**理想贾鹏GTC 2025 VLA**

出处：

<https://www.bilibili.com/video/BV11yX5Y9EEj/?vd_source=115911bd71b74bfcc0cad43e576887e4>

首先我会花几分钟和大家一起看一下自动驾驶在中国面临的挑战。

其次会总结一下2024年我们如何使用端到端+VLM双系统去尝试解决这些挑战的，以及最终呈现给用户什么样的产品体验。

第三，我会详细介绍我们的最新思考，以及在VLA技术上的突破。

最后会通过几个实车demo来展示一下VLA技术可能给用户带来哪些革命性的产品体验。

在开始深入探讨之前，我想先和大家一起来看一下自动驾驶在中国面临的独特挑战。作为全球自动驾驶的典范，在tesla 2024的财报会议上，马斯克也提到FSD在中国的落地上也遇到了很大的挑战。其中之一就是复杂而多样的公交车道，的确除了到处乱窜的电动车和高强度的人车博弈外，中国的道路结构本身也足够复杂。为了提升出行效率，公交车道被广泛使用。然而，各地的标识方法和使用规则非常多样，采用了比如地面的文字标识、空中指示牌或者路边标牌，同时会以不同的文字形式说明这些车道的时段限制。这些多变的规则和文字表达为自动驾驶带来了巨大的挑战。

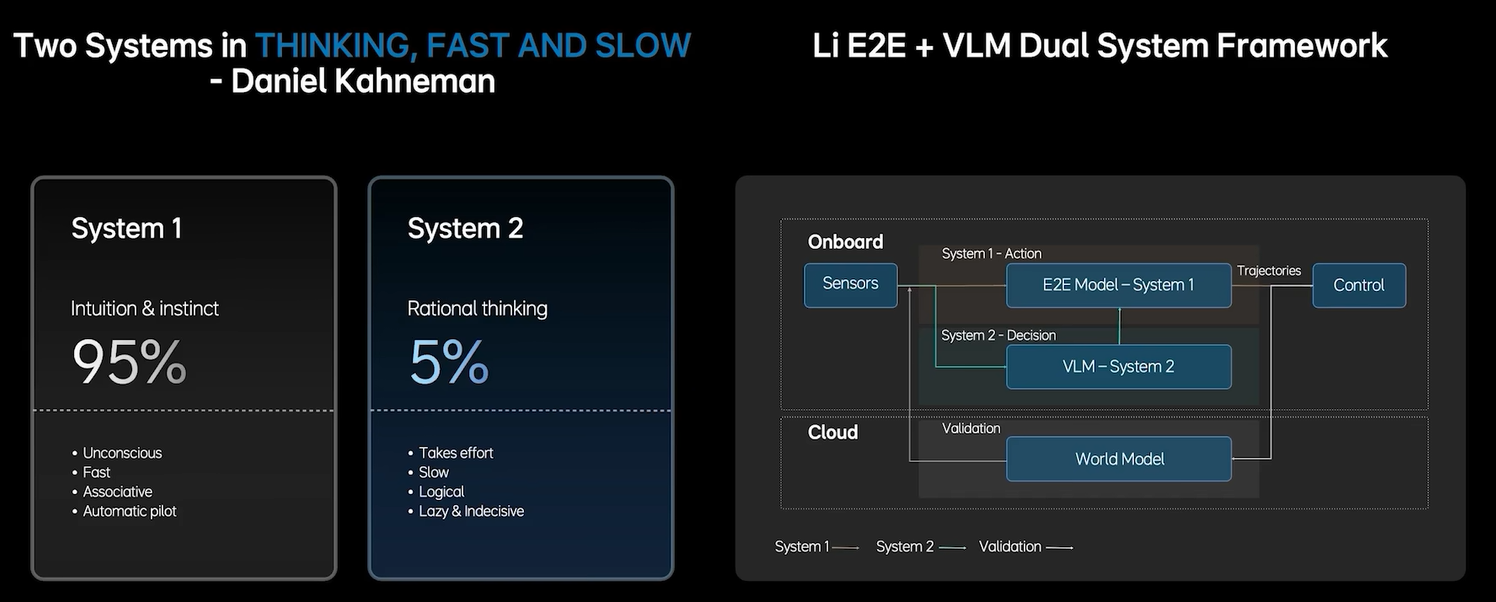


有朋友可能会说，能够通过地图或者先验信息来解决，但其实挑战也很大。中国城市快速发展，常常出现新增的公交车道，或者因为施工所导致部分标识的不清晰或者重刷，这样的话，任何先验信息都会面临着鲜度不足的问题。如果想从根本上去解决公交车道的挑战，车端需要具备实时识别和理解这些文字的能力。

公交车道还只是诸多挑战之一，随着数字城市的不断深入，很多车道和区域被赋予了动态变化的能力，越来越多的城市出现了动态可调的可变车道和潮汐车道，同时为了充分利用入口的空间，也增加了如待选区、待行区，这些车道和待行区域的进入时机也是有多样化的信号灯或者LED文字牌来控制的。同时咱们城市的建设日新月异，每天都会面临着设备的新增、故障和维护。自动驾驶系统需要时刻保持对这些变化的理解，,如果想打造一套无缝的点到点的智能驾驶体验，自动驾驶车辆还需要能够顺利通过ETC和收费站，从而也要求系统能够识别和理解全国各地的各种各样的ETC标识和支付方式标识。

综上所述，在中国，自动驾驶系统不仅要应对高强度的人车博弈，还需要能够读懂文字，,以及具备常识和很强的逻辑推理能力。特斯拉向中国用户推送了FSD的功能，我们也看到FSD的公交车道、待行区等特殊场景上的表现确实是有所不足的。

那么，理想汽车是如何应对这些挑战的呢？在去年的GTC 2024大会上，我有幸介绍了我们的自动驾驶框架。

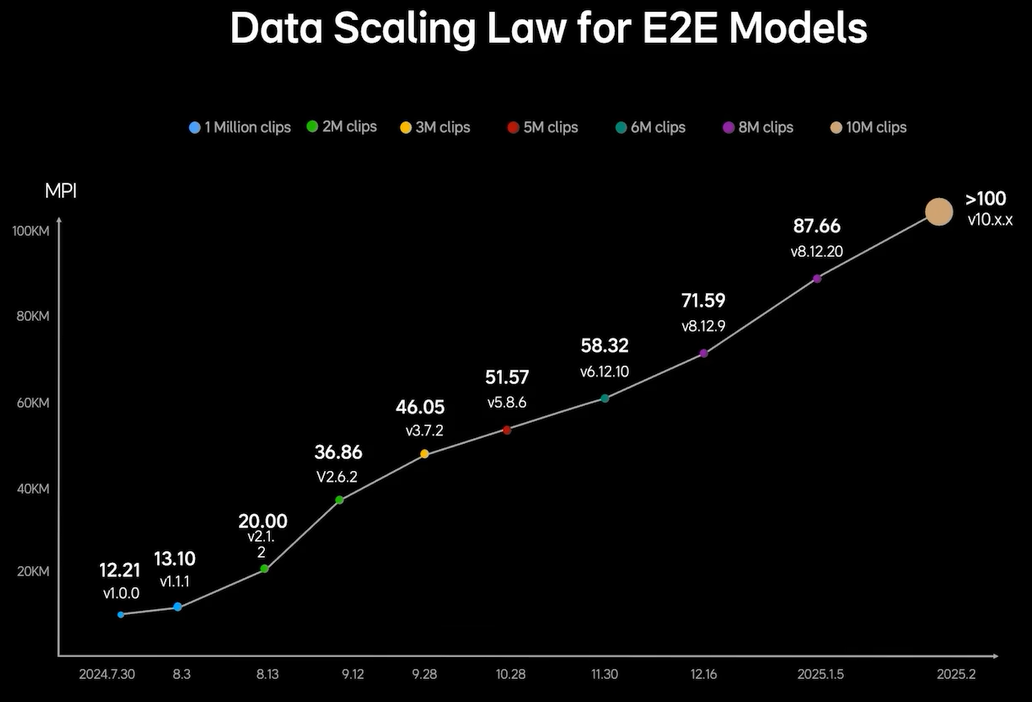


这个框架是基于诺贝尔奖获得者丹尼尔卡尼曼提出的快慢思考双系统理论。简单地说，人的思维可以分为两个系统，快思考系统1和慢思考系统2。快系统依赖于直接判断，大多数情况下，人类日常角色都使用该系统；而当我们面临复杂问题时才会调用我们慢系统去想一想、思考一下再行动。

我们也在车端实现了这样的双系统。首先，车端通过端到端的模型实现了快思考，这个单一的模型实现了根据传感器的输入来直接输出轨迹。类比于人类的直觉反应，该系统通过模仿人类的驾驶行为来应对各种各样的场景，完全基于数据驱动，中间无需人为设定规则，而且不使用任何高精地图或者先验信息，故该系统的训练和执行效率都很高。另一方面，慢系统则依托于一个22亿参数规模的VLM，在需要文字理解能力、常识和逻辑推理的场景中，VLM会通过思维链COT进行复杂的逻辑分析，给出驾驶决策并指挥快系统去执行。端到端模型和VLM模型分别跑在一颗Orin-X芯片上。这套双系统的一些技术细节，我们也发表在了CORL 2024的Drive VLM这篇论文中：

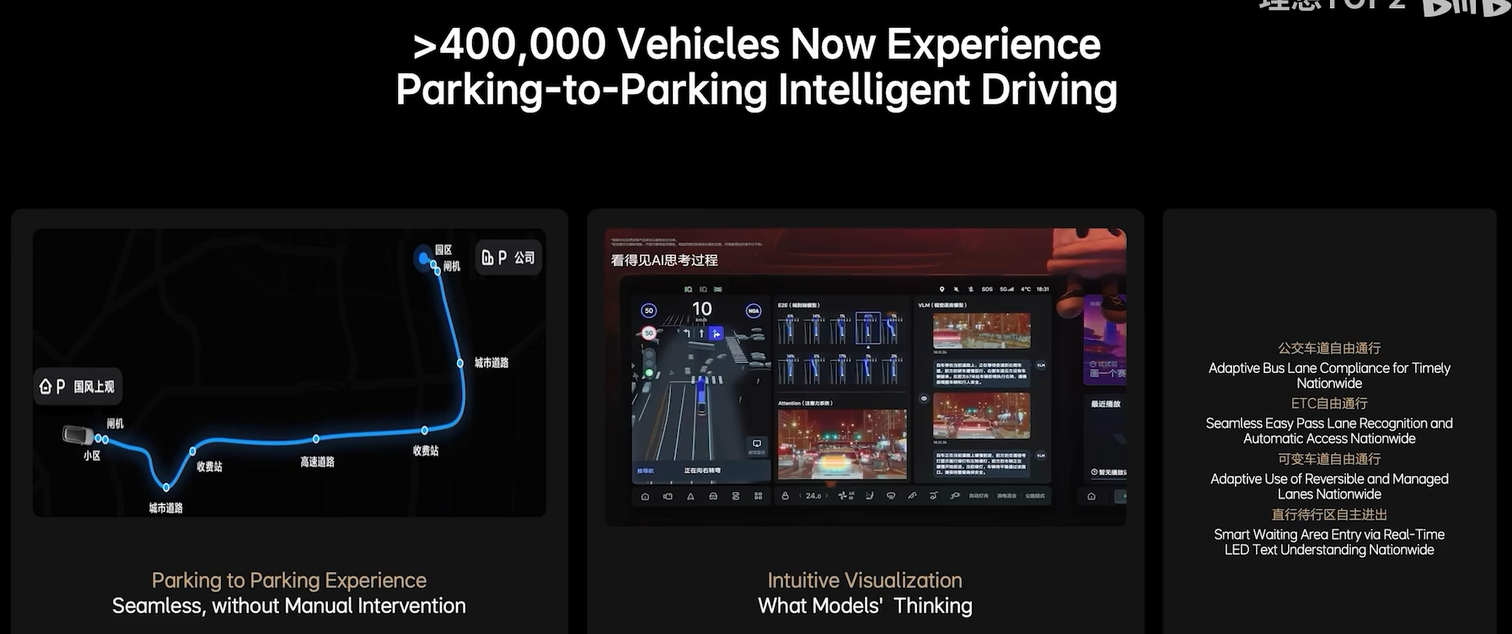
<https://arxiv.org/abs/2402.12289>

系统采用了数据驱动的范式，从而提高了迭代效率。在过去半年的量产实践中，我们实现了每周3~4次的模型发版。然而，如何快速地进行测试和验证也逐渐成为了我们的瓶颈。考虑到中国拥有数百万公里的道路，通过实车来进行测试的话，这样做既不现实，而且效率和成本也难以接受。因此，我们在云端构建了一个世界模型，它提供了一个3D的交互环境，使得双系统可以在此环境中进行高效的闭环仿真测试。以上三个模型就构成了理想自动驾驶的整个体系，简洁而高效。在这一过程中，我们非常惊喜地发现，双系统同样遵循scaling law：



随着训练数据的增加，模型在实际驾驶场景中的表现逐步提升。

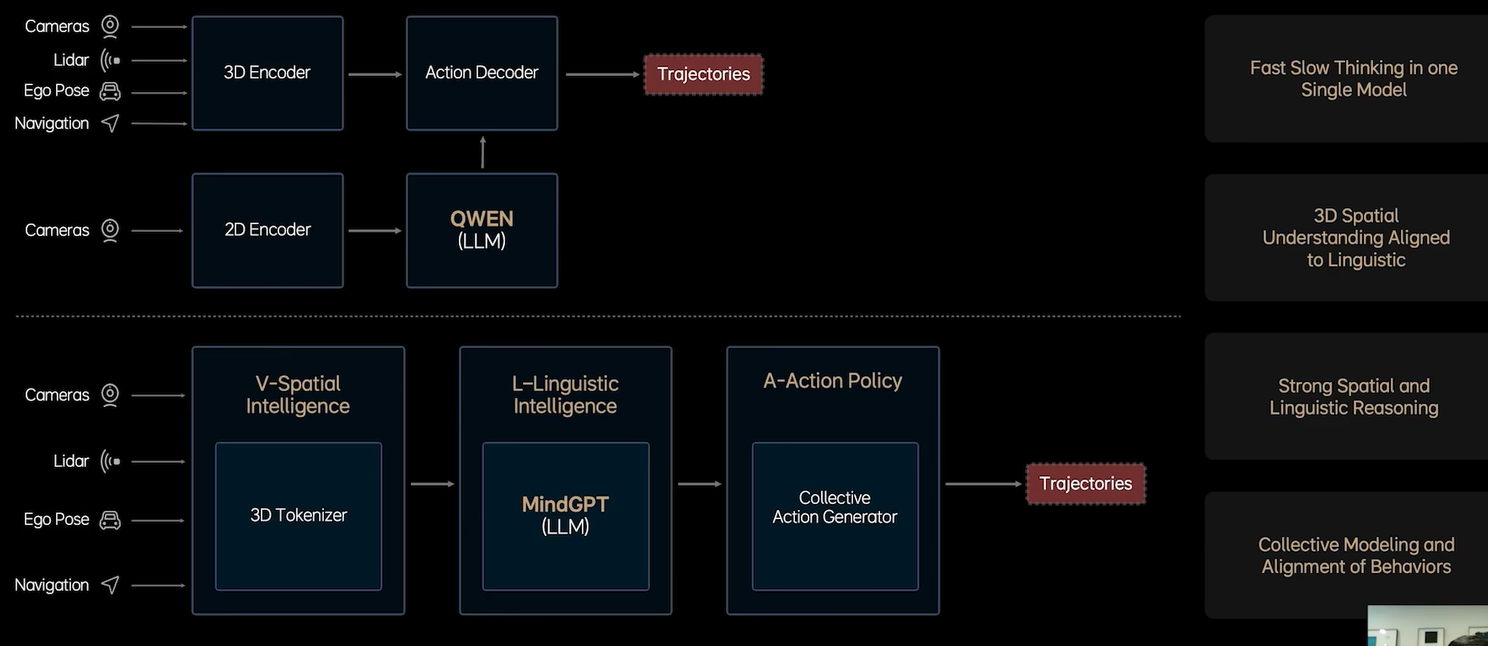
基于双系统架构，我们成功在英伟达Orin-X平台上推出了全球首个车位到车位的智能驾驶产品，目前已经推送了超过40万台车。所谓车位到车位，指的是从停车位出发，穿越车库经过园区进入公共道路，直到通过ETC进入高速公路的整个过程中，没有任何的系统降级和退出。



同时，借助VLM的通识能力，我们在这一产品中实现了多个行业首创的功能，例如，实现了不依赖于地图或者先验信息的全国ETC自由通行；全国潮汐车道和可变车道的自由通行；在待转区、待行区的自主进出，以及在坑洼路面积、水路面积、水路面的自动减速。这些突破性功能目前在其他车上都尚未实现，极大地提升了用户体验，赢得了广泛的好评，也推动了我们的销量在持续增长。

同时，我们知道用户可能会担心端到端模型是个黑盒子。比如说，有的用户会想知道模型在想什么，下一步的动作是什么。为了解决这一问题，我们创新性地引入了AI推理交互功能。用户可以直观地看到模型在关注哪些点、即将执行的动作以及整体的思考过程。这个透明化的设计显著提升了用户的信心，让用户用得更加安心、放心，也得到了大量的积极反馈。

在2024年10月理想汽车量产双系统方案后，端到端+VLM的模式逐渐成为了行业的标杆。许多同行开始采用这一路线，不仅在自动驾驶领域，在通用机器人领域也得到了应用。例如上个月Figure.AI发布机器人系统也使用了类似的架构。在过去几年的实践中，自动驾驶团队形成了一种高效的工作模式。

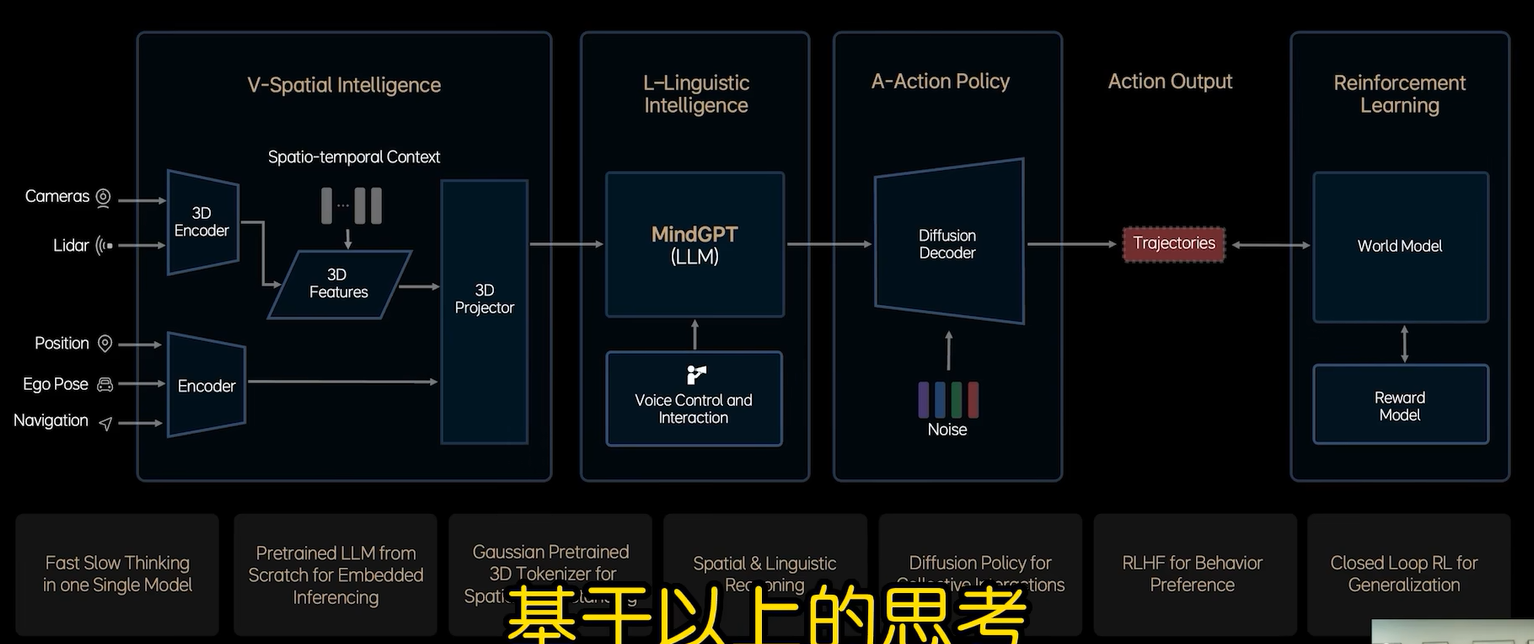


在产品交付的同时，我们不断反思和总结现有架构的不足，并留出一部分资源用于探索下一代技术。在端到端+VLM的量产过程中，我们也发现了一些需要改进的地方。首先，虽然我们可以通过异步联合训练，让端到端和VLM协同工作。但由于它们是两个独立的模型，而且运行于不同的频率，整体的联合训练和优化是非常困难的。其次，我们的VLM模型是基于开源的LLM，它使用了海量的互联网二级图文数据来进行预训练，而在3D空间理解和驾驶知识方面是有所不足的。虽然可以通过后训练进行一定程度的弥补，但是它的上限还不是很高。第三，自动驾驶芯片如Orin-X和Thor-U，它的内存带宽和算力是不及服务器GPU的。如何进一步提升模型参数量和能力的同时还能够保证高效推理的实现，这将是一个巨大的挑战。第四，目前驾驶行为的学习更多地依赖于transformer的回归建模，但这种方法难以处理人类驾驶行为的多模态性。这里的多模态性是指在相同的场景下，不同人的选择是不同的。即使是同一个驾驶员不同心情的时候，驾驶行为也是不同的。

那么有什么方法能够进一步提升模型的上限，让用户体验到更加丝滑的产品呢？同时，我不仅要提升模型的上限，而且还要提升模型的下限。如何对齐到人类的价值观也是急需解决的问题。

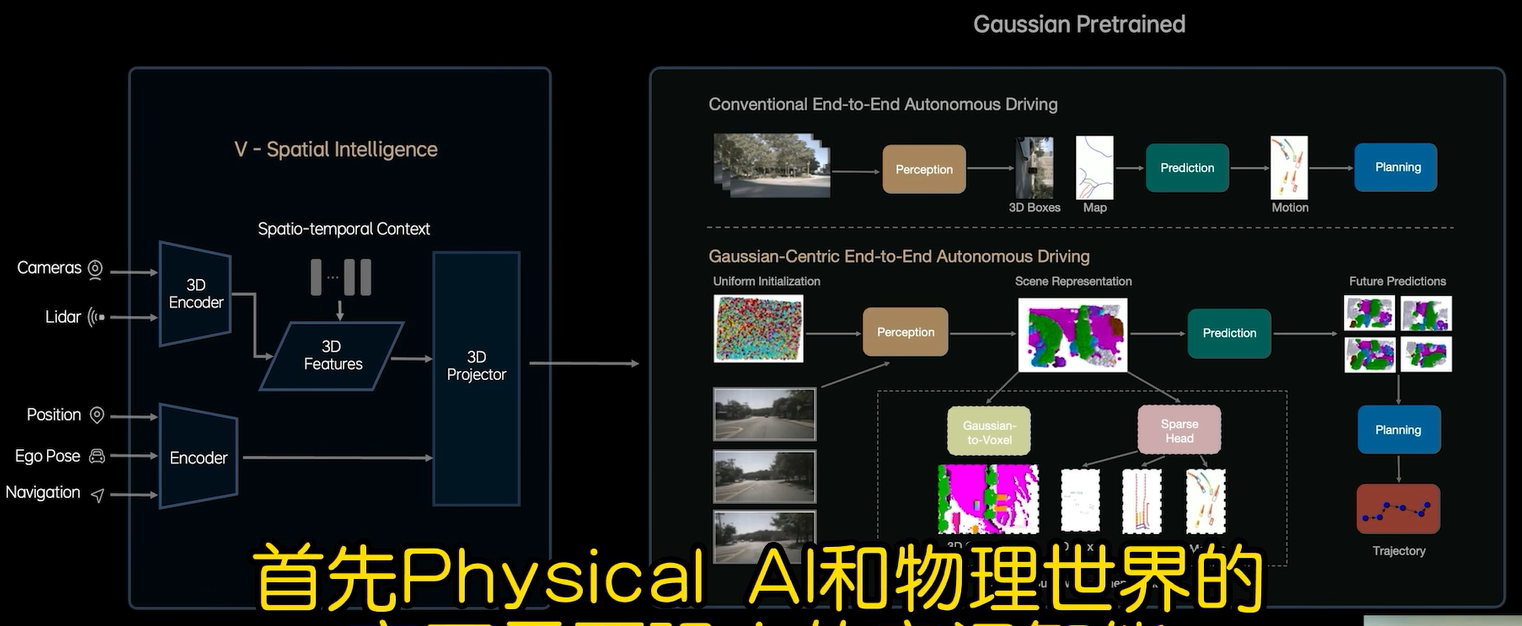
幸运的是在我们交付端到端+VLM期间，空间智能、LLM、AIGC和具身智能都有了快速的发展，也给了我们很多的启示。我们在思考能否将端到端模型和VLM模型合二为一，像GPT o1和DeepSeek R1一样，模型自己学会快慢思考，同时赋予模型强大的3D空间理解能力和行为生成能力，将双系统天花板进一步打开。

基于以上的思考，我们提出了理想的VLA模型MindVLA，其中，VLA是视觉语言行为大模型，它将空间智能、语言智能和行为智能统一在一个模型里。VLA是physical AI的最新范式，它赋予自动驾驶这样的物理系统感知思考和适应环境的能力。MindVLA不是简单地将端到端模型和VLM模型结合在一起，而是全新设计了所有的模块。



3D空间编码器通过语言模型和逻辑推理结合在一起后给出合理的驾驶决策并输出一组action token，最终通过diffusion进一步优化出最佳的驾驶轨迹，其中，action token是周围环境和自车驾驶行为的编码。整个模型推理过程都要发生在车端，而且要做到实时运行。

接下来我会从六个关键技术点去详细介绍MindVLA的设计和训练过程中，包括MindVLA强大的3D空间理解能力是如何获得的？我们是如何从0开始设计和训练语言模型，使其具备驾驶知识和推理能力的。Diffusion是如何与语言模型结合在一起的，以及我们是如何解决VLA在车端芯片上的实时推理。好，咱们一起看一下这六个关键技术。

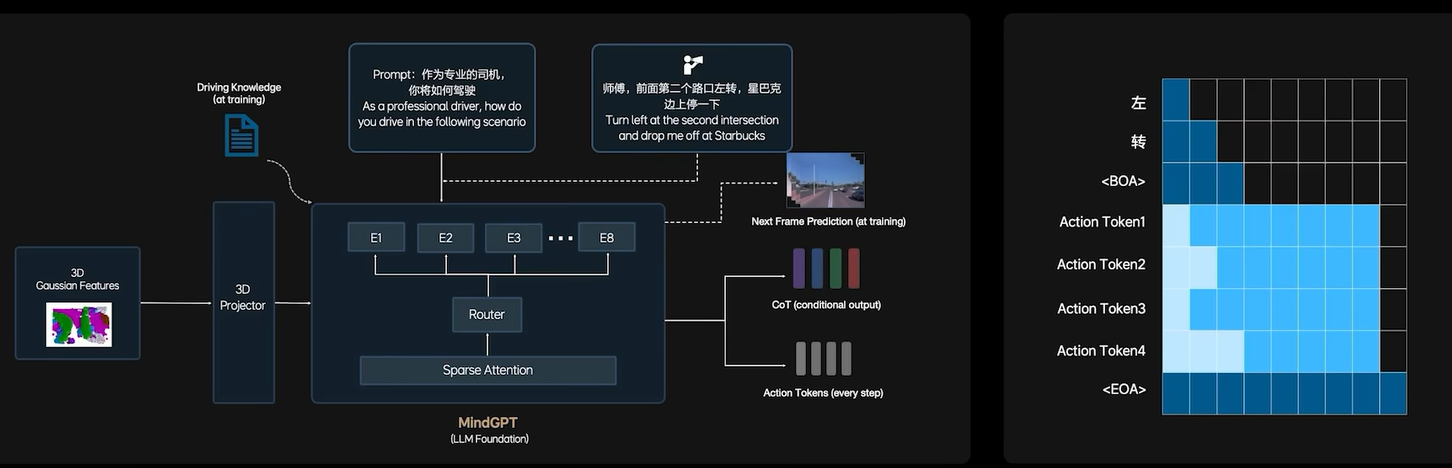


首先Physical AI和物理世界的交互需要强大的空间智能，也就是对3D物理世界的感知、理解和推理能力。回顾自动驾驶技术的发展历程，对空间的理解，我们经历了从单目2D特征到单目3D特征，再到多相机的鸟瞰图BEV特征和占用网格特征等不同的阶段。这个演变过程也体现出咱们工程师对物理世界几何和语义信息精确提取的不懈追求。然而这些方法大多依赖于监督学习，需要精确的标注数据，从而效率和数据利用率都很低下。

我们前面提到在双系统的实践中，我们欣喜地发现了数据量的能力。随着数据量的提升，系统的表现会同步提升。如何充分利用我们海量的数据？在研发世界模型时，我们发现3D高斯是一个极其优良的中间标准。它不仅具备出色的多尺度3D几何表达能力，同时也能够承载丰富的语义。最为关键的是可以通过图片RGB进行自监督训练，这使得我们有机会通过充分利用海量的真实数据来获得一个优秀的3D表征。我们的研究成果显示，采用自监督训练得到的高斯表征，能够极大地促进下游任务的性能提升，部分实验结果和能力可参见我们的论文Gaussian AD（https://arxiv.org/abs/2412.10371）里，大家如果有兴趣可以进一步查阅。

解决了3D表征后，接下来需要考虑的是如何将它和语言智能结合在一起呢？

LLM已经被证明是一个强大的通用模型，它可以兼容视觉、语言等多种模态。但是如果想要LLM同时具备3D的空间理解能力、3D的空间推理能力及强大的语言能力，那么就需要在模型的预训练阶段就要加入大量的相关数据。然而，车载芯片如Orin-X和Thor-的算力和内存带宽都有限，因此，如何设计模型架构，让模型参数进一步提升，还能在有限的资源下实现实时推理，这里解释一下为什么我们还要进一步增加模型参数量，因为数据参数规模和能力强弱可以画等号，越大越好。为了解决这个问题，我们需要从零开始设计和训练一个适合VLA的基座模型，因为任何开源的LLM模型都不具备这样的能力。在这个过程中稀疏化是模型设计的关键。它可以实现模型容量扩容的同时不会大幅度增加推理负担。我们通过两个维度来实现技术化。一个是我们采用了MoE的架构，通过多个专家实现模型扩容，同时还能够激活参数量不会大幅度增加。另一个是我们引入了sparse attention来进一步提高稀疏化率，从而提升端侧的推理效率。

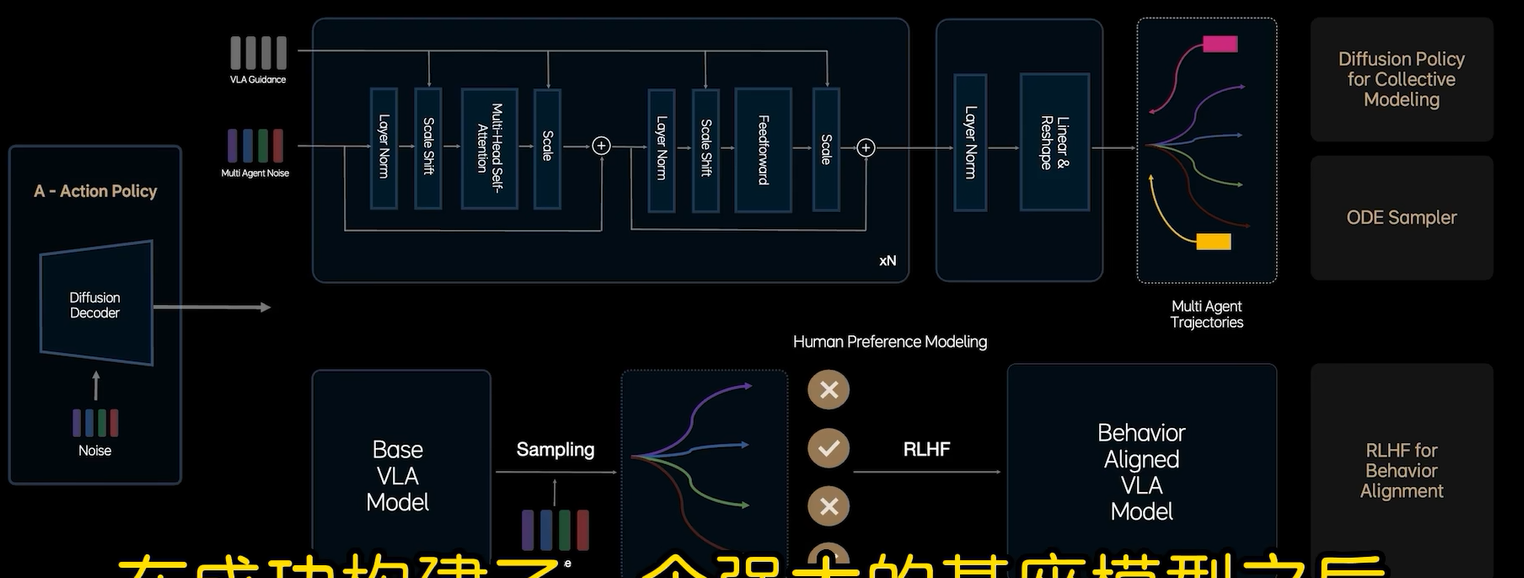


在这个新的基座模型训练过程中，我们花了很多时间去找到最佳的数据配比，融入了大量的3D数据和自动驾驶相关的图文数据，并减少了文史类数据的比例。最后为了进一步激发模型的3D空间理解和推理能力，我们加入了未来帧的预测生成和稠密深度的预测等训练任务。

LLM在获得3D空间智能的同时，在逻辑推理方面需要进一步的提升。我们训练模型去学习人类的思考过程，并自主切换快思考和慢思考。在慢思考模式下模型会经过思维链CoT来输出action token。由于自动驾驶不需要冗长的CoT，同时也因为实时性的要求，所以我们使用了固定且简短的CoT模板。而在快思考模式下，模型则不需要经过CoT就直接输出action token。这就是我们将快慢思考有机结合在同一个模型中的体现。

很多人会问LLM是token by token的输出，推理速度能够支撑自动驾驶吗？的确，即便有了上述的结构设计和优化，想要实现VLA超过10赫兹的推理速度是具有挑战性的。我们做了大量的工程工作去压榨Orin-X和Thor-U的性能的。针对CoT过程，我们采用了小词表和投机推理，大幅提升CoT的效率。针对action token的推理，我们采用了创新性的并行解码的方法，即是在同一个transformer模型中加入了两种推理模式，语言逻辑的推理通过因果注意力机制逐次输出，而action token则采用双向注意力机制一次性全部输出。经过上面一系列的设计和优化，我们终于实现了模型的参数规模与实时推理性能之间的平衡。最后，在上述的架构图中还有一个亮点。VLA的强大之处在于用户可以直接与模型对话，给它下达指令，模型会自动拆解并执行任务。我稍后会介绍这些特性是如何改变自动驾驶产品形态的。

在成功构建了一个强大的基座模型之后，我们利用diffusion将action token解码成最终的驾驶轨迹。在日常的驾驶过程中，车辆与周围的交通参与者，如车辆、行人、骑行人存在着密切的交互关系，并会最终影响自车的行为。因此在diffusion模型中我们不仅生成自车的轨迹，还预测其他车辆和行驶轨迹，从而大大提升了VLA模型在复杂交通环境中的博弈能力。此外，diffusion还有一个巨大的优势，就是可以根据外部的条件输入改变生成结果，这在图像生成领域被称为多风格生成。有了这样的特性，类似“理想同学开快点，我赶时间”这样的功能就很容易实现了。我们在Diffusion模块中使用的是多层的DiT。然而，Diffusion模型面临着一个显著的挑战，那就是其生成效率极低，即需要很多个step才能成功生成稳定的轨迹。为了解决这一问题，我们采用了基于常微分方程的ODE采样器来大幅度加速Diffusion的生成过程，使其在2~3步内即可生成稳定的轨迹。

至此我已经完整介绍了VLA的架构和基础的训练过程。

VLA模型在绝大多数场景下能够接近人类的价值水平，然而在某些长尾case下，VLA仍然存在着不符合人类价值观的问题。为解决这个问题，我们增加了后训练的阶段，希望能够对齐人类驾驶员在这些case下的行为。在过去几年里，我们不仅积累了大量的人类司机的驾驶数据，也有很多接管数据，这些接管数据都是不符合人类预期的表现。我们筛选了大量的接管数据，建立了一个人类偏好的数据集，应用RLHF去微调模型的采样过程，使模型能够从这些偏好数据中学习和对齐人类的行为。这一创新性的步骤让我们在模型性能上取得了进一步的提升。随着偏好数据的逐步丰富，模型的表现也逐步接近了专业司机的水平，安全下限也得到了巨大的提升。有关RLHF与diffusion结合的具体细节，我们也在近期的论文中进行了详细的阐述，欢迎大家查阅。

要实现真正的自动驾驶，仅仅达到人类司机的水平还是不够的。那么如何让系统有机会超越人类驾驶水平呢？或许大家第一反应是强化学习。强化学习在自动驾驶领域已经不算新鲜事儿了，但是过去的尝试都没有取得很好的效果。我认为这里面有两个主要的限制因素。第一，早期的车端架构未能实现端到端的可训。强化学习作为一种稀疏的弱监督过程，若无法实现高效地、无损地信息传递的话，强化学习的效果就会大打折扣。第二，physical AI需要与真实世界进行交互以获取奖励信号。因此，自动驾驶作为physical AI最直接的应用，它的强化学习也高度依赖于良好的交互环境。

然而，过去的尝试都是基于3D的游戏引擎，场景真实度不足，限制了强化学习在真实价值场景中的应用。同时，因为场景建设效率低下、场景规模小，模型很容易学偏，从而导致强化出来的模型完全不可用。我们已经获得了一个端到端可训的VLA模型，它解决了第一个限制。至于良好的交互环境，我们的做法是结合场景重建与生成。纯生成模型的优势在于其良好的泛化能力能够生成多变的场景，但是可能会出现不符合物理规律的幻觉，难以满足自动驾驶的严格要求。相反，纯重建模型是依托于真实数据来实现3D场景的重建，但是在大视角变换下可能会出现空洞和变形，也没办法满足自动驾驶的要求。为了解决这些问题，我们选择以真实数据的3D重建为基础，同时特意在不同的视角下添加噪音来训练生成模型以便恢复这些模糊的视角。这样做之后，生成模型就具备了多视角的生成能力。在与3D重建联合优化后，可以获得一个各个视角下都接近真实世界的3D环境。这在很大程度上解决了上面提到的第二个限制，关于生成和重建是如何结合的。很多细节发表在我们的论文里，其中四篇还中了今年的CVPR 2025：：*StreetCrafter、DrivingSphere、DriveDreamer4D*与*ReconDreamer*。

突破了这两个限制之后，我们终于有机会去尝试大规模的自动驾驶强化学习。但规模化需要解决效率的问题，无论重建和生成效率都不高。过去一年里，我们与英伟达团队密切合作，进行了大量的工程优化，显著提升了场景生成和重建的效率。其中一项工作是将3DGS的训练速度提高了七倍以上，这项工作也已经投到SIGGRAPH 2025上了。

好，以上就是按照VIA最关键的六个基础点。

总结下来MindVLA成功整合了空间智能、语言智能和行为智能，可以说这是一个巨大的突破。并且，通过创新性的预训练和后训练方法，我们发现VLA实现了卓越的泛化性能和涌现特性，它不仅在驾驶场景下表现良好，在室内环境中也展示出了一定的适应性和延展性，这是Physical AI能够大规模落地的关键。一旦跑通这套范式，理想将有望为更多行业赋能。

下面我们给出了related papers：



那么在技术分享之后，让我们转向用户更关心的产品形态和体验。VLA模型究竟能够为用户带来什么不同的产品体验呢？一句话来讲，有MindVLA赋能的车不再只是一个驾驶工具，而是一个能够与用户沟通，理解用户意图的司机。最后我们也展示一下在研发过程中的三个实车demo，也对应了我们日常生活中用车的三个场景。



首先来看第一个，当你来到一个陌生园区，想去某个特地商店，比如星巴克，但是具体的位置你不知道。在这种情况下，您只需要告诉车辆带我去星巴克，车辆现在没有导航信息的情况下，通过自主漫游找到目的地。在执行任务的过程中，你还可以随时进行人工干预，比如说“开太快了，开慢点吧”，或者说“我觉得你应该走左边，感觉右边会绕远”，通过这样的自然对话来改变它的路线和行为。Mind VLA能够理解并执行这些指令，而且在认出目标商店后将你放在附近。

还有一个大家经常遇到的情景，在一个陌生的城市打车时，你不知道如何描述你的位置，因此，你找不到司机，同时司机也找不到你。当你拥有Mind VLA赋能的车辆，就没有让这样的烦恼了。你不需要描述你的位置，只需要拍一张附近环境的照片发给汽车，让车子自己来找你。他可以自己从车库开出来，开出园区，经过一些城市道路，找到图片中的位置，把你接走。

当你驾车来到商场地库找车位是一件非常麻烦的事儿。有了Mind VLA后，你只需要在超市门口下车，然后对车说自己去找停车位挺好。Mind VLA能够理解你的指令，并利用它强大的空间推理能力，自主寻找停车位来进行停车。即使遇到了死胡同，他也会自如地倒车，然后重新寻找合适的车位停下来。这整个过程中不依赖于地图或导航信息，只依赖于Mind VLA强大的空间和逻辑推理能力。

这些场景展示了自动驾驶车辆如何从单纯的运输工具转变为贴心的专职司机。他能听得懂，看得见，找得到。想象一下，将来每个人拥有一个这样的司机，你可以让他接孩子，带老人去菜市场买菜，那将是一个多么令人愉悦的体验。我们希望Mind VLA能为自动驾驶车辆赋予类似人类的认知和适应能力，将它转变为能够思考的智能体。就像iphone重新定义了手机，Mind VLA也将重新定义自动驾驶。

关于这个report的更多解读还可参见如下网址：

<https://zhuanlan.zhihu.com/p/1885988337225032557>

**贾鹏的google scholar:**

<https://scholar.google.com/citations?hl=zh-CN&user=Z_QY_VwAAAAJ&view_op=list_works&sortby=pubdate>

**DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models**

原文出处：<https://arxiv.org/abs/2402.12289>

日期：2024.06

**Motivation:**

（Abstract）

A primary hurdle of autonomous driving in urban environments is understanding complex and long-tail scenarios, such as challenging road conditions and delicate human behaviors.

（Introduction）

Existing autonomous driving systems, typically comprising 3D perception, motion prediction, and planning, struggle with these scene understanding challenges. Specifically, 3D perception [2, 3, 4, 5] is limited to detecting and tracking familiar objects, omitting rare objects and their unique attributes; motion prediction [6, 7, 8, 9, 10] and planning [11, 12, 13] focus on trajectory-level actions, often neglecting the decision-level interactions between objects and vehicles.

**Contribution:**

（Abstract）

We introduce DriveVLM, an autonomous driving system leveraging Vision-Language Models (VLMs) for enhanced scene understanding and planning capabilities. DriveVLM integrates a unique combination of reasoning modules for scene description, scene analysis, and hierarchical planning. Furthermore, recognizing the limitations of VLMs in spatial reasoning and heavy computational requirements, we propose DriveVLM-Dual, a hybrid system that synergizes the strengths of DriveVLM with the traditional autonomous driving pipeline.

（Introduction）

We introduce DriveVLM, a novel autonomous driving system that aims at these scene understanding challenges, capitalizing on the recent Vision-Language Models (VLMs) [14, 15, 16, 17] which have demonstrated exceptional prowess in visual comprehension and reasoning. Specifically, **DriveVLM contains a Chain-of-Though (CoT) process with three key modules: scene description, scene analysis, and hierarchical planning.** **The scene description module** linguistically depicts the driving environment and identifies critical objects in the scene; **the scene analysis module** delves into the characteristics of the critical objects and their influence on the ego vehicle; **the hierarchical planning module** formulates plans step-by-step, from meta-actions and decision descriptions to waypoints. These modules respectively correspond to the components of the traditional perception-prediction-planning pipeline, but the difference is that these modules tackle object perception, intention-level prediction and task-level planning, which were extremely challenging to cope with in the past.

While VLMs excel in visual understanding, they have limitations in spatial grounding and reasoning, and their computational intensity poses challenges for onboard inference speed. Therefore we further propose **DriveVLM-Dual, a hybrid system that combines the strengths of both DriveVLM and traditional systems.** **DriveVLM-Dual optionally integrates DriveVLM with traditional 3D perception and planning modules**, such as 3D object detectors, occupancy networks, and motion planners, enabling the system to achieve 3D grounding and high-frequency planning abilities. **This dual system design, akin to the human brain’s slow and fast thinking processes**, **adapts efficiently to varying complexity in driving scenarios.**

🡺 Contribution 1: We introduce DriveVLM, a novel autonomous driving system that leverages VLMs for effective scene understanding and planning. We further introduce DriveVLM-Dual, a hybrid system that incorporates DriveVLM and a traditional autonomous pipeline, which achieves improved spatial reasoning and real-time planning capabilities.

Meanwhile, **we formally define the scene understanding and planning (SUP) task, and propose new evaluation metrics to assess the scene analysis and meta-action planning capabilities of DriveVLM and DriveVLM-Dual.** **We carry out a comprehensive data mining and annotation pipeline to construct an in-house SUP-AD dataset for the SUP task.** Extensive experiments on both the nuScenes dataset and our own dataset demonstrate the superior performance of DriveVLM, particularly in fewshot scenarios. Furthermore, DriveVLM-Dual exceeds state-of-the-art end-to-end motion planning methods. We have also deployed the model on a production vehicle, confirming that DriveVLM-Dual is effective in real-world autonomous driving environments. Additionally, we have included a demo in the supplementary materials.

🡺Contribution 2: We present a comprehensive data mining and annotation pipeline to construct a scene understanding and planning dataset (SUP-AD), together with metrics for evaluation.

Contribution 3: We have successfully deployed DriveVLM-Dual system in a production vehicle and test various effective strategies for accelerating VLM deployment in real driving scenarios.

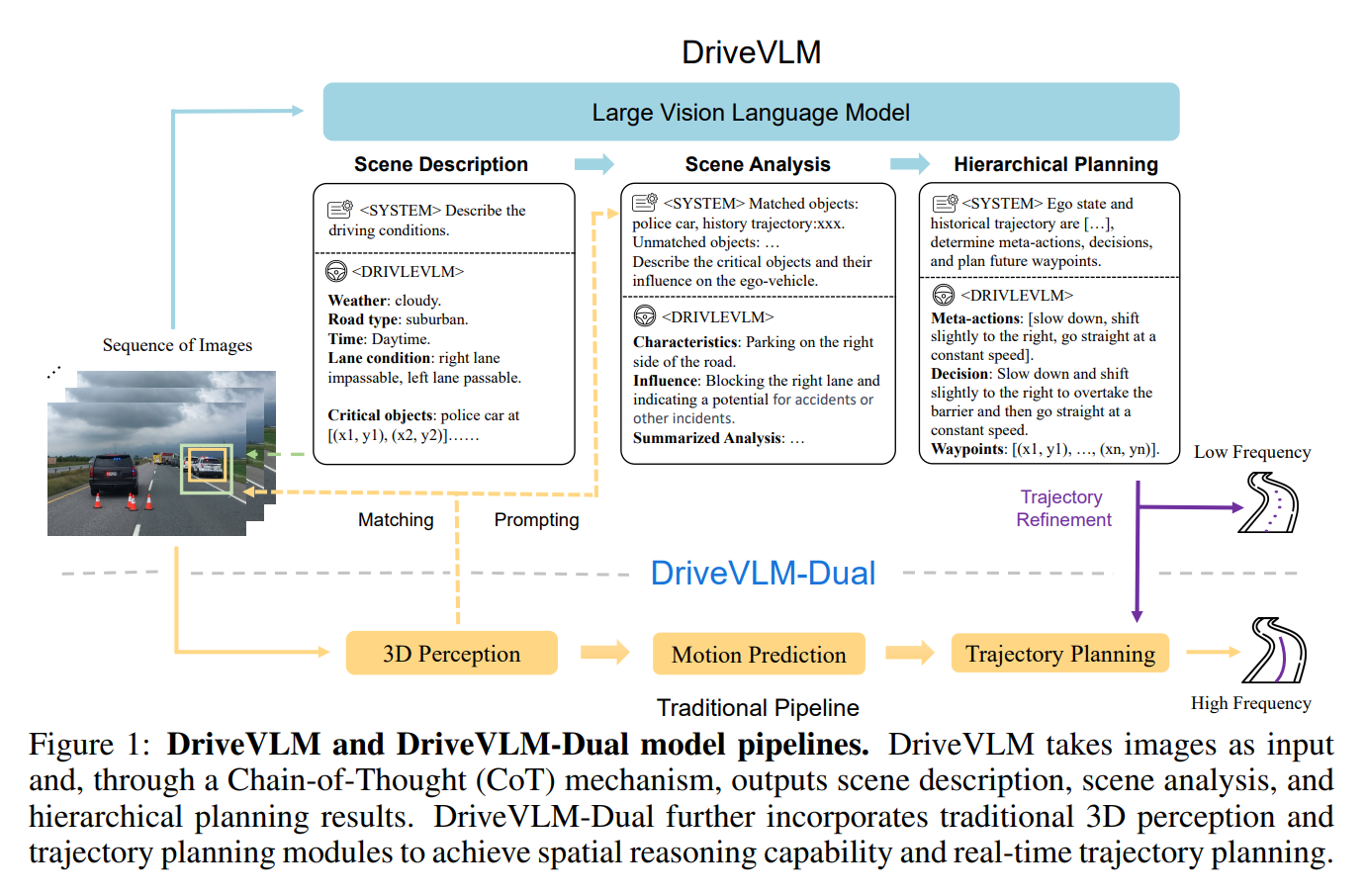
**Related Works:**

**Vision-Language Models (VLMs).** Recently, there has been a surge in research on large VisionLanguage Models (VLMs), exemplified by works such as **MiniGPT-4** [16], **LLaVA** [17], **QwenVL** [18], and others [19, 14, 20, 21]. VLMs can be used in various scenarios, especially robotics [22, 23, 24, 25, 26], where VLMs output corresponding actions that can be high-level instructions [22] or low-level robot actions [24]. DriveVLM focuses on utilizing VLMs to assist in autonomous driving, thereby establishing a novel framework. A Concurrent work [15] shares a similar motivation.

**Learning-based Planning.** The integration of learning frameworks into motion planning has been an active area of research since Pomerleau [11] pioneering contributions. One promising line of work is **Reinforcement learning** and **imitation learning** [27, 28, 29]. These methods can learn an end-to-end planning policy that directly maps raw sensory inputs to control actions [29]. Several works [30, 31, 32, 33] improve interpretability by explicitly building dense cost maps derived from learning-based modules. **A recent trend involves training multiple blocks in an end-to-end fashion** [32, 33, 34, 35]. **These methods enhance overall performance, but rely on backpropagation from future trajectory predictions loss in a less interpretable decision-making process** [36].

**DriveVLM-Overview:**

The overall pipeline of DriveVLM is given as follow:



**DriveVLM-Scene Description:**

The scene description module identifies driving environment description and critical objects.

**Environment Description.** Driving environments, such as weather and road conditions, have a non-negligible impact on driving difficulty. Therefore, the model is first prompted to output a linguistic description *E* of the driving environment, including several conditions: ***E* = {*E*weather, *E*time, *E*road, *E*lane}**, each representing a crucial aspect of the driving environment. The weather component, ***E*weather, spans conditions from sunny to snowy**, affecting visibility and traction. The time component, ***E*time, distinguishes between daytime and nighttime**, impacting driving strategies due to visibility changes. Road types, ***E*road, such as urban or highway**, introduce different challenges, while lane conditions, ***E*lane**, **focus on current lane positioning and possible maneuvers**, crucial for safe driving decisions.

**Critical Object Identification.** In addition to environmental conditions, various objects in driving scenarios significantly influence driving behaviors. **Unlike traditional autonomous driving perception modules, which detect all objects within a specific range, we solely focus on identifying critical objects that are most likely to influence the current scenario, inspired by human cognitive processes during driving.** **Each critical object, denoted as *Oc*, contains two attributes: the object category *c* and its approximate bounding box coordinates *b*(*x*1, *y*1, *x*2, *y*2) on the image. The category and coordinates are mapped to their corresponding language token id in the language modality**, enabling seamless integration into the following modules. Moreover, **taking advantage of the pre-trained vision encoder,** DriveVLM can identify long-tail critical objects that may elude typical 3D object detectors, such as road debris or unusual animals.

**DriveVLM-Scene Analysis:**

The emergence of advanced vision-language models has provided us with the ability to perform a more comprehensive analysis of the current scene. The scene-level analysis summarizes all the critical objects together with the environmental description. This summary gives a comprehensive understanding of the scene, and is fed into the following planning module.

**Critical Object Analysis.** DriveVLM characterizes critical objects in three aspects: **static attributes *Cs***, **motion states *Cm***, and **particular behaviors *Cb***. Static attributes ***Cs*** **describe inherent properties of objects**, **such as a roadside billboard’s visual cues or a truck’s oversized cargo**, which are critical in preempting and navigating potential hazards. Motion states ***Cm*** **describe an object’s dynamics over a period**, **including position, direction, and action—characteristics that are vital in predicting the object’s future trajectory and potential interactions with the ego vehicle.** Particular behaviors ***Cb*** **refer to special actions or gestures of an object that could directly influence the ego vehicle’s next driving decisions.** **We do not require the model to analyze all three characteristics for all objects. In practice, only one or two characteristics apply to a critical object.** Upon analyzing these characteristics, DriveVLM then predicts the potential influence I of each critical object on the ego vehicle.

**DriveVLM-Hierarchical Planning:**

The scene-level summary is then combined with the route, ego pose and velocity to form a prompt for planning. Finally, DriveVLM progressively generates driving plans, in three stages: **meta-actions**, **decision description**, and **trajectory waypoints**.

**Meta-actions *A*.** **A meta-action, denoted as *ai*, represents a short-term decision of the driving strategy.** **These actions fall into 17 categories, including but not limited to acceleration, deceleration, turning left, changing lanes, minor positional adjustments, and waiting.** To plan the ego vehicle’s future maneuver over a certain period, we generate a sequence of meta-actions.

**Decision Description *D*.** Decision description *D* **articulates the more fine-grained driving strategy the ego vehicle should adopt.** **It contains three elements: Action *𝓐*, Subject *𝓢*, and Duration 𝓓.** **Action** pertains to meta actions **such as ‘turn’, ‘wait’, or ‘accelerate’**. **Subject** refers to the interacting object, **such as a pedestrian, a traffic signal, or a specific lane.** **Duration** indicates the temporal aspect of the action, **specifying how long it should be carried out or when it should start.**

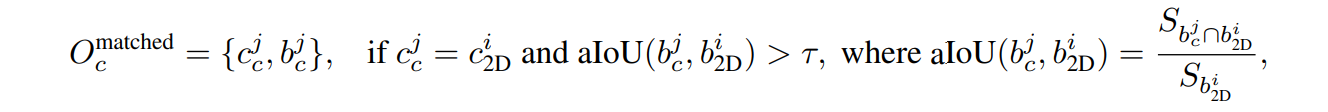
**Trajectory Waypoints *W*.** Upon establishing the decision description *D*, our next phase involves the generation of corresponding trajectory waypoints. These waypoints, denoted by ***W* = {*w*1, *w*2, ..., *w*n}, *wi* = (*xi* , *yi*)**, **depict the vehicle’s path over a certain future period with predetermined intervals *∆*t**. **We map these numerical waypoints into language tokens for auto-regressive generation.**

**DriveVLM-Dual:**

To mitigate the challenges of high latency and imprecise spatial and motion understanding in VLMs, we propose DriveVLM-Dual, a collaboration between DriveVLM and the traditional autonomous driving system.

This novel approach involves two key strategies: **incorporating 3D perception for critical object analysis**, and **high-frequency trajectory refinement**.

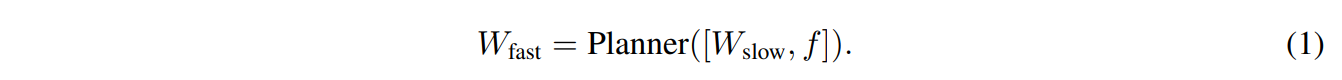
**Integrating 3D Perception.** We represent objects detected by a 3D detector as *O*3D = {*ci*3D, *bi*3D}, where *bi*3D denotes the *i*-th bounding box and *ci*3D denotes its category. These 3D bounding boxes are then back-projected onto 2D images to derive corresponding 2D bounding boxes *bi*2D. We conduct IoU matching between these 2D bounding boxes *bi*2D and *bjc*. *bjc* are the bounding boxes of previously identified critical objects *O*critical = {*cjc*, *bjc*}. We classify critical objects that meet a certain approximate IoU threshold and belong to the same category as matched critical objects *Omatchedc*, defined as



Those critical objects without a corresponding match in the 3D data are noted as *Ounmatchedc*.

In the scene analysis module, for *Omatchedc*, the center coordinates, orientations, and historical trajectories of the corresponding 3D objects are used as language prompts for the model, assisting in object analysis. Conversely, for *Ounmatchedc*, analysis relies solely on the language tokens derived from the image. This design enables DriveVLM-Dual to understand the locations and motions of critical objects more accurately, enhancing the overall performance.

**High-frequency Trajectory Refinement.** To achieve real-time, high-frequency inference capabilities, we integrate it with a conventional planner to form a slow-fast dual system, combining the advanced capabilities of DriveVLM with the efficiency of traditional planning methods. After obtaining a trajectory from DriveVLM at low frequency, denoted as *W*slow, we take it as a reference trajectory for a classical planner for high-frequency trajectory refinement. In the case of an optimization-based planner, *W*slow serves as the initial solution for the optimization solver. For a neural network-based planner, *W*slow is used as an input query, combined with additional input features *f*, and then decoded into a new planning trajectory denoted as *W*fast. The formulation of this process can be described as:



This refinement step ensures that the trajectory produced by DriveVLM-Dual (1) achieves higher trajectory quality, and (2) meets real-time requirements. In practice, the two branches operate asynchronously in a slow-fast manner, where the planner module in the traditional autonomous driving branch can selectively receive trajectory from the VLM branch as additional input.

**Task Definition:**

**The Scene Understanding for Planning task is defined as follows.** **The input** comprises multi-view videos *V* from surrounding cameras and optionally 3D perception results *P* from a perception module. **The output** includes the following components:

**Scene Description *E*:** Composed of weather condition *E*weather, time *E*time, road condition *E*road, and lane conditions *E*lane.

**Scene Analysis *S*:** Including object-level analysis and scene-level summary *S*.

**Meta Actions *A*:** A sequence of actions representing task-level maneuvers.

**Decision Description *D*:** A detailed account of the driving decisions.

**Trajectory Waypoints *W*:** The waypoints outlining the planned trajectory of the ego vehicle.

**Evaluation Metrics—见 Appendix B:**

*Scene Description/Analysis Evaluation.*

We compare the generated scene description with a human-annotated ground truth description.

The ground truth description encompasses structured data such as environmental conditions, navigation, lane information, and critical events with specific objects, verbs, and their influences.

The LLM assesses and scores the generated descriptions based on their consistency with the ground truth.

**Scene Description/Analysis Evaluation的计算见Appendix B中的Table 10。**

*Meta-action Evaluation.*

A driving decision is formulated as a sequence of meta-actions.

We employ a dynamic programming algorithm to compare the model-generated sequences with a manually annotated ground truth sequence.

**Meta-action Evaluation的计算见Appendix B中的Table 11、公式(3)和公式(4)。**

*Sr*, *c*是一个2D矩阵，其中，每一行代表的actions in the ground truth action sequence；每一列代表的是actions in the model output action sequence。

*Sr*, *c*matching代表的是the *r*-th position in the reference sequence == the *c*-th position in the model-generated sequence。

*Sr*, *c*missing代表的是由于the *r*-th position in the reference sequence != the *c*-th position in the model-generated sequence，故看下the (*r*-1)-th position in the reference sequence与the *c*-th position in the model-generated sequence之间的匹配结果即*Sr*-1, *c*，然后将这个匹配结果减去惩罚值*p*missing。

*Sr*, *c*redundant代表的是由于the *r*-th position in the reference sequence != the *c*-th position in the model-generated sequence，故看下the *r*-th position in the reference sequence与the (*c*-1)-th position in the model-generated sequence之间的匹配结果即*Sr*, *c*-1，然后将这个匹配结果减去惩罚值*p*redundant。

**Datasets:**

*SUP-AD*: The SUP-AD dataset is a dataset built by our proposed data mining and annotation pipeline. It is divided into train, validation, and test splits with a ratio of 7.5 : 1 : 1.5.

uScenes: The nuScenes dataset is a large-scale driving dataset of urban scenarios with 1000 scenes, where each scene lasts about 20 seconds. Following previous works [34, 51], we adopt Displacement Error (DE) and Collision Rate (CR) as metrics to evaluate models’ performance.

**DriveAction: A Benchmark for Exploring Human-like Driving Decisions in VLA Models**

原文出处：<https://arxiv.org/abs/2506.05667>

日期：2025.06

**Motivation:**

（Abstract）

A primary hurdle of autonomous driving in urban environments is understanding complex and long-tail scenarios, such as challenging road conditions and delicate human behaviors.

**World4Drive: End-to-End Autonomous Driving via Intention-aware Physical Latent World Model**

原文出处：<https://arxiv.org/abs/2507.00603>

日期：2025.07

**Motivation:**

（Abstract）

A primary hurdle of autonomous driving in urban environments is understanding complex and long-tail scenarios, such as challenging road conditions and delicate human behaviors.

**OmniReason: A Temporal-Guided Vision-Language-Action Framework for Autonomous Driving**

原文出处：<https://arxiv.org/abs/2509.00789>

日期：2025.08

**Motivation:**

（Abstract）

A primary hurdle of autonomous driving in urban environments is understanding complex and long-tail scenarios, such as challenging road conditions and delicate human behaviors.

**The Better You Learn, The Smarter You Prune: Towards Efficient Vision-language-action Models via Differentiable Token Pruning**

原文出处：<https://arxiv.org/abs/2509.12594>

日期：2025.09

**Motivation:**

（Abstract）

A primary hurdle of autonomous driving in urban environments is understanding complex and long-tail scenarios, such as challenging road conditions and delicate human behaviors.