SAFE-SIM: Safety-Critical Closed-Loop Traffic Simulation with Controllable Adversaries

Wei-Jer Chang[1] Francesco Pittaluga[2] Masayoshi Tomizuka[1] Wei Zhan[1] Manmohan Chandraker[2,3]

[1] UC Berkeley [2] NEC Labs America [3]UC San Diego

Abstract

Evaluating the performance of autonomous vehicle planning algorithms necessitates simulating long-tail safety-critical traffic scenarios. However, traditional methods for generating such scenarios often fall short in terms of controllability and realism, and neglect the dynamics of agent interactions. To mitigate these limitations, we introduce SAFE-SIM*, a novel diffusion-based controllable closed-loop safety-critical simulation framework. Our approach yields two distinct advantages: 1) the generation of realistic long-tail safety-critical scenarios that closely emulate real-world conditions, and 2) enhanced controllability, enabling more comprehensive and interactive evaluations. We develop a novel approach to simulate safety-critical scenarios through an adversarial term in the denoising process, which allows an adversarial agent to challenge a planner with plausible maneuvers, while all agents in the scene exhibit reactive and realistic behaviors. Furthermore, we propose novel guidance objectives and a partial diffusion process that enables a user to control key aspects of the generated scenarios such as the collision type and aggressiveness of the adversarial driver while maintaining the realism of the behavior. We validate our framework empirically using the NuScenes dataset, demonstrating improvements in both realism and controllability. These findings affirm that diffusion models provide a robust and versatile foundation for safety-critical, interactive traffic simulation, extending their utility across the broader landscape of autonomous driving.*

1. Introduction

A key safety feature of autonomous vehicles (AVs) is their ability to navigate near-collision events in real-world scenarios. However, these events rarely occur on roads and testing AVs in such high-risk situations on public roads is unsafe. Therefore, simulation is indispensable in the development and assessment of AVs, providing a safe and reliable means to study their safety and dependability.

Recent studies have primarily created static scenarios that challenge planners, neglecting dynamic, *closed-loop* simulations. This oversight fails to account for the adaptive responses of other agents, crucial for detailed safety evaluations. Additionally, such simulations lack *controllability*, limiting the exploration to a single adversarial outcome per scenario.

Table 1. **Comparison of methods.** Our contribution is the development of a framework for (a) safety-critical (b) closed-loop (c) controllable adversarial simulations. These aspects are not concurrently present in previous frameworks. We are the first to enable an ego planner to be tested against controllable adversaries with varied behavior patterns.

Method	Safety-Critical	Controllable	Controllable Adversary	Evaluate Planner	Closed-Loop	Real-World
CTG [1]	×	✓	×	×	✓	✓
CTG++ [2]	×	✓	×	×	✓	✓
STRIVE [3]	✓	×	×	✓	✓	✓
DiffScene [4]	✓	✓	×	✓	×	×
SAFE-SIM (Ours)	✓	✓	✓	✓	✓	✓

In this work, we introduce SAFE-SIM, a *closed-loop* simulation framework for generating safety-critical scenarios, with a particular emphasis on *controllability* and *realism* for the behavior of agents, which allows simulations over a long-horizon as needed to evaluate AV planning algorithms (Fig. 1). Different from prior works [1–4] that primarily adhere to rule-constraint satisfaction, our approach enhances *controllability* by modulating adversarial vehicle behaviors within identical scenarios, thereby facilitating a broader exploration of potential outcomes. See Tab. 1 for a comprehensive comparison of these approaches.

Our approach builds upon recent developments in controllable diffusion models [1, 5, 6]. Specifically, we adopt a test-time guidance to direct the denoising phase of the diffusion process, using the gradients from differentiable objectives to enhance scenario generation, enabling generation of adversarial scenarios in which and adversarial agent collides with the ego agent behaving according to specific planning policy. Additionally, we develop an novel approach, which we refer to as Partial Diffusion that introduces trajectory proposals into the diffusion process to provide a high degree of controllability over the type of collision scenario.

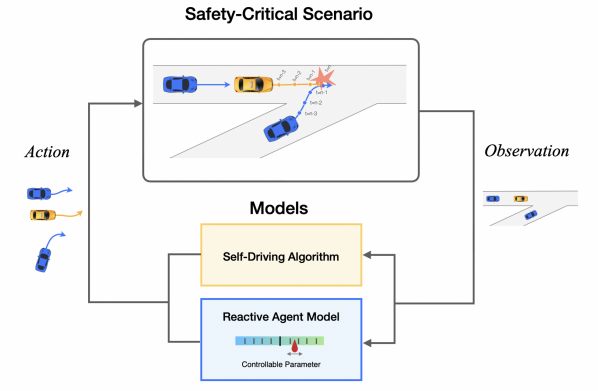

Figure 1. **Overview of Controllable Safety-Critical Closed-Loop Simulation Framework.** This framework involves evaluating a planner within scenarios featuring multiple reactive agents. These agents possess two distinct qualities: they are trained using real-world driving data, ensuring realistic behavior, and their actions are controllable particularly their adversarial behavior.

Overall, our balanced integration of adversarial objectives with regularization during the guidance phase combined with Partial Diffusion allows for refined control over the conditions of the generated scenarios, ensuring both their realism and relevance to safety-critical testing.

In our study, we conduct experiments on the nuScenes dataset [7]. Our results demonstrate a marked improvement in the controllability and realism of scenarios compared to traditional adversarial scenario generation methods. Furthermore, we showcase the advantage of our proposed methods in controlling variations of adversarial behavior to test AV planners. These attributes make our approach particularly well-suited for the closed-loop simulation of AVs, providing a more reliable and comprehensive framework for safety evaluation.

2. Problem Formulation

We consider a simulated interactive traffic scenario consisting of N agents; one is the ego vehicle controlled by the *planner* π, and the remaining $N-1$ are reactive agents modeled by a function g. Our objective is to create a safety-critical *closed-loop* collision simulation, where reactive agents demonstrate *realistic*, *controllable* behavior. Of the $N-1$ reactive agents, one is considered the adversarial agent (denoted as agent a) meant to collide with the adversarial agent.

The adversarial agent, formulated within the reactive agent model g, is governed by an adversarial term designed (detailed in Sec. 4) to be both controllable and adversarial to the planner π. This setup allows the adversarial agent to pose direct challenges to π, testing its resilience in complex scenarios. The dual role of the adversarial agent and non-adversarial agents ensures that while it challenges π, the overall simulation environment plausibly represents real-world driving conditions.

At any given timestep t, the states of the N vehicles are represented as $\mathbf{s}_t = [\mathbf{s}_t^1, \ldots, \mathbf{s}_t^N]$, where $\mathbf{s}_t^i = (x_t^i, y_t^i, v_t^i, \theta_t^i)$ indicates the 2D position, speed, and yaw of vehicle i. The corresponding actions for each vehicle are $\mathbf{a}_t = [\mathbf{a}_t^1, \ldots, \mathbf{a}_t^N]$, with $\mathbf{a}_t^i = (\dot{v}_t^i, \dot{\theta}_t^i)$ representing the acceleration and yaw rate. To predict the state at the next timestep $t+1$, a transition function f is used, which computes $\mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{a}_t)$ based on current state and action. We adopt unicycle dynamics as the transition function.

Each agent's decision context is $\mathbf{c}_t^i$, which includes the agent-centric map I^i and the T_{hist} historical states of neighboring vehicles from time $t - T_{\text{hist}}$ to t, defined as $\mathbf{s}_{t-T_{\text{hist}}:t} = \{\mathbf{s}_{t-T_{\text{hist}}}, \ldots, \mathbf{s}_t\}$. In closed-loop traffic simulation, each agent continuously generates and updates its trajectory based on the current decision context $\mathbf{c}_t^i$. After generating a trajectory, the simulation executes the first few steps of the planned actions before updating $\mathbf{c}_t^i$ and re-planning.

Planner π : The planner π determines the ego vehicle's future trajectory over a time horizon t to $t+T$. The planned state sequence is denoted by $s_{t:t+T}^1 = \pi(\mathbf{c}_t^1)$, where $\pi(\mathbf{c}_t^1)$ processes the historical states and map data within $\mathbf{c}_t^1$ to plan future states based on the current scene context.

Reactive Agents g : The reactive agent model g, parameterized by θ, is designed to simulate the behavior of the $N-1$ non-ego vehicles, represented by the set $\{s_{t:t+T}^i\}_{i=2}^N$. Each vehicle's state sequence, $s_{t:t+T}^i$, is generated by $g_\theta(\mathbf{c}_t^i, \psi_i)$, which incorporates the decision context $\mathbf{c}_t^i$ and a set of control parameters ψ_i unique to each agent. These parameters ψ_i enable the fine-tuning of individual behaviors within the simulation. In our approach, we train the model g on real-world driving data to ensure the trajectories it produces are not only controllable, supporting the generation of various safety-critical scenarios, but also realistic.

3. Diffusion Models for Traffic Simulation

For closed-loop safety-critical traffic simulation, the reactive agents, especially the adversarial agent, should be 1) controllable, and 2) realistic. With recent advances in controllable diffusion models [1, 5, 6], we adopt trajectory diffusion models to generate realistic simulations.

We define the model's operational trajectory as τ, which comprises both action and state sequences: $\tau := [\tau_a, \tau_s]$. Specifically, $\tau_a := [a_0, \ldots, a_{T-1}]$ represents the sequence of actions, while $\tau_s := [s_1, \ldots, s_T]$ denotes the corresponding sequence of states. Following the approach described in [1], our model predicts the action sequence τ_a, and the state sequence τ_s can be derived starting from the initial state s_0 and dynamic model f.

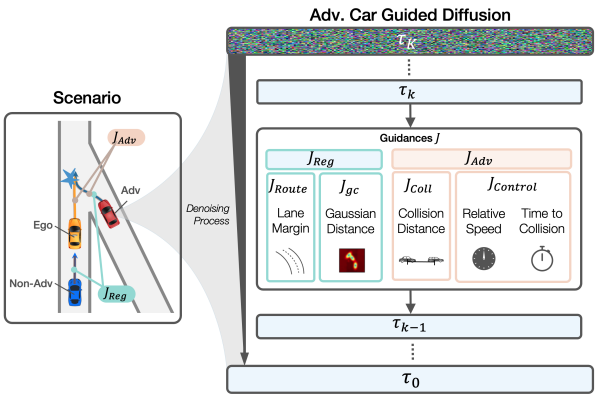

Figure 2. **Guided Diffusion Process for the Adversarial Agent.** This process optimizes the adversarial agent's trajectory using the adversarial cost function J_{adv} in relation to the ego vehicle. Simultaneously, it applies regularization through J_{reg} for maintaining realism.

4. Diffusion Models for Safety-Critical Traffic Simulation

The diffusion model, once trained on realistic trajectory data, inherently reflects the behavioral patterns present in its training distribution. However, to effectively simulate and analyze safety-critical scenarios, there is a crucial need for a mechanism that allows for the controlled manipulation of agent behaviors [1, 5]. This is particularly important for generating adversarial behaviors and ensuring long-term scene consistency in simulations.

4.1. Guiding Reactive Agents

Our approach specifically introduces guidance to the sampled trajectories at each denoising step, aligning them with predefined objectives $J(\tau)$. The concept of guidance involves using the gradient of J to subtly perturb the predicted mean of the model at each denoising step [8][6]:

$$\tilde{\tau}_0 = \hat{\tau}_0 - \alpha \Sigma_k \nabla_{\tau_k} J(\hat{\tau}_0) \tag{1}$$

In practice, diversifying the behavior of adversarial agents within the *same* scenarios is crucial for a thorough assessment of AVs. Despite the significance of this challenge, it remains largely unexplored in previous works [1][3].

The loss function for the non-reactive agents, $J(\tau)$, consists of a collision term J_{coll}, which encourages collisions between the adversarial agent and the ego agent, two control terms J_{v} and J_{ttc}, which control the relative speed and time-to-collision between the ego and adversarial agent respectively, a regularization term J_{Gauss}, which discourages collisions between the reactive agents, and a route guidance term J_{route}, which discourages the reactive agents from going outside the road:

$$J(\tau) = \rho \underbrace{(J_{\text{coll}} + J_{\text{v}} + J_{\text{ttc}})}_{J_{\text{adv}}(\tau)} + \underbrace{J_{\text{route}} + J_{\text{Gauss}}}_{J_{\text{reg}}(\tau)}, \tag{2}$$

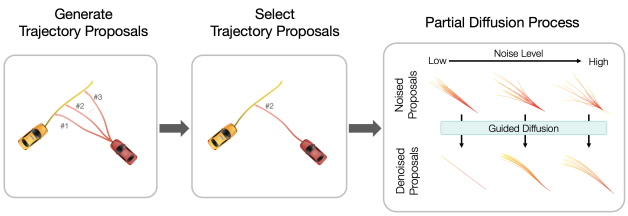

Figure 3. **Framework for Partial Diffusion.**

where ρ denotes a scalar weight that determines whether a reactive agent behaves adversarially towards the ego agent, i.e., whether it attempts to collide with the ego agent.

Safety Criticality of Collisions: We control the relative speed J_v between the ego and adversary at each time step (v_t^1 and v_t^a) and the time-to-collision (TTC) cost J_{ttc} [9] to control the safety criticality of potential collisions, with the latter given by:

$$J_{\text{ttc}} = \sum_{t=1}^{T} -\exp\left(-\frac{\tilde{t}_{\text{col}(t)}^2}{2\lambda_t} - \frac{\tilde{d}_{\text{col}(t)}^2}{2\lambda_d}\right), \tag{3}$$

where $\tilde{t}_{\text{col}(t)}$ is the time to collision at time t, $\tilde{d}_{\text{col}(t)}$ is the distance to collision and λ_t and λ_d are bandwidth parameters for time and distance. This formula uses a constant velocity assumption. Intuitively, The time-to-collision cost favors scenarios with high relative speeds and challenging collision angles for the ego vehicle to avoid. For a detailed explanation of J_{ttc}, see supplementary.

4.2. Partial Diffusion: Controlling Collision Types

We introduce a novel approach through a partial diffusion process, utilizing trajectory proposals to initiate the diffusion process. This methodology enables the variation in collision types by the adversarial agent within the diffusion, tailoring the adversarial outcomes to specific evaluation needs, the results are discussed in Sec. 5.2.

Figure 3 illustrates our framework, which is divided into three main steps to generate trajectory proposals for various collision scenarios. First, we create initial trajectory proposals (τ_0) aimed at capturing different types of collisions. The next critical step involves setting the partial diffusion ratio γ, which defines the specific point in the process, $k_p = \gamma \cdot K$, at which we start modifying the trajectory. Starting from step k_p, we adjust the trajectory by adding a precise level of Gaussian noise $\epsilon \sim N(0, I)$: $\hat{\tau}_{k_p} = \sqrt{\bar{\alpha}_{k_p}}\tau_0 + \sqrt{1 - \bar{\alpha}_{k_p}}\epsilon$. The final stages include removing noise and using guided diffusion for the rest of the k_p steps to refine the trajectory into a realistic path that suits our collision scenario goals.

To generate the trajectory proposals, we develop a rule-based approach in which we first identify the centerlines of the ego and adversarial agent and then search for potential

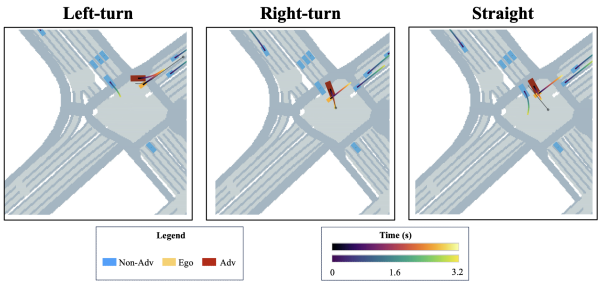

Figure 4. **Qualitative Results of Controllability of Partial Diffusion.** The depicted safety-critical scenarios demonstrate the framework's capability to generate controllable adversarial behavior for comprehensive planner assessment.

Method	Collision (%) ↑	Other Offroad (%) ↓	Adv Offroad (%) ↓	Collision Rel Speed (m/s) ↓	Realism ↓	Time (s) ↓
STRIVE	36.4	2.2	**11.4**	5.52	0.85	427.2 ± 169.8
SAFE-SIM	**43.2**	**1.9**	**11.4**	**-0.12**	**0.38**	104.5 ± 17.7

Table 2. **Safety-critical Traffic Simulation.** We compare the our approach against STRIVE [3] for safety-critical traffic simulation with a rule-based planner. Our outperforms STRIVE on all metrics.

TTC Cost Weight	TTC Cost	Coll Speed (m/s)	Coll Angle (deg)	Coll Rate (%) ↑	Realism ↓
0.0	0.18	2.45	-7.43	48.2	0.76
1.0	0.21	2.30	0.43	53.6	0.79
2.0	0.26	3.78	-17.0	60.7	0.81

Table 3. **Controlling Time to Collision (TTC).** The table shows the impact of different TTC Cost weights on collision scenarios. Increasing the TTC Cost weight results in an increase in collision rate, suggesting a heightened challenge for the ego vehicle in avoiding collisions.

intersections of their respective centerlines. This method allows for precise control over the diffusion trajectory, enabling the creation of customized collision scenarios by adversarial agents. Users can adjust γ to fine-tune the balance between explicit control and the model's inherent guidance, thus enhancing the scope of evaluation for autonomous vehicle technologies.

5. Experiments

We validate the efficacy of our proposed framework via experiments employing real-world driving data on nuScenes [10]. These experiments are conducted using a rule-based planner, as delineated in [3]. The findings reveal that our framework successfully induces controllable adversarial behavior in realistic safety-critical situations, a crucial aspect for exhaustive testing of autonomous vehicles (AVs).

5.1. Evaluation of Safety-Critical Traffic Simulation

We initiated our evaluation by comparing our method with STRIVE [3], recognized for its proficiency in generating adversarial safety-critical scenarios using a learned traffic model and adversarial optimization in the latent space. Our comparison focused on the collision rates between the ego and adversarial agents ("Collision"), the collision rates, the off-road rate of the adversarial agent ("Adv Offroad") and the other agents ("Other Offroad"), the speed of the adversarial vehicle ("Adv Speed"), the realism of the generated scenario ("Realism"), as proposed in [1], and the simulation time per scenario ("Time). The results are presented in Tab. 2. Our method excels in all metrics, especially collisions and realism.

5.2. Evaluation: Controlling Safety-Criticality and Collision Types

Our method demonstrates enhanced controllability in generating adversarial scenarios compared to previous approaches. Specifically, we focus on controlling two critical aspects: time-to-collision before the interaction between the ego and the adversarial vehicle and the collision types.

We manipulate the scenario's safety-criticality by adjusting the relative weight of the TTC cost. To assess the impact of these adjustments, we measure the average TTC cost shortly before a collision occurs (0.5 seconds). Our observations, detailed in Tab. 3, show that increasing the TTC weight raises the TTC cost. Notably, while the relative collision speed remains fairly consistent, the collision angle shifts, indicating a higher difficulty in avoiding ego-adversary collisions. Based on qualitative examples in the supplementary material, these changing angles could potentially lead to more challenging cases for the ego vehicle, thereby enhancing the safety-critical aspect of the simulation.

As shown in Fig. 4, we demonstrates how our proposed partial diffusion process can generate a variety of collision types by different trajectory proposals. For more qualitative results, please see the supplementary materials.

6. Conclusion

In this work, we present a closed-loop simulation framework that employs guided diffusion models and partial diffusion for generating diverse, safety-critical scenarios to evaluate AV algorithms. A key aspect of our method lies in its ability to vary the types of adversarial behavior within collision scenarios. By integrating adversarial objectives and partial diffusion into the diffusion model's architecture, we achieve detailed control over the spectrum of adversarial actions. This versatility enables our framework to produce a broader range of realistic and manageable scenarios, setting a new standard in adversarial scenario generation beyond the limitations of existing approaches.

References

[1] Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. Guided conditional diffusion for controllable traffic simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3560–3566. IEEE, 2023. 1, 2, 3, 4

[2] Ziyuan Zhong, Davis Rempe, Yuxiao Chen, Boris Ivanovic, Yulong Cao, Danfei Xu, Marco Pavone, and Baishakhi Ray. Language-guided traffic simulation via scene-level diffusion. *arXiv preprint arXiv:2306.06344*, 2023. 1

[3] Davis Rempe, Jonah Philion, Leonidas J Guibas, Sanja Fidler, and Or Litany. Generating useful accident-prone driving scenarios via a learned traffic prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17305–17315, 2022. 1, 3, 4

[4] Chejian Xu, Ding Zhao, Alberto Sangiovanni-Vincentelli, and Bo Li. Diffscene: Diffusion-based safety-critical scenario generation for autonomous vehicles. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023. 1

[5] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022. 1, 2, 3

[6] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13756–13766, 2023. 1, 2, 3

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2

[8] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 3

[9] Haruki Nishimura, Jean Mercat, Blake Wulfe, Rowan Thomas McAllister, and Adrien Gaidon. Rap: Risk-aware prediction for robust planning. In *Conference on Robot Learning*, pages 381–392. PMLR, 2023. 3

[10] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 4