CAT: Closed-loop Adversarial Training for Safe End-to-End Driving

Anonymous Author(s)
Affiliation
Address
email

Abstract: Driving safety is a top priority for autonomous vehicles. Orthogonal to prior work handling accident-prone traffic events by algorithm designs at the policy level, we present a general iterative learning framework called **C**losed-loop **A**dversarial **T**raining (CAT) for safe end-to-end driving. CAT aims to continuously improve safety performance by training the driving agent on safety-critical scenarios that are dynamically generated over time. A novel resampling technique is developed to turn normal real-world driving scenarios into safety-critical ones through probabilistic factorization, where the adversarial traffic flow is cast as the product of standard motion prediction sub-problems. Consequently, CAT is able to utilize pre-trained motion forecasting models to launch more effective physical attacks with significantly less computational cost compared to existing safety-critical scenario generation methods. We incorporate CAT into the MetaDrive simulator and validate our approach on hundreds of driving scenarios imported from real-world driving datasets. Experimental results demonstrate that CAT can generate effective safety-critical scenarios countering the agent being trained. After training, the agent can achieve superior driving safety in both normal and adversarial traffic scenarios on the hold-out test set. The demo video is available in the supplementary materials.

Keywords: Safety-Critical Scenario Generation, Adversarial Training, End-to-End Driving

1 Introduction

While end-to-end driving has achieved promising performance in urban piloting [1] and track racing [2], safely handling accident-prone traffic events is still one of the crucial capabilities for both human driving and autonomous driving (AD). It is important to ensure AI driving safety in risky situations before real-world deployment [3]. However, it is insufficient to train or evaluate the safe end-to-end driving agents on safety-critical scenarios only collected from real-world traffic datasets [4, 5] since such events of interest are extremely rare [6, 7].

Prior work improves the driving agent against safety-critical scenarios through rule-based reasoning [8], motion verification [9], constrained reinforcement learning [10], etc. Orthogonal to the elaborate algorithm designs at the policy level, recent studies obtain robust driving policies at the environmental level by creating accident-prone scenarios as augmented training samples [11, 12]. Nevertheless, the learned policy may easily overfit a fixed set of safety-critical events but fail to handle unknown hazards. The alternative is to dynamically generate challenging scenarios that match the current capability of the driving agent in a closed-loop manner. However, the state-of-the-art safety-critical scenario generation methods [11, 12, 13] are not yet applicable for that purpose due to the following reasons: (i) *Scene generalizability*: probabilistic graph methods like CausalAF [11] require human prior knowledge of each scene graph and thus cannot scale to large and complex driving datasets; (ii) *Model dependency*: kinematics gradient methods like KING [12] relies on the

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

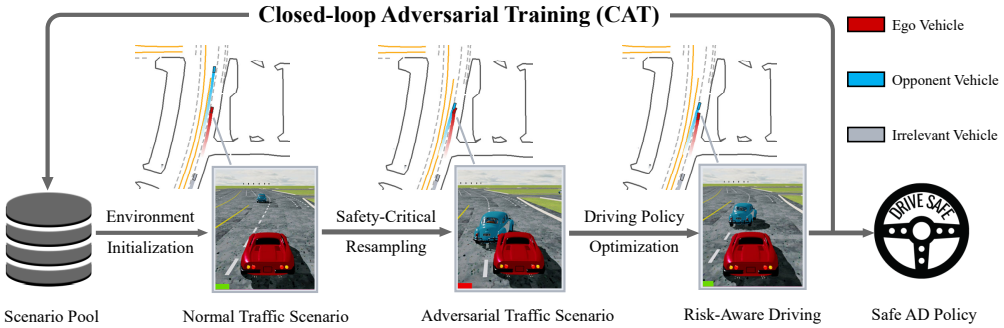

Figure 1: CAT iterates over safety-critical scenario generation and driving policy optimization in a closed-loop manner. In this example, the safety-critical resampling technique alters the behavior of the opponent vehicle (blue car) such that it suddenly cut into the lane of the ego vehicle (red car), enforcing the agent to learn risk-aware driving skills such as deceleration and yielding.

forward simulation of the running policy and the backward propagation based on the vehicle kinematics, which might not be accessible in the model-free end-to-end driving; (iii) *Time efficiency:* autoregression-based generation methods like STRIVE [13] take minutes to optimize the adversarial traffic per scenario, which is time prohibitive for large-scale training with millions of episodes.

In this paper, we present the Closed-loop Adversarial Training (CAT) framework for safe end-to-end driving. As shown in Fig. 1, CAT imports driving scenarios from real-world driving logs and then generates safety-critical counterparts as adversarial training environments tailored to the current driving policy. The agent continuously learns to address emerging challenges and improves risk awareness in closed-loop training. Given that CAT directly launches physical attacks against the estimated ego trajectory, the proposed framework is thus agnostic to the policy used by the agent and is compatible with a wide range of end-to-end learning approaches, including reinforcement learning (RL) [14], imitation learning (IL) [15], human-in-the-loop feedback (HF) [16], etc.

One crucial component of the proposed framework is a novel factorized safety-critical resampling technique that efficiently turns normal driving scenarios into safety-critical ones during training. Specifically, we cast the safety-critical traffic generation as the risk-conditioned Bayesian probability maximization and decompose it into the multiplication of standard motion forecasting sub-problems. Thus, we can utilize off-the-shelf motion forecasting models [17, 18] as the learned prior to generate adversarial scenarios with high fidelity, diversity, and efficiency. Compared to previous safety-critical traffic generation methods, the proposed technique obtains a higher attack success rate while significantly reducing the computational cost, making the CAT framework effective and efficient for end-to-end training.

To demonstrate the efficacy of our approach, we incorporate the proposed CAT framework into the MetaDrive simulator [19] and compose adversarial traffic environments from a hundred complex driving scenarios in a closed-loop manner to train RL-based driving agents without any ad hoc safety designs. Experimental results show that CAT brings realistic and challenging physical attacks during training, and the resulting agent obtains superior driving safety in both normal and adversarial traffic scenarios on the hold-out test set. The contributions of this paper are summarized as follows:

i) We present the closed-loop adversarial training framework for end-to-end safe driving, which is agnostic to the policy learning method and the policy function design.

ii) We propose an efficient safety-critical scenario generation technique tailored to end-to-end policy learning, which balances attack success rate and computation cost by resampling the learned traffic prior.

iii) We incorporate our approach into the MetaDrive simulator and demonstrate it generates effective adversarial samples during training and substantially improves driving safety in complex testing scenarios imported from the real world.

2 Related Work

Adversarial Training for Autonomous Driving. Deep neural networks (DNNs), pervasively used in learning-based AD systems, are found vulnerable to adversarial attacks [20, 21]. Recent studies tend to manipulate the physical environment to generate realistic yet adversarial observation sequences from LiDAR inputs [22], camera inputs [23], and other physical-world-resilient objectives [24]. Compared to the above work focusing on perception, adversarial training for AD decision-making is much less explored. Ma et al. [25] first investigate the adversarial RL on an autonomous driving scenario. Wachi [26] employ the multi-agent DDPG algorithm [27] to enforce the competition between player and non-player vehicles. In addition to algorithmic level designs, a more natural but less explored approach is to iteratively propose challenging scenarios during training [28]. There is a line of works on evolving training environments in RL [29, 30]. However, existing approaches are evaluated only in simplified environments like bipedal walker and heuristically modify the terrain or static barriers, which is not meaningful for AD tasks. In this work, we focus on generating realistic and safety-critical traffic scenarios to facilitate closed-loop adversarial training for end-to-end driving.

Safety-critical Traffic Scenario Generation. Safety-critical traffic scenario generation is of great value in adaptive stress testing [31] and corner case analysis [32] for the research and development of autonomous vehicles. L2C [33] learns to place and trigger a cyclist to collide with the target vehicle via RL algorithms, but it is insufficient to model complex vehicle interactions in real-world scenes. For robust imitation learning, kinematics gradients [12] and black-box optimization [22] can be used to magnify traffic risks. However, it relies on the forward simulation of the running policy and the backward propagation based on the vehicle kinematics, which might not be accessible in model-free end-to-end driving. CausalAF [11] builds scenario causal graphs to uncover behavior of interest and generates additional training samples to improve the robustness of driving policies. Nevertheless, the evaluations are limited to three scenarios since it requires human prior knowledge of each scene and thus hardly scale to the larger dataset. STRIVE [13] constructs a latent space to constrain the traffic prior and searches for the best responsive mapping via gradient-based optimization on that dense representation. Despite its impressive results on realistic traffic flows, the autoregression on raster maps takes several minutes to optimize the adversarial traffic for each scene, which brings about a costly computational burden for periodic policy optimization. We refer to the survey [34] for more details. Different from the above literature, we propose a novel adversarial traffic generation algorithm for real-world scenarios with an admissible time consumption, making it viable for large-scale policy iterations involving millions of episodes.

3 Closed-loop Adversarial Training Framework

We present the Closed-loop Adversarial Training (CAT) framework for safe end-to-end driving. As shown in Fig. 1, CAT iterates over safety-critical scenario generation and driving policy optimization in a closed-loop manner. In this section, we first formulate the closed-loop adversarial training as a min-max problem and then introduce the factorization of adversarial traffic and the practical implementation of CAT.

3.1 Problem Formulation

Although CAT is designed to accommodate a range of driving policies, we focus on RL-based AD in this work which is formulated as Markov Decision Process (MDP) [35] in the form of (S, A, R, f). S and A denote the state and action spaces, respectively. The reward function $R = d - \alpha c$ wherein d is the displacement toward the destination and c is a boolean indicating collision with other objects. α is a hyper-parameter for the reward shaping. f is the transition function to describe the dynamics of the traffic scenario. The goal is to maximize the expected return $J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T} R(s_t, a_t)\right]$ the driving policy π receives within the time horizon T, where $\tau \sim \pi$ is short handed for $a_t \sim \pi(\cdot|s_t), s_{t+1} \sim f(\cdot|s_t, a_t)$.

When importing a real-world traffic scenario, CAT manipulates original traffic trajectories to magnify the possibility of traffic collisions with the agent itself ($\mathbb{E}[c] \uparrow$). Consequently, the modified adversarial traffic dynamics $s_{t+1}^{Adv} \sim f^{Adv}(\cdot|s_t, a_t)$ naturally hinders total rewards the agent receives ($\mathbb{E}[\Sigma R] \downarrow$). CAT aims to enhance the robustness of the learning agent via the following adversarial optimization:

$$\max_{\pi} \min_{f^{Adv}} \; J(\pi, f^{Adv}). \tag{1}$$

3.2 Factorized Safety-Critical Resampling

The fundamental problem is to construct f^{Adv} by generating compliant future traffic trajectories that are prone to collisions with the agent's rollout. To formalize the traffic collisions, we denote the vehicle controlled by the learning agent as the ego vehicle (EV) and other vehicles as opponent vehicles (OVs) and represent a traffic scenario as a tuple $(M, S_{1:T}^{\text{EV}}, \boldsymbol{S}_{1:T}^{\text{OV}})$ with duration T time steps. Here, the High-Definition (HD) road map M consists of road shapes, traffic signs, traffic lights, etc. $S_{1:t}^{\text{EV}}$ denotes the past states of the EV. $\boldsymbol{S}_{1:t}^{\text{OV}}$ is an N-element array $[S_{1:t}^{\text{OV}_1}, ..., S_{1:t}^{\text{OV}_N}]$, wherein each element stands for the past states of the corresponding OV. For simplicity, we denote $X = (M, S_{1:t}^{\text{EV}}, \boldsymbol{S}_{1:t}^{\text{OV}})$ as the information cutoff by step t and $Y^{\text{EV}} = S_{t:T}^{\text{EV}}$, $\boldsymbol{Y}^{\text{OV}} = \boldsymbol{S}_{t:T}^{\text{OV}}$ are the future trajectories of EV and OVs starting from t, respectively. Y^{EV} is conditioned on the RL agent π. The cutoff step t is fixed. We define a binary random variable $Coll = \{True, False\}$ to denote whether Y^{EV} collides with $\boldsymbol{Y}^{\text{OV}}$. Consequently, the optimization of f^{adv} can be cast as trajectory posterior probability maximization under the condition of any collision:

$$\min_{f^{Adv}} J(\pi, f^{Adv}) \Leftrightarrow \max_{\boldsymbol{Y}^{\text{OV}}} \; \mathbb{P}(\boldsymbol{Y}^{\text{OV}}|Coll = True, X). \tag{2}$$

Considering that the opponent vehicle must launch effective attacks based on the potential ego behavior while the agent's future action sequence is also responsive and even defensive to the malicious traffic flow, the opponents' trajectories $\boldsymbol{Y}^{\text{OV}}$ and the ego vehicle's trajectory Y^{EV} are not independent. Therefore, it only makes sense to model $\boldsymbol{Y}^{\text{OV}}$ and Y^{EV} simultaneously and estimate the joint traffic distribution of safety-critical scenarios:

$$\mathbb{P}(Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}}|Coll = True, X). \tag{3}$$

Under some mild assumptions in Theorem 1, we can factorize Eq. (3) with the Bayesian formula.

Theorem 1. *Suppose that the EV's reaction depends on the future traffic unidirectionally, then we have* $\mathbb{P}(Y^{EV}, \boldsymbol{Y}^{OV}|Coll = True, X) \propto \mathbb{P}(\boldsymbol{Y}^{OV}|X)\mathbb{P}(Y^{EV}|\boldsymbol{Y}^{OV}, X)\mathbb{P}(Coll = True|Y^{EV}, \boldsymbol{Y}^{OV})$.

Proof. See the Appendix. □

Note that the safety-critical scenario generation objective of CAT, namely $\min_{f^{Adv}} J(\pi)$, is to magnify the probability of traffic collisions with the agent as possible. Thus, after the factorization, we can search the best responsive ${}^*\boldsymbol{Y}^{\text{OV}}$ through the marginal distribution given as:

$$\begin{aligned} & \max_{\boldsymbol{Y}^{\text{OV}}} \; \mathbb{P}(\boldsymbol{Y}^{\text{OV}}|Coll = True, X) \\ = \; & \max_{\boldsymbol{Y}^{\text{OV}}} \; \sum_{Y^{\text{EV}}} \mathbb{P}(Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}}|Coll = True, X) \\ = \; & \max_{\boldsymbol{Y}^{\text{OV}}} \underbrace{\mathbb{P}(\boldsymbol{Y}^{\text{OV}}|X)}_{\text{1st Term}} \sum_{Y^{\text{EV}}} \underbrace{\mathbb{P}(Y^{\text{EV}}|\boldsymbol{Y}^{\text{OV}}, X)}_{\text{2nd Term}} \underbrace{\mathbb{P}(Coll = True|Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}})}_{\text{3rd Term}}. \end{aligned} \tag{4}$$

It is beneficial to perform the above safety-critical traffic probability factorization since each term in Eq. (4) features a specific meaning and is tractable to handle. Each term is interpreted as follows:

i) **Traffic prior.** The 1st term is the standard motion prediction problem in which we can leverage arbitrary probabilistic traffic models [17, 36, 37, 38] to portray the multi-modal trajectory distribution. Taking the pre-trained model as the traffic prior enables the attack plausibility in complex scenarios without human specifications.

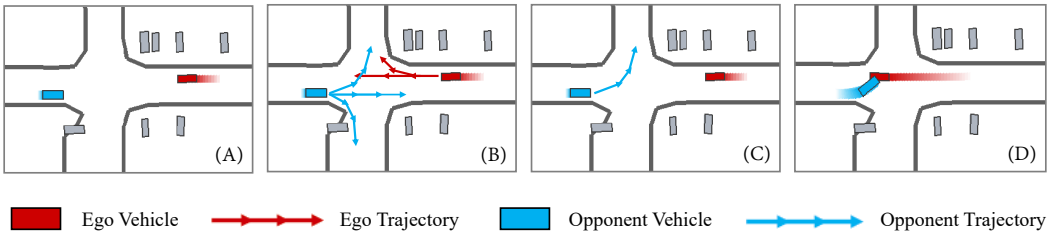

Figure 2: Illustration of Factorized Safety-Critical Resampling. (**A**) We initialize 1s traffic history with the dense map representation. (**B**) We then predict the traffic prior as well as the agent's reaction. (**C**) The most accident-prone trajectory of the opponent vehicle is selected. (**D**) The generated scene is thus expected to be safety-critical.

ii) **Ego estimation.** The 2nd term denotes the interactive ego trajectory yielding to the current state and upcoming traffic flow. The transition can be deterministic if the world model is learned or accessible under model-based settings [12]. As for the inference of real-world-compliant traffic flows, we can employ an interactive motion predictor [18] conditioned on known surrounding vehicles' trajectories to better reflects the ego compliance under risky interactions.

iii) **Collision likelihood.** The 3rd term reflects the likelihood of a collision in the compositional future, which can be treated as a typical binary classifier to fit [39].

As shown in Fig. 2, it is possible to approach the near-optimal adversarial trajectory via numerical optimization after each term is calculated.

3.3 Implementation Details

We summarize the overall implementation of the CAT framework for safe end-to-end driving in **Algorithm 1**. Recalling the training objective of CAT in Eq. (1), we need to perform iterative optimization of policy learning and adversarial environment generation synchronously in a closed loop. The policy optimization can be achieved by arbitrary end-to-end driving policy learning approaches, and we employ a vanilla RL algorithm.

Below, we focus on the adversarial environment generation, where we utilize the proposed factorized safety-critical resampling in Eq. (4). Note that we make a simplification in CAT by enforcing a single rival to launch the attack in each generated scene while simply maneuvering the other vehicles to avoid self-collisions. This is reasonable since most traffic accidents are caused by two traffic participants rather than involving multiple vehicles.

We first predict the traffic prior $\mathbb{P}(Y^{\mathrm{OV}}|X)$ using a pre-trained probabilistic traffic forecasting model $\mathcal{G}$. Considering the strong performance and the ease of sampling, we adopt DenseTNT [17], an anchor-free goal-based motion predictor, in this work. Specifically, we propose M possible candidates $\{(Y_i^{\mathrm{OV}}, P_i^{\mathrm{OV}})\}_{i=1}^{M}$ in parallel. The component $Y_{i,k}^{\mathrm{OV}}$ in the k-th time step consists of the predicted position and yaw of the opponent vehicle. The probability of the trajectory P_i^{OV} coincides with the probability of the corresponding destination goal.

We then tackle the ego estimation term $\mathbb{P}(Y^{\mathrm{EV}}|Y^{\mathrm{OV}}, X)$. Considering the non-stationary policy during training, we notice that the ego behavior does not necessarily match the logged behavior in the dataset. Consequently, directly utilizing the pre-trained traffic estimator derived from natural traffic flows [18] to provide ego trajectory probability has a severe bias. Alternatively, we record the latest N rollouts of EV in each scenario formed as $\{(Y_j^{\mathrm{EV}}, P_j^{\mathrm{EV}})\}_{j=1}^{N}$ and recompute the likelihood of visited state sequences deduced by the current policy π: $P_{j,k+1}^{\mathrm{EV}} = P_{j,k}^{\mathrm{EV}} \cdot \pi(a_k|s_k)$.

At last, we empirically estimate the collision likelihood $\mathbb{P}(Coll|Y^{\mathrm{EV}}, Y^{\mathrm{OV}})$. Given the specific compositional future of Y_j^{EV} and Y_i^{OV}, we compute the minimal distance between their bounding boxes in the following steps and set the collision likelihood as $P_{i,j}^{Coll} = \alpha^k$ if the closest gap is 0 at timestep k. Here, $\alpha \in (0, 1]$ is a heuristic decay factor to reflect the uncertainty of traffic models with the increasing prediction horizon.

Algorithm 1: Closed-loop Adversarial Training (CAT) for Safe End-to-End Driving.

Input: Initial driving policy π, Learning algorithm $\mathcal{T}$, Traffic Motion Predictor $\mathcal{G}$
Output: Robust driving policy π^*

Initialize the scenario pool $\mathcal{D} = \{X_1, X_2, ...X_{|\mathcal{D}|}\}$ from real-world datasets.
while π *is not converged* **do**
  Randomly sample normal traffic X from the scenario pool $\mathcal{D}$
  $\{(Y_i^{\text{OV}}, P_i^{\text{OV}})\}_{i=1}^{M} \sim \mathcal{G}(X)$ `// Compute the traffic prior.`
  for *i in* $1, 2, ..., M$ **do**
    for *j in* $1, 2, ..., N$ **do**
      $P_{ij}^{Coll} = \alpha^k \cdot \mathbb{I}[\text{BBox}(Y_{j,k}^{\text{EV}}) \cap \text{BBox}(Y_{i,k}^{\text{OV}}) \neq \emptyset | \ \exists k]$
    $P(Y_i^{\text{OV}}|Coll, X) = P_i^{\text{OV}} \sum_{j=1}^{N} P_j^{\text{EV}} P_{ij}^{Coll}$ `// Compute the posterior probability.`
  ${}^*Y^{\text{OV}} = \arg\max_{Y_i^{\text{OV}}} P(Y_i^{\text{OV}}|Coll, X)$ `// Select the best response.`
  obs = simulator.reset(X, ${}^*Y^{\text{OV}}$) `// Compose the adversarial environment.`
  for *t in* $1, 2, 3..., |T|$ **do**
    act $\sim \pi(\cdot|\text{obs})$
    obs = simulator.step(act) `// Policy execution.`
    $Y_{1:t}^{\text{EV}} = Y_{1:t-1}^{\text{EV}} \oplus Y_t^{\text{EV}}$
    $P_{1:t}^{\text{EV}} = P_{1:t-1}^{\text{EV}} \cdot \pi(\text{act}|\text{obs})$
  $\pi \leftarrow \mathcal{T}(\pi)$ `// Policy optimization.`
  $\{(Y_i^{\text{EV}}, P_i^{\text{EV}})\}_{i=1}^{N} = \{(Y_i^{\text{EV}}, P_i^{\text{EV}})\}_{i=2}^{N} \oplus (Y^{\text{EV}}, P^{\text{EV}})$ `// Update ego rollout queue.`

4 Experiments

4.1 Experiment Setup

We import 100 real-world traffic scenarios involving complex vehicle interactions from the Waymo Open Motion Dataset (WOMD) [4] as the raw data. Each scene in WOMD contains a traffic participant labeled as *Object of Interest*, which is also designated as the opponent vehicle (OV) in our experiments. All the experiments are conducted in MetaDrive [19], an open-source and lightweight AD simulator. The detailed hyper-parameter settings can be referred to the Appendix. Here, we point out some pivotal parameters. Each scene lasts 9s, in which we take the first 1s traffic history as X and manipulate the following 8s to generate the adversarial trajectory Y^{OV}. We set $M = 32$ as the number of OV trajectory candidates, $N = 5$ as the length of ego rollout queue during training and $\alpha = 0.99$ to penalize the uncertainty of motion forecasting.

4.2 Evaluation of Safety-critical Traffic Generation in CAT

The factorized safety-critical resampling is the crucial component of CAT to generate adversarial training environments. We provide qualitative and quantitative comparisons with the following baselines: **(A) Raw Data**: Replaying the recorded real-world traffic. **(B) M2I (adv)** [18]: The interactive traffic motion prediction is similar to our factorized formulation and thus can be modified as an adversarial scenario generator. **(C) STRIVE** [13]: The state-of-the-art safety-critical scenario generation methods performing gradient-based optimization on the latent code.

Qualitative analysis. In Fig. 3, we present 9 different types of safety-critical scenarios that CAT generates from raw scenes, according to the pre-crashed traffic categorized by the National Highway Traffic Safety Administration (NHTSA). It can be concluded that CAT is able to generate adversarial traffic given arbitrary real-world raw scenes. Meanwhile, the generated trajectories are in line with human driver behavior, even though we don't specify prior knowledge of that scene. In Fig. 4, we compare the generated adversarial traffic of the four methods on the same intersection. In the raw scene, the leading vehicle turns preferentially and does not cross the path of the ego vehicle. The opponent attempts to collide with the agent at the intersection through the safety-critical generation. However, M2I (adv) has a bias in estimating the reaction of the ego vehicle, which does not cause the expected accident. STRIVE finds the solution to enforce a crash, but it is still cumbersome to tweak the multinomial loss function to balance the goal of colliding as soon as possible and reasonable

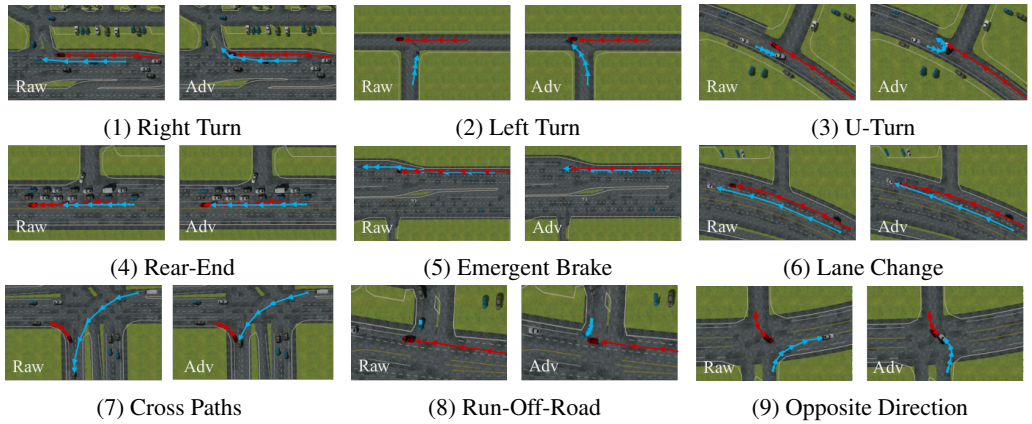

Figure 3: Qualitative results on the diversity of safety-critical scenarios generated by CAT. In each subfigure, the left and right are the raw scene and the adversarial counterpart. The ego and adversarial trajectories are highlighted with red and blue arrows, respectively.

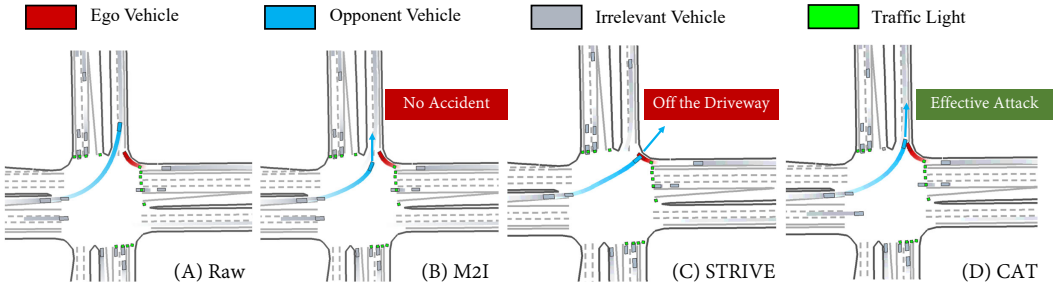

Figure 4: Qualitative results on the plausibility of safety-critical scenarios generated by CAT. The attack is regarded as effective only if leading traffic accidents are consistent with real-world events.

driving behavior, like keeping the vehicle in the driveway. By contrast, our factorized safety-critical resampling leverages the learned motion prior to regularize the opponent's trajectory, magnifying the traffic risk while preserving its plausibility. More visualization can be found in Appendix.

Quantitative analysis.

In Table 1, we conduct the comparative study mainly on two metrics. The first metric of interest is the attack success rate as the driving policies are responsive and even defensive to the traffic flow. We adopt three kinds of agents with fixed policies to validate: *(i) Replay Agent*: Replay the original trajectory of the ego vehicle logged in real-world data-set. *(ii) IDM Agent*: A heuristic controller well-adopted in AD tasks [40]. *(iii) Pre-trained Agent*: A pre-trained RL policy on WOMD. We find that M2I (adv) is insufficient for ego prediction and attacks less effectively especially against low-level policy, which is fatal for end-to-end driving. CAT collects ego rollouts to enhance the confidence of ego estimation during training ($N = 5$) and testing ($N = 1$) which significantly improves the attack success rate and is competitive with the SOTA method STRIVE. The second metric of interest is the time consumption per scene, which is non-negligible considering the large number of scenario iterations during training. We find that STRIVE generally requires 2-3 minutes to process a single scene due to its autoregression procedure on the raster map, which means it takes days to train the agent in a closed loop involving thousands of episodes. By contrast, our approach best balance the attack success rate and computational time compared and enjoys a privileged advantage in closed-loop adversarial training for end-to-end driving.

Table 1: Comparison of adversarial traffic generation algorithms on 100 scenes.

Methods	Attack Success Rate ↑			Per Scene Creating Time ↓
	Replay	IDM	Pretrained	
Raw Data	0%	34%	14%	/
M2I (adv)	47%	41%	19%	$\mathbf{0.41 \pm 0.03}$s
STRIVE	85%	82%	66%	153.10 ± 47.33s
CAT ($N = 1$)	91%	71%	62%	0.66 ± 0.09s
CAT ($N = 5$)	**91%**	**86%**	**69%**	3.34 ± 0.41s

Table 2: Performance of end-to-end driving policies with different training pipelines.

Metrics	No Adv	Heuristic	Open-loop	Closed-loop
Train Attack Num ↑	7546 ± 506	9881 ± 810	18541 ± 2172	$\mathbf{24997 \pm 2437}$
Test Crash Rate (Raw) ↓	$19.7\% \pm 1.37\%$	$16.8\% \pm 0.43\%$	$15.1\% \pm 1.45\%$	$\mathbf{11.2\% \pm 2.48\%}$
Test Crash Rate (Adv) ↓	$49.6\% \pm 2.11\%$	$41.4\% \pm 1.73\%$	$29.9\% \pm 2.08\%$	$\mathbf{20.0\% \pm 3.11\%}$

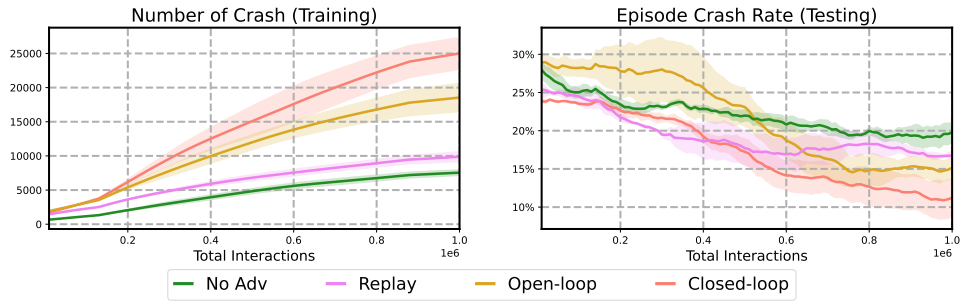

Figure 5: The learning curves with different training pipelines.

4.3 Evaluation of Closed-loop Adversarial Training in CAT

We show how CAT improves AI driving safety in accident-prone driving scenarios. We split the 100 scenes into 70 training and 30 testing scenarios. We train a TD3 [41] driving policy from scratch with 4 types of training pipelines: **(A) No Adv**: Remove the opponent vehicle. **(B) Replay**: Replay the human behaviors stored in the dataset. **(C) Open-loop**: Use CAT to manipulate the opponent trajectory against the log-replayed ego rollout, instead of the ego trajectory of RL agent. **(D) Closed-loop**: Use CAT to generate adversarial scenario dynamically against the learning agent. We evaluate the driving policies trained from different pipelines with three metrics. The first metric is the number of effective attacks occurred during adversarial training, describing the total number of collision with the surrounding vehicles. We also evaluate the crash rate, the ratio of the episodes that the ego vehicle crashes into others, on the hold-out testing scenarios with log-replay traffic (*Raw*) or with CAT generated traffic (*Adv*).

As shown in Table 2 and Fig. 5, we find that CAT substantially increases safety-critical events compared with other baselines during training, showing that CAT can generate challenging collision-prone scenarios. On the other hand, the agent trained with CAT demonstrate superior safety performance in testing time.

5 Conclusion and Discussion

In this paper, we propose the closed-loop adversarial training (CAT) framework for safe end-to-end driving. The crucial component of CAT is an efficient adversarial traffic generation technique. Empirical results demonstrate that CAT can provide realistic physical attacks during training and enhance AI driving safety in the test time.

Limitation: Following limitations wait to be addressed in future work: (i) we only consider adversarial vehicles in this work but the safety-critical behaviors of pedestrians and cyclists are also of importance for safe driving and yet to be done, it requires the access to a different motion forecasting model; (ii) Experiment on one hundred scenes cannot cover all the accident-prone situations, thus there are other possible failure modes in the resulting agent; (iii) we only investigate the RL-based driving policy but the adversarial scenarios should also benefit the human-in-the-loop imitation learning [16, 42].

Transferring to real-world driving: The proposed adversarial training method and the comparison with prior methods are evaluated in the simulation of one hundred complex traffic scenarios imported from real-world driving dataset [4]. Thus, the evaluation contains realistic and complex vehicle interactions and shows promise for transferring to real-world settings.

References

[1] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[2] J. Herman, J. Francis, S. Ganju, B. Chen, A. Koul, A. Gupta, A. Skabelkin, I. Zhukov, M. Kumskoy, and E. Nyberg. Learn-to-race: A multimodal control environment for autonomous racing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9793–9802, 2021.

[3] C. Xu, W. Ding, W. Lyu, Z. Liu, S. Wang, Y. He, H. Hu, D. Zhao, and B. Li. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles. *arXiv preprint arXiv:2206.09682*, 2022.

[4] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021.

[5] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

[6] F. M. Favarò, N. Nader, S. O. Eurich, M. Tripp, and N. Varadaraju. Examining accident reports involving autonomous vehicles in california. *PLoS one*, 12(9):e0184952, 2017.

[7] A. Sinha, S. Chand, V. Vu, H. Chen, and V. Dixit. Crash and disengagement data of autonomous vehicles on public roads in california. *Scientific data*, 8(1):298, 2021.

[8] B. Mirchevska, C. Pek, M. Werling, M. Althoff, and J. Boedecker. High-level decision making for safe and reasonable autonomous lane changing using reinforcement learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2156–2162. IEEE, 2018.

[9] D. Isele, A. Nakhaei, and K. Fujimura. Safe reinforcement learning on autonomous vehicles. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–6. IEEE, 2018.

[10] L. Wen, J. Duan, S. E. Li, S. Xu, and H. Peng. Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–7. IEEE, 2020.

[11] W. Ding, H. Lin, B. Li, and D. Zhao. Causalaf: Causal autoregressive flow for safety-critical driving scenario generation. In *Conference on Robot Learning*, pages 812–823. PMLR, 2023.

[12] N. Hanselmann, K. Renz, K. Chitta, A. Bhattacharyya, and A. Geiger. King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 335–352. Springer, 2022.

[13] D. Rempe, J. Philion, L. J. Guibas, S. Fidler, and O. 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.

[14] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

[15] Z. Zhu and H. Zhao. A survey of deep rl and il for autonomous driving policy learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):14043–14065, 2021.

[16] Z. Peng, Q. Li, C. Liu, and B. Zhou. Safe driving via expert guided policy optimization. In *Conference on Robot Learning*, pages 1554–1563. PMLR, 2022.

[17] J. Gu, C. Sun, and H. Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15303–15312, 2021.

[18] Q. Sun, X. Huang, J. Gu, B. C. Williams, and H. Zhao. M2i: From factored marginal trajectory prediction to interactive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6543–6552, 2022.

[19] 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*, 2022.

[20] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[21] Q. Zhang, S. Hu, J. Sun, Q. A. Chen, and Z. M. Mao. On adversarial robustness of trajectory prediction for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15159–15168, 2022.

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9909–9918, 2021.

[23] A. Boloor, X. He, C. Gill, Y. Vorobeychik, and X. Zhang. Simple physical adversarial examples against end-to-end autonomous driving models. In *2019 IEEE International Conference on Embedded Software and Systems (ICESS)*, pages 1–7. IEEE, 2019.

[24] Z. Kong, J. Guo, A. Li, and C. Liu. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14254–14263, 2020.

[25] X. Ma, K. Driggs-Campbell, and M. J. Kochenderfer. Improved robustness and safety for autonomous vehicle control with adversarial reinforcement learning. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1665–1671. IEEE, 2018.

[26] A. Wachi. Failure-scenario maker for rule-based agent using multi-agent adversarial reinforcement learning and its application to autonomous driving. *arXiv preprint arXiv:1903.10654*, 2019.

[27] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[28] L. Anzalone, P. Barra, S. Barra, A. Castiglione, and M. Nappi. An end-to-end curriculum learning approach for autonomous driving scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):19817–19826, 2022.

[29] R. Wang, J. Lehman, J. Clune, and K. O. Stanley. Poet: open-ended coevolution of environments and their optimized solutions. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 142–151, 2019.

[30] R. Wang, J. Lehman, A. Rawal, J. Zhi, Y. Li, J. Clune, and K. Stanley. Enhanced poet: Open-ended reinforcement learning through unbounded invention of learning challenges and their solutions. In *International Conference on Machine Learning*, pages 9940–9951. PMLR, 2020.

[31] Z. Zhong, Y. Tang, Y. Zhou, V. d. O. Neves, Y. Liu, and B. Ray. A survey on scenario-based testing for automated driving systems in high-fidelity simulation. *arXiv preprint arXiv:2112.00964*, 2021.

[32] S. Riedmaier, T. Ponn, D. Ludwig, B. Schick, and F. Diermeyer. Survey on scenario-based safety assessment of automated vehicles. *IEEE access*, 8:87456–87477, 2020.

[33] W. Ding, B. Chen, M. Xu, and D. Zhao. Learning to collide: An adaptive safety-critical scenarios generating method. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2243–2250. IEEE, 2020.

[34] 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*, 2023.

[35] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

[36] T. Gilles, S. Sabatini, D. Tsishkou, B. Stanciulescu, and F. Moutarde. Home: Heatmap output for future motion estimation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 500–507. IEEE, 2021.

[37] B. Varadarajan, A. Hefny, A. Srivastava, K. S. Refaat, N. Nayakanti, A. Cornman, K. Chen, B. Douillard, C. P. Lam, D. Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7814–7821. IEEE, 2022.

[38] S. Shi, L. Jiang, D. Dai, and B. Schiele. Motion transformer with global intention localization and local movement refinement. *arXiv preprint arXiv:2209.13508*, 2022.

[39] X. Wang, J. Liu, T. Qiu, C. Mu, C. Chen, and P. Zhou. A real-time collision prediction mechanism with deep learning for intelligent transportation system. *IEEE transactions on vehicular technology*, 69(9):9497–9508, 2020.

[40] M. Treiber, A. Hennecke, and D. Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2):1805, 2000.

[41] S. Fujimoto, H. Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[42] S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

A Proof of Theorem 1

Theorem. *Suppose that the EV's reaction depends on the future traffic unidirectionally, then we have* $\mathbb{P}(Y^{EV}, \boldsymbol{Y}^{OV}|Coll = True, X) \propto \mathbb{P}(\boldsymbol{Y}^{OV}|X)\mathbb{P}(Y^{EV}|\boldsymbol{Y}^{OV}, X_t)\mathbb{P}(Coll = True|Y^{EV}, \boldsymbol{Y}^{OV})$.

Proof. According to Bayes theorem, we have

$$\mathbb{P}(Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}}|Coll = True, X) \propto \mathbb{P}(Coll = True|Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}}, X)\mathbb{P}(Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}}, X) \tag{A.1}$$

Since $Coll$ merely depends on $Y^{\text{EV}}_{t:t+l}$ and $\boldsymbol{Y}^{\text{OV}}_{t:t+l}$, (A.1) is equivalent to

$$\mathbb{P}(Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}}|Coll = True, X) \propto \mathbb{P}(Coll = True|Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}})\mathbb{P}(Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}}, X) \tag{A.2}$$

Suppose that the AV's reaction depends on the future traffic unidirectionally; continuing with Bayes theorem, we have

$$\begin{aligned}
&\mathbb{P}(Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}}|Coll = True, X) \\
&\quad\propto\ \mathbb{P}(Coll = True|Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}})\mathbb{P}(Y^{\text{EV}}|\boldsymbol{Y}^{\text{OV}}, X)\mathbb{P}(\boldsymbol{Y}^{\text{OV}}, X) \\
&\quad\propto\ \mathbb{P}(Coll = True|Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}})\mathbb{P}(Y^{\text{EV}}|\boldsymbol{Y}^{\text{OV}}, X)\mathbb{P}(\boldsymbol{Y}^{\text{OV}}|X)\mathbb{P}(X)
\end{aligned} \tag{A.3}$$

Since the past state X is given, we can omit the last item $\mathbb{P}(X)$ in (A.3). Therefore, it holds that

$$\mathbb{P}(Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}}|Coll = True, X) \propto \mathbb{P}(\boldsymbol{Y}^{\text{OV}}|X)\mathbb{P}(Y^{\text{EV}}|\boldsymbol{Y}^{\text{OV}}, X)\mathbb{P}(Coll = True|Y^{\text{EV}}, \boldsymbol{Y}^{\text{OV}}) \tag{A.4}$$

The proof of Theorem 1 is completed. □

B Hyper-parameter Settings

Table 3: CAT

Hyper-parameter	Value
Scenario Horizon T	9s
History Horizon t	1s
# of OV candidates M	32
# of EV candidates N	5
Penalty Factor α	0.99
Policy Training Steps	10E6

Table 4: TD3

Hyper-parameter	Value
Discounted Factor γ	0.99
Train Batch Size	256
Critic Learning Rate	3E-4
Actor Learning Rate	3E-4
Policy Delay	2
Target Network τ	0.005

Table 5: DenseTNT and M2I

Hyper-parameter	Value
Train Batch size	256
Train Epoches	30
Sub Graph Depth	3
Global Graph Depth	1
NMS Threshold	7.2
Number of Mode	32

C Qualitative Results of Safety-critical Traffic Generation

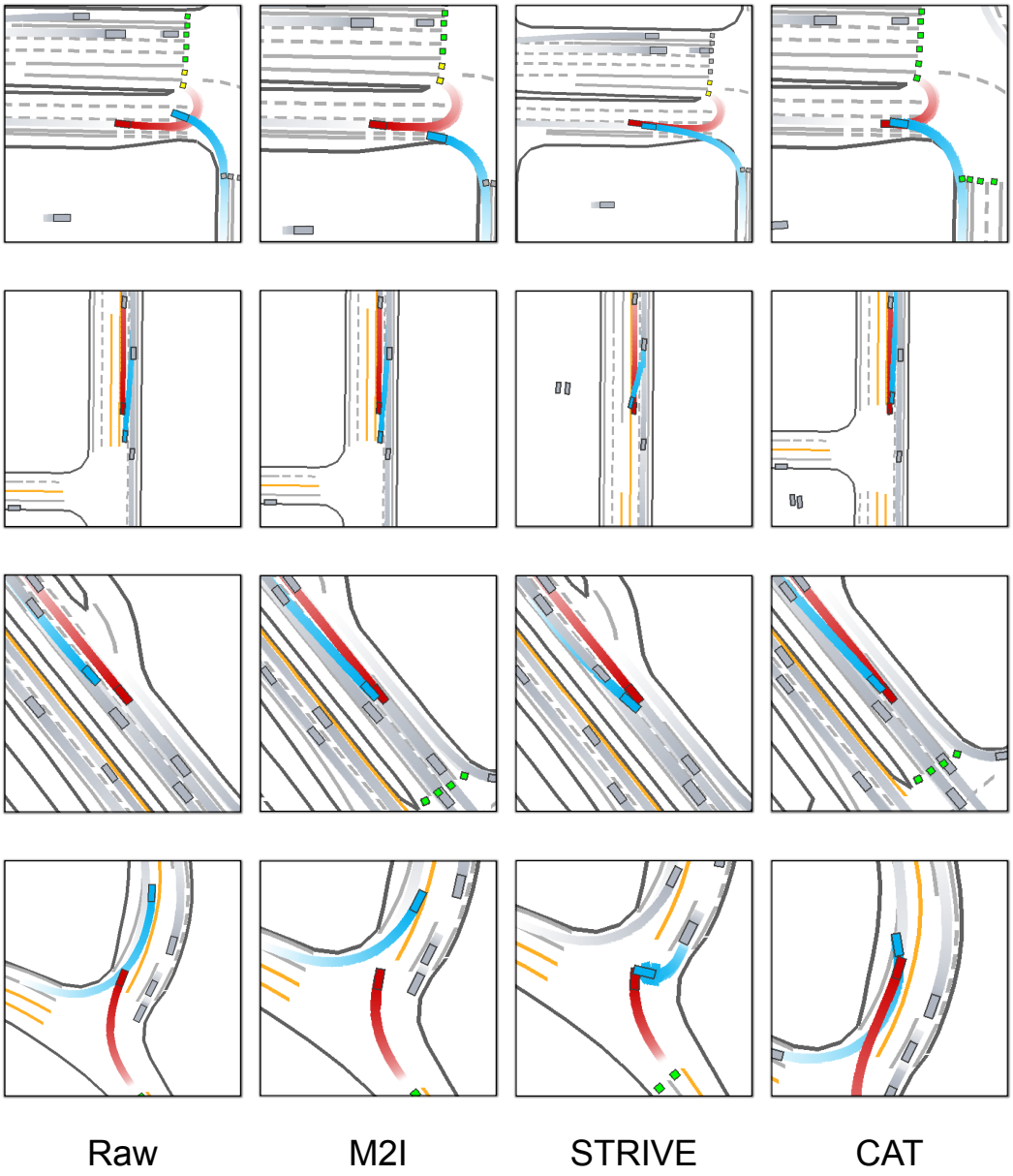

Figure 6: Comparing the different scenario generation methods. M2I and CAT both can determine the object of interest while STRIVE select the closest vehicle as the opponent. The red car is the ego vehicle and the blue car is the opponent vehicle.

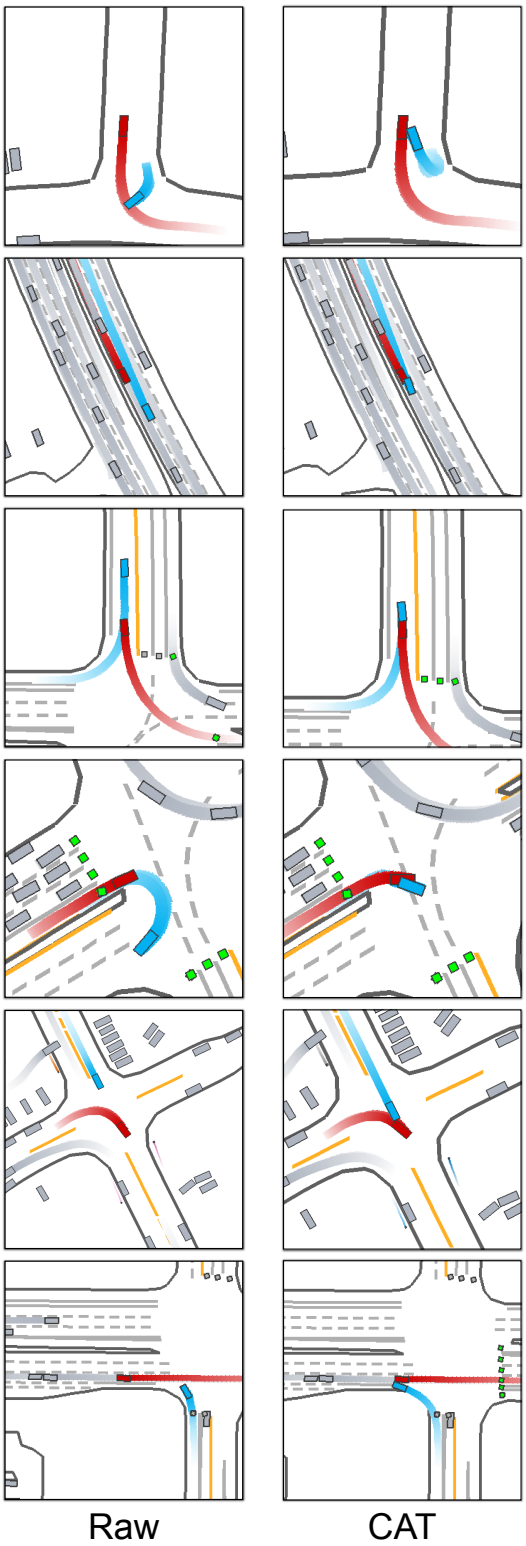

Figure 7: More comparison between the original scenarios in raw datasets and the safety-critical scenarios generated by our method. The red car is ego vehicle and the blue car is the opponent vehicle.