RLFTSim: Realistic and Controllable Multi-Agent Traffic Simulation via Reinforcement Learning Fine-Tuning

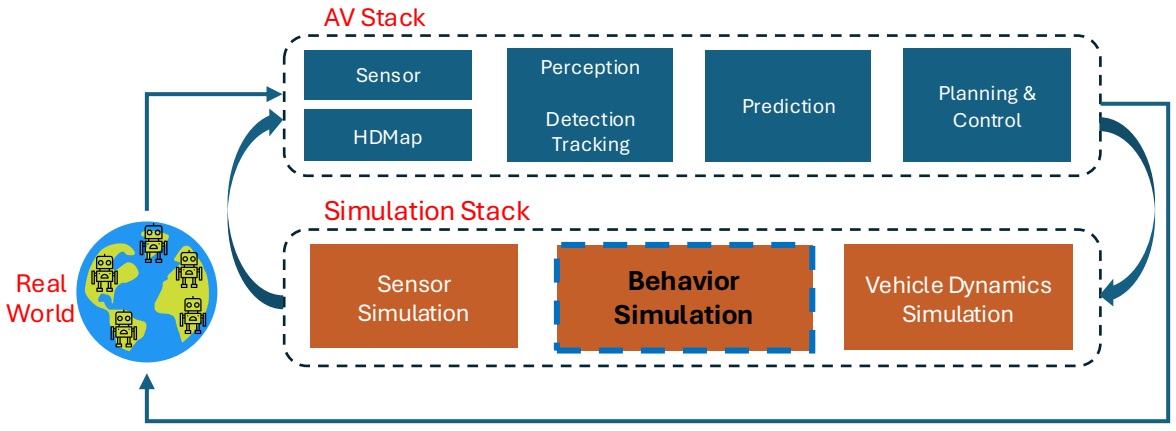
Ehsan Ahmadi[1,2] Hunter Schofield[2,3] Behzad Khamidehi[2]

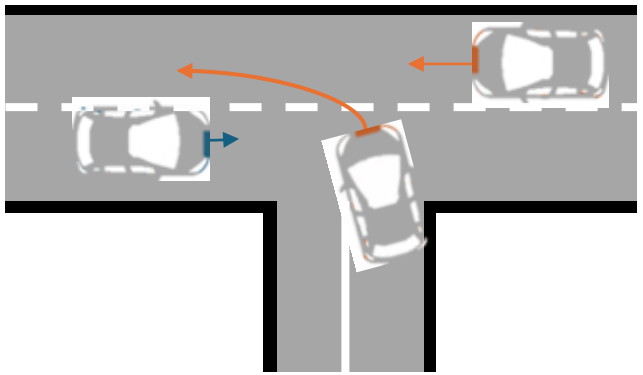
Fazel Arasteh[2] Jinjun Shan[3] Lili Mou[1,4] Dongfeng Bai[2] Kasra Rezaee[2]

[1]University of Alberta [2]Huawei Technologies Canada [3]York University [4]Canada CIFAR AI Chair, Amii

Preliminaries: Overview

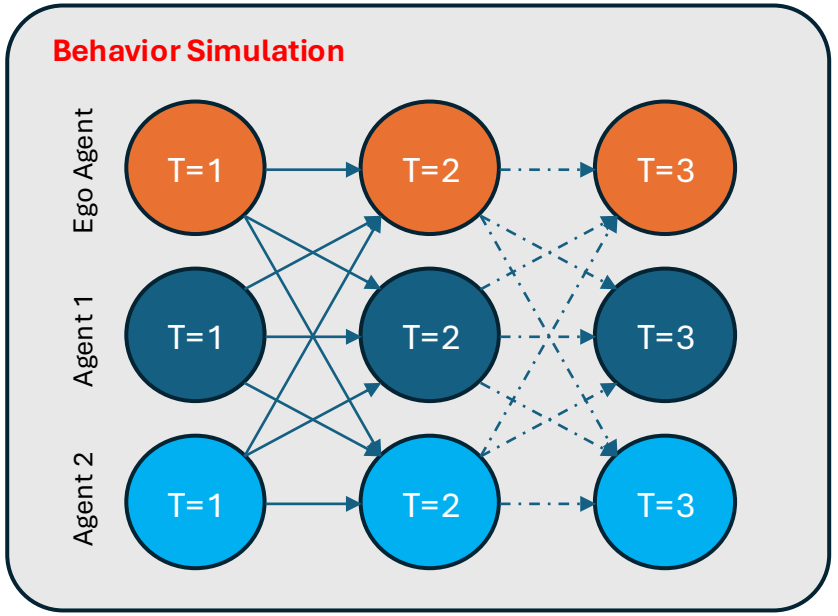
- Training & testing Autonomous Driving policies using simulation
- AD policy controls the ego agent
- Behavior Simulator controls the motion for social agent





Problem Setup: Traffic Agent Behavior Simulation

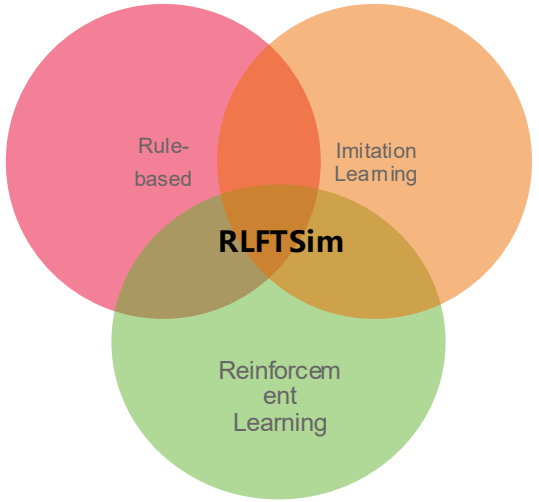
- **What:** Predicting microscopic actions of all traffic agents
- **Inputs:** past trajectory of traffic agents (1.1 seconds), and map
- **Outputs:** future trajectory of traffic agents (8 seconds)
- **Constraint:** Temporally autoregressive
- **Settings**:
  - Goal-free
  - Goal-conditioned (controllability)



Prior Work on Traffic Simulation

Rule-based	Imitation Learning	Reinforcement Learning
✓ Easy to enforce rules ✓ Light compute	✓ Strong realism ✓ Efficient training	✓ Generalizable given a good reward
✗ Unrealistic ✗ Low capacity for complex interaction	✗ Drift / error propagation ✗ Open-loop; no corrective behavior	✗ Hard to capture human intent ✗ Realism gap
Examples: IDM[1], Waymax[2], nuPlan[3], CARLA[4]	Examples: MTVE[5], Trajeglish[6], SMART[7]	Examples: GPU-Drive[8], HR-PPO[9]

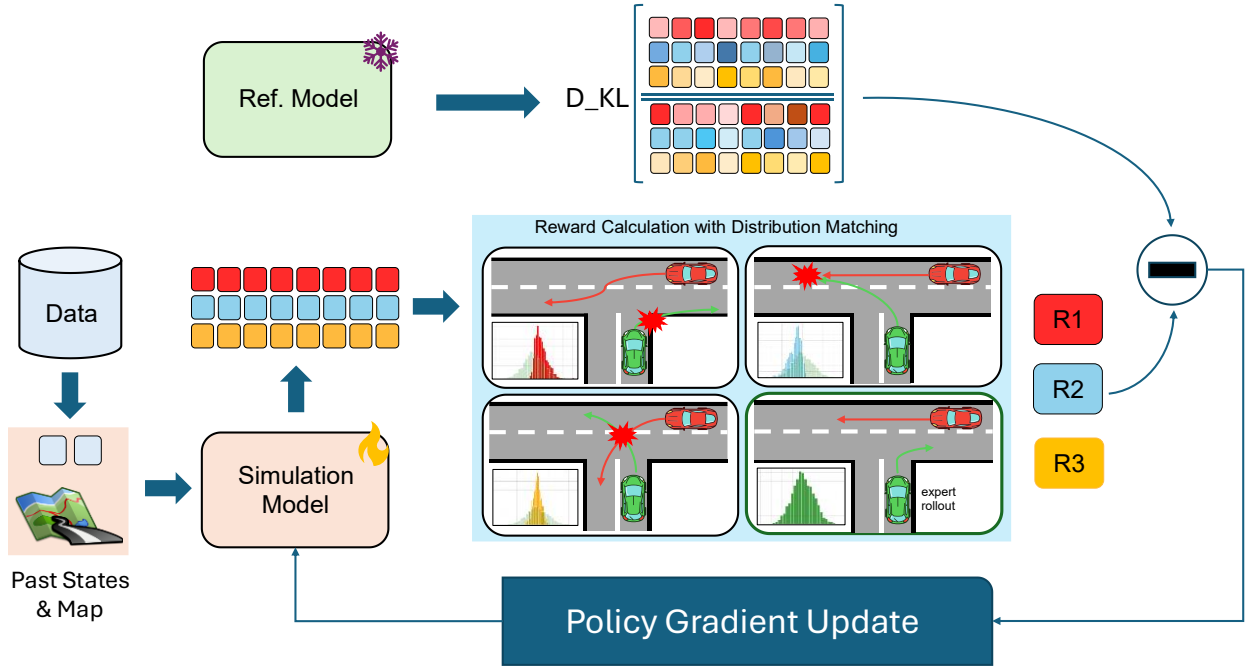
RLFTSim: *Reinforcement Learning Fine-Tuning on top of pre-trained model (Imitation Learning) using a Rule-based Reward.*



Method: Post-training Framework

RL Post-training with:

- Objective: Realism Meta-Metric
- Autoregressive base model (SMART)
- Chucked Trajectory Tokenization
- Anchoring to the reference model (KL divergence regularization term)



RL Fine-Tuning
Reward: Realism Meta-Metric (RMM)

- Calculate:
  - Histogram of the features of 32 rollouts
  - Histogram of the features of the ground truth

  (Over 8 seconds)

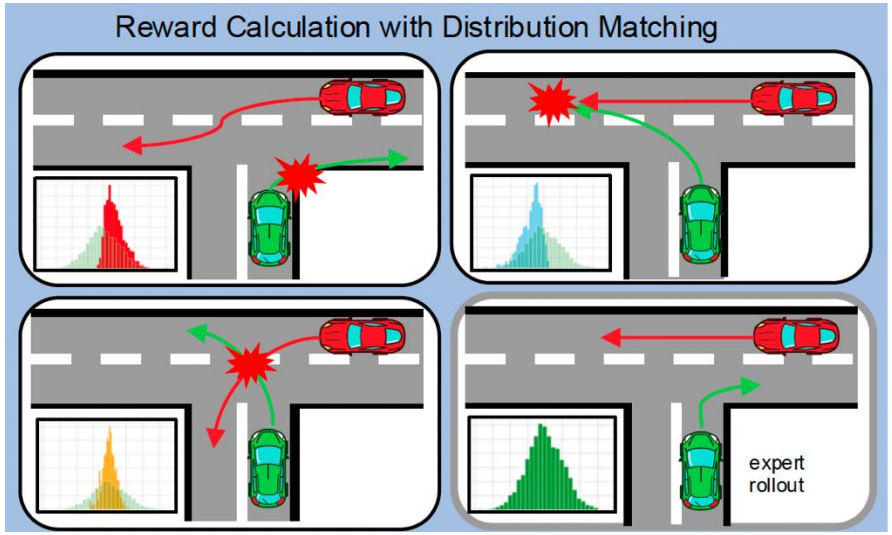
$$\hat{P}_{d,a,t}(k) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{f_{d,a,t}^{(i)} \in \mathcal{B}_{d,a,k}\}$$

$$\hat{P}_{d,a}(k) = \frac{1}{T}\sum_{t=1}^{T} \hat{P}_{d,a,t}(k)$$

- Geometric mean of Likelihood of rollouts under the ground truth distribution

$$\text{RMM} = \sum_{d=1}^{D} w_d \left[\prod_{(a,t_a)\in V} \hat{P}_{d,a}(k^*_{d,a,t_a}) \right]^{\frac{1}{|V|}}$$

$$V = \{(a, t_a); a \in \text{eval. agents}, t_a \in \text{valid time steps}\}.$$



Improved version: RMM Leave-One-Out

- Drawbacks of naively applying RMM reward:
  - It is sparse: K=32 rollouts per-reward scalar reward
  - High computation time

- Our solution: Meta-metric Leave-One-Out reward (MLOO)

Meta-metric Leave-One-Out (MLOO). To achieve a better density-variance trade-off, we introduce a dense per-rollout reward signal, $\mathrm{RMM}_i^{\mathrm{MLOO}}$, defined as:

$$\mathrm{RMM}_i^{\mathrm{MLOO}} = \frac{1}{N}\sum_{j=1}^{N} \mathrm{RMM}_{-j} - \mathrm{RMM}_{-i}, \qquad (2)$$

Analysis of Variance

- Unbiased Gradient estimation with MLOO
- Variance scaling of RMM

Proposition 1 (Unbiased gradient estimation with MLOO). *Let $\tau_{1:N} = (\tau_1, \ldots, \tau_N)$ be N i.i.d. rollouts sampled from the policy π_θ. Applying REINFORCE with per-rollout reward $\mathrm{RMM}_i^{\mathrm{MLOO}}$ as defined in Eq. 2, the policy-gradient estimator*

$$g = \sum_{i=1}^{N} \nabla_\theta \log \pi_\theta(\tau_i)\, \mathrm{RMM}_i^{\mathrm{MLOO}} \tag{3}$$

is an unbiased estimator of $\nabla_\theta \mathbb{E}[\mathrm{RMM}(\tau_{1:N-1})]$.

Proposition 2 (Variance Scaling with Simulator Bias). *Under Assumption 1, for N rollouts of length T, the realism meta-metric Eq. 1 satisfies*

$$Var(\mathrm{RMM}) = O\Big((\hat{N}_{\mathrm{eff}} \cdot T)^{-1}\Big), \tag{4}$$

where $\hat{N}_{\mathrm{eff}} = N/\hat{\kappa}$ is the effective sample size, $\hat{\kappa} = \max_d \kappa_d \geq 1$, and $\kappa_d = \sum_{k=1}^{K} \alpha_{k,d}^2 / q_{k,d}$ measures the mismatch between the simulator bin probabilities $q_{k,d}$ and the ground-truth bin frequencies $\alpha_{k,d}$ for feature dimension d.

Analysis of Variance

- MLOO vs. RLOO variance scaling

Proposition 3

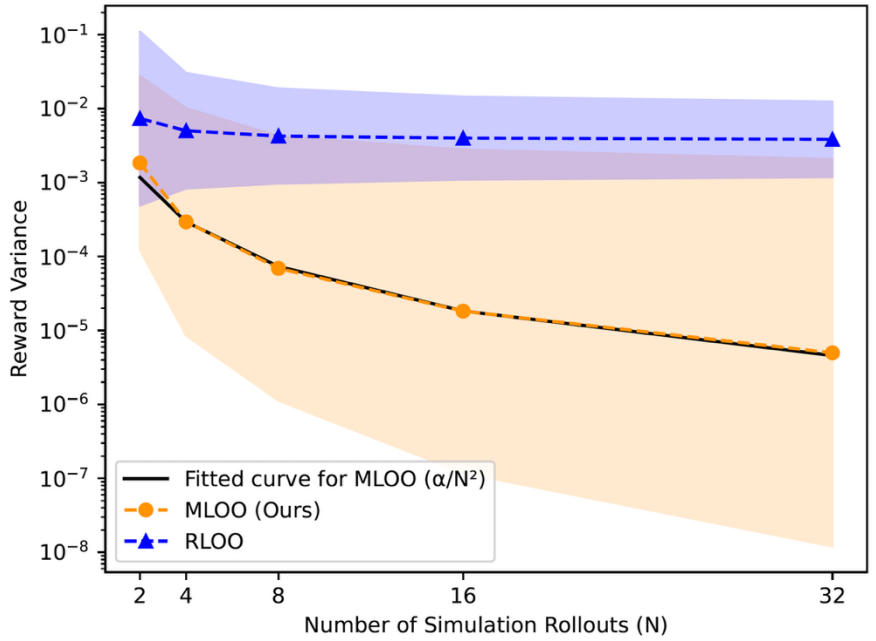
$$\mathrm{RMM}_i^{MLOO} = \frac{1}{N}\sum_{j=1}^{N}\mathrm{RMM}_{-j} - \mathrm{RMM}_{-i}, \tag{5}$$

$$\mathrm{RMM}_i^{RLOO} = \mathrm{RMM}_i - \frac{1}{N-1}\sum_{j\neq i}\mathrm{RMM}_j. \tag{6}$$

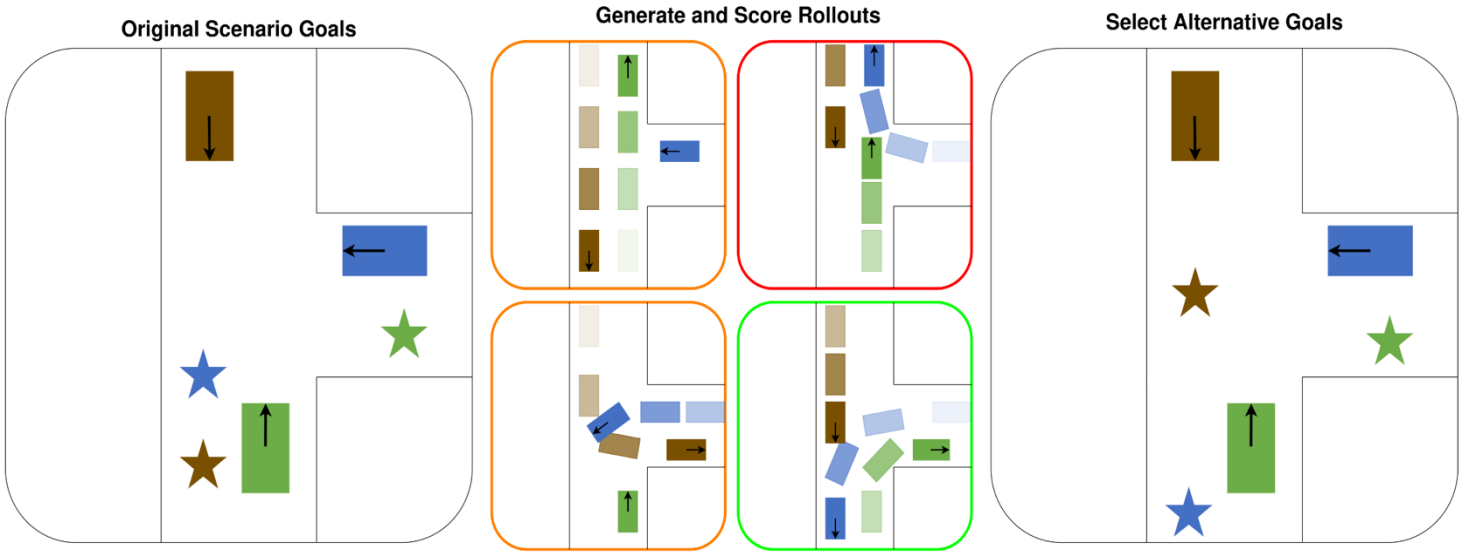
Then the variances satisfy:

$$Var(\mathrm{RMM}_i^{MLOO}) = O\left(\frac{1}{N^2 T}\right), \tag{7}$$

$$Var(\mathrm{RMM}_i^{RLOO}) = O\left(\frac{1}{T}\right). \tag{8}$$



Controllability



- Add goals as either vector observations or map-to-agent relationships via cross attention.
- Hindsight Experience Replay aids controllability learning since goal reaching events can be rare during exploration.
- To allow policy gradient updates to support HER, we use importance sampling to rescale the policy ratio.

$$\hat{r}_{i,t}(\theta) = \frac{\pi_\theta(S^i_{>=t} \mid S^i_{<t}, \hat{x}_g, C)}{\pi_\theta(S^i_{>=t} \mid S^i_{<t}, x_g, C)}$$

Experiments: WOSAC Benchmark

- Waymo Open Sim Agent Challenge
- Private test set
- Evaluates models on realistic traffic behavior generation.

Table 1. Traffic simulation benchmarking results.

Model	RMM↑	Kinematic↑	Interactive↑	Map-based↑
TrafficBotsV1.5 [37]	0.7167	0.4304	0.7114	0.8871
VBD [13]	0.7375	0.4169	0.7819	0.8636
MVTE [31]	0.7469	0.4503	0.7706	0.8859
Trajeglish [22]	0.7409	0.4166	0.7845	0.8703
KiGRAS [42]	0.7761	0.4691	0.8064	0.9126
DRoPE-Traj [41]	0.7786	0.4779	0.8065	0.9144
GUMP [11]	0.7596	0.4780	0.7887	0.8832
BehaviorGPT [44]	0.7637	0.4333	0.7997	0.9064
UniMM [17]	0.7839	0.4914	0.8089	0.9188
TrajTok [38]	0.7861	0.4887	0.8116	**0.9231**
SMART-tiny [34]	0.7755	0.4759	0.8039	0.9102
SMART-tiny [34] (ref. model)†	0.7824	0.4854	0.8089	0.9180
SMART-tiny CAT-K [39]	0.7856	**0.4931**	0.8106	0.9205
RLFTSim (ours)	**0.7867**	0.4927	**0.8129**	0.9210

Experiments:
Ablation Study on Reward Function

Metametric Leave-One-Out (MLOO) has higher:

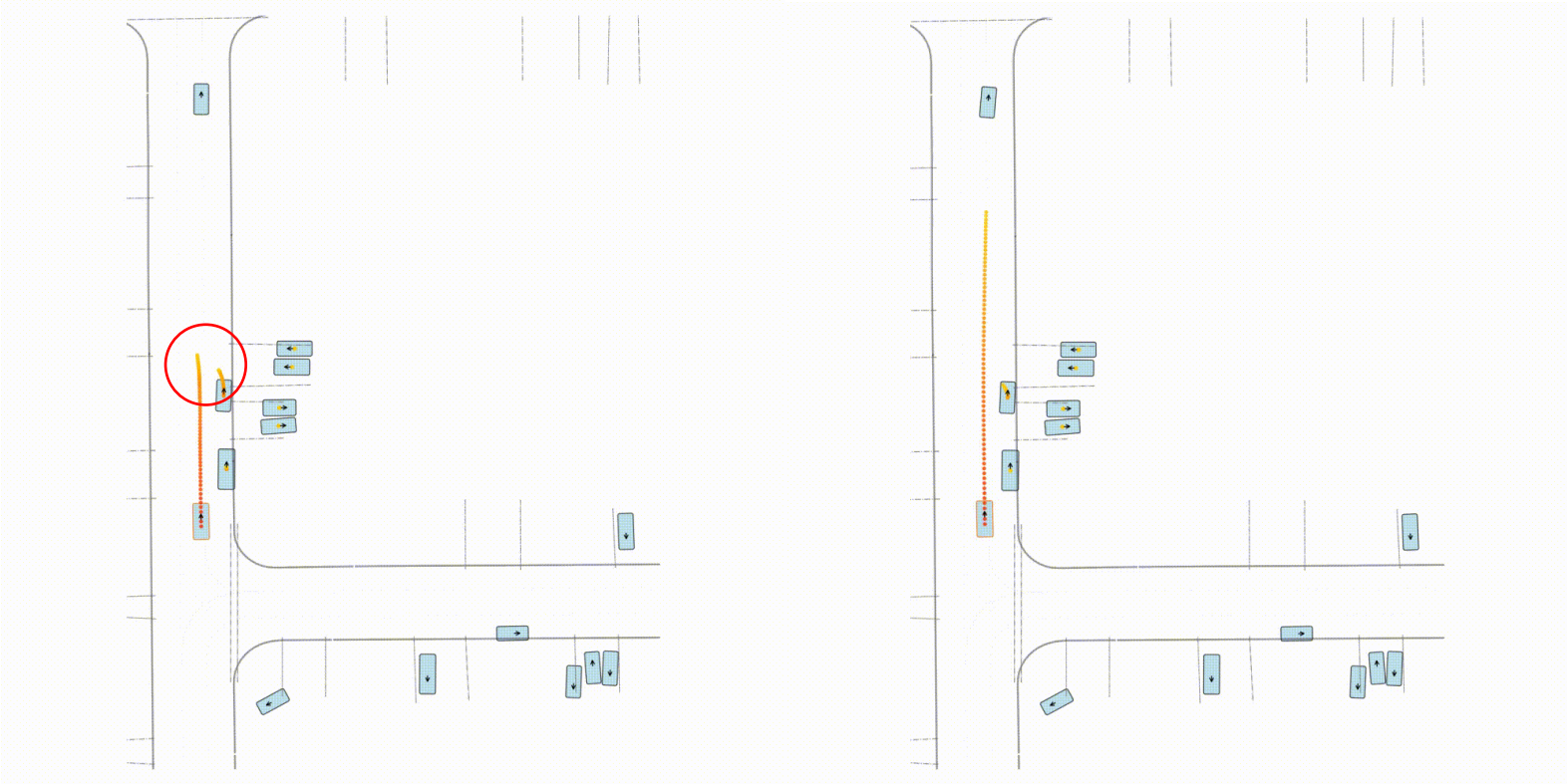
- RMM
- Kinematic Realism
- Interactive Realism

Table 2. Ablation study on the reward function on the full validation set. Standard errors are shown in parentheses.

Reward	RMM↑	Kinematic↑	Interactive↑	Map-based↑	minADE↓
SMART-tiny [34] (ref. model)	0.7804 (3.2e-4)	0.4904 (5.2e-4)	0.8032 (4.1e-4)	0.9167 (5.6e-4)	**1.3016** (4.2e-3)
$\text{minADE}^{\text{RLOO}}$	0.7801 (3.3e-4)	0.4897 (5.2e-4)	0.8032 (4.1e-4)	0.9161 (5.8e-4)	1.3202 (4.5e-3)
RMM^{RLOO}	0.7821 (3.3e-4)	0.4913 (5.1e-4)	0.8065 (4.2e-4)	0.9169 (6.0e-4)	1.3229 (4.4e-3)
RMM^{MLOO}	**0.7830** (3.3e-4)	**0.4924** (5.0e-4)	**0.8070** (4.1e-4)	**0.9182** (5.7e-4)	1.3150 (4.4e-3)
Col.+Off.+ADE	0.7803 (3.3e-4)	0.4896 (5.2e-4)	0.8039 (4.1e-4)	0.9162 (5.9e-4)	1.3313 (4.5e-3)
Collision+Offroad	0.7786 (3.5e-4)	0.4891 (5.2e-4)	0.8037 (4.2e-4)	0.9117 (6.4e-4)	1.3461 (4.3e-3)

MLOO: Metameric Leave-One-Out
RLOO : REINFORCE Leave-One-Out
ADE: Average Distance Error
RMM: Realistic Metametric

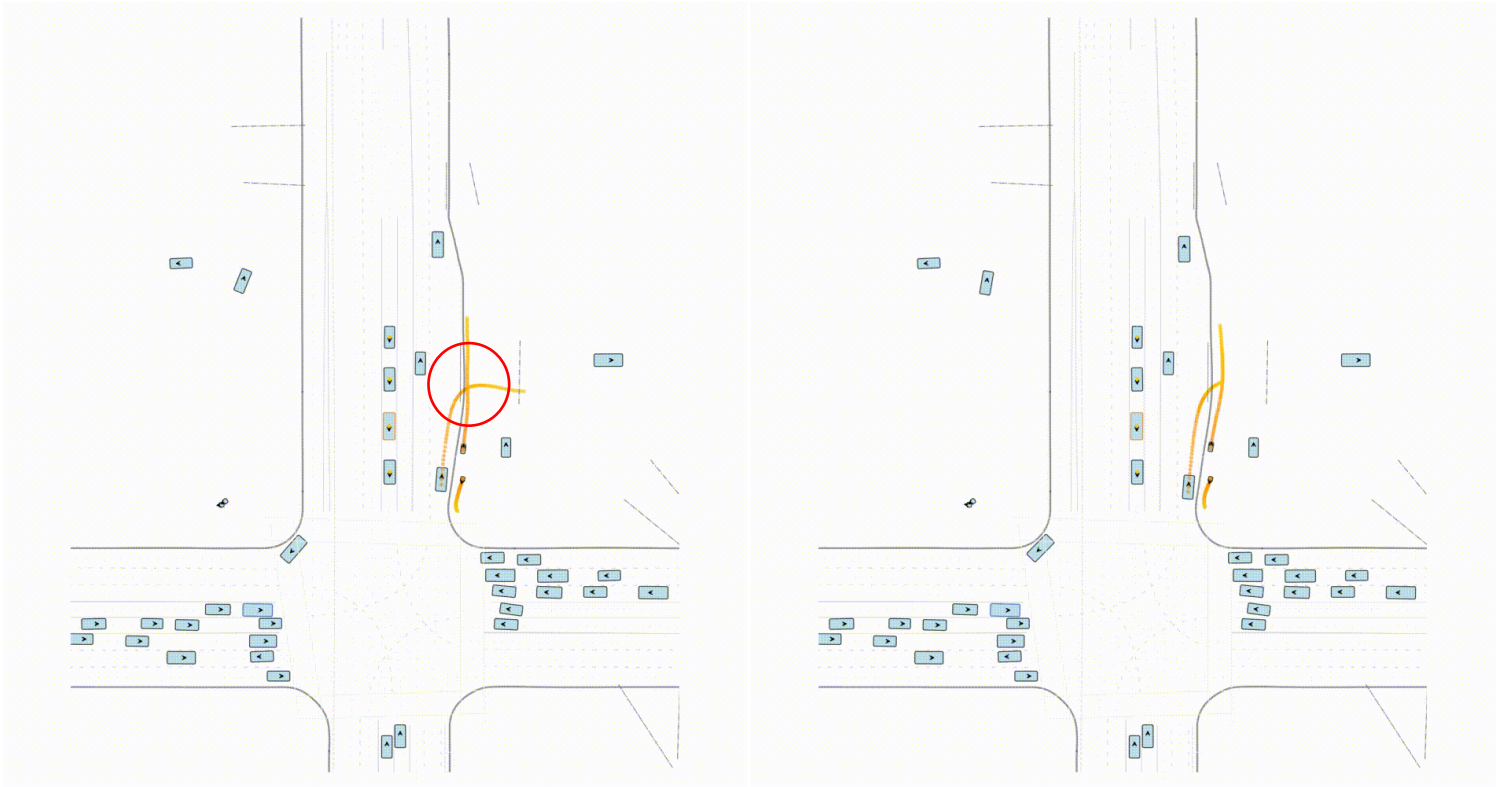
Qualitative Samples I: Collision



Pre-train

RLFTSim (Ours)

Qualitative Samples II: Collision

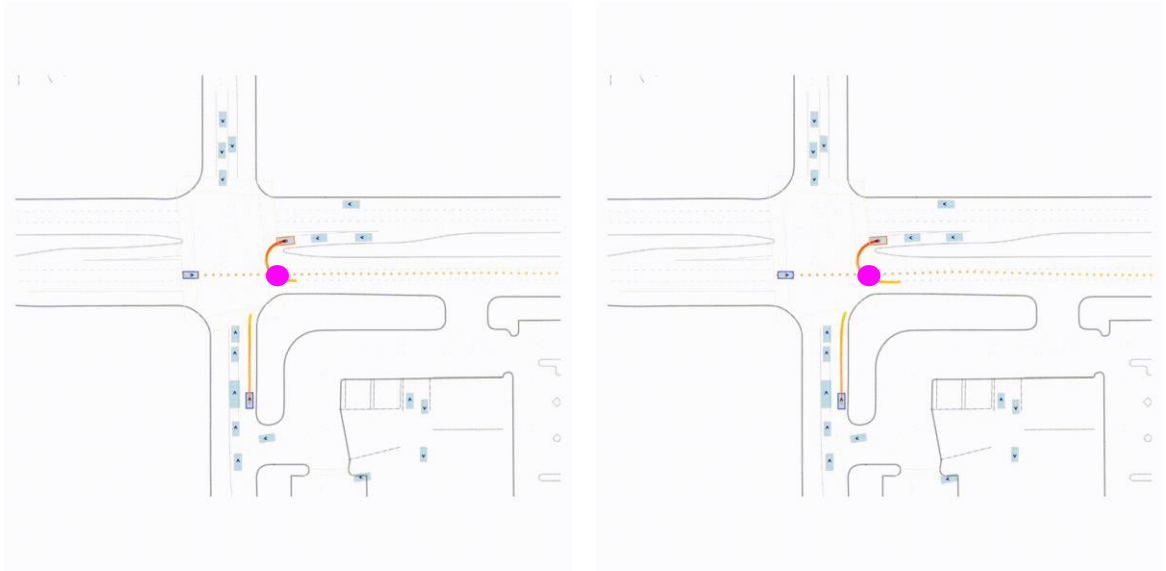


Pre-train

RLFTSim (Ours)

Qualitative Samples III: Controllability

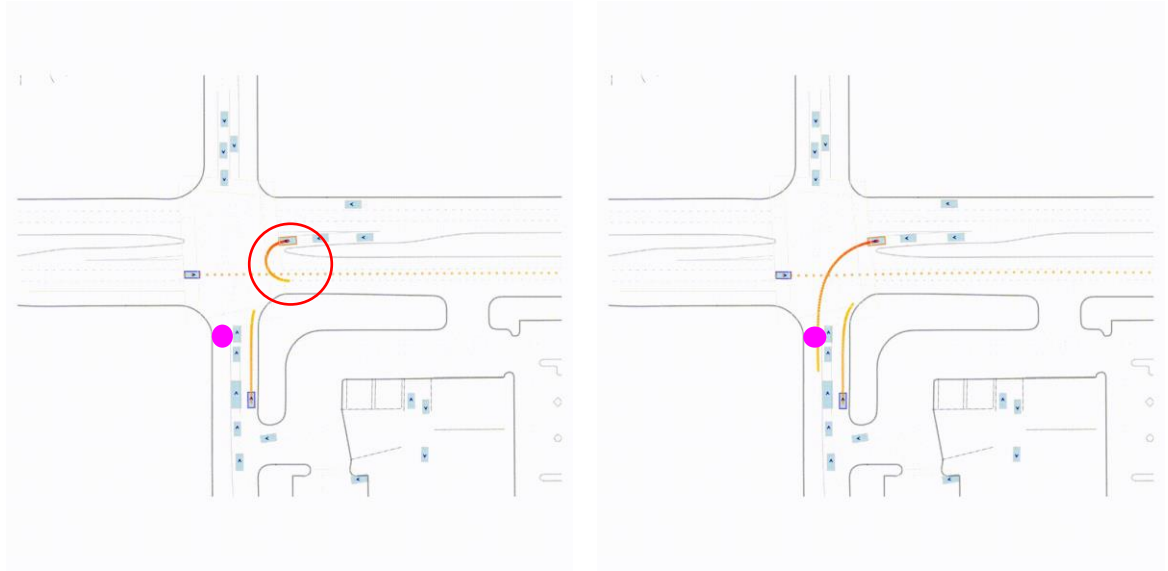
Goal: U-turn



Pre-train
Successful U-turn

RLFTSim (Ours)
Successful U-turn

Goal: Left turn



Pre-train
Failed left turn

RLFTSim (Ours)
Successful left turn

Goal Conditioned Fine-Tuning (GCFT) Visualization. The goal point is shown with a magenta colored circle for the focus vehicle in center. The simulation is shown for various o goals. We have a U-turn goal in the left, and a left turn goal in the right.

Conclusion

Takeaways

- Imitation Learning is not enough in traffic simulation
- MLOO as a **low-variance and dense** reward signal
- RLFTSim as a **sample efficient** and **effective** post-training method to enhance realism
- Adaptation of Hindsight Experience Replay to **distill controllability** via goal point conditioning

Next Steps

- RMM limitations
- Study of potential reward hacking issues

References

[1] Treiber, M., Hennecke, A., Helbing, D. Congested traffic states in empirical observations and microscopic simulations. *Phys. Rev. E*, 2000.

[2] Gulino, C., Fu, J., Luo, W., et al. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *NeurIPS*, 2023.

[3] Caesar, H., Kabzan, J., Tan, K.S., et al. nuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles. arXiv:2106.11810, 2022.

[4] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V. CARLA: An open urban driving simulator. *CoRL*, 2017.

[5] Wang, Y., Zhao, T., Yi, F. Multiverse Transformer: 1st place solution for Waymo Open Sim Agents Challenge 2023. arXiv:2306.11868, 2023.

[6] Philion, J., Peng, X.B., Fidler, S. Trajeglish: Traffic modeling as next-token prediction. *ICLR*, 2024.

[7] Wu, W., Feng, X., Gao, Z., Kan, Y. SMART: Scalable multi-agent real-time motion generation via next-token prediction. *NeurIPS*, 2024.

[8] Kazemkhani, S., et al. GPUDrive: Data-driven, multi-agent driving simulation at 1M FPS. arXiv:2408.01584, 2024.

[9] Cornelisse, D., Vinitsky, E. Human-compatible driving agents through data-regularized self-play reinforcement learning. *RLJ*, 2024.

Questions?

Please find us in **Poster Session 6**
15:30 - 17:30 ExHall A, Poster #331

Project page:
https://ehsan-ami.github.io/rlftsim