RoboTron-Sim: Improving Real-World Driving via Simulated Hard-Case

Supplementary Material

A. Experimental Details

A.1. Datasets

RoboTron-Sim is trained using a hybrid data strategy combining:

- **Real-world Data**: 28,130 samples from nuScenes[1].
- **Simulated Data**: 47,553 purpose-built samples from our Hard-case Augmented Synthetic Scenarios(HASS) dataset, generated in CARLA simulator[2], designed to address the inherent imbalance in real-world data distribution. While the dataset covers a broad range of driving situations, it places particular emphasis on addressing challenging cases, including H2D scenarios and Long-Tail scenarios. Partial results are illustrated in Figure S1.

A.2. Evaluation Metrics

Following BEV-Planner [5], we evaluate via L2 Distance, Collision Rate, and Boundary Violation Rate.

- **Trajectory Accuracy (L2 Distance)**:

$$L2 = \frac{1}{T}\sum_{t=1}^{T} \|\hat{p}_t - p_t^{gt}\|_2 \tag{1}$$

where $\hat{p}_t$ and p_t^{gt} denote the predicted and ground-truth positions at timestep t over a $T = 3s$ horizon.

- **Safety Metrics (Collision Rate)**:
Computes the percentage of predicted trajectories that result in collisions with other agents or obstacles.

$$Collision = \frac{1}{T}\sum_{t=1}^{T} \frac{N_{\text{collision},t}}{N_{\text{total},t}} \times 100\% \tag{2}$$

where $N_{\text{collision}}$ is the number of predicted trajectories leading to collisions, and N_{total} is the total number of evaluated trajectories at timestep t over a $T = 3s$ horizon.

- **Boundary Violation Rate**:

$$Boundary = \frac{1}{T}\sum_{t=1}^{T} \frac{N_{\text{violation},t}}{N_{\text{total},t}} \times 100\% \tag{3}$$

where $N_{\text{violation}}$ counts trajectories exceeding road boundaries, and N_{total} is the total evaluated trajectories at timestep t over $T = 3s$. Calculated by comparing ego segmentation masks with drivable area labels.

B. More Results

B.1. Robustness of HASS

We investigate the performance trend divergence between simulated data augmentation and real data-only scenarios across multiple orders of magnitude in real data volume in RoboTron-Sim, with quantitative comparisons presented in Table S1 and Table S2. Table S1 presents quantitative results with full simulated data integration, while Table S2 provides detailed metrics when trained without any simulated data, using real-world data exclusively. The experimental results demonstrate enhanced stability of overall performance through simulated data augmentation.

nuScenes	HASS	L2(m)	Collision(%)
0%	100%	1.24	0.99
10%	100%	0.87	0.89
25%	100%	0.67	0.64
50%	100%	0.63	0.57
75%	100%	0.61	0.54
100%	100%	0.56	0.58

Table S1. Performance variation with nuScenes blending ratio under full HASS integration.

nuScenes	HASS	L2(m)	Collision(%)
10%	0%	2.41	3.22
25%	0%	1.24	1.46
50%	0%	1.15	1.54
75%	0%	1.03	1.03
100%	0%	0.72	0.67

Table S2. Performance scaling with nuScenes blending ratio (HASS Excluded).

B.2. Effectiveness of HASS

We generate two distinct datasets based on nuScenes scenarios: General Augmented Synthetic Scenarios (GASS) for common driving conditions and Hard-case Augmented Synthetic Scenarios (HASS) for challenging situations, aiming to investigate which synthetic data generation mechanisms yield more meaningful performance improvements. The evaluation results categorized by individual scenarios are presented in Table S3, while the aggregated metrics for H2D scenarios (Night+Turn+Rain) are summarized in Table S4.

B.3. Model Generalization

To verify the model generalization in the **planning task**, we further evaluate model performance on the NAVSIM (NV) benchmark using the predictive driver model score (PDMS), which is based on five factors: no at-fault col-

Data	Day		Night		Straight		Turn		Sunny		Rainy	
	L2	Col	L2	Col	L2	Col	L2	Col	L2	Col	L2	Col
nuScenes	0.59	0.50	1.40	2.71	0.59	0.55	1.32	1.80	0.64	0.63	1.15	0.81
nuScenes + GASS	0.55	0.42	1.00	2.53	0.50	0.49	1.21	1.89	0.52	0.58	0.99	0.79
nuScenes + HASS	0.54	0.47	0.81	1.56	0.55	0.52	0.64	1.01	0.56	0.64	0.56	0.32

Table S3. Performance comparison across various training data in each scenario.

Training Data	E2D		H2D	
	L2 (m)	Collision (%)	L2 (m)	Collision (%)
nuScenes	0.61	0.56	1.29	1.77
nuScenes + GASS	0.52	0.50	1.07	1.74
nuScenes + HASS	0.55	0.54	0.67	0.96

Table S4. Performance comparison in H2D and E2D scenarios.

Method	Data	NC↑	DAC↑	TTC↑	Comf.↑	EP↑	PDMS↑
Human	-	100.0	100.0	100.0	99.9	87.5	94.8
Ego-MLP	NV	93.0	77.3	83.6	**100.0**	62.8	65.6
UniAD	NV	97.8	91.9	92.9	**100.0**	78.8	83.4
ParaDrive	NV	97.9	92.4	93.0	99.8	79.3	84.0
RoboTron-Sim	NV	98.0	93.0	93.3	99.8	79.9	84.6
RoboTron-Sim	NV+HASS	**98.2**	**93.6**	**93.8**	99.9	**81.1**	**85.6**

Table S5. Performance on NAVSIM benchmark, †indicates that RoboTron-Sim is trained without HASS.

lisions (NC), drivable area compliance (DAC), time-to-collision (TTC), comfort (Comf.), and ego progress (EP). Table S5 demonstrates that RoboTron-Sim delivers comparable or superior performance compared to existing methods, with the integration of HASS achieving a PDMS of 85.6 and setting a new SOTA result on NV benchmark.

We also explore the robustness of the model on the **VQA task**. The VQA data curated for HASS encompasses three categories of questions: (1) Descriptive questions, such as "What is the color of the traffic light ahead?", are answered directly using data generated by the simulator; (2) Hypothetical questions, such as "If you turn right at this intersection, what would you encounter?", are annotated using GPT-4o based on environment visuals and predefined rules; (3) Reasoning questions, such as "Why are you slowing down here?", are generated by GPT-4o based on driving videos and trajectories to enhance the understanding of vehicle behavior. We conduct separate validations on the BDD-X and LingoQA datasets. As shown in Table S6, with HASS integration, RoboTron-Sim achieves SOTA performance on both benchmarks (e.g., improving METEOR from 52.23 to 56.30 on BDD-X, and increasing CIDEr from 61.3 to 62.2 on LingoQA).

Method	Data	BLEU	METEOR	CIDEr
QwenVL	BDD-X	25.89	46.54	19.91
LLaVA-1.5	BDD-X	25.97	45.08	21.62
Senna	BDD-X	31.04	50.44	34.31
RoboTron-Sim	BDD-X	32.54	52.23	37.19
RoboTron-Sim	BDD-X+HASS	**33.25**	**56.30**	**38.17**
LLaVA	LingoQA	12.5	18.5	57.0
Vicuna-7B	LingoQA	10.1	15.2	51.0
BLIP-2	LingoQA	13.0	17.4	60.1
LingoQA	LingoQA	15.0	18.6	59.5
RoboTron-Sim	LingoQA	15.5	18.5	61.3
RoboTron-Sim	LingoQA+HASS	**16.6**	**19.0**	**62.2**

Table S6. Performance on NAVSIM benchmark, †indicates that RoboTron-Sim is trained without HASS.

B.4. Model Compatibility

We conduct a comparative analysis of three models: VAD [3] (representing classical end-to-end models), LLaVA-OneVision [4] (as a representative multimodal large language model), and our RoboTron-Sim, evaluating their performance gains when augmenting real-world data with simulated data. To systematically investigate model compatibility with simulated data augmentation, we conduct cross-architecture evaluations on L2 distance. As evidenced in Table S7, VAD exhibits fundamental compatibility limitations, with marginal L2 reductions (↓2.5% E2D, ↓1.1% H2D). Although MLLM demonstrates preliminary compatibility, showing gradual improvements, the gains remain constrained in the hard cases (↓9.0% E2D, ↓7.3% H2D). In stark contrast, our RoboTron-Sim achieves breakthrough enhancements (↓48.1%) in H2D case while maintaining stable performance in E2D case. This empowers knowledge transfer from synthetic domains while preserving real-world physical constraints, unlocking the model's untapped potential.

B.5. Deployment Costs

We compare the key deployment metrics for the models on RTX-4090, as shown in Table S8. It shows that RoboTron-Sim is applicable to smaller models like RoboTron-Sim-0.5B (replacing the LLM from Qwen2-7B to Qwen1.5-0.5B), achieving comparable performance to RoboTron-

Method	Data	L2 Distance (m)	
		E2D	H2D
VAD	nuScenes	0.78	0.88
	nuScenes + HASS	0.76(↓ 2.5%)	0.87(↓ 1.1%)
MLLM	nuScenes	1.00	1.23
	nuScenes + HASS	0.91(↓ 9.0%)	1.14(↓ 7.3%)
RoboTron-Sim	nuScenes	0.61	1.29
	nuScenes + HASS	0.57(↓ 6.6%)	0.67(↓ 48.1%)

Table S7. L2 Distance performance gains of HASS across different models in E2D and H2D scenarios. To rigorously evaluate the model's inherent capability to comprehend dynamic environments without relying on ego-pose dependencies, we conducted ablation studies by removing ego-pose inputs from both MLLM and RoboTron-Sim architectures.

Model	Latency	E2D			H2D		
		Day	Straight	Sunny	Night	Turn	Rainy
VAD	115.3ms	0.77	0.78	0.78	0.94	0.87	0.83
RoboTron-Sim-7B	612.8ms	0.54	0.55	0.56	0.81	0.64	0.56
RoboTron-Sim-0.5B	141.4ms	0.57	0.62	0.60	0.81	0.69	0.64

Table S8. Comparison of deployment costs.

Sim-7B and exhibiting deployment efficiency akin to traditional end-to-end model. This alignment of low deployment costs and performance improvement makes RoboTron-Sim practical for a wide range of real-world applications.

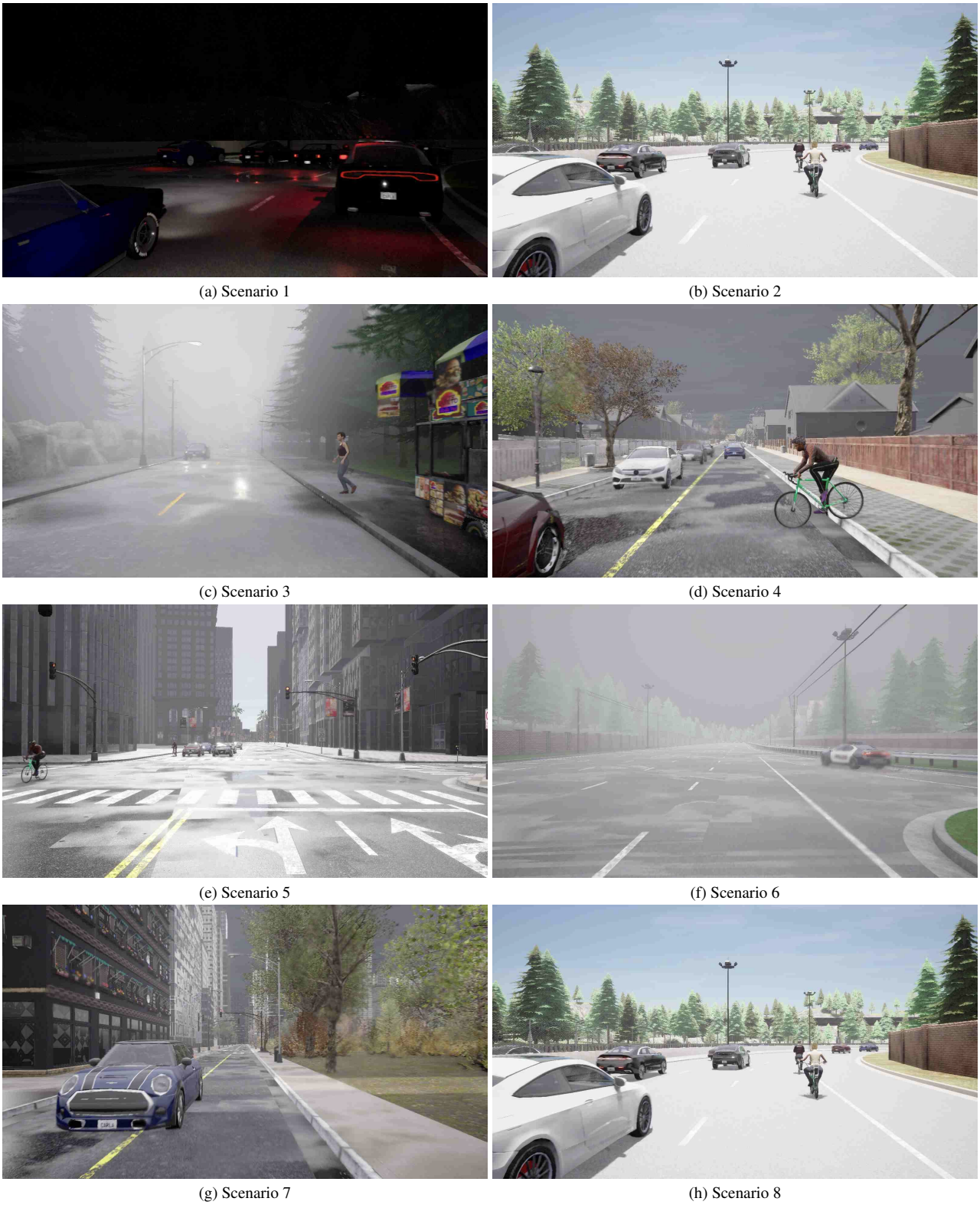
C. Visualization

C.1. In Hard-to-Drive(H2D) Scenarios

We conduct trajectory visualization comparisons among the baseline method, RoboTron-Sim, and ground truth (GT) using representative long-tail cases from the nuScenes test set (including turn, night, and similar challenging scenarios), as shown in Figure S2.

C.2. In Long-Tail Scenarios

We conduct trajectory visualization comparisons among the baseline method, RoboTron-Sim, and ground truth (GT) using representative long-tail cases from the nuScenes test set (including lane invasion, temporary parking ahead, and similar challenging scenarios), as shown in Figure S3.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

(e) Scenario 5

(f) Scenario 6

(g) Scenario 7

(h) Scenario 8

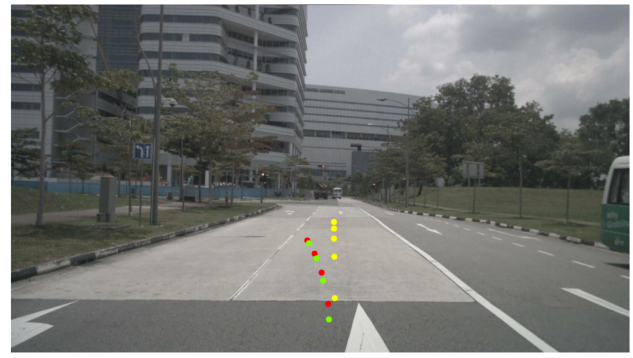
Figure S1. Visualization of HASS.

(a) Scenario 1

(b) Scenario 2

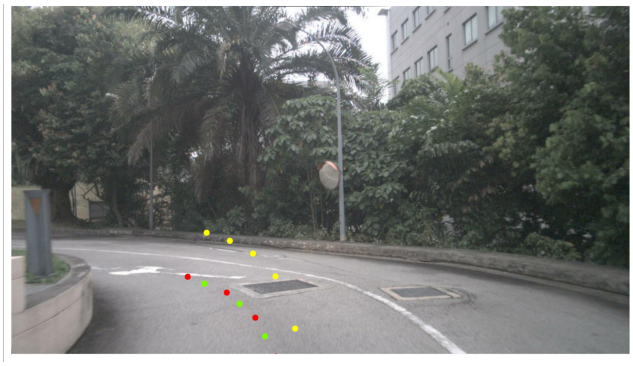
(c) Scenario 3

(d) Scenario 4

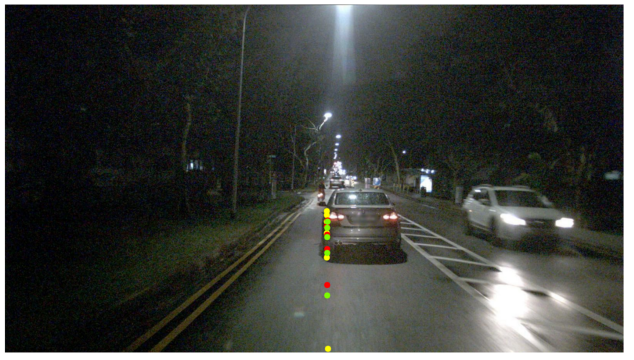


(e) Scenario 5

(f) Scenario 6



(g) Scenario 7



(h) Scenario 8

Figure S2. Visual comparison of planning trajectories in H2D scenarios. Ground-truth trajectories are marked in red, baseline predictions in yellow, and RoboTron-Sim's predictions in green.

(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

(e) Scenario 5

(f) Scenario 6

Figure S3. Visual comparison of planning trajectories in Long-Tail scenarios. Ground-truth trajectories are marked in red, baseline predictions in yellow, and RoboTron-Sim's predictions in green.

References

[1] 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 *CVPR*, pages 11621–11631, 2020. 1

[2] 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. 1

[3] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023. 2

[4] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2

[5] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14864–14873, 2024. 1