SHRED: Synthesizing Rule-based Environments for Driving

Micha Fauth[1] Long Nguyen[1] Bernhard Jaeger[1] Daniel Dauner[1]
Maximilian Igl[2] Andreas Geiger[1] Kashyap Chitta[1,2]
[1]University of Tübingen, Tübingen AI Center [2]NVIDIA Research

Abstract

We provide several simple, interpretable rule-based baselines for the 2025 Waymo Scenario Generation Challenge. The challenge addresses the task of generating realistic simulation scenarios, which are crucial for developing autonomous vehicles. Our best baseline, SHRED, achieves a high score of 0.624, notably close to the theoretical upper bound of 0.692, and ranking third on the leaderboard. Our results demonstrate the strong performance of simple methods and help contextualize more complex approaches.

1. Introduction

Simulation has emerged as a crucial tool for scaling the development of autonomous driving systems. It enables fast, repeatable, and cost-effective evaluation of self-driving policies [1]. However, for simulation-based evaluation to provide meaningful insights, the simulation needs to be realistic, repeatable (low variance) and large-scale.

The Waymo Scenario Generation Challenge addresses this need by tasking participants with developing generative models that, given a map and the number agents involved in the scene as inputs, propose multiple diverse agent placements and simulated behaviors. Given the data-driven nature of this task, the best performance is likely achieved by a learning-based approach. Nevertheless, rule-based methods provide an important baseline as they are interpretable and computationally efficient. More importantly, high performing rule-based baselines can also identify tasks and datasets for which the more complex learning-based methods are unnecessary. Thus, they can help the community to better understand which open problems have the highest real-world impact and are most pressing to address [3, 10].

In this work, we propose multiple simple baselines for the Waymo Scenario Generation Challenge. We also evaluate multiple approaches that use privileged information, such as the logged future on a given scenario. Together, these experiments allow us to better understand the dataset and the complex aggregate "Realism Meta Metric" based on which the challenge is scored. We find that our best simple baseline, SHRED, achieves a score of 0.624 on the validation set, while the upper bound is only 0.692 - putting into perspective other learning-based submissions to the leaderboard. Fig. 1 provides a comparison between ground truth and SHRED-generated scenarios.

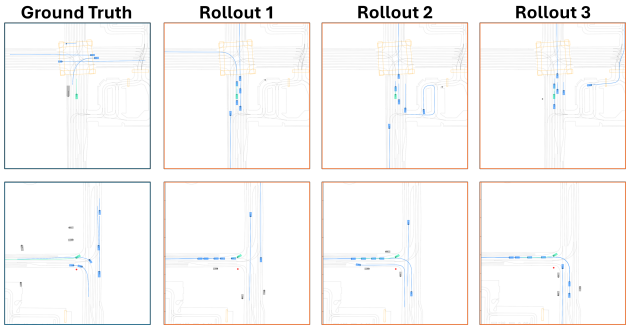

Figure 1. **Rule-based driving scenarios.** The first column shows the ground truth for two representative scenarios from the dataset. Subsequent columns present individual rollouts generated by our method for each scenario. Off-road objects are shown in gray, on-road objects in blue, and the autonomous driving vehicle in green.

We also find some limitations in the current evaluation metrics. In particular, the collision and off-road labels, the two largest contributors to the overall metric, are subject to significant label noise, which could incentivize conformity to metrics over optimizing scenario realism.

2. Method

As shown in Fig. 2, we decompose the problem into two steps: (1) generating initial states, and (2) simulating the behavior of the dynamic objects over the required horizon.

Following the rules of the Scenario Generation Challenge, we use the dynamic and static map features, past and current timestep of the ego vehicle, and the number of agents to generate for each object type (vehicle, cyclist, and pedestrians) as inputs. More specifically, from the map, we extract the polylines of road edges, center lines, as well as statistics like number of lanes and average speed limit.

To generate the initial states of all objects, we first es-

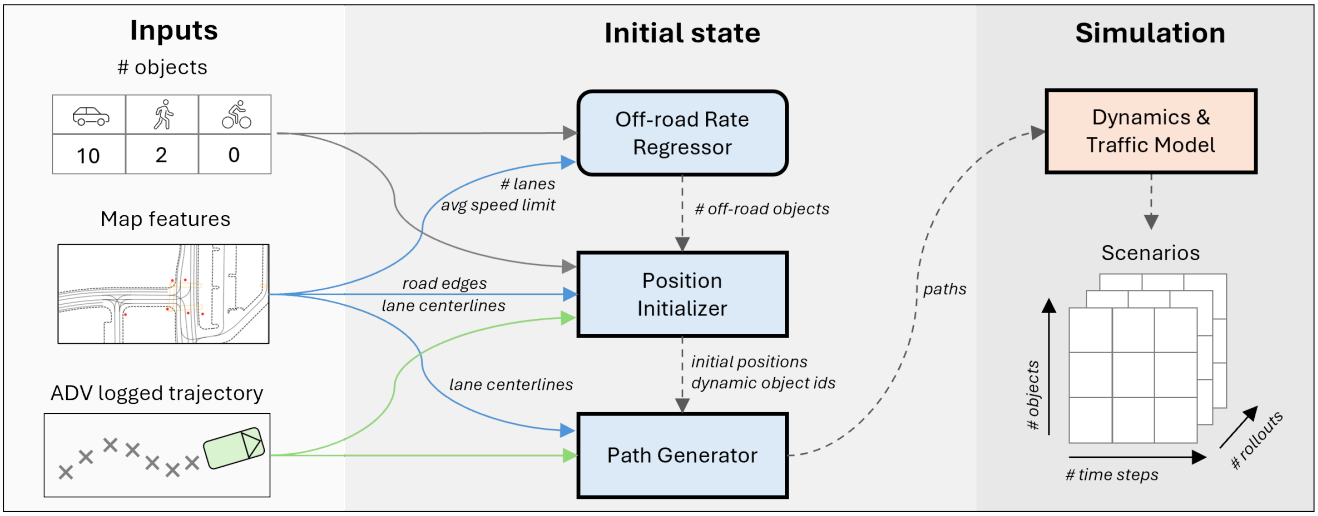

Figure 2. **SHRED**. Starting from the inputs specified by the challenge (left), we first regress an estimate of the number of off-road objects, then initialize positions in unoccupied areas and generate paths for all dynamic objects (center). Finally, a dynamics and traffic model is used for simulation, producing trajectories over 91 time steps (right).

timate the number of objects that are positioned offroad. Subsequently, we use rejection sampling to find valid initial locations for all objects. Lastly, we generate future paths for all non-stationary on-road agents.

Off-road Rate Regressor: Accurately predicting the number of off-road objects is a key factor in achieving a high overall score. We explore using a constant off-road ratio (e.g. 40% off-road), as well as using a simple 3-Layer Multi-Layer-Perceptron (MLP) regressor taking number of objects, number of lanes and the avg. speed limit as input.

Position Initializer: For placing objects, we make the simplifying assumptions that all pedestrians are placed off-road, all cyclists are placed on-road, and vehicles are placed such that the off-road ratio is satisfied. Furthermore, we assume that all off-road objects are static and all on-road objects are dynamic and controlled by the IDM traffic model. First, the autonomous driving vehicle (ADV) is initialized at its known location. Then, on-road objects are randomly placed on lane centerlines within a certain distance around the ADV using rejection sampling. A sampled location is rejected if it is within $d = {}^{\max(l_1,w_1)}/_2 + {}^{\max(l_2,w_2)}/_2$ of another object, where l_i, w_i are the length and width of object i. Velocities are initialized to that of the ADV for cars, and to a fixed value for cyclists. Off-road objects and on-road objects that failed to be placed successfully after n attempts are placed off-road with a similar procedure, with their velocities fixed at 0.

Path Generator: Similar to [2, 6], we generate and sample a fixed length path for each dynamic object along lane centerlines starting from the object's initial position. For the ADV, which follows the logged trajectory (available as a task input) for the first 11 timesteps, we instead identify the closest point on a lane centerline immediately after this interval. The rest of the path is then constructed from this point onward along the lane centerline.

IDM Traffic Model: For the simulation of dynamic objects, we modify the Intelligent Driver Model (IDM) [9] available in the Waymax simulator [5]. It follows the centerline paths and controls the acceleration of vehicles by taking into account a target speed, as well as their distance and relative velocity to leading objects, hence promoting collision avoidance. To prevent traffic light violations, we additionally override acceleration outputs when a vehicle approaches a red traffic light.

3. Experiments

This section details the dataset, evaluation metrics, experimental setup, and our results.

Dataset: We use traffic scenarios from the Waymo Open Motion Dataset [4]. This dataset includes static map features, traffic signal states, and object trajectories. The training and validation splits contain 91 time steps at 10 Hz, covering past, present, and future states, whereas the test split includes only the past and current time steps (11 steps at 10 Hz). To comply with the rules of the Scenario Generation Challenge, we remove all past and future states of all objects except for the ADV's past and current states.

Evaluation Split: To reduce computational costs, we report results on *miniVal*, a subset of 430 samples from the validation set, i.e. approximately one percent of the total validation data. This subset provides a representative sample of the full validation set, as confirmed by the comparable results obtained from the official Waymo evaluation server.

Scenario Generation Metrics: The Scenario Generation

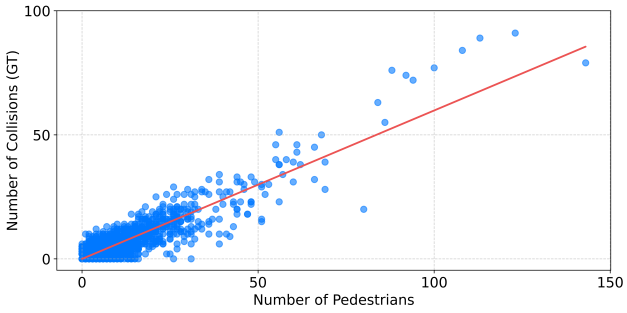

Figure 3. Correlation between the number of pedestrians to be simulated and the number of collisions in the ground truth.

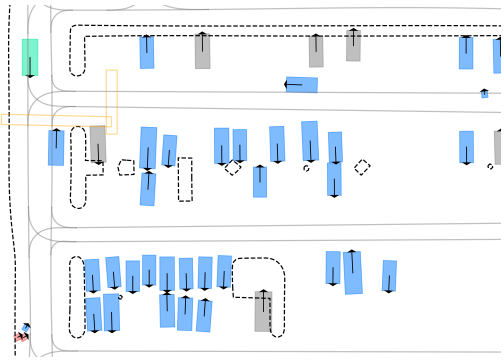

Figure 4. Example scenario where many static parked cars are classified as on-road in ground truth. Vehicles colored gray represent off-road objects, blue indicates on-road objects, red marks objects involved in collisions, and the ADV is shown in green.

Challenge employs a modified evaluation metric originally introduced in the Waymo Open Sim Agents Challenge [8]. Submissions are evaluated using the approximate negative log-likelihood (NLL) of the ground truth samples under the distribution of simulated scenario rollouts. This metric incentivizes scenario generation that more closely matches the *distribution* of the ground truth data. The final "Meta" metric weights and aggregates ten such distributional submetrics which can be grouped into three categories: i) **Kinematics:** linear speed (LINS) and acceleration (LINA), angular speed (ANGS) and acceleration (ANGA) ii) **Interactive:** collision indication (COLL), distance to nearest object (DTNO), time to collision (TTC) iii) **Map-based:** off-road indication (OFFR), traffic light violation (TLV), distance to road edge (DTRE). In contrast to the Sim Agents Challenge, the Scenario Generation Challenge evaluates all valid samples of all objects that need to be simulated, rather than a selected subset of objects.

Heuristics and Hyperparameters: We conduct an in-depth analysis on a subset of the training data, from which we derive several key insights used as heuristics in our method. Firstly, we find that about 60 percent of the objects are placed on-road and 40 percent off-road, which we also try as a fixed ratio for object placement. Pedestrians are predominantly found off-road in the dataset.

Secondly, we discover a strong correlation between the number of pedestrians and the number of collisions in the ground truth, as illustrated in Fig. 3. This motivates the implementation of a simple forced collision mechanism. Specifically, when the number of pedestrians exceeds four, we enforce collisions for every second pedestrian to reflect the likelihood of collisions seen in the real-world data. We also identify frequent cases of vehicle collisions in the ground truth, often caused by sensor noise or erroneous pedestrian detections inside vehicles.

Regarding static vehicles, we find no consistent pattern indicating whether such objects should be on- or off-road, as there are numerous examples like Fig. 4 in the dataset. However, given that most off-road objects are static, we choose to classify all off-road objects as static in our setup. To estimate the off-road object ratio more systematically, we train a multi-layer perceptron (MLP) regressor using four features: the number of vehicles, number of pedestrians, number of road lanes, and average speed limit. The model is trained on approximately 9,000 randomly sampled training scenarios and achieves a mean absolute error (MAE) of about 5.4. Additionally, we prioritize sampling from the ADV's lane and neighboring lanes before considering all lanes within the ADV's radius (50 meter), leading to improved DTNO submetric scores.

For object bounding box sizing, we simplify the process by computing the mean length, width, and height for each object type across the analyzed subset of training data and use these constant dimensions across all timesteps. For behavior modeling, we employ the default parameters of the IDM policy as implemented in the Waymax simulator. The only exception is the desired velocity, which is set to the speed limit of the ADV's lane.

Results: We evaluate multiple configurations of our method to understand the impact of different modifications (Table 1). For efficient evaluation, all experiments are conducted using a single rollout per scenario. However, we

ID	Position	Lat. Con.	Lon. Con.	Inter.	Map.	Kin.	META
A1	40/60	CV	CV	0.719	0.541	0.275	0.568
A2	40/60	Paths	IDM	0.705	0.530	0.484	0.599
A3	40/60	Paths	IDM w/ Lights	0.705	0.553	**0.491**	0.609
A4	40/60 w/ Col.	Paths	IDM w/ Lights	**0.723**	0.554	**0.491**	0.617
B1	MLP	Paths	IDM w/ Lights	0.705	**0.573**	0.490	0.616
B2	MLP w/ Col.	Paths	IDM w/ Lights	**0.723**	0.572	0.490	**0.624**

Table 1. **Ablation study.** Each row represents a different configuration of our method, varying in initial position handling, lateral control (Lat. Con.) and longitudinal control (Lon. Con.). META shows the aggregated performance score as a weighted sum of the interactive, map-based, and kinematic metrics.

Method	COLL	OFFR	DTNO	TTC	LINS	LINA	ANGS	ANGA	TLV	DTRE	META
	0.25	0.25	0.1	0.1	0.05	0.05	0.05	0.05	0.05	0.05	1.00
Logged oracle	0.921	0.625	0.331	0.887	0.416	0.552	0.638	0.727	0.994	0.353	0.692
Logged oracle first 3 frames	0.921	0.625	0.281	0.861	0.356	0.492	0.590	0.670	0.994	0.328	0.672
Logged avg. distributions	0.880	0.520	0.281	0.869	0.333	0.513	0.597	0.699	0.990	0.231	0.633
IDM logged trajectories	0.874	0.608	0.304	0.869	0.345	0.471	0.602	0.680	0.870	0.334	0.653
IDM logged trajectories w/ traffic lights	0.882	0.615	0.303	0.870	0.351	0.484	0.601	0.680	0.979	0.337	0.663
SHRED	0.860	0.565	0.254	0.850	0.326	0.434	0.556	0.645	0.970	0.214	0.624

Table 2. **Upper bounds**. This table presents an evaluation of several upper-bound baselines on the *miniVal* set and compares them with our method. The metametric is computed using the official challenge weights shown below each submetric's name.

observe a slight but consistent improvement in the metrics when evaluating 32 rollouts. As a baseline, we begin with a constant velocity (CV) model using a fixed initial distribution of 40% off-road and 60% on-road objects (A1). This configuration exhibits poor kinematic performance due to its simplistic motion model, which assumes a constant velocity along the initial heading. Moreover, it leads to off-road trajectories and collisions. Introducing more realistic longitudinal control via IDM (A2), aligned with lateral control by following lane centerlines, substantially improves kinematic scores. Further enhancing IDM with traffic light awareness leads to additional gains (A3), also in the map-based metrics. Interestingly, the constant velocity baseline still yields a higher interactive score. Upon closer analysis, we find that our approach tends to simulate fewer collisions than there are in the ground truth. To address this, we introduce a forced collision mechanism in pedestrian-heavy scenarios (A4). This increases the interactive score by approximately 0.02. Furthermore, we replace the fixed off-road rate initialization with a learned MLP-based regressor (B1). This leads to a notable improvement in the map-based metric, primarily due to higher scores in the OFFR metric. Combining the learned initialization with the forced collision mechanism (B2), we could achieve a META score of 0.624, our best overall performance on *miniVal* set.

Logged Oracles: In Table 2, we define 5 oracles for evaluation. The first oracle approximates an upper bound for *miniVal* by computing the Negative Log-Likelihood (NLL) of all valid ground truth samples under their corresponding ground truth distributions. The second oracle leverages the distribution derived from the first three ground truth frames as the model output and evaluates the NLL of the ground truth under this distribution. This highlights the critical role of scene initialization. Third, we utilize the average distributions computed across all scenarios in *miniVal* as a common output distribution for all scenarios and calculate the NLL of the ground truth samples accordingly. Lastly, we simulate the Waymax IDM policy following logged trajectories with and without traffic light awareness.

The most appropriate upper-bound comparison for our approach would be the results of IDM following logged trajectories with traffic light considerations. With a delta of 0.06, the largest notable performance drop appears in the OFFR metric, especially if we consider the relatively high weight assigned to this metric in the META score. This discrepancy suggests that our method, even when using an MLP-based regressor, struggles to accurately predict the correct number of off-road objects. We believe that this is a key area for improvement in future work.

We also observe a consistent performance gap across all kinematic metrics. We attribute this to our simplifying assumption that off-road objects are static while on-road objects are dynamic. Due to inaccuracies in off-road prediction, as well as static on-road parked cars in the ground truth, these assumptions likely lead to lower kinematic scores. Comparing the interactive metrics, our COLL score almost aligns with the score from the IDM upper bound. Although the map-based metrics of the IDM upper bound closely match those of the general upper bound (logged oracle), noticeable differences remain in the interactive and kinematic metrics. The latter may be mitigated through hyperparameter tuning of the IDM. However, the gap in the interactive metrics likely reflects a fundamental limitation of the IDM model (at least for meaningful parameterizations): its inability to capture the collision noise present in the ground-truth data, as IDM is inherently designed to avoid collisions whenever possible.

4. Conclusion

We introduce SHRED, a simple and mainly rule-based baseline for the Waymo Scenario Generation Challenge 2025. Despite its simplicity, SHRED achieves competitive leaderboard scores. We also identify some limitations of the current evaluation metrics. Labeled collisions in the ground truth likely are entirely noise and incentivize undesirable behaviors such as forced collisions. Similarly, the off-road label is not semantically meaningful and is particularly noisy for parked vehicles. These issues indicate that current metrics assess conformity to labeling statistics rather than the realism of generated scenarios. Future work should aim to address these issues, potentially through a two-stage evaluation framework as explored in [7].

Acknowledgments

Bernhard Jaeger and Andreas Geiger were supported by the ERC Starting Grant LEGO-3D (850533) and the DFG EXC number 2064/1 - project number 390727645. Daniel Dauner was supported by the German Federal Ministry for Economic Affairs and Climate Action within the project NXT GEN AI METHODS (19A23014S). We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Bernhard Jaeger, Daniel Dauner, and Kashyap Chitta.

References

[1] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024. 1

[2] Kashyap Chitta, Daniel Dauner, and Andreas Geiger. Sledge: Synthesizing driving environments with generative models and rule-based traffic, 2024. 2

[3] Daniel Dauner, Marcel Hallgarten, Andreas Geiger, and Kashyap Chitta. Parting with misconceptions about learning-based vehicle motion planning. 2023. 1

[4] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Benjamin Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aur'elien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *ICCV*, 2021. 2

[5] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, John D. Co-Reyes, Rishabh Agarwal, Rebecca Roelofs, Yao Lu, Nico Montali, Paul Mougin, Zoey Yang, Brandyn White, Aleksandra Faust, Rowan McAllister, Dragomir Anguelov, and Benjamin Sapp. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. In *NeurIPS*, 2023. 2

[6] Napat Karnchanachari, Dimitris Geromichalos, Kok Seang Tan, Nanxiang Li, Christopher Eriksen, Shakiba Yaghoubi, Noushin Mehdipour, Gianmarco Bernasconi, Whye Kit Fong, Yiluan Guo, and Holger Caesar. Towards learning-based planning:the nuplan benchmark for real-world autonomous driving, 2024. 2

[7] Reza Mahjourian, Rongbing Mu, Valerii Likhosherstov, Paul Mougin, Xiukun Huang, Joao Messias, and Shimon Whiteson. Unigen: Unified modeling of initial agent states and trajectories for generating autonomous driving scenarios, 2024. 4

[8] Nico Montali, John Lambert, Paul Mougin, Alex Kuefler, Nick Rhinehart, Michelle Li, Cole Gulino, Tristan Emrich, Zoey Yang, Shimon Whiteson, Brandyn White, and Dragomir Anguelov. The waymo open sim agents challenge. *arXiv.org*, 2305.12032, 2023. 3

[9] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62(2), 2000. 2

[10] Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv.org*, 2305.10430, 2023. 1