2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Reinforcement Learning with Human Feedback for Realistic Traffic Simulation

Yulong Cao[1] Boris Ivanovic[1] Chaowei Xiao[1,2] Marco Pavone[1,3]

***Abstract*— In light of the challenges and costs of real-world testing, autonomous vehicle developers often rely on testing in simulation for the creation of reliable systems. A key element of effective simulation is the incorporation of realistic traffic models that align with human knowledge, an aspect that has proven challenging due to the need to balance realism and diversity. Towards this end, in this work we develop a framework that employs reinforcement learning from human feedback (RLHF) to enhance the realism of existing traffic models. This work also identifies two main challenges: capturing the nuances of human preferences on realism and unifying diverse traffic simulation models. To tackle these issues, we propose using human feedback for alignment and employ RLHF due to its sample efficiency. We also introduce the first dataset for realism alignment in traffic modeling to support such research. Our framework, named TrafficRLHF, demonstrates its proficiency in generating realistic traffic scenarios that are well-aligned with human preferences through comprehensive evaluations on the nuScenes dataset.**

I. Introduction

Due to the significant expenses and risks associated with conducting large-scale real-world tests [1], autonomous vehicle (AV) developers rely heavily on comprehensive testing in simulation to ensure the development of reliable systems [2]. To maximize the effectiveness of simulators, it is crucial for them to offer *realism* that is aligned with human knowledge. Accordingly, *realistic* traffic models are essential to ensure that insights gained from simulation testing apply seamlessly to real-world scenarios [3], [4]. However, developing traffic models that balance realism and diversity remains an ongoing challenge. To tackle this challenge, we leverage recent advancements in alignment research for traffic modeling; moreover, such alignment is model-agnostic and can improve various existing traffic models. Consequently, our primary research objective is to develop a framework based on reinforcement learning with human feedback (RLHF) to widely improve the realism of existing traffic models.

In order to align human preferences for generating realistic traffic simulations, two significant challenges need to be addressed: (1) the limited expressiveness of existing methods in capturing human preferences for realism in traffic simulations, and (2) unifying diverse traffic simulation models. The first challenge relates to traditional approaches relying on predefined rules or statistical models that fail to encompass the wide range of human preferences and subjective perceptions of realistic traffic scenarios. For example, consider a vehicle temporarily merging left into an oncoming traffic lane to nudge around a stopped vehicle. This is realistic behavior (especially on smaller two-lane roads), but a violation of commonly-handcrafted off-lane driving heuristics. The second challenge relates to the wide variety of traffic simulation models, each with their own strengths and limitations, underlying assumptions, data requirements, and simulation algorithms, making it challenging to incorporate human preferences seamlessly.

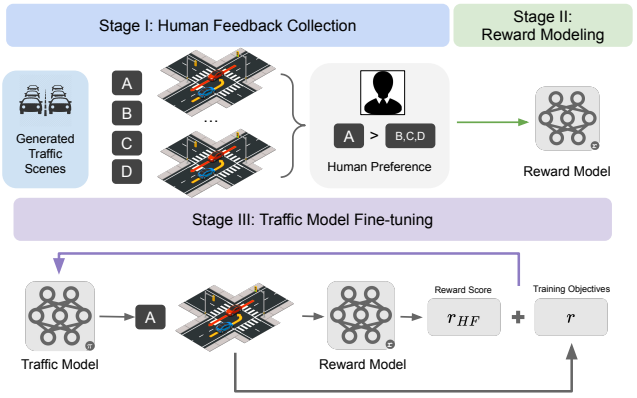


Fig. 1: An overview of our 3-stage TrafficRLHF approach. Stage I entails the collection of human feedback on traffic scenarios to train a reward model. In Stage II, this reward model is utilized to rank the realism of traffic scenarios. Lastly, Stage III involves the fine-tuning of the traffic models based on the reward model, thereby improving the realism of the generated traffic scenarios.

To address these challenges, we introduce TrafficRLHF. Visualized in Fig. 1, it is a three-stage framework utilizing human feedback to enhance traffic models for realistic scenario generation. The framework tackles key challenges in the field of traffic simulation, leveraging recent advancements in reinforcement learning with human feedback (RLHF) and a state-of-the-art autoregressive backbone model.

To tackle the first challenge—capturing human preferences on realism—we propose using human feedback as a tool for alignment. RLHF has recently demonstrated its effectiveness in conjunction with large language models. Its impressive sample efficiency makes it a prime candidate for realism alignment in traffic models. We will further show that, through the labeling of a modest quantity of data with human preferences, one can establish a reward model suitable for refining a wide range of traffic models. Addressing the sec-

[1]Yulong Cao and Boris Ivanovic are with NVIDIA Research {yulongc,bivanovic}@nvidia.com
[2]Chaowei Xiao is with School of Computer, Data & Information Sciences, University of Wisconsin-Madison, and with NVIDIA Research chaoweix@nvidia.com
[3]Marco Pavone is with Department of Aeronautics and Astronautics, Stanford University, and with NVIDIA Research {pavone@stanford.edu,mpavone@nvidia.com}

ond challenge—the need to unify diverse traffic simulation models—requires a method that can be applied across diverse model architectures. This is achieved using our RLHF-based method, which only requires a universal input interface for a reward model. To fulfill this requirement, we employ a state-of-the-art autoregressive backbone model, CTG [5], with a preferred roll-out length to optimize simulating the near future. Experiments on the real-world nuScenes autonomous driving dataset [6] confirm TrafficRLHF's ability to generate realistic trajectories that align with human preferences. As one result, TrafficRLHF is able to reduce unrealistic collisions or off-road driving by up to 80%.

Contributions. The main contributions of our work are threefold: **(1)** We introduce the first dataset designed for realism alignment in traffic modeling, **(2)** we formulate a reward model that quantifies realism according to human preferences, and **(3)** we propose a model-agnostic framework, leveraging RLHF, that effectively enhances the realism of a broad range of existing traffic models.

II. Related Work

Traffic simulation. Traffic simulation techniques can be broadly categorized into rule-based and learning-based approaches. Rule-based methods, employing analytical models such as cellular automata and intelligent driver models [7], tend to set fixed routes for vehicles and separate longitudinal from lateral agent movements. This inflexibility restricts their capacity to mimic the variety of real-world driving behaviors.

On the other hand, learning-based approaches utilize deep generative models trained on trajectory datasets to mimic real-world driving behaviors [8], [9], [10], [3], [4]. However, they often lack the flexibility for users to dictate specifications for generated traffic behaviors during inference, thereby falling short in producing *desired* traffic scenarios.

A parallel line of work focuses on generating adversarial or safety-critical scenarios, crafting trajectories that provoke AV misbehavior [11], [12], [13], [14]. Methods like STRIVE [15], CTG [5], and TrafficGen [16] have made strides in this direction, but they struggle to generate traffic scenarios aligned with human perceptions of realism.

Despite advancements in traffic simulation methodologies, a significant gap persists in the field: the ability to produce traffic scenarios that closely align with human perceptions of realism. Accordingly, a key focus of our work is integrating human preferences seamlessly into traffic simulation models.

Reinforcement Learning from Human Feedback. Research into Reinforcement Learning from Human Feedback (RLHF) spans at least a decade, with significant contributions from numerous researchers [17], [18], [19], [20], [21]. RLHF is a component of a larger paradigm known as the *human-in-the-loop learning process* [22]. In scenarios where reward engineering for RL proves challenging or costly, human feedback data emerges as a crucial asset.

Recent advances in the field have seen the use of human feedback for fine-tuning LLMs [23], [21], [24]. A notable example is InstructGPT [25], which leverages both human demonstrations and preferences to achieve significant improvements over the GPT-3 baseline in terms of preferences from human annotators. This demonstrates the scalability and utility of RLHF for tuning large models with human feedback.

Our proposed RLHF approach offers unique advantages for generating realistic traffic scenarios with human feedback, particularly in terms of realism which are subjective and lack explicit formulations. This approach minimizes data collection and feedback time, avoiding the need for costly labeled data.

III. TrafficRLHF

In this section, we formally specify the traffic simulation problem and delineate the three core components of TrafficRLHF. As shown in Fig. 1, these components include: data collection, comprised of accumulating human preferences for realism using a blend of genuine traffic scenes and those generated from traffic models; reward model training, which entails training a model capable of evaluating the realism of a traffic scenario based on human preferences; and fine-tuning, wherein we harmonize the generative model with the reward model trained in the previous stage. Ultimately, this methodology enables a wide array of traffic models to generate more authentic traffic scenarios.

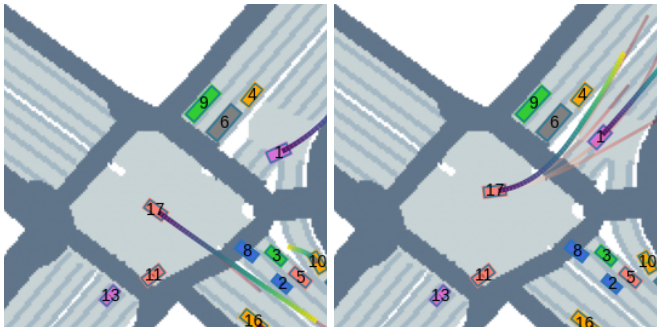
A. Traffic Simulation Formulation

We formulate the problem of traffic simulation similar to [5]. For a scenario with M vehicles, their state (comprised of x, y position, speed, and yaw) at timestep t is represented as $s_t = [s_t^1, ..., s_t^M]$, where $s_t^i = (x_t^i, y_t^i, v_t^i, \theta_t^i)$, and their action (*i.e.*, control) as $a_t = [a_t^1, ..., a_t^M]$, where $a_t^i = (\dot{v}t, \dot{\theta}t)$ (acceleration and yaw rate). We designate $\mathbf{c} = (I, s_{t-T_{\text{hist}}:t+1})$ as decision-relevant context, comprising local semantic maps for all agents $I = \{I^1, ..., I^M\}$, as well as their current and T_{hist} preceding states $s_{t-T_{\text{hist}}:t+1} = s_{t-T_{\text{hist}}}, \ldots, s_t$. The state s_{t+1} at time $t+1$ is derived from a transition function f that computes $s_{t+1} = f(s_t, a_t)$ given the previous state s_t and control a_t following unicycle dynamics.

B. Human Feedback Collection

Our proposed framework TrafficRLHF aligns generated traffic scenarios with human preferences for realism. It requires only a moderate amount of trajectory data collection and minimal human feedback. Since all trajectories are collected in simulation using existing traffic models, a considerable amount of data can be generated and enhanced iteratively. Furthermore, any human with driving experience can serve as a labeler after a brief familiarization with the user interface, eliminating the need for specialized expertise.

To garner human feedback, pairs of traffic scenarios (S_1, S_2) are generated from the same initial context $\mathbf{c}$. Each scenario $S_i = [s_{t-T_{\text{hist}}:t+T_{\text{fut}}}]_i$, where $i = 1, 2$, represents stacked state sequences of length T_{fut}. Human annotators are shown videos of these paired traffic scenarios, with maps and vehicles appropriately labeled. The annotators are then tasked with specifying which scenario appears more realistic.



(a) Driving straight (b) Turning left

Fig. 2: Illustration of multi-modal futures in generated traffic scenarios. It is considered reasonable for *vehicle 17* to either go straight (a) or turn left (b). Directly comparing such scenarios could be infeasible during the human feedback collection process.

As current traffic models often yield less realistic data, we increase the total amount of generated traffic scenarios to boost the likelihood of generating at least one realistic scenario. Now, since the generated scenarios usually comprise multi-modal futures, they may not always be directly comparable (e.g., one scenario has a vehicle turning right, whereas in another the vehicle moves straight). Accordingly, we ask labelers to select the most realistic example instead of conducting a pairwise comparison. This coarse-grained labeling method (compared to pairwise preference labeling) also reduces human bias towards specific scenarios.

Concretely, we generate N traffic scenarios $S_1, \ldots, S_N$ and request human annotators to select the most realistic one from the set. This process results in $N-1$ scenario pairs with preference relationships $(S_i \succ S_j)$, where S_i is the scenario chosen by the human annotator and S_j is any other scenario. In cases where the traffic model fails to generate realistic scenarios, annotators are provided the option to declare that none of the presented scenarios appear realistic. If this option is selected, we will use the ground truth scenario S_{gt} as the preferred scenario, denoted as $(S_{\text{gt}} \succ S_j)$.

C. Reward Modeling

The training methodology for the reward model $r_{\text{HF}}(\cdot;\phi)$ is inspired by Abramson et al. [26]. Two traffic scenarios, generated under the same initial conditions, are compared and the reward model is trained with the following loss:

$$l_{\text{RM}} = \mathbb{E}_{S_1,S_2\sim\mathcal{D}}[-\log\sigma(r_{\text{HF}}(S_1;\phi) - r_{\text{HF}}(S_2;\phi))], S_1 \succ S_2. \quad (1)$$

Here, $r_{\text{HF}}(S_i;\phi)$ represents the reward model, which can be transformed from existing trajectory prediction models at ease. The training dataset $\mathcal{D}$ is compiled using human preferences over traffic scenarios, generated with the existing traffic models (without any improvements). And $S_1, S_2 \sim \mathcal{D}$ is a pair of scenarios sampled from the human-labeled dataset. Note that r_{HF} can be either an iterative model or an encoder-decoder model taking a fixed-length sequence. To maximize TrafficRLHF's ability to adapt a variety of traffic models during the final fine-tuning stage, we choose an iterative model that can take as input an arbitrary sequence length. To calculate the reward, we compute the average of the reward over a sequence S_i.

D. Traffic Model Fine-tuning

Similar to the fine-tuning method proposed by Abramson et al. [26], our objective is:

$$\tilde{r} = r + \alpha \cdot r_{\text{HF}}(S),$$

where r is the original loss for training the provided traffic model and α is a scaling term. This mixed objective is used to avoid overfitting to the reward score and degrading task performance. In RLHF for training large language models (LLMs, e.g., ChatGPT [27]), the majority of the model's parameters are usually frozen for computational efficiency. However, existing traffic models are usually much smaller than LLMs. In our experiments, we will show that fine-tuning the entire motion prediction model can lead to higher performance improvements.

IV. Experiments

We conduct experiments to validate that (1) TrafficRLHF can generate more realistic traffic behaviors, and (2) the learned reward model is generalizable and can be reused for a diverse range of traffic models.

A. Experimental Setup

Datasets. The nuScenes dataset [6] contain 5.5 hours of human-annotated vehicle trajectories in two unique urban settings. It captures a myriad of driving scenarios, including instances of high traffic density. In our study, we use mainly the training split of the nuScenes dataset for data collection, reward model training, and traffic model fine-tuning. Evaluation of the traffic models' performance is carried out on a randomly-sampled 100-scene subset of the validation split.

Metrics. Following [5], [4], we evaluate the *stability* (*i.e.*, avoiding collisions and off-road driving) and *realism* of generated trajectories from each model. We evaluate stability by reporting overall failure rate (**fail**), measured as the average percentage of agents encountering a critical failure (*i.e.*, a collision or road departure) in a scene, which can be automatically detected with a set of predefined rules. To assess realism, we compare motion distributions between generated traffic simulations and ground truth trajectories from the dataset by calculating the Wasserstein distance between normalized histograms of their motion profiles. We measure realism using comfort as a proxy (**real**), which is the average of realism values for three properties: longitudinal acceleration magnitude, lateral acceleration magnitude, and jerk. Also, since the reward model naturally captured the human preference and we measure the human preference score of a scene using the reward cost score (**reward cost**) from the reward model.

Traffic Model Benchmarks. Recent advancements in the field of traffic generation include three particularly noteworthy models. First, Conditional Traffic Generation (**CTG**) is a model that leverages conditional diffusion to generate

simulations that match user specifications [5]. Second, the Bi-Level Imitation Learning System (**BITS**) is a model that utilises a two-tiered imitation learning strategy to approach traffic prediction tasks [4]. Lastly, **TrafficGen** is an autoregressive generative model that leverages an encoder-decoder architecture, enabling the diverse initialization of traffic scenarios [16].

As mentioned in Section III-D, TrafficRLHF uses pre-trained models for fine-tuning. Given that both CTG and BITS were originally trained on the nuScenes dataset [6], we use their original parameters. For TrafficGen, we reproduced its performance on nuScenes using the original implementation provided by the authors.

B. Experimental Details

Traffic Generation and Data Collection. To generate traffic scenarios, we utilize the state-of-the-art Conditional Traffic Generation (CTG) model [5]. CTG is built with an encoder for traffic context and a diffusion model for traffic generation. In particular, CTG guides the diffusion process with user specifications, enabling the sampling of trajectories from an unconditional diffusion model to achieve a specific preset objective, yielding controllable scenario generation. For data collection, we use CTG to generate traffic scenarios, operating under the guidance of a "no collision" principle, a feature that is evident in the output scenarios. We sample 500 scenes from the nuScenes training split [6] and process each of them with CTG to generate five unique scenarios. To ascertain human preferences, we present these five scenarios to labelers, tasking them with identifying the most realistic scenario (indicating "none" if none are realistic). While the only requirements for the labelers are driving experience, we select 5 experts for the simplicity and additional training costs to collect such dataset. Subsequently, we construct our human preference dataset by pairing either the most realistic sample (or the ground truth if no sample was realistic) with another sample. This curated dataset is then used to train the reward model, with the loss calculated according to Eq. (1).

Using our human feedback collection method, we observe that bias appears to be limited: 96.6% of scenarios were selected by at least four out of five human evaluators. Upon manual review of the inconsistent examples, we found that the presence of multi-modal futures was a common factor.

Reward Model and Training. In our experiments, the Reward Model (RM) is parameterized using the CTG encoder and fully-connected neural networks for the output layer. We chose CTG's encoder for the RM as it performantly encodes traffic scenarios. Furthermore, the simplicity of the CTG encoder facilitates a straightforward training schema, thereby reducing overhead when compared to other models that incorporate more complex components, such as RL policy networks [4], [9]. Note that the choice of reward model architecture is not restricted to our current selection and can be explored in future work.

Traffic Simulation. Following Zhong et al. [5] in CTG, to perform closed-loop traffic simulation of a scene, our model is applied for all agents in a standard control loop. At each step, the model generates a guided trajectory and executes the initial few actions before re-planning at a set frequency. In all our experiments, each scene is rolled out for 10 seconds, starting from a ground truth driving log, and re-planned at a frequency of 2 Hz.

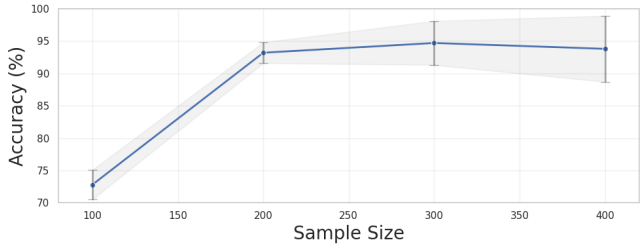


Fig. 3: Reward model accuracy quickly improves with more data, reaching a high plateau after only 200 data samples.

C. Reward Model Validation

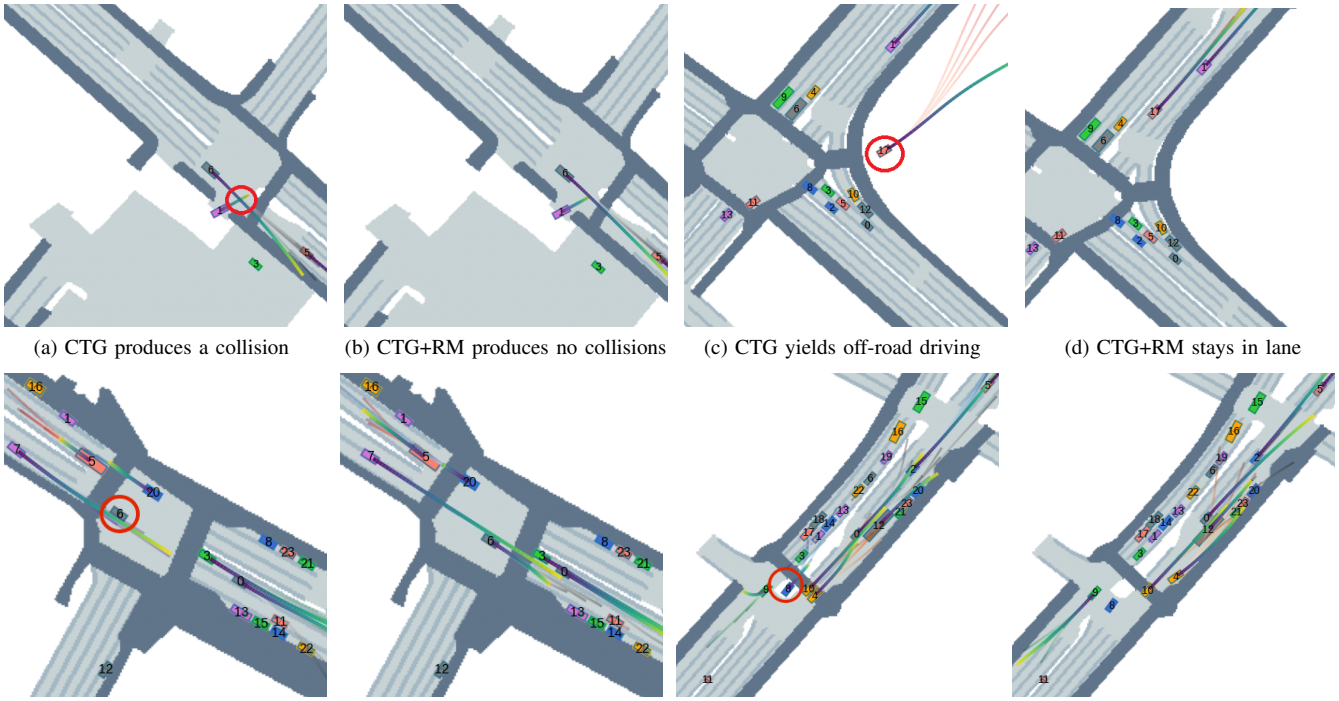
Before proceeding to fine-tune the traffic model, we first assess the RM's performance on the collected human preference data. This evaluation involves comparing the reward scores with human preferences. For the evaluation, we partition the collected human preferences into a training subset and a validation subset, comprising 400 and 100 scenes, respectively. As shown in Fig. 3, the accuracy of the RM rapidly increases with growing sample sizes. However, the variance of the accuracy tends to broaden with further increments. This increasing variance may indicate overfitting during the training process [28], since the nuScenes dataset is relatively small in size (totaling 850 scenes). To optimize performance, we deploy the RM trained with 200 samples for fine-tuning.

D. Evaluation of Traffic Model

CTG+RM. Having evaluated the learned RM, we use it to fine-tune the traffic model as in Section III-D. As depicted in Fig. 4, fine-tuning with the RM significantly enhances realism by decreasing failure rates (collisions and driving off-road). Furthermore, as illustrated in Figs. 4f and 4g, fine-tuning using our RM also significantly improves realism in nuanced cases that involve vehicles errantly stopping.

For quantitative results, since CTG can guide its generation process using user specifications, we evaluate the model with varying output specifications as in the original authors' setup. As shown in Table I, fine-tuning yields reduced failure rates in the the generated traffic scenarios. We observe that the realism metrics deteriorate when guidance is related to the goal positions/speeds, possibly because the guidance in the diffusion process of CTG overpowers the loss provided by the RM penalty losses.

Generality of Learned RM. To highlight the versatility of our trained RM in generalizing and enhancing a variety of models, we apply fine-tuning to two recent traffic models—BITS [4] and TrafficGen [16]—using the learned RM. Given the complex designs of BITS and TrafficGen, we restrict fine-tuning to the motion forecasting components, leaving other model parameters static. For BITS, we only



(a) CTG produces a collision (b) CTG+RM produces no collisions (c) CTG yields off-road driving (d) CTG+RM stays in lane

(e) CTG stops in the middle of road (f) CTG+RM produces flowing traffic (g) CTG ignores boundaries (h) CTG+RM stops at boundaries

Fig. 4: CTG [5] fine-tuned with our reward model (+ RM) avoids unrealistic collisions, off-road conditions, and stops. Unrealistic generated behaviors are marked with red circles.

TABLE I: Fine-tuning with our reward model significantly improves CTG's output quality. Bolded is best.

	Speed limit		**Target speed**		**No collision**		**No off-road**		**Goal waypoint**	
	Real ↓	Fail ↓	Real ↓	Fail ↓	Real ↓	Fail ↓	Real ↓	Fail ↓	Real ↓	Fail ↓
CTG [5]	**0.36**	0.17	0.86	0.18	0.57	0.27	0.50	0.46	**0.56**	0.39
+ TrafficRLHF (Ours)	0.36	**0.15**	**0.73**	**0.16**	**0.38**	**0.05**	**0.39**	**0.21**	0.73	**0.34**

TABLE II: TrafficRLHF improves the performance of many traffic models.

Model	**Real ↓**	**Fail ↓**	**Cost↓**
CTG [5]	0.57	0.27	13.21
+ TrafficRLHF (Ours)	**0.38**	**0.05**	**3.70**
BITS [4]	1.22	0.31	12.40
+ TrafficRLHF (Ours)	**0.83**	**0.23**	**9.50**
TrafficGen [16]	1.68	0.39	15.20
+ TrafficRLHF (Ours)	**1.17**	**0.22**	**11.30**

adjust the spatial goal network, while the parameters for the goal-conditional policy remain unaltered. Though the goal-conditional policy network determines the actions for each agent within the traffic simulation, employing the reward model for policy network fine-tuning necessitates a different strategy, potentially resulting in an uneven comparison. Thus, we opt to fine-tune solely the spatial goal network, which also plays a critical role in guiding the traffic simulation. For TrafficGen, we only fine-tune the motion forecasting components. Notably, TrafficGen also employs a traffic initialization network. For a fair comparison, we maintain a fixed initialization during the fine-tuning process.

As depicted in Fig. 5, fine-tuning with our RM decreases the frequency of collisions and off-road driving in both

TABLE III: Performance of different fine-tuning strategies.

	Real ↓	**Fail ↓**	**Cost↓**
CTG [5]	0.57	0.27	13.21
CTG [5] + RM (encoder)	0.43	0.18	9.10
CTG [5] + RM (decoder)	0.39	0.10	5.20
CTG [5] + RM (full)	**0.38**	**0.05**	**3.70**

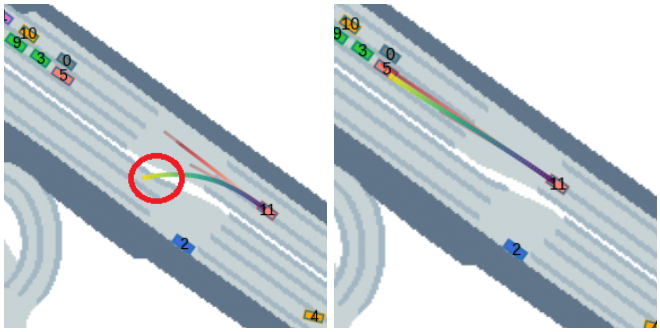
models. We also present quantitative evaluations in Table II. When fine-tuned with the reward model—which was trained on CTG-generated traffic scenarios—there were measurable improvements across all three metrics for both BITS and TrafficGen. Although the degree of improvement is not as substantial as that seen with CTG, it underscores the broad applicability of our proposed method.

E. Ablation Studies

In our research, we delve further into the performance capabilities of our proposed method by executing an ablation study involving the fine-tuning of the CTG model.

Fine-tuning components. We experiment with fine-tuning specific components of the model and present the results in Table III. First, we tweak the encoder portion of the CTG model, incorporating the reward model (RM) in the process. This step leads to substantial enhancements across all performance measures, suggesting the pivotal role of

(a) BITS produces a collision (b) BITS+RM has no collisions



(c) TrafficGen exits lanes (d) TrafficGen+RM stays in lane

Fig. 5: BITS [4] and TrafficGen [16] fine-tuned with our reward model (+ RM) avoid unrealistic behaviors. Unrealistic generated behaviors are marked with red circles.

human feedback, captured via the reward model, in refining the model's ability to generate realistic traffic scenarios. For our next variant, we focus on the decoder portion of the CTG model, integrating the reward model into it. Interestingly, this setup outperforms the previous variant, reinforcing the idea that the diffusing model (decoder) might play a more integral role in the generation of more realistic traffic scenarios. Lastly, we carry out an all-encompassing fine-tuning, involving both the encoder and decoder of the CTG model, while also adding the reward model. This configuration performs the best among all, suggesting the importance of simultaneous fine-tuning of all model components alongside the incorporation of human feedback via the reward model. All things considered, while full fine-tuning yields the most promising results, the decoder (diffusing model) appears to be more responsive to modifications, likely owing to its inherent role in CTG's mechanism for generating realistic traffic scenarios [5].

Human feedback. We experiment with training reward model using a diverse range of human feedback on traffic scenarios generated by different models. To do so, we repeated the process of collecting human feedback on generated traffic scenarios on both TrafficGen and BITS. The rest of the process, including reward model training and traffic model fine-tuning, remain the same. As the result in Table IV (highlighted in purple), we observe a slight improvement of realism for the fine-tuned model. This suggests that the need to amass a large dataset of human feedback from scenarios generated by diverse traffic models may be less critical. The model-agnostic nature of our method confers an advantage, requiring only moderate labeling efforts focused on scenarios generated from a single traffic model.

TABLE IV: Ablation study on different reward models (RM) and human feedback (HF) examples. Results for models that were fine-tuned with RM using the same backbone are highlighted in cyan. Results of models fine-tuned with the RM trained with HF examples from the same model is highlighted in purple.

Traffic Model	RM	HF	Real ↓	Fail ↓
TrafficGen	-	-	1.68	0.39
	CTG	CTG	1.17	0.22
	CTG	TrafficGen	0.92	0.18
	TrafficGen	CTG	**0.83**	**0.12**
BITS	-	-	1.22	0.31
	CTG	CTG	0.83	0.23
	CTG	BITS	**0.73**	**0.21**
	TrafficGen	CTG	0.81	0.25

Reward model. We also experiment with training reward model on different architectures. We use TrafficGen encoder as the backbone and different human feedback to train a reward model and fine-tune the other traffic models. We didn't use the BITS model due to its complicated architecture for enabling bi-level optimization. As the result shown in Table IV (highlighted in cyan), we observe a more significant improvements on realism compared to using human feedback on different models shown earlier. Based on the findings above that fine-tuning decoder being more effective, one hypothesis is that the latent representation for the traffic model and reward model should be similar and RLHF does not significantly help with learning better representations [27]. This underscores the significance of employing a Reward Model (RM) capable of learning robust and comprehensive representations. Similar observations have been made in other applications as well [27].

V. Conclusion

In conclusion, this work demonstrates the value of RLHF in developing more realistic traffic models. Our method, TrafficRLHF, addresses the identified challenges of capturing nuanced human preferences and unifying diverse traffic simulation models. By using human feedback for alignment, we harness the data efficiency of RLHF and use an autoregressive backbone model to provide a generalizable reward interface. Eperiments on the large-scale, real-world nuScenes dataset demonstrate that our resulting TrafficRLHF model can generate trajectories closely aligned with human preferences.

Our contributions not only provide the first dataset of realism alignment for traffic modeling, but also offer a versatile RLHF-based framework that enhances the realism of a wide range of existing traffic models. This lays a robust foundation for future work in the field and has the potential to significantly enhance the utility of traffic simulations for autonomous vehicle development.

REFERENCES

[1] N. Kalra and S. M. Paddock, *Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?* RAND Corporation, 2016. [Online]. Available: http://www.jstor.org/stable/10.7249/j.ctt1btc0xw

[2] Waymo, "Waymo safety report," https://storage.googleapis.com/waymo-uploads/files/documents/safety/2021-03-waymo-safety-report.pdf, February 2021.

[3] 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 (CVPR)*, June 2021, pp. 10 400–10 409.

[4] 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.

[5] 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.

[6] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[7] E. Brockfeld, R. D. Kühne, A. Skabardonis, and P. Wagner, "Toward benchmarking of microscopic traffic flow models," *Transportation Research Record*, vol. 1852, no. 1, pp. 124–129, 2003.

[8] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," in *Conference on Robot Learning (CoRL)*. PMLR, 2020, pp. 86–99.

[9] Y. Chen, B. Ivanovic, and M. Pavone, "Scept: Scene-consistent, policy-based trajectory predictions for planning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 17 103–17 112.

[10] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 683–700.

[11] 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.

[12] Y. Cao, C. Xiao, A. Anandkumar, D. Xu, and M. Pavone, "Advdo: Realistic adversarial attacks for trajectory prediction," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 36–52.

[13] Y. Abeysirigoonawardena, F. Shkurti, and G. Dudek, "Generating adversarial driving scenarios in high-fidelity simulators," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8271–8277.

[14] B. Chen, X. Chen, Q. Wu, and L. Li, "Adversarial evaluation of autonomous vehicles in lane-change scenarios," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2021.

[15] D. Rempe, J. Philion, L. J. Guibas, S. Fidler, and O. Litany, "Generating useful accident-prone driving scenarios via a learned traffic prior," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[16] L. Feng, Q. Li, Z. Peng, S. Tan, and B. Zhou, "Trafficgen: Learning to generate diverse and realistic traffic scenarios," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3567–3575.

[17] R. Akrour, M. Schoenauer, and M. Sebag, "Preference-based policy learning," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11*. Springer, 2011, pp. 12–27.

[18] ——, "April: Active preference learning-based reinforcement learning," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*. Springer, 2012, pp. 116–131.

[19] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz, "Policy shaping: Integrating human feedback with reinforcement learning," *Advances in neural information processing systems*, vol. 26, 2013.

[20] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[21] N. Jaques, A. Ghandeharioun, J. H. Shen, C. Ferguson, A. Lapedriza, N. Jones, S. Gu, and R. Picard, "Way off-policy batch deep reinforcement learning of implicit human preferences in dialog," *arXiv preprint arXiv:1907.00456*, 2019.

[22] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, 2022.

[23] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, 2019.

[24] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3008–3021, 2020.

[25] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[26] J. Abramson, A. Ahuja, F. Carnevale, P. Georgiev, A. Goldin, A. Hung, J. Landon, J. Lhotka, T. Lillicrap, A. Muldal *et al.*, "Improving multimodal interactive agents with reinforcement learning from human feedback," *arXiv preprint arXiv:2211.11602*, 2022.

[27] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities," Microsoft, Tech. Rep. MSR-TR-2023-8, February 2023. [Online]. Available: https://www.microsoft.com/en-us/research/publication/chatgpt-for-robotics-design-principles-and-model-abilities/

[28] L. Gao, J. Schulman, and J. Hilton, "Scaling laws for reward model overoptimization," 2022.