ACCIDENT PRONE TRAFFIC TRAJECTORY GENERATION USING SOCIALVAE

ZHANQIAN WU [ZHANQIAN@SEAS], BOWEN JIANG [JBWJOY@SEAS], JIE MEI [JIEMEI@SEAS],

ABSTRACT. This project focuses on generating accident-prone traffic scenarios using SocialVAE to assess and train autonomous vehicles (AVs). SocialVAE uses a timewise variational autoencoder (VAE) to generate future vehicle trajectories, leveraging a recurrent neural network (RNN) architecture with Long Short-Term Memory units. By conditioning latent variables at each time step and incorporating a backward RNN for navigation inference, the model captures the uncertainty and multimodality of vehicle decision-making. An attention mechanism encodes neighboring vehicles' states and interactions to enhance prediction accuracy. Experiments on the INTERACTION dataset validated this framework's ability to refine AV safety and reliability, providing an accessible solution that operates on standard PC hardware.

1. INTRODUCTION

In the realm of autonomous vehicle (AV) technology, ensuring the safety and reliability of these systems in diverse and unpredictable traffic conditions remains a paramount challenge. Traditional datasets often fall short in providing the variety of complex, real-world scenarios needed to thoroughly test and train AV systems. This limitation is acute in the case of near-collision events, which are rare yet critical for assessing an AV's ability to handle hazardous situations.

In this study, we introduce a streamlined approach for generating challenging traffic scenarios to evaluate the robustness of AV systems in complex environments. Central to our methodology is the use of a generative model of traffic movement to assess the plausibility of scenarios during optimization by their likelihood. Our Accident Prone Traffic Trajectory Generation method employs a graph-based SocialVAE, which incorporates a vanilla recurrent neural network (RNN) architecture with Long Short-Term Memory (LSTM) units for sequential predictions. This model is enhanced with latent variables that capture the intricate dynamics of vehicle movements, offering a realistic simulation of demanding traffic situations. Additionally, a backward RNN structure is used to infer navigation patterns from complete trajectories, and an attention mechanism dynamically encodes the states of neighboring vehicles, considering their social interactions. These features allow for a detailed representation of traffic dynamics, facilitating the generation of maps that include trajectories with potential accidents. The generated scenarios are then analyzed for qualitative performance, demonstrating the system's capability to operate efficiently on standard PC hardware.

1.1. Contributions.

(1) **Innovative Use of SocialVAE**: Our approach incorporates RNNs equipped with LSTM units [2]. 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) [4] and Houston et al. (2020) [7] 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) [6] 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) [3] 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 [10] 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 [10] 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 [8] 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 [9], 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.

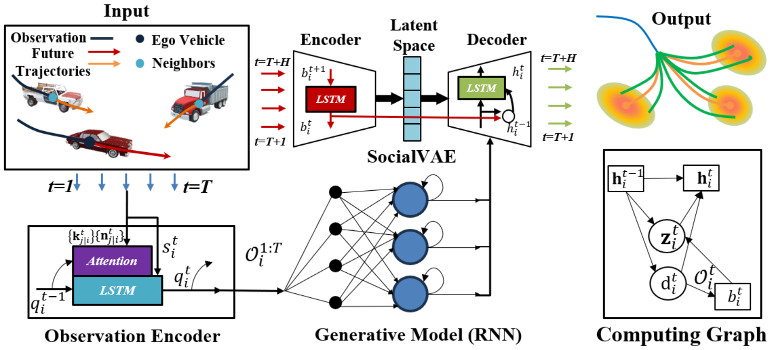

FIGURE 1. An overview of accident prone traffic trajectory generation with SocialVAE, which incorporates a recurrent neural network (RNN)-based VAE operating in a timewise manner with stochastic latent variables generated sequentially for predicting trajectories. The observation encoder's attention mechanism 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.

4.2. **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.

$$p(\mathbf{d}_i^{T+1:T+H}|\mathcal{O}_i^{1:T}) = \prod_{t=T+1}^{T+H} \int_{\mathbf{z}_i^t} p(\mathbf{d}_i^t|\mathbf{d}_i^{T:t-1}, \mathcal{O}_i^{1:T}, \mathbf{z}_i^t) p(\mathbf{z}_i^t|\mathbf{d}_i^{T:t-1}, \mathcal{O}_i^{1:T}) \mathbf{d}\mathbf{z}_i^t. \tag{1}$$

To implement the sequential generative model $p(d_{t+1}^i|o_{1:T}^i, z_t^i)$, we use LSTM where the state variable h_t^i is updated recurrently by $\mathrm{h}_i^t = \overrightarrow{g}(\psi_{\mathbf{z}\mathrm{d}}(\mathrm{z}_i^t, \mathrm{d}_i^t), \mathrm{h}_i^{t-1})$, 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.

$$p(\mathbf{z}_i^t|\mathbf{d}_i^{T:t-1}, \mathcal{O}_i^{1:T}) := p_\theta(\mathbf{z}_i^t|\mathbf{h}_i^{t-1}) \tag{2}$$

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_t^i and incorporates both observations and earlier displacements as reflected by h_{t-1}^i. Thus, $\mathrm{d}_i^t \sim p_\xi(\cdot|\mathrm{z}_i^t, \mathrm{h}_i^{t-1})$ 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 $\mathbf{x}_i^t = \mathbf{x}_i^T + \sum_{\tau=T+1}^{t} \mathbf{d}_i^\tau$ 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 $O_{1:T+H}^i$ is utilized. This is denoted by Eq. 3, where t ranges from $T+1$ to $T+H$, and the initial state $b_{T+H+1}^i = 0$. The backward state b_t^i 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.

$$\mathbf{b}_i^t = \overleftarrow{g}(\mathcal{O}_i^t, \mathbf{b}_i^{t+1}) \tag{3}$$

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 $t \leq T$.

$$\mathcal{O}_i^t := \left[f_{\mathbf{s}}(\mathbf{s}_i^t), \textstyle\sum_j w_{j|i}^t f_{\mathbf{n}}(\mathbf{n}_{j|i}^t)\right] \tag{4}$$

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

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

$$\mathcal{L}_{\text{kl}} = w_{KL} D_{KL}\left[q_\phi(\mathbf{z}_i^t|\mathbf{b}_i^t, \mathbf{h}_i^{t-1}) || p_\theta(\mathbf{z}_i^t|\mathbf{h}_i^{t-1})\right] \tag{5}$$

- **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. $\mathbf{e}_{i,j} = \hat{\mathbf{y}}_i - \mathbf{n}_{i,j}$.

$$\mathcal{L}_{\text{adv}} = \sum_{i=1}^{N}\sum_{j=1}^{M} \exp\left(-\sqrt{\|\mathbf{e}_{i,j}\|_2}\right) \cdot \frac{1}{\sum_{k=1}^{M} \exp\left(-\sqrt{\|\mathbf{e}_{i,k}\|_2}\right)} \tag{6}$$

- **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.

$$\mathcal{L}_{\text{mse}} = \frac{\sum_{t=1}^{T} w_t \sum_{i=1}^{N} (\hat{\mathbf{y}}_{t,i} - \mathbf{y}_{t,i})^2}{\sum_{t=1}^{T} w_t} \tag{7}$$

- **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.

$$\mathcal{L}_{\text{kin}} = \sum_{t=1}^{T-1} \|\hat{\mathbf{d}}_{t+1} - \hat{\mathbf{d}}_t\|_2 + \sum_{t=1}^{T-2} \|\Delta\theta_{t+1} - \Delta\theta_t\|_2 \tag{8}$$

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. 2. This approach reduces prediction bias by avoiding the over-representation of trajectories from high-density regions.

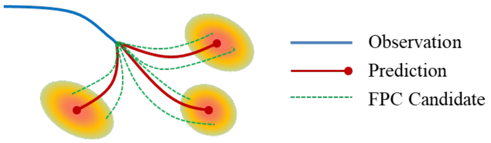

FIGURE 2. An example of FPC to extract 3 predictions from 9 candidates

5. EXPERIMENTS

5.1. Implementation Details.

Dataset. The INTERACTION Dataset [11] provides a diverse range of global driving scenarios crucial for autonomous vehicle research. It includes detailed annotations of dynamic behaviors and complex interactions, supporting studies in motion prediction and behavior modeling to enhance vehicle safety and efficacy.

Training. In [8], 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. 3.

TABLE 1. SocialVAE Training and Hyperparameters

Hardware			
Parameter	**Value**	**Parameter**	**Value**
Computing Platform	NVIDIA RTX 2060 GPU	CPU	Intel i7-9750H @ 2.6GHz
GPU Memory	6GB	RAM	16GB
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	128
Learning Rate	1×10^{-4}	Weight Scaling Factor	0.1

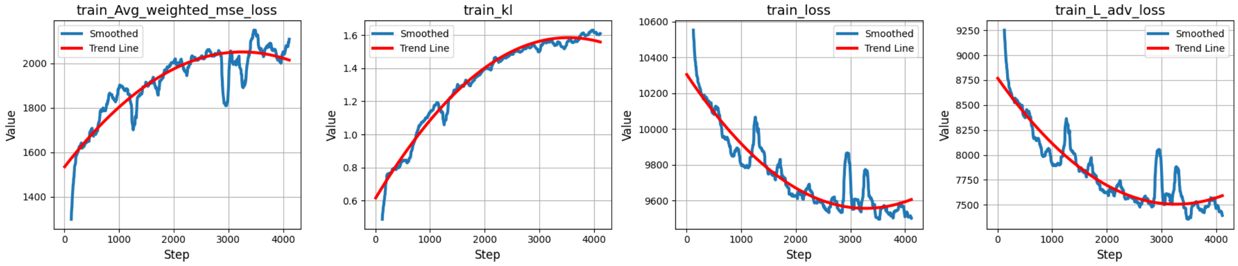

FIGURE 3. Losses

5.2. Qualitative Results of Scenario Generation.

Compared to rolling out a given planner on unmodified scenarios, challenging scenarios from our approach produce collisions and less comfortable driving. Fig. 4 presents the generated trajectories of the ego vehicle in the Intersection EP1 scenario, depicting head-on collisions and rear-end collisions with surrounding vehicles in the top and bottom rows, respectively. Fig. 5 illustrates the generated trajectory of the ego vehicle in a head-on collision with a pedestrian and the trajectory of the ego vehicle side-impacting surrounding vehicles in the USA Roundabout FT scenario.

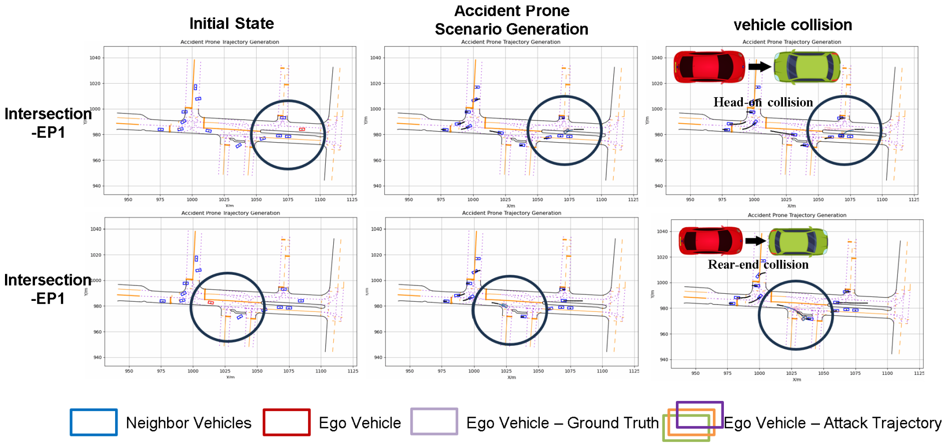

FIGURE 4. Generated Trajectory in Intersection Scenario

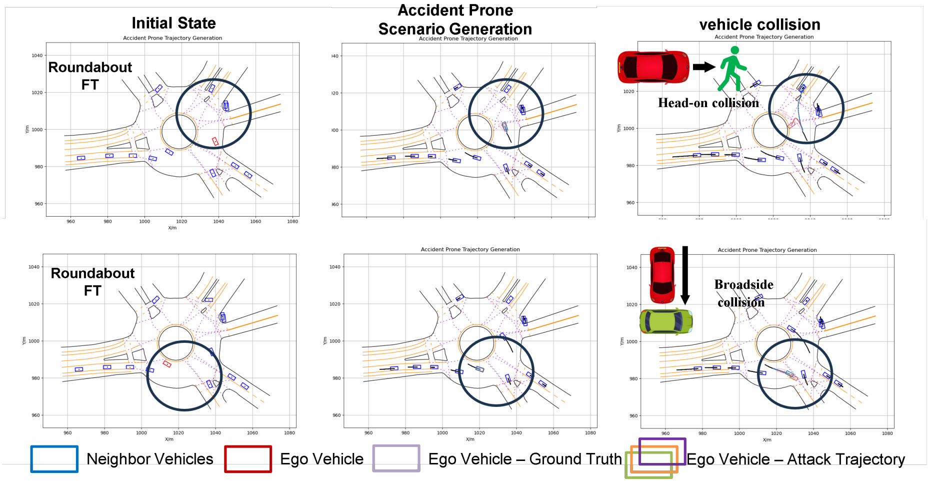

FIGURE 5. Generated Trajectory in Roundabout Scenario

6. DISCUSSION

By integrating the SocialVAE architecture into STRIVE's framework, our project successfully generated complex traffic accident scenarios for testing AV planners. SocialVAE's timewise latent variable generation and attention mechanism achieved good fidelity of accident-prone scenarios. This integration brings forth more challenging scenarios that better reflect real-world traffic interactions, allowing rigorous evaluation and refinement. Good generation outcome using simplified structure and dataset also provide promising potential for implementing the SocialVAE approach with complete settings as STRIVE. This project enables autonomous vehicle systems to address a broader spectrum of potential risks, improving their robustness and safety.

Further explorations for this project includes: (1) **Quantitative analysis and comparison.** Although our experiments produced promising collision trajectories, further quantitative evaluation of the generation process and trajectory quality is still needed, especially compared to the results in [8, 1, 5]. Future research could employ the same datasets as these papers, such as using rule-based planners and larger network. (2) **Constraint on trajectories to improve collision rate.** During the generation process, some trajectories didn't result in effective collisions due to the low collision likelihood in the original scenarios and partly because some trajectories prematurely terminate due to distortion before a collision occurs. By designing constraints, we can reduce the manual screening required after generation. (3) **More diverse and complex scenarios.** In real-world traffic scenarios, significant numbers of non-motorized participants (e.g., cyclists, pedestrians) also demand careful attention from autonomous driving planners. Future research can incorporate these elements, like temporary changes in road topology, to better refine planners' performance under extreme conditions.

References

[1] Yasasa Abeysirigoonawardena, Florian Shkurti, and Gregory Dudek. Generating adversarial driving scenarios in high-fidelity simulators. pages 8271–8277, 05 2019.

[2] 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.

[3] Luca Bergamini, Yawei Ye, Oliver Scheel, Long Chen, Chih Hu, Luca Del Pero, Błazej Osiński, 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.

[4] 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.

[5] Wenhao Ding, Baiming Chen, Bo Li, Kim Ji Eun, and Ding Zhao. Multimodal safety-critical scenarios generation for decision-making algorithms evaluation. *IEEE Robotics and Automation Letters*, 6(2):1551–1558, 2021.

[6] 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.

[7] 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.

[8] 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.

[9] Jingkang Wang, Ava Pun, James Tu, Sivabalan Manivasagam, Abbas Sadat, Sergio Casas, Mengye Ren, and Raquel Urtasun. Advsim: Generating safety-critical scenarios for self-driving vehicles, 2023.

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

[11] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kümmerle, Hendrik Königshof, 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.