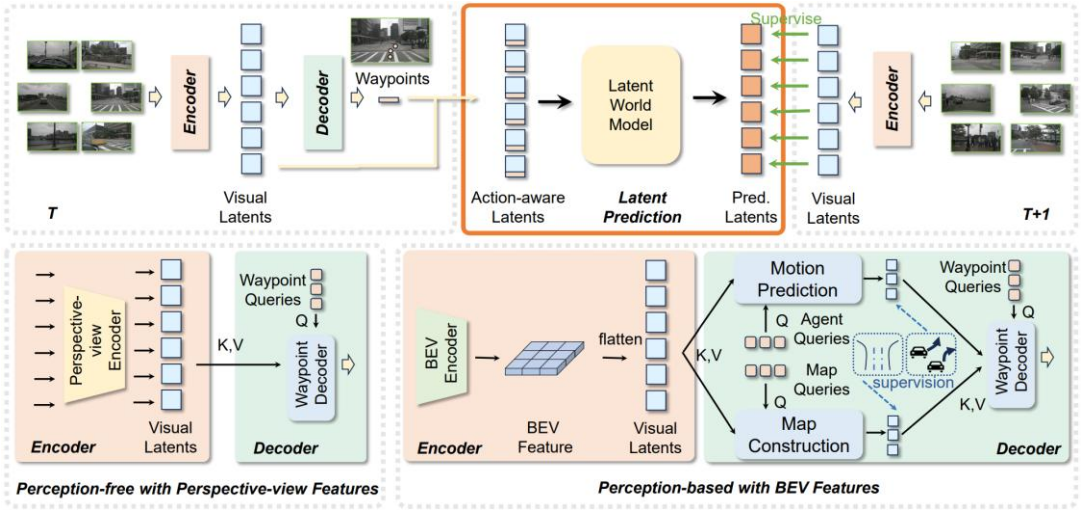
GAIA-1: A Generative World Model for Autonomous Driving



DriveDreamer: Towards Real-world-driven World Models for Autonomous Driving

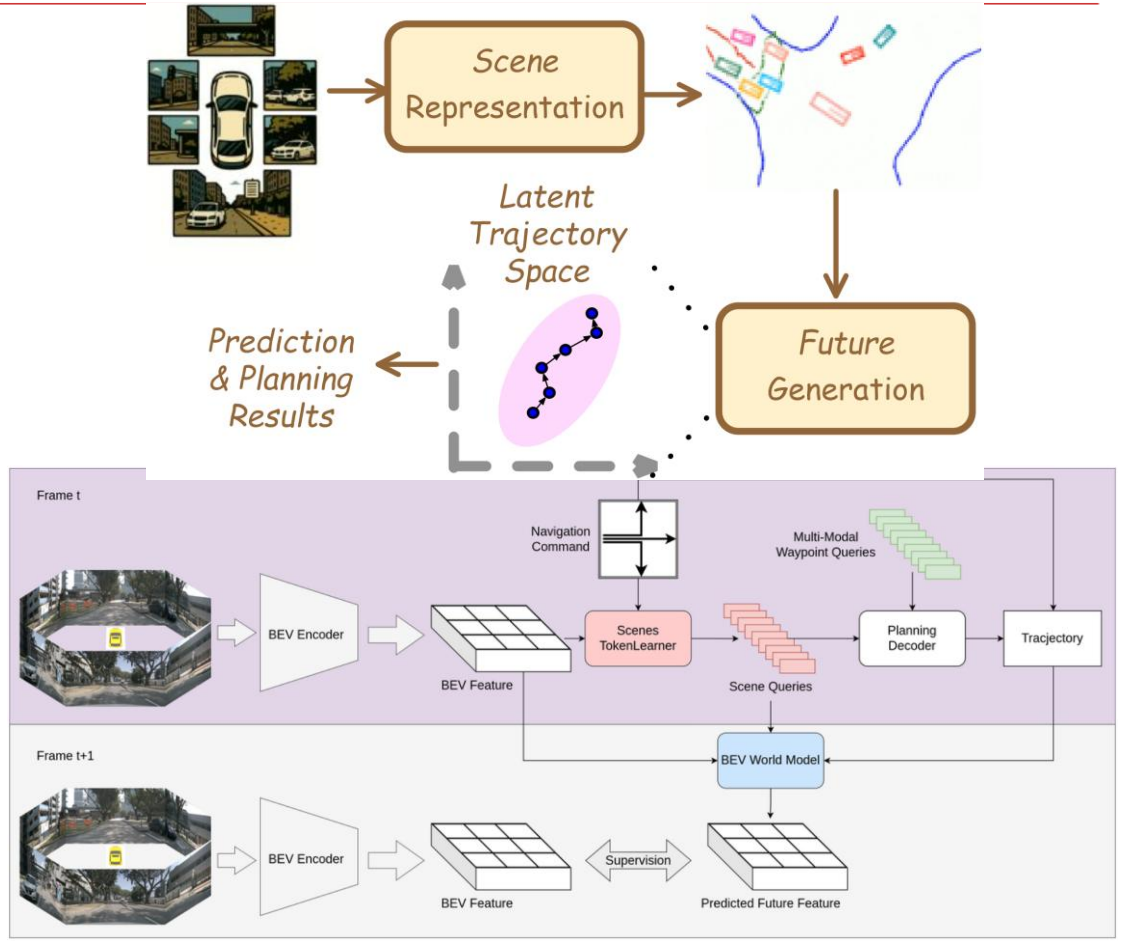
世界模型 (WM)--增强端到端

模块化端到端的中间感知任务需要大量标注成本，直接端到端仅靠轨迹优化难以精确理解场景表示。**世界模型则能够通过未来场景预测任务以自监督的方式学习到良好的场景表示，可以辅助增强各种端到端范式的场景表示能力，进而提高轨迹规划的准确性。**



Enhancing End-to-End Autonomous Driving with Latent World Model

LAW利用当前场景特征和预测轨迹生成未来帧的潜在特征，并通过未来实际观测特征进行自监督学习，从而提升场景特征表示和轨迹预测性能。

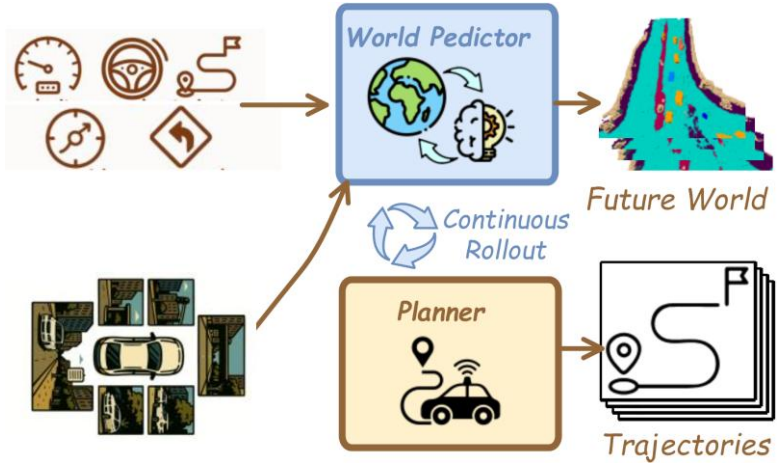


Navigation-Guided Sparse Scene Representation for End-to-End Autonomous Driving

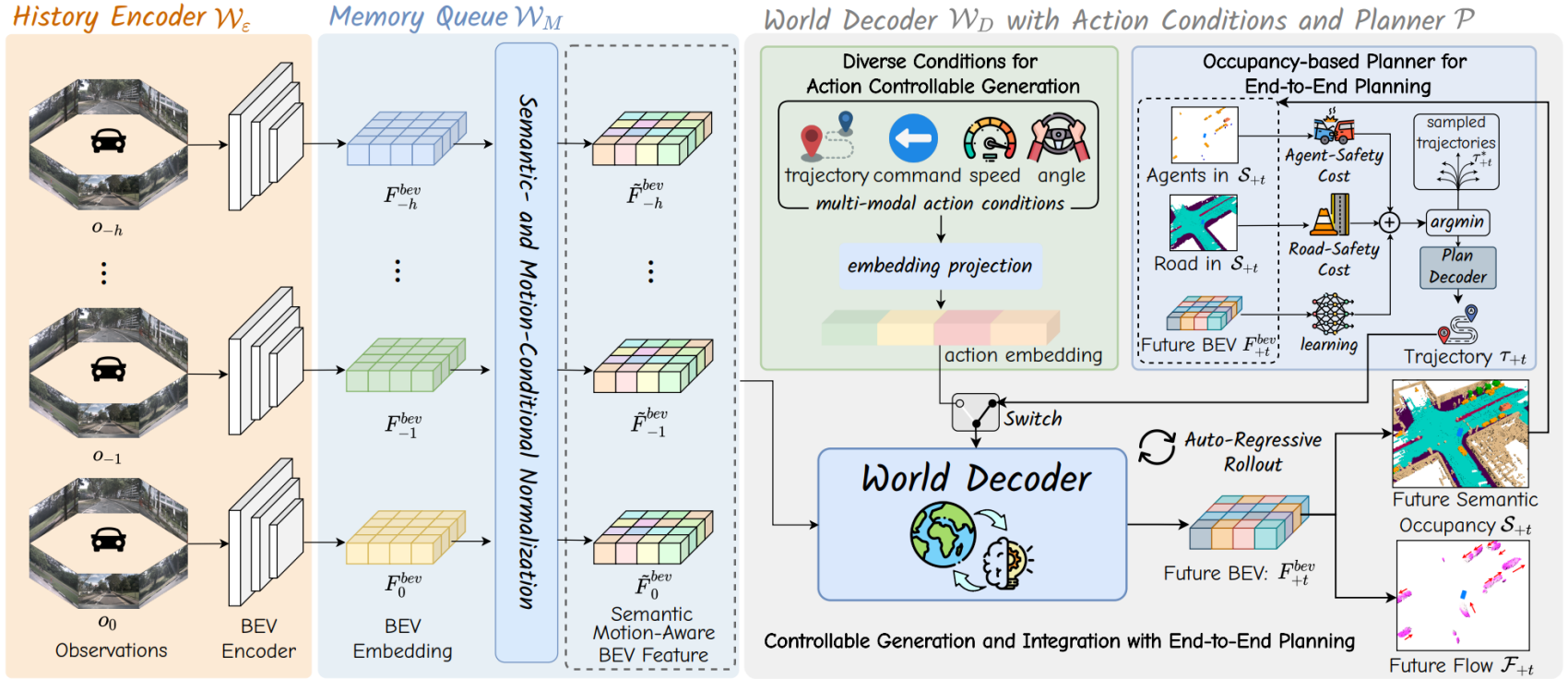
SSR在LAW的基础上进一步通过稀疏场景表示和端到端学习提升自动驾驶系统效率，核心在于用少量导航引导的标记替代传统密集感知任务。

世界模型 (WM)--直接轨迹规划

世界模型可以通过预测未来场景来直接优化轨迹。世界模型需为下游任务提供统一接口——例如为行为规划模块输出可解释的环境状态，为轨迹生成提供动态障碍物约束。



比如Drive-OccWorld将DWM 与规划器结合，先预测未来3秒的3D占用，再据此生成“无碰撞轨迹”。



Driving in the Occupancy World: Vision-Centric 4D Occupancy Forecasting and Planning via World Models for Autonomous Driving

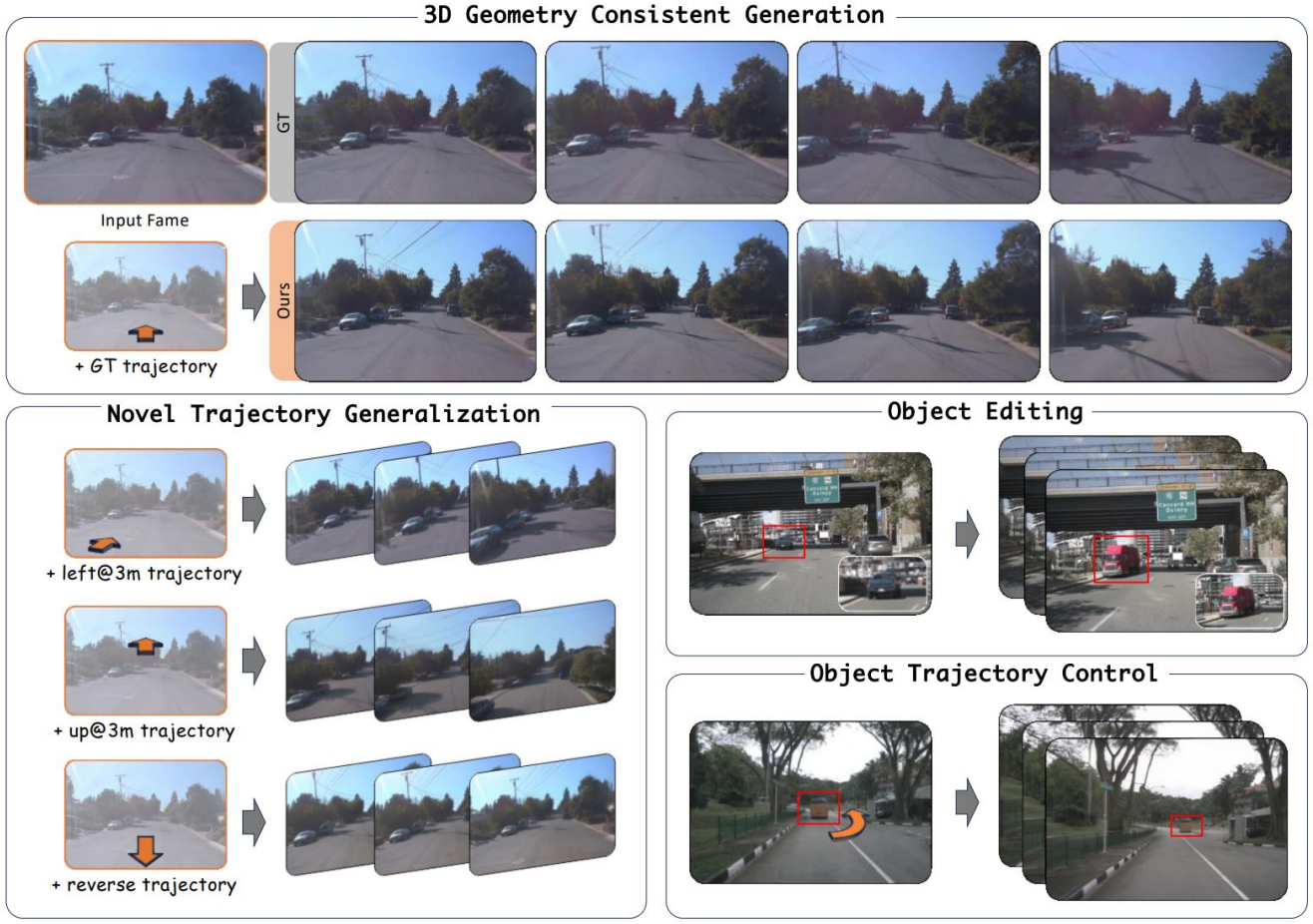
世界模型 (WM)--最新相关工作

GeoDrive: 3D Geometry-Informed Driving World Model with Precise Action Control, arXiv 2025.5.28. (北大+理想)

自动驾驶世界模型通过模拟三维动态环境，使以下关键能力成为可能：轨迹一致的视角合成、符合物理规律的运动预测，以及安全感知的场景重建和生成。

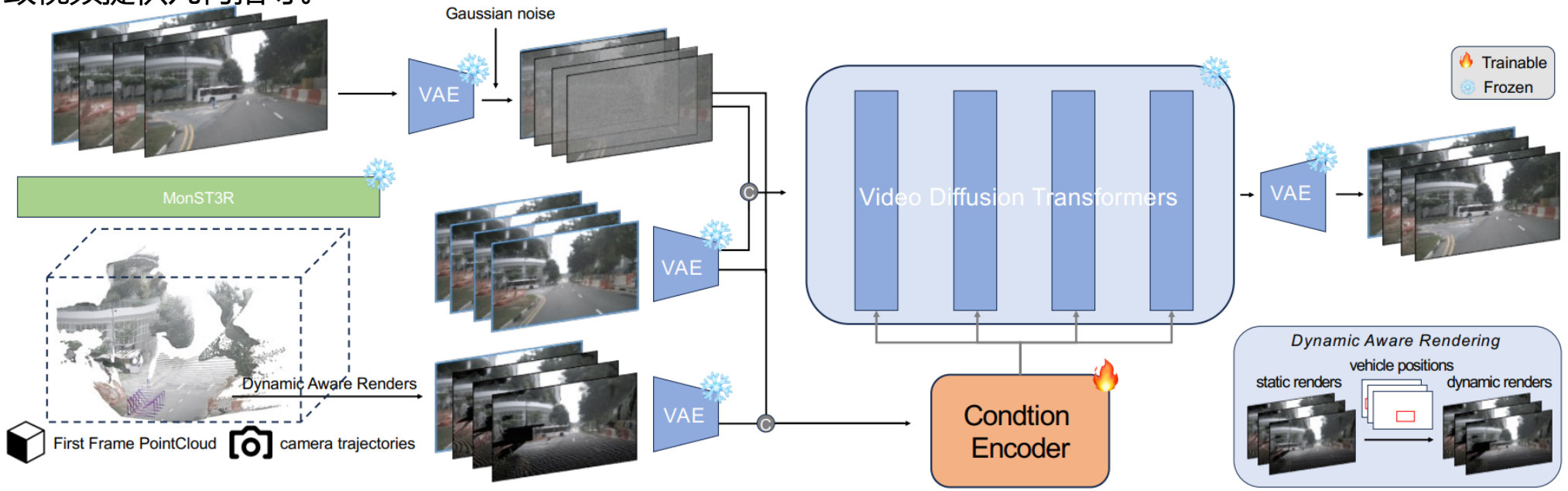
现有方法普遍依赖二维建模、缺乏三维空间感知，从而导致轨迹不合理、动态交互失真的问题。

GeoDrive首创性地将三维点云渲染过程纳入生成范式，在每一帧生成中显式注入空间结构信息，在动作准确性和3D空间感知方面明显优于现有模型，从而为更安全的自动驾驶实现了更逼真、适应性更强且更可靠的场景建模。此外，模型可以泛化到新的轨迹，并且提供交互式场景编辑功能，例如目标编辑和目标轨迹控制。



世界模型 (WM)--最新相关工作

给定初始参考图像和自车轨迹，框架合成遵循输入轨迹的真实未来帧。利用参考图像中的3D几何信息来指导世界建模。首先，重建3D表示，然后沿着用户指定的轨迹渲染视频序列，并进行动态物体处理。渲染后的视频为生成遵循输入轨迹的时空一致视频提供几何指导。



当前的方法在保持强大的 3D 几何一致性或在遮挡处理期间累积伪影方面表现出不足，这两者对于自动导航任务中的可靠安全评估都至关重要。为了解决这个问题，GeoDrive 将强大的 3D 几何条件明确地集成到驾驶世界模型中，以增强空间理解和动作可控性。具体而言，首先从输入帧中提取 3D 表示，然后根据用户指定的自车轨迹获取其 2D 渲染。为了实现动态建模，提出一个训练期间的动态编辑模块，通过编辑车辆的位置来增强渲染。大量实验表明，该方法在动作精度和 3D 空间-觉察方面均显著优于现有模型，从而能够构建更逼真、适应性更强、更可靠的场景建模，从而实现更安全的自动驾驶。

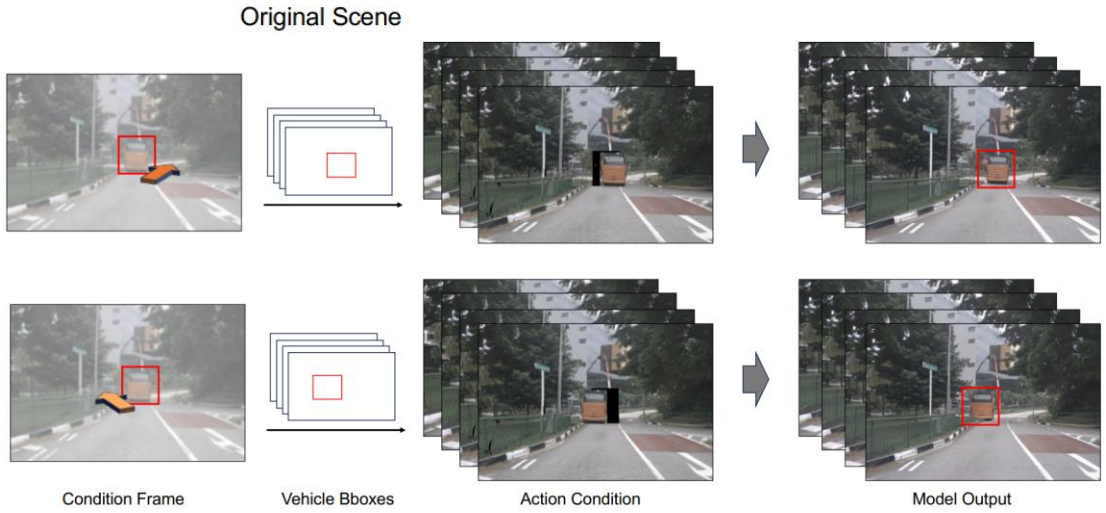
GeoDrive: 3D Geometry-Informed Driving World Model with Precise Action Control, arXiv 2025.5.28.

世界模型 (WM)--最新相关工作

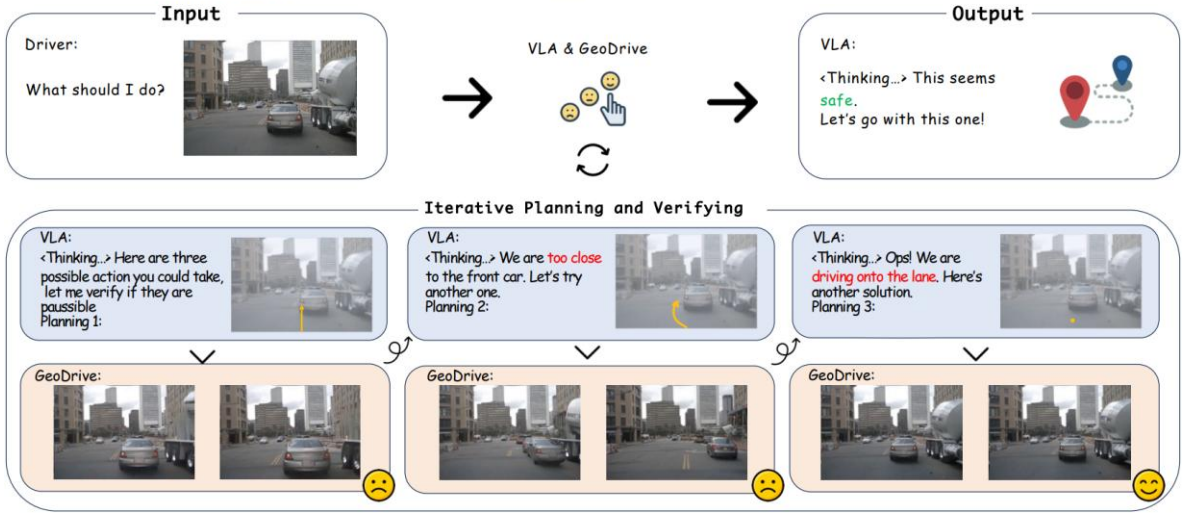
不仅支持高级场景编辑（例如目标编辑和目标轨迹控制），还可以作为交互式环境来辅助高级规划模型（例如VLA）

目标编辑

目标轨迹控制



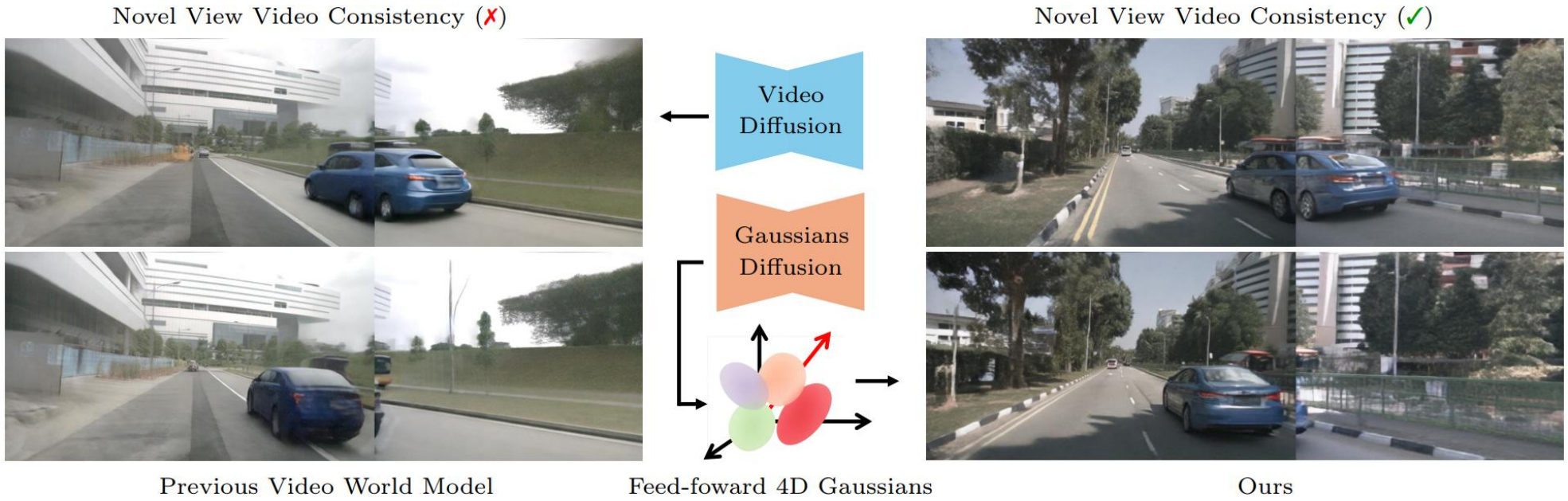
辅助高级规划



GeoDrive: 3D Geometry-Informed Driving World Model with Precise Action Control, arXiv 2025.5.28.

世界模型 (WM)--最新相关工作

WorldSplat: Gaussian-Centric Feed-Forward 4D Scene Generation for Autonomous Driving, arXiv 2025.9.27. (南开+小米)

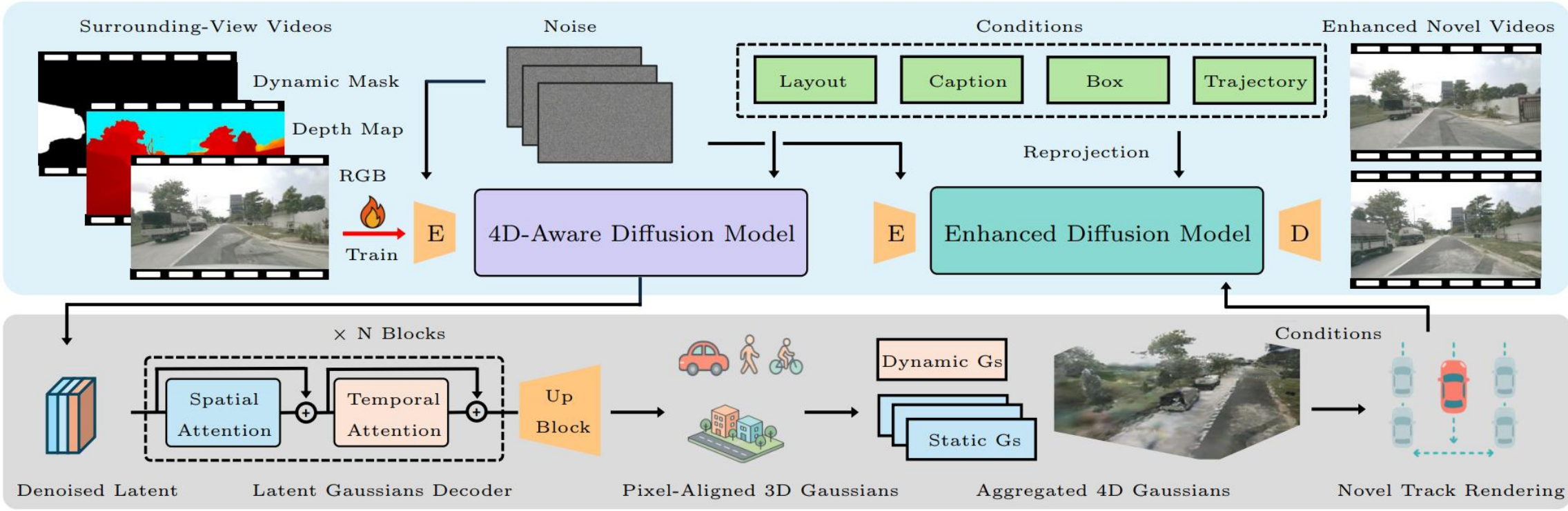


- 世界模型是一种生成式的方法，现有方法主要聚焦于生成多样化、真实的的驾驶视频；然而由于3D一致性有限且视角覆盖稀疏，这些方法难以支持便捷、高质量的新视角合成。与之相反，近年来的3D/4D重建方法虽大幅提升了真实驾驶场景的重建效果，却天生缺乏生成新视角能力。
- 所以一种可能的方式是利用生成+重建结合的形式，来建模自动驾驶场景。3DGS用于重建原始场景，生成方法用于优化新视角，两者强强联合以构建自动驾驶中的世界模型。

- WorldSplat是一种全新的前馈式框架，该框架整合了生成式方法与重建式方法的优势，用于4D驾驶场景合成。通过将4D感知潜在扩散模型与增强型扩散网络相结合，能够生成显式4D高斯分布，并将其优化为高保真、具备时间与空间一致性的多轨迹驾驶视频。

世界模型 (WM)--最新相关工作

框架包含三个关键模块：用于多模态潜变量生成的4D感知潜在扩散模型、用于前馈式4D高斯预测及实时轨迹渲染的潜在高斯解码器，以及用于视频质量优化的增强型扩散模型。



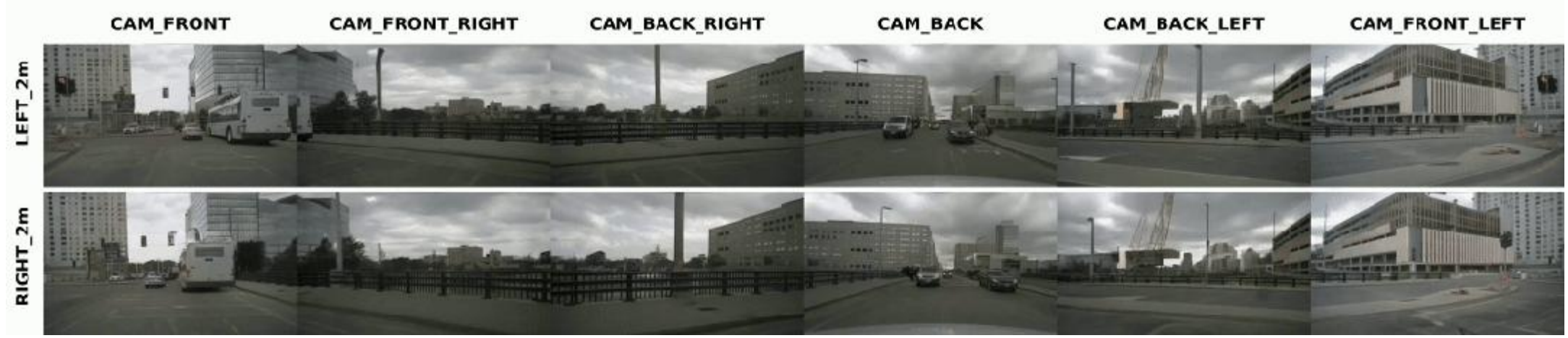
给定噪声潜变量与细粒度条件（即边界框、道路草图、文本描述及自车轨迹），**4D感知潜在扩散模型**通过去噪生成包含RGB、深度与动态目标信息的多模态潜变量，这些潜变量随后用于4D高斯预测。

潜在4D高斯解码器从多模态潜变量L中预测像素对齐的3D高斯分布，随后利用潜变量中的语义信息区分动态与静态目标，并基于3D高斯分布重建4D场景。

增强型扩散模型对基于4D高斯分布渲染的RGB视频进行优化，生成过程同时以原始输入与渲染视频为条件。该优化过程可丰富空间细节并增强时间连贯性，最终输出高保真新视角序列。

WorldSplat: Gaussian-Centric Feed-Forward 4D Scene Generation for Autonomous Driving, arXiv 2025.9.27.

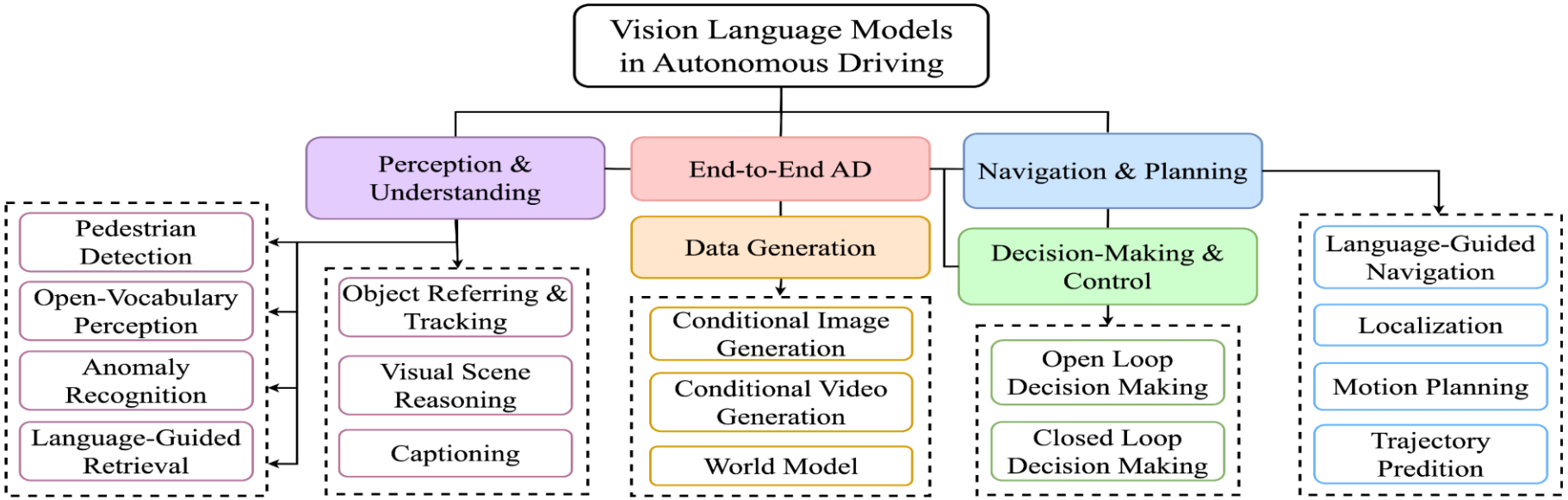
世界模型 (WM)--最新相关工作



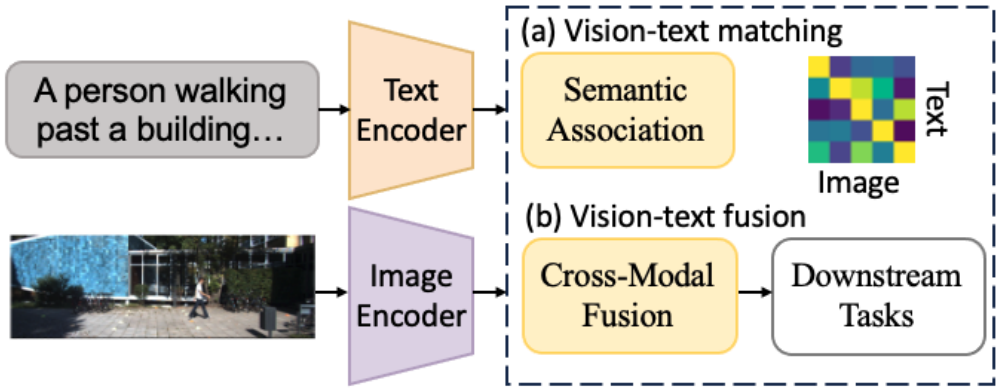
WorldSplat: Gaussian-Centric Feed-Forward 4D Scene Generation for Autonomous Driving, arXiv 2025.9.27.

视觉-语言-动作模型(VLA)

自动驾驶中的视觉-语言



自动驾驶中的视觉语言任务

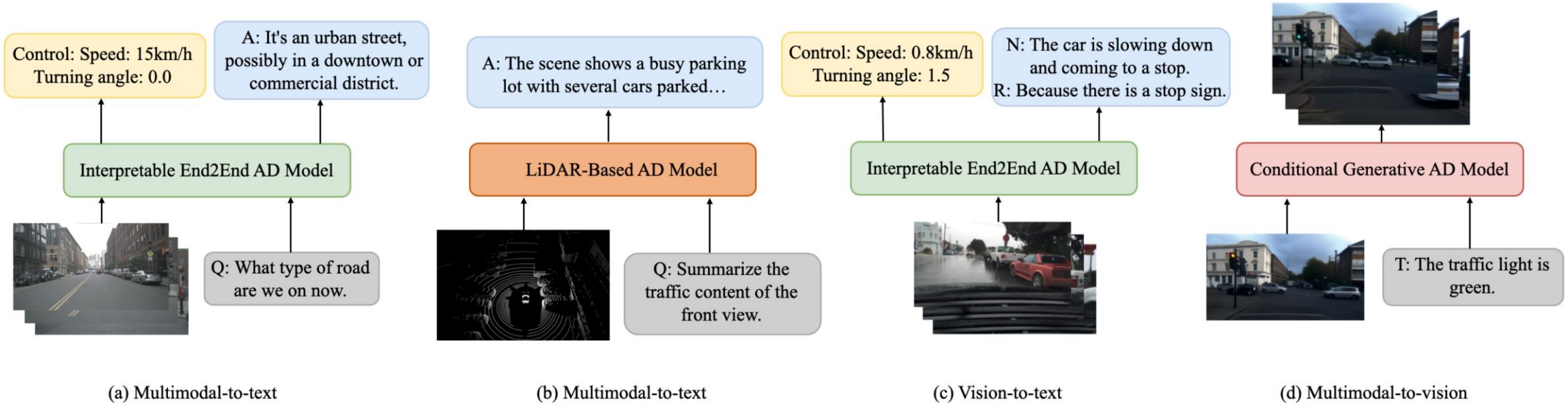


自动驾驶中视觉、语言模态连接方法

- 视觉–文本匹配：语义相似度匹配
- 视觉–文本融合：融合后的特征可用于下游任务

Vision Language Models in Autonomous Driving: A Survey and Outlook

自动驾驶中主流视觉语言模型

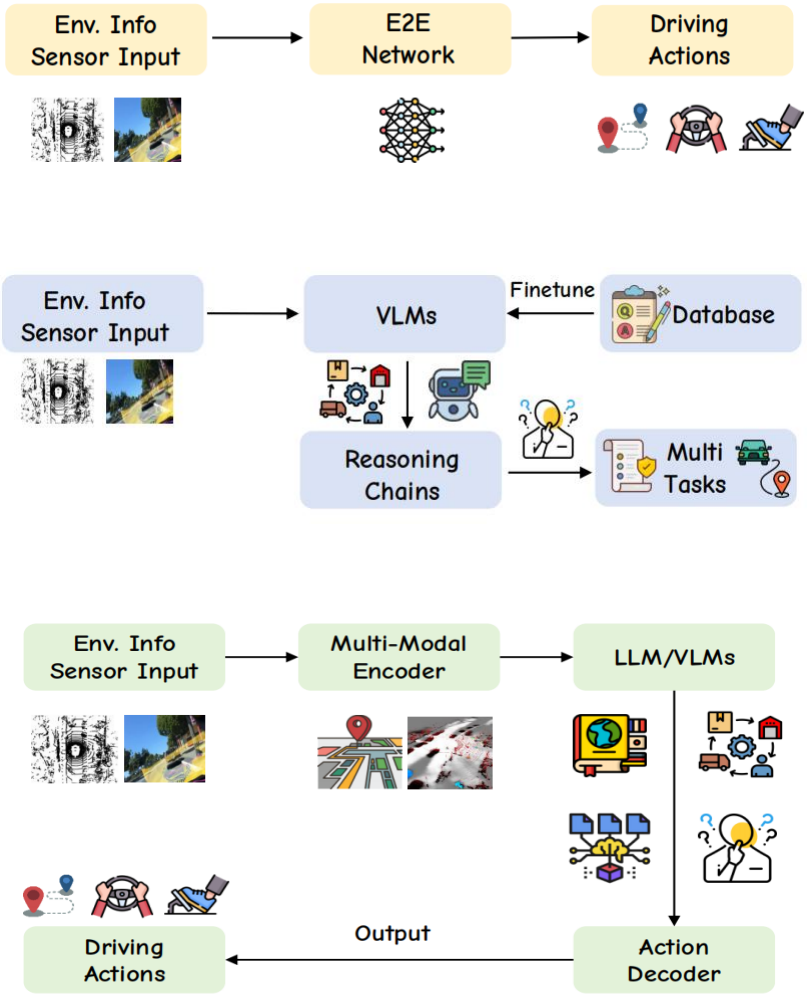


(a) Multimodal-to-text (b) Multimodal-to-text (c) Vision-to-text (d) Multimodal-to-vision

- 多模态输入到文本（基于视觉）：以文本和图像或视频作为输入，生成文本。
- 多模态输入到文本（基于LiDAR）：以文本和点云作为输入，生成文本。
- 视觉输入到文本：以图像或视频作为输入，生成文本。
- 多模态输入到视觉：以图像或视频作为输入，生成图像或视频。

Vision Language Models in Autonomous Driving: A Survey and Outlook

端到端自动驾驶架构的进化



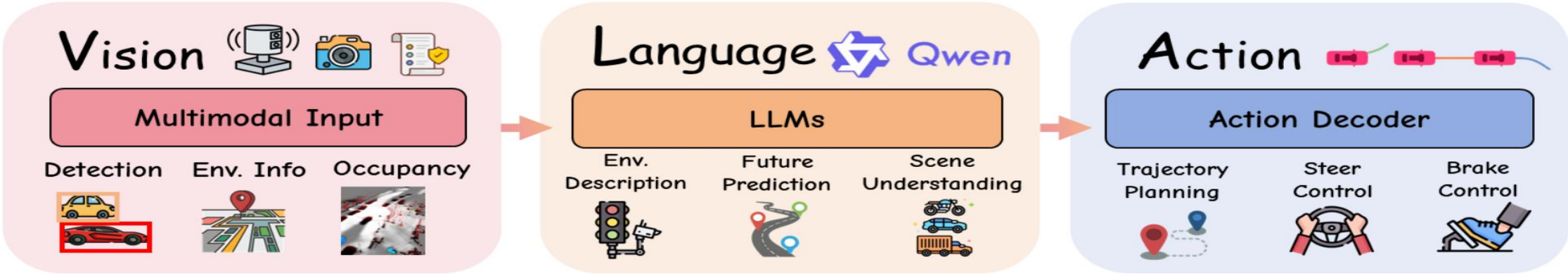
黑箱端到端：从“感知”直接映射到“驾驶动作”

VLM引入了自然语言推理，增强了可解释性
早期工作以“感知”部分为中心

集成了感知、推理和驾驶动作完整VLA架构

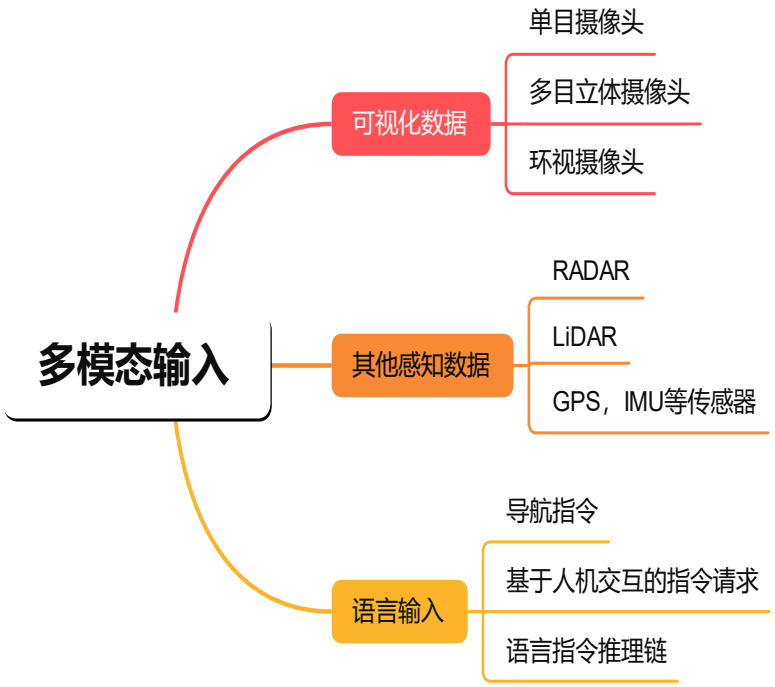
A Survey on Vision-Language-Action Models for Autonomous Driving

VLA用于自动驾驶的范式



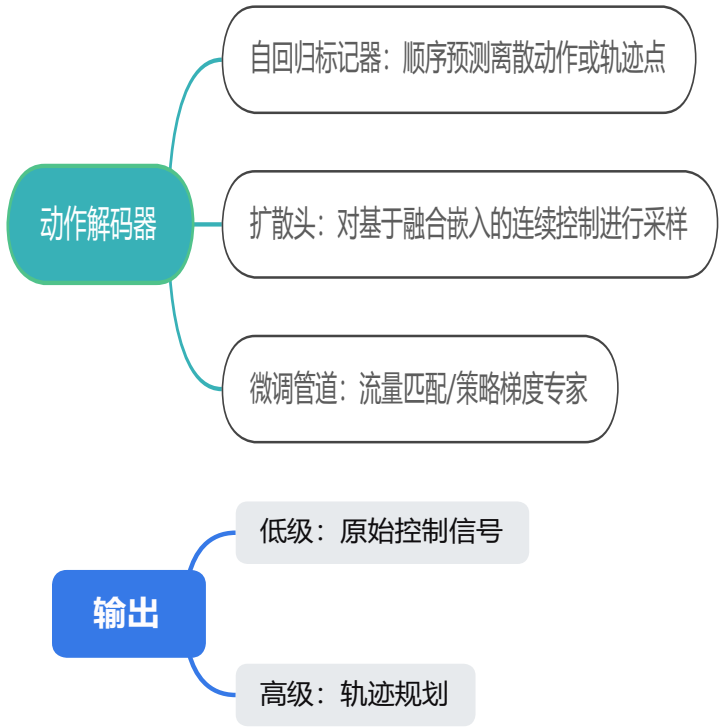
VLA4AD 的基本架构将视觉感知、语言理解和行动生成集成在一个连贯的管道中

编码：将原始传感器数据转换为潜在表示，例如BEV，点云，像素模块。



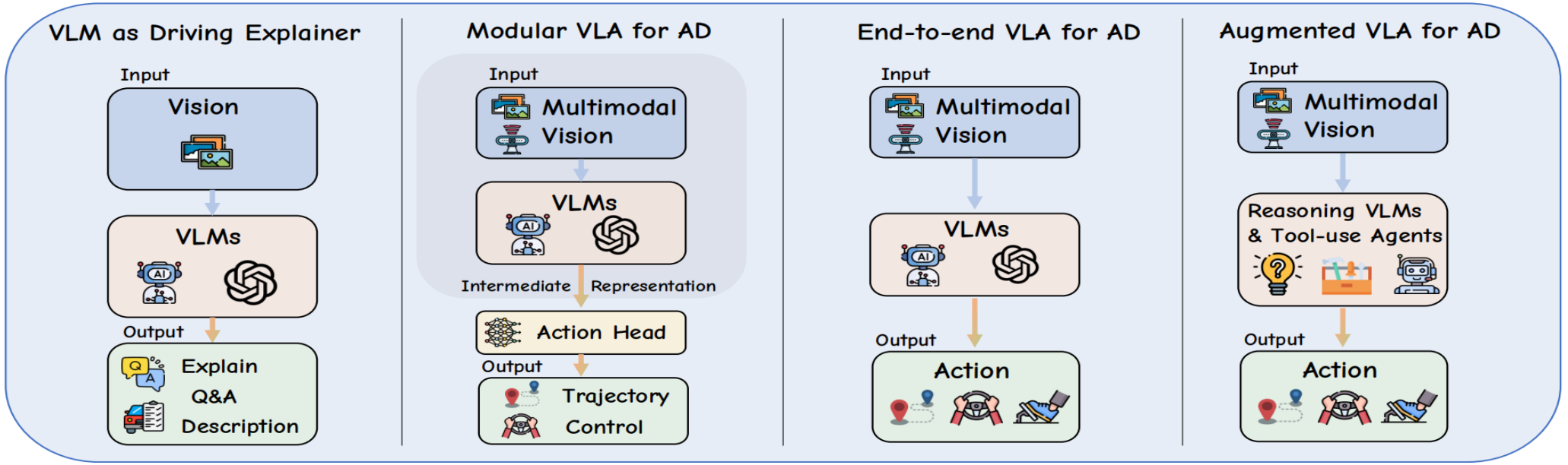
语言处理器：使用预训练的解码器进行自然语言的处理。

- 指令调优变体和检索增强提示：注入领域知识。
- 轻量级微调：高效自适应。



A Survey on Vision-Language-Action Models for Autonomous Driving

VLA架构的进化

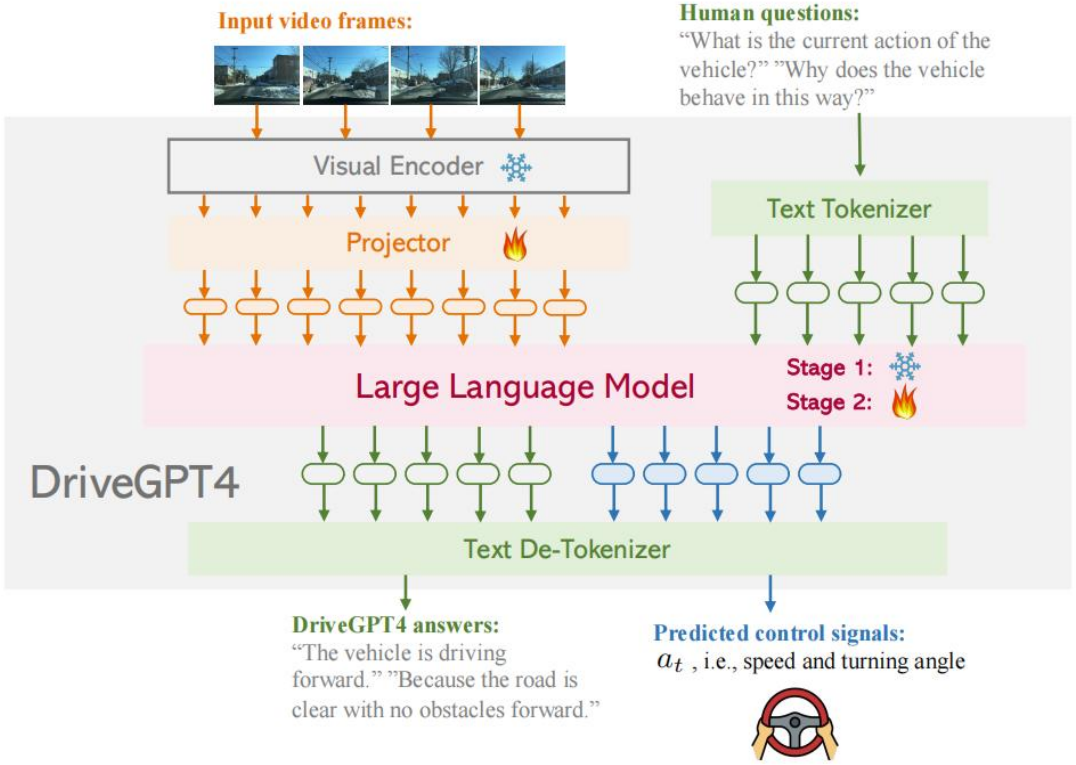


VLM作为语言头 → 模块化 (action解耦) → 端到端VLA → 增强的VLA

VLM作为语言头	模块化 (action解耦)	端到端VLA	增强的VLA
VLM进行场景描述并解释驾驶行为，缓解端到端“黑箱”问题	模块化端到端，VLM生成中间表示，action模块解耦	端到端VLA，直接从多模态输入到驾驶动作的全流程决策模型	运用推理增强技术的先进VLA模型

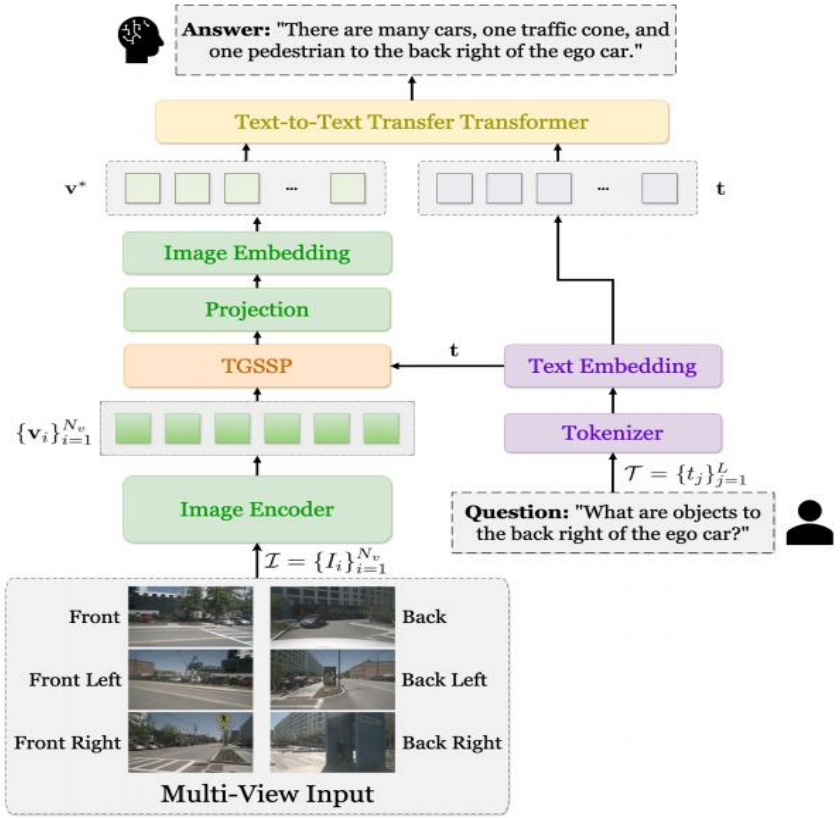
A Survey on Vision-Language-Action Models for Autonomous Driving

早期VLA--增强可解释性



Drivegpt4: Interpretable end-to-end autonomous driving via large language model

- 视觉编码器：提取帧级特征。
- 视频tokenizer：将时序帧特征转为一系列token使语言模型能够理解。
- LLaMA语言头：负责多模态融合后的推理与生成。

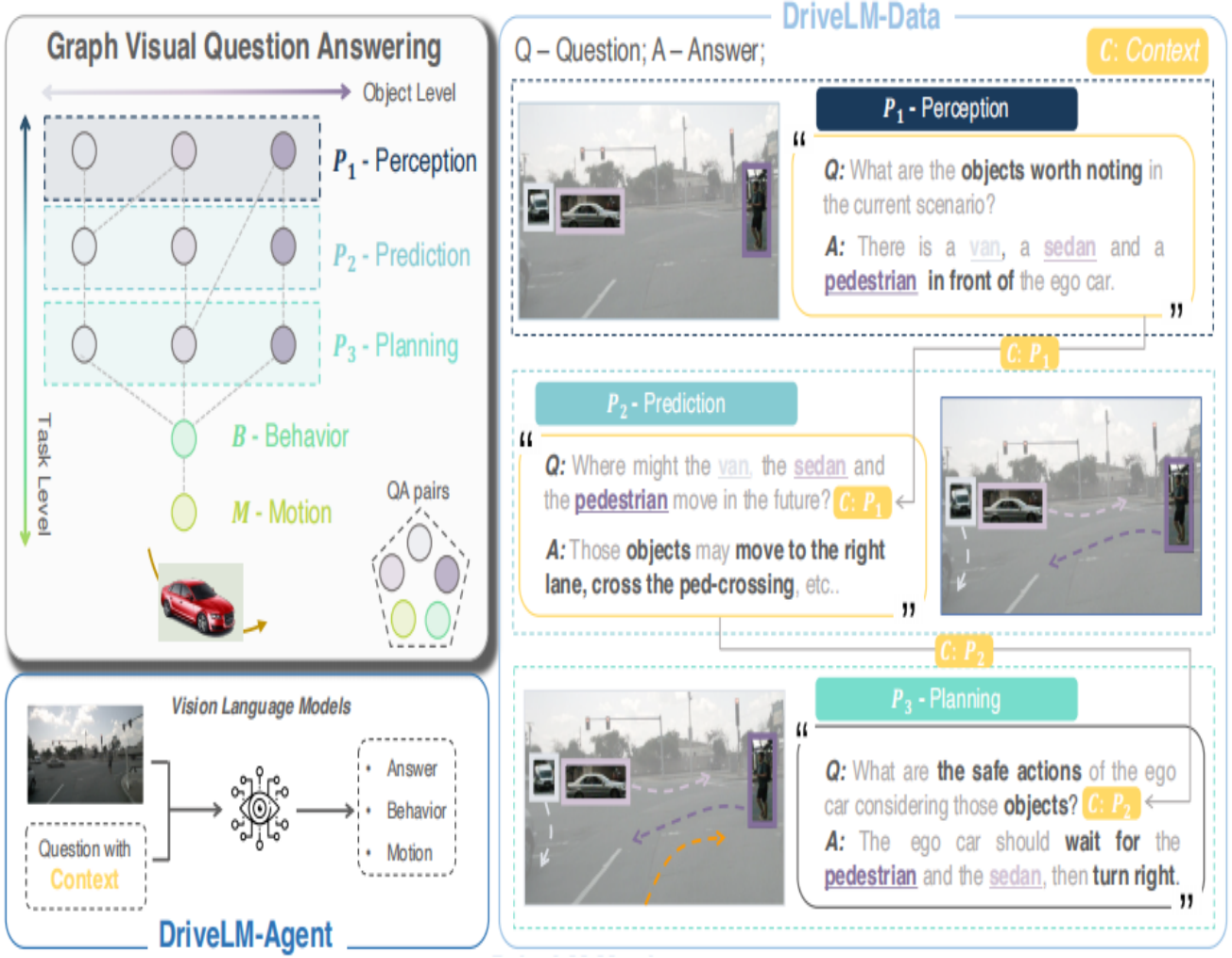


TS-VLM: Text-Guided SoftSort Pooling for Vision-Language Models in Multi-View Driving Reasoning

提出轻量级多视角VLM，用于多摄像头输入下的场景理解与语言解释。

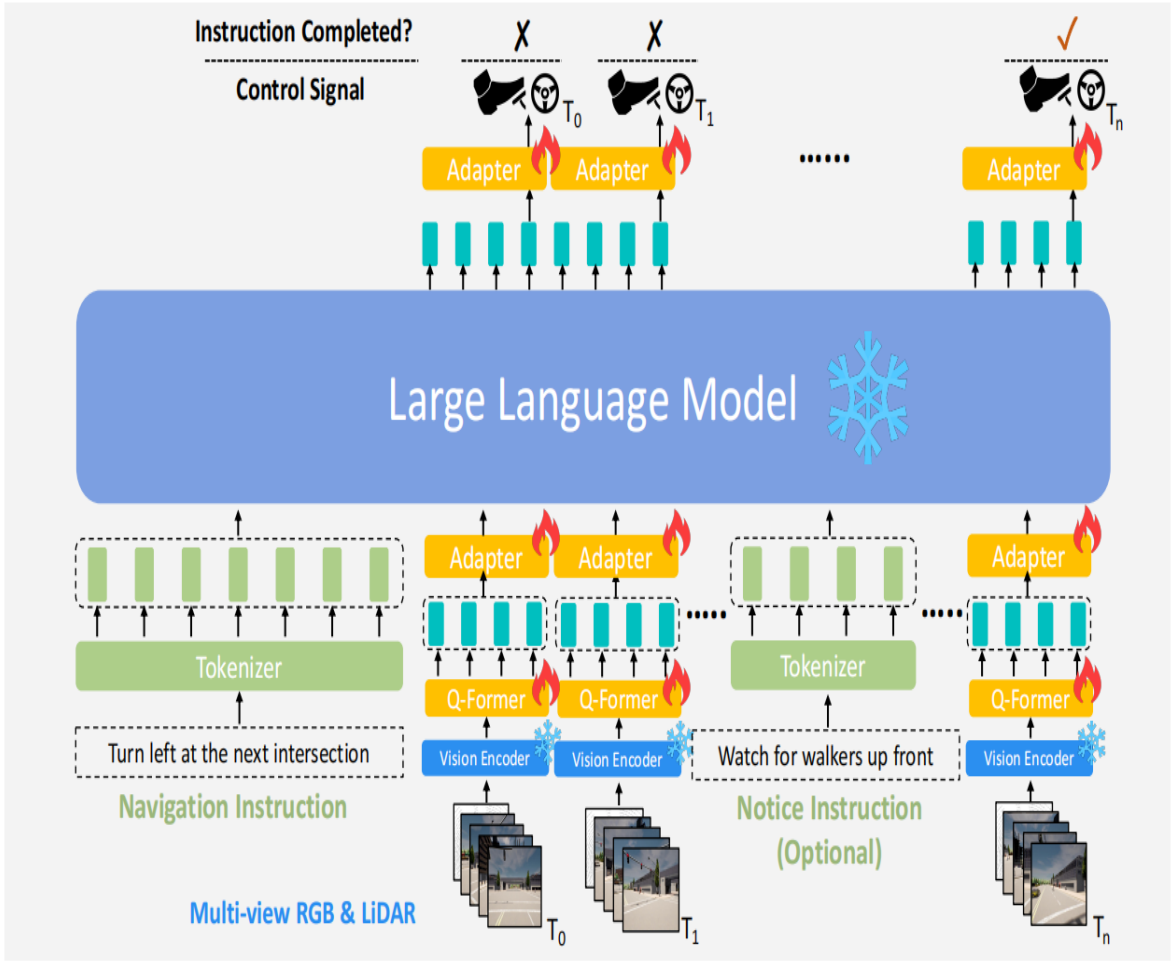
- 根据输入文本引导多视角特征排序与融合。
- 根据问题语义聚焦相关视角，实现可解释、语义驱动的视觉聚合。

中期VLA--模块化（action解耦）



Drivelm: Driving with graph visual question answering

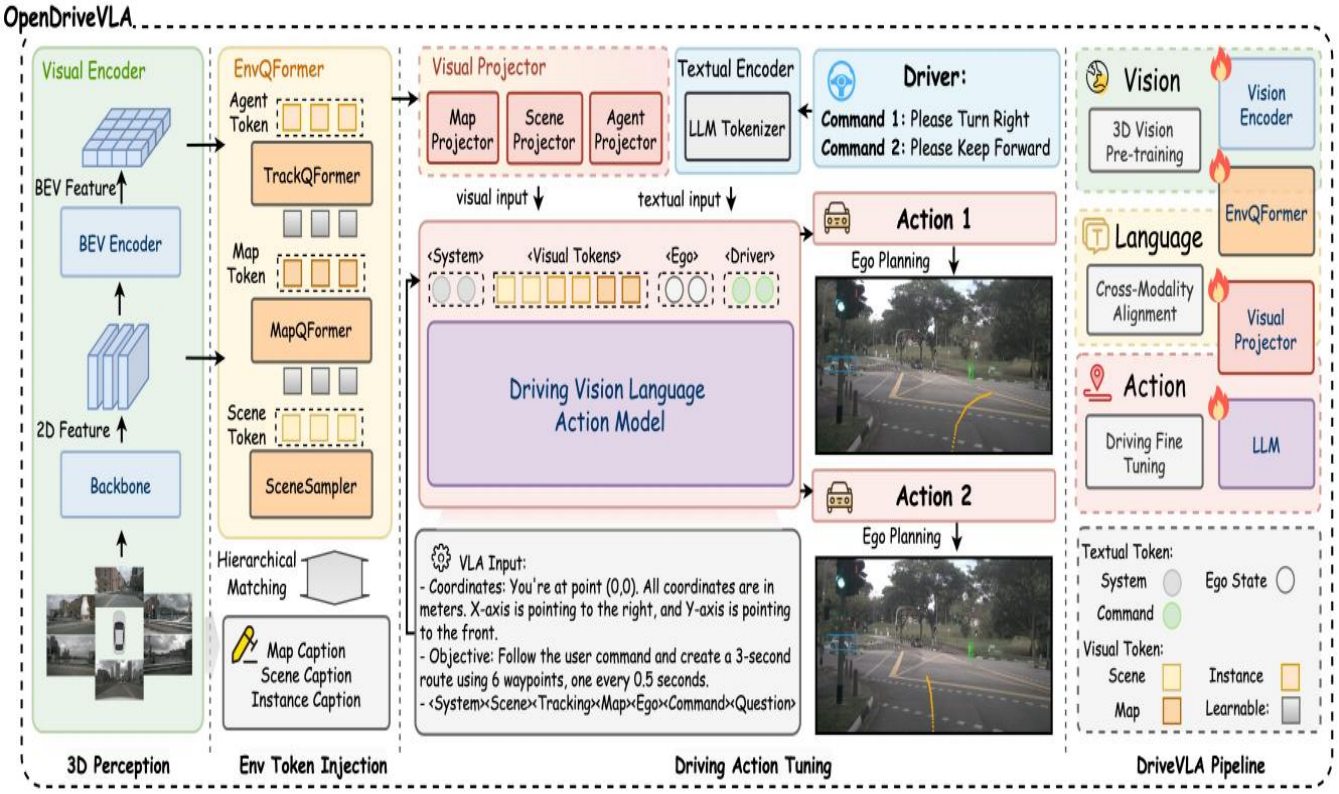
模块化VLA：拆解为Graph，Visual，Question，Answering四大任务。



Lmdrive: Closed-loop end-to-end driving with large language models

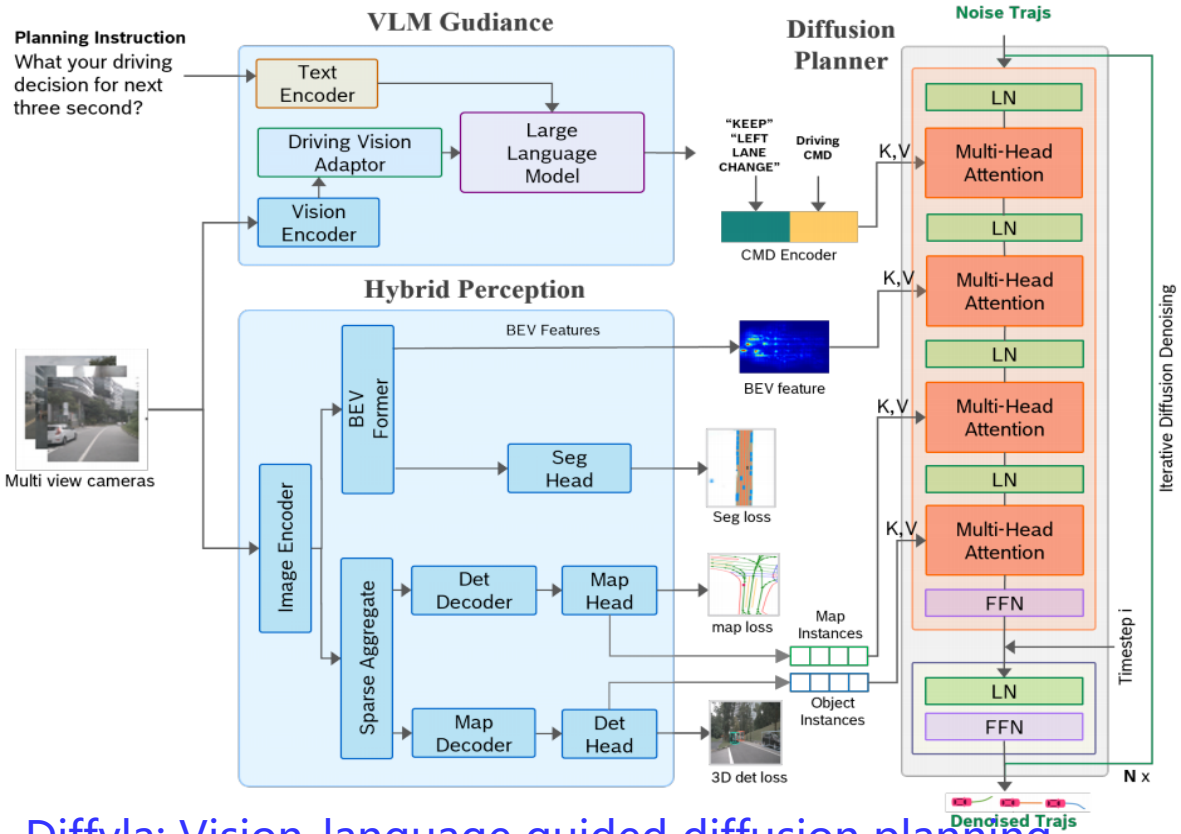
解耦输出头(Action Adapter)将LLM的输出映射为：未来waypoints(连续动作)；是否完成指令(逻辑控制)。

近期VLA--端到端



Opendrivevla: Towards end-to-end autonomous driving with large vision language action model

- 分层特征对齐：将2D、3D视觉token映射到语义空间，解决模态鸿沟。
- 交互：自回归建模自车、智能体和静态环境的交互，提升轨迹可靠性。



Diffvla: Vision-language guided diffusion planning for autonomous driving

- 混合感知：稀疏与密集感知并行，提供全面的环境表征。
- VLM引导：生成高级驾驶指令，指导扩散过程。
- 扩散规划：采用截断扩散策略，利用多锚点先验，高效生成多条候选轨迹。

VLA的不足

- **物理不可行或结构复杂的动作生成**

VLM擅长语义理解和语言推理，但直接用VLM输出文本化的动作或轨迹，可能不符合物理约束，导致车辆动作不连贯或不可执行

- **推理方式单一，复杂推理速度慢**

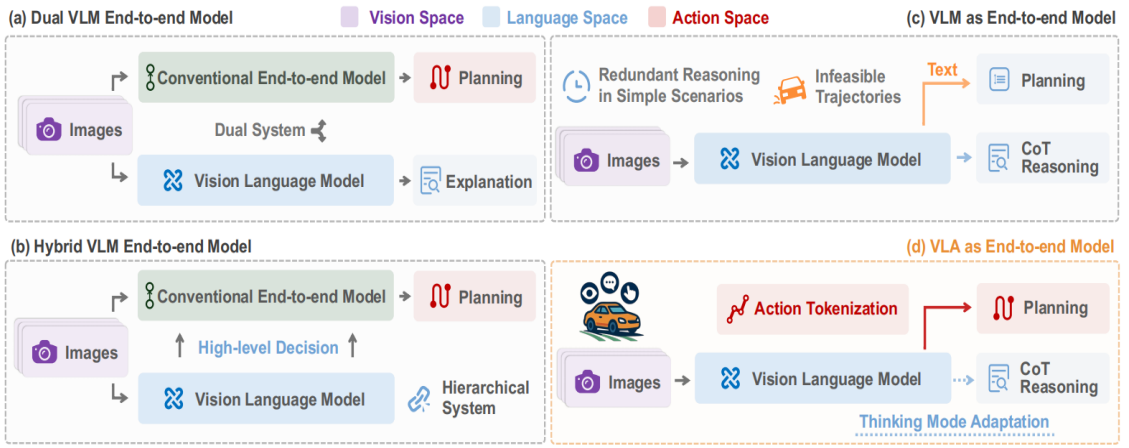
多数模型采用固定的思维链(CoT)，产生过多冗余文本，推理效率低。

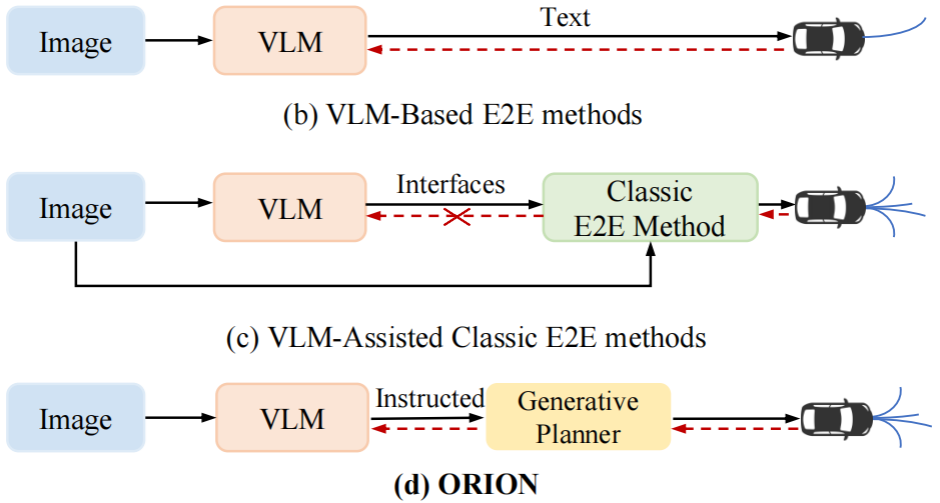
- **时序建模能力不足**

现有VLM通常简单拼接多帧图像捕捉时间信息，受限于token长度难以有效利用历史上下文。

- **“接口式”融合割裂优化**

让VLM输出中间动作指导传统端到端模型（即action解耦），限制了联合优化。





VLA--最新相关工作

当前基于VLA的端到端自动驾驶方法存在明显缺陷，AutoVLA 的提出正是为了从根本上解决这些问题。其核心创新理念是：在一个单一的自回归生成模型中，统一语义推理（Reasoning）和物理动作生成（Action），并赋予模型自适应切换“快思考”和“慢思考”模式的能力。

- **物理动作令牌化**

从真实数据聚类2048个离散动作令牌，每个令牌编码0.5秒运动，覆盖主流驾驶模式。

连续轨迹规划

↓

离散令牌序列预测问题

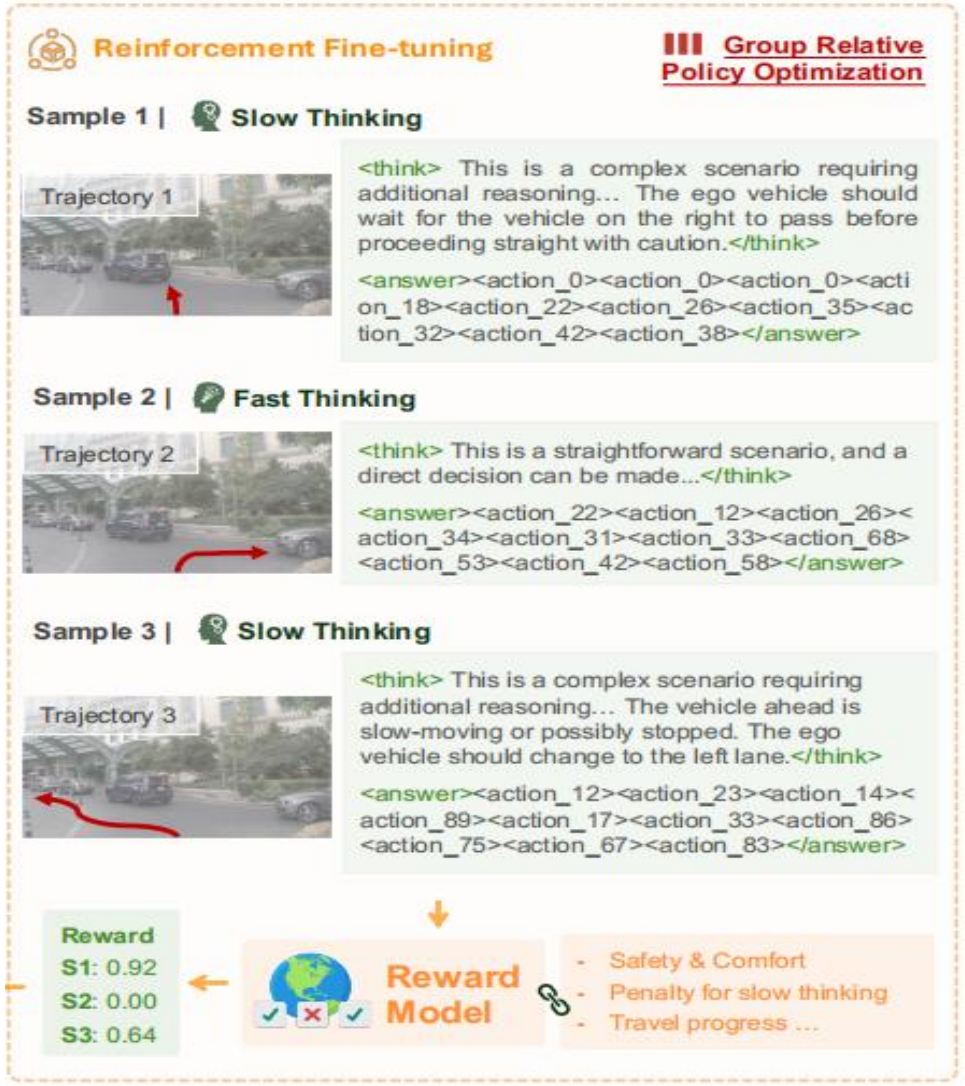
- **轨迹编码为令牌序列**

输入5秒连续轨迹，拆分为10个片段，每个片段匹配码本中最接近的动作令牌。

输出：自回归生成混合序列（文本推理令牌+动作令牌）

AutoVLA: A Vision-Language-Action Model for End-to-End Autonomous Driving with Adaptive Reasoning and Reinforcement Fine-Tuning

VLA--最新相关工作



双思维推理机制

- Fast Thinking

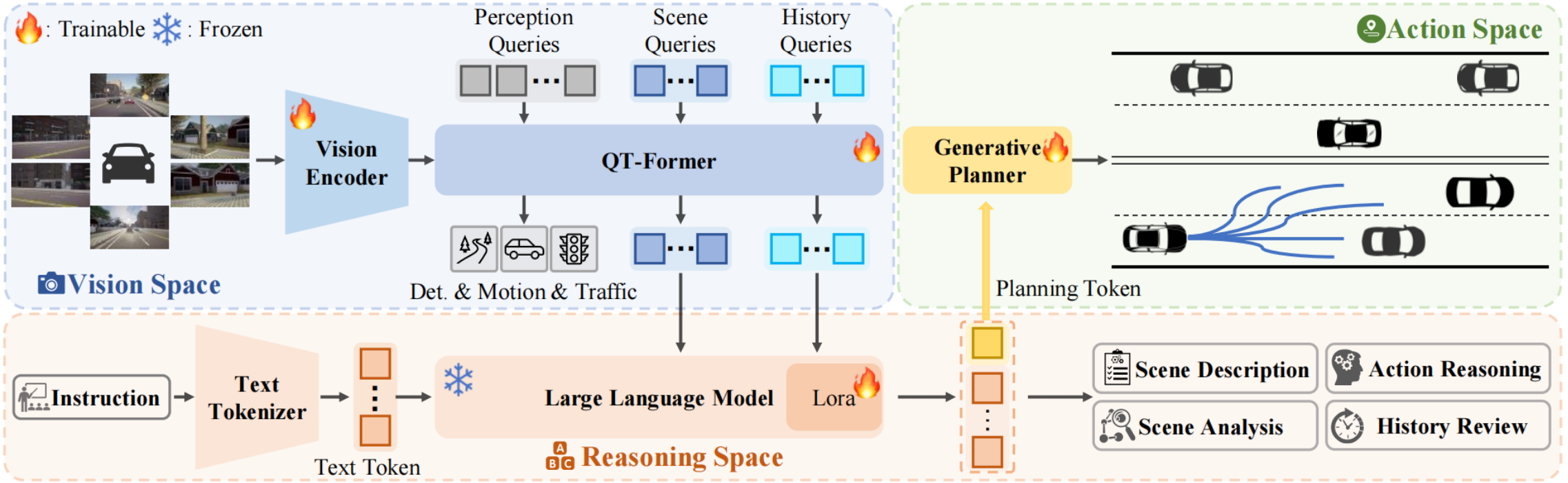
适配场景：直行无干扰等简单场景，无需复杂的语义推理
输出：物理动作令牌（无文本推理令牌）
优势：延迟低，占用内存少

- Slow Thinking

适配场景：绕行，多车博弈，无保护左转等。
输出：混合序列（文本推理令牌+动作令牌）

AutoVLA: A Vision-Language-Action Model for End-to-End Autonomous Driving with Adaptive Reasoning and Reinforcement Fine-Tuning

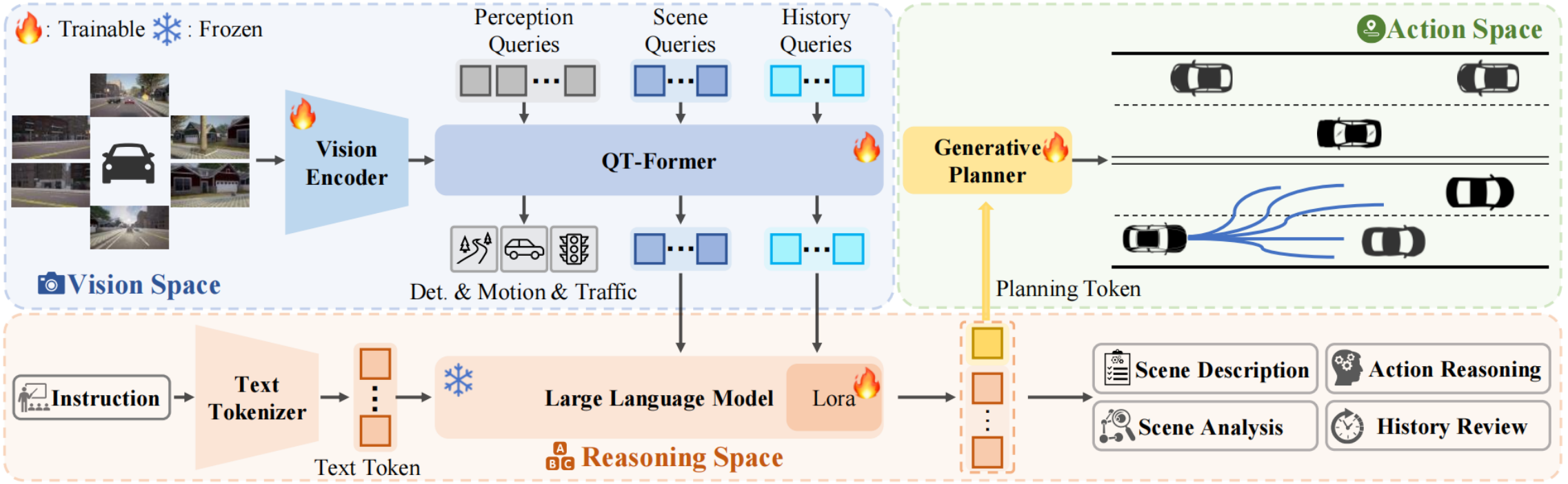
VLA--最新相关工作



- 视觉-语言模型的强大推理能力来指导自动驾驶的轨迹生成，并通过一个创新的生成式规划器（Generative Planner）弥合语义推理空间与数值动作空间之间的鸿沟，从而实现更安全、更智能的端到端自动驾驶。
- **性能突破**：它首次证明了将大型VLM深度融合进端到端驾驶框架可以带来闭环性能的质的飞跃，为自动驾驶的性能上限提供了新的答案。
- **框架创新**：提供了如何将VLM的通用能力与自动驾驶的专业任务相结合的全新范式，具有重要的启发性和可扩展性。
- **可解释性**：LLM的推理过程以自然语言呈现，使得自动驾驶系统的决策过程变得透明、可理解、可信任，这对于安全和调试至关重要。

Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation

VLA--最新相关工作



QT-Former：

查询式时空融合模块

场景查询：提取当前帧关键信息 → 图像特征

感知查询：多任务感知（目标检测，运动预测） → 自注意力

历史查询：结合长时记忆库，通过相对时间戳检索历史信息 → 历史Token

交叉注意力，生成场景Token

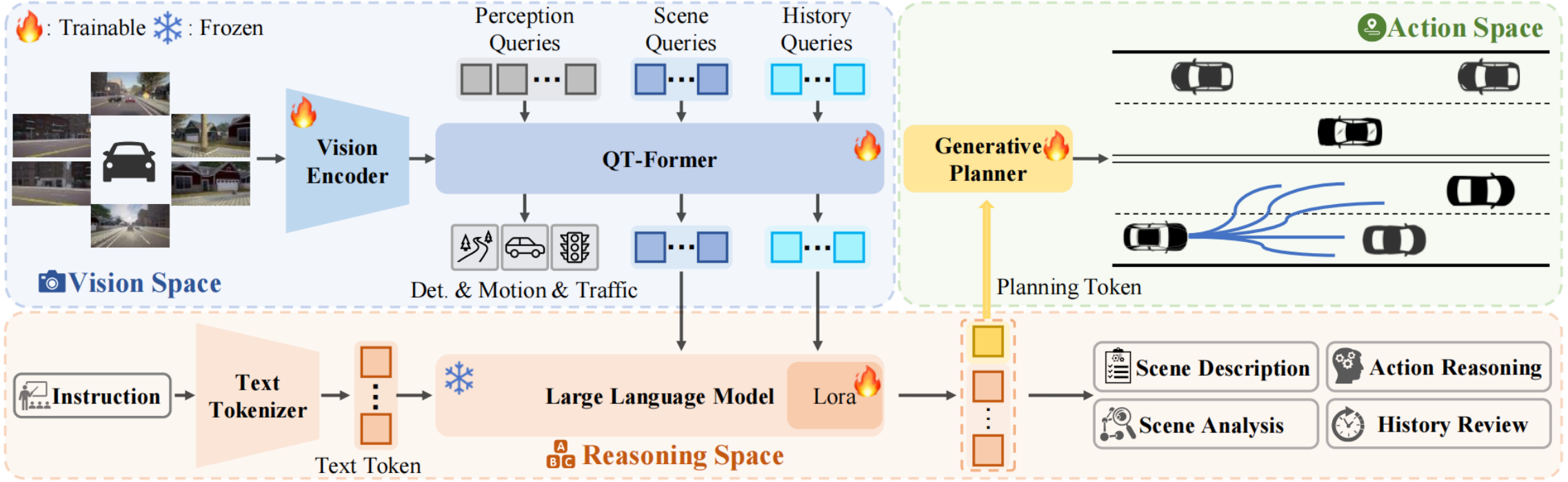
交叉注意力

映射到LLM

LLM

Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation

VLA--最新相关工作



LLM：驾驶场景推理引擎

- 在完成一系列问答后，LLM会输出一个特殊的 “规划令牌” 。这个令牌不是一个具体的指令，而是整个推理过程所有信息的浓缩和总结，代表了LLM对 “当前应该做什么” 的最终判断。

生成式规划器：弥合鸿沟的桥梁，最关键的创新

- 它成功地将**语义推理（LLM的输出）** 和**数值动作（轨迹坐标）** 在潜空间中对齐。LLM不再需要直接输出数值，而是输出一个高级的 “意图编码”，由生成式规划器将这个意图 “翻译” 成精确的轨迹。这就好比LLM是 “大脑” （决定做什么），生成式规划器是 “小脑” （精确地执行怎么做）。

Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation

世界模型 VS VLA模型

世界模型 VS VLA模型

对比维度	世界模型	视觉-语言-动作模型
核心目标	**预测未来**：学习环境的状态转移动力学。	**决策当下**：基于当前观测和常识，直接输出最优动作。
信息流	当前状态 + 智能体动作 → **未来状态**	当前观测 + 语言上下文 → **当前动作**
依赖能力	对物理世界的**动力学建模**能力。	对开放世界的**认知与推理**能力。
可解释性	较高。可以通过检查其预测的未来状态来判断其决策依据。	较高。**VLM的推理过程以自然语言呈现**，决策逻辑相对直观。
数据效率	理论上更高。学会动力学后，可通过内部推演应对大量未见过的场景。	相对较低。严重依赖大规模、高质量的标注数据（尤其是驾驶动作数据）。
实时性挑战	高。需要进行多次内部模拟推演才能做出一个决策，计算开销大。	相对较低。本质上是前向推理，一次前向传播即可输出动作。
对长尾问题的处理	潜力大。如果动力学模型准确，可以推演出罕见场景的演变。	依赖数据。如果数据中没有类似场景，模型可能无法正确反应。但大模型的强泛化能力有助于缓解此问题。

世界模型 VS VLA模型

对比维度	世界模型（World Model）	VLA 模型
目标	学习并预测车辆周围的“世界” —— 环境状态与未来演化	让车辆理解自然语言指令，并将视觉信息转化为行动
输入	视觉、激光雷达、车辆状态等传感器数据	视觉输入 + 语言指令（如“靠右车道行驶”）
输出	对环境的**未来预测**（轨迹、时空状态）或潜在世界表征	**控制信号或轨迹**（油门/方向盘）
关注重点	世界如何变化？（What happens next?）	我应该怎么做？（What should I do given human intent?）
任务类型	感知 + 预测 + 世界建模 + 规划辅助	指令理解 + 多模态对齐 + 决策生成
典型代表	UniAD、Dreamer、World Models、SeerNet、Tesla Occupancy Network	DriveGPT、VLA-Drive、GPT-Driver、LLaDA、DriveLM
是否依赖语言?	不依赖语言，偏底层认知模型	强依赖语言，是人机交互和高层决策接口
适用阶段	自动驾驶核心能力（环境感知/预测）	高级驾驶辅助 + 自然语言交互 + 高层规划

世界模型 VS VLA模型

1. 世界模型是 VLA 的“基础认知层”

- VLA 想完成“听懂 + 会做”，前提是必须“看懂世界”。
- 所以许多 VLA 模型会内置或调用世界模型（环境记忆与预测模块）。

2. 两者端到端自动驾驶路线

路线	E2E 自动驾驶方式
世界模型路线	感知→预测→（内部模拟）→规划→控制
VLA路线	视觉 + 语言 → 模态融合 → 输出动作控制

3. Tesla、NVIDIA 等厂商正在融合两者

公司	技术策略
Tesla	构建 Occupancy World Model（占据网格世界模型），同时发展 FSD + 语言指令（车内 GPT 助理）
NVIDIA	DRIVE Sim 中内置世界模型，同时推出 VLA API 让车回应语音/语言指令
Waymo/Google DeepMind	GeoSim 世界建模 + PaLM-E/LangChain for driving agent

世界模型 VS VLA模型

VLA优势

VLA模型的优势主要体现在与人类认知和交互的对齐上：

- **可解释性与信任：** VLA具备生成思维链解释的能力，极大地提高了系统的透明度，这是其在人机交互方面的重要优势。
- **语义灵活性：** VLA在将自由形式指令与自我中心视觉环境关联方面表现出色。
- **通用机器人潜力：** VLA模型本质上是通用策略，使其不仅适用于驾驶，还可用于更广泛的具身AI任务。

VLA挑战：延迟与安全保障

VLA的复杂结构带来了实际部署的挑战：

- **推理限制和延迟：** 集成多个复杂模态（视觉、语言）会显著增加计算复杂度，可能导致在实时、低级控制系统中产生延迟问题。
- **安全关键性与意外动作：** VLA的一个关键限制是可能因高风险行为或识别错误而造成物理伤害。标准VLA在导航和抓取过程中曾表现出危险行为，因此需要像SafeVLA这样的专业算法来明确集成安全约束，从而平衡任务性能和安全。
- **对抗性漏洞（LLM风险）：** 语言赋予VLA力量的同时，也引入了新的安全漏洞。VLA容易受到基于语言的攻击，例如“提示注入”或“角色扮演攻击”，这些攻击可能迫使模型绕过安全约束并执行不安全动作。由于VLA旨在遵循指令，恶意或模糊的提示可能劫持系统，构成灾难性的现实世界风险。

世界模型 VS VLA模型

WA优势

WA模型的核心优势在于其预测能力和对环境动力学的掌握：

- **预测稳健性：** WA模型通过评估反事实情况（即如果采取不同行动会发生什么），提供理性的规划，从而实现主动决策而非被动纠正。
- **通过模拟实现训练效率：** WA能够模拟逼真的驾驶环境（例如Waymo的SceneDiffuser++），允许在数十亿虚拟里程中进行快速测试、验证和训练，这对于捕获罕见或复杂的极端事件至关重要。

WA挑战：混乱、计算与反事实失败

虽然WA模型提供了卓越的预测能力，但其可靠性受到根本性数学限制的制约：

- **计算需求：** 世界模型架构对计算资源要求极高，无论是在训练还是实时执行阶段。这催生了 Waymo 的“教师/学生”模型等优化技术，以满足低延迟需求 。
- **混沌动力学问题：** 现实世界是一个复杂的非线性系统。有研究表明，在具有混沌和不确定性特征的环境中应用反事实推理，可能导致预测轨迹与真实结果之间出现**巨大偏差**。
- **推理的复杂性：** 逻辑反事实推理任务难以扩展，尤其是在系统复杂性增加时。

这种“可靠性悖论”表明，WA模型只能预测合理的未来，而不能保证真正的未来

世界模型 VS VLA模型

世界模型和VLA它们有很强的互补性，走向融合

- **VLA作为世界模型的“先验”**：一个精确的世界模型需要巨大的计算量来模拟所有物理细节。VLM所具有的丰富常识可以作为一个强大的**先验知识**，帮助世界模型进行更高效、更准确的预测，避免在无关紧要的细节上浪费算力。
- **世界模型作为VLA的“推演工具”**：目前的VLA主要基于**当前和过去**的信息进行决策。如果能集成一个轻量级的世界模型，VLA就可以在输出决策前，先进行一步快速的“思想实验”（“如果我执行这个动作，周围车辆会如何反应？”），从而做出更长远、更安全的规划。这将使VLA从“直觉反应”升级为“深谋远虑”。
- **统一于“理解”之下**：无论是预测动力学还是进行常识推理，最终目的都是让机器**更好地理解驾驶场景**。一个真正智能的驾驶系统，必然同时具备深厚的物理世界理解和丰富的语义世界知识。

世界模型 VS VLA模型

混合架构：VLWM的诞生

未来的基础模型，通常被称为视觉-语言-世界模型（VLWM）或3D VLA生成式世界模型，旨在结合VLA的语义推理能力与WA的预测能力 。

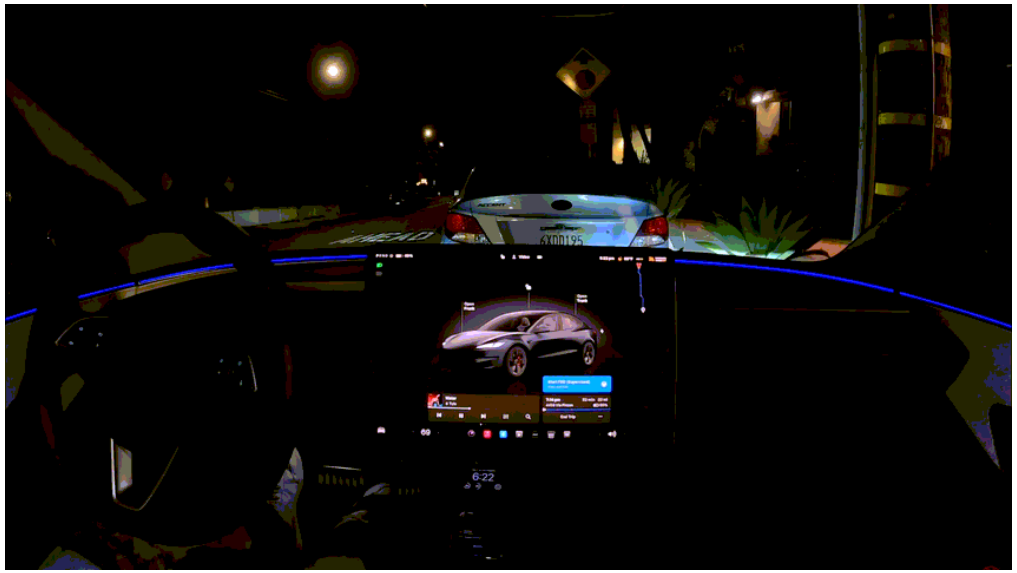
VLWM的关键进步在于结构化未来预测。它不生成嘈杂、高容量的原始视频预测，而是预测未来世界状态的抽象表示（例如，目标、交织的动作和状态变化，以“标题树”等结构化文本表示）。这使得模型能够同时具备快速反应式动作和通过基于语义距离的成本最小化进行反思式、战略性规划的能力。

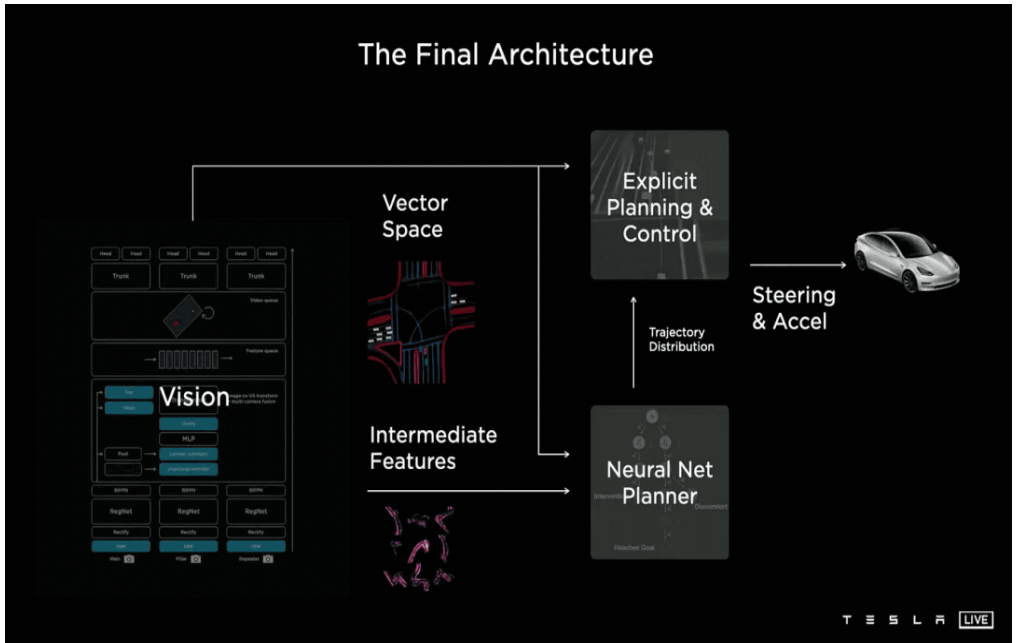
架构预测

目前，虽然世界模型（WA）在工业界部署中占据主导地位，但在推理、稳定性和可解释性方面仍面临挑战。因此，最稳健的L4/L5系统将很可能是一个以**世界模型为核心**（提供卓越的动力学预测和环境鲁棒性），并辅以**VLA接口**（提供卓越的可解释性、人机交互和高级语义推理）的混合体。VLA组件是实现人机交互和监管机构接受所必需的矫正机制。

企业技术案例分析

特斯拉--传统模块化过渡到端到端



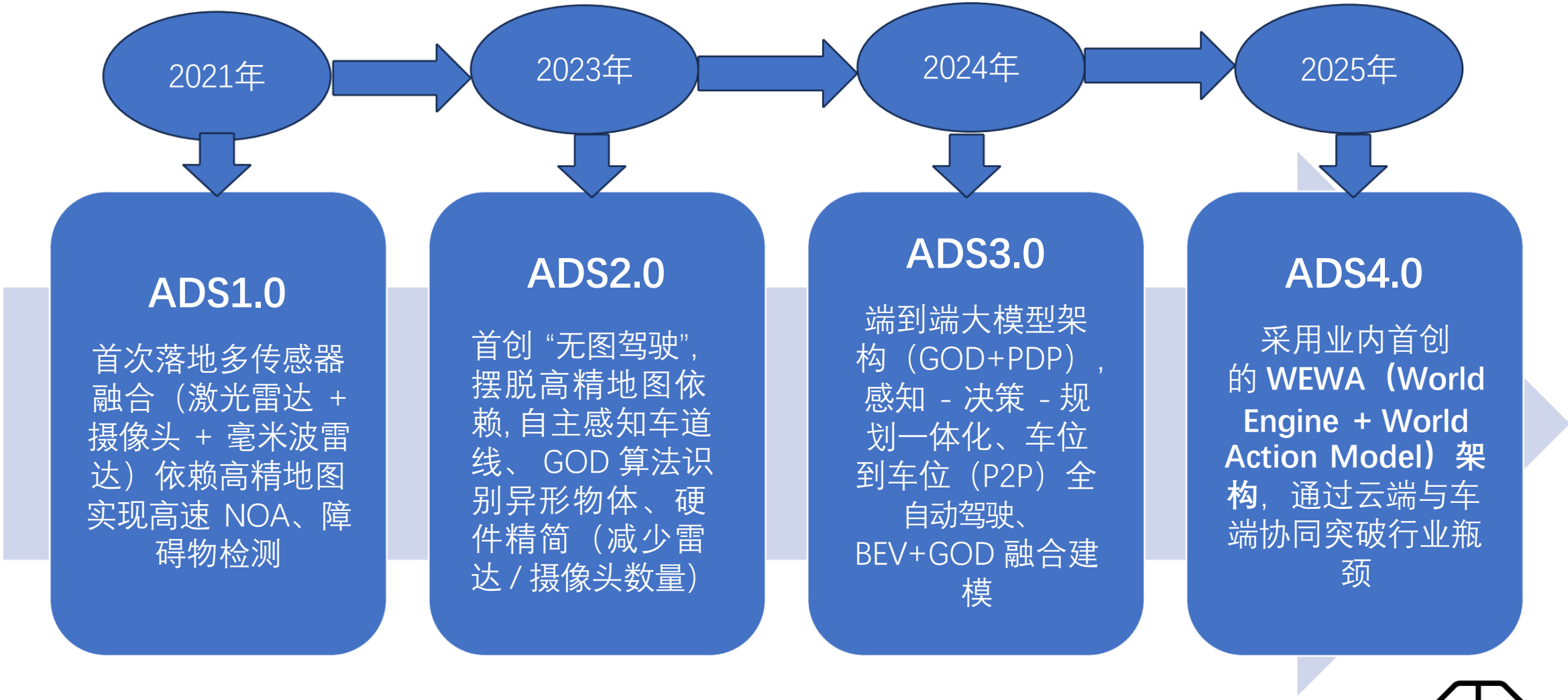


2014-2020	2020-2021	2021 AI Day	2022 AI Day	2024 FSD V12
传统模块化结构 Autopilot基础辅助驾驶	模块化+模仿学习 FSD Beta	模块化端到端 Transformer融合多个摄像头映射到3D BEV空间	模块化端到端 将环境划分为体素并预测占用概率	模块化端到端 完全端到端，移除手写规则逻辑

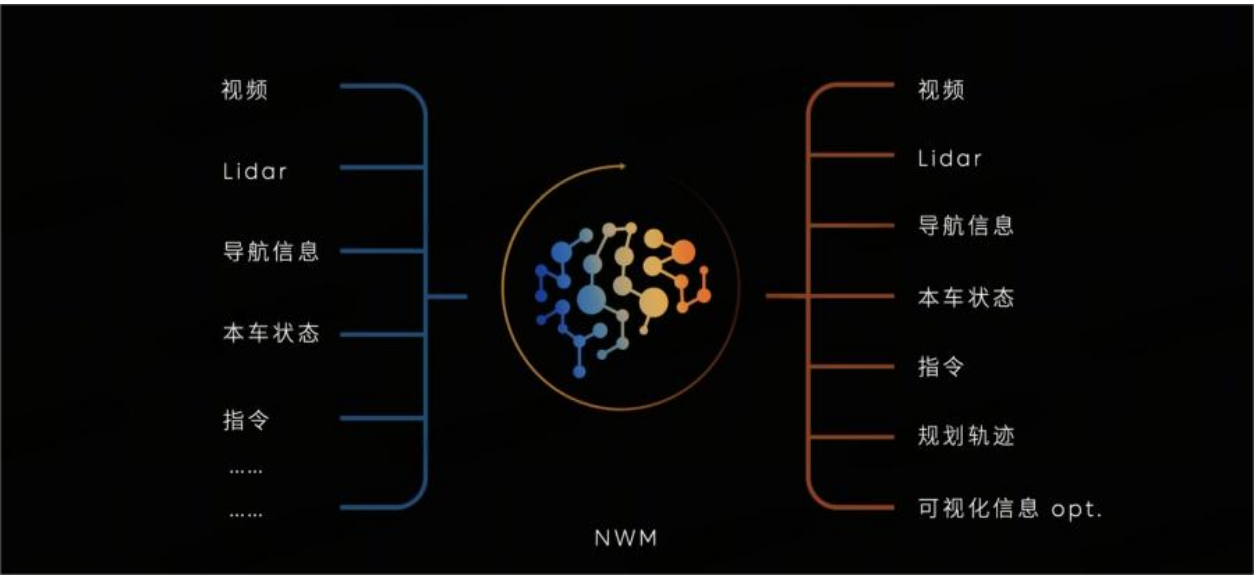
FSD V12 标志着从传统规则式自动驾驶向神经网络端到端驾驶策略的转变：系统通过单一深度学习模型，从多摄像头视频直接输出方向、油门和刹车控制，几乎完全移除了手写逻辑。

FSD V14（2025 年）则在此基础上进一步引入更强的时空 Transformer 架构与自监督闭环训练，实现持续自我学习与场景泛化能力提升，使车辆在复杂城市环境中的决策更加自然、接近人类驾驶。

华为--世界模型


2021年
2023年
2024年
2025年
ADS1.0
首次落地多传感器融合（激光雷达 + 摄像头 + 毫米波雷达）依赖高精地图实现高速 NOA、障碍物检测
ADS2.0
首创 “无图驾驶”，摆脱高精地图依赖，自主感知车道线、 GOD 算法识别异形物体、硬件精简（减少雷达 / 摄像头数量）
ADS3.0
端到端大模型架构（GOD+PDP），感知 - 决策 - 规划一体化、车位到车位（P2P）全自动驾驶、BEV+GOD 融合建模
ADS4.0
采用业内首创的 WEWA（World Engine + World Action Model）架构，通过云端与车端协同突破行业瓶颈

蔚来--世界模型



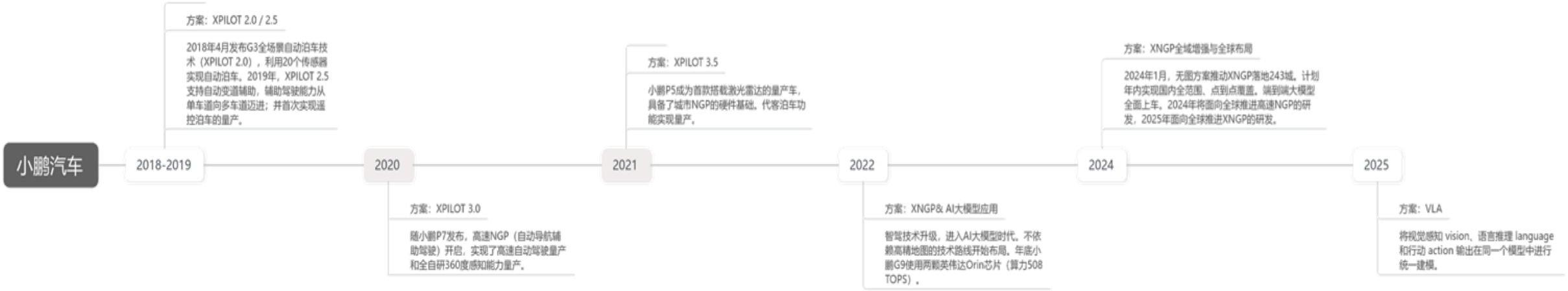
蔚来于2024年创新科技日发布中国首个智能驾驶世界模型 NWM（NIO WorldModel），即可以全量理解信息、生成新的场景、预测未来可能发生的多元自回归生成模型。

NWM可以在100毫秒内，推演216种可能发生的轨迹、寻找最优路径；还能基于3秒钟视频的Prompt输入，生成120秒想象的视频。

NWM相比常规端到端架构，实现了三点优化：

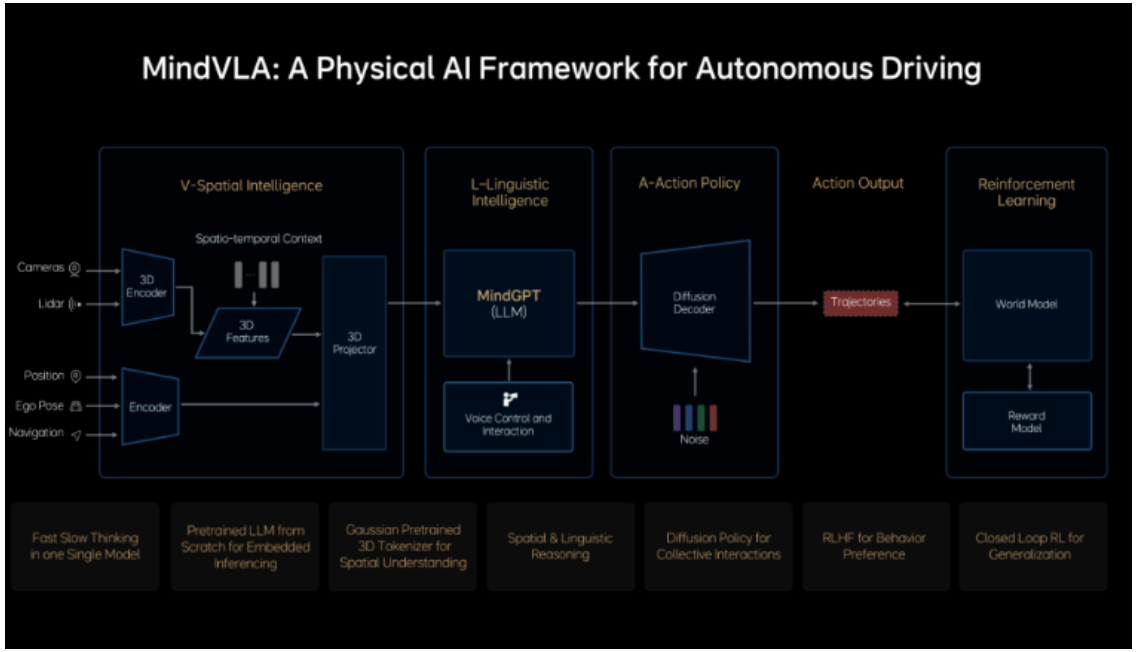
- 全量理解信息，空间认知能力更强
- 能够预测接下来的场景
- 生成式无监督的方式，对数据的利用更加高效

小鹏--VLA



小鹏的技术架构由三大核心模块构成：神经网络XNet、规划大模型XPlanner和大语言模型XBrain。三者协同运作，使自动驾驶系统能够快速适应多样化场景，并通过持续的数据迭代不断提升智能水平。

理想--VLA与世界模型



理想的MindVLA将空间智能、语言智能和行为智能进行了整合，并通过独特设计的预训练和后训练方法，实现了优秀的泛化能力和涌现特性。

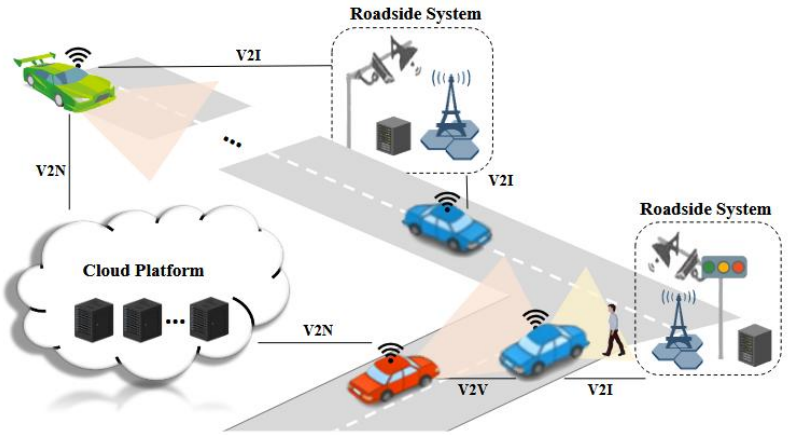
MindVLA采用世界模型进行强化学习，基于真实驾驶道路数据生成大量的多变的道路数据。

未来展望

未来展望--协同感知端到端

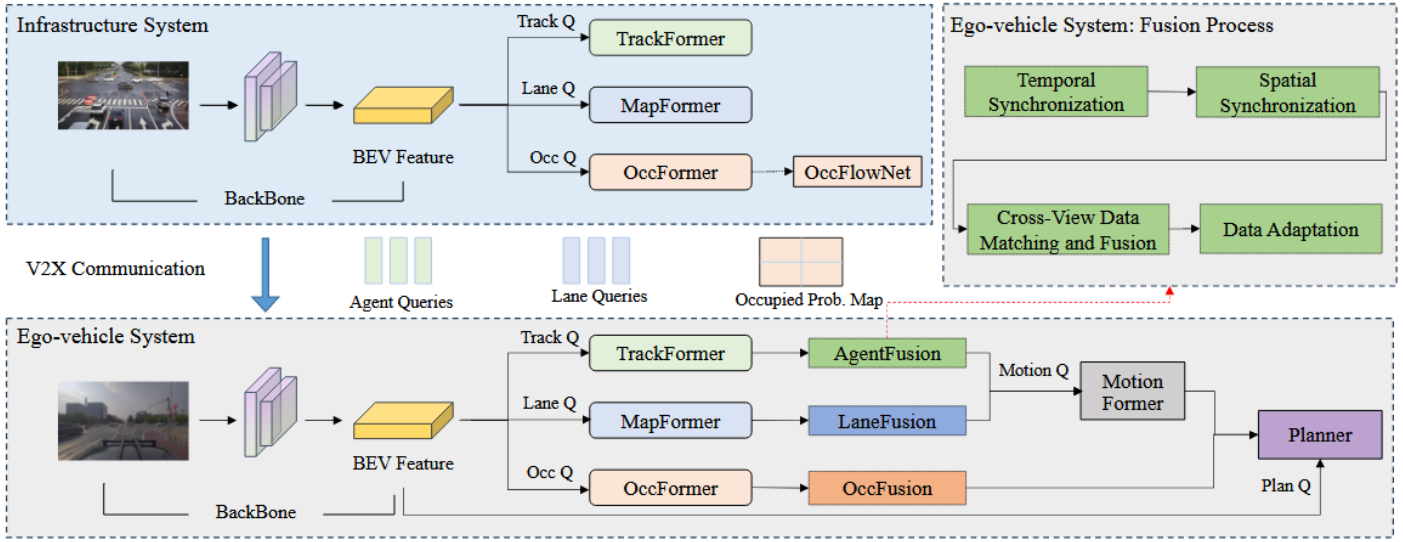
全面地感知周围环境是自动驾驶的一项基本任务。但由于从单一观察角度获得的信息有限，单车感知仍然面临重大挑战。在**车联网**(V2X)通信的帮助下，**协同感知**提供了一种有前途的解决方案。

协同感知旨在通过多个智能体之间的信息交换来解决单车自动驾驶系统固有的局限性。



Towards Vehicle-to-everything Autonomous Driving: A Survey on Collaborative Perception

V2X场景示意图

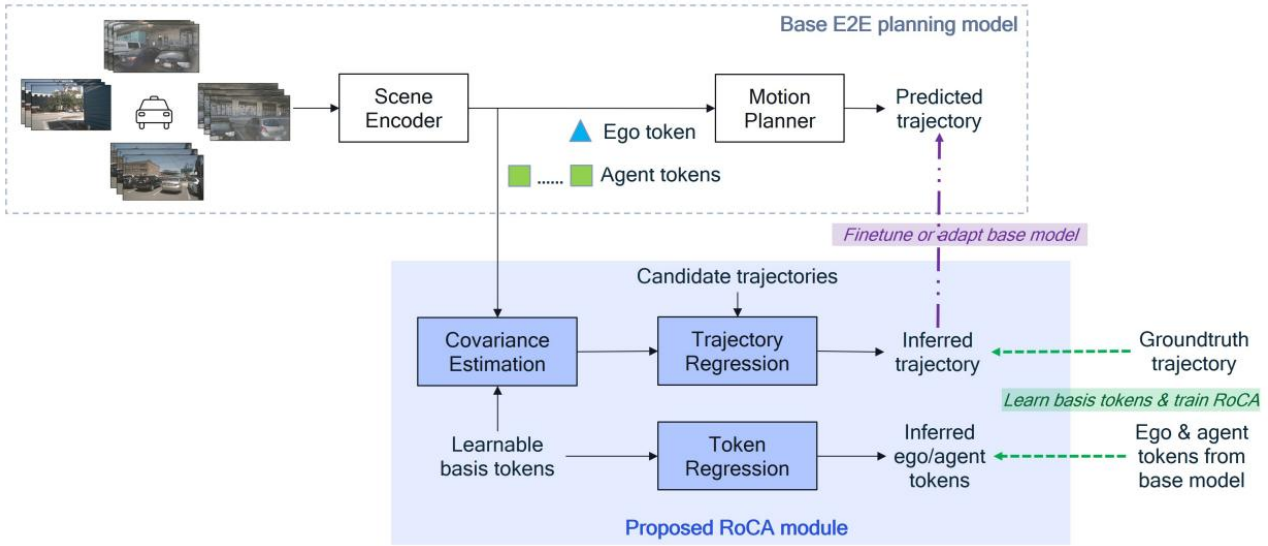


End-to-End Autonomous Driving through V2X Cooperation

UniV2X是首个用于车辆与基础设施协同自动驾驶的全栈协同端到端自动驾驶框架。它将协同感知、中间表示学习、占用预测和规划等多个关键模块集成到一个协调的架构中，通过V2X通信获取并充分利用路端或其他车辆传感器传输的数据，实现了从感知到规划的端到端学习。

未来展望--零（少）样本学习

自动驾驶模型难免遇到超出训练分布的真实场景。因此，需要将端到端驾驶的领域自适应任务形式化，并融合零样本和小样本学习技术，让模型在标注数据有限甚至缺失的情况下，成功适应未知领域。

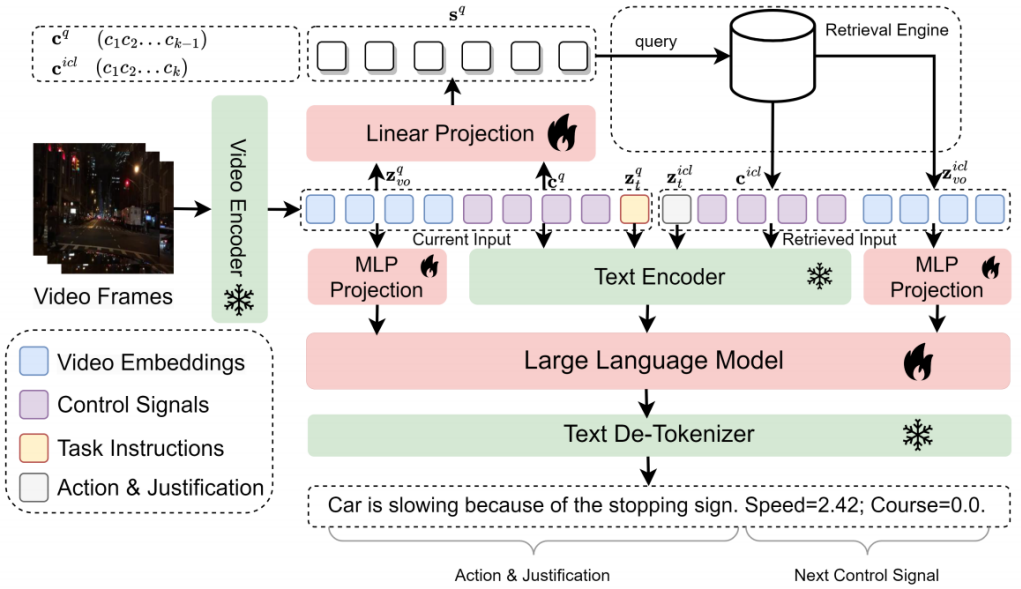


RoCA: Robust Cross-Domain End-to-End Autonomous Driving

RoCA模块通过“令牌优化 + 概率建模 + 轨迹优化”的逻辑，弥补基础端到端模型在鲁棒性、适应性上的不足，让自动驾驶系统能更可靠地应对复杂、多样的行驶场景。

未来展望--跨模态智能交互

未来的系统必须将手势、声音和标识作为“语言”通道的一部分进行解析。例如识别警察的手势或行人的挥手，然后产生明确的、人类可读的回应（灯光、显示器、喇叭）。



RAG-Driver: Generalisable Driving Explanations with Retrieval-Augmented In-Context Learning in Multi-Modal Large Language Model

RAG-Driver提出了一条途径：将现场感知与符号规则和上下文融合到基础非语言线索中。

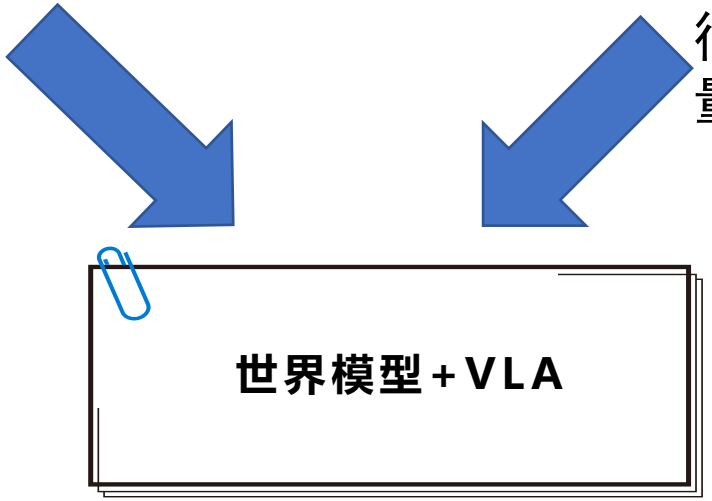
未来展望--世界模型与VLA结合

世界模型

优势：模拟环境动态，长期预测与规划能力，可减少真实交互成本。
劣势：训练复杂，高维建模存在挑战，学习效率低。

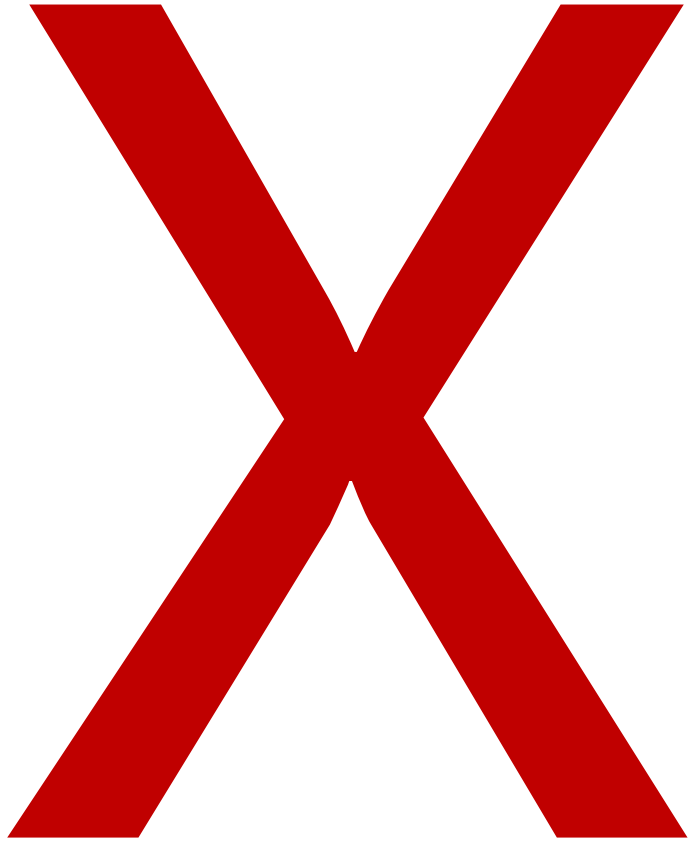
VLA

优势：跨模态理解能力，强泛化能力，学习效率高。
劣势：缺乏对环境物理的深层建模，行为一致性与可解释性弱，依赖大量数据。



- 用世界模型提供可预测的环境反馈，增强VLA的决策一致性。
- 用VLA的语义理解与任务泛化能力指导世界模型学习重点。
- 构建“语言驱动的世界模拟器”，实现端到端可解释的自动驾驶系统。

未来展望



谢谢大家！
请批评指正！

jiawei@hfut.edu.cn