Ethical Decision-Making for Autonomous Driving: Scenario Modeling, Loss Function Design, and Simulation Implementation

Zhongyuan Jiang

Guangdong Country Garden School, Foshan, China

Zhongyuan.Jiang66@gmail.com

Keywords: Autonomous Driving, Ethical Decision-Making, Social Variational Autoencoder (Social-VAE), CARLA Simulator, Loss Function Design, INTERACTION Dataset

Abstract: Ethical decision-making in autonomous driving remains a fundamental challenge, particularly in dilemma scenarios where vehicles must weigh passenger safety against potential harm to pedestrians or property. In this work, we propose a reinforcement learning framework enhanced with a Social Variational Autoencoder (Social-VAE) to capture interactive behaviors among traffic participants. Ethical dilemmas are modeled through parametric two-lane scenarios, incorporating sensitive variables such as pedestrian identity, passenger composition, and traffic signal states. A nonlinear loss function balances vehicle self-damage and pedestrian harm, enabling continuous control over steering, acceleration, and braking. Training and evaluation are conducted in the CARLA simulator using the INTERACTION dataset to ensure realistic multi-agent dynamics. For benchmarking, we adopt the PCLA leaderboard evaluation protocol, which provides standardized comparison across safety, efficiency, and ethical trade-offs. Our results demonstrate that the proposed framework achieves improved decision consistency and robustness in ethically challenging scenarios, bridging moral reasoning with practical control policies in autonomous vehicles.

1. Introduction

The deployment of autonomous vehicles (AVs) in real-world traffic environments raises not only technical challenges but also profound ethical concerns. Unlike conventional control tasks, ethical decision-making in AVs requires balancing safety, efficiency, and moral responsibility in situations where no outcome is cost-free. Classical examples include dilemma scenarios in which the vehicle must choose between colliding with pedestrians or avoiding them by risking the safety of passengers through a sudden lane change or barrier collision. Designing decision-making frameworks that are both ethically principled and technically feasible remains an open research problem.

Existing studies have largely relied on rule-based systems or simplified binary-choice models, which struggle to generalize to complex, continuous control settings. Moreover, ethical dilemmas are inherently multi-agent, involving pedestrians, other vehicles, and dynamic environments where social interactions strongly influence outcomes. This motivates the integration of data-driven learning methods with simulation environments capable of reproducing realistic physical and social dynamics.

In this work, we propose a reinforcement learning framework augmented with a Social Variational Autoencoder (Social-VAE) to explicitly model the stochastic and interactive nature of human behaviors in traffic. Ethical dilemmas are formalized as a value-loss minimization problem, where both pedestrian harm and self-damage are parameterized in a nonlinear loss function. To bridge theory and practice, we employ the CARLA simulator for environment construction and the INTERACTION dataset to train and validate socially-aware control policies under realistic driving patterns.

To ensure reproducibility and benchmarking, we adopt the PCLA leaderboard evaluation protocol, which measures performance across safety, efficiency, and ethical trade-offs. Our contributions can be summarized as follows:

(1) We design a parametric ethical dilemma modeling framework that captures pedestrian identity, passenger composition, traffic signals, and vehicle state.

  DOI: 10.25236/iwmecs.2025.001

(2) We introduce a Social-VAE reinforced learning approach that generates socially consistent driving behaviors in multi-agent environments.

(3) We implement and evaluate the framework in CARLA using the INTERACTION dataset, with standardized benchmarking via the PCLA leaderboard.

This work bridges the gap between abstract ethical reasoning and practical control in autonomous driving, offering a reproducible pathway toward ethically-aware AV decision-making.

(1) Innovative Use of SocialVAE: Our approach incorporates RNNs equipped with LSTM units [1]. This setup enables the generation of complex traffic scenarios that closely mimic real-world interactions among vehicles.

(2) Attention Mechanism for Neighbor Encoding: An attention mechanism to encode the states of neighboring vehicles considers the social features exhibited by these entities. This development is critical in scenarios with dense traffics, ensuring model accuracy among vehicles in proximity.

(3) Practical Implementation on PCs: The methodology is designed to run on standard PCs, enhancing accessibility and broadening testing capabilities.

2. Background

Caesar et al. (2020) [3] and Houston et al. (2020) [5] have noted that traditional datasets primarily sourced from real-world driving are significantly limited due to the rarity of near-collision scenarios, which are crucial for testing autonomous vehicle (AV) systems. While simulation platforms like CARLA (Dosovitskiy et al., 2017) [4] and NVIDIA's DRIVE Sim have addressed these issues by providing controlled environments where diverse and uncommon scenarios can be tested, these tools still struggle with replicating the dynamic complexity of real-world conditions.

Innovations by Bergamini et al. (2021) [2] have advanced the field by using deep learning techniques such as variational autoencoders (VAEs) and GANs to generate more plausible and challenging traffic scenarios. Despite these advancements, existing simulations often fail to adjust in response to the evolving behaviors of AV systems during testing, limiting their application in developing robust decision-making frameworks for AVs.

To surmount these challenges, our approach incorporates the SocialVAE [8] to create adaptive traffic scenarios that more effectively test AV systems. This method simulates a broad range of adversarial conditions within a learned traffic model, dynamically generating scenarios that provoke specific undesirable behaviors from the AV. Unlike previous methods, our approach does not rely on a set adversarial strategy; instead, it continuously adapts to the AV's reactions, ensuring that the scenarios are both realistic and tailored to test the AV's unique capabilities thoroughly. This technique aims to enhance the safety and reliability of AVs by providing a more comprehensive testing framework that reflects the unpredictable nature of real-world driving.

3. Related Work

Xu et al.'s work [8] contributes to understanding pedestrian dynamics within traffic systems using a sophisticated timewise VAE. This focus on pedestrian behavior. However, SocialVAE primarily addresses pedestrian trajectories and does not extend to the intricacies of vehicular dynamics. Our approach enhances the scope of traffic management systems to better predict complex traffic interactions.

On the other hand, Rempe et al.'s research [6] emphasizes the generation of challenging vehicular scenarios, employing a graph-based conditional VAE (CVAE) to create challenging traffic conditions. This is pivotal for testing the limits of predictive capabilities under potential collision scenarios. Although it is highly effective, STRIVE's model training and simulation largely depends on hardware requirements. Our approach replaced the CVAE with SocialVAE, which has a simpler configuration. Thus, personal implementation on a PC or laptop becomes available.

4. Approach

4.1. Overall Structure

Following previous research [7], the scenario generation is approached as an optimization problem that modifies agent trajectories in a baseline scenario derived from real-world data. The SocialVAE method estimates the distribution of future trajectories for each agent in a scene, using historical observations. It predicts each agent (ego vehicle)'s future independently and can handle scenes with any number of agents. Undesirable outcomes include collisions, uncomfortable driving conditions, and violations of traffic laws. The computation graph shows the state transfer inside the VAE. The overall structure is shown in Fig. 1.

4.2. Ethical Scenario Construction

While the SocialVAE module provides the ability to generate diverse and socially consistent trajectories, it does not by itself impose ethical considerations. To explicitly model moral dilemmas, we construct a set of parametric ethical scenarios in which the ego vehicle must resolve trade-offs between passenger safety, pedestrian protection, and compliance with traffic rules. These scenarios extend the baseline INTERACTION dataset with additional ethically sensitive parameters and are instantiated within the CARLA simulator.

Scenario Design. We adopt the classical two-lane dilemma as a core template: the ego vehicle encounters an unavoidable hazard and must either continue in its current lane, colliding with one or more pedestrians, or swerve into a roadside barrier, endangering its passengers. To enrich this template, we introduce the following parameter categories (also illustrated in Fig. 2):

• Pedestrian attributes: Annotated by age, social role, and group size, affecting ethical evaluation weights.

• Passenger composition: Includes the number and type of occupants, impacting $\mathcal{L}_{\text{self}}$.

• Traffic signals and right-of-way: Encodes legality of pedestrian crossing, influencing $\mathcal{L}_{\text{rule}}$.

• Vehicle initial state: Parameterizes lane position, speed, and heading to determine feasible evasive maneuvers.

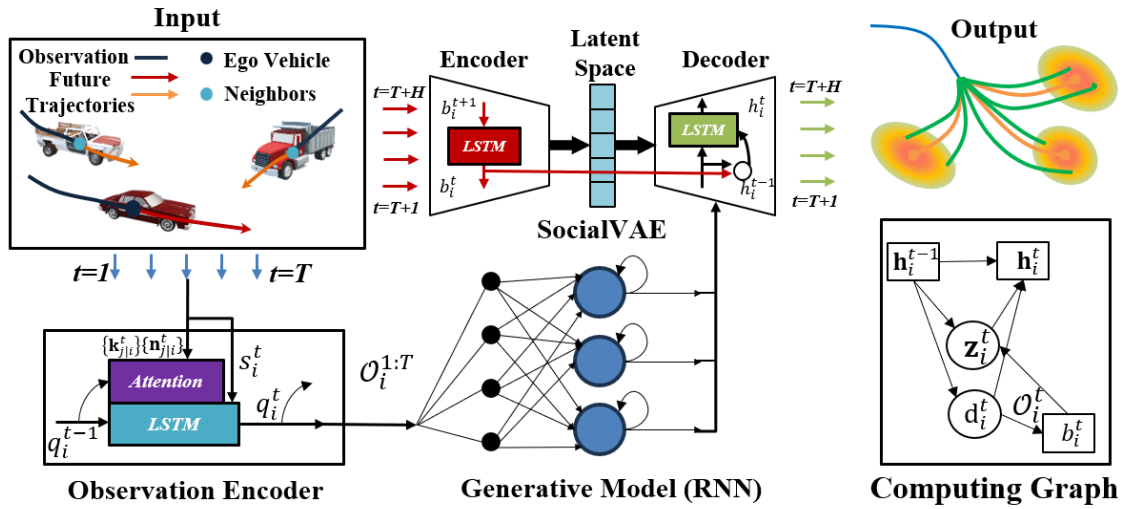

Figure 1 An overview of accident prone traffic trajectory generation with SocialVAE

It incorporates a recurrent neural network (RNN)-based VAE operating in a timewise manner with stochastic latent variables generated sequentially for predicting trajectories. The observation encoder's attention mechanism takes into account the state $n_{j|i}$ and social features $k_{j|i}$ of each neighboring entity. The diagram on the right illustrates the flow of states within the timewise VAE.

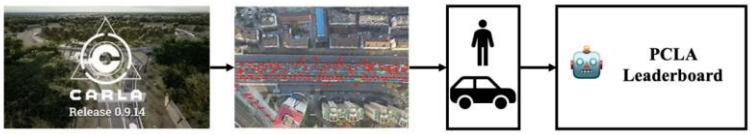

Figure 2 Pipeline of Ethical Scenario Construction.

A baseline traffic scene from the INTER-ACTION dataset is augmented with ethical parameters, including pedestrian attributes, passenger composition, traffic signals, and vehicle initial state. SocialVAE generates plausible trajectories for surrounding agents, while the ego vehicle's actions are evaluated using a multi-component ethical loss.

Value Quantification and Loss. Each outcome is mapped to a value-cost system to balance ethical priorities. The total ethical loss is:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{self}} + \beta \cdot \mathcal{L}_{\text{pedestrian}} + \gamma \cdot \mathcal{L}_{\text{rule}}, \tag{1}$$

where α, β, γ weight the importance of passenger safety, pedestrian harm, and traffic rule compliance, respectively. This loss guides both the generation of ego trajectories and evaluation of scenario difficulty.

Scenario Generation Mechanism. The pipeline (Fig. 2) combines:

(1) Retrieval of a baseline scene from the INTERACTION dataset.

(2) Injection of ethically annotated pedestrians and obstacles.

(3) Trajectory generation for surrounding agents using SocialVAE to maintain realistic social interactions.

(4) Ego vehicle control evaluation under continuous actions (steering, acceleration, braking) in CARLA, using the multi-component ethical loss.

CARLA Implementation. Ethical parameters are dynamically adjustable through a scenario parameter interface. This modular design ensures reproducibility, supports batch simulation, and allows systematic benchmarking of AV planners under ethically challenging situations.

4.3. SocialVAE

Generative Model: Using the LSTM Structure,instead of directly predicting the absolute coordinates, we define a displacement sequence $d^i_{t+1:t+H}$. The generative model is defined as Eq. 1, where z^i_t, d^i_t and $O^i_{1:T}$ denote the latent variables introduced at time step t, the displacement sequence and the observation sequence, respectively.

$$p(d_i^{T+1:T+H}|O_i^{1:T}) = \prod_{t=T+1}^{T+H} \int_{z_i^t} p\,(d_i^t|d_i^{T:t-1}, O_i^{1:T}, z_i^t) p(z_i^t|d_i^{T:t-1}, O_i^{1:T}) dz_i^t \tag{2}$$

To implement the sequential generative model $p(d^i_{t+1}|o^i_{1:T}, z^i_t)$, we use LSTM where the state variable h^i_t is updated recurrently by $\mathbf{h}^t_i = \vec{g}(\psi_{\mathbf{zd}}(z^t_i, \mathbf{d}^t_i), \mathbf{h}^{t-1}_i)$, where $t=T+1,...,T+H$. The prior distribution of SocialVAE is conditioned and can be obtained from the LSTM state variable. The second term of Eq. 1 can be expressed as Eq. 2, where θ are parameters for a neural network to be optimized.

$$p(z_i^t|d_i^{T:t-1}, O_i^{1:T}) := p_\theta(z_i^t|h_i^{t-1}) \tag{3}$$

Latent Space Sampling: The first component of the integral shown in Eq. 1 suggests that new displacements are sampled from the prior distribution p, which depends on the latent variable z^i_t and incorporates both observations and earlier displacements as reflected by h^i_{t-1}. Thus, $d^t_i \sim p_\xi(\cdot\,|z^t_i, h^{t-1}_i)$ represents the sampled displacement. where z^i_t, h^i_{t-1} and ξ denote conditioned latent variables, previous displacements and the observation sequence, respectively. Therefore, we can obtain $\mathbf{x}^t_i = \mathbf{x}^T_i + \sum_{\tau=T+1}^{t} \mathbf{d}^\tau_i$ as a stochastic estimation for the spatial position at time t.

Inference Model: To estimate the posterior distribution q over the latent variables, the entire GT observation sequence from $O^i_{1:T+H}$ is utilized. This is denoted by Eq. 3, where t ranges from T + 1 to T + H, and the initial state $b^i_{T+H+1}=0$. The backward state b^i_t transmits GT trajectory data from T + H

down to t, forming the posterior by combining information from both the backward state b_t^i and the forward state h_t^i.

$$b_i^t = \overleftarrow{g}(O_i^t, b_i^{t+1}) \tag{4}$$

Observation Encoding: If there are multiple neighboring agents in the scene during the prediction process. We need to treat the local observation from agent i to the scene at time t = 2, ..., T as Eq. 4. This includes data from agent and a combined representation of all its neighboring agents. s_i^t is the self-state of agent i, $\mathbf{n}_{j|i}^t$ is the local state of neighbor agent j,f_s,f_n are learnable feature extraction neural networks and $w_j^t|i$ is the attention mechanism weight if t ≤ T.

$$O_i^t := \left[f_s(s_i^t), \sum_j w_{j|i}^t \, f_n(n_{j|i}^t) \right] \tag{5}$$

Training Loss: The VAE calculates the loss for backpropagation and network weight updates. The loss is a combination
of several components: $\min_\theta L_{kl} + L_{mse} + L_{adv} + L_{kin}$

- **KL Loss**: Measures the difference between the encoded distribution and a standard normal distribution.

$$L_{kl} = w_{KL} D_{KL} \left[q_\phi(z_i^t | b_i^t, h_i^{t-1}) \| p_\theta(z_i^t | h_i^{t-1}) \right] \tag{6}$$

- **Adversarial Loss**: Penalizes predicted trajectories that come too close to neighboring trajectories, using Euclidean distance between the i-th predicted point and the j-th neighbor's position, i.e. $\mathbf{e}_{i,j} = \hat{y}_i - \mathbf{n}_{i,j}$.

$$L_{adv} = \sum_{i=1}^{N} \sum_{j=1}^{M} \exp\left(-\sqrt{\|e_{i,j}\|_2} \right) \cdot \frac{1}{\sum_{k=1}^{M} \exp\left(-\sqrt{\|e_{i,k}\|_2} \right)} \tag{7}$$

- **Average Weighted MSE**: Weighted version of the mean squared error between original and reconstructed data. Let $w_t = \exp(-\alpha t)$ be the weight for time step t, where α is the decay rate.

$$L_{mse} = \frac{\sum_{t=1}^{T} w_t \sum_{i=1}^{N} (\hat{y}_{t,i} - y_{t,i})^2}{\sum_{t=1}^{T} w_t} \tag{8}$$

- **Kinematic Loss**: Penalizes deviations in velocities and angular velocities of the predicted trajectories, where $\hat{\mathbf{d}}_t, \Delta\theta_t$ are the displacement and angular velocity at time t.

$$L_{kin} = \sum_{t=1}^{T-1} \| \hat{d}_{t+1} - \hat{d}_t \|_2 + \sum_{t=1}^{T-2} \| \Delta\theta_{t+1} - \Delta\theta_t \|_2 \tag{9}$$

Final Position Clustering (FPC): FPC is implemented to improve the diversity of trajectories. For each cluster, FPC selects the trajectory closest to the center, generating a diverse set of predictions, as shown in Fig. 3. This approach reduces prediction bias by avoiding the over-representation of trajectories from high-density regions.

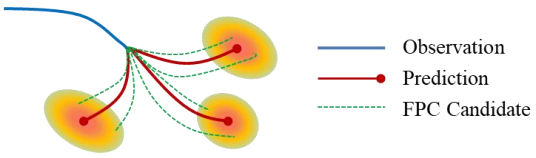

Figure 3 An example of FPC to extract 3 predictions from 9 candidates

5. Experiments

5.1. Implementation Details

Dataset. To comprehensively evaluate autonomous vehicle (AV) decision-making in accident-prone and ethically challenging situations, we utilize both real-world and synthetic data sources. The INTERACTION Dataset [9] provides diverse traffic scenarios from global intersections, roundabouts,

and highways, annotated with vehicle, pedestrian, and cyclist trajectories as well as traffic signal states. This allows realistic modeling of agent interactions, social compliance, and rare near-collision events. To extend scenario coverage to ethically sensitive and high-risk situations that rarely occur in real-world datasets, we construct synthetic scenarios using RoadRunner [10], a high-fidelity scene authoring tool. RoadRunner enables precise control over road topology, lane geometry, dynamic agent behaviors, static obstacles, and traffic signal configurations.

Ethical scenarios are parameterized by agent attributes such as pedestrian age, mobility, and priority, vehicle passenger composition, initial lane positions, and compliance with traffic rules. This enables generation of forced-choice dilemmas, for instance, where the AV must select between colliding with a pedestrian, another vehicle, or a static obstacle. Generated scenarios capture variations in agent behavior, including sudden crossings, lane changes, and braking maneuvers, reflecting the stochastic nature of real traffic and potential ethical conflicts. SocialVAE [8] is then used to generate multimodal trajectories conditioned on historical observations and ethical parameters, incorporating attention mechanisms to model social interactions among agents and latent variables to represent uncertainty and multimodality.

All scenarios, both real and synthetic, are executed in CARLA [4] to evaluate AV planners. Metrics such as collision rates, near-misses, traffic rule adherence, and comfort indices are recorded, and the impact of ethical parameters on AV decision-making is analyzed. Fig. 4 illustrates the full pipeline: real-world data from INTERACTION informs baseline scenarios, RoadRunner constructs synthetic ethical scenarios, SocialVAE generates agent trajectories, and CARLA executes the scenarios for evaluation. This hybrid approach enables rigorous testing of AV systems under realistic, accident-prone, and ethically complex conditions.

Algorithm 1 SocialVAE Structure

```
class SocialVAE:
    function init_():
        Initialize model parameters
        Define sub-modules and RNNs
    function attention(q, k, mask):
        Compute & return attention weights
    function enc(x, neighbor, y):
        Compute social features
        Update RNN state
        Return final state
    function forward(x, neighbor, n_predictions):
        if training then Call learn function
        Generate & return predictions
    function learn(x, y, neighbor):
        Encode inputs
        Compute & return errors and losses
    function loss(err, kl, L_adv_loss, avg_weighted_mse_loss, Knematic_loss):
        Compute & return total and individual losses
```

Figure 4 Hybrid pipeline for ethical and accident-prone scenario generation.

The real-world trajectories from the INTERACTION Dataset provide baseline data. RoadRunner

is used to construct synthetic scenarios with ethical annotations. SocialVAE generates multimodal agent trajectories conditioned on observations and ethical parameters. Scenarios are executed in CARLA for AV evaluation.

Training. In [6], training was conducted on a computing cluster comprising an NVIDIA Titan RTX GPU and 12 Intel i7-7800X @3.5GHz CPUs, offering significantly greater computational power and memory than a personal computer. We utilized SocialVAE and a smaller dataset, making training feasible on a personal computer, while still achieving favorable results with the generated trajectories within hours. Hardware and parameters we used are listed in Tab. 1. Training losses are plotted in Fig. 6.

Training. Training was conducted on a high-performance computing server equipped with an NVIDIA A100 GPU and dual Intel Xeon Gold 6230 CPUs @ 2.1GHz, providing substantial computational power and memory to efficiently handle SocialVAE training on the full dataset. This setup allowed training of the model with larger batch sizes and longer prediction horizons while maintaining reasonable runtime. Hardware specifications and hyperparameters are summarized in Tab. 1. Training losses are illustrated in Fig. 6.

Table 1 SocialVAE Training and Hyperparameters on Professional Server.

Hardware			
Parameter	**Value**	**Parameter**	**Value**
Computing Platform	NVIDIA A100 GPU	CPU	2x Intel Xeon Gold 6230 @ 2.1GHz
GPU Memory	40GB	RAM	256GB
Hyperparameters			
Parameter	**Value**	**Parameter**	**Value**
Utilized Model	SocialVAE	Observation Radius	10000
Prediction Time Steps	25	Observation Time Steps	10
RNN Hidden Layer Dim	512	Latent Variable Dim	32
Embedding Layer Dim	128	Input Dim	2
Feature Dim	256	Batch Size	512
Learning Rate	1 × 10 -4	Weight Scaling Factor	0.1

5.2. Qualitative Results of Scenario Generation with Ethical Considerations

In this section, we present the qualitative results of our accident-prone traffic scenario generation, emphasizing ethical decision-making aspects. By combining SocialVAE-based trajectory generation with RoadRunner-constructed scenarios, our framework produces complex situations that challenge autonomous vehicle planners with respect to both safety and ethical considerations. The training reconstruction loss over epochs for SocialVAE is shown in Fig. 5.

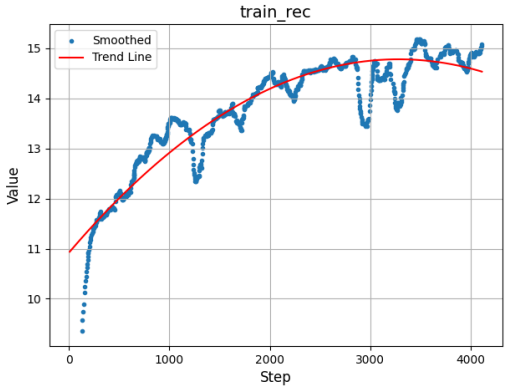

Figure 5 The training reconstruction loss over epochs for SocialVAE

The plot illustrates the convergence of the model during training on the INTERACTION and RoadRunner datasets, demonstrating stable optimization and gradual reduction in reconstruction error.

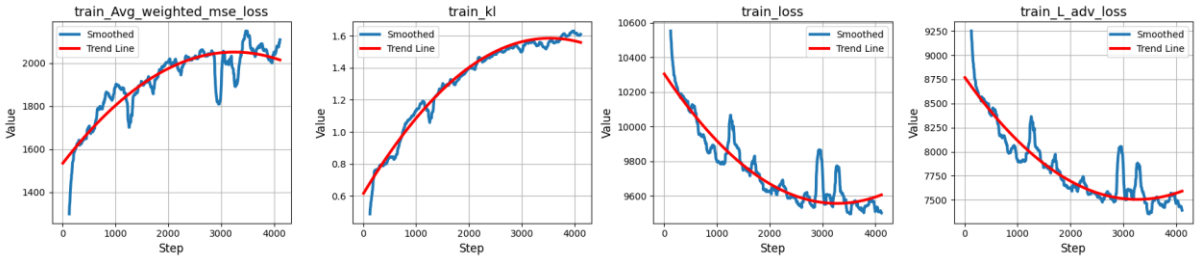

Figure 6 Training losses of SocialVAE

The variation trends of each loss term during training are shown in Fig. 6, where the average weighted MSE loss, KL loss, and adversarial loss all steadily decrease with training steps, and the total loss exhibits a continuous convergence trend, verifying the stability and effectiveness of the model optimization.

5.2.1. Intersection EP1 Scenario

Fig. 7 shows the ego vehicle's trajectory in the Intersection EP1 scenario, which is a multi-lane urban intersection with moderate traffic density. RoadRunner was used to construct dynamic traffic participants, including other vehicles with realistic acceleration and deceleration profiles.

• **Ethical Context:** The ego vehicle faces a potential head-on collision with an oncoming vehicle if it maintains its planned lane. Alternatively, it could brake to avoid the collision, risking a rear-end impact with the trailing vehicle. This scenario tests the AV's ability to weigh passenger safety against potential harm to other road users, a classic ethical dilemma in AV planning.

• **Generated Trajectories:** SocialVAE outputs multiple plausible trajectories, ranging from aggressive maneuvers that maintain lane priority to conservative braking trajectories. The multi-modal nature of SocialVAE captures the uncertainty in surrounding vehicles' behavior and the resulting trade-offs in ego decisions.

• **Ethical Implication:** The head-on collision trajectory highlights situations where strict adherence to traffic rules may endanger other participants, whereas the rear-end collision demonstrates the AV's compromise between risk reduction and operational feasibility. These trajectories provide rich test cases for evaluating AV decision-making under ethical constraints.

5.2.2. Roundabout FT Scenario

Fig. 8 illustrates trajectories generated in a complex roundabout, featuring interactions with both other vehicles and pedestrians. RoadRunner was used to introduce dynamic pedestrians crossing the roundabout unexpectedly, enabling the generation of ethical conflict scenarios.

• **Ethical Context:** The AV must balance pedestrian safety, compliance with traffic rules, and smooth traffic flow. In this scenario, the vehicle may need to perform sudden braking or lane adjustments to avoid a pedestrian, potentially causing minor collisions with other vehicles.

• **Generated Trajectories:** SocialVAE generates a diverse set of trajectories, including:

(1) *Pedestrian avoidance*: the ego vehicle decelerates sharply to prevent hitting a pedestrian, possibly leading to a rear-end collision.

(2) *Aggressive lane-maintaining*: the vehicle maintains its lane and speed, resulting in a side-impact with other vehicles or a near-miss with pedestrians.

(3) *Balanced maneuvering*: the vehicle adjusts its trajectory smoothly, partially sacrificing lane adherence to minimize harm to all participants.

• **Ethical Implication:** These trajectories illustrate the AV's trade-offs among minimizing harm, following traffic laws, and maintaining operational efficiency. They allow researchers to evaluate the planner's ethical reasoning and robustness in dynamic, multi-agent environments.

5.2.3. Multiple Scenario Trajectory Visualization

Fig. 9 shows a set of generated trajectories for both scenarios. Each color corresponds to a different predicted trajectory sampled from SocialVAE's latent space, demonstrating the diversity of possible outcomes.

• **Trajectory Diversity:** The figure emphasizes the multi-modality of the generated predictions. Even within the same scenario, the AV may choose distinct maneuvers based on latent variables capturing uncertainties in human driver behaviors and environmental conditions.

• **Ethical Scenario Representation:** By visualizing multiple trajectories, we can analyze which ethical trade-offs are made under varying conditions, such as prioritizing pedestrian safety, reducing collisions with vehicles, or preserving smooth traffic flow.

• **Scenario Utility:** These visualizations provide insight into edge cases where AVs face conflicting ethical objectives, making them valuable for evaluating planners and refining policy design.

Overall, integrating SocialVAE with RoadRunner-generated scenarios produces ethically challenging traffic situations with realistic multi-agent interactions. This approach facilitates rigorous testing of autonomous vehicles, ensuring that planners are evaluated not only on collision avoidance but also on their decision-making under ethical uncertainty.

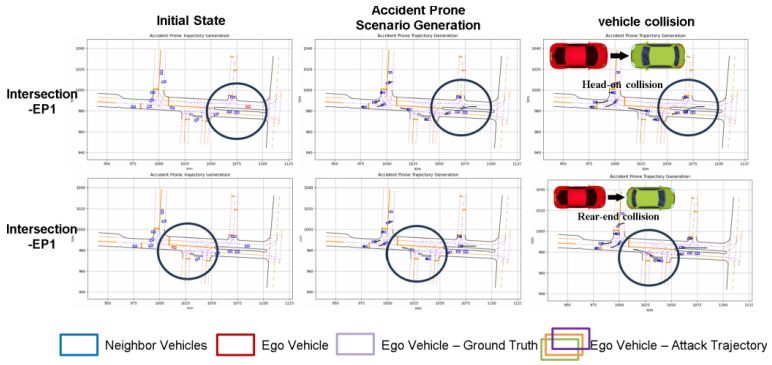

Figure 7 Generated Trajectory in Intersection Scenario

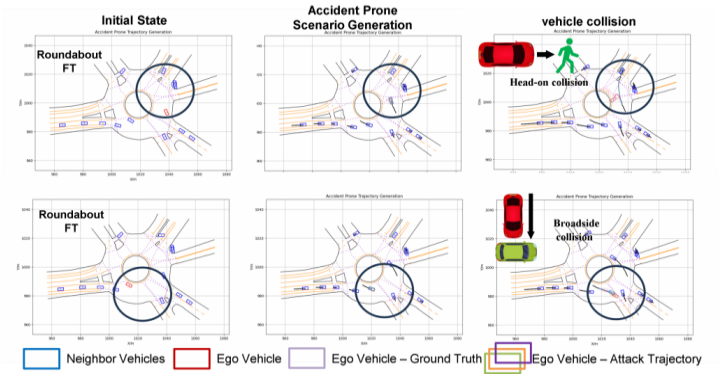

Figure 8 Generated Trajectory in Roundabout Scenario

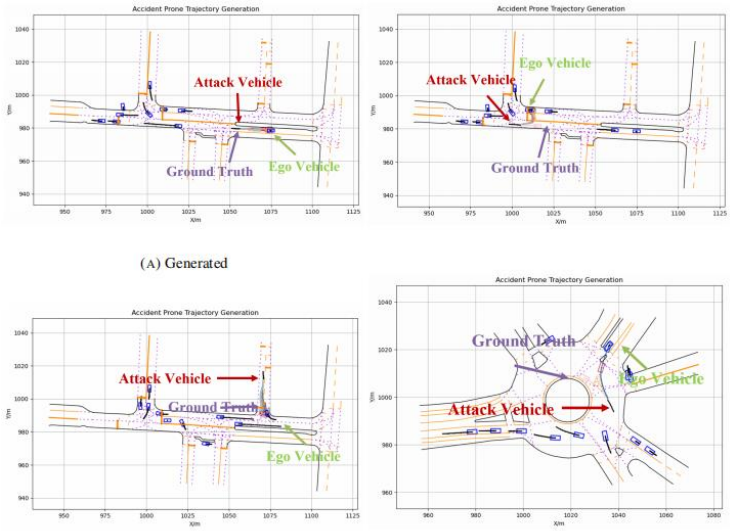

Figure 9 Examples of generated accident-prone traffic scenarios.

The (a) and (b) in Figure 9 depict head-on and rear-end collision trajectories of the ego vehicle in an intersection scenario. The (c) and (d) in Figure 9 show collisions involving pedestrians and side-impact interactions in a roundabout scenario. These scenarios are constructed to evaluate autonomous vehicle decision-making under challenging ethical and safety-critical conditions.

6. Discussion

In this work, we have presented a framework for generating ethically challenging traffic scenarios using SocialVAE and RoadRunner, evaluated on real-world driving data from the INTERACTION Dataset. The qualitative results demonstrate that our approach can produce a diverse set of trajectories that highlight potential collisions, uncomfortable driving conditions, and ethical dilemmas, such as prioritizing pedestrian safety versus vehicle occupant safety.

6.1. Ethical Implications

The generated scenarios provide a systematic method to explore edge cases in autonomous driving where ethical decision-making is critical. Unlike conventional datasets that primarily focus on common driving situations, our approach intentionally produces rare and challenging interactions. By doing so, planners and reinforcement learning policies can be stress-tested for situations requiring moral trade-offs, such as deciding between potential harm to pedestrians versus nearby vehicles. This capability is crucial for the deployment of AVs in real-world conditions, where rare but high-impact events may determine public trust and regulatory acceptance.

6.2. Trajectory Diversity and SocialVAE Efficacy

SocialVAE demonstrates strong capability in modeling the multi-modal nature of agent trajectories. The stochastic latent variables allow sampling of multiple plausible futures for each agent, capturing uncertainties in human driving behavior and environmental interactions. The generated trajectories vary significantly in both ego vehicle behavior and outcomes, providing a rich testbed for evaluating AV decision-making under uncertainty.

6.3. Scenario Construction with RoadRunner

The integration of RoadRunner enables the creation of realistic, interactive, and repeatable scenarios. By designing environments with dynamic agents and ethical conflict points, we can systematically investigate planner robustness. The ability to programmatically manipulate agent behaviors and scenario layouts ensures reproducibility while still reflecting the stochastic nature of real traffic.

6.4. Limitations

Despite the promising results, our framework has limitations. First, the ethical evaluation is qualitative; quantifying moral decisions remains challenging. Second, SocialVAE predictions are constrained by the historical data distribution and may not fully capture extremely rare behaviors. Third, RoadRunner scenarios, while realistic, may not account for all real-world variability, such as weather effects or sensor noise.

6.5. Future Work

Future research could extend this framework in several directions. One avenue is integrating explicit ethical reasoning modules into the trajectory planning stage, enabling the AV to weigh ethical trade-offs quantitatively. Another direction is combining additional datasets, including diverse traffic cultures and conditions, to improve generalization. Finally, closed-loop testing in simulated environments with full AV control could validate the practical effectiveness of the generated ethical scenarios, bridging the gap between scenario generation and autonomous vehicle deployment.

Overall, our study highlights the importance of ethically-aware scenario generation for autonomous vehicles. By leveraging SocialVAE, RoadRunner, and real-world datasets, we provide a methodology to systematically stress-test planners and reinforce safe and socially acceptable AV behavior.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016.

[2] Luca Bergamini, Yawei Ye, Oliver Scheel, Long Chen, Chih Hu, Luca Del Pero, Błazej Osinski, Hugo Grimmet, and Peter Ondruska. Simnet: Learning reactive self-driving simulations from real-world observations. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[4] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on Robot Learning*, pages 1–16. PMLR, 2017.

[5] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska. One thousand and one hours: Self-driving motion prediction dataset. https://level-5.global/level5/data/, 2020.

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

[7] Wang J, Pun A, Tu J, et al. Advsim: Generating safety-critical scenarios for self-driving vehicles[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9909-9918.

[8] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. *SocialVAE: Human Trajectory Prediction Using Timewise Latents*, page 511–528. Springer Nature Switzerland, 2022.

[9] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Ku… mmerle, Hendrik Ko… nigshof, Christoph Stiller, Arnaud de La Fortelle, and Masayoshi Tomizuka. INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in

Interactive Driving Scenarios with Semantic Maps. *arXiv:1910.03088 [cs, eess]*, 2019.

[10] Pandolfino E R, Douglas L A. Historical and Current Status of the Greater Roadrunner in the Central Valley and Surrounding Foothills of California[J]. Central Valley Birds, 2024, 27(4): 124-151.

Appendix

Appendix A. The Video Frame Sampling Visualization is shown in Figure 10.

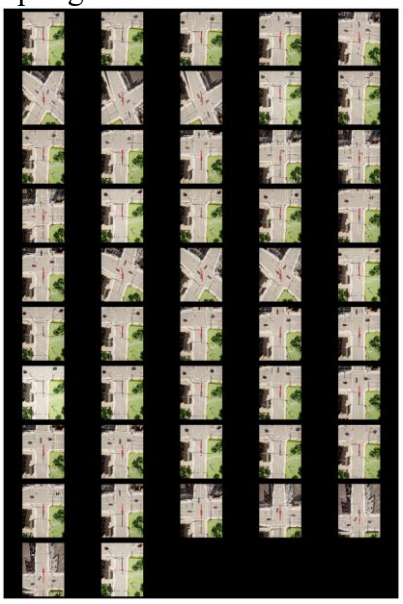

Figure 10 Visualization of frames sampled from the video at 2 frames per second. Each row contains 5 frames. This provides an overview of the temporal evolution of the scenario and highlights key moments captured for scenario generation analysis.