A. Few-Shot Prompts

A.1. Prompt for generating snippets

In this section, we present an example to illustrate our process for collecting snippets. This particular example focuses on generating more adversarial behaviors of surrounding vehicles within the base scenario of a *"Straight Obstacle"*. The specific prompt employed in this process is detailed in Table 4. Upon receiving responses from the GPT-4.0, we meticulously review each description-snippet pair for adversarial behavior. This review includes manual verification and correction of any errors, such as the use of unusable APIs, ensuring the accuracy and applicability of the snippets.

A.2. Prompt for extracting descriptions

In this section, we present the few-shot prompt, as detailed in Table 5, designed for extracting component-specific descriptions from a comprehensive input description. While for extracting the specified descriptions for each component from the text, we will employ regular expressions.

B. Detailed Scenario descriptions

In this section, we provide the detailed descriptions for the scenarios generated by our method for each base scenario, the full descriptions are shown in Table 6 and Table 7.

C. Additional Experiment Details and Results

We run all our experiments on one NVIDIA RTX A6000; we consistently use the online version of GPT-4 as the underlying LLM for manually monitoring the generation quality, and it can be easily adapted to use the API with version `gpt-4-1106-preview` for making it more automatic.

C.1. Detailed Metric

We adopt the metrics from Safebench [46], and we provide the intuition for each metric here:

- CR (collision rate): Evaluates the rate of collisions, reflecting the autonomous system's accident avoidance capability.
- RR (frequency of running red lights): Measures the frequency of running red lights, an essential aspect of traffic law adherence.
- SS (frequency of running stop signs): Assesses how often the vehicle fails to stop at stop signs, a key traffic rule compliance metric.
- OR (average distance driven out of road): Quantifies the average distance the vehicle deviates from its intended roadway, indicating lane discipline.
- RF (route following stability): Examines the stability with which the vehicle follows its planned route.
- Comp (average percentage of route completion): Represents the average percentage of the planned route successfully completed by the vehicle.
- TS (average time spent to complete the route): Evaluates the time efficiency of the vehicle in completing its assigned routes.
- ACC (average acceleration): Measures the average acceleration, providing insights into the smoothness of the vehicle's operation.
- YV (average yaw velocity): Assesses the average yaw velocity, indicating the vehicle's turning and handling characteristics.
- LI (frequency of lane invasion): Quantifies the frequency of lane invasions, a measure of lane-keeping accuracy.

While overall score (OS) is an aggregated metric that combines all metrics above to provide a comprehensive performance overview. We also adopt the same weights in Safebench [46].

C.2. Detailed Performance for Differently Trained ego vehicles

We provide a detailed assessment of adversarial performance in the scenes generated by different methods in each test ego vehicle (trained with SAC, PPO, TD3, respectively). Detailed statistics on the collision rate can be found in Table 8, while the overall score is reported in Table 9. The results demonstrate that our method exhibits greater generalizability across diverse training algorithms for ego vehicles, consistently achieving the best average performance.

C.3. Adversarial Finetuning Details

We utilize a pretrained surrogate SAC-trained model from Safebench [46] for our experiment. The fine-tuning details are as follows:

- We finetune the model with 500 epochs (one epoch just represents one simulation for one scene).
- The policy learning rate and Q-value learning rate are both set at 0.0001.
- The experience replay buffer is divided into two: one for non-collision cases with a size of $20,000$, and another for collision cases with a size of 200.
- The batch size is fixed at 512. In each batch, 80% of the samples are drawn from the non-collision buffer, while the remaining 20% are from the collision buffer. If collision cases are less than $20\% \times 512$, all available collision cases are selected.
- The entropy regularization coefficient, analogous to the inverse of the reward scale in the original SAC paper, is set at 0.01.
- The discount factor is 0.99, and the Q-ensemble critic comprises 2 models.

During evaluation, it was observed that prolonged training epochs led the model to adopt a stopping strategy. Therefore, to measure the performance accurately, we assess all checkpoints post 100 epochs at intervals of every 50 epochs. The optimal performance is reported based on the checkpoint exhibiting the lowest collision rate while maintaining a reasonable route completion rate (> 0.3), as adversarial events typically occur after 30% completion of the total route.

```
Your task is to provide descriptions of adversarial behaviors exhibited by surrounding objects
(pedestrians, cars, cyclists, motorcycles) that may lead to a collision with the ego vehicle on a
straight road. The behaviors should be safety-critical. Ensure that the provided code snippets adhere
to the Scenic API without creating new APIs. Here are some refined examples:

Behavior Description: A pedestrian suddenly starts crossing the road without looking.
Snippet:
behavior AdvBehavior():
    do CrossingBehavior(ego, globalParameters.ADV_SPEED, globalParameters.ADV_DISTANCE)
param ADV_SPEED = Range(0, 5)
param ADV_DISTANCE = Range(0, 20)

Behavior Description: A pedestrian steps onto the road right in front of the ego vehicle and stops.
Snippet:
behavior AdvBehavior():
    try:
        do CrossingBehavior(ego, globalParameters.ADV_SPEED, globalParameters.ADV_DISTANCE)
    interrupt when network.laneAt(self) is network.laneAt(ego):
        take SetWalkingSpeedAction(0)
param ADV_SPEED = Range(0, 5)
param ADV_DISTANCE = Range(0, 20)

Behavior Description: A car in an adjacent lane suddenly merges into the ego's lane.
Snippet:
behavior AdvBehavior():
    laneChangeCompleted = False
    try:
        do FollowLaneBehavior(target_speed=globalParameters.ADV_SPEED)
    interrupt when withinDistanceToAnyCars(self, globalParameters.ADV_DISTANCE) and not
    laneChangeCompleted:
        current_laneSection = network.laneSectionAt(self)
        leftLaneSec = current_laneSection._laneToLeft
        do LaneChangeBehavior(
            laneSectionToSwitch=leftLaneSec,
            target_speed=globalParameters.ADV_SPEED)
        laneChangeCompleted = True
param ADV_SPEED = Range(0, 10)
param ADV_DISTANCE = Range(0, 30)

Behavior Description: An adversarial cyclist sprints from behind a bus stop onto the road and stops in
front of the ego vehicle.
Snippet:
behavior AdvBehavior():
    do CrossingBehavior(ego, globalParameters.OPT_ADV_SPEED, globalParameters.OPT_ADV_DISTANCE) until
    (distance from self to network.laneAt(ego)) < globalParameters.OPT_STOP_DISTANCE
    while True:
        take SetWalkingSpeedAction(0)
param OPT_ADV_SPEED = Range(0, 10)
param OPT_ADV_DISTANCE = Range(0, 15)
param OPT_STOP_DISTANCE = Range(0, 1)

Now, based on these examples, your task is to provide additional Scenic code snippets that simulate
adversarial behaviors in traffic scenarios. Each snippet should be accompanied by a concise and clear
behavior description, similar to the provided examples. Your code must follow the existing Scenic
repository's API structure without introducing new APIs.
```

Table 4. A few-shot prompt for generating adversarial behavior in “Straight Obstacle” base scenario.

```
Your task is to decompose full descriptions of safety-critical scenarios into sub-descriptions for the
following distinct components:

Behavior: Describe the behavior of the adversarial object (you should also indicate the type of the
object like pedestrians, cars, cyclists, and motorcycles).
Geometry: Specify the road condition where the scenario occurs (e.g., straight road, three-way
intersection).
Spawn Position: Indicate the initial relative position of the adversarial object to the ego vehicle.

Here are refined examples:
Scenario: The ego vehicle is driving on a straight road, and the car in front brakes suddenly as the
ego approaches.
Behavior: The adversarial car suddenly brakes when the ego approaches.
Geometry: A straight road.
Spawn Position: The adversarial car is directly in front of the ego vehicle.

Scenario: The ego vehicle attempts a right turn at a four-way intersection, and an adversarial
pedestrian steps onto the road in front of it.
Behavior: The adversarial pedestrian deliberately steps onto the road right in front of the ego
vehicle.
Geometry: Lanes for turning right on a four-way intersection.
Spawn Position: The adversarial pedestrian is on the right front of the ego.

Scenario: The ego vehicle navigates around a parked car, and an oncoming car suddenly turns into its
path.
Behavior: The adversarial car suddenly turns into the ego's path without signaling.
Geometry: A two-lane road.
Spawn Position: The adversarial car is oncoming from the left lane of the ego.

Scenario: The ego vehicle is traveling along a straight road when a pedestrian, initially hidden behind
a bus stop on the sidewalk to the right, unexpectedly dashes onto the road directly in front of the ego
vehicle and comes to an abrupt stop.
Behavior: The adversarial pedestrian suddenly sprints from right, stopping abruptly in front of the ego
vehicle.
Geometry: A straight road.
Spawn Position: The adversarial pedestrian is initially stationed behind a bus stop on the right front.

Scenario: The ego vehicle is changing to the right lane when an emergency vehicle approaches rapidly
from behind.
Behavior: The adversarial car approaches rapidly from behind.
Geometry: A lane with right lanes on a straight road.
Spawn Position: The adversarial car is approaching from the rear on the right lane.

Scenario: The ego vehicle is turning right at an intersection, and a crossing car from left suddenly
brakes.
Behavior: The adversarial car suddenly brakes near the ego vehicle.
Geometry: Lanes for turning right on an intersection
Spawn Position: The adversarial car is crossing the intersection from the left.
```

Table 5. A few-shot prompt for extracting descriptions for different component given the full description of the safety-critical scenarios.

Table 6. The descriptions of scenarios generated in our method for the first four base scenarios.

Base Scenario	Description of Scenarios
Straight Obstacle	1. The ego vehicle is driving on a straight road; the adversarial pedestrian suddenly crosses the road from the right front and suddenly stops in front of the ego. 2. The ego vehicle is driving on a straight road; the adversarial pedestrian stands behind a bus stop on the right front, then suddenly sprints out onto the road in front of the ego vehicle and stops. 3. The ego vehicle is driving on a straight road; the adversarial pedestrian appears from a driveway on the left and suddenly stop and walk diagonally. 4. The ego vehicle is driving on a straight road; the adversarial pedestrian suddenly appears from behind a parked car on the right front and suddenly stop. 5. The ego vehicle is driving on a straight road; the adversarial pedestrian is hidden behind a vending machine on the right front, and abruptly dashes out onto the road, and stops directly in the path of the ego vehicle.
Turning Obstacle	1. The ego vehicle is turning left at an intersection; the adversarial motorcyclist on the right front pretends to cross the road but brakes abruptly at the edge of the road, causing confusion. 2. The ego vehicle is turning left at an intersection; the adversarial pedestrian on the opposite sidewalk suddenly crosses the road from the right front and stops in the middle of the intersection. 3. The ego vehicle is turning right at an intersection; the adversarial pedestrian on the left front suddenly crosses the road and stops in the middle of the intersection, blocking the ego vehicle's path. 4. The ego vehicle is turning left at an intersection; the adversarial cyclist on the left front suddenly stops in the middle of the intersection and dismounts, obstructing the ego vehicle's path. 5. The ego vehicle is turning right at an intersection; the adversarial motorcyclist on the opposite sidewalk abruptly crosses the road from the right front and comes to a halt in the center of the intersection.
Lane Changing	1. The ego vehicle is attempting to change lanes to avoid a slow-moving leading vehicle; the adversarial car in the target lane suddenly merges into the ego vehicle's original lane, blocking the ego from returning to its initial position. 2. The ego vehicle is changing to the right lane; the adversarial car is driving parallel to the ego and blocking its path. 3. The ego vehicle is performing a lane change to evade a slow-moving vehicle; the adversarial car in the target lane on the right front suddenly brakes, causing the ego vehicle to react quickly to avoid a collision. 4. The ego vehicle is preparing to change lanes to evade a slow-moving leading vehicle; the adversarial car in the target lane starts weaving between lanes, making it difficult for the ego to predict its position and safely execute the lane change. 5. The ego vehicle is attempting to change lanes to avoid a slow-moving leading vehicle; the adversarial car in the target lane suddenly slows down, matching the speed of the leading vehicle, and effectively blocks the ego from completing the lane change.
Vehicle Passing	1. The ego approaches a parked car that is blocking its lane and must use the opposite lane to bypass the vehicle, cautiously monitoring oncoming traffic, and suddenly encounters a jaywalking pedestrian, requiring the ego to quickly assess the situation and respond appropriately to avoid a collision. 2. The ego encounters a parked car blocking its lane and must use the opposite lane to bypass the vehicle, carefully assessing the situation and yielding to oncoming traffic, when an oncoming motorcyclist swerves into the lane unexpectedly, necessitating the ego to brake or maneuver to avoid a potential accident. 3. The ego approaches a parked car obstructing its lane and must use the opposite lane to go around when an oncoming car suddenly turns into the ego's path without signaling, requiring the ego to react quickly and take evasive action to prevent a collision. 4. The ego encounters a parked car blocking its lane and must use the opposite lane to bypass the vehicle when an oncoming car suddenly accelerates, closing the gap for the ego to safely return to its lane, necessitating the ego to quickly decide whether to accelerate or brake to avoid a collision. 5. The ego encounters a parked car blocking its lane and must use the opposite lane to bypass the vehicle when an oncoming pedestrian enters the lane without warning and suddenly stops, necessitating the ego to brake sharply or steer to avoid hitting the pedestrian.

Table 7. The descriptions of scenarios generated in our method for the last four base scenarios.

Base Scenario	**Description of Scenarios**
Red-light Running	1. The ego is driving straight through an intersection when a crossing vehicle runs the red light and unexpectedly accelerates, forcing the ego to quickly reassess the situation and perform a collision avoidance maneuver. 2. The ego vehicle is moving straight through the intersection; the adversarial vehicle, initially on the left front, runs the red light and makes an abrupt right turn, forcing the ego vehicle to perform a collision avoidance maneuver. 3. The ego vehicle is moving straight through the intersection; the adversarial vehicle, initially on the left front, runs the red light and makes an abrupt left turn, forcing the ego vehicle to perform a collision avoidance maneuver. 4. The ego moves straight at an intersection when a crossing vehicle runs the red light from the right and brakes abruptly, causing the ego to rapidly adapt its trajectory and perform a collision avoidance maneuver. 5. The ego vehicle is going straight through the intersection; the adversarial vehicle approaches from the left front and cuts off the ego vehicle.
Unprotected Left-turn	1. The ego starts an unprotected left turn at an intersection while yielding to an oncoming car when the oncoming car's throttle malfunctions, leading to an unexpected acceleration and forcing the ego to quickly modify its turning path to avoid a collision. 2. The ego attempts an unprotected left turn at an intersection while yielding to an oncoming car when the oncoming car suddenly brakes, necessitating the ego to rapidly reassess the situation and adjust its turn. 3. The ego vehicle is making an unprotected left turn; the adversarial vehicle approaches the intersection at a normal speed but then suddenly attempts to make a last-second right turn. 4. The ego attempts an unprotected left turn at an intersection while yielding to an oncoming car when the oncoming car veers erratically. 5. The ego commences an unprotected left turn at an intersection while yielding to an oncoming car when the adversarial car, coming from the right, blocks multiple lanes by driving extremely slowly, forcing the ego vehicle to change lanes.
Right-turn	1. The ego is performing a right turn at an intersection when the crossing car suddenly speeds up, entering the intersection and causing the ego to brake abruptly to avoid a collision. 2. The ego vehicle is turning right; the adversarial car (positioned ahead on the right) blocks the lane by braking suddenly. 3. The ego vehicle is turning right; the adversarial car (positioned ahead on the right) reverses abruptly. 4. The ego vehicle is turning right; the adversarial car (positioned behind on the right) suddenly accelerates and then decelerates. 5. The ego vehicle is turning right; the adversarial vehicle enters the intersection from the left side, swerving to the right suddenly.
Crossing Negotiation	1. The ego vehicle is approaching the intersection needs crossing negotiation; the adversarial car (on the left) suddenly accelerates and enters the intersection first and suddenly stops. 2. The ego vehicle is approaching the intersection needs crossing negotiation; the adversarial car (on the right) suddenly accelerates and enters the intersection first and suddenly stops. 3. The ego vehicle is entering the intersection needs crossing negotiation; the adversarial vehicle comes from the opposite direction and turns left and stops, causing a near collision with the ego vehicle. 4. The ego vehicle is entering the intersection needs crossing negotiation; the adversarial vehicle comes from the right and turns left and stops, causing a near collision with the ego vehicle. 5. The ego vehicle is entering the intersection needs crossing negotiation; the adversarial car, coming from the right, blocks multiple lanes by driving extremely slowly, forcing the ego vehicle to change lanes.

Table 8. **Collision Rate (CR) Performance Across Different Models**. This table presents the detailed analysis of the *collision rate* (CR) for various test ego vehicles, each trained with distinct RL algorithms. We showcase the mean CR for each model, demonstrating how they perform in the selected scenes under the same base scenario. Bold values denote the best performance for each scenario. Algorithms include: LC: Learning-to-collide, AS: AdvSim, CS: Carla Scenario Generator, AT: Adversarial Trajectory Optimization. Higher values of CR (↑) is preferable here.

Model	**Algo.**	**Base Traffic Scenarios**								**Avg.**
		Straight Obstacle	Turning Obstacle	Lane Changing	Vehicle Passing	Red-light Running	Unprotected Left-turn	Right-turn	Crossing Negotiation	
SAC (4D)	LC	0.40	0.11	0.60	0.80	0.92	0.86	0.70	0.75	0.642
	AS	0.53	0.40	0.75	**0.90**	0.62	0.85	0.21	0.53	0.599
	CS	0.53	0.68	0.67	**0.90**	0.76	0.90	0.69	0.68	0.725
	AT	0.73	0.48	0.77	**0.90**	**1.00**	**0.97**	0.76	**0.90**	0.814
	ChatScene	**0.94**	**0.73**	**0.92**	0.81	0.70	0.88	**0.84**	0.78	**0.825**
PPO (4D)	LC	0.09	0.11	**1.00**	**1.00**	0.22	0.20	0.28	0.00	0.363
	AS	0.39	0.19	**1.00**	**1.00**	0.67	**0.61**	0.62	**0.92**	0.675
	CS	0.22	**0.61**	**1.00**	**1.00**	0.47	0.37	0.52	0.44	0.579
	AT	0.10	0.11	0.98	0.87	0.13	0.21	0.03	0.05	0.310
	ChatScene	**0.87**	**0.61**	0.97	0.98	**0.89**	0.52	**0.74**	0.91	**0.811**
TD3 (4D)	LC	0.42	0.06	**1.00**	0.70	**1.00**	**1.00**	0.79	**1.00**	0.746
	AS	0.60	0.39	0.83	0.70	0.41	0.65	0.03	0.26	0.484
	CS	0.61	0.53	**1.00**	0.70	0.67	0.80	0.83	0.67	0.726
	AT	0.67	0.35	0.59	0.70	**1.00**	0.85	**0.99**	0.90	0.756
	ChatScene	**0.87**	**0.75**	0.96	**0.99**	0.77	0.86	0.75	0.90	**0.856**

Table 9. **Overall Score (OS) Performance Across Different Models**. This table presents the detailed analysis of the *overall score* (OS) for various test ego vehicles, each trained with distinct RL algorithms. We showcase the mean OS for each model, demonstrating how they perform in the selected scenes under the same base scenario. Bold values denote the best performance for each scenario. Algorithms include: LC: Learning-to-collide, AS: AdvSim, CS: Carla Scenario Generator, AT: Adversarial Trajectory Optimization. Lower values of OS (↓) is preferable here.

Model	**Algo.**	**Base Traffic Scenarios**								**Avg.**
		Straight Obstacle	Turning Obstacle	Lane Changing	Vehicle Passing	Red-light Running	Unprotected Left-turn	Right-turn	Crossing Negotiation	
SAC (4D)	LC	0.716	0.824	0.617	0.515	0.491	0.521	0.497	0.500	0.585
	AS	0.663	0.673	0.552	**0.466**	0.650	0.538	0.746	0.617	0.613
	CS	0.661	0.532	0.569	**0.466**	0.578	0.508	0.503	0.539	0.544
	AT	0.565	0.633	0.546	**0.466**	**0.462**	**0.475**	0.461	**0.423**	0.504
	ChatScene	**0.450**	**0.505**	**0.451**	0.489	0.582	0.492	**0.426**	0.461	**0.482**
PPO (4D)	LC	0.858	0.823	0.457	0.445	0.850	0.858	0.702	0.887	0.735
	AS	0.726	0.780	0.457	0.444	0.617	**0.646**	0.535	**0.408**	0.577
	CS	0.806	0.570	0.468	0.444	0.722	0.771	0.582	0.656	0.627
	AT	0.853	0.819	**0.439**	0.487	0.896	0.852	0.826	0.861	0.754
	ChatScene	**0.497**	**0.578**	0.444	**0.421**	**0.503**	0.705	**0.504**	0.417	**0.509**
TD3 (4D)	LC	0.708	0.843	0.442	0.561	**0.461**	**0.466**	0.447	0.378	0.538
	AS	0.631	0.668	0.511	0.561	0.757	0.638	0.834	0.755	0.669
	CS	0.627	0.599	0.430	0.561	0.622	0.559	0.430	0.542	0.546
	AT	0.587	0.689	0.629	0.561	0.462	0.534	**0.348**	0.423	0.529
	ChatScene	**0.462**	**0.483**	**0.407**	**0.410**	0.527	0.483	0.493	**0.385**	**0.456**