arXiv:2602.01916v1 [cs.RO] 2 Feb 2026

ForSim：用于交通策略微调的逐步前向仿真

Keyu Chen[1], Wenchao Sun[1],Hao Cheng[1],Zheng Fu[1],Sifa Zheng[1]

摘要—作为自动驾驶闭环训练与评估的基础，交通仿真仍面临两项根本挑战：开放环模仿学习引入的协变量偏移，以及难以反映现实交通中观察到的多模态行为。尽管 RIFT 等近期框架通过组相对优化在一定程度上解决了这些问题，但其前向仿真过程仍基本缺乏响应性，导致虚拟域中的智能体交互不切实际，最终限制了仿真保真度。为解决这些问题，我们提出 *ForSim*，一种逐步闭环前向仿真范式。在每个虚拟时间步，交通智能体通过基于物理原理的运动动力学，推进在时空上与参考轨迹最匹配的虚拟候选轨迹，从而在确保模态内一致性的同时保留多模态行为多样性。其他智能体则根据逐步预测进行更新，从而实现连贯且能感知交互的演化。将*ForSim* 集成到 RIFT 交通仿真框架后，它与组相对优化协同运作，以微调交通策略。大量实验表明，这种集成在保持效率、真实感和舒适性的同时，持续提升安全性。这些结果凸显了在前向仿真中对闭环多模态交互进行建模的重要性，并提升了自动驾驶交通仿真的保真度与可靠性。项目页面：
https://currychen77.github.io/ForSim/

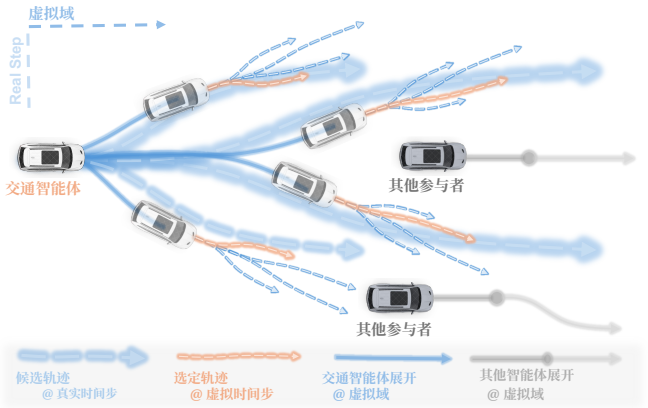

图 1. *ForSim* 在闭环动力学下引入多模态候选轨迹的逐步展开。在每个虚拟时间步，交通智能体选择时空对齐轨迹以保持模态一致性，同时通过 PID 控制器和运动学自行车模型进行传播，确保物理合理性。其他智能体遵循逐步预测，从而实现交互且连贯的演化。

I. 引言

交通仿真构成了现代自动驾驶的基础，为闭环训练与评估提供了依据 [1]。它要求以能够反映闭环环境中真实世界驾驶行为的方式，对多智能体行为进行随时间推移的仿真。这带来了两项核心挑战：协变量偏移 和多模态性。

协变量偏移是开环模仿学习（IL）中长期存在的问题。基于离线专家示范训练的模型在部署到闭环场景时，往往会因累积误差和分布漂移而难以应对。经典的解决方案是在线学习 [2]，该方法会展开策略并查询专家，以生成新的示范数据 [3]。然而，在多智能体交通仿真中，频繁查询专家的成本高得令人难以承受，也难以扩展。为应对这一挑战，近期研究在闭环范式下将模仿学习（IL）与强化学习（RL）相结合 [4]–[6]。这种混合方法利用了两种范式的互补优势：模仿学习通过提炼专家演示实现样本高效初始化，而强化学习则使策略能够通过闭环执行中的交互驱动反馈进行调整。通过综合利用这些优势，此类方法将策略更新建立在观测、动作与奖励之间的因果联系之上，从而减少协变量偏移，同时借助专家数据加快学习逼真的驾驶行为。与模仿学习—强化学习混合方法不同，CAT-K [7] 采用纯监督策略进行闭环微调，将每次推演与前k条专家轨迹中最接近的一条对齐。这种对齐确保推演与真实标注（GT）示范保持一致，从而在整个闭环过程中维持有效的监督信号。然而，尽管有所改进，此类方法通常只利用缓冲区中最可能的推演来优化策略，而该推演在整个展开时域内仅对应一种行为模态。因此，它们难以捕捉现实交通中的多模态驾驶行为。要解决这一局限，不仅需要覆盖多种驾驶模态的数据集，还需要一种能够在闭环执行期间保持探索、避免模式坍塌的策略。这对数据多样性和策略的多模态建模能力都提出了很高要求。

为发挥模仿学习和强化学习各自的优势，RIFT [8] 首先在真实世界数据集上进行开环模仿学习预训练，以捕捉逼真的驾驶行为；随后通过组相对优化对策略进行微调，以缓解协变量偏移，同时保留多模态性。然而，RIFT采用非反应式前向仿真：只有第一步是闭环的，之后所有候选轨迹均以开环方式展开。这会导致虚拟时间域中的智能体交互不切实际，限制前向仿真的

*本研究得到国家自然科学基金委员会资助，项目编号为 52572497 和 52221005。[1]Keyu Chen、Wenchao Sun、Hao Cheng、Zheng Fu 和 Sifa Zheng 均就职于中国北京清华大学车辆与运载学院。

保真度。

基于上述观察，我们提出*ForSim*，这是一种逐步闭环前向仿真范式，旨在捕捉虚拟前向仿真中的交互。在实时域中，每个交通智能体通过交通策略生成多个候选轨迹，代表不同的合理模态。*ForSim* 针对每个候选轨迹独立执行前向仿真，并将其视为相应虚拟域中的参考轨迹。在每个虚拟时间步，智能体传播与该参考轨迹在时空上对齐程度最高的轨迹，确保整个推演过程中的模态内一致性。轨迹传播遵循基于物理的动力学，从而确保物理合理性和动态连贯性。其他智能体则沿着预测轨迹传播，这些轨迹在每个虚拟时间步都会更新，以确保能够响应不断变化的交互。这一迭代过程确保所有候选模态在闭环条件下连贯演化，同时保留多模态行为多样性、模态内一致性和物理合理性。随后，在RIFT框架中逐步评估生成的多模态推演，并通过组相对目标实现对交通策略的精细优化。我们的贡献总结如下：

- 我们提出*ForSim*，这是一种逐步闭环前向仿真范式，能够确保行为多模态性、模态内一致性和物理合理性，从而支持在 RIFT 框架内对交通策略进行细粒度优化。
- 我们证明，*ForSim*通过对闭环多模态交互进行建模，优于现有的前向仿真范式，同时进一步提升了交通仿真的真实感和可靠性。

II. 相关工作

交通仿真。 近期，交通仿真领域的进展探索了多种生成架构，包括变分自编码器 [9], [10], 扩散模型 [11], [12], 和下一令牌预测 [13], [14], 以提升真实感 [15]和长时域鲁棒性 [16], [17]。与此同时，研究者通过基于代价的条件化 [18], 语言提示 [19], [20], 引导采样 [21], 和基于检索的生成 [22], 来实现用户对齐的交通场景

生成 [23], [24]。然而，大多数方法都受到协变量偏移问题的困扰，该问题源于开环训练与闭环部署之间的不匹配，进而进一步削弱了长期稳定性。闭环微调策略提供了部分解决方案，但各自都存在权衡。混合模仿学习/强化学习方法 [4]–[6]提升了鲁棒性，但由于奖励设计方面的挑战，通常会牺牲真实感。CAT-K [7] 等监督式方法取得了出色的性能，但依赖稀缺的专家演示。基于人类反馈的强化学习 [25] 改善了人类对齐，但需要成本高昂的反馈，并且受到奖励模型不稳定性的影响。RIFT [8]通过组相对优化缓解了协变量偏移，同时保留了多模态性，但其对非反应式

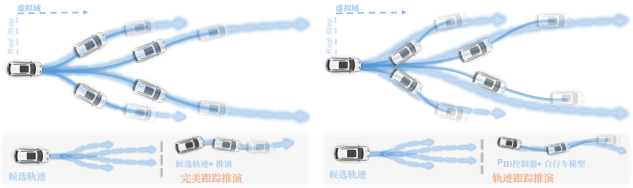

图2. **典型推演范式**。左图展示了完美跟踪，即车辆严格沿规划轨迹行驶。右图展示了轨迹跟踪，即使用控制器和运动学自行车模型，在动态约束下跟踪轨迹。

前向仿真的依赖限制了虚拟域内闭环智能体交互的真实感。

前向仿真。在自动驾驶中，前向仿真是指在给定轨迹的条件下，虚拟展开中心车辆及其他智能体未来的演变过程，用于评估规划轨迹，并衔接规划与执行。如图2所示，中心车辆的典型推演范式包括完美跟踪，即车辆严格沿规划轨迹行驶；以及轨迹跟踪，即使用控制器和运动学自行车模型 [26]在动态约束下跟踪轨迹。实践中，轨迹跟踪是主流范式，现有方法的主要差异在于对其他智能体响应的建模方式。PDM-Closed [27]通过使用轨迹跟踪逐一展开候选轨迹，并假设其他智能体保持恒定速度进行预测，从而从中选择候选轨迹 [28]。PlanTF [29] 和 Pluto [30]将前向仿真融入训练或后处理，以更好地对齐规划轨迹与执行展开，但假设其他智能体遵循日志回放。NAVSIM [31]通过结合轨迹展开与日志回放 [32]或基于规则的更新 [31] ，在 nuPlan 数据集上实现了伪闭环评估 [33]。尽管取得了这些进展，大多数方法本质上仍是开环的：它们展开固定轨迹而不进行重新规划，并依据静态预测推演其他智能体，忽略了动态的智能体间交互。这与现实世界中的规划—执行—再规划循环存在差异，限制了前向仿真的保真度和可靠性。

III. 方法

如第 I节所述，本研究旨在缓解交通仿真中的协变量偏移，同时保留多模态多样性和前向仿真的可靠性。RIFT 是实现这一目标的合适基础 [8], ，它将能够捕捉真实行为并具备路线级可控性的模仿学习预训练，与在高保真模拟器中进行的强化学习微调相结合，以减少协变量偏移并增强风格可控性。

尽管 RIFT 具备这些优势，但它采用非反应式前向仿真范式，相当于 PDM-Closed [27], ，其中只有第一步处于闭环状态，而后续步骤仍处于开环状态。这忽略了智能体间交互，限制了长期真实感。我们通过改进前向仿真组件，支持对候选轨迹进行闭环、交互感知展开，从而确保更忠实地评估多模态候选

轨迹，并最终提升交通仿真的真实感、可靠性和可控性。

A. 预备知识

遵循 RIFT [8], , 我们在 Carla 中构建了以自动驾驶车辆为中心的闭环仿真 [34]。关键背景车辆（CBV）——即最有可能与自动驾驶车辆交互的车辆——由交通策略控制；在下文中，我们将每辆关键背景车辆统称为交通智能体。对于每个交通智能体，策略的输入包括其当前特征F_{cbv}、邻近车辆的历史特征F_{neighbor}以及矢量化地图特征F_{map}。这些特征被编码为嵌入$E_{\text{cbv}} \in \mathbb{R}^{1\times D}$、$E_{\text{neighbor}} \in \mathbb{R}^{N_{\text{neighbor}}\times D}$和$E_{\text{map}} \in \mathbb{R}^{N_{\text{map}}\times D}$，其中$N_{\text{neighbor}}$和$N_{\text{map}}$分别表示邻近车辆和地图元素的数量，而$D$表示嵌入维度。随后，这些嵌入经由 Transformer 编码器模块聚合，形成以关键背景车辆为中心的场景嵌入E_{enc}。

为捕捉多模态驾驶行为，策略进一步采用纵向–横向解耦机制 [30]。具体而言，高层横向查询$Q_{\text{lat}} \in \mathbb{R}^{N_{\text{ref}}\times D}$由参考线信息构建，而可学习的纵向查询$Q_{\text{lon}} \in \mathbb{R}^{N_{\text{lon}}\times D}$则表示多种纵向锚点。二者结合形成多模态导航查询$Q_{\text{nav}} \in \mathbb{R}^{N_{\text{ref}}\times N_{\text{lon}}\times D}$，并通过 Transformer 解码器模块与场景嵌入E_{enc}交互。解码器的最终输出Q_{dec}经由多层感知机头处理，生成一组候选轨迹$\mathcal{T} \in \mathbb{R}^{N_{\text{ref}}\times N_{\text{lon}}\times T\times 6}$及其置信度分数$\mathcal{S} \in \mathbb{R}^{N_{\text{ref}}\times N_{\text{lon}}}$，其中每个轨迹点$\tau_i^t$编码$[p_x, p_y, \cos\theta, \sin\theta, v_x, v_y]$。在这些候选轨迹中，置信度分数最高的轨迹由 PID 控制器执行，确保关键背景车辆与环境保持闭环交互。

与此同时，所有候选轨迹均通过前向仿真进行展开~，以生成展开轨迹$\mathcal{T} = \{\tilde{\tau}_i\}_{i=1}^{N_{\text{ref}}\times N_{\text{lon}}}$，为模态评估奠定基础，并随后通过组相对优化微调交通策略。由于我们的工作聚焦于前向仿真，核心挑战在于以忠实反映真实世界交互与转移的方式展开所有候选轨迹，从而为展开轨迹评估和策略优化提供可靠基础。

为应对这一挑战，我们将前向仿真分解为两个组成部分：（i）交通智能体的逐步虚拟展开，以及（ii）其他智能体的逐步预测传播；二者共同定义了在闭环交互下如何展开候选轨迹。

B. 交通智能体逐步虚拟展开

与完美跟踪直接复用规划轨迹，以及轨迹跟踪在不作调整的情况下严格跟踪固定轨迹不同——这两种方法均如图 2 所示——我们采用逐步规划—执行—重新规划流程，动态适应实时状态变化。该框架更真实地刻画了真实世界驾驶动力学，在这种环境中，智能体会随着状态和交互不断变化而持续修订计划。具体而言，在初始真实世界时间步 ($t = 0$)，我们生成 $N_{\text{ref}} \times N_{\text{lon}}$ 条候选轨迹，并对每条轨迹进行一步前向模拟，从而在虚拟时间 t=1 得到 $N_{\text{ref}} \times N_{\text{lon}}$~个推演分支。随后，在每个虚拟时间步 $\tilde{t}$，每个分支都会根据指定范式选取一个候选索引 $(i_{\tilde{t}}, j_{\tilde{t}})$，并通过使用 PID 控制器与运动学自行车模型的组合来跟踪选定轨迹，从而推进其状态：

$$x_{\tilde{t}+1} = \mathcal{F}_{\text{PID+Bike}}\big(x_{\tilde{t}},\ \tau_{(i_{\tilde{t}},j_{\tilde{t}})}(x_{\tilde{t}})\big), \quad \tilde{t} = 1, \ldots, T, \quad (1)$$

~其中 $x_{\tilde{t}}$ 表示 t 处的虚拟状态，且 $\mathcal{F}_{\text{PID+Bike}}$表示通过 PID 跟踪执行单步动作，随后进行运动学自行车模型传播。为具体实现这一框架，我们探索了三种逐步展开范式，如图 3所示：分别是最大似然展开、模式一致展开和轨迹对齐展开。

~最大似然展开。在每个虚拟时间步 t，每个分支都会选择在当前虚拟状态下预测的置信度最高的候选项：

$$(i_{\tilde{t}}, j_{\tilde{t}}) = \arg\max_{i,j}\ s_{(i,j)}(x_{\tilde{t}}). \quad (2)$$

该策略确保每个分支在整个推演过程中始终遵循最可能的模式。然而，该策略不可避免地会使各分支向主导模式收敛。由于策略对微小的状态扰动具有鲁棒性，由多样化候选项引入的初始偏差通常会被忽略，最终使预期的多模态行为坍缩。

模式一致展开。另一种策略要求每个分支从初始时间步在 $t = 0$ 到 $t = 1$ 中选定候选模式后，始终保持在同一候选模式中。具体而言，如果某个分支以 $\tilde{t} = 1$ 处的候选索引 $(i^{\text{ref}},\ j^{\text{ref}})$ 作为种子，那么在所有后续虚拟时间步中，它都会保持 $(i_{\tilde{t}},\ j_{\tilde{t}}) = (i^{\text{ref}},\ j^{\text{ref}})$，并通过 $\mathcal{F}_{\text{PID+Bike}}$ 相应地进行传播。尽管这种方法保证了明确的模式一致性，但也引入了错位：对应于不同虚拟状态下同一候选模式的轨迹并不完全连贯，由此产生的偏差累积可能导致非预期的模式转移，最终损害推演稳定性。

轨迹对齐推演。为了解决先前策略的局限性，我们提出了轨迹对齐~范式。每个分支仍在虚拟时间 t=1处初始化，并使用特定的候选索引（i^{ref}，j^{ref}），该索引对应于从 t=0~到 t=1的初始转移过程中选定的轨迹。该参考轨迹在整个~推演过程中保持固定。在后续的每个虚拟时间步 $t \geq 2$，分支都会动态选择一个候选项：该候选项的轨迹经过时间对齐后，与固定参考轨迹的距离最近，距离通过平均位移误差（ADE）衡量：

$$(i_{\tilde{t}}, j_{\tilde{t}}) = \arg\min_{(i,j)}\ \text{ADE}_{\text{align}}\Big(\tau_{(i,j)}(x_{\tilde{t}}),\ \tau_{(i^{\text{ref}},j^{\text{ref}})}\Big), \quad (3)$$

其中，$\text{ADE}^{\text{align}}$ 表示将参考轨迹时间对齐至虚拟评估窗口后得到的平均位移误差。

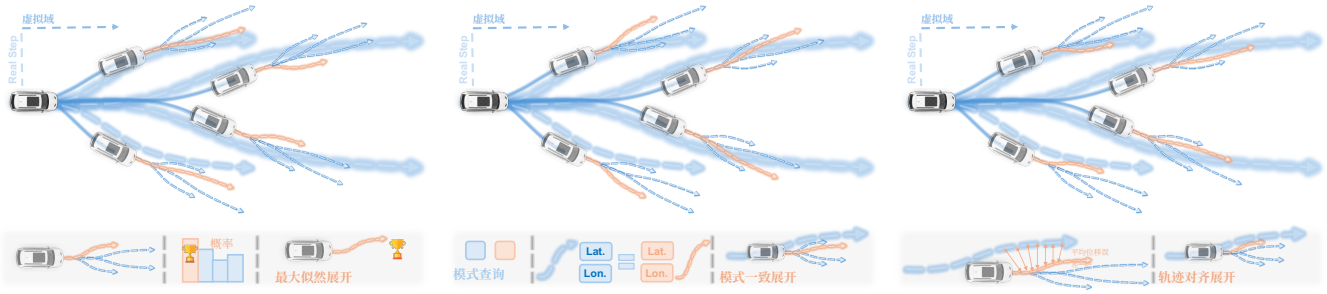

图3. **交通智能体的三种逐步展开范式示意图：**最大似然展开、模式一致展开和轨迹对齐展开。最大似然展开会迅速丧失多模态多样性，而模式一致展开则会因虚拟状态间错位不断累积而受到影响。相比之下，轨迹对齐展开会在每一步选择与其初始模式最接近的候选轨迹（以平均位移误差（ADE）衡量），从而保持多模态保真度，并生成物理一致的展开。

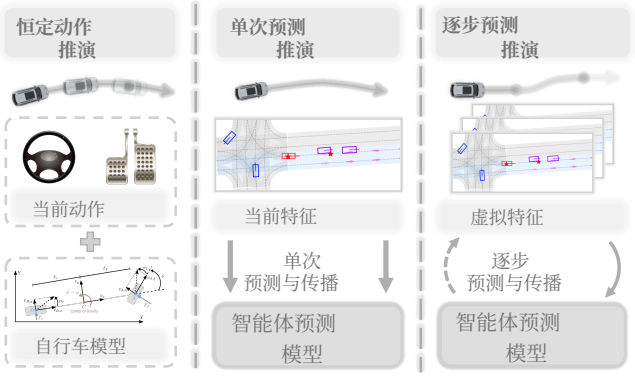

图4. **其他智能体展开范式：**恒定动作展开以开环方式传播当前动作；单次预测展开能够生成更合理的轨迹，但仍是开环方式；逐步预测展开则可实现闭环、交互感知的推演。

这种形式既保证了对原始模态的保真度，又避免了向最大似然展开发生模式坍塌，以及模式一致展开中固有的时间漂移。结合通过 $\mathcal{F}_{\text{PID+Bike}}$ 进行的状态传播，轨迹对齐展开能够保持多模态一致性，并确保展开轨迹在物理上合理、在动态上连贯，因此是最有效的范式。

C. 其他智能体逐步预测展开 t

除了展开交通智能体外，准确模拟其他智能体对于实现真实交互也至关重要。我们探索了三种范式，如图4所示。

恒定动作展开。最常用的基线假设每个其他智能体在整个展开时域内都保持当前动作不变。利用运动学自行车模型，将当前时间步记录的控制输入向前传播。虽然这种方法确保了物理可行性，但它完全采用开环方式，无法捕捉交互引起的行为变化，往往会导致不真实的智能体响应。

单次预测展开。 一种更先进的替代方案利用预测器，根据当前状态预测智能体轨迹。随后，该预测轨迹直接沿时间向前传播。与恒定动作展开相比，该方法能给出更合理的未来状态估计。由于预测仅在初始化时生成，仿真保持开环，因而无法捕捉交互应对不断变化的交通动态时的行为。

逐步预测展开。 为克服这些局限，我们采用逐步预测范式：其他智能体在每个虚拟时间步根据不断演化的虚拟状态更新其预测轨迹。每次更新后的预测再向前传播一步。该设计使其他智能体的规划频率与交通智能体保持同步，从而实现响应式行为以及感知交互的更新。因此，它有助于形成闭环动力学，并显著提升长时程内智能体间交互的真实感。

这一范式的核心是一个智能体预测模型，它为所有其他智能体生成时域为T 的轨迹，记为 $\hat{\mathbf{y}} \in \mathbb{R}^{N\times T\times 6}$，其中每个状态都包含智能体的位置、朝向和速度。真实值监督依据前 T_f个时间步的未来观测构建，形成目标轨迹$\mathbf{y}^\star$。为了兼顾与真实值数据的保真度和物理真实性，我们将多个互补的损失项整合为一个复合目标函数：

$$\mathcal{L}_{\text{pred}} = w_{\text{anchor}}\mathcal{L}_{\text{anchor}} + w_{\text{kin}}\mathcal{L}_{\text{kin}} + w_{\text{smooth}}\mathcal{L}_{\text{smooth}}, \quad (4)$$

其中各损失项定义如下

$$\mathcal{L}_{\text{anchor}} = \text{SmoothL1}\left(\hat{\mathbf{y}}_{0:T_f-1}, \mathbf{y}^\star\right), \quad (5)$$

$$\mathcal{L}_{\text{kin}} = \text{SmoothL1}\left(\mathbf{v}_t, \frac{\mathbf{p}_{t+1}-\mathbf{p}_{t-1}}{2\Delta t}\right), \quad (6)$$

$$\mathcal{L}_{\text{smooth}} = \|\Delta\mathbf{v}_t\|_1 + \|\Delta^2\mathbf{v}_t\|_1, \quad (7)$$

其中 $\mathbf{p}_t = (x_t,\ y_t)$、$\mathbf{v}_t = (v_t^x,\ v_t^y)$、$\Delta\mathbf{v}_t = \mathbf{v}^{t+1} - \mathbf{v}_t$，以及$\Delta^2\mathbf{v}_t = \Delta\mathbf{v}_{t+1} - \Delta\mathbf{v}_t$。该目标兼顾短期准确性、运动学一致性和时间平滑性，从而使预测结果在物理上合理且稳定，适用于长时域闭环仿真。我们设定 $w_{\text{anchor}} = 1.0$、$w_{\text{kin}} = 0.5$和 $w_{\text{smooth}} = 0.5$。

D. 组相对策略微调

基于前向仿真的结果，我们利用 RIFT [8]中提出的组相对优化原则，以闭环方式微调交通策略。对于每个交通智能体的真实世界状态 s, $G = N_{\text{ref}} \times N_{\text{lon}}$，对候选轨迹进行前向仿真以生成展开分支，随后使用逐步奖励模型对其进行评估，从而获得展开级回报 $\mathcal{R} = \{r_i\}_{i=1}^{G}$。形式化地，给定一组展开轨迹 $\widetilde{\mathcal{T}} = \{\tilde{\tau}_i\}_{i=1}^{G}$，

算法 1闭环交通策略优化

输入：经过模仿学习预训练的交通策略 $\pi_{\theta_{\text{init}}}$、经过模仿学习预训练的预测器 $\mathcal{M}_{\text{pred}}$、缓冲区 $\mathcal{D}$
策略 $\pi_\theta \leftarrow \pi_{\theta_{\text{init}}}$
对于迭代 $= 1, \ldots\ldots, I$ **执行** ▷ 强化学习微调
  更新旧策略 $\pi_{\theta_{old}} \leftarrow \pi_\theta$
  当 $\mathcal{D}$ 未完成 **时** ▷ 收集推演数据
    对于 步骤 $= 1, \ldots, T$ **执行**
      生成 $\{\tau_i\}_{i=1}^{G}$ 自 $\pi_{\theta_{old}}$ ▷ 策略推断
      通过算法推导 $\{\tilde{\tau}_i\}_{i=1}^{G}$ 。2 ▷ 前向仿真
      使用公式（$\{r_i\}_{i=1}^{G}$，$\{\hat{A}_i\}_{i=1}^{G}$ 针对每个 $\tilde{\tau}_i$ 式（8）
      将转移存入缓冲区 $\mathcal{D}$
    结束循环
  结束 while 循环
  执行 *ForSim* 迭代 $= 1, \ldots, \mu$ **执行** ▷ 微调
    从缓冲区 $\mathcal{D}$ 中采样小批量转移样本
    更新预测器 $\mathcal{M}_{\text{pred}}$ 依据公式（4） ▷ 训练预测器
    更新交通策略 π_θ 依据公式（9） ▷ 训练策略
  结束循环
结束循环
输出：强化学习微调的交通策略

算法 2*ForSim*：逐步前向仿真

输入：参考轨迹 $\{\tau_i\}_{i=1}^{G}$、交通策略 $\pi_{\theta_{\text{init}}}$、预测器 $\mathcal{M}_{\text{pred}}$
对于 $\tau_i \in \{\tau_i\}_{i=1}^{G}$ **执行**：通过 τ_i 经由 $\mathcal{F}_{\text{PID+Bike}}$ 传播，得到 x_1^{center}：通过 $\mathcal{M}_{\text{pred}}$ 预测其他智能体，得到 x_1^{agents}：记录初始展开状态 $x_1 = (x_1^{\text{center}}, x_1^{\text{agents}})$：**对于** 虚拟时间步 $\tilde{t} = 1, \ldots\ldots, T$ **执行** ▷ 逐步展开：从 $x_{\tilde{t}}$ 生成轨迹，通过公式（3）进行选择，然后通过公式（1）传播，得到 $x_{\tilde{t}+1}^{\text{center}}$ ▷ 交通智能体：通过 $\mathcal{M}_{\text{pred}}(x_{\tilde{t}})$ 预测 $x_{\tilde{t}+1}^{\text{agents}}$ ▷ 其他智能体：记录下一步展开状态 $x_{\tilde{t}+1} = (x_{\tilde{t}+1}^{\text{center}}, x_{\tilde{t}+1}^{\text{agents}})$
结束循环：**结束循环**：**输出**：展开轨迹 $\{\tilde{\tau}_i\}_{i=1}^{G}$

我们定义：

$$r_i = \sum_{t=0}^{T} \gamma^t \left[\text{RM}(\tilde{\tau}_i^t)\right], \hat{A}_i = \frac{r_i - \text{mean}(\mathcal{R})}{\text{std}(\mathcal{R})}, \forall i = 1, \ldots, G \quad (8)$$

其中，$\hat{A}_i$ 表示其归一化的组相对优势。该归一化方式能够促进高回报行为，同时保留候选集固有的多模态性。

为更新策略，我们采用 RIFT 的组相对优化框架，该框架运用双重裁剪机制来限制策略更新、稳定训练，并保持与用户偏好风格的一致性。由此得到的优化目标为：

$$\mathcal{J}(\theta) = \mathbb{E}_{s\sim\mathcal{D}}\left[\frac{1}{G}\sum_{i=1}^{G}\psi(\rho_i(\theta), \hat{A}_i)\right], \quad \rho_i(\theta) = \frac{\pi_\theta(\tau_i|s)}{\pi_{\theta_{\text{old}}}(\tau_i|s)},$$
$$\psi(\rho, \hat{A}) = \begin{cases} \min\left(\rho\hat{A}, \text{clip}(\rho, 1-\epsilon, 1+\epsilon)\hat{A}\right), & \hat{A} \geq 0, \\ \max\Big(\min\left(\rho\hat{A}, \text{clip}(\rho, 1-\epsilon, 1+\epsilon)\hat{A}\right), c\hat{A}\Big), & \hat{A} < 0 \end{cases}, \quad (9)$$

其中，$\rho_i(\theta)$表示更新后的交通策略与旧交通策略之间的轨迹级似然比。

通过这种方式，直接基于前向仿真展开轨迹对交通策略进行微调，从而提升可控性，缓解协变量偏移，并在闭环设置下增强训练稳定性。整体训练流程（算-法1）会反复调用前向仿真模块（算法2），生成用于策略更新的虚拟展开。

IV. 实验

本节通过实证评估*ForSim* 在提升交通仿真的真实感和可控性方面的效果。我们的实验旨在回答以下关键研究问题：**Q1**：与基线方法相比，*ForSim*中的闭环前向仿真是否能提升交通仿真质量？**Q2**：*ForSim* 中的展开范式分别对整体性能有何贡献？**Q3**：所提出的展开范式为何在实践中能有效发挥作用？

我们首先介绍实验设置，包括仿真平台、评估指标和基线方法。随后，我们进行全面的实证比较，展示 *ForSim* 在各种定量和定性标准下的优越性。最后，我们分析*ForSim*中推演范式的设计，探讨其优于现有方法的原因，以及它们如何提升闭环仿真保真度。

A. 实验设置

a) 仿真环境：所有实验均在 CARLA 仿真器中进行 [34],，并遵循 RIFT 框架 [8], 。该框架通过为可能在自动驾驶车辆（AV）沿指定路线行驶时与其交互的关键背景车辆（CBV）分配学习得到的交通策略，实现以自动驾驶车辆为中心的交通仿真。RIFT 采用两阶段训练流程：通过开环模仿学习预训练建立轨迹级真实性和路线级可控性，随后通过强化学习微调处理风格级多样性和协变量偏移缓解。

为隔离推演范式的影响，整体训练流程（包括奖励设计、学习协议和评估流程）均与 RIFT 保持一致。唯一的改动在于前向仿真组件，该组件由我们提出的逐步闭环范式取代。这种受控设计使我们能够系统地单独考察推演范式如何影响交通流的真实感和可控性。

b) 基线方法：为确保比较公平，我们采用与 RIFT 相同的基线方法，包括将 RIFT 本身作为强有力的参照。所有方法均在评分头上进行微调，以隔离前向仿真的影响。对于基于预测的推演，还会对预测头进行微调，以确保推演与策略优化保持一致。

纯强化学习/模仿学习：仅通过强化学习或模仿学习训练的模型，包括 *Pluto* [30]和 *FREA* [24]。

强化学习微调/监督微调：基于预训练*Pluto* 模型的微调方法，包括 *RIFT-Pluto* 、 *PPO-Pluto* 、 *REINFORCE-Pluto* 、 *GRPO-Pluto* 和 *SFT-Pluto*。

混合方法：结合强化学习与监督微调的方法，包括 *RTR-Pluto* 和 *RS-Pluto*。

所有基于强化学习的方法都使用常规风格奖励进行训练，并遵循 RIFT 中定义的标准配置。

表 I**可控性与真实感对比。** 指标在以下条件下进行评估：PDM-Lite [35] 自动驾驶 设置，使用三个随机种子，其中 **最佳** 和 次佳- 结果 分别以高亮标示。

方法	Type	运动学			交互				Map	舒适性
		S-SW ↑	S-WD ↓	A-SW ↑	CPK ↓	RP ↑	二维碰撞时间 ↑	ACT ↑	ORR ↓	UC ↓
Pluto [30]	IL	0.88 ± 0.01	5.81 ± 0.06	0.90 ± 0.01	5.06 ± 2.69	564.14 ± 114.41	2.50 ± 1.48	2.44 ± 1.39	0.24 ± 0.15	**56.45** ± 4.14
FREA [24]	RL	0.93 ± 0.01	5.10 ± 0.14	0.93 ± 0.01	30.42 ± 5.28	292.81 ± 68.54	2.71 ± 1.40	2.67 ± 1.41	9.01 ± 2.09	72.40 ± 1.72
SFT-Pluto	SFT	0.88 ± 0.02	6.01 ± 0.19	0.87 ± 0.02	6.33 ± 2.23	780.48 ± 41.05	2.20 ± 1.64	2.12 ± 1.51	**0.06** ± 0.07	68.14 ± 4.91
RS-Pluto [6]	监督微调+强化学习微调	0.93 ± 0.00	5.40 ± 0.15	0.92 ± 0.01	**4.11** ± 3.90	819。40 ± 74.07	2.27 ± 1.45	2.23 ± 1.43	1.05 ± 0.31	70.31 ± 4.07
RTR-Pluto [5]	监督微调+强化学习微调	0.85 ± 0.00	6.24 ± 0.16	0.81 ± 0.03	6.98 ± 2.59	481.60 ± 70.19	2.55 ± 1.60	2.47 ± 1.51	0.08 ± 0.09	**55.58** ± 4.76
PPO-Pluto	RLFT	0.95 ± 0.01	4.96 ± 0.31	0.90 ± 0.02	6.89 ± 3.19	683.57 ± 38.12	2.66 ± 1.50	2.60 ± 1.43	**0.07** ± 0.13	58.29 ± 2.70
REINFORCE-Pluto	RLFT	0.92 ± 0.01	5.63 ± 0.19	0.90 ± 0.02	6.98 ± 0.86	813.70 ± 24.76	2.39 ± 1.64	2.30 ± 1.55	1.37 ± 1.13	68.10 ± 1.22
GRPO-Pluto [36]	RLFT	0.94 ± 0.04	4.96 ± 0.89	**0.96** ± 0.00	7.24 ± 4.04	892.65 ± 65.27	2.65 ± 1.44	2.61 ± 1.48	0.10 ± 0.08	78.58 ± 0.59
RIFT-Pluto [8]	RLFT	**0.97** ± 0.01	**4.46** ± 0.43	0.93 ± 0.01	6.83 ± 2.62	**995。33** ± 84.62	**2.74** ± 1.30	**2.71** ± 1.32	0.36 ± 0.20	76.90 ± 2.82
ForSim-Pluto（我们的方法）	RLFT	**0.96** ± 0.01			**3.95** ± 1.57	**1005.15** ± 32.28				

表 II**消融中心滚动范式的消融实验。** 带 * 的条目对周围智能体采用恒定动作。

范式	效率指标		违规指标	
	BR ↓	RP ↑	ORR ↓	CPK ↓
完美跟踪 *	**0.00** ± 0.00	993.63 ± 48.01	**0.26** ± 0.22	**4.00** ± 0.83
轨迹跟踪*	3.33 ± 5.77	**1087.56** ± 36.41	0.67 ± 0.57	4.00 ± 1.47
最大似然	**0.00** ± 0.00	910.97 ± 9.51	**0.11** ± 0.05	4.76 ± 3.53
模式一致	3.33 ± 5.77	940.94 ± 73.70	0.28 ± 0.29	4.55 ± 1.34
轨迹对齐	**0.00** ± 0.00			

表 III**其他推演范式的消融。**

范式	效率指标		违规指标	
	BR ↓	RP ↑	ORR ↓	CPK ↓
恒定动作	**0.00** ± 0.00	**1128.77** ± 96.24	**0.91** ± 0.30	4.97 ± 1.57
单次预测	**0.00** ± 0.00	**1170.28** ± 29.17	2.31 ± 0.29	**3.41** ± 1.44
逐步预测	**0.00** ± 0.0			

B. 交通仿真质量（*Q1*）

a) 指标：按照 WOSAC 评估框架 [37], ，我们使用四类指标评估仿真质量：运动学、交互、地图和舒适性。

运动学指标用于评估智能体运动的分布真实性，包括对速度（S-SW）和加速度（A-SW）进行 Shapiro–Wilk检验，以及对速度进行 Wasserstein距离计算（S-WD）。

交互指标包括每公里碰撞次数（CPK）、路线进度（RP）以及安全关键指标：2D-TTC [38]和 ACT [39]。

地图指标通过驶离道路率（ORR）衡量空间合规性，该指标反映了车辆处于可行驶区域之外的时间。

舒适性指标依据 RIFT [8]，通过不适率（UC）衡量驾驶平顺性；该指标定义为加速度超过舒适阈值的时间占比。

b) 主要结果与分析：*ForSim*在安全相关指标上展现出显著改进，尤其是降低了 CPK、二维碰撞时间和 ACT，如表 I所示。这些结果凸显了对闭环交互进行建模的有效性，使智能体能够更有效地预判并响应动态交通状况。重要的是，这种安全性提升是在不牺牲效率或真实感的情况下实现的：运动学特性与 RP 得以保持，或甚至相对于 RIFT 略有提升。与此同时，ORR 的降低表明行为在空间上更加合规。此外，*ForSim*显著提升了舒适性，这得益于采用 PID 控制器以及可确保物理合理性的运动学自行车模型。

C. 推演范式分析（*Q2* 和 *Q3*）

a）推演范式消融：如表II 和III所示，我们针对交通智能体和其他智能体的推演范式开展了消融研究。对于交通智能体（表II），我们比较了三种闭环变体：最大似然推演、模式一致展开和轨迹对齐推演，同时将其他智能体固定为逐步预测展开。相比之下，完美跟踪和轨迹跟踪在前向仿真期间绕过交通策略，因此不受其他智能体转移的影响。因此，我们对其他智能体保留标准的恒定动作展开，这与先前的工作一致 [27]。

在交通智能体范式中，轨迹对齐展开在安全性与效率之间实现了最有利的权衡，以具有竞争力的 RP 和 ORR 达到了最低的 BR 和 CPK。通过选择与初始参考最匹配的时空对齐轨迹，该策略确保了物理模态一致性。与逐步展开相结合时，它能够捕捉具有交互感知能力的转移，从而更忠实地模拟现实世界中的多智能体交互。

对于其他参与者（表 III），我们将交通智能体固定为轨迹对齐展开，并评估三种推演范式：恒定动作展开、单次预测展开和逐步预测展开。恒定动作展开作为一种开环策略，忽略了不断变化的交通动态，导致响应能力较差且 CPK 较高。单次预测展开提高了短时域精度，但缺乏长时域监督，导致误差累积和频繁越界偏离，表现为 ORR 较高。相比之下，逐步预测展开以闭环方式利用短时域精度，持续根据动态情境调整智能体行为，因此在保持较高 RP 并消除阻塞的同时，实现了有竞争力的 ORR 和 CPK。这些发现凸显了闭环建模在捕捉响应迅速、具有交互性的多智能体行为方面的有效性。

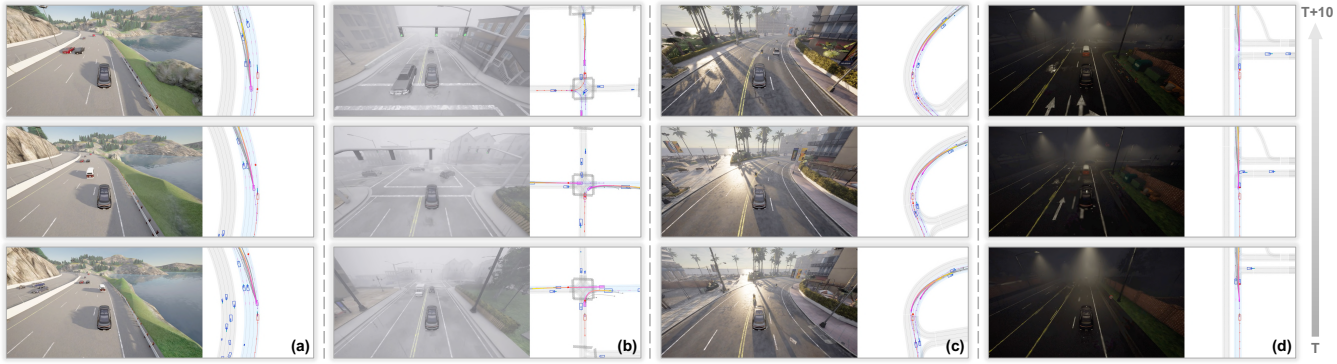

图5。 **代表性场景 o f *ForSim*。** 交通智能体（CBV）标记为 紫色，自 （PDM-Lite [35]）标记为 ，其他智能体标记为 蓝色.

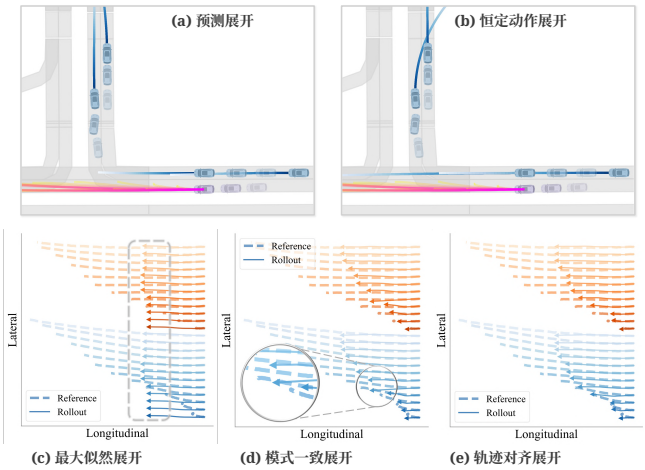

图 6。**展开范式对比。** 对于其他智能体，（a）–（b）：预测展开相比恒定动作展开提升了长时域稳定性。对于交通智能体，（c）–（e）：轨迹对齐展开保持了多模态一致性，而最大似然展开和模式一致展开则受到模式坍塌和时间不一致性的影响。

综合来看，这些结果表明，将轨迹对齐展开与逐步预测展开相结合，是最有效的前向仿真配置，能够在保持效率的同时显著提升安全性。

b）代表性场景分析： 图5 展示了代表性场景，突出了*ForSim* 在所提出的前向仿真范式下的有效性。通过对闭环虚拟交互进行建模，交通策略学会选择具有上下文感知能力且风险敏感的行为。

在图 5（a）中，智能体在低曲率转弯过程中从外侧车道变换至内侧车道，执行高效率的变道策略，展现了其预判并优化未来运动的能力。图 5（b）说明，尽管交叉路口附近存在多个驶离道路的候选轨迹，智能体仍会选择遵循空间约束的参考对齐轨迹。图 5（c）展示了两名骑行者在转弯过程中从偏离中心的位置恢复，体现了对协变量偏移的鲁棒性以及执行纠正行为的能力。图 5（d）展示了因前车减速而发起的超车操作；在此过程中，智能体避免了潜在碰撞，同时保持了驾驶效率。

这些案例共同表明，在前向仿真过程中对闭环交互进行建模，能够使交通策略在各种场景中内化兼顾风险、效率和真实行为的决策方式。

c) 推演范式的定性分析：图6展示了推演范式的定性比较。图 6(a)–(b) 对比了针对其他智能体的两种推演范式：预测展开和恒定动作展开。前者能够生成更稳定、更准确的长时域展开轨迹，因为经训练的预测器可以生成物理上合理的未来状态，且对瞬态控制噪声不那么敏感。相比之下，恒定动作展开会在整个预测时域内仅传播初始动作，导致误差严重累积，真实感下降。

图6(c)–(e) 展示了采用三种范式的交通智能体展开轨迹。如第 III-B 节所述，最大似然展开会反复选择概率最高的轨迹，从而使多样化的意图模式坍缩，限制多样性并降低组内优势区分度。模式一致展开通过固定所选模式来确保显式的模态一致性，但会遭遇时间不一致问题——来自同一模式、但对应不同虚拟状态的轨迹在物理上可能无法对齐，导致模态漂移。相比之下，轨迹对齐展开通过选择时空距离上与参考轨迹最匹配的候选轨迹来确保物理一致性。这种方法既保留了多模态多样性，又增强了物理合理性和模态内一致性。

V. 结论

在这项工作中，我们提出了*ForSim*，一种逐步闭环前向仿真范式，旨在解决现有交通仿真框架中的关键局限，尤其是以往展开过程中缺乏响应能力和交互保真度的问题。*ForSim*在每个虚拟时间步为交通智能体采用轨迹对齐展开，并通过基于物理的动力学推进它们，从而保留多模态多样性、模态内一致性和物理合理性。同时，其他智能体通过逐步预测展开进行更新，使其能够在闭环条件下实现响应式、交互感知的演化。将*ForSim*集成到 RIFT 框架后，可实现交互感知仿真，并通过组相对目标进行细粒度策略优化。实验表明，*ForSim* 在不牺牲效率、真实感或

舒适性的前提下提升了安全性相关指标，验证了在虚拟仿真中对闭环多模态交互进行建模的重要性。这些结果表明，*ForSim*是一种通用且有效的范式，可提升交通仿真的保真度和可靠性。

参考文献

[1] D. Chen, M. Zhu, H. Yang, X. Wang, and Y. Wang, 《数据驱动交通仿真：全面综述》，IEEE Transactions on Intelligent Vehicles, 2024 1 。S. Ross, G. Gordon, and D. Bagnell, 《将模仿学习和结构化预测归约为无遗憾在线学习》，载于第十四届人工智能与统计国际会议论文集。JMLR Workshop and Conference Proceedings，2011年，第627–635页。1[3] Z. Peng, W. Mo, C. Duan, Q. Li, and B. Zhou, 《通过代理价值传播从主动人类参与中学习》，神经信息处理系统进展，2023 1 。Y. Lu, J. Fu, G. Tucker, X. Pan, E. Bronstein, R. Roelofs, B. Sapp, B. White, A. Faust, S. Whiteson 等，《模仿还不够：通过强化学习增强模仿，以应对具有挑战性的驾驶场景》，载于2023年IEEE/RSJ智能机器人与系统国际会议（IROS）。IEEE，2023年，第7553–7560页。1, 2 C. Zhang, J. Tu, L. Zhang, K. Wong, S. Suo, and R. Urtasun, 《在闭环中学习逼真的交通智能体》，载于机器人学习会议。PMLR, 2023, 800 821 1, 2, 6 第 – 页。Z. Peng, W. Luo, Y. Lu, T. Shen, C. Gulino, A. Seff, and J. Fu, 《通过强化学习微调改进自动驾驶智能体行为》，载于欧洲计算机视觉会议。Springer，2024年，第165 181 1, 2, 6 –页。Z. Zhang, P. Karkus, M. Igl, W. Ding, Y. Chen, B. Ivanovic, and M. Pavone, 《令牌化交通模型的闭环监督微调》，载于计算机视觉与模式识别会议论文集, 2025, 5422 5432 1, 2 第 – 页。K. Chen, W. Sun, H. Cheng, and S. Zheng, 《RIFT：面向逼真且可控交通仿真的闭环强化学习微调》，arXiv预印本 arXiv: 2505 03344, 2025 1, 2, 3, 4, 5, 6 。。S. Suo, S. Regalado, S. Casas, and R. Urtasun, 《Trafficsim：学习模拟逼真的多智能体行为》，载于IEEE/CVF计算机视觉与模式识别会议论文集, 2021, 10 400 10 409 2 第 – 页。D. Xu, Y. Chen, B. Ivanovic, and M. Pavone, 《Bits：用于交通仿真的双层模仿》，载于2023年IEEE机器人与自动化国际会议（ICRA）。IEEE, 2023, 2929 2936 2 第 – 页。K. Chitta, D. Dauner, and A. Geiger, 《告别对基于学习的车辆运动规划的误解》，载于欧洲计算机视觉会议。Springer，2024年，第57–74页。2[12] Y. Zhou, N. Ye, W. Ljungbergh, T. Li, J. Yang, Z. Yang, H. Zhu, C. Petersson, and H. Li, 《解耦扩散激发自适应场景生成》，arXiv预印本 arXiv: 2504 10485, 2025 2 。。W. Wu, X. Feng, Z. Gao, and Y. Kan, 《Smart：通过下一令牌预测实现可扩展的多智能体实时运动生成》，神经信息处理系统进展，第37卷，114 071，2024年 114 048 2 第 – 页。[14] J. Philion, X. B. Peng, and S. Fidler, 《Trajeglish：将交通建模为下一令牌预测》，载于第十二届国际学习表征会议, 2023 2 。S. Tan, K. Wong, S. Wang, S. Manivasagam, M. Ren, and R. Urtasun, 《Scenegen：学习生成逼真的交通场景》，载于IEEE/CVF计算机视觉与模式识别会议论文集, 2021, 892 901 2 第 – 页。X. Yang, S. Tan, and P. Krähenbühl, 《通过交错的自回归运动与场景生成进行长期交通仿真》，arXiv预印本 arXiv: 2506 17213, 2025 2 。。Z. Peng, Y. Liu, and B. Zhou, 《Infgen：将场景生成作为下一令牌组预测》，arXiv预印本 arXiv: 2506 23316, 2025 2 。。Z. Zhong, D. Rempe, D. Xu, Y. Chen, S. Veer, T. Che, B. Ray, and M. Pavone, 《用于可控交通仿真的引导式条件扩散》，载于2023年IEEE机器人与自动化国际会议（ICRA）。IEEE, 2023, 3560 3566 2 第 – 页。Z. Zhong, D. Rempe, Y. Chen, B. Ivanovic, Y. Cao, D. Xu, M. Pavone, and B. Ray, 《通过场景级扩散实现语言引导的交通仿真》，arXiv预印本 arXiv:2306.06344，2023年。2

[20] S. Tan, B. Ivanovic, Y. Chen, B. Li, X. Weng, Y. Cao, P. Krähenbühl, and M. Pavone, 《可提示的闭环交通仿真》，载于第八届机器人学习年会, 2024 2 。J. Lu, K. Wong, C. Zhang, S. Suo, and R. Urtasun, 《Scenecontrol：用于可控交通场景生成的扩散模型》，载于2024年IEEE机器人与自动化国际会议（ICRA）。IEEE, 2024, 16 908 16 914 2 第 – 页。W. Ding, Y. Cao, D. Zhao, C. Xiao, and M. Pavone, 《Realgen：用于可控交通场景的检索增强生成》，载于欧洲计算机视觉会议。Springer, 2024, 93 110 2 第 – 页。H. Lin, X. Huang, T. Phan-Minh, D. S. Hayden, H. Zhang, D. Zhao, S. Srinivasa, E. M. Wolff, and H. Chen, 《用于闭环交通生成的因果组合扩散模型》，arXiv预印本 arXiv: 2412 17920, 2024 2 。。K. Chen, Y. Lei, H. Cheng, H. Wu, W. Sun, and S. Zheng, 《FREA：通过可行性引导生成具有合理对抗性的安全关键场景》，载于机器人学习会议。PMLR，2025年，第566 586 2, 5, 6 –页。Y. Cao, B. Ivanovic, C. Xiao, and M. Pavone, 《基于人类反馈的强化学习用于逼真交通仿真》，载于2024年IEEE机器人与自动化国际会议（ICRA）。IEEE, 2024, 14 428 14 434 2 第 – 页。P. Polack, F. Altché, B. d'Andréa Novel, and A. de La Fortelle, 《运动学自行车模型：一种用于自动驾驶车辆可行轨迹规划的一致模型？》，载于2017年IEEE智能车辆研讨会（IV）。IEEE, 2017, 812 818 2 第 – 页。D. Dauner, M. Hallgarten, A. Geiger, and K. Chitta, 《摒弃对基于学习的车辆运动规划的误解》，载于机器人学习会议。PMLR, 2023, 1268 1281 2, 6 第 – 页。C. Schöller, V. Aravantinos, F. Lay, and A. Knoll, 《匀速模型能让我们从行人运动预测中学到什么》，IEEE机器人与自动化快报，第5卷，第2期，第1696–1703页，2020年。2[29]J. Cheng, Y. Chen, X. Mei, B. Yang, B. Li, and M. Liu, 《重新思考用于自动驾驶的基于模仿的规划器》，载于2024年IEEE机器人与自动化国际会议（ICRA）。IEEE, 2024, 14 123 14 130 2 第 – 页。J. Cheng, Y. Chen, and Q. Chen, 《Pluto：突破基于模仿学习的自动驾驶规划上限》，arXiv预印本 arXiv: 2404 14327, 2024 2, 3, 5, 6 。。D. Dauner, M. Hallgarten, T. Li, X. Weng, Z. Huang, Z. Yang, H. Li, I. Gilitschenski, B. Ivanovic, M. Pavone 等，《NAVSIM：数据驱动的非反应式自动驾驶车辆仿真与基准测试》，神经信息处理系统进展，第37卷，第 28 706 28 719, 2024 2 –页。H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari, 《nuplan：面向自动驾驶车辆的闭环机器学习规划基准》，arXiv预印本 arXiv: 2106 11810, 2021 2 。。W. Cao, M. Hallgarten, T. Li, D. Dauner, X. Gu, C. Wang, Y. Miron, M. Aiello, H. Li, I. Gilitschenski 等，《面向自动驾驶的伪仿真》，arXiv预印本 arXiv: 2506 04218, 2025 2 。。A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, 《Carla：一个开放的城市驾驶模拟器》，载于机器人学习会议。PMLR, 2017, 1 16 3, 5 第 – 页。J. Beißwenger, 《PDM-Lite：面向Carla排行榜2.0的基于规则的规划器》，https://github.com/OpenDriveLab/DriveLM/blob/DriveLM-CARLA/docs/report.pdf, 2024, 2025 04 09 6, 7 访问日期：--。Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu 等，《Deepseekmath：突破开放语言模型数学推理的极限》，arXiv预印本 arXiv:2402.03300，2024年 6。。[37] N. Montali, J. Lambert, P. Mougin, A. Kuefler, N. Rhinehart, M. Li, C. Gulino, T. Emrich, Z. Yang, S. Whiteson 等，《Waymo开放仿真智能体挑战赛》，神经信息处理系统进展，第36, 59 151 59 171, 2023 6 卷，第 – 页。H. Guo, K. Xie, and M. Keyvan-Ekbatani, 《安全关键变道过程中驾驶员规避行为建模：三维碰撞时间与深度强化学习》，事故分析与预防，第186, 107063, 2023 6 卷，第p页。。S. P. Venthuruthiyil and M. Chunchu, 《预期碰撞时间（ACT）：一种基于轨迹的主动安全评估二维替代安全指标》，《交通研究C辑：新兴技术》，第139卷，第103655页，2022年。