安全关键交通模拟与引导潜在扩散模型

1st Mingxing Peng
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
mpeng060@connect.hkust-gz.edu.cn

2nd Ruoyu Yao
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
ryao092@connect.hkust-gz.edu.cn

3rd Xusen Guo
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
xguo796@connect.hkust-gz.edu.cn

4th Yuting Xie
School of Computer Science and Engineering
Sun Yat-sen University
Guangzhou, China
xieyt8@mail2.sysu.edu.cn

5th Xianda Chen
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
xchen595@connect.hkust-gz.edu.cn

6th Jun Ma
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
jun.ma@ust.hk

摘要—安全关键的交通模拟在评估自主驾驶系统在罕见和具有挑战性的场景中的表现中发挥着至关重要的作用。然而，现有的方法通常由于对物理合理性的考虑不足而生成不现实的场景，并且存在生成效率低下的问题。为了解决这些问题，我们提出了一种引导式潜在扩散模型（LDM），能够生成物理上真实且对抗的安全关键交通场景。具体而言，我们的模型采用基于图的变分自编码器（VAE）来学习一个紧凑的潜在空间，以捕捉复杂的多智能体交互并提高计算效率。在这个潜在空间中，扩散模型执行去噪过程以产生真实的轨迹。为了实现可控和对抗性的情景生成，我们引入了新颖的引导目标，驱动扩散过程向生成对抗性和行为上真实驾驶行为的方向发展。此外，我们基于物理可行性检查开发了一个样本选择模块，进一步增强了生成场景的物理合理性。在 nuScenes 数据集上的广泛实验表明，我们的方法与现有基线相比，在保持高度现实性的同时实现了更优的对抗效果和生成效率。我们的工作提供了一种有效的工具来进行真实的安全关键情景模拟，为自主驾驶系统的更强大评估铺平了道路。

***Index Terms*—扩散模型，交通仿真，安全关键仿真。**

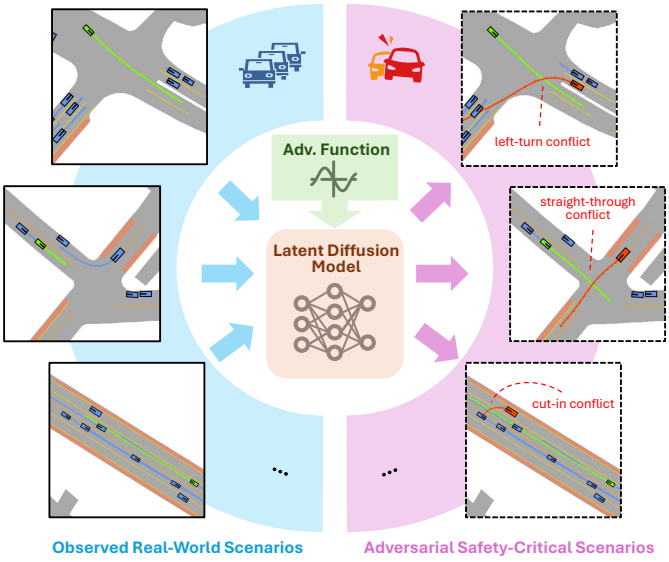

图 1. 我们指导的 LDM 在关键交通安全模拟中的概述。我们的模型通过利用由对抗目标引导的 LDM 生成逼真且具有挑战性的驾驶场景，有效测试 AV 系统。

I. 介绍

随着自动驾驶汽车（AV）技术的快速发展，对自动驾驶汽车的安全性能进行严格且高效的评估已成为关键的研究重点，因为可靠的评估对于加速自主驾驶系统的发展至关重要 [1], [2]。然而，在现实世界的道路测试中，交通情况主要为正常状态，而安全关键事件发生得非常罕见，这使得收集足够数量的安全关键事件数据既耗时又昂贵。因此，安全关键的交通模拟已成为现代自动驾驶汽车验证流程不可或缺的一部分 [3]。

传统方法通常依赖于基于规则的模拟器，如 CARLA [4] 和 SUMO [5]，这些模拟器允许用户通过明确指定代理状态手动设计关键安全场景。然而，这种手动脚本编写过程需要大量的专业知识和广泛的参数调整。即使经过仔细的手动设计，生成的场景往往也难以再现自然交通中多智能体之间的交互以及类似人类的驾驶行为，从而限制了它们在大规模评估中的实用性。为了克服这些局限性，近期的数据驱动方法首先从大型轨迹数据集中学习现实的驾驶行为，然后在测试时进行优化以合成关键安全场景 [6], [7]。例如，Strive [7] 使用图神经网络（GNN）构建交通先验模型来进行可信的交通建模，并随后执行对抗性优化。尽管这类方法实现了现实主义方面的适度改进，但仍难以准确地模拟复杂的轨迹交互，并且缺乏物理可行性的保证 [6], [8]。此外，它们的迭代搜索程序需要大量的计算资源，导致生成效率低下 [6], [7]。

最近，扩散模型在生成逼真的顺序数据方面表现出色，包括模拟现实世界驾驶行为的互动交通场景 [9], [10]。扩散模型中固有的噪声到数据生成范式捕捉到了复杂的多智能体依赖关系，并且在推理过程中提供了灵活的指导，从而促进了可控生成 [11], [12]。此外，最近在潜在扩散模型（LDMs）方面的进展表明，学习紧凑且表达能力强的驾驶轨迹潜在表示有助于建模它们的联合分布，提升了生成交通场景的真实感和多样性 [13], [14]，同时提高了计算效率 [15]。总的来说，这些工作突显了扩散模型在生成现实且可控的安全关键场景方面的强大潜力。

受这些工作的启发，我们提出了一种用于模拟物理上真实且对抗性的安全关键交通场景的引导式 LDM 框架，如图 1 所示。具体来说，我们利用基于 GNN 的编码器将过去和未来的轨迹转换为紧凑的潜在表示，以捕捉复杂的多智能体交互。在过去的场景信息潜在表示条件下，我们的模型逐步去噪一个噪声未来潜在表示来重建真实的驾驶轨迹。为了有效引导生成对抗性的安全关键场景，我们引入了新颖的指导目标，这些目标在推理过程中逐步扰动采样过程。此外，我们设计了一个基于物理可行性检查的样本选择模块，以确保生成的场景符合物理约束。通过使用配备基于规则的规划器的自我车辆进行闭环模拟来评估生成的安全关键场景的有效性和真实性。这个安全关键交通仿真框架有助于揭示自动驾驶系统的限制和脆弱性，为提高 AVs 的安全性和鲁棒性提供了宝贵的见解。

总结，本文的主要贡献包括：

- 我们提出了一种用于关键安全交通模拟的引导式 LDM 框架，在一个紧凑的潜在空间中执行去噪过程，该方法能够有效捕捉复杂的多智能体交互并提高计算效率。
- 我们设计了新颖的引导目标和基于物理可行性检查的样本选择模块，这些共同实现了对抗性和物理上真实的驾驶轨迹的可控生成。
- 在 nuScenes 数据集上进行的大量实验表明，我们的模型在驾驶场景生成方面显著优于基线方法，在对抗有效性、多样性和生成效率方面表现更优，同时保持物理合理和行为真实的轨迹。

II. 相关工作

本节回顾了两个领域的相关工作：安全关键的交通模拟和基于扩散的交通场景生成。

A. 安全关键交通仿真

安全关键的交通模拟对于评估自动驾驶系统在罕见和高风险场景下的表现至关重要 ~ [3]。传统方法通常依赖于像 CARLA ~ [4] 和 SUMO ~ [5] 这样的模拟器，通过修改代理状态来手动设计这些场景。然而，这些方法需要大量的领域专业知识，并且经常产生不现实或不可扩展的设计 ~ [16], [17]。为了提高真实性，最近的研究利用大规模的驾驶数据集学习真实的交通模式，并在测试时通过基于优化的技术生成安全关键场景 ~ [6], [7], [18]。特别地，AdvSim [6] 直接扰动标准轨迹空间以诱导对抗性轨迹，而 Strive [7] 则使用基于图的表示在潜在空间中执行对抗优化 [19]。MixSim [18] 学习路线条件策略以实现反应性和可控重模拟，支持混合现实交通场景中的真实变化和安全关键交互。尽管这些方法在现实感方面取得了适度的改进，但它们仍受效率低

下的限制，并继续难以生成既合理又可控制的驾驶行为，从而限制了其在大规模场景生成方面的实用性。

B. 基于扩散的交通场景生成

扩散模型 [20] 由于能够模拟复杂的交通模式并生成高保真的仿真，这些仿真与现实世界场景非常接近，因此受到了广泛关注。此外，它们提供了增强的可控性，可以基于特定条件或指导 [21] 定制交通场景。例如，DJINN [10] 利用无分类器扩散模型生成了交通场景中所有代理的联合交互轨迹，这些轨迹依赖于一组灵活的代理状态。此外，一些研究通过在推理时间 [11]–[13], [22], [23] 将引导机制纳入扩散模型进一步提高了可控性。具体而言，CTG [11] 采用信号时态逻辑（STL）公式作为引导来生成符合规则的轨迹，而 MotionDiffuser [12] 则提出了一些可微分的成本函数作为指导，使生成的轨迹能够满足物理约束。DiffScene [22] 和 Safe-Sim [23] 提出了基于安全的目标函数来模拟关键驾驶场景。其他研究利用 LDMs 学习更有效的驾驶轨迹表示，并对这些轨迹进行联合分布建模 [13], [14], [24]。然而，这些方法在可控性和灵活性方面仍然存在局限性。例如，AdvDiffuser [13] 无法轻松应用于生成特定对抗车辆的场景，并且需要训练不同的分类器以适应各种对抗策略。

III. 方法论

本节介绍了我们提出的用于生成逼真和对抗性关键交通安全场景的模型。我们提出模型的整体框架如图 2 所示。首先，我们提供了一个正式的关键场景模拟问题定义。然后，描述了用于场景生成的 LDM。最后，引入了新颖的引导目标以及基于物理可行性的样本选择模块，这些被用来指导生成向逼真和对抗性驾驶场景发展。

A. 问题公式化

我们专注于涉及 N 个代理的安全关键交通模拟，其中一个代理是由规划器 π 控制的主车辆，其余 $N-1$ 辆车的未来轨迹由我们的模型生成。在这些车辆中，有一辆是敌对车辆，其目标是与主车辆发生碰撞，而其他车辆则保持现实的轨迹。

一个驾驶场景 S 包含 N 个代理状态和一个地图 $\boldsymbol{m}$。在每个时间步 t，每个代理 $s_t^i = (x_t^i, y_t^i, \theta_t^i, v_t^i)$ 的状态表示二维位置、航向和速度。每个代理的相应动作表示为 $a_t^i = (\dot{v}i^t, \dot{\theta}i^t)$，指示加速度和偏转率。我们将所有代理在过去 T_{hist} 个时间步的轨迹表示为 $\boldsymbol{x} = \{s_{t-T_{hist}}, s_{t-T_{hist}+1}, ..., s_t\}$。规划器 π 确定了自车在未来从 t 到 $t+T$ 时间段内的轨迹，记作 $s_{t:t+T}^0 = \pi(\boldsymbol{m}, \boldsymbol{x})$。

我们提出的 LDM g，由 θ 参数化，模拟未来 $N-1$ 非自我车辆的轨迹，表示为 $\boldsymbol{\tau} = \{s_{t:t+T}^i\}_{i=1}^{N-1}$。该模型包括一个编码器 $\mathcal{E}$，它将历史轨迹数据和地图信息编码成紧凑的潜在表示，并且有一个解码器 $\mathcal{D}$，它将去噪后的潜在表示解码为非自我代理的预测未来轨迹。在训练过程中，模型从现实世界的数据中学习真实的交通行为，在推理过程中，对抗目标函数引导生成安全关键场景。

B. 潜扩散模型在交通模拟中的应用

我们提出了一种 LDM，通过迭代去噪过程生成现实且可控的对抗性安全关键驾驶场景。与直接在轨迹空间 [11], [23], [25] 进行操作的传统扩散方法不同，我们的方法在潜在空间中执行去噪过程，从而减少了计算开销并增强了特征表达力 [7], [13], [15], [24]。我们的模型基于预训练的图卷积变分自编码器 Strive [7]，其中编码器捕获复杂的多智能体交互，解码器使用运动学自行车模型自回归生成未来轨迹以确保生成轨迹的真实性。

架构。如图 2 所示，我们的 LDM 由三个组件组成：两个冻结的基于 GNN 的编码器、一个可学习的 U-Net 去噪网络和一个冻结的基于 GNN 的解码器。先验编码器 $\mathcal{E}_\theta(\boldsymbol{x}, \boldsymbol{m})$ 对过去的代理轨迹和本地地图特征进行编码以生成条件输入 $\boldsymbol{c}$，而后验编码器则额外结合未来轨迹来产生潜变量 $\boldsymbol{z}$。前向过程从干净的潜在变量 $\boldsymbol{z}^0 \sim q(\boldsymbol{z}^0)$ 开始，并通过以下过渡逐步注入高斯噪声：

$$q(\boldsymbol{z}^k \mid \boldsymbol{z}^{k-1}) = \mathcal{N}\left(\boldsymbol{z}^k; \sqrt{1-\beta_k}\boldsymbol{z}^{k-1}, \beta_k \boldsymbol{I}\right) \quad (1)$$

其中 β_k 表示预定义的方差计划，用于控制每一步的噪声水平。对于足够大的步数 K，分布 $\boldsymbol{z}^K$ 收敛到各向同性高斯分布，即 $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$。

为了加速推理，我们采用了去噪扩散隐式模型（DDIM）采样策略 [26]，这使得非马尔可夫逆扩散成为可能，并通过跳过中间步骤支持高效采样而无需重新训练。逆过程定义为：

$$\boldsymbol{z}^{k-1} = \sqrt{\alpha_{k-1}} \cdot \tilde{\boldsymbol{z}}^0 + \sqrt{1-\alpha_{k-1}} \cdot \epsilon_\theta(\boldsymbol{z}^k, k, \boldsymbol{c}) \quad (2)$$

其中 $\epsilon_\theta(\boldsymbol{z}^k, k, \boldsymbol{c})$ 表示条件为 $\boldsymbol{c}$ 的噪声预测模型，而 $\alpha_k = \prod_{i=1}^{k}(1-\beta_i)$ 代表噪声调度的累积乘积。预测的干净潜

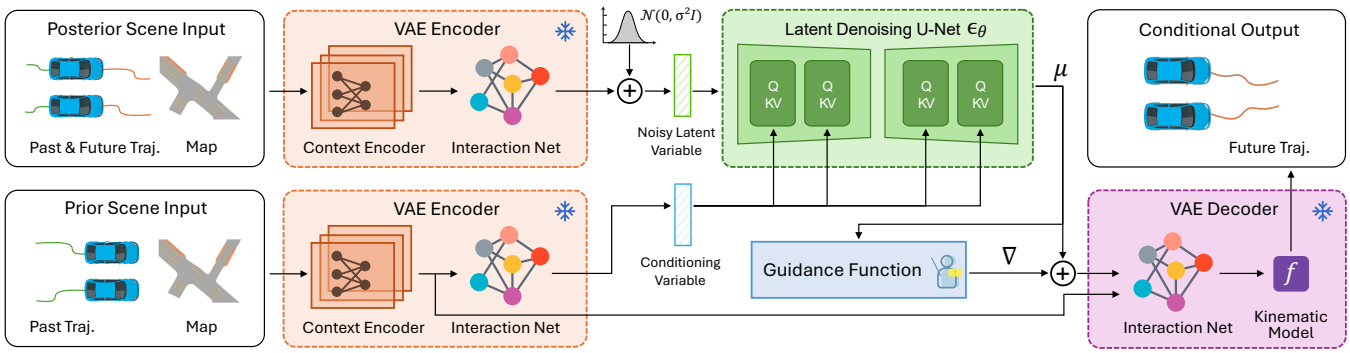

图 2. 我们提出的引导 LDM 的整体框架。基于图的 VAEs 将后验和先验场景输入编码为潜在表示，这些表示通过受先验潜条件约束的 U-Net 进行去噪。在采样阶段，我们引入了我们提出的指导目标来引导扩散过程生成对抗性的安全关键驾驶场景。

变量 $\tilde{\boldsymbol{z}}^0$ 被估计为：

$$\tilde{\boldsymbol{z}}^0 = \left(\frac{\boldsymbol{z}^k - \sqrt{1-\alpha_k} \cdot \epsilon_\theta(\boldsymbol{z}^k, k, \boldsymbol{c})}{\sqrt{\alpha_k}} \right) \tag{3}$$

从 $\boldsymbol{z}^K$ 开始迭代应用此逆过程得到最终去噪潜变量 $\hat{\boldsymbol{z}}^0$。

最后，解码器 $\mathcal{D}_\theta(\hat{\boldsymbol{z}}^0, \boldsymbol{x}, \boldsymbol{m})$ 基于去噪的潜变量和历史上下文自回归生成代理动作，并将生成的动作通过运动自行车模型传播，以确保结果轨迹的物理合理性。

训练。我们固定预训练的 VAE，仅通过最小化噪声预测损失来训练 LDM：

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}^k, \epsilon \sim \mathcal{N}(0,I), k, \boldsymbol{c}} \left[\|\epsilon - \epsilon_\theta(\boldsymbol{z}^k, k, \boldsymbol{c})\|^2 \right] \tag{4}$$

其中 $\epsilon \sim \mathcal{N}(0, I)$ 是高斯噪声，$\epsilon_\theta(\boldsymbol{z}^k, k, \boldsymbol{c})$ 是噪声预测模型，该模型依赖于过去的上下文潜变量 $\boldsymbol{c}$。

C. 可控生成与引导函数

为了实现关键安全驾驶场景的可控生成，我们引入了一个目标函数 $\mathcal{J}(\boldsymbol{\tau})$，它引导去噪过程。由于去噪过程在潜在空间中进行，每个引导迭代步骤首先使用解码器 $\mathcal{D}$ 将潜向量 $\boldsymbol{z}$ 解码为相应的轨迹 $\boldsymbol{\tau}$，公式表示为 $\boldsymbol{\tau} = \mathcal{D}_\theta(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{m})$。在每次反向步骤 t 中，将引导目标的梯度注入预测噪声，遵循 [27]：

$$\tilde{\boldsymbol{\epsilon}}_\theta = \boldsymbol{\epsilon}_\theta - s \, \nabla_{\boldsymbol{z}_t} \mathcal{J}(\mathcal{D}_\theta(\boldsymbol{z}^t, \boldsymbol{x}, \boldsymbol{m})), \tag{5}$$

其中 s 表示引导尺度。然后使用在 (2) 中定义的 DDIM 公式更新扰动潜变量 $\boldsymbol{z}^{k-1}$。

详细来说，目标 $\mathcal{J}(\boldsymbol{\tau})$ 由三个组件组合而成：

$$\mathcal{J}(\boldsymbol{\tau}) = w_b \mathcal{J}_{\text{br}}(\boldsymbol{\tau}) \; + \; w_{ar} \mathcal{J}_{\text{ar}}(\boldsymbol{\tau}) \; + \; w_a \mathcal{J}_{\text{adv}}(\boldsymbol{\tau}) \tag{6}$$

其中 $\mathcal{J}_{br}$ 是非对抗车辆的现实约束条件，确保它们避免相互碰撞和偏离道路。同样地，$\mathcal{J}_{ar}$ 是针对对抗车辆的现实约束条件，确保对抗代理维持合理的行为，不与非对抗车辆相撞或离开道路。对抗目标 $\mathcal{J}_{adv}$ 控制对抗车辆的行为以诱导其与自车发生碰撞。w_b、$w_a r$ 和 w_a 是三个超参数，用于控制三个不同目标的权重。

车辆之间的碰撞惩罚定义如下：

$$\text{veh\_coll\_pens}_{ij}(t) = \begin{cases} 1 - \frac{d_{ij}(t)}{p_{ij}}, & \text{if } d_{ij}(t) \le p_{ij} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

其中，$d_{ij}(t)$ 表示在时间 t 时车辆 i 和 j 之间的欧几里得距离，而 $p_{ij} = r_i + r_j + d_{\text{buffer}}$ 代表由车辆半径之和与预定义的安全缓冲区确定的碰撞阈值。

类似地，地图碰撞惩罚定义为：

$$\text{env\_coll\_pens}_i(t) = \begin{cases} 1 - \frac{d_i(t)}{p_i}, & \text{if } d_i(t) \le p_i \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

其中 $d_i(t)$ 表示从车辆中心到最近的不可行驶区域的距离，而 t 表示时间，p_i 表示发生碰撞前允许的最大位移。

在实践中，$\mathcal{J}_{br}$ 和 $\mathcal{J}_{ar}$ 通过累积所有相关代理和时间步长的车辆碰撞和地图碰撞惩罚来计算。相比之下，对抗目标定义为：

$$\mathcal{J}_{\text{adv}}(\boldsymbol{\tau}) = \sum_{t=1}^{T} min(0, \; d(t) \; - \; p) \tag{9}$$

其中 $d(t)$ 是当前时刻 t 对抗车辆与自我车辆之间的中心到中心距离，p 表示对抗车辆与自我车辆之间的碰撞阈值。相应地，这三个指导目标分别针对非对抗性和对抗性车辆进行单独计算，并独立更新相应的潜在变量。具

体来说，$\mathcal{J}_{br}$ 的梯度被应用于非对抗车辆的潜在表示，而包含 $\mathcal{J}_{ar}$ 和 $\mathcal{J}_{adv}$ 的组合目标的梯度则用于更新与对抗车辆对应的潜在变量，从而确保对其行为进行有针对性的控制。

样本选择模块。在推理阶段，我们为给定场景中的非自车代理生成多个候选未来轨迹，并使用加权分数对它们进行排序：

$$\mathcal{C} = w_g \, \mathcal{J}(\boldsymbol{\tau}) + w_p \, (1 - \Phi(\boldsymbol{\tau})), \tag{10}$$

其中，$\Phi(\boldsymbol{\tau})$ 是一个物理可行性指示函数，定义为：

$$\Phi(\boldsymbol{\tau}) = a_{\text{lon}}(\boldsymbol{\tau}) \wedge a_{\text{lat}}(\boldsymbol{\tau}), \tag{11}$$

其中 $a_{\text{lon}}(\boldsymbol{\tau})$ 表示符合纵向加速度约束，$a_{\text{lat}}(\boldsymbol{\tau})$ 表示符合横向加速度约束。选择分数最低的候选轨迹 $\mathcal{C}$。

IV. 实验结果

本节展示了实验结果，验证了我们提出模型的有效性。我们首先介绍了研究中使用的数据集和评估指标。然后，我们在实际交通模拟和安全关键场景生成任务上将我们的方法与代表性基线进行了比较。此外，我们还进行了一项消融研究，以分析所提出的引导机制的贡献。

A. 数据集

我们在 nuScenes 数据集 [28] 上进行实验，该数据集包含 1,000 个驾驶场景，每个场景持续时间为 20 秒，采样频率为 2 Hz。模型在提供的训练分割上进行训练，并在验证分割上进行评估。遵循 nuScenes 预测挑战的指南，我们使用过去 2 秒（4 个时间步）的运动来预测接下来的 6 秒（12 个时间步）。

B. 评估指标

我们从三个角度评估生成的场景：对抗性、真实性和多样性。对抗性通过诸如 Adv-Ego 碰撞率和 Adv 加速等指标来衡量为自车创建安全关键场景的有效性，数值越高表示对抗行为越强。真实性通过车辆碰撞率（Veh Coll）和地图碰撞率（Env Coll）等指标评估场景的合理性，并进一步细分为不同代理类型之间的碰撞率（Adv-Other, Other-Ego, Other-Other）以及不同类型代理的地图碰撞率（Adv Offroad, Other Offroad），数值越低表示行为越真实。多样性使用最终位移多样性（FDD）[13], [18] 进行量化，更高的 FDD 值表示场景变化更大。此外，我们还采用平均位移误差（ADE）、最终位移误差（FDE）和最小场景最终位移误差（minSFDE）等指标来衡量整体轨迹准确性，并使用推理时间评估生成效率。

C. 实验设置

基线。我们将我们的方法与两个有代表性的基线进行比较。AdvSim [6] 优化了预定义的对抗车辆的加速，以引发碰撞，初始状态由 SimNet [29] 生成。Strive [7] 使用其官方实现，在基于交通模型的学习潜在空间中进行对抗优化。

实现细节。我们的模型是用 PyTorch 实现的，并在四块 NVIDIA RTX 4090 GPU 上训练了六个小时。扩散模型使用 Adam 进行了 200 个纪元的训练，学习率为 5×10^{-4}，扩散步数为 20，测试样本数量为 10。

为了确保在可控性方面的公平比较，选择初始状态下距离自车最近的对抗车辆，并且必须满足可行性约束 [7]。与 Strive 不同，在 Strive 中对抗车辆可能会动态变化，我们在整个场景中固定选定的对抗代理。此外，在所有实验中，自车由基于规则的规划器控制，以确保不同方法之间的一致性。

D. 模拟真实交通

为了评估我们的 LDM 在模拟真实世界交通场景中的有效性，我们在 nuScenes 数据集上进行了对比实验。具体来说，我们将基于无引导扩散的模型与两个代表性基线进行了比较：基于 VAE 的 Strive 模型和基于模仿学习的 AdvSim 模型，在没有对抗优化的情况下进行评估。如表 I 所示，我们的模型实现了逼真且多样化的轨迹生成，证明了其在捕捉复杂交通行为的同时保持多样性的能力。特别地，我们的模型表现出更强的可能性，车辆碰撞率（0.31%）低于 AdvSim（0.78%）和 Strive（0.36%）。这一结果表明，我们的方法更有效地捕获多智能体交互并生成一致的交通行为。此外，与 AdvSim 的确定性生成过程不同，Strive 和我们的模型都采用了基于采样的生成过程，这内在地支持了更多样化的轨迹生成。值得注意的是，我们的模型在保持逼真性的同时实现了具有竞争力的多样性，展示了其同时提高场景合成保真度和灵活性的能力。

表 I

nuScenes 数据集上无引导扩散模型与 AdvSim [6] 和 Strive [7] 的真实交通模拟结果比较。我们将我们的无引导扩散模型与其他两个未经过对抗优化评估的模型进行对比。

模型	现实主义					Diversity
	Veh Coll (%) ↓	Env Coll ↓	ADE (m) ↓	FDE (m) ↓	minSFDE (m) ↓	FDD (m) ↑
AdvSim	0.78	**14.92**	**2.24**	**5.39**	5.39	0.0
Strive	0.36	16.23	2.74	6.65	3.68	**13.82**
Ours	**0.31**	15.99	2.41	5.94	**3.66**	10.25

表 II

nuScenes 数据集上的安全关键交通仿真结果对比。我们将我们的方法与基于规则的规划器进行的安全关键交通仿真中的 AdvSim [6] 和 Strive [7] 进行了比较。

模型	对抗性		现实主义						Efficiency
	Adv-Ego Coll (%) ↑	Adv Acc ↑	Adv Offroad (%) ↓	Other Offroad (%) ↓	Adv-Other Coll (%) ↓	Other-Ego Coll (%) ↓	Other-Other Coll (%) ↓	Other Acc ↓	Infer time (s) ↓
AdvSim	24.72	0.90	15.60	**14.85**	**0.56**	**0.91**	0.11	0.38	338.35
Strive	22.69	0.88	18.94	16.64	0.90	1.08	**0.05**	0.39	609.72
Ours	**38.17**	**1.14**	**11.49**	16.83	5.68	1.47	0.63	**0.33**	**171.60**

E. 模拟安全关键场景

安全关键交通模拟结果的比较显示在表 II 中。与基线模型相比，我们提出的模型在生成对抗场景的同时保持了高度的真实性和高效的生成能力方面具有显著优势。我们的模型达到了最高的对抗有效性，对抗性自我碰撞率为 38.17%，远远超过了 AdvSim（24.72%）和 Strive（22.69%）。这表明该模型有能力生成有效的挑战自动驾驶系统的安全关键场景。同时，对抗行为仍然具有行为合理性。对抗车辆的离路率为 11.49%，显著低于 AdvSim（15.60%）和 Strive（18.94%），表明我们的模型生成了真实且激进的行为而未违反道路约束。在生成效率方面，基于扩散的方法实现了平均推理时间为 171.60 秒，明显快于测试时间优化方法如 AdvSim（338.35 秒）和 Strive（609.72 秒）。总体而言，我们的方法实现了强大的对抗性能，并同时保持了高水平的真实性和生成效率，使其非常适合大规模自动驾驶车辆的安全验证。

F. 消融研究

为了研究提出的引导机制的有效性，我们进行了一项消融研究，比较了三种设置：回复，该设置从数据集中重播原始非自我车辆轨迹；无指导，该设置运行不带引导的基于扩散的模型；以及有指导，该设置在扩散过程中包含所提出的引导模块。

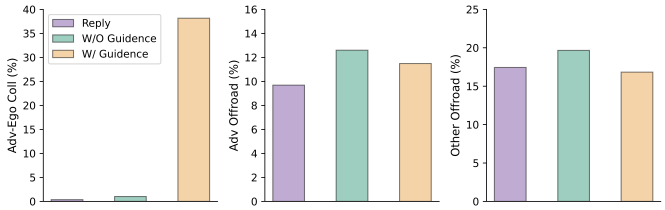

图 3. 消融研究关于指导组件。回复指使用数据集中的重放非自我车辆轨迹进行的模拟。无指导表示我们的无引导扩散模型，而有指导代表我们提出的基于引导的扩散模型。

如图 3 所示，回复设置产生最小的对抗有效性，其对抗性自我碰撞率接近于零，因为代理只是重放数据集轨迹。无指导模型的结果展示了模拟现实交通场景的强大能力，实现了合理的车辆行为和适度的离道率。随着指导模块的集成，有指导模型进一步提高了对抗性，达到最高的 Adv-Ego 碰撞率，证明了对抗性引导的有效性。此外，Adv Offroad 和 Other Offroad 有所减少，表明基于现实的引导有助于提升生成场景的真实感而不降低对抗有效性。这些发现表明，所提出的指导机制在

生成安全关键场景时发挥着至关重要的作用，并保持了高度的真实性。

V. 结论

本文介绍了一个用于模拟安全关键交通场景的引导 LDM 框架。通过在一个从基于图的 VAE 学习到的紧凑潜在空间中执行扩散，我们的方法有效地捕捉了复杂的多代理互动并提升了计算效率。为了实现对安全关键场景的可控生成，我们提出了创新性的指导目标来引导扩散过程以生成对抗性和行为上现实的驾驶行为。此外，我们引入了一个基于物理可行性检查的简单但有效的样本选择模块，进一步增强了生成场景的物理合理性。在 nuScenes 数据集上的广泛实验表明，我们的方法在对抗有效性、真实感、多样性和生成效率方面超过了现有的基线方法。未来，我们计划探索将大型语言模型（LLMs）或 AI 代理集成进来，以进一步提升安全关键交通模拟的可控性和性能。

参考文献

[1] J. Guo, U. Kurup, and M. Shah, "Is it safe to drive? an overview of factors, metrics, and datasets for driveability assessment in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3135–3151, 2019.

[2] Y. Kang, H. Yin, and C. Berger, "Test Your Self-Driving Algorithm: An overview of publicly available driving datasets and virtual testing environments," *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 2, pp. 171–185, 2019.

[3] W. Ding, C. Xu, M. Arief, H. Lin, B. Li, and D. Zhao, "A survey on safety-critical driving scenario generation—a methodological perspective," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 6971–6988, 2023.

[4] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *the Proceedings of Conference on Robot Learning*, pp. 1–16, 2017.

[5] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *the Proceedings of 2018 21st International Conference on Intelligent Transportation Systems*, pp. 2575–2582, 2018.

[6] J. Wang, A. Pun, J. Tu, S. Manivasagam, A. Sadat, S. Casas, M. Ren, and R. Urtasun, "AdvSim: Generating safety-critical scenarios for self-driving vehicles," in *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9909–9918, 2021.

[7] D. Rempe, J. Philion, L. J. Guibas, S. Fidler, and O. Litany, "Generating useful accident-prone driving scenarios via a learned traffic prior," in *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17305–17315, 2022.

[8] N. Hanselmann, K. Renz, K. Chitta, A. Bhattacharyya, and A. Geiger, "KING: Generating safety-critical driving scenarios for robust imitation via kinematics gradients," in *the Proceedings of European Conference on Computer Vision*, pp. 335–352, 2022.

[9] W. Mao, C. Xu, Q. Zhu, S. Chen, and Y. Wang, "Leapfrog diffusion model for stochastic trajectory prediction," in *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5517–5526, 2023.

[10] M. Niedoba, J. Lavington, Y. Liu, V. Lioutas, J. Sefas, X. Liang, D. Green, S. Dabiri, B. Zwartsenberg, A. Scibior, *et al.*, "A diffusion-model of joint interactive navigation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[11] Z. Zhong, D. Rempe, D. Xu, Y. Chen, S. Veer, T. Che, B. Ray, and M. Pavone, "Guided conditional diffusion for controllable traffic simulation," in *the Proceedings of IEEE International Conference on Robotics and Automation*, pp. 3560–3566, 2023.

[12] C. Jiang, A. Cornman, C. Park, B. Sapp, Y. Zhou, D. Anguelov, *et al.*, "MotionDiffuser: Controllable multi-agent motion prediction using diffusion," in *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9644–9653, 2023.

[13] Y. Xie, X. Guo, C. Wang, K. Liu, and L. Chen, "AdvDiffuser: Generating adversarial safety-critical driving scenarios via guided diffusion," in *the Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 9983–9989, 2024.

[14] E. Pronovost, M. R. Ganesina, N. Hendy, Z. Wang, A. Morales, K. Wang, and N. Roy, "Scenario diffusion: Controllable driving scenario generation with diffusion," *Advances in Neural Information Processing Systems*, vol. 36, pp. 68873–68894, 2023.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

[16] J. M. Scanlon, K. D. Kusano, T. Daniel, C. Alderson, A. Ogle, and T. Victor, "Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain," *Accident Analysis & Prevention*, vol. 163, p. 106454, 2021.

[17] Q. Li, Z. Peng, L. Feng, Q. Zhang, Z. Xue, and B. Zhou, "MetaDrive: Composing diverse driving scenarios for generalizable reinforcement learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3461–3475, 2022.

[18] S. Suo, K. Wong, J. Xu, J. Tu, A. Cui, S. Casas, and R. Urtasun, "Mixsim: A hierarchical framework for mixed reality traffic simulation," in *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9622–9631, 2023.

[19] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.

[20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[21] M. Peng, K. Chen, X. Guo, Q. Zhang, H. Lu, H. Zhong, D. Chen, M. Zhu, and H. Yang, "Diffusion models for intelligent transportation systems: A survey," *arXiv preprint arXiv:2409.15816*, 2024.

[22] C. Xu, D. Zhao, A. Sangiovanni-Vincentelli, and B. Li, "DiffScene: Diffusion-based safety-critical scenario generation for autonomous vehicles," in *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

[23] W.-J. Chang, F. Pittaluga, M. Tomizuka, W. Zhan, and M. Chandraker, "SAFE-SIM: Safety-critical closed-loop traffic simulation with diffusion-controllable adversaries," in *the Proceedings of European Conference on Computer Vision*, pp. 242–258, 2024.

[24] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu, "Executing your commands via motion diffusion in latent space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18000–18010, 2023.

[25] Z. Zhong, D. Rempe, Y. Chen, B. Ivanovic, Y. Cao, D. Xu, M. Pavone, and B. Ray, "Language-guided traffic simulation via scene-level diffusion," in *the Proceedings of Conference on Robot Learning*, pp. 144–177, 2023.

[26] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[27] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[28] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

[29] L. Bergamini, Y. Ye, O. Scheel, L. Chen, C. Hu, L. Del Pero, B. Osiński, H. Grimmett, and P. Ondruska, "SimNet: Learning reactive self-driving simulations from real-world observations," in *the Proceedings of IEEE International Conference on Robotics and Automation*, pp. 5119–5125, 2021.