

DrivePI:面向统一自动驾驶理解、感知、预测与规划的空间感知4D MLLM

刘哲¹、黄润辉¹、杨瑞¹、严思明²、
王子宁²、侯璐²、林迪³、白翔⁴、赵恒爽^{1, e}

¹香港大学, ²银网智能科技有限公司,
³天津大学, ⁴华中科技大学

摘要

尽管多模态大语言模型 (MLLMs) 已在多个领域展现出强大能力, 但其在自动驾驶领域生成精细三维感知与预测输出的应用仍鲜有探索。本文提出 DrivePI——一种新型空间感知四维 MLLM, 该框架作为统一的视觉语言-动作 (VLA) 架构, 同时兼容视觉-动作 (VA) 模型。我们的方法通过端到端优化并行执行空间理解、三维感知 (即三维空间占用)、预测 (即占用流) 和规划 (即动作输出)。为获取精确几何信息与丰富视觉表征, 本方法将点云、多视角图像和语言指令整合至统一 MLLM 架构中。我们进一步开发数据引擎, 生成用于四维空间理解的文本-占用与文本-流问答对。值得注意的是, 仅以 0.5B Qwen2.5 模型作为 MLLM 骨干, DrivePI 作为单一统一模型即可匹配甚至超越现有 VLA 模型与专用 VA 模型。具体而言, 与 VLA 模型相比, DrivePI 的性能比 OpenDriveVLA-7B 高出 2 倍。在 nuScenes-QA 数据集上平均准确率达到 5%, 并在 nuScenes 上将碰撞率较 Orion 降低 70% (从 0.37% 降至 0.11%)。相较于专业视觉对齐模型, DrivePI 在 OpenOcc 数据集上以 10.3 的 RayIoU 超越 FB-OCC, 将 OpenOcc 上占用流的平均绝对值误差 (mAVE) 从 0.591 降至 0.509, 并在 nuScenes 规划任务中实现比 VAD 低 32% 的 L2 误差 (从 0.72 米降至 0.49 米)。代码将发布于 <https://github.com/happiness1z/DrivePI>。

1. 介绍

在端到端自动驾驶系统中, 视觉行动 (VA) 模型 [14, 20] 以视觉信息 (激光雷达点云、图像) 作为输入, 并输出动作信号。

^e通讯作者

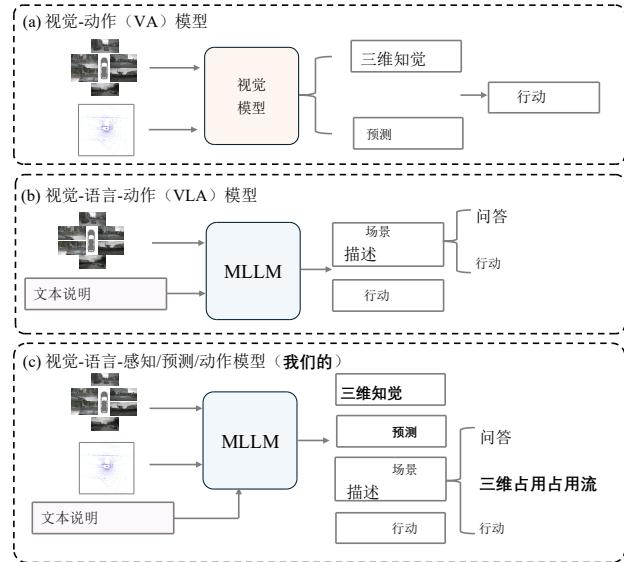


图1. (a)展示了主流视觉-动作 (VA) 模型在端到端自动驾驶中的应用流程。(b)呈现了主流视觉-语言-动作 (VLA) 模型的架构。(c)展示我们自主研发的DrivePI系统, 该系统将粗粒度语言理解与精细的三维感知预测相结合, 既继承了现有 VA 模型的优势, 又融合了 VLA 模型的特性。

nals 方面取得了实质性进展。具体而言, 如图 1(a) 所示, UniAD [14] 和 VAD [20] 采用了一种模块化框架, 从三维感知逐步推进到预测, 随后整合这些信息以生成最终驾驶动作。此外, FusionAD [64] 通过结合激光雷达点云和相机图像进一步提升了 UniAD 的性能。尽管这些方法凭借其精确的空间感知能力和模块化设计取得了令人鼓舞的结果, 但它们在基于语言的场景交互方面存在局限, 这降低了用户友好性。

为解决这些局限性, 研究者 [9, 53, 71] 探索了利用多语言语言模型 (MLLMs) 强大的推理能力和类人决策能力。具体-

具体来说，OpenDriveVLA[71]和orion[9]采用了视觉语言-动作（VLA）框架，该框架以多视角图像和语言指令作为输入生成动作，如图1(b)所示。这些基于VLA的方法在场景交互能力上表现优异，显著提升了用户参与度。然而，由于缺乏与VA模型模块化设计相匹配的精细中间3D感知和预测输出，这些VLA方法通常难以保证可靠输出，从而影响了可解释性和安全性的保障。

因此，一个自然的问题随之产生：我们能否开发出一个统一框架，将视觉感知模型（VA）的精确空间感知与基于视觉语言交互（VLA）方法的自然语言交互相结合？本文提出DrivePI——一种新型空间感知的四维MLLM，作为自动驾驶领域的统一视觉-语言-动作（VLA）框架，如图1(c)所示。我们将其命名为4DMLLM，因其同时输出三维空间占用和流动信息，捕捉精细的时空动态特征。与现有将VA和VLA方法视为独立范式的传统做法不同，DrivePI构建了一个统一架构，将视觉感知模型的空间精度与视觉语言交互框架的可解释性和交互能力无缝整合。

具体而言，DrivePI展现出四项显著特征使其区别于现有方法。首先，与仅依赖相机图像作为输入的主流VLA方法不同，DrivePI引入激光雷达作为辅助感知模态，提供精确的三维几何信息，从而更好地激发多模态激光雷达（MLLMs）的空间理解能力。其次，DrivePI生成中间级细粒度的三维感知（例如三维空间占用）和预测（例如占用流）表征，确保MLLM输出特征保持可靠的空间感知能力，从而提升自动驾驶系统的可解释性和安全保障。第三，我们开发了增强型数据引擎，将三维空间占用和占用流表征无缝整合到自然语言场景描述中，使模型能够通过文本理解推理复杂的时空动态。第四，DrivePI作为统一模型，采用涵盖三维感知、预测、规划及场景理解等所有任务的端到端联合优化。综上所述，我们的贡献如下：

- 我们提出DrivePI，这是首个统一的空间感知4DMLLM框架，它将粗粒度的语言空间理解与精细的3D感知能力无缝融合，弥合了基于视觉辅助（VA）与基于视觉语言分析（VLA）的自动驾驶范式之间的差距，同时继承了两种方法的互补优势；

- 我们采用激光雷达作为补充感知模态与相机图像协同工作，提供高精度三维几何信息，从而更充分激发机器学习模型的空间理解能力。此外，DrivePI可实现精准的三维感知（例如占用预测）与预测（例如占用流预测），有效提升模型的可解释性与安全保证。

- 我们基于数据引擎开发了三个互补的空间理解基准，通过构建多个问答（QA）对实现：静态场景理解的3D占用感知、动态运动分析的占用流预测，以及决策评估的轨迹规划。这些基准共同评估了语言空间推理能力在时间和空间维度上的不同方面。

- 尽管仅采用紧凑的0.5B参数MLLM架构，DrivePI在三维空间占用率和占用流方面仍优于现有视觉辅助模型，同时在自动驾驶领域保持与现有视觉辅助驾驶框架相当的交互能力。

2. 相关作品

多模态大语言模型。随着大语言模型（LLMs）的快速发展[2, 31, 46, 61]，大量研究尝试将额外模态整合到LLMs中，从而扩展其应用能力。目前，多模态大语言模型（MLLMs）已成功应用于各类任务，包括图像理解[3, 7, 32, 34, 37]、视频理解[30, 55, 56]以及语义理解[4, 22, 42, 51, 67, 69]，这充分展现了LLMs在下游应用中的巨大潜力。因此，越来越多的研究[1, 38, 39, 58, 62, 70]开始探索如何通过MLLMs实现空间智能。具体而言，VSI-Bench[62]建立了一个综合基准，该基准收集多样化的室内图像并生成一系列问题，旨在评估三维空间理解能力，从而验证MLLMs的空间推理能力。Gemini RoboticsER[1]从单张图像预测度量3D边界框，并进一步实现开放词汇3D目标检测。沿此方向，Seed1.5-VL[12]通过统一MLLM架构，同时处理2D或3D定位任务与其他任务（例如OCR、空间理解）的图像数据，从而增强MLLMs的感知能力。然而，这些方法主要实现粗粒度空间感知，例如描述物体间关系或预测物体级3D边界框。

端到端自动驾驶。由于

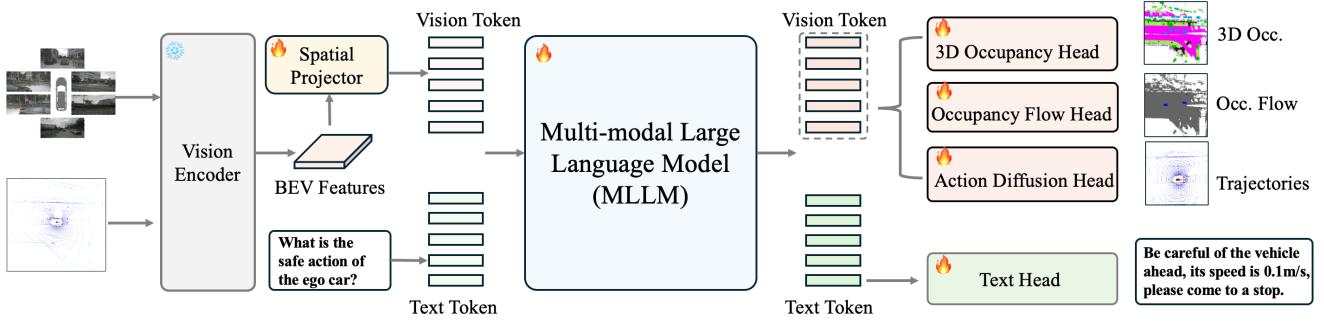


图2. DrivePI的流程包含以下步骤：首先，我们使用视觉编码器从图像和激光雷达数据中提取特征，获得潜在的BEV特征，随后通过空间投影器将其转换为视觉标记。接着，将视觉标记和文本标记输入 MLLM 生成输出标记。该 MLLM 通过四个专用模块生成响应：文本模块以自回归方式理解场景，三维占用模块实现精准空间感知，占用流模块进行像素级运动预测，动作扩散模块则负责轨迹规划。

端到端模型的开发。许多研究[14, 20, 27, 45, 57, 64]探索了如何在端到端模型中执行感知、预测和规划任务，极大提升了规划性能的上限并降低了自动驾驶系统的复杂性。UniAD[14]采用模块化架构，通过不同模块处理各项任务，并以端到端方式联合训练这些任务。VAD [20]提出向量化表征，并通过不同的Transformer解码器实现感知、地图构建和规划查询之间的交互[50]。尽管这些方法展示了有前景的规划结果，但它们缺乏通过自然语言描述与用户交互的能力，导致用户友好性降低。

视觉-语言-动作模型。利用MLLM的强大功能能力，多项研究[43、44、48、53、54、59、60]已成功为自动驾驶系统开发MLLM。这些方法利用MLLM的高级推理能力生成丰富的场景描述和高级驾驶指令，从而提升端到端规划的可解释性。此外，为充分发挥MLLM的理解能力，近期研究[9、18、19、23、24、71]将规划模块直接集成到MLLM中，实现视觉-语言-动作（VLA）框架，使推理与规划任务得以整合。例如，OpenDriveVLA[71]采用鸟瞰视角（BEV）表示法生成对应于智能体、静态地图和场景上下文的独立标记。这些标记随后被投射到统一的语义空间，并由MLLM处理，以促进信息交互，从而为规划解码器生成轨迹。本文提出DrivePI，这是一种新型VLA框架，能够无缝整合粗粒度和细粒度信息。DrivePI在保持文本表征对感知和规划推理的可解释性的同时，采用细粒度解码机制生成精确的三维物体占据关系。

泛量与轨迹预测

3. 方法

多模态大语言模型凭借其强大的用户交互能力和类人决策能力，在自动驾驶领域备受关注。然而现有基于LLM的方法难以通过下一轮词预测直接输出精细感知结果（例如三维空间占用率和占用流），而视觉感知（VA）方法却能轻松实现这些功能。为解决这一局限，我们提出DrivePI——一种新型视觉语言感知（VLA）框架，它既能实现粗粒度语言理解，又能达成精细空间感知，完美融合了VA模型与VLA框架的互补优势。接下来我们将详细介绍DrivePI的具体实现。

3.1. 概述

如图2所示，我们展示了DrivePI的流程。首先，为确保MLLMs的输入包含准确的几何信息，我们引入了激光雷达点云作为补充输入（注：nuScenes数据集中的激光雷达点云包含时间信息），相较于单独的相机图像，其能提供更精确的三维空间信息。这一增强对探索MLLMs的空间感知能力至关重要。其次，我们采用先进的多模态视觉编码器[36]处理多视角图像和激光雷达点云，将其转化为紧凑的潜在BEV特征表示。随后，通过空间投影器将潜在BEV特征映射到语言空间并生成视觉标记。这些视觉标记与文本标记随后输入MLLM。最后，我们使用四个专用头：文本头以自回归方式生成场景理解响应，3D占用头实现精准三维感知，占用流头进行细粒度运动预测，动作扩散头用于轨迹规划。需注意的是，MLLM的训练-

能够，且所有任务在训练期间均被联合优化。**空间投影器**。给定潜在BEV特征 $F_{bev} \in \mathbb{R}^{H \times W \times C}$ ，其中 $H \times W$ 通常超过 100×100 的分辨率，若直接在像素级别将这些特征输入MLLM将产生不可承受的计算成本。此处 H, W, C 分别为 F_{bev} 的高度、宽度和通道维度。为解决这一挑战，我们首先将 F_{bev} 拼接为 N 个大小为 $K \times K$ 的拼接块，生成视觉特征 $F_{patch} \in \mathbb{R}^{N \times K^2 \times C}$ ，其中 $N = \frac{H}{K} \times \frac{W}{K}$ 。传统方法涉及应用池化操作将 $K \times K$ 空间特征聚合为单一表征，生成池化特征 $F_{pool} \in \mathbb{R}^{N \times 1 \times C}$ 。然而，这种方法通常会导致细粒度空间信息的丢失。因此，遵循[17]在二维MLLM中处理高分辨率图像输入的方法，我们采用交叉注意力机制，其中 F_{pool}, F_{patch} 和 F_{patch} 分别作为查询、键和值。该设计保留了更详细的空间信息。最后，我们使用线性层将处理后的特征转换为与MLLM输入隐藏状态的通道维度 C_l 匹配，生成最终视觉标记 $F_v \in \mathbb{R}^{N \times C_l}$ 。

3.2. 粗粒度空间理解

本文将粗粒度空间理解定义为基于文本的描述，这种描述源于语言相较于丰富视觉信息（例如图像、激光雷达点云）的高度压缩特性。尽管多语言语言模型（MLLMs）在多个领域展现出卓越能力，但探索其空间理解能力仍是关键挑战，特别是在自动驾驶领域，精准的空间推理至关重要。为此，我们基于数据引擎开发了三个互补的空间理解基准测试：通过构建多组问答（QA）对实现静态场景理解的三维空间感知、动态运动分析的占用流预测，以及决策评估的轨迹规划。这些基准测试共同评估了语言空间推理能力在时空维度上的不同方面。

数据引擎。如图3所示，我们的流程包含三个主要阶段：场景理解的标题标注、四维空间理解标注以及规划推理标注。在第一阶段，我们采用InternVL3-78B[72]分别生成场景的正视图和背面视图描述。这种策略能有效避免多语言语言模型（MLLM）在区分不同视角时可能出现的混淆。随后将两视角的描述文字合并，构建完整的场景描述，并通过进一步优化确保描述质量。第二阶段旨在为MLLM配备四维空间理解能力。具体而言，我们利用真实标注的占用数据

并生成多样化的文本-占用与文本流问答对。这些问答对聚焦于关键任务，例如判断给定位置是否被占用、识别对应物体类别以及预测速度信息。这使得Drive-PI能够以文本形式探索精细的三维占用与流动占用，相较于先前方法[24, 43, 59]。在最终阶段，我们生成文本规划问答对，基于自我车辆的未来轨迹注释来增强规划可解释性。这些问答对要求MLLM分析周围环境，并提供高级驾驶指令及建议轨迹。这种多阶段流程既保证了生成语言数据集的通用性，又确保了其高质量，使MLLM能够开发四维空间理解与规划能力。关于数据引擎的更多细节（例如提示词设计），请参阅我们的补充材料。

3.3. 细粒度空间学习

本文将细粒度空间学习定义为整合显式空间能力的系统，涵盖三维空间占用、占用流分析及轨迹规划三大核心模块。虽然语言描述能有效捕捉高层次语义概念和全局空间布局，但其固有的局限性使其难以满足自动驾驶任务所需的精细空间定位与几何理解精度。为此，我们创新性地引入细粒度视觉头来解决这一技术瓶颈。

细粒度视觉头。我们采用三个细粒度视觉头以实现精确的空间能力：用于体素场景理解的3D占用头、用于时序动态建模的占用流头，以及用于轨迹规划的动作扩散头。具体而言，为实现这些细粒度头，我们首先从多模态表征中提取对应的视觉标记 $F^*v \in \mathbb{R}^{N \times C_l}$ 。随后，我们使用线性投影层将 F^*v 转换为 $F_{outv} \in \mathbb{R}^{N \times K^2 C}$ 。接着将 F_{outv} 重塑为空间特征图 $F_{out} \in \mathbb{R}^{H \times W \times C}$ ，其中 H 和 W 分别表示特征图的高度和宽度。基于空间组织的特征 F_{out} ，我们可以按照现有VA模型无缝集成三个预测头。关于各头的更多细节，请参阅补充材料。

3.4. 损失函数

我们将总损失定义为四个组件的加权和： L_{llm} 用于文本头， L_{occ} 用于占用头， L_{flow} 用于占用流头，以及 L_{action} 用于动作扩散头。因此，总损失 L_{total} 的公式为：

$$L_{total} = \lambda_1 L_{llm} + \lambda_2 L_{occ} + \lambda_3 L_{flow} + \lambda_4 L_{action}, \quad (1)$$

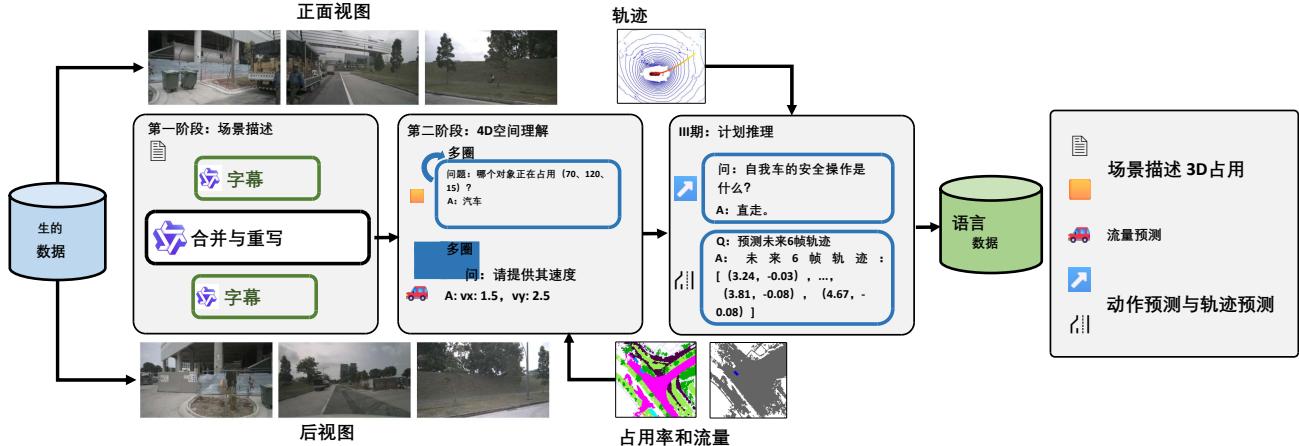


图3. 我们多阶段数据处理流程的示意图。首先分别生成前视图和后视图的描述文字，随后采用InternVL3-78B模型（基于Qwen 2.5-72B[3]作为语言模型）将这些描述文字整合，从而优化生成的场景描述。此外，我们通过多轮对话生成基于占用率和流动数据的真实标注的文本-占用率与文本-流动问答对，以提升四维空间理解能力。最后生成文本-规划问答对，使MLLM能够预测本车未来的动作。

其中 λ_1 、 λ_2 、 λ_3 、 λ_4 分别为文本场景理解、3D占用感知、占用流预测和轨迹规划的平衡权重。所有任务均以端到端的方式联合优化。

4. 实验

4.1. 数据集与评估指标

在本文中，我们对nuScenes[5]数据集进行了全面实验，这是一个大规模自动驾驶基准数据集。该数据集包含750个训练场景、150个验证场景和150个测试场景。nuScenes提供了包含LiDAR点云和6个摄像头多视角图像的同步多模态传感器数据。对于占用预测，我们采用OpenOcc[49]作为主要占用评估指标。此外，为了与大多数3D占用方法进行全面比较，我们在Occ3D [47]上评估结果。对于占用流，我们利用OpenOcc[49]提供的占用流标注来评估DrivePI在细粒度运动预测中的性能。对于轨迹规划评估，我们遵循先前研究[53, 71]的既定协议，并在nuScenes数据集上采用开环评估指标。在文本理解方面，我们基于nuScenes-QA[41]数据集（包含37.7万条训练问答对）开展实验，通过自主研发的数据生成流程，成功产出8.4万条场景描述、56万条用于4D空间推理的问答对及2.4万条规划推理用问答对，最终构建出规模超100万条问答对的完整训练数据集。

在评估指标方面，我们采用已建立的指标来评估DrivePI的性能。对于三维占用率和占用流，我们使用RayIoU[33]和平均Av-

误差速度误差（mAVE）[5]。在规划方面，我们遵循先前的方法[45, 71]，并采用L2距离误差和碰撞率指标。为评估文本理解能力，我们采用nuScenes-QA基准测试中的官方评估指标[41]。针对4D空间理解（见表6），我们采用以下指标：

- 1) 1) 占用状态、占用分类及动作状态的准确度；2) 占用流的平均绝对误差（mAVE）；3) 规划过程中的L2距离误差与碰撞率。

4.2. 实施细节

为平衡计算效率，我们采用Qwen2.5-0.5B[3]作为基础模型。DrivePI的训练包含两个阶段：第一阶段冻结视觉编码器和MLLM，然后使用数据引擎生成的标题对空间投影器进行1个周期的训练，使其将视觉表征与语言嵌入空间对齐；第二阶段在仅冻结视觉编码器的情况下，联合优化空间投影器、MLLM及所有任务特定头（即文本头、3D占用头、占用流头和动作扩散头）1个周期。为实现损失平衡，我们采用 $\lambda_1=\lambda_2=\lambda_3=\lambda_4=1$ 的平衡权重。所有实验均在8块×NVIDIA L40S GPU上进行。

4.3. 主要结果

我们从文本理解、三维空间占用、占用流分析和轨迹规划四个关键任务对DrivePI进行全面评估。在三维空间占用与占用流评估方面，我们采用OpenOcc[49]基准测试进行对比，该基准测试为nuScenes场景提供了三维空间占用和占用流的标注数据。对于轨迹规划任务，我们直接采用官方提供的标注数据。

表1. OpenOcc验证集上的三维占用率及占用流性能

方法	VLM 的	OccScore↑	RayIoU (3D Occup.)↑	MAVE (血流) ↓	RayIoU _{1m}	RayIoU _{2m}	RayIoU _{4m}
OccNeRF [66]		28.5	31.7	—	16.6	29.3	49.2
RenderOcc [40]		33.0	36.7	—	20.3	32.7	49.9
LetOccFlow [35]		36.4	40.5	—	25.5	39.7	56.3
OccNet [49]		35.7	39.7	—	29.3	39.7	50.0
BEVDetOcc-SF [15]		33.0	36.7	1.420	31.6	37.3	41.1
FB-Occ [26]		39.2	39.0	0.591	32.7	39.9	44.4
F-Occ [68]		41.0	39.9	0.491	33.9	40.7	45.2
CascadeFlow [29]		40.9	39.6	0.470	33.5	40.3	45.0
ALOcc-Flow-3D [6]		43.0	41.9	0.556	35.6	42.8	47.4
驱动程序接口 (驱动程序接口)	✓	49.3	49.3	0.509	45.0	50.0	52.9

表2. nuScenes验证集上的规划性能。需注意，我们的统一模型DrivePI在训练过程中默认不纳入自我状态信息，以避免潜在的捷径学习。

方法	VLM 的	自我状态	L2 (m) ↓				结肠 (%) ↓			
			1s	2s	3s	平均数	1s	2s	3s	平均数
ST-P3 [] FF [13] EO [21]			1.33 0.55 0.67	2.11 1.20 2.54	2.90 1.43	2.11	0.23 0.06 0.04	0.62 0.17 0.09	1.27 1.07 0.88	0.71 0.43 0.33
UniAD [14]			0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
VAD [20]			0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
VAD [20]	✓	✓	0.17	0.34	0.60	0.37	0.07	0.10	0.24	0.14
OmniDrive [53] 猎户座[9] OpenDriveVLA-7B [71]	✓ ✓ ✓	✓ ✓ ✓	0.14 0.17 0.20	0.29 0.31 0.58	0.55 0.55 1.21	0.33 0.34 0.66	0.00 0.05 0.00	0.13 0.25 0.22	0.78 0.80 0.55	0.30 0.37 0.25
驱动程序接口 (驱动程序接口) 驱动程序接口 (驱动程序接口)	✓ ✓		0.24 0.19	0.46 0.36	0.78 0.64	0.49 0.40	0.38 0.00	0.27 0.05	0.48 0.28	0.38 0.11

我们基于nuScenes数据集开展研究。为验证DrivePI的文本理解能力，我们在nuScenes-QA基准测试中取得显著成果——该测试集包含多种针对自动驾驶场景设计的视觉问答对。除特别说明外，所有实验均采用统一的0.5B模型架构，确保各任务间参数权重共享。

OpenOcc上的3D占用率与占用流。如表1所示，我们对比了DrivePI与现有方法在3D占用率和占用流方面的表现。DrivePI在多项指标上均表现出色，其OccScore为49.3%，RayIoU为49.3%，mAVE为0.509。DrivePI在3D占用率上比代表性方法FB-OCC[26]高出10.3 RayIoU，并将流mAVE从0.591降至0.509。值得注意的是，DrivePI在OccScore上比先前的最先进方法ALOcc-Flow-3D[6]高出6.3%，RayIoU高出7.4%，并将mAVE降低了0.047，仅使用0.5B MLLM骨干网络就取得了新的最先进结果。这证明了DrivePI在VLA框架内实现卓越4D细粒度空间感知能力的有效性。

nuScenes规划。为深入评估DrivePI作为VLA模型的规划能力，我们进行了实验-

在nuScenes基准测试中的表现[5]。如表2所示，我们将DrivePI的性能与传统的端到端VA方法和基于VLM的方法进行对比。DrivePI在包含自我状态时，L2误差为0.40，碰撞率为0.11%，在碰撞率方面分别比端到端VA方法VAD [20]和VLA方法OpenDriveVLA-7B[71]低0.03%和0.14%。值得注意的是，与近期的orion[9]相比，DrivePI将碰撞率降低了70%（从0.37%降至0.11%）。在不包含自我状态时，DrivePI的L2误差比VAD低32%（从0.72米降至0.49米）。这些结果证明了DrivePI作为VLA模型在规划任务中的有效性。

nuScenes-QA上的文本理解。文本理解能力对VLA模型至关重要，因为它使自动驾驶系统能够基于自然语言进行解释和推理，从而支持更类人的决策。我们在nuScenes-QA[41]基准上验证了DrivePI的文本理解性能。如表3所示，仅使用0.5B模型规模的DrivePI实现了60.7%的准确率，比OpenDriveVLA-7B[71]高出2.5%。这些结果表明DrivePI具有令人期待的文本理解能力。

3D Occupancy on Occ3D.超越统一模型

表3. nuScenes-QA验证集上的文本理解性能。Ext.、Cnt.、Obj.、Sts.、Cmp.和Acc.分别代表存在、计数、对象、状态、比较和总体准确率。

方法	Ext. \uparrow	Cnt. \uparrow	Obj. \uparrow	Sts. \uparrow	Cmp. \uparrow	Acc. \uparrow
LLaMA-AdapV2 [11]	19.3	2.7	7.6	10.8	1.6	9.6
LLaVA1.5 [34]	45.8	7.7	7.8	9.0	52.1	26.2
激光雷达-激光雷达[63]	74.5	15.0	37.8	45.9	57.8	48.6
BEVDet+BUTD [41]	83.7	20.9	48.8	52.0	67.7	57.0
OpenDriveVLA-0.5B [71]	83.9	22.0	50.2	57.0	68.4	58.4
OpenDriveVLA-3B [71]	84.0	22.3	50.3	56.9	68.5	58.5
OpenDriveVLA-7B [71]	84.2	22.7	49.6	54.5	68.8	58.2
驱动程序接口（驱动程序接口）	85.3	22.4	57.5	59.1	68.3	60.7

表4. Occ3D-nuScenes验证集上的三维占用性能。*表示DrivePI仅在Occ3D-nuScenes的三维占用任务上进行训练。

方法	VLM 的	RayIoU \uparrow	RayIoU _{1m}	RayIoU _{2m}	RayIoU _{4m}
RenderOcc [40]		19.5	13.4	19.6	25.5
SimpleOcc [10]		22.5	17.0	22.7	27.9
BEVFormer [25]		32.4	26.1	32.9	38.0
BEVDet-Occ [16]		32.6	26.6	33.1	38.2
FB-Occ [26]		33.5	26.7	34.1	39.7
SparseOcc [33]		36.1	30.2	36.8	41.2
OPUS [52]		41.2	34.7	42.1	46.7
驱动程序*（我们）	✓	46.0	42.2	46.7	49.2

表5. DrivePI中文本头与视觉头的消融研究

#	文本标题	视网膜头	3D Occup.	偶发性流量	规划	质量保证
			RayIoU \uparrow	mAVE \downarrow	L2 \downarrow Col. \downarrow	
<i>I</i>	✓	-	-	-	- -	61.2
	-	✓	47.5	0.69	1.02 0.39	-
	✓	✓	49.3	0.51	0.50 0.38	60.7

在对所有任务进行联合训练后，我们还专门针对Occ3D的3D占用任务对DrivePI进行单独训练，以便在Occ3D基准测试中与现有方法进行全面比较[47]。如表4所示，DrivePI以46.0%的射线交并比（RayIoU）实现了最先进的性能，较先前最佳方法OPUS[52]显著提升了4.8%。值得注意的是，DrivePI基于MLLM架构构建，这凸显了其在细粒度3D感知方面的强大能力，尽管该架构主要设计用于多模态理解。

4.4. 消融研究

在本节中，我们通过消融实验来验证DrivePI架构设计的有效性。除非特别说明，实验默认使用DrivePI-0.5B版本。**文本理解头与视觉头的消融实验。**DrivePI在统一的视觉语言架构（VLA）框架下，整合了文本理解、三维感知、预测及规划功能。为简化表述，我们将精细的三维空间占用头、空间占用流头和轨迹规划头的组合统称为视觉头。为评估各组件的协同效应，我们...

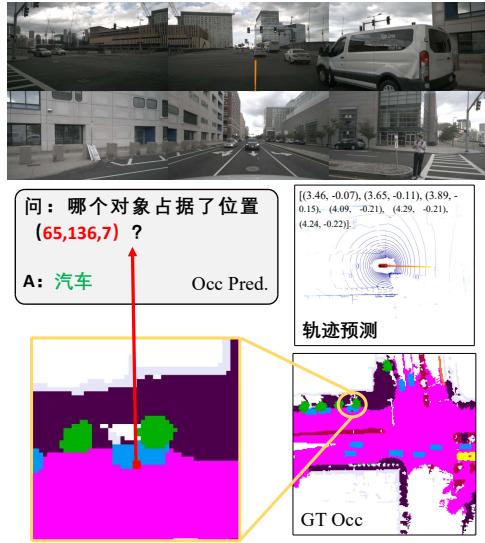
在独立测试中，我们通过移除单个模块进行消融研究，定量结果汇总于表5。当仅启用文本模块（I）时，DrivePI在文本理解任务上达到61.2%的准确率，表现具有竞争力。当仅启用视觉模块（即占用、流动和规划）时（II），DrivePI仍取得良好结果：3D占用任务的RayIoU为47.5%，占用流动任务的mAVE为0.69，轨迹规划任务的L2误差为1.02，碰撞率为0.39（训练过程中未使用本车信息）。当同时启用文本和视觉模块（III）时，DrivePI在多数任务中表现更优。与仅视觉设置（II）相比，统一模型（III）使RayIoU提升1.8%，mAVE降低0.18，L2误差减少0.52。主要原因在于文本理解有助于更好地适配视觉任务的特征空间。此外，其文本理解性能保持在60.7%的准确率，接近纯文本设置（I）。这些结果验证了在单一VLA框架内将文本理解与三维感知、预测及规划任务统一的有效性。

文本数据规模扩展。为探究不同规模文本数据的影响，我们通过调整指令调优数据量在多个任务中开展消融实验，包括基于数据引擎预测占用状态、占用类别、动作状态及轨迹规划，以及nuScenes-QA的问答任务。如表6所示，实验主要在仅使用文本头的Qwen-2.5B模型上进行。当仅使用112K样本（84K字幕+28K占用问答对）训练时，DrivePI在占用状态预测准确率仅为73%，占用类别预测准确率仅为14.3%。当占用问答对从28K扩展至560K时，性能显著提升：占用状态预测准确率提升14%，占用类别预测准确率提升44.9%。进一步增加占用流程、动作及官方nuScenes-QA的问答对后，占用状态与类别预测准确率分别提升1.6%和0.6%。此外，

表6. DrivePI数据缩放的消融研究。发生状态、发生类别、发生流、动作状态列分别表示占用状态（即“是”或“否”）、占用类别、占用流、动作指令（即“直行”、“右转”、“左转”和“停止”）。

模型尺寸	字幕	训练数据量 OCC. 流程操作					偶发状态 Acc. \uparrow	偶发性 Acc. \uparrow	偶发性流 mAVE \downarrow	操作状态 Acc. \uparrow	规划		质量保证 Acc. \uparrow
		质量保证	L2 \downarrow	Col. \downarrow									
0.5B	84k	420k	140k	24k	377k	86.0	50.4	0.91	83.8	0.79	0.63	60.7	
3B	84k	28k	—	—	—	73.0	14.3	—	—	—	—	—	—
	84k	56k	—	—	—	74.2	22.4	—	—	—	—	—	—
	84k	280k	—	—	—	86.8	54.7	—	—	—	—	—	—
	84k	560k	—	—	—	87.0	59.2	—	—	—	—	—	—
	84k	420k	140k	24k	377k	88.6	59.8	0.69	83.3	0.86	0.62	63.0	

粗粒度理解



细粒度理解

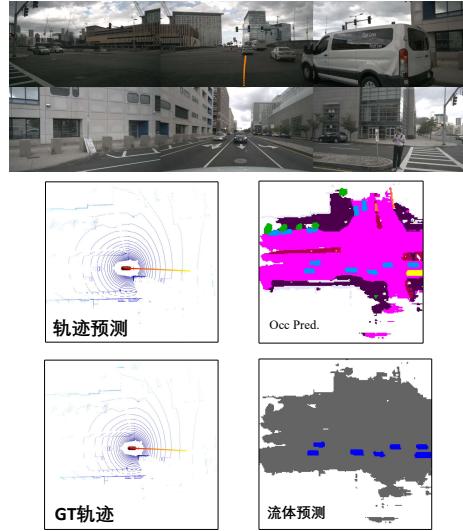


图4. DrivePI粗粒度与细粒度理解的可视化呈现。我们通过场景描述、三维占用率及轨迹预测结果展示DrivePI的粗粒度理解能力。

在占用流、动作状态和规划方面的对应性能分别为0.69的平均绝对误差（mAVE）、83.3%的准确率和0.62的碰撞率。最后，我们还评估了采用文本与视觉头联合训练的0.5B模型。研究发现，尽管QA准确率较低，但我们的0.5B模型在动作状态预测和L2误差方面表现更优，这证明了DrivePI在整合文本理解、三维感知、预测和规划方面的有效性。

4.5. 可视化

本节通过定性可视化演示，展示DrivePI的多粒度理解能力。如图4所示，我们展示了DrivePI在场景描述、三维空间占用、占用流、动作及轨迹预测方面的可视化效果。在粗粒度层面，DrivePI能生成详细的场景描述并基于不同指令提供合理答案；而在细粒度层面，则可输出精细结果（例如三维空间占用、占用流等）。

通过专用解码头实现轨迹规划，从而获得更精准的预测结果。此外，我们发现两个重要现象：首先，在场景描述方面，DrivePI展现出理解详细外观信息的能力（例如“天空部分多云”），这表明视觉编码器在将正面视角图像转换为鸟瞰视角表示时，仍能有效保留关键视觉信息。其次，我们观察到粗粒度与细粒度预测之间存在强一致性。例如，粗粒度语言描述的对象类别与网格坐标(65,136,7)处的细粒度三维空间占用预测相对应。同样，预测轨迹也表现出相似的一致性。这种跨层次理解的一致性验证了DrivePI的有效性——它将粗粒度语言空间理解与细粒度三维感知能力相结合，从而提升了自动驾驶系统的可解释性与决策可解释性。

5. 结论

本文提出了一种创新的VLA框架DrivePI，该框架不仅实现了文本格式的粗粒度空间理解，还具备与视觉表征（VA）模型相当的细粒度空间感知能力，从而兼具现有VA模型与VLA框架的优势。值得注意的是，DrivePI仅采用0.5B参数的大型语言模型（LLM）作为骨干网络，展现出卓越的效率。尽管仅使用紧凑的0.5B参数MLLM架构，DrivePI在三维空间占用及占用流预测方面仍优于现有VA模型，同时在自动驾驶领域的交互能力与现有VLA框架保持相当水平。我们期待这一新型VLA框架能为未来研究提供启发，通过语言推理和精细三维输出，提升自动驾驶系统的可解释性与决策能力。

局限性。我们的研究方法仍存在两个主要局限性，值得未来深入探讨。首先，我们采用简单的多任务学习策略来平衡不同任务间的损失权重，这种做法可能无法在相互竞争的目标间取得最佳平衡。其次，我们未采用强化学习技术——这类技术能通过试错学习提升推理能力，尤其在复杂规划场景中效果更佳。

参考文献

- [1] Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, Michael Bloesch, et al. Gemini机器人1.5:通过先进的具身推理，思维和运动转移，推动通用机器人的前沿.*arXiv预印本 arXiv:2510.03342,2025.2*
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, 等. Qwen技术报告.*arXiv预印本 arXiv:2309.16609,2023.2*
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, 等. Qwen2.5-vl 技术报告.*arXiv预印本 arXiv:2502.13923,2025.2,5*
- [4] 白泽晨、何彤、梅海阳、王皮超、高子腾、陈乔雅、张正和郑绍迈。一个代号来概括他们：语言指导视频中的推理分割。在NeurIPS 2024年会议上。*2*
- [5] 霍尔格·凯撒、瓦伦·班基蒂、亚历克斯·H·朗、苏拉布·沃拉、威尼斯·艾琳·李昂、徐强、阿努什·克里希南、潘宇、詹卡洛·巴尔丹和奥斯卡·贝伊博姆。场景：自动驾驶的多模态数据集。在CVPR, 2020。*5, 6, 3*
- [6] 杜冰陈、金芳、韩文成、程新静、尹俊波、徐成忠、法哈德·沙巴兹·汗、建-沈斌。Alocc: 基于自适应提升的三维语义占用与基于成本体积的流预测。在CVPR会议上，2025。*6*
- [7] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, 等. Internvl: 视觉基础模型的扩展与通用视觉-语言任务的对齐. 在CVPR,2024.*2*
- [8] 程志、许振佳、冯思远、Eric Cousineau、杜一伦、Benjamin Burchfiel、Russ Tedrake和宋书然。扩散策略：通过动作扩散实现视觉运动策略学习。国际机器人研究杂志，44 (10-11) :1684–1704,2025。*3*
- [9] 付浩宇、张电坤、赵宗创、崔建峰、梁定康、张冲、张定远、谢宏伟、王冰、白翔。Orion: 基于视觉-语言指导动作生成的端到端自动驾驶整体框架。在ICCV, 2025。*1, 2, 3, 6*
- [10] 甘万水、莫宁凯、徐洪斌和横尾直人。自动驾驶中三维占用估计的简易尝试。*arXiv预印本 arXiv:2303.10076, 2023.7*
- [11] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: 参数高效的视觉指令模型.*arXiv预印本 arXiv:2304.15010,2023.7*
- [12] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, 等. Seed1. 5-vl 技术报告.*arXiv预印本 arXiv:2505.07062,2025.2*
- [13] 胡培云、黄亚伦、约翰·多兰、大卫·赫尔德和德瓦·拉马南。基于自监督自由空间预测的安全局部运动规划。发表于CVPR, 2021。*6*
- [14] Yihan Hu, Jiazhai Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhui Wang, 等. 计划导向的自动驾驶. 在CVPR,2023.*1,3,6*
- [15] 黄俊杰和关黄。Bevdet4d: 利用时间线索进行多摄像头3D目标检测。*arXiv预印本 arXiv:2203.17054, 2022.6*
- [16] 黄俊杰、黄冠、朱正、叶云和杜大龙。摘要：鸟瞰视角下的高性能多相机3D物体检测。*arXiv预印本 arXiv:2112.11790, 2021.7*
- [17] 黄润辉、丁新鹏、王春伟、韩建华、刘玉龙、赵恒爽、徐航、侯璐、张伟和梁晓丹。Hires-llava: 高分辨率大型视觉语言模型中碎片化输入的恢复。在CVPR, 第29814–29824页, 2025。*4*
- [18] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, 等. Emma: 自动驾驶的端到端多模态模型.*arXiv预印本 arXiv:2410.23262,2024.3*
- [19] Anqing Jiang, Yu Gao, Zhigang Sun, Yiru Wang, Jijun Wang, Jinghao Chai, Qian Cao, Yuweng Heng, Hao Jiang,