IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

SEAL: Towards Safe Autonomous Driving via Skill-Enabled Adversary Learning for Closed-Loop Scenario Generation

Benjamin Stoler[1] Ingrid Navarro[1] Jonathan Francis[1,2] Jean Oh[1]

***Abstract*—Verification and validation of autonomous driving (AD) systems and components is of increasing importance, as such technology increases in real-world prevalence. Safety-critical scenario generation is a key approach to robustify AD policies through closed-loop training. However, existing approaches for scenario generation rely on simplistic objectives, resulting in overly-aggressive or non-reactive adversarial behaviors. To generate diverse adversarial yet realistic scenarios, we propose SEAL, a scenario perturbation approach that leverages learned objective functions and adversarial, human-like skills. SEAL-perturbed scenarios are more realistic than SOTA baselines, leading to improved ego task success across real-world, in-distribution, and out-of-distribution scenarios, of more than 20%. To facilitate future research, we release our code and tools: https://navars.xyz/seal/**

***Index Terms*—Intelligent Transportation Systems, Autonomous Vehicle Navigation, Performance Evaluation and Benchmarking**

I. Introduction

With the growing deployment of autonomous driving (AD) technologies in real-world settings, ensuring the safety of such systems has only increased in importance and public concern [3], [4]. As AD verification and validation approaches continue to evolve, scenario-based testing via datasets and simulation has emerged as a core methodology, where alternatives such as on-road testing via a sufficiently large number of miles driven can be prohibitively expensive, risky, and infeasible [5], [6]. While validation of system behavior under normal operating circumstances is valuable, testing AD behavior under *safety-critical* and other corner-case circumstances is vital for Safety of the Intended Functionality (SOTIF) standards [7]–[9].

Scenarios are often curated in the form of large datasets of real-world recorded driving traces, providing a basis for assessing human behaviors and for training machine learning models [10]–[12]. AD subsystems are then asked to perform tasks such as forecasting the future motion of various road users or controlling the behavior of certain vehicles in a simulated reconstruction [13]–[16]. However, the presence of critical scenarios in collected datasets is exceedingly low, a problem identified as the "curse-of-rarity" in autonomous driving [17]–[19]. Thus, programmatically *generating* safety-critical scenarios is necessary. To ensure that generated scenarios retain realistic properties, it is appealing to perturb the behavior of one or more agents in a principled way, rather than using first principles to painstakingly assemble a scenario from scratch [1], [20]–[22]. In this setting, one agent is referred to as the *ego* agent, while the modified background traffic participants are *adversary* agent(s), who attempt to attack the ego in some way.

State-of-the-art (SOTA) approaches in perturbation-based scenario generation have coupled a dynamic scenario generation framework with an ego control policy being trained with closed-loop objectives [1], [23], [24], in contrast with previous less-efficient staged approaches [25], [26]. These approaches can still be sub-optimal, however, in that they can struggle to provide *useful* training stimuli to a closed-loop agent. In particular, we identify three key issues in recent SOTAs: 1) they have a limited view of safety-criticality, e.g., focusing only on inducing collisions or near-misses; 2) they lack reactivity to an ego agent's behavior diversity; and 3) their optimization objectives tend to maximize "unrealistic" and overly-aggressive adversarial behavior, limiting their usefulness for balanced model training.

Therefore, in this paper, we propose and evaluate a method for **S**kill-**E**nabled **A**dversary **L**earning (SEAL), which yields significantly improved downstream ego behavior, in closed-loop training with safety-critical scenario generation. Our method addresses the identified limitations in prior art by introducing two novel components, as shown in Figure 1. First, we introduce a learned objective function to *anticipate* how a reactive ego agent will respond to a candidate adversarial agent behavior. We quantify both collision closeness and induced ego behavior deviation, thus providing a broadened understanding of safety criticality. Second, we develop a skill-enabled, reactive adversary policy; in particular, inspired by human cognition, we leverage a hierarchical framework that is akin to how humans operate vehicles [27] and we create an adversarial prior that selects human-like *skill primitives* to increase criticality while maintaining realism.

Furthermore, we argue that safety-critical scenario generation should be evaluated based on behavior realism and usefulness for ego policy improvement, not just induced criticality. Prior work often assesses ego policies on generated scenarios where safety-critical behavior remains effectively *in-distribution* with respect to training data and heuristic

Manuscript received: February, 17, 2025; Revised May, 30, 2025; Accepted June, 28, 2025.

This paper was recommended for publication by Editor Ashis Banerjee upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the Korean Ministry of Trade, Industry, and Energy (MOTIE; grant #P0026022), and by the Korea Institute of Advancement of Technology (KIAT), through the International Cooperative R&D program (#P0019782): Embedded AI Based fully autonomous driving software and Maas technology development.

[1]Benjamin Stoler, Ingrid Navarro, and Jean Oh are with the School of Computer Science, Carnegie Mellon University {bstoler, ingridn, jeanoh}@cs.cmu.edu

[2]Jonathan Francis is with the Bosch Center for Artificial Intelligence and also with the School of Computer Science, Carnegie Mellon University jon.francis@us.bosch.com

Digital Object Identifier (DOI): see top of this page.

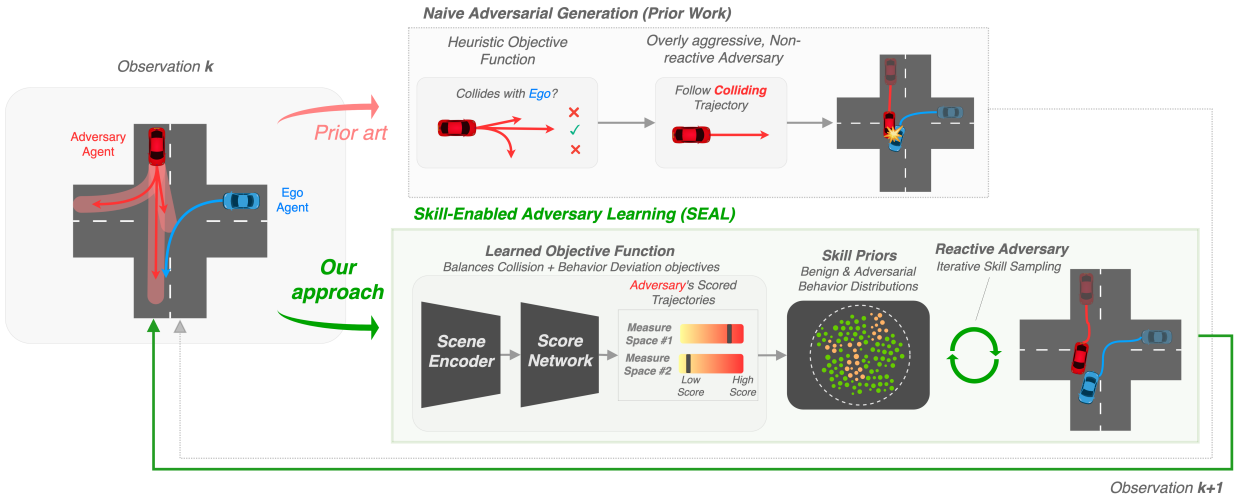


Fig. 1: An overview of SEAL. Our scenario generation approach leverages a learned objective function and an adversarial skill-based, reactive policy, for improved adversary realism and more effective closed-loop training, leading to safer autonomous driving agents, compared to previous approaches such as CAT [1] and GOOSE [2].

perturbations [1], [28]. To address this, we build on recent scenario characterization work, SafeShift [18], to identify real (non-generated) but safety-relevant scenarios, enabling a more realistic, out-of-distribution evaluation. While in-distribution performance is informative, real-world performance on challenging scenarios is ultimately most important.

In summary, our paper comprises three main contributions:

1) We propose two novel techniques for safety-critical perturbation: (i) a learned objective function to select candidate trajectories; and (ii) an adversarial skill-based, reactive policy for more realism in adversary behavior.
2) We design an improved evaluation setting for closed-loop training, utilizing real-world safety-relevant scenarios in contrast to just in-distribution generated scenarios.
3) We provide results on several key experiments, showing an increase of more than 20% in ego task success rate over SOTA baselines, across scenarios generated closed-loop by our proposed framework, across scenarios generated closed-loop by previous SOTA baseline frameworks, *and* across real-world safety-relevant scenarios.

II. Related Work

A. Scenario Generation in Autonomous Driving

Approaches for generating scenarios that reproduce the distribution of *normal* driving behavior have been extensively explored. Some methods ensure the diversity of generated traffic behavior [29], [30], while others aim for controllability through rule-based or language-driven specifications [31]–[33]. However, due to the rarity of safety-critical events in recorded data [17]–[19], other approaches have focused on directly generating corner-case scenarios by injecting adversarial behaviors. Earlier works in safety-critical scenario generation relied on gradient-based optimization approaches with access to vehicle dynamics [25], [28], [34], a limitation in model-free settings. Other methods, such as diffusion-based approaches [26], [35], are compute-intensive and impractical to be used in a closed-loop manner. Efficient methods like CAT [1] and GOOSE [2], which leverage trajectory prediction priors and reinforcement learning (RL) respectively, prioritize simple collision objectives and are non-reactive to the ego agent. Similarly, [36] employs reactive adversaries but focuses only on collisions for criticality and defines realism via proximity to ground-truth trajectories, making it sensitive to distribution shifts. In contrast, our approach efficiently generates reactive, nuanced adversarial behavior across multiple axes of criticality, providing a stronger closed-loop training signal.

B. Robust Training and Evaluation in Autonomous Driving

Several techniques for robustifying AD policies against safety-critical and out-of-distribution scenarios have been explored. Formal methods, such as Hamilton-Jacobi (HJ) reachability, have been utilized in various driving tasks, but struggle with dimension scaling [37], [38]. Similarly, domain randomization has been used as a form of data augmentation (e.g., randomizing vehicle control parameters [39] or scenario initial states [40]) but requires excessive sampling to cover a sufficient domain size. Thus, adaptive stress testing [17], [41] and adversarial training have been increasingly used, either as a fine-tuning scheme [25], [26] or in a fully closed-loop training pipeline [1], [34], providing continuous feedback to an ego agent. However, these approaches still tend to optimize for naive collision objectives alone.

Evaluation of robust training and scenario generation approaches is crucial. Many works evaluate generated scenarios against fixed rule-based or replay ego planners alone [2], [14], [25], [35], [42], offering limited insights into the efficacy of adversarial agents against more sophisticated ego agents. Additionally, adversarially-trained ego policies are often tested on scenarios perturbed by the same adversarial method used in training [1], [26], [28], [34], leading to in-distribution evaluations. Conversely, we focus on out-of-distribution evaluation of well-trained, reactive ego policies, in both adversarial

scenarios perturbed by *other* SOTA approaches, as well as real safety-relevant scenarios.

Out-of-distribution evaluation has been well-explored in AD trajectory prediction [16], [18], [43], [44], but these approaches often aim to characterize an entire scenario without focusing on a single ego driver or identifying a specific adversary. In AD control tasks, some prior work has explored out-of-distribution settings, such as CARNOVEL [45], [46], which tests unseen scenario types like roundabouts. Additionally, Lu et al. [47] evaluate across real-world scenarios of various difficulty levels, but do not hold out the hardest scenarios during training. Our approach thus addresses this gap by offering a more comprehensive and rigorous evaluation, across a wide set of adversarial and real-world scenarios.

III. Preliminaries

In this section, we define relevant notation and task definitions used in the rest of this paper. Let $(x,\ y)^{(t)}$ represent the location of an agent (i.e., vehicle, pedestrian, or cyclist) in the ground plane at some given time t. We then define an agent's trajectory as the ordered set $X = ((x\ , y)^{(t)} \mid t \in \{1, 2, ..., T\})$ over T timesteps at some fixed time delta.

Base Scenario: We define a base scenario, $\mathbf{S}$, as the tuple $(\mathbf{X},\ \mathbf{M},\ \texttt{ego},\ \texttt{adv})$, with $\mathbf{X} = \{X_i \mid i \in \{1, 2, ..., N\}\}$ consisting of the set of all agent trajectories observed, where X_i denotes the trajectory of an agent with the ID of i, and N is the total number of agents. All relevant map and scenario meta information (such as lane connectivity, traffic light locations, etc.) is given as $\mathbf{M}$. Finally, `ego` and `adv` refer respectively to the agent IDs of the ego vehicle (to be controlled in simulation) and the adversarial vehicle (to be perturbed to induce criticality).

Scenario Perturbation Task: For this task, K re-simulations of a base scenario $\mathbf{S}$ are performed as episodes, where agents start from the same state as the base scenario and follow a behavior prescribed by some policy (i.e., a reactive policy or predefined trajectory), which may be different than their original trajectory. Let $\widetilde{\mathbf{X}}^{(k)}$ represent the observed trajectories in the k-th re-simulation of $\mathbf{S}$. The perturbation-based safety critical scenario generation task is thus assigning behaviors to roll-out for all non-ego agents, conditioned on the base scenario $\mathbf{S}$ and K previous episodes, $\{\widetilde{\mathbf{X}}^{(k)} \mid k \in \{1, 2, ..., K\}\}$, such that the resulting $\widetilde{\mathbf{X}}^{(K+1)}$ satisfies some specified desired properties of criticality. Importantly, we treat the ego agent's behavior as a *black box*: while we are able to observe previous behavior as $\widetilde{\mathbf{X}}^{K}_{\text{ego}}$, we have no access to the model or any privileged information on `ego`'s decision-making process. In practice, during training, we maintain a queue of the most recent K perturbation roll-outs for each base scenario; during evaluation, we instead run K sequential perturbation–simulation steps and use the final roll-out as the adversarial scenario.

IV. Approach: Skill-Enabled Adversary Learning for Scenario Generation

To increase scenario criticality while preserving realism, we propose the **S**kill-**E**nabled **A**dversary **L**earning (SEAL) approach for perturbation-based scenario generation. Similar to CAT [1], SEAL employs a probabilistic trajectory predictor π_{gen} to sample candidate adversary futures conditioned on $\mathbf{S}$. However, directly executing these samples has three main limitations: 1) it measures criticality only via collisions, ignoring, e.g., forced ego hard braking or swerving; 2) it prevents reactivity to ego decisions; and 3) it often generates non-human-like behavior, driving straight at the ego with no avoidance. SEAL addresses these with a learned objective for flexible trajectory selection and an adversarial skill policy for more human-like and reactive behavior.

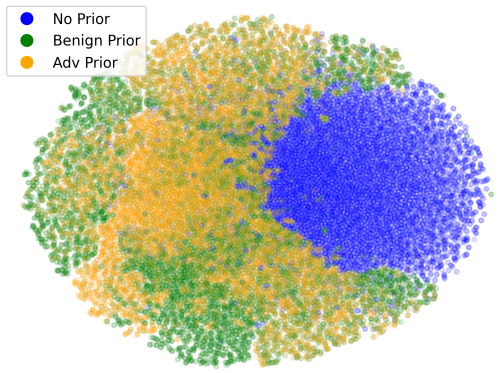


Fig. 2: Skill space visualized with t-SNE [48]. Benign and adversarial priors map to several regions representing useful, human-like skills, with meaningful separation and overlap.

A. Learned Objective Function

Many previous works rely on heuristic approaches to select the *best* trajectory from a candidate set to be assigned to the behavior of the adversary agent, $\widetilde{X}^{(K+1)}_{\text{adv}}$. For instance, CAT [1] compares bounding box overlaps across the previous K episodes in all candidate routes, selecting the one which collides with the most previous ego roll-outs at the earliest time step or is closest to a collision, otherwise. We instead aim to select among candidate trajectories in a more flexible way that captures both closeness to collision as well as likelihood of anticipated ego behavior deviation (e.g., causing the ego to swerve or execute a hard-brake maneuver).

We frame the problem as a supervised regression task. First, we build a dataset of simulated outcomes, where we roll out and observe all trajectory pairs of ego and adversarial agents, $(\widetilde{X}^{(K+1)}_{\text{ego}},\ \widetilde{X}^{(K+1)}_{\text{adv}})$. To keep ego behavior as a black-box in downstream closed-loop training, we have the ego follow a reactive heuristic policy during this stage. We then obtain ground truth values from the collected demonstrations, using the following scoring functions, similar to measure functions used in prior work [18], [21]:

$$f_{\text{coll}} = \exp\left(-\frac{1}{b}\min_t \left|\left|\widetilde{X}^{(k),t}_{\text{ego}} - \widetilde{X}^{(k),t}_{\text{adv}}\right|\right|_2\right) \tag{1}$$

$$f_{\text{diff}} = 1 - \exp\left(-\frac{1}{b}\sum_t \left|\left|\widetilde{X}^{(k-1),t}_{\text{ego}} - \widetilde{X}^{(k),t}_{\text{ego}}\right|\right|_2\right), \tag{2}$$

where $b \in \mathbb{R}$ is a hyperparameter controlling sensitivity to distance values. Both Equation (1) and Equation (2) map to $[0, 1]$, where 1 indicates maximal criticality and 0 indicates

minimal. Equation (1) captures collision closeness between the ego and adversary over a given roll-out, while Equation (2) captures ego behavior difference between two episodes. However, instead of only assessing past episodes, we propose to *predict* these measures for a roll-out yet to happen by training a neural network, π_{score} (detailed in Section IV-C). This π_{score} network aims to predict f_{coll} and f_{diff} conditioned on a previous $\widetilde{X}_{\text{ego}}^{(k)}$ and the proposed $\widetilde{X}_{\text{adv}}^{(K+1)}$. The final score for ranking candidate trajectories is the sum of the predicted f_{coll} and f_{diff} values from π_{score}, averaged over the K previous ego roll-outs.By using π_{score} in place of additional heuristic simulation, we enable scoring to be conditioned on actual ego policy roll-outs and substantially reduce runtime overhead.

B. Adversarial Skill Learning

We design a reactive policy π_{adv}, to guide the adversary's behavior, unlike recent works [1], [2], where the selected adversary follows a predefined trajectory. This adversarial policy observes and acts in a closed-loop simulator alongside the ego policy. In this context, skill-based hierarchical policies are appealing approaches as they capture maneuvers at a higher abstraction, compared to the low-level actions of a simulator, corresponding more closely to how humans operate vehicles [27].

We build upon prior work, ReSkill [49], which utilizes expert demonstrations to extract paired observation and action sequences as state-conditioned "skills" which are then embedded using a Variational AutoEncoder (VAE). Additionally, a state-conditioned prior network is trained to map from a state to a useful location in the VAE's latent space to be decoded into a reconstructed skill for the agent to follow.

In our work, we separate the demonstrated skills into adversarial (i.e., those ending in a collision or near-miss) and benign skills (i.e., those avoiding a collision while staying on road). We use a sliding-window partitioning scheme that excludes segments starting within twice the skill horizon before an out-of-road event, and labels as adversarial those within the same window before a collision. This reflects the intuition that not only the final skill but also preceding behavior contributes to unsafe outcomes. We then train two prior networks in parallel with a shared-skill VAE: benign skills flow through a "benign" prior while adversarial skills flow through an analogous "adversarial" prior. In this way, the adversarial agent policy, π_{adv}, leverages the adversarial prior to select skills likely to lead to safety-critical outcomes. Furthermore, because each prior is implemented as a real-valued non-volume preserving transformation trained on observed data (as in ReSkill), sampled noise vectors bijectively correspond to plausible, in-distribution behaviors. Figure 2 visualizes the learned skill spaces over uniformly sampled states; regions of overlap correspond to skills which may be useful to both an adversarial and benign agent (e.g., lane-keeping, smooth kinematics, etc.) while distinct regions correspond to skills only useful for that particular agent (e.g., for an adversary: cutting-off another vehicle, hard-braking in a dangerous way, etc.).

To integrate this skill module with the trajectory generation and ranking discussed in Section IV-A, we first select the highest ranking candidate trajectory, $\widetilde{X}_{\text{adv}}^{(K+1)}$. We derive goals and subgoals from this selected trajectory to provide to π_{adv} as navigation information. Skills are then executed in a hierarchical manner as in [49]: at the start of the episode or when a skill has completed, a new skill is selected based on the current observation and adversarial prior. The agent then decodes that skill, in a closed-loop manner, into raw actions. To further increase safety-criticality, the adversary initially exactly follows $\widetilde{X}_{\text{adv}}^{(K+1)}$ before switching to this adversarial skill policy at a fixed offset before the anticipated point of maximal collision risk.

C. SEAL Implementation Details

For training and validating both the learned objective function and skill spaces, we leverage the well-established Waymo Open Motion Dataset (WOMD) [10] dataset, as well as a subset of scenarios therein labeled by Waymo as containing interacting agents. A further subset of 500 of these scenarios has been used by prior work, and we henceforth refer to this set as `WOMD-Normal` [1], [50]. We split these scenarios into 400 training and 100 evaluation examples.

For π_{gen}, we utilize a pre-trained DenseTNT [13] trajectory prediction model, as used by CAT. π_{gen} takes as input the first one second of $\mathbf{X}$, as well as the static meta information $\mathbf{M}$, and produces 32 candidate eight-second future adversary paths. We use the MetaDrive simulator [51] and its included IDM policy [52] as the heuristic reactive agent to collect imperfect demonstration data, described and utilized in both Section IV-A and Section IV-B. For data augmentation, **all** agents in the scenario follow the IDM policy and produce useful demonstrations, rather than collecting examples from solely the ego. We extract subgoals from each trajectory using MetaDrive's default waypoint logic, placing checkpoints every 8 meters as navigation input for π_{adv}.

We implement π_{score} as a VectorNet-style polyline encoder [53], followed by a multilayer perceptron decoder to the predicted values of f_{coll} and f_{diff}. We use an MSE loss objective on the sum of the two values, ensuring equal weight to both predicted measures. For π_{adv}, we leverage the skill embedding framework from [49], with identical architectures and loss functions across our two parallel prior networks. We empirically set the hyperparameter b in Equation (1) and Equation (2) to 8, use a skill time horizon of 10 steps, and fix K to 5 (consistent with CAT).

V. Experimental Setup

We leverage SEAL to generate scenarios for two primary purposes: providing data augmentation during closed-loop training of reinforcement learning (RL) agent policies, and providing a means of evaluating such agents' capabilities.

A. Policy Training

For closed-loop training of an ego agent policy, we leverage the `WOMD-Normal` set along with the MetaDrive simulator [51], described in Section IV-C. Then, we follow the curriculum training approach proposed by CAT [1], where a

TABLE I: Ego performance on adversarially-perturbed (a, b, c) and unmodified, real-world (d, e) scenarios. `WOMD-Normal` are WOMD [10] scenarios with basic interactive agents labeled by Waymo; `WOMD-SafeShift-Hard` refers to SafeShift-mined [18] real scenarios in WOMD. Adversarially-perturbed scenarios use `WOMD-Normal` as base scenarios, in both training and evaluation settings. *Higher* success rates and lower crash and out of road rates are better. Ego realism scores are shown in (f), averaged over settings (a–e) using Wasserstein distance (WD); lower is better.

(a) WOMD-Normal, GOOSE-Gen [2]

Training	Success	Crash	Out of Road
None (Replay)	0.59 (0.00)	0.41 (0.00)	0.00 (0.00)
No Adv	0.41 (0.06)	0.37 (0.02)	0.23 (0.04)
GOOSE	0.37 (0.07)	0.35 (0.09)	0.30 (0.17)
CAT	0.35 (0.03)	0.27 (0.02)	0.39 (0.06)
SEAL	**0.44** (0.04)	0.27 (0.00)	0.27 (0.00)

(b) WOMD-Normal, CAT-Gen [1]

Training	Success	Crash	Out of Road
None (Replay)	0.18 (0.00)	0.82 (0.00)	0.00 (0.00)
No Adv	0.32 (0.01)	0.46 (0.02)	0.22 (0.01)
GOOSE	0.25 (0.10)	0.47 (0.02)	0.31 (0.04)
CAT	0.32 (0.03)	0.32 (0.03)	0.40 (0.00)
SEAL	**0.42** (0.02)	0.32 (0.04)	0.24 (0.02)

(c) WOMD-Normal, **SEAL**-Gen

Training	Success	Crash	Out of Road
None (Replay)	0.32 (0.00)	0.68 (0.00)	0.00 (0.00)
No Adv	0.33 (0.03)	0.50 (0.05)	0.21 (0.04)
GOOSE	0.26 (0.08)	0.46 (0.00)	0.27 (0.06)
CAT	0.31 (0.00)	0.34 (0.04)	0.36 (0.02)
SEAL	**0.38** (0.04)	0.36 (0.01)	0.25 (0.06)

(d) WOMD-Normal, Real Scenarios

Training	Success	Crash	Out of Road
None (Replay)	1.00 (0.00)	0.00 (0.00)	0.00 (0.00)
No Adv	0.48 (0.02)	0.21 (0.01)	0.28 (0.04)
GOOSE	0.44 (0.13)	0.23 (0.03)	0.34 (0.10)
CAT	0.50 (0.02)	0.15 (0.06)	0.36 (0.10)
SEAL	**0.59** (0.01)	0.15 (0.00)	0.27 (0.01)

(e) WOMD-SafeShift-Hard, Real Scenarios

Training	Success	Crash	Out of Road
None (Replay)	0.97 (0.00)	0.01 (0.00)	0.02 (0.00)
No Adv	0.28 (0.05)	0.38 (0.05)	0.33 (0.02)
GOOSE	0.19 (0.04)	0.42 (0.06)	0.36 (0.04)
CAT	0.24 (0.00)	0.38 (0.03)	0.37 (0.05)
SEAL	**0.38** (0.02)	0.29 (0.02)	0.33 (0.04)

(f) Aggregate Realism

Training	Yaw WD	Acc WD	Road WD
None (Replay)	0.014	0.269	0.004
No Adv	0.147	**3.041**	**0.252**
GOOSE	0.152	3.052	0.312
CAT	0.154	3.050	0.374
SEAL	**0.146**	3.074	0.270

TABLE II: Scenario generation quality; results are averaged over all tested ego models. WD measures are Wasserstein distances over adversary behavior; a lower value indicates greater realism. Lower collision velocities (m/s) and head-on rates are better. A *lower* ego `Success` is better, as this table assesses safety-critical effectiveness.

Eval Scenario Type	Ego Success (↓)	Realism WD (↓)	Yaw WD (↓)	Acc WD (↓)	Road WD (↓)	Coll. Vel. (↓)	Head-On (↓)	Head-On (Severe) (↓)
WOMD-Normal, Real Scenarios	60.0%	0.056	0.120	0.020	0.027	2.849	06.7%	04.2%
WOMD-SafeShift-Hard, Real Scenarios	41.3%	0.069	0.116	0.044	0.043	2.206	00.0%	00.0%
WOMD-Normal, GOOSE-Gen	43.0%	0.401	0.124	0.601	0.482	4.744	13.2%	11.7%
WOMD-Normal, CAT-Gen	**29.6%**	0.167	0.123	0.305	0.074	4.136	**07.1%**	05.9%
WOMD-Normal, **SEAL**-Gen	31.9%	**0.108**	**0.121**	**0.157**	**0.049**	**2.950**	09.2%	**03.6%**

random base scenario $\mathbf{S}$ from the train split is selected and has a random chance of being perturbed; this perturbation chance increases throughout the training process. Agents observe the environment via simulated LiDAR returns and navigation information based on their original destination in $\mathbf{X}$. Agents act on the environment with normalized steering and acceleration forces as $\mathbf{a}$; the ego and adversarial agents follow either a policy or a predefined trajectory, while all other agents follow their original trajectory in $\mathbf{X}$.

We utilize ReSkill [49] as our underlying RL algorithm, a recent SOTA approach in hierarchical RL. We use our skill space built in Section IV-B, utilizing the benign prior rather than the adversarial one. The low-level action learned by the ReSkill agent is a remediating $\Delta\mathbf{a}$ adjustment to the action decoded based on the current skill and state pair, $\mathbf{a}'$, while the high-level action selects the noise vector to be passed to the prior. Thus, the action sent to the environment is $\mathbf{a} = \mathbf{a}' + \Delta\mathbf{a}$. Actions are performed at a 10Hz rate, and all agents are trained for one million timesteps in total, empirically sufficient for consistent policy convergence.

B. Evaluation Settings

Many previous works evaluate agent performance, in-distribution, on a held-out subset of their *own* generated scenarios [1], [26], [28], [34]. For additional comprehensiveness, we propose to utilize a recent scenario characterization approach, SafeShift [18], for identifying real-world safety-relevant base scenarios, denoted as `WOMD-SafeShift-Hard`. We start by identifying scenarios containing interacting agents labeled by Waymo. We then apply SafeShift's hierarchical scoring to these agents and select scenarios where the *interacting* agents have trajectory scores in the top 20th percentile across WOMD, randomly sampling 100 scenarios therein. The ego and adversary agents are assigned to the interacting agents with the higher and lower trajectory score, respectively.

We baseline SEAL against two recent SOTA safety critical scenario generation approaches, that can be utilized in a closed-loop manner: CAT [1] and GOOSE [2]. CAT heuristically chooses a trajectory from π_{gen} to apply to the adversarial agent; we use the same π_{gen} function for both CAT and SEAL, for fairness. GOOSE learns to iteratively modify control points of a NURBS [54] curve fit to the original adversary's trajectory, observing the outcome of each roll-out. We train GOOSE against the MetaDrive IDM agent using the `WOMD-Normal` training set and GOOSE's "deceleration" task goal—induce a collision while maintaining kinematic feasibility. For consistency, we limit the number of GOOSE policy steps (i.e., observed roll-outs) to $K = 5$.

C. Metrics

Within MetaDrive, episodes are terminated when the ego agent either arrives safely at its goal (`Success`), collides with another agent (`Crash`), or violates an off-road constraint (i.e., crosses a road edge or yellow median; `Out of Road`). As such, we report these corresponding rates as the key metrics for ego performance, following prior work [1].

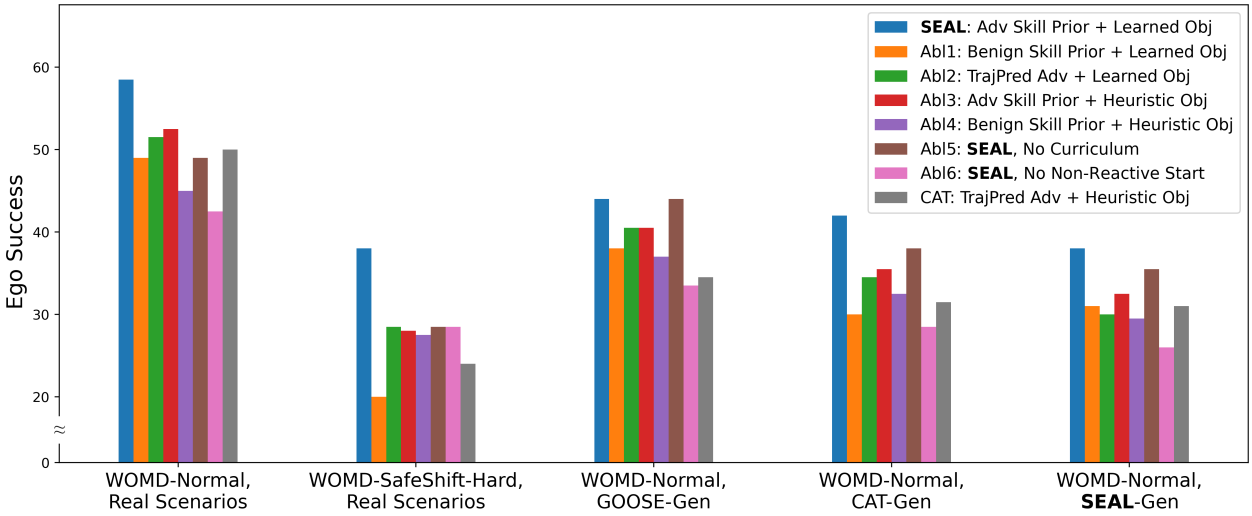
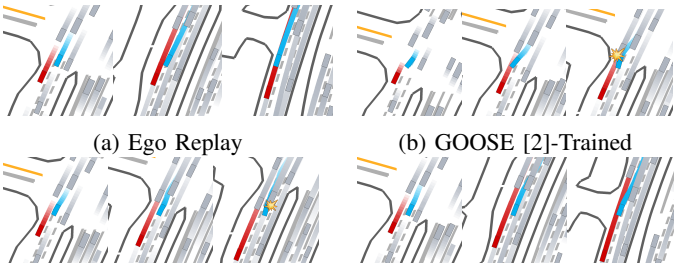


Fig. 3: Ablation study on SEAL scenario generation training pipelines. Our full approach with learned objectives (Section IV-A) and adversarial skill policies (Section IV-B) produces the strongest downstream agents, across all five evaluation settings.



(a) Ego Replay (b) GOOSE [2]-Trained

(c) CAT [1]-Trained (d) **SEAL**-Trained

Fig. 4: Qualitative examples of **driving policies**; the **blue** ego is a learned agent while the **red** adversary is fixed. (a) shows the original human trajectory from `WOMD-SafeShift-Hard`, while (b), (c), and (d) show ego behaviors learned in different pipelines.

For evaluating generated scenario quality, we examine the induced ego `Success` rate, across all tested ego methods. We derive a realism metric based on distributional measures, following MixSim [14] and other prior work [15], [33]. In particular, we utilize the Wasserstein distance (WD) over adversarial "profiles"—normalized histograms constructed from the adversary's yaw rates, acceleration values, and out-of-road rates. All WD values are computed via comparison to profiles derived from the original X_{adv} in $\mathbf{S}$, which we average to compute an overall `Realism` meta-metric. We also report relative collision velocities along the contact normal, as in [25], along with head-on collision rates and severe head-on rates (where severity is defined as collision velocity exceeding 5 m/s, thereby filtering out low-speed, glancing incidents).

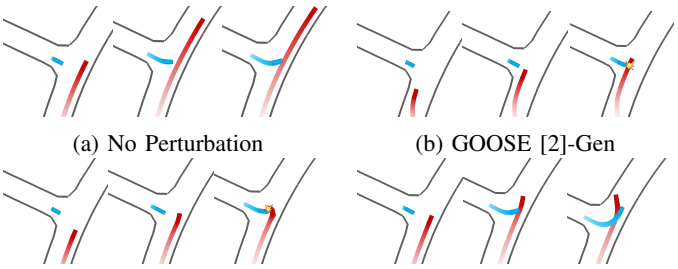
VI. Results

We report the median and interquartile range (IQR) over four seeds, for greater statistical robustness. These statistical summaries are computed independently over each metric, so `Success`, `Crash`, and `Out of Road` may not sum to 100%. We also evaluate a non-reactive ego replay policy (`Replay`), which rolls out the original X_{ego} trajectory, as well as a ReSkill [49] agent trained without any adversarial scenario generation (`No Adv`). Note that due to re-simulation limitations, `Replay` in `WOMD-Normal` and `WOMD-SafeShift-Hard` may have a nonzero failure rate.

Downstream Performance. Our closed-loop training results are summarized in Table I. SEAL-trained policies average a $\mathbf{21.5\%}$ **increase** in `Success` rate relative to the top baseline in each setting, achieving a strong balance between `Crash` and `Out of Road` rates. While a baseline-trained policy may have slightly better performance on one failure type, it is achieved by sacrificing performance against the other. Compared to GOOSE and CAT, SEAL training yields more realistic yaw and road compliance but less realistic acceleration, indicating stronger braking and more disciplined in-lane maneuvering to manage criticality. Despite high kinematic realism, `No Adv` egos crash frequently due to lack of experience in challenging scenarios and resulting poor reactivity.

We highlight qualitative examples of ego behavior in Figure 4, showcasing how different training regimes influence the execution of the same *benign* skills, in a scenario drawn from the `WOMD-SafeShift-Hard` set of real, safety-relevant scenarios. While all ego policies operate within the same offline-learned skill space, their online adaptation differs across training methods. The `Replay` ego depicts the ground-truth human trajectory, which merges safely into the right lane. The GOOSE-trained ego initiates the merge too early and fails to recover in time, resulting in a collision. The CAT-trained ego begins to merge later but hesitates under pressure, slows down, and is rear-ended. In contrast, the SEAL-trained ego merges with a sufficient gap while accommodating a close-following tail vehicle, resulting in a smooth and safe maneuver. These differences highlight how SEAL's more realistic and nuanced adversarial training scenarios better prepare ego policies to navigate challenging interactions effectively.

Scenario Generation Quality. To directly assess scenario generation quality, we aggregate metrics in Table II, averaged over all ego methods. Although CAT scenarios induce a lower ego `Success` rate and raw head-on rate than SEAL scenarios,



(a) No Perturbation (b) GOOSE [2]-Gen

(c) CAT [1]-Gen (d) **SEAL**-Gen

Fig. 5: Qualitative examples of **scenario perturbation**; the **blue** ego follows a fixed replay policy while the **red** adversary is modified. (a) shows the original `WOMD-Normal` scenario, while (b), (c), and (d) show perturbations generated by GOOSE, CAT, and SEAL, respectively.

SEAL scenarios exhibit the highest `Realism` among scenario generation approaches, a **35.3% improvement**, contributing to SEAL-trained policies' superior downstream performance.Furthermore, SEAL's collision velocities and severe head-on rates are far lower than baseline approaches.

We also showcase qualitative examples of the tested scenario generation approaches in Figure 5, using a fixed ego `Replay` policy to isolate differences in adversary behavior. CAT and GOOSE both produce aggressive trajectories that lead to collisions: CAT stops and turns directly into the ego, while GOOSE swerves across the lane and slows in the ego's path to force a t-bone. In contrast, the SEAL adversary exhibits more nuanced behavior, slowing down to let the ego catch up, moving away at the last moment to induce a near-miss, and thus demonstrating interesting *adversarial* skill behavior.

Ablation Studies. To further investigate how different components of SEAL affect downstream training, we perform extensive ablation studies shown in Figure 3, as well as comparing against CAT as it is a slightly stronger baseline than GOOSE. We study the effect of our learned objective function by comparing it to the heuristic, bounding box overlap approach used by CAT (`Learned Obj` and `Heuristic Obj`, respectively). Similarly, we compare our adversarial skill policy (`Adv Skill Prior`) with a benign prior variant (`Benign Skill Prior`) and a predefined trajectory following policy (`TrajPred Adv`). We also compare SEAL against two additional ablations: one trained *without* curriculum, and another *without* the initial non-reactive start. Our full SEAL approach performs best across all settings; both the learned objective function and adversarial skill policy are essential, while the curriculum and non-reactive start further improve performance.

VII. Conclusion

As autonomous driving (AD) systems advance, ensuring safety remains essential. While recent safety-critical scenario generation techniques show promise, they often lack the realism, reactivity, and nuance needed to provide strong training signals for closed-loop agents. We thus introduced Skill-Enabled Adversary Learning (SEAL) as a perturbation-based safety-critical scenario generation approach, combining a learned objective function and an adversarial skill policy. In all test settings—across both real-world challenging scenarios and generated scenarios by SEAL and other SOTA methods—SEAL-trained policies achieved significantly higher success rates, with a more than 20% relative increase. Upon deeper analysis, SEAL-generated scenarios contain less aggressive but more realistic adversaries, helping to explain the observed ego agent improvements. We argue that realism metrics, downstream task utility, and out-of-distribution evaluation settings are vital in assessing adversarially-perturbed scenarios.

While SEAL is quite effective, further improvements are still possible. Incorporating finer-grained metrics into the objective function could enable more adaptive and controllable generation beyond safety criticality alone. Additionally, enhancing realism metrics to reflect human decision-making at the skill-level could provide deeper insights into scenario quality. We encourage future work to explore these topics.

Acknowledgment

This work was partially performed during Benjamin Stoler's internship at Stack AV; the authors thank Stack for their mentorship. The authors additionally thank Evan Lohn for many valuable discussions throughout the design process.

References

[1] L. Zhang, Z. Peng, Q. Li, and B. Zhou, "Cat: Closed-loop adversarial training for safe end-to-end driving," in *Conference on Robot Learning*. PMLR, 2023, pp. 2357–2372.

[2] J. Ransiek, J. Plaum, J. Langner, and E. Sax, "Goose: Goal-conditioned reinforcement learning for safety-critical scenario generation," in *2024 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2024.

[3] M. Cummings, "Assessing readiness of self-driving vehicles," in *The 103rd Transportation Research Board (TRB) Annual Meeting, Washington, DC*, 2024.

[4] Y. Xing, H. Zhou, X. Han, M. Zhang, and J. Lu, "What influences vulnerable road users' perceptions of autonomous vehicles? a comparative analysis of the 2017 and 2019 pittsburgh surveys," *Technological Forecasting and Social Change*, vol. 176, p. 121454, 2022.

[5] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 182–193, 2016.

[6] G. Lou, Y. Deng, X. Zheng, M. Zhang, and T. Zhang, "Testing of autonomous driving systems: where are we and where should we go?" in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2022, pp. 31–43.

[7] *Road Vehicles—Safety of the Intended Functionality*, International Organization for Standardization Std., 2022, standard ISO 21448:2022. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:77490:en

[8] Q. Song, E. Engström, and P. Runeson, "Industry practices for challenging autonomous driving systems with critical scenarios," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 4, pp. 1–35, 2024.

[9] X. Zhang, J. Tao, K. Tan, M. Törngren, J. M. G. Sánchez, M. R. Ramli, X. Tao, M. Gyllenhammar, F. Wotawa, N. Mohan *et al.*, "Finding critical scenarios for automated driving systems: A systematic mapping study," *IEEE Transactions on Software Engineering*, vol. 49, no. 3, pp. 991–1026, 2022.

[10] 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*, 2021, pp. 9710–9719.

[11] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes *et al.*, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," *arXiv preprint arXiv:2301.00493*, 2023.

[12] 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.

[13] 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*, 2021, pp. 15 303–15 312.

[14] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9622–9631.

[15] N. Montali, J. Lambert, P. Mougin, A. Kuefler, N. Rhinehart, M. Li, C. Gulino, T. Emrich, Z. Yang, S. Whiteson *et al.*, "The waymo open sim agents challenge," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[16] S. H. Park, G. Lee, J. Seo, M. Bhat, M. Kang, J. Francis, A. Jadhav, P. P. Liang, and L.-P. Morency, "Diverse and admissible trajectory forecasting through multimodal context understanding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 282–298.

[17] S. Feng, H. Sun, X. Yan, H. Zhu, Z. Zou, S. Shen, and H. X. Liu, "Dense reinforcement learning for safety validation of autonomous vehicles," *Nature*, vol. 615, no. 7953, pp. 620–627, 2023.

[18] B. Stoler, I. Navarro, M. Jana, S. Hwang, J. Francis, and J. Oh, "Safeshift: Safety-informed distribution shifts for robust trajectory prediction in autonomous driving," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 1179–1186.

[19] H. X. Liu and S. Feng, "Curse of rarity for autonomous vehicles," *nature communications*, vol. 15, no. 1, p. 4808, 2024.

[20] 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.

[21] P. Huang, W. Ding, J. Francis, B. Chen, and D. Zhao, "Cadre: Controllable and diverse generation of safety-critical driving scenarios using real-world trajectories," *arXiv preprint arXiv:2403.13208*, 2024.

[22] Y. Cao, C. Xiao, A. Anandkumar, D. Xu, and M. Pavone, "Advdo: Realistic adversarial attacks for trajectory prediction," in *European Conference on Computer Vision*. Springer, 2022, pp. 36–52.

[23] X. Yang, L. Wen, Y. Ma, J. Mei, X. Li, T. Wei, W. Lei, D. Fu, P. Cai, M. Dou *et al.*, "Drivearena: A closed-loop generative simulation platform for autonomous driving," *arXiv preprint arXiv:2408.00415*, 2024.

[24] H. Tian, K. Reddy, Y. Feng, M. Quddus, Y. Demiris, and P. Angeloudis, "Enhancing autonomous vehicle training with language model integration and critical scenario generation," *arXiv preprint arXiv:2404.08570*, 2024.

[25] 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*, 2022, pp. 17 305–17 315.

[26] 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.

[27] N. Medeiros-Ward, J. M. Cooper, and D. L. Strayer, "Hierarchical control and driving." *Journal of experimental psychology: General*, vol. 143, no. 3, p. 953, 2014.

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

[29] D. Xu, Y. Chen, B. Ivanovic, and M. Pavone, "Bits: Bi-level imitation for traffic simulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2929–2936.

[30] S. Suo, S. Regalado, S. Casas, and R. Urtasun, "Trafficsim: Learning to simulate realistic multi-agent behaviors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 400–10 409.

[31] J. Lu, K. Wong, C. Zhang, S. Suo, and R. Urtasun, "Scenecontrol: Diffusion for controllable traffic scene generation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 16 908–16 914.

[32] 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 *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3560–3566.

[33] 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 *Conference on Robot Learning*. PMLR, 2023, pp. 144–177.

[34] 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*, 2021, pp. 9909–9918.

[35] W.-J. Chang, F. Pittaluga, M. Tomizuka, W. Zhan, and M. Chandraker, "Safe-sim: Safety-critical closed-loop traffic simulation with diffusion-controllable adversaries," 2024. [Online]. Available: https://arxiv.org/abs/2401.00391

[36] C. Zhang, S. Biswas, K. Wong, K. Fallah, L. Zhang, D. Chen, S. Casas, and R. Urtasun, "Learning to drive via asymmetric self-play," in *European Conference on Computer Vision*. Springer, 2024, pp. 149–168.

[37] B. Chen, J. Francis, J. Oh, E. Nyberg, and S. L. Herbert, "Safe autonomous racing via approximate reachability on ego-vision," *arXiv preprint arXiv:2110.07699*, 2021.

[38] Z. Qin, T.-W. Weng, and S. Gao, "Quantifying safety of learning-based self-driving control using almost-barrier functions," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 12 903–12 910.

[39] K. L. Voogd, J. P. Allamaa, J. Alonso-Mora, and T. D. Son, "Reinforcement learning from simulation to real world autonomous driving using digital twin," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 1510–1515, 2023.

[40] W. Huang, H. Liu, Z. Huang, and C. Lv, "Safety-aware human-in-the-loop reinforcement learning with shared control for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[41] M. Koren, S. Alsaif, R. Lee, and M. J. Kochenderfer, "Adaptive stress testing for autonomous vehicles," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1–7.

[42] 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)*. IEEE, 2020, pp. 2243–2250.

[43] L. Ye, Z. Zhou, and J. Wang, "Improving the generalizability of trajectory prediction models with frenet-based domain normalization," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 562–11 568.

[44] M. Itkina and M. Kochenderfer, "Interpretable self-aware neural networks for robust trajectory prediction," in *Conference on Robot Learning*. PMLR, 2023, pp. 606–617.

[45] A. Filos, P. Tigkas, R. McAllister, N. Rhinehart, S. Levine, and Y. Gal, "Can autonomous vehicles identify, recover from, and adapt to distribution shifts?" in *International Conference on Machine Learning*. PMLR, 2020, pp. 3145–3153.

[46] J. Francis, B. Chen, W. Yao, E. Nyberg, and J. Oh, "Distribution-aware goal prediction and conformant model-based planning for safe autonomous driving," *arXiv preprint arXiv:2212.08729*, 2022.

[47] Y. Lu, J. Fu, G. Tucker, X. Pan, E. Bronstein, R. Roelofs, B. Sapp, B. White, A. Faust, S. Whiteson *et al.*, "Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7553–7560.

[48] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[49] K. Rana, M. Xu, B. Tidd, M. Milford, and N. Sünderhauf, "Residual skill policies: Learning an adaptable skill-based action space for reinforcement learning for robotics," in *Conference on Robot Learning*. PMLR, 2023, pp. 2095–2104.

[50] Z. Huang, Z. Zhang, A. Vaidya, Y. Chen, C. Lv, and J. F. Fisac, "Versatile scene-consistent traffic scenario generation as optimization with diffusion," *arXiv preprint arXiv:2404.02524*, 2024.

[51] 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.

[52] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, no. 2, p. 1805, 2000.

[53] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, and C. Schmid, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 525–11 533.

[54] W. Ma and J. P. Kruth, "Nurbs curve and surface fitting for reverse engineering," *The International Journal of Advanced Manufacturing Technology*, vol. 14, pp. 918–927, 1998.