# CarPlanner：大规模自动驾驶强化学习中的一致自回归轨迹规划

张东坤[1,2] 梁嘉铭[2] 郭可[2] 陆莎[1] 王启[2] 熊蓉[1,✉] 缪振伟[2,†] 王跃[1] [1]浙江大学 [2]菜鸟网络 [1] {zhangdongkun, lusha, rxiong, ywang24}@zju.edu.cn
[2] {liangjiaming.ljm, muguo.gk, ruifeng.wq, zhenwei.mzw}@alibaba-inc.com

## 摘要

*Trajectory planning is vital for autonomous driving, ensuring safe and efficient navigation in complex environments. While recent learning-based methods, particularly reinforcement learning (RL), have shown promise in specific scenarios, RL planners struggle with training inefficiencies and managing large-scale, real-world driving scenarios. In this paper, we introduce **CarPlanner**, a **C**onsistent **a**uto-**r**egressive **Planner** that uses RL to generate multi-modal trajectories. The auto-regressive structure enables efficient large-scale RL training, while the incorporation of consistency ensures stable policy learning by maintaining coherent temporal consistency across time steps. Moreover, CarPlanner employs a generation-selection framework with an expert-guided reward function and an invariant-view module, simplifying RL training and enhancing policy performance. Extensive analysis demonstrates that our proposed RL framework effectively addresses the challenges of training efficiency and performance enhancement, positioning CarPlanner as a promising solution for trajectory planning in autonomous driving. To the best of our knowledge, we are the first to demonstrate that the RL-based planner can surpass both IL- and rule-based state-of-the-arts (SOTAs) on the challenging large-scale real-world dataset nuPlan. Our proposed CarPlanner surpasses RL-, IL-, and rule-based SOTA approaches within this demanding dataset.*

## 1. 引言

轨迹规划[36]在自动驾驶中至关重要，它利用感知和轨迹预测模块的输出，为自车生成未来位姿。控制器会跟踪这个规划好的轨迹，为闭环驾驶生成控制指令。近年来，基于学习的轨迹规划方法因其优势而受到关注。

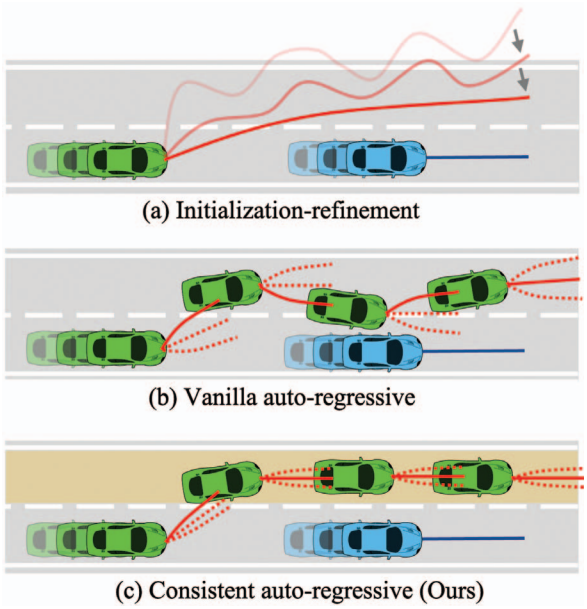


图1. 多步轨迹生成框架。(a) 初始化-优化框架：生成初始轨迹并迭代优化。(b) 标准自回归模型：顺序解码后续姿态。(c) 我们的一致性自回归模型：整合时间一致的模式信息。

具备自动化算法迭代、消除繁琐规则设计，并在多样化的现实场景中确保安全性和舒适性的潜力[36]。

现有大多数研究[11, 16, 29]采用模仿学习（IL）来使规划轨迹与人类专家的轨迹保持一致。然而，这种方法存在分布偏移[28]和因果混淆[8]的问题。强化学习（RL）提供了一个潜在的解决方案，通过奖励函数应对这些挑战并提供更丰富的监督信号。尽管RL在游戏[34]、机器人[19]和语言模型[25]等领域展现出有效性，但在大规模驾驶任务中仍面临训练效率低下和性能问题的困扰。

[†]Project lead. ✉ Corresponding author.

据我们所知，目前还没有强化学习方法能在nuPlan [2]等大规模开放数据集上取得有竞争力的结果，该数据集以多样化的真实场景为特色。

因此，本文旨在解决轨迹规划中强化学习面临的两大关键挑战：1）训练效率低下 2）性能不佳。训练效率低下源于强化学习通常在无模型环境中运行，需要依赖在CPU上运行的低效模拟器反复执行策略以收集数据。为克服这一挑战，我们提出一种高效的基于模型的方法，利用神经网络作为转移模型。该方法针对GPU等硬件加速器进行优化，使我们的时间成本与基于模仿学习的方法相当。

为了将强化学习应用于轨迹规划问题，我们将其构建为利用马尔可夫决策过程的多步序贯决策任务。现有分步生成轨迹{v*}的方法主要分为两类：初始化-优化[17, 20, 33, 45]和自回归模型[27, 32, 41, 46]。

第一类方法如图1(a)所示，首先生成初始轨迹估计，随后通过强化学习的迭代应用进行优化。然而包括Gen-Drive[18]在内的最新研究表明，该方法仍落后于最先进的模仿学习与基于规则的规划器。该方法的显著缺陷在于忽略了轨迹规划任务中固有的时序因果关系。此外，在高维轨迹空间直接进行优化的复杂性也会制约强化学习算法的性能。第二类方法为自回归模型（图1(b)），其通过转移模型中的单步策略递归生成自车位姿。此类方法将所有时间步的自车位姿整合形成完整规划轨迹。由于考虑了时序因果关系，当前自回归模型能够实现交互行为。但这类方法的共同局限在于依赖从动作分布中自回归随机采样以生成多模态轨迹。这种基础的自回归流程可能损害长期一致性，并不必要地扩大强化学习的探索空间，从而导致性能不佳。

为解决自回归模型的局限性，我们推出了CarPlanner——一种专为高效大规模基于强化学习的规划器训练而设计的连贯自回归模型（见图1(c)）。该模型的核心创新在于采用连贯模式表征作为自回归模型的条件输入。具体而言，我们采用纵向-横向解耦的模式表征方法，其中纵向模式为标量，用于捕捉平均...

[†]In this paper, the term “trajectory” refers to the future poses of the ego vehicle or traffic agents. To avoid confusion, we use the term “state (action) sequence” to refer to the “trajectory” in the RL community.

纵向模式关注速度，而横向模式涵盖了从自我车辆当前状态及地图信息中衍生的所有可能路径。该模式在不同时间步长中保持不变，在策略采样过程中提供稳定一致的引导。

此外，我们提出了一种适用于大规模多样化场景的通用奖励函数，无需针对特定场景设计奖励机制。该函数由专家引导项和任务导向项组成：首项通过量化智能体规划轨迹与专家轨迹之间的位移误差，结合一致性模式表征，有效缩小策略探索空间；次项融合了驾驶任务中的常识性要求，包括碰撞规避与可行驶区域遵守。我们还引入了不变视角模块（IVM），通过将智能体、地图及路径信息转换至当前自车坐标系，并裁剪远离自车的冗余信息，为策略提供与时间维度解耦的视角输入，从而简化特征学习并增强泛化能力。

***To our knowledge, we are the first to demonstrate that RL-based planner outperforms state-of-the-art (SOTA) IL and rule-based approaches on the challenging large-scale nuPlan dataset***综上所述，本文的主要贡献如下：

- 我们提出CarPlanner，一种能够生成一致多模态轨迹的一致性自回归规划器，通过训练强化学习策略来实现。
- 我们引入了一种专家指导的通用奖励函数和IVM，以简化强化学习训练并提升策略泛化能力，从而优化闭环性能。
- 我们对模仿学习与强化学习的训练特性进行了严谨分析，既揭示了两者的优势与局限，又凸显了强化学习在应对分布偏移和因果混淆等挑战时的独特优势。
- 我们的框架展现出卓越的性能，在nuPlan基准测试中超越了所有基于强化学习、模仿学习和规则的先进方法。这彰显了强化学习在复杂现实驾驶场景中的巨大潜力。

## 2. 相关工作

### 2.1. 基于模仿的规划

基于人类示范的模仿学习（IL）在训练规划器方面的应用近来引起了广泛关注。该方法充分利用经验丰富驾驶员的驾驶专业技能——他们能够安全舒适地应对各种现实场景，同时具备大规模轻松采集驾驶数据的附加优势[2,9,15]。

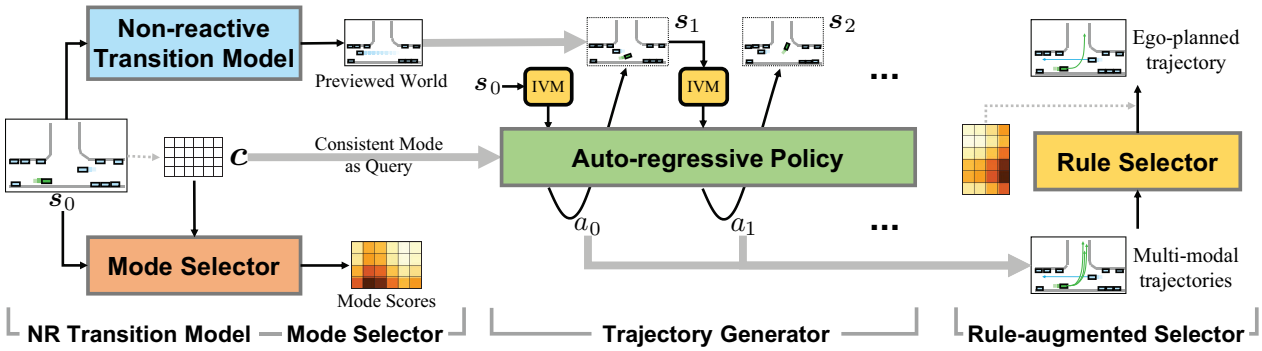


图2展示了CarPlanner的四个组成部分。(1) 非反应式转换模型以初始状态$\boldsymbol{s}_0$为输入，预测交通参与者的未来轨迹。(2) 模式选择器根据初始状态和模式$\boldsymbol{c}$输出评分。(3) 轨迹生成器采用自回归结构，在保持模式一致性的条件下生成与模式对齐的多模态轨迹。(4) 规则增强选择器通过安全性、舒适性和行进效率指标对模式评分进行补偿。

众多研究[5,17,29]致力于开发创新网络以提升该领域的开环性能。然而自动驾驶的终极挑战在于实现闭环操作，这需要通过驾驶导向的指标进行评估，例如安全性、交通规则遵守度、舒适性和行进效率。这揭示了规划器在训练与测试阶段存在的显著差距。此外，模仿学习尤其容易受到分布偏移[28]和因果混淆[8]等问题的影响。前者会导致系统在遇到训练数据分布未覆盖的场景时做出次优决策；后者则因网络依赖专家示范的模仿损失，可能无意中捕捉错误关联并基于输入信息形成捷径解决方案。尽管多项研究[1,3,4,42]已着力解决这些挑战，但训练与测试间的鸿沟依然显著。

### 2.2. 自动驾驶中的强化学习

在自动驾驶领域，强化学习已证明其在处理特定场景方面的有效性，例如高速公路驾驶[22,39]、车道变更[14,23]以及无保护左转[22,40]。大多数方法直接在控制空间上学习策略，包括油门、刹车和转向指令。由于控制指令执行频率高，仿真过程可能耗时且探索行为可能不一致[40]。部分研究[40,44]提出通过学习轨迹规划器来定义智能体规划轨迹作为行动，这既扩展了探索空间的时间维度，又提升了训练效率。但如ASAP-RL[40]所指出的，轨迹规划时长与训练性能之间存在权衡：增加轨迹时长会导致行为反应迟钝和数据量减少，而缩短轨迹时长则会遇到与控制空间类似的挑战。此外，这些方法通常采用无模型设定，使——由于它们难以应用于大规模驾驶数据集中复杂多样的现实场景，本文提出采用基于模型的框架，以促进在大规模数据集上的强化学习训练。在此框架下，我们旨在通过使用转移模型来克服轨迹跨度权衡问题，该模型能够在测试阶段为策略提供环境预览，从而实现多步决策。

## 3. 方法

### 3.1. 预备知识

MDP用于建模序列决策问题，其形式化定义为元组$\langle \mathcal{S}, \mathcal{A}, P_\tau, R, \rho_0, \gamma, T \rangle$。$\mathcal{S}$表示状态空间，$\mathcal{A}$表示动作空间。$P_\tau$:$\mathcal{S}\times\mathcal{A}\to \Delta(\mathcal{S})$ [†]描述状态转移概率，$R$:$\mathcal{S}\times\mathcal{A}\to \mathbb{R}$表示有界奖励函数。$\rho_0 \in \Delta(\mathcal{S})$是初始状态分布，$T$为时间跨度，$\gamma$是未来奖励的折扣因子。状态-动作序列定义为$\tau = (\boldsymbol{s}_0, a_0, \boldsymbol{s}_1, a_1, \ldots, \boldsymbol{s}_T)$，其中$\boldsymbol{s}_t \in \mathcal{S}$和$a_t \in \mathcal{A}$分别表示时间步$t$的状态与动作。强化学习的目标是最大化期望回报：

$$\max_{\pi} \mathbb{E}_{\boldsymbol{s}_t \sim P_\tau, a_t \sim \pi}[\sum_{t=0}^{T} \gamma^t R(\boldsymbol{s}_t, a_t)]. \tag{1}$$

向量化状态表示。状态 $\boldsymbol{s}_t$ 包含地图和智能体信息的向量化表示[10]。地图信息 $m$ 包括道路网络、交通信号灯等，通过折线和多边形进行表征。智能体信息包含自车及其他交通参与者的当前与历史位姿，这些信息通过折线表示。自车索引号为0，交通参与者索引号范围为1至 $N$。对于每个智能体 $i$，其历史轨迹记为 $s^i_{t-H:t}, i \in \{0, 1, \ldots, N\}$，其中 $H$ 表示历史时间跨度。

[†]$\Delta(\mathcal{X})$ denotes the set of probability distribution over set $\mathcal{X}$.

### 3.2. 问题表述

我们将轨迹规划任务建模为一个序列决策过程，并将自回归模型解耦为策略模型和转移模型。连接轨迹规划与自回归模型的关键在于将动作定义为自车下一时刻的位姿，即$a_t = s^0_{t+1}$。因此，在运行自回归模型后，解码得到的位姿会被收集为自车规划轨迹。具体而言，在此定义和向量化表示下，我们可以将状态-动作序列简化为状态序列：

$$\begin{aligned} &P(\boldsymbol{s}_0, a_0, \boldsymbol{s}_1, a_1, \ldots, \boldsymbol{s}_T) \\ &= P(m, s^{0:N}_{-H:0}, s^0_1, m, s^{0:N}_{1-H:1}, s^0_2, \ldots, m, s^{0:N}_{T-H:T}) \\ &= P(m, s^{0:N}_{-H:0}, m, s^{0:N}_{1-H:1}, \ldots, m, s^{0:N}_{T-H:T}) \\ &= P(\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T). \end{aligned} \tag{2}$$

状态序列可以进一步以自回归的方式建模，并分解为策略模型和转移模型：

$$\begin{aligned} P(\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T) &= \rho_0(\boldsymbol{s}_0) \prod_{t=0}^{T-1} P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t) \\ &= \rho_0(\boldsymbol{s}_0) \prod_{t=0}^{T-1} P(s^0_{t+1}, s^{1:N}_{t+1}|\boldsymbol{s}_t) \\ &= \rho_0(\boldsymbol{s}_0) \prod_{t=0}^{T-1} \underbrace{\pi(a_t|\boldsymbol{s}_t)}_{\text{Policy}} \underbrace{P_\tau(s^{1:N}_{t+1}|\boldsymbol{s}_t)}_{\text{Transition Model}}. \end{aligned} \tag{3}$$

从方程（3）中可以清晰看出典型自回归方法存在的固有问题：策略分布导致的时间步间行为不一致，该分布依赖于从动作分布中进行的随机采样。

为了解决上述问题，我们在自回归模式中引入了跨时间步保持一致的模态信息$\boldsymbol{c}$：

$$\begin{aligned} &P(\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T) = \int_{\boldsymbol{c}} P(\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T, \boldsymbol{c}) d\boldsymbol{c} \\ &= \rho_0(\boldsymbol{s}_0) \int_{\boldsymbol{c}} P(\boldsymbol{c}|\boldsymbol{s}_0) P(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T|\boldsymbol{c}) d\boldsymbol{c} \\ &= \rho_0(\boldsymbol{s}_0) \prod_{t=0}^{T-1} \underbrace{P_\tau(s^{1:N}_{t+1}|\boldsymbol{s}_t)}_{\text{Transition Model}} \int_{\boldsymbol{c}} \underbrace{P(\boldsymbol{c}|\boldsymbol{s}_0)}_{\text{Mode Selector}} \prod_{t=0}^{T-1} \underbrace{\pi(a_t|\boldsymbol{s}_t, \boldsymbol{c})}_{\text{Policy}} d\boldsymbol{c}. \end{aligned} \tag{4}$$

由于我们专注于自车轨迹规划，一致模式$\boldsymbol{c}$不会影响转移模型。

式（4）中定义的一致性自回归框架揭示了一个生成-选择机制：模式选择器基于初始状态$s_0$对每个模式进行评分，轨迹生成器则通过从模式条件策略中采样来生成多模态轨迹。

非反应式转换模型。公式（4）中构建的转换模型需要在每个时间步中使用。因为它根据当前状态$\boldsymbol{s}_t$生成交通参与者时间步$t+1$的位姿。实际应用中，这一过程较为耗时，且我们并未观察到使用该转换模型带来的性能提升，因此我们采用轨迹预测器$P(s^{1:N}_{1:T}|\boldsymbol{s}_0)$作为非反应式转换模型——该模型可根据初始状态$\boldsymbol{s}_0$一次性生成交通参与者的所有未来位姿。

### 3.3. 规划器架构

我们提出的CarPlanner框架如图2所示，包含四个关键组件：1）非反应式转换模型，2）模式选择器，3）轨迹生成器，以及4）规则增强选择器。

我们的规划器在生成-选择框架下运行。给定初始状态$s_0$和所有可能的$N_{\text{mode}}$模式，轨迹选择器会对每种模式进行评估并分配分数。随后轨迹生成器会生成与对应模式相匹配的$N_{\text{mode}}$条轨迹。对于轨迹生成器，初始状态$s_0$会被复制$N_{\text{mode}}$次，每次复制对应$N_{\text{mode}}$种模式中的一种，从而有效创建$N_{\text{mode}}$个并行世界。策略将在这些预演世界中执行。在策略推演过程中，轨迹预测器作为状态转移模型，生成所有时间范围内交通参与者的未来位姿。

#### 3.3.1. 非反应式转换模型

该模块以初始状态 $s_0$ 作为输入，并输出交通参与者的未来轨迹。初始状态通过智能体与地图编码器处理后，再经过自注意力Transformer编码器[38]融合智能体与地图特征。最终将智能体特征解码为未来轨迹。

智能体与地图编码器。状态$s_0$同时包含地图与智能体信息。地图信息$m$由$N_{m,1}$条折线与$N_{m,2}$个多边形构成。折线描述车道中心线与边界线，每条折线包含3$N_p$个点（分别对应车道中心、左边界与右边界）。每个点具有$D_m$ =9维特征，包含以下属性：x坐标、y坐标、航向角、限速值与类别。将左右边界点与中心点拼接后，形成$N_{m,1} \times N_p \times 3D_m$维特征。我们采用PointNet[26]从每条折线的点集中提取特征，得到$N_{m,1} \times D$维输出（其中$D$表示特征维度）。多边形则表征交叉口、人行横道、停止线等区域，每个多边形包含$N_p$个点。我们使用另一个PointNet从多边形点集提取特征，生成$N_{m,2} \times D$维输出。随后将折线与多边形的特征拼接形成完整地图特征，最终维度为$N_m \times D$。智能体信息$A$包含$N$个智能体，每个智能体记录过去$H$个时间步的位姿。每位姿具有$D_a$ =10维特征，包含

具有以下属性：x坐标、y坐标、航向角、速度、边界框、时间步长和类别。因此，智能体信息维度为$N \times H \times D_a$。我们应用另一个PointNet从每个智能体的姿态中提取特征，得到智能体特征维度为$N \times D$。

3.3.2. 模式选择器

该模块以$s_0$和纵横分解的模式信息作为输入，输出每种模式的概率。模式数量为$N_{\text{mode}} = N_{\text{lat}} N_{\text{lon}}$。

路由-速度分解模式。为捕捉纵向行为，我们生成代表每种模式对应轨迹平均速度的$N_{\text{lon}}$个纵向模式。每个纵向模式$c_{\text{lon},j}$定义为标量值$\frac{j}{N_{\text{lon}}}$，沿维度$D$重复扩展。因此纵向模式的维度为$N_{\text{lon}} \times D$。针对横向行为，我们通过图搜索算法从地图中识别出$N_{\text{lat}}$条可行路径，这些路径对应自车可用的车道。路径的维度为$N_{\text{lat}} \times N_r \times D_m$。我们采用另一个PointNet聚合每条路径上$N_r$个点的特征，生成维度为$N_{\text{lat}} \times D$的横向模式。为构建完整的模式表示$\boldsymbol{c}$，我们将横向与纵向模式组合，形成$N_{\text{lat}} \times N_{\text{lon}} \times \times 2D$的联合维度。为使模式信息与其他特征维度对齐，我们将其传入

通过一个线性层，将其映射回 $N_{\text{lat}} \times N_{\text{lon}} \times D$。
基于查询的Transformer解码器。该解码器用于将模式特征与从{v*}中提取的地图和智能体特征进行融合。在此框架中，模式作为查询，而地图和智能体信息则充当键和值。更新后的模式特征通过多层感知机（MLP）解码，生成每种模式的得分，最后通过softmax算子进行归一化处理。

3.3.3. 轨迹生成器

该模块以自回归方式运行，根据当前状态$\boldsymbol{s}_t$和一致的模式信息$\boldsymbol{c}$，循环解码自车$a_t$的下一姿态。

不变视角模块（IVM）。在将模式与状态输入网络之前，我们对其进行预处理以消除时间信息。针对状态$\boldsymbol{s}_t$中的地图与智能体信息，我们选取与自车当前位置最近的$K$个近邻（KNN）仅输入策略中。$K$分别设置为地图与智能体元素数量的一半。对于表征横向行为的路线，我们过滤掉最靠近自车当前位置的点为起点的路段，保留$K_r$个点。此处$K_r$设置为单条路线中$N_r$点数的四分之一。最后，我们将路线、智能体及地图坐标转换到当前时间步$t$的自车坐标系下。通过从当前时间步$t$中减去历史时间步$t - H$：$t$，得到时间步范围$-H$：0。

基于查询的Transformer解码器。我们采用与模式选择器相同的主干网络架构，但查询维度不同。由于IVM以及不同模式会产生不同状态，地图和智能体信息无法在模式间共享。因此，我们为每个独立模式分别进行信息融合。具体而言，查询维度为$1 \times D$，而键与值的维度为$(N + N_m) \times D$。输出特征维度保持$1 \times D$。需要注意的是，Transformer解码器能够并行处理多模式信息，无需按顺序逐个处理模式。

策略输出。模式特征由两个独立的头部处理：策略头部和价值头部。每个头部包含其自身的MLP，用于生成动作分布的参数和相应的价值估计。我们采用高斯分布来建模动作分布，在训练期间从该分布中采样动作。相反，在推理过程中，我们利用分布的均值来确定动作。

3.3.4. 规则增强选择器

该模块仅在推理过程中使用，输入初始状态$s_0$、多模态自我规划轨迹以及预测的智能体未来轨迹。它会计算驾驶导向的指标，如安全性、行进度和舒适度。通过基于规则的分数与模式选择器提供的模式分数进行加权求和，获得综合评分。最终选择得分最高的自我规划轨迹作为规划器的输出。

## 3.4. 训练

我们首先训练非反应式过渡模型，并在模式选择器和轨迹生成器的训练过程中冻结权重。我们采用赢家通吃策略，而非将所有模式输入生成器——该策略会根据自车真实轨迹指定一个正向模式，并将其作为轨迹生成器的条件输入。

模式分配。对于横向模式，我们将最接近自车真实轨迹终点的路线指定为正横向模式。对于纵向模式，我们将纵向空间划分为$N_{\text{lon}}$个区间，并将包含真实轨迹终点的区间指定为正纵向模式。

奖励函数。为处理多样化的场景，我们使用自车未来位姿与真实轨迹之间的负位移误差（DE）作为通用奖励。同时引入额外项以提升轨迹质量：碰撞率与可行驶区域合规性。若未来位姿发生碰撞或超出可行驶区域，奖励值设为-1；否则设为0。

模式丢弃。在某些情况下，自车没有可用的路线可供遵循。然而，由于路线在Transformer中充当查询，缺少路线可能导致不稳定或危险的输出。为了缓解这个问题，我们实施了

在训练过程中，模式丢弃模块会随机屏蔽路由路径，以防止对此信息的过度依赖。

损失函数。对于选择器，我们采用交叉熵损失，即正样本模式的负对数似然，以及一个回归自车真实轨迹的辅助任务。对于生成器，我们使用PPO[31]损失，该损失包含三部分：策略改进、价值估计和熵。完整描述可在补充材料中找到。

# 4. 实验

## 4.1. 实验设置

数据集与模拟器。我们采用nuPlan [2]——一个用于研究自动驾驶轨迹规划的大规模闭环平台——来评估方法的有效性。该数据集包含1,500小时由人类专家驾驶员在4个不同城市采集的行车日志数据，涵盖车道保持与变更、左右转弯、通过交叉路口及公交站点、环岛通行、行人交互等复杂多样场景。作为闭环平台，nuPlan提供的模拟器以数据集场景初始化仿真环境：交通参与者由日志回放（非交互式）或IDM[37]策略（交互式）控制，本车则由用户提供的规划器控制。模拟器以10Hz频率运行15秒，每个时间戳向规划器查询轨迹方案，并通过LQR控制器跟踪轨迹生成车辆控制指令。

基准与指标。我们采用两个基准：Test14-Random和Reduced-Val14，用于与其他方法进行比较并分析我们方法内部的设计选择。PlanTF [4]提供的Test14-Random包含261个场景，PDM [7]提供的Reduced-Val14包含318个场景。

我们使用官方nuPlan开发工具包[†]提供的闭环分数（CLS）来评估所有方法的性能。该CLS分数涵盖安全（S-CR、S-TTC）、可行驶区域合规性（S-Area）、进度（S-PR）、舒适度等多个维度。根据交通参与者的不同行为类型，CLS进一步细分为CLS-NR（非反应型）和CLS-R（反应型）。

实现细节。我们遵循PDM [7]的方法构建训练和验证集。训练集规模为176,218个场景，涵盖所有可用场景类型，每种类型包含4,000个场景。验证集规模为1,118个场景，从14种类型中各选取100个场景。所有模型均在2张NVIDIA 3090 GPU上训练50个周期，单GPU批处理大小为64。采用AdamW优化器，初始学习率为1e-4，当验证损失停止下降时立即以0.3的衰减系数降低学习率（耐心值设为0）。对于强化学习训练，

[†]https://github.com/motional/nuplan-devkit

| Type | Planners | CLS-NR | CLS-R |
|---|---|---|---|
| Rule | IDM [37] | 70.39 | 72.42 |
| | PDM-Closed [7] | 90.05 | **91.64** |
| IL | RasterModel [2] | 69.66 | 67.54 |
| | UrbanDriver [29] | 63.27 | 61.02 |
| | GC-PGP [12] | 55.99 | 51.39 |
| | PDM-Open [7] | 52.80 | 57.23 |
| | GameFormer [17] | 80.80 | 79.31 |
| | PlanTF [4] | 86.48 | 80.59 |
| | PEP [42] | 91.45 | 89.74 |
| | PLUTO [3] | <u>91.92</u> | 90.03 |
| RL | CarPlanner (Ours) | **94.07** | <u>91.1</u> |

表1. 在Test14-Random中与SOTAs的对比。根据轨迹生成器的类型，所有方法被分为规则、模仿学习和强化学习三类。最佳结果以**粗体**标出，次佳结果以下划线标示。

| Type | Planners | CLS-NR | S-CR | S-PR |
|---|---|---|---|---|
| Rule | PDM-Closed [7] | <u>91.21</u> | **97.01** | <u>92.68</u> |
| IL | GameFormer [17] | 83.76 | 94.73 | 88.12 |
| | PlanTF [4] | 83.66 | 94.02 | 92.67 |
| | Gen-Drive (Pretrain) [18] | 85.12 | 93.65 | 86.64 |
| RL | Gen-Drive (Finetune) [18] | 87.53 | 95.72 | 89.94 |
| | CarPlanner (Ours) | **91.45** | <u>96.38</u> | **95.37** |

表2. 在非反应式交通代理的Reduced-Val14中与SOTAs的对比。

我们将折扣系数$\gamma$ =设为0.1，GAE参数$\lambda$ =设为0.9。价值函数、策略函数和熵损失的权重分别设置为3、100和0.001。纵向模态数量设定为12个，横向模态最大数量设定为5个。

## 4.2. 与先进技术的比较

SOTAs。我们根据轨迹生成器的类型将方法分为规则法、模仿学习法和强化学习法。（1）PDM [7] 赢得了2023年nuPlan挑战赛，其基于模仿学习和基于规则的变体分别记为PDM-Open和PDM-Closed。PDM-Closed遵循生成-选择框架，使用IDM生成多条候选轨迹，并采用基于规则的选择器（综合考虑安全性、行进度和舒适性）来筛选最优轨迹。（2）PLUTO [3] 同样遵循生成-选择框架，通过对比式模仿学习融合多种数据增强技术来训练生成器。（3）GenDrive [18] 是同期提出的工作，采用预训练-微调流程：先通过模仿学习预训练基于扩散模型的规划器，再基于通过AI偏好训练的奖励模型，使用强化学习对去噪过程进行微调。

结果。我们在Test14-Random和Reduced-Val14基准测试中将我们的方法与SOTAs进行比较，结果如表1和表2所示。总体而言，我们的CarPlanner展现出卓越性能，尤其在非反应式环境中表现突出。

在非反应式设置中，我们的方法在所有指标上均取得了最高分，相较于PDM-Closed和PLUTO分别提升了4.02和2.15，这证实了强化学习的潜力及其卓越性能。

| Design Choices | | | | Closed-loop metrics (↑) | | | | | Open-loop losses (↓) | |
|---|---|---|---|---|---|---|---|---|---|---|
| Reward DE | Reward Quality | Coord Trans | KNN | CLS-NR | S-CR | S-Area | S-PR | S-Comfort | Loss Selector | Loss Generator |
| ✗ | ✓ | ✓ | ✓ | 31.79 | 95.74 | 98.45 | 33.10 | 48.84 | 1.03 | **30.3** |
| ✓ | ✗ | ✓ | ✓ | 90.44 | 97.49 | 96.91 | 93.33 | 90.73 | **0.99** | 1221.6 |
| ✓ | ✓ | ✗ | ✓ | 90.78 | 96.92 | 98.46 | 91.37 | **94.23** | 1.00 | 2130.7 |
| ✓ | ✓ | ✓ | ✗ | 92.73 | 98.07 | 98.46 | 94.69 | 93.44 | 1.03 | 2083.6 |
| ✓ | ✓ | ✓ | ✓ | **94.07** | **99.22** | **99.22** | **95.06** | 91.09 | 1.03 | 1624.5 |

表3：强化学习训练中设计选择的消融研究。结果基于Test14随机非反应基准测试。

我们提出的框架。此外，CarPlanner在进度指标S-PR上相比Tab.2中的PDM-Closed展现出显著提升，同时保持可比的碰撞指标S-CR，这表明我们的方法在提升驾驶效率的同时能够维持安全驾驶。值得注意的是，我们未采用模仿学习中常用的数据增强[3,4]或自我历史掩码[11]等技术，这凸显了我们方法在解决闭环任务时的内在能力。

在反应式设置中，虽然我们的方法表现良好，但略逊于PDM-Closed。这种差异的出现是因为我们的模型仅在非反应式设置中进行训练，未曾与反应式设置所使用的IDM策略产生交互；因此，在测试过程中，我们的模型对反应式智能体产生的干扰鲁棒性较弱。

### 4.3. 消融实验

我们研究了强化学习训练中不同设计选择的影响，结果如表3所示。

奖励项的影响。当仅使用质量奖励时，规划器倾向于生成静态轨迹，导致进度指标较低。这是因为自车起始于安全可行驶状态，但前进时可能面临碰撞或驶离可行驶区域的风险。另一方面，当质量奖励与DE奖励结合使用时，相较于单独使用DE奖励，闭环指标实现显著提升。例如S-CR指标从97.49升至99.22，S-Area指标从96.91升至99.22。这些改进表明质量奖励能够激励安全舒适的驾驶行为。

IVM的有效性。结果表明，IVM中的坐标变换和KNN技术显著提升了闭环指标与生成器损失。例如，采用坐标变换技术后，整体闭环分数从90.78提升至94.07，S-PR值从91.37增至95.06。这些改进源于强化学习中价值估计精度的提升，从而实现了闭环场景下的泛化驾驶。

### 4.4. 扩展到IL

除了为强化学习训练设计之外，我们还扩展了CarPlanner以整合模仿学习。我们通过严谨的分析比较模仿学习中不同设计选择的影响，并

强化学习训练过程总结在表4中。我们的研究结果表明，虽然模式丢弃和选择器辅助任务对模仿学习和强化学习都有促进作用，但在模仿学习中常显效的自我历史丢弃和骨干网络共享机制，却不太适用于强化学习训练。

自我历史丢弃。先前的研究[1, 3, 4, 11]表明，通过模仿学习训练出的规划器可能过度依赖过去位姿而忽略环境状态信息。为解决这一问题，我们将ChauffeurNet[1]与PlanTF[4]的技术结合为自我历史丢弃模块，通过随机掩码自我历史位姿与当前速度{v*}来缓解因果混淆问题。

我们的实验证实，自我历史丢弃技术有助于模仿学习训练，因为它能提升闭环指标（如S-CR和S-Area）的表现。然而在强化学习训练中，我们观察到自我历史丢弃对优势估计产生了负面影响，这显著影响生成器损失中的价值部分，最终导致闭环性能下降。这表明强化学习能通过揭示与奖励信号相一致的因果关系，自然解决模仿学习中固有的因果混淆问题——该奖励信号显式编码了任务导向的偏好。这种能力凸显了强化学习在推动基于学习的规划边界方面的潜力。

骨干共享。这一选择常用于基于模仿学习的多模态规划器，旨在通过跨任务特征共享提升泛化能力。尽管骨干共享能通过平衡轨迹生成器与选择器的损失来助力模仿学习，但我们发现它会反向影响强化学习训练。具体而言，骨干共享会导致强化学习中轨迹生成器和选择器的损失同时升高，表明各任务梯度存在相互干扰。轨迹生成与选择任务在强化学习中的目标差异会产生冲突，从而降低整体策略性能。因此，我们在强化学习框架中避免使用骨干共享，以保持任务特定的梯度流并提升策略质量。

### 4.5. 定性结果

我们提供了如图3所示的定性结果。该场景要求自车在右转过程中绕行行人。在这种情况下，我们的方法展现出平滑高效的性能。从$t_{\text{sim}} = 0s$到$t_{\text{sim}} = 9s$，所有方法都等待行人通过路口。至$t_{\text{sim}} = 10s$时，一名行人突然折返准备重新过街。PDM-Closed方案则

| | Design Choices | | | | Closed-loop metrics (↑) | | | | | Open-loop losses (↓) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Loss Type | Mode Dropout | Selector Side Task | Ego-history Dropout | Backbone Sharing | CLS-NR | S-CR | S-Area | S-PR | S-Comfort | Loss Selector | Loss Generator |
| IL | ✗ | ✗ | ✗ | ✗ | 90.82 | 97.29 | 98.45 | 92.15 | 94.57 | 1.04 | **147.5** |
| | ✓ | ✗ | ✗ | ✗ | 91.21 | 96.54 | 98.46 | 91.44 | **96.92** | 1.07 | 153.0 |
| | ✓ | ✓ | ✗ | ✗ | 91.51 | 96.91 | 98.46 | **95.30** | 96.91 | 1.04 | 162.3 |
| | ✓ | ✓ | ✓ | ✗ | 92.72 | 98.06 | 98.84 | 94.88 | 95.35 | 1.04 | 167.5 |
| | ✓ | ✓ | ✓ | ✓ | 93.41 | 98.85 | 98.85 | 93.87 | 96.15 | 1.04 | 174.3 |
| RL | ✗ | ✗ | ✗ | ✗ | 91.67 | 98.84 | 98.84 | 91.69 | 90.73 | 1.04 | 1812.6 |
| | ✓ | ✗ | ✗ | ✗ | 93.46 | 98.07 | **99.61** | 94.26 | 92.28 | 1.09 | 2254.6 |
| | ✓ | ✓ | ✗ | ✗ | **94.07** | **99.22** | 99.22 | 95.06 | 91.09 | **1.03** | 1624.5 |
| | ✓ | ✓ | ✓ | ✗ | 89.51 | 97.27 | 98.44 | 90.93 | 83.20 | 1.05 | 5424.3 |
| | ✓ | ✓ | ✓ | ✓ | 88.66 | 95.54 | 98.84 | 92.82 | 86.05 | 1.21 | 1928.1 |

表4. 使用我们的CarPlanner时不同组件对IL和RL损失的影响。结果基于Test14-random非交互基准测试得出。

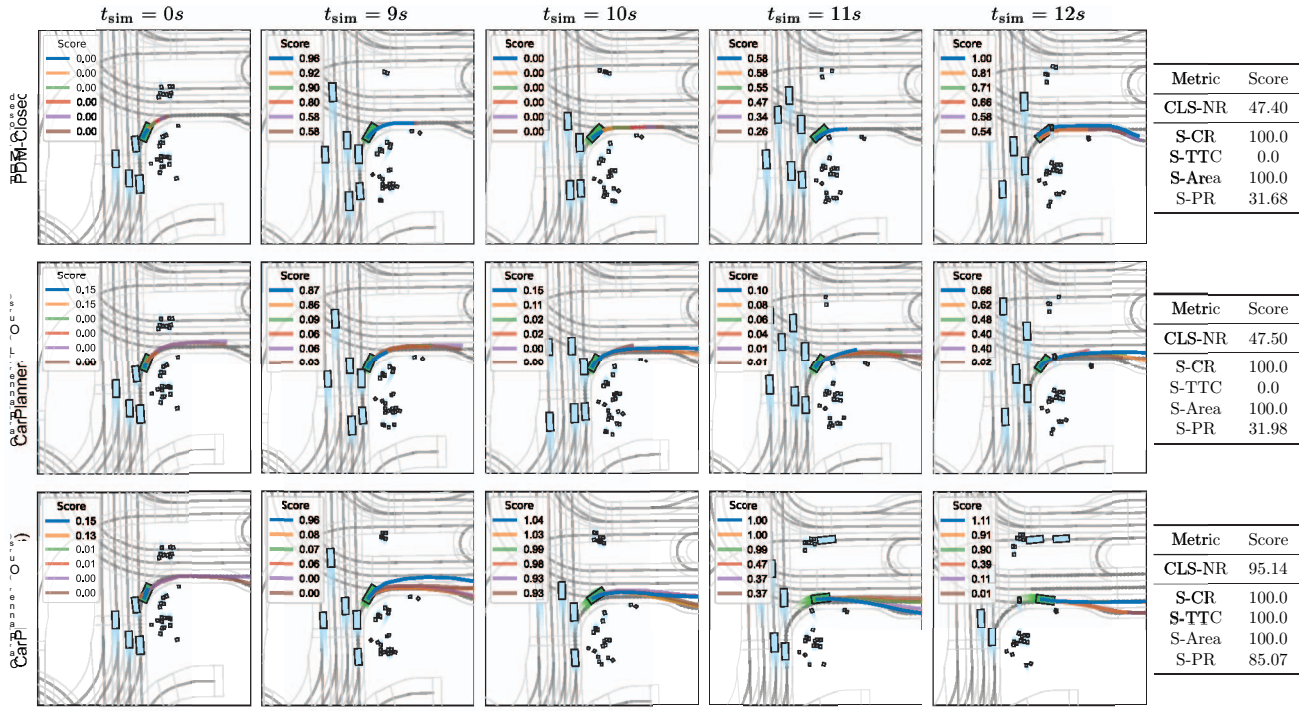


图3. 在非反应性环境中PDM-Closed与我们方法的定性对比。该场景被标注为等待行人通过。每个帧画面中，自车标记为绿色。交通参与者标记为天蓝色。蓝色线状图为自车规划轨迹。

不了解这一情况并采取了紧急制动，但仍然与这名行人发生了交汇。相比之下，我们的IL变体能够感知行人的移动趋势，因此实施了制动操作，但始终与行人保持较近距离。我们的RL方法通过提前启动（最高速度达到$t_{\text{sim}}$ =9$s$）成功规避了该风险，同时实现了最高的行进效率与安全指标。

## 5. 结论

本文介绍了CarPlanner——一种面向大规模强化学习训练的一致性自回归规划器。得益于所提出的框架，我们训练的基于强化学习的规划器在性能上超越了现有基于强化学习、模仿学习及规则的先进方法。此外，我们还通过分析表明IL和RL的特点，突显了RL在向基于学习的规划迈出更进一步的潜力。

局限性与未来工作。强化学习需要精细设计，且易受输入表示影响。该方法可能过度拟合训练环境，并在未见环境中出现性能下降[21]。我们利用专家辅助的奖励设计来引导探索，但这种方式可能限制强化学习的全部潜力——因其本质上依赖专家示范，可能阻碍发现超越人类专业知识的解决方案。未来工作将致力于开发能够克服这些局限的鲁棒强化学习算法，实现在多元环境中的自主探索与泛化。

## 6. 致谢

衷心感谢王静柯在有益讨论中的贡献以及所有审稿人对本文改进的帮助。本研究由浙江省自然科学基金（项目编号：LD24F030001）和国家自然科学基金（项目编号：62373322）资助。

## 参考文献

[1] Mayank Bansal、Alex Krizhevsky与Abhijit Ogale。ChauffeurNet：通过模仿最佳路径与合成最差场景学习驾驶。发表于*Proc. of Robotics: Science and Systems*，2019年。3, 7 [2] Holger Caesar、Juraj Kabzan、Kok Seang Tan、Whye Kit Fong、Eric Wolff、Alex Lang、Luke Fletcher、Oscar Beijbom与Sammy Omari。NuPlan：基于闭环机器学习的自动驾驶规划基准。*arXiv preprintarXiv:2106.11810*，2021年。2, 6, 12 [3] 程杰、陈英兵与陈启峰。Pluto：突破基于模仿学习的自动驾驶规划极限。*arXiv preprint arXiv:2404.14327*，2024年。3, 6, 7, 12, 13 [4] 程杰、陈英兵、梅晓东、杨博文、李博与刘明。重新思考基于模仿学习的自动驾驶规划器。载于*Proc. of the IEEE Intl. Conf. onRobotics & Automation*，页14123–14130。IEEE，2024年。3, 6, 7 [5] Kashyap Chitta、Aditya Prakash、Bernhard Jaeger、余泽浩、Katrin Renz与Andreas Geiger。TransFuser：基于Transformer传感器融合的自动驾驶模仿学习。*IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*，2022年。3 [6] Ignasi Clavera、傅遥与Pieter Abbeel。模型增强行动者-评论家：通过路径反向传播。发表于*Proc. of the Int. Conf. on Learning Representations*，2020年。15 [7] Daniel Dauner、Marcel Hallgarten、Andreas Geiger与Kashyap Chitta。破除基于学习的车辆运动规划误区。载于*Proc. of the Conf. onRobot Learning*，2023年。6 [8] Pim De Haan、Dinesh Jayaraman与Sergey Levine。模仿学习中的因果混淆问题。*Proc. of the Advances in Neural Information Processing Systems*，32卷，2019年。1, 3 [9] Scott Ettinger、程舒阳、Benjamin Caine、刘晨曦、赵航、Sabeek Pradhan、柴宇宁、Ben Sapp、齐晨睿、周寅等。自动驾驶的大规模交互式运动预测：Waymo开放运动数据集。载于*Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*，页9710–9719，2021年。2, 12

[10] 高继阳、孙晨、赵航、沈毅、德拉戈米尔·安格洛夫、李聪聪、科迪莉亚·施密德。《Vectornet：基于矢量化表示的高清地图与智能体动态编码》。见*Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*，第11525–11533页，2020年。3

[11] 郭可、景伟、陈军波、潘佳。CCIL：面向城市驾驶的上下文条件模仿学习。发表于《*Proc. of Robotics: Science and Systems*》，2023年。第1、7页 [12] Marcel Hallgarten、Martin Stoll、Andreas Zell。从预测到规划：基于目标条件的车道图遍历。《*arXiv preprint arXiv:2302.07753*》，2023年。第6页 [13] Nicklas A Hansen、苏浩、王晓龙。模型预测控制中的时序差分学习。收录于《*Proc. of the Intl. Conf. on Machine Learning*》，第8387–8406页。PMLR，2022年。第15页 [14] 何祥坤、杨浩晗、胡忠旭、吕辰。自动驾驶车辆鲁棒换道决策：基于观测对抗强化学习的方法。《*IEEE Transactions on Intelligent Vehicles*》，第8卷第1期，第184–193页，2022年。第3页 [15] John Houston、Guido Zuidhof、Luca Bergamini、叶亚威、陈龙、Ashesh Jain、Sammy Omari、Vladimir Iglovikov、Peter Ondruska。一千零一小时：自动驾驶运动预测数据集。收录于《*Proc. of the Conf. on Robot Learning*》，第409–418页，2021年。第2页 [16] 胡艺涵、杨嘉植、陈立、李科宇、司马崇皓、朱熙洲、柴思琪、杜森尧、林天威、王文海等。面向规划任务的自动驾驶。收录于《*Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*》，第17853–17862页，2023年。第1页 [17] 黄智宇、刘浩宸、吕辰。Gameformer：基于博弈论的自动驾驶交互预测与规划Transformer建模与学习。收录于《*Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*》，第3903–3913页，2023年。第2、3、6、13页 [18] 黄智宇、Weng Xinshuo、Maximilian Igl、陈宇霄、曹玉龙、Boris Ivanovic、Marco Pavone、吕辰。Gen-drive：通过奖励建模与强化学习微调增强扩散生成驾驶策略。《*arXiv preprint arXiv:2410.05582*》，2024年。第2、6、12页 [19] Julian Ibarz、Jie Tan、Chelsea Finn、Mrinal Kalakrishnan、Peter Pastor、Sergey Levine。深度强化学习训练机器人方法论：《*Intl. Journal of Robotics Research*》第40卷第4-5期，第698–721页，2021年。第1页 [20] 蒋驰宇、Andre Cornman、Cheolho Park、Benjamin Sapp、周寅、Dragomir Anguelov等。MotionDiffuser：基于扩散的可控多智能体运动预测。收录于《*Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*》，第9644–9653页，2023年。第2页 [21] Robert Kirk、Amy Zhang、Edward Grefenstette、Tim Rocktäschel。深度强化学习中零样本泛化研究综述。《*Journal of Artificial Intelligence Re- search*》第76卷，第201–264页，2023年。第8页 [22] Edouard Leurent、Jean Mercat。密集交通中自主决策的社会注意力机制。《*arXiv preprint arXiv:1911.12250*》，2019年。第3页 [23] 李国法、杨逸凡、李申、曲行达、吕能超、李升波。风险感知的自动驾驶换道场景决策：深度强化学习方法。《*Transportation research part C: emerging technologies*》第134卷，第103452页，2022年。第3页

[24] 李全意、彭正浩、冯岚、刘志政、段晨达、莫文杰、周博磊。ScenarioNet：大规模交通场景仿真与建模的开源平台。*Proc. of the Advances in Neural Information Processing Systems*，第36卷：3894–3920，2023年。12 [25] 欧阳龙、吴杰弗里、江旭、阿尔梅达·迪奥戈、温赖特·卡罗尔、米什金·帕梅拉、张冲、阿加瓦尔·桑迪尼、斯拉玛·卡塔琳娜、雷·亚历克斯等。通过人类反馈训练语言模型遵循指令。*Proc. of the Advances in Neural Information Processing Systems*，第35卷：27730–27744，2022年。1 [26] 祁驰航、苏浩、莫开春、吉巴斯·莱奥尼达斯。PointNet：基于点集的3D分类与分割深度学习模型。见*Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*，第652–660页，2017年。4 [27] 尼古拉斯·莱因哈特、罗恩·麦卡利斯特、克里斯·北谷、谢尔盖·莱文。Precog：视觉多智能体环境中基于目标的预测。见*Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*，第2821–2830页，2019年。2 [28] 罗斯·斯蒂芬、戈登·杰弗里、巴格内尔·德鲁。模仿学习与结构化预测到无遗憾在线学习的归约。见*Proceedings of the fourteenth international conference on artificial intelligence and statistics*，第627–635页，2011年。1,3 [29] 奥利弗·舍尔、卢卡·贝加米尼、马切伊·沃尔奇克、布拉泽伊·奥辛斯基、彼得·翁德鲁斯卡。Urban Driver：通过策略梯度从真实世界演示中学习驾驶。见*Proc. of the Conf. on Robot Learning*，第718–728页，2022年。1,3,6 [30] 约翰·舒尔曼、菲利普·莫里茨、谢尔盖·莱文、迈克尔·乔丹、彼得·阿比尔。基于广义优势估计的高维连续控制。*arXiv preprint arXiv:1506.02438*，2015年。11,12 [31] 约翰·舒尔曼、菲利普·沃尔斯基、普拉富拉·达里瓦尔、亚历克·拉德福德、奥列格·克利莫夫。近端策略优化算法。*arXiv preprint arXiv:1707.06347*，2017年。6,11 [32] 阿里·塞夫、布莱恩·塞拉、陈典、吴梅森、周奥里克、纳亚坎蒂·尼格马、里法特·哈立德、阿尔鲁富·拉米、萨普·本杰明。MotionLM：作为语言建模的多智能体运动预测。见*Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*，第8579–8590页，2023年。2 [33] 史少帅、蒋立、戴登新、希勒·伯恩特。具有全局意图定位与局部运动细化的运动Transformer。*Proc. of the Advances in Neural Information Processing Systems*，第35卷：6531–6543，2022年。2 [34] 戴维·西尔弗、黄阿贾、麦迪逊·克里斯、盖兹·亚瑟、西弗雷·洛朗、范登德里斯切·乔治、施里特维泽·朱利安、安东oglou·约阿尼斯、潘尼尔斯尔瓦姆·维达、兰克托特·马克等。基于深度神经网络与树搜索掌握围棋游戏。*nature*，第529卷(第7587期)：484–489，2016年。1 [35] 索西蒙、雷加拉多·塞巴斯蒂安、卡萨斯·塞尔吉奥、乌尔塔松·拉克尔。TrafficSim：学习模拟真实多智能体行为。见*Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*，第10400–10409页，2021年。13 [36] Ardi Tampuu, Tambet Matiisen, Maksym Semikin, Dmytro Fishman, Naveed Muhammad. 端到端驾驶技术综述：架构与训练方法. *IEEE Transactions on Neural Networks and Learning Systems*, 33(4): 1364–1384, 2020. 1 [37] Martin Treiber, Ansgar Hennecke, Dirk Helbing. 经验观测与微观模拟中的拥堵交通状态. *Physical review E*, 62(2):1805, 2000. 6 [38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, Illia Polosukhin. 注意力机制即是全部. *Proc. of the Advances in Neural Information Processing Systems*, 30, 2017. 4 [39] 王焕杰, 袁士华, 郭梦雨, 李学远, 蓝威. 基于深度强化学习的高速公路匝道合流自动驾驶方法. *Proceedings of the Institution of Mechanical engineers, Part D: Journal of Automobile engineering*, 235(10-11):2726–2739, 2021. 3 [40] 王乐天, 刘杰, 邵昊, 王文硕, 陈若冰, 刘宇, Steven L Waslander. 参数化技能与先验驱动的自动驾驶高效强化学习. 见 *Proc. of Robotics: Science and Systems*, 韩国大邱, 2023. 3 [41] 吴炜, 冯晓欣, 高子严, 阚宇恒. SMART：基于下一令牌预测的可扩展多智能体实时仿真. *arXiv preprint arXiv:2405.15677*, 2024. 2 [42] 张东坤, 梁家铭, 卢莎, 郭珂, 王琦, 熊蓉, 苗振伟, 王悦. PEP：嵌入式策略的自动驾驶轨迹规划. *IEEE Robotics and Automation Letters*, 2024. 3, 6 [43] 张哲俊, Alexander Liniger, Christos Sakaridis, Fisher Yu, Luc V Gool. 基于相对位姿编码的异构折线变换器实时运动预测. *Proc. of the Advances in Neural Information Processing Sys- tems*, 36, 2024. 13 [44] 周通, 王乐天, 陈若冰, 王文硕, 刘宇. 利用任务无关的以自我为中心运动技能加速自动驾驶强化学习. 见 *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 页11289–11296. IEEE, 2023. 3 [45] 周扬, 邵昊, 王乐天, Steven L Waslander, 李洪胜, 刘宇. SmartRefine：面向高效运动预测的场景自适应优化框架. 见 *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 页15281–15290, 2024. 2 [46] 周子康, 胡海波, 陈鑫宏, 王建平, 关楠, 吴魁, 李永辉, 黄昱凯, 薛春杰. BehaviorGPT：基于下一区块预测的自动驾驶智能体仿真. *arXiv preprint arXiv:2405.17372*, 2024. 2