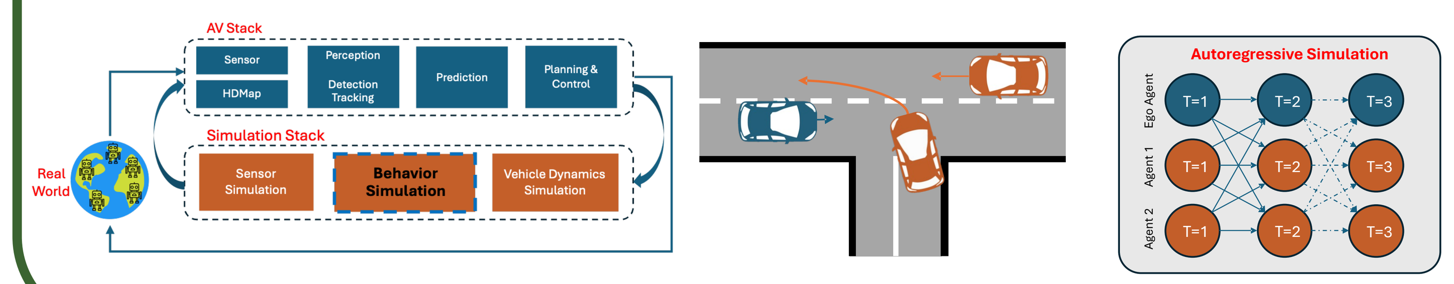
RLFTSim: Realistic and Controllable Multi-Agent Traffic Simulation via Reinforcement Learning Fine-Tuning

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

CVPR DENVER COLORADO JUNE 3-7, 2026

Realistic Closed-Loop Simulation

- **Imitation drifts.** Small errors compound in closed-loop rollouts, pushing agents into states never seen in training.
- **Realism is the bottleneck.** AV safety validation needs simulators whose distributions match real driving — not just plausible trajectories.
- **Post-train, don't restart.** Aligning a strong pre-trained simulator beats training from scratch — given the right reward.

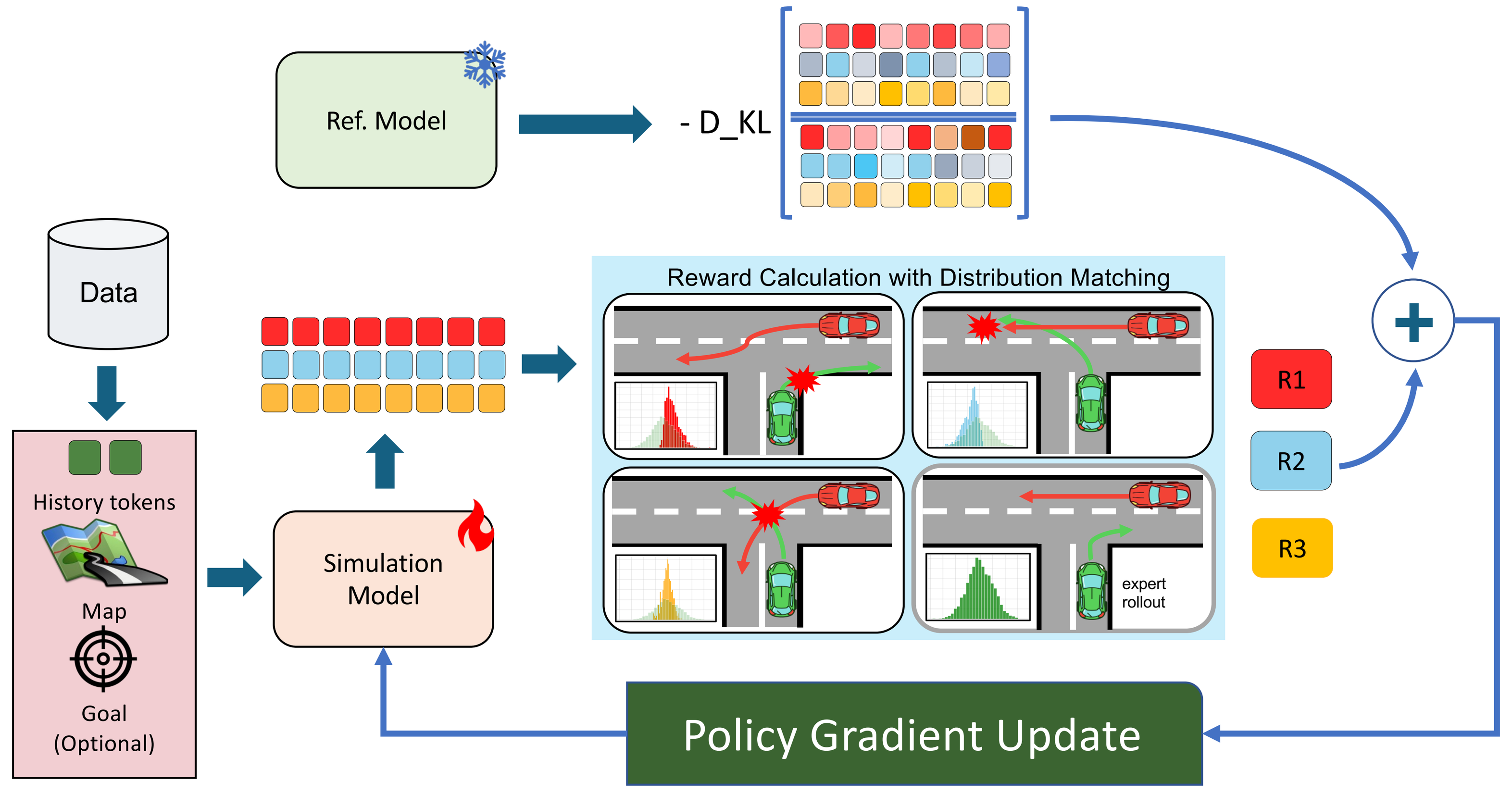


Challenges & Contributions

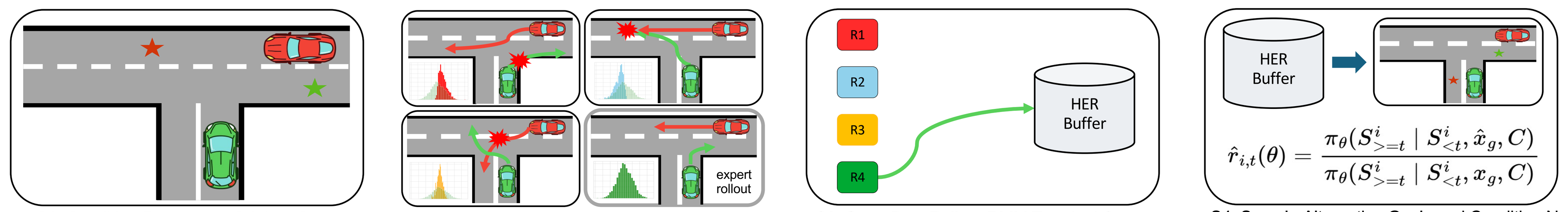
Rule-based	Imitation Learning	Reinforcement Learning
✓ Easy to enforce rules ✓ Light compute	✓ Strong realism ✓ Efficient training	✓ Generalizable given a good reward
✗ Unrealistic ✗ Low capacity for complex interaction	✗ Drift / error propagation ✗ Open-loop; no corrective behavior	✗ Hard to capture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

- **RLFTSim:** on-policy RL fine-tuning that aligns simulator rollouts with real-world driving distributions.
- **MLOO:** a dense, low-variance per-rollout reward aligned with the WOSAC realism meta-metric, with unbiased gradients and $O(1/N^2T)$ variance scaling.
- **Goal-conditioned controllability:** distilled via Hindsight Experience Replay and a balanced reward, steerable without sacrificing realism.

RLFTSim

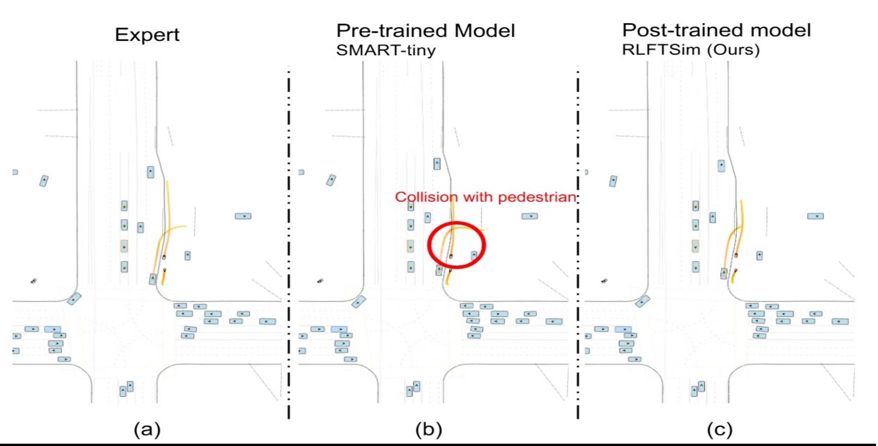


GCFT

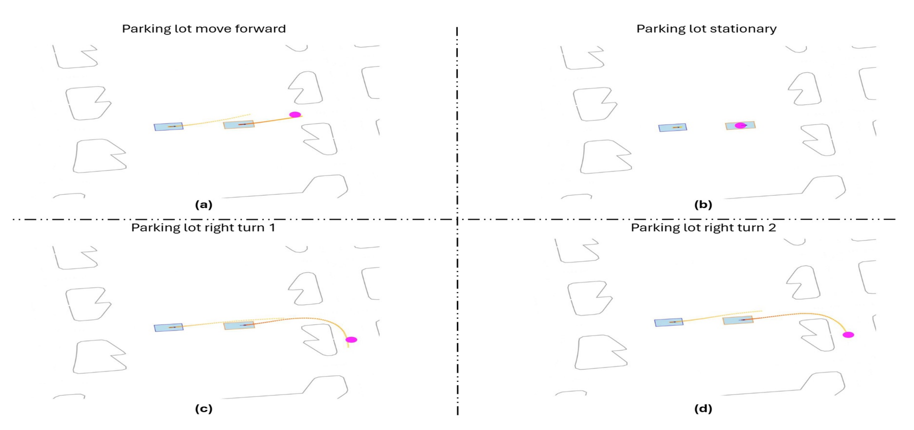


S1: Condition Rollouts on Original Goals
S2: Generate Rollouts and Compute Rewards
S3: Store Best Rollout RMM Terminal States
S4: Sample Alternative Goals and Condition New Rollouts. Scale Policy Ratio to Improve Stability

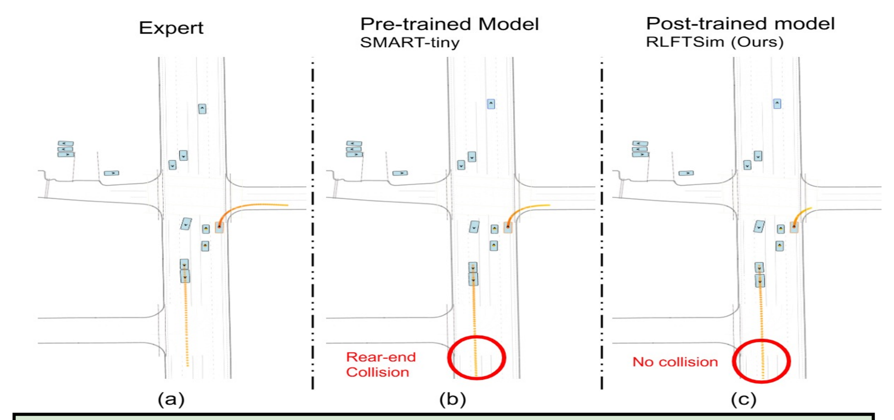
Qualitative Samples



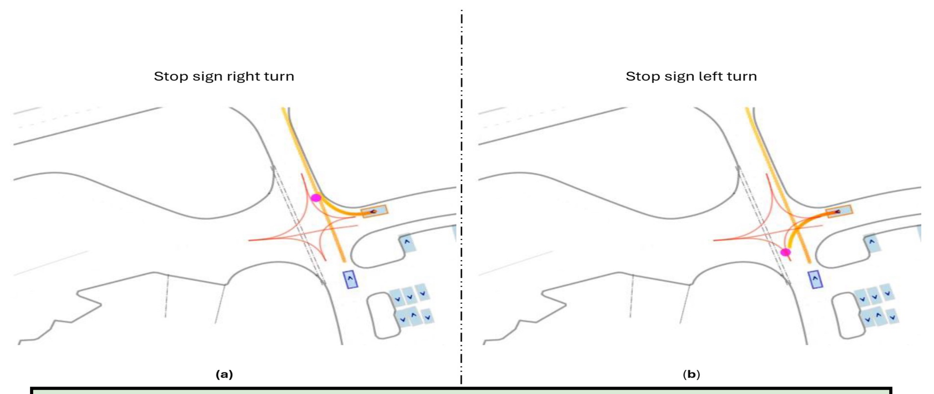
Comparison of generated trajectories for a **right-turn** into a parking lot. SMART-tiny results in a **collision** while **RLFTSim** safely completes the maneuver.



Controllability analysis in a **parking lot scenario**. After goal-conditioned fine-tuning, we can generate **novel behaviors**.



Comparison of generated trajectories along a **freeway**. SMART-tiny results in a **rear-end collision** while **RLFTSim** adheres to safe driving behavior.



Controllability analysis in a **stop sign** intersection scenario. After goal-conditioned fine-tuning, we can generate rollouts with **different driving intents**.

MLOO: Dense and Low-variance Reward

Challenge: RMM as an RL Reward

WOSAC Realism Meta-Metric (RMM):

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

Distribution-matching score over $N{=}32$ rollouts vs. ground truth.

Two issues block direct use as a reward:

- **Sparse:** 32 rollouts $\rightarrow$ a single scalar.
- **Trade-off:** shrinking N to densify *raises* variance.

MLOO: Leave-One-Out Reward

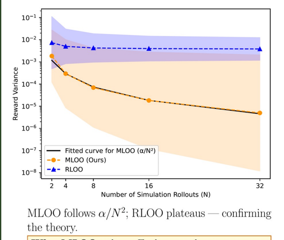
Per-rollout reward by leave-one-out subtraction:

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

RMM_{-j}: meta-metric on $N-1$ rollouts excluding rollout j.

- **Dense:** one reward per rollout.
- **Zero-sum:** $\sum_i \mathrm{RMM}_i^{\mathrm{MLOO}} = 0$.
- Trained with **REINFORCE + KL** vs. pre-trained reference.

Variance Scaling



MLOO follows α/N^2; RLOO plateaus — confirming the theory.

Why MLOO wins. Each reward compares one rollout against $N-1$ correlated leave-one-out estimates — shared structure makes variance *cancel* as N grows.

Theoretical Guarantees

Prop. 1: Unbiased Policy Gradient

The REINFORCE estimator

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

is **unbiased** for $\nabla_\theta \mathbb{E}[\mathrm{RMM}(\tau_{1:N-1})]$ $\tau_{1:N} \overset{\text{i.i.d.}}{\sim} \pi_\theta$.

Prop. 2: RMM Variance Scaling

For N rollouts of length T:

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

$\hat{\kappa} \geq 1$: simulator–ground-truth mismatch. Perfect simulator $\Rightarrow O(1/NT)$.

Prop. 3: MLOO vs. RLOO Variance

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

Quadratic reduction in N for MLOO vs. **flat** for RLOO.

Experiments

RQ1: Does MLOO-based RL fine-tuning **improve simulation realism**?

Model	RMM↑	Kinematic↑	Interactive↑	Map-based↑
TrafficBotsV1.5 [36]	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 [41]	0.7761	0.4691	0.8064	0.9126
DRoPE-Traj [40]	0.7786	0.4779	0.8065	0.9144
GUMP [11]	0.7596	0.4780	0.7887	0.8832
BehaviorGPT [43]	0.7637	0.4333	0.7997	0.9064
UniMM [17]	0.7839	0.4914	0.8089	0.9188
TrajTok [37]	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 [38]	0.7856	**0.4931**	0.8106	0.9205
RLFTSim (ours)	**0.7867**	0.4927	**0.8129**	0.9210

RQ2: Are **alternative reward functions** as effective as MLOO in post-training?

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)
$\mathrm{minADE}^{\mathrm{RLOO}}$	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)
$\mathrm{RMM}^{\mathrm{RLOO}}$	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)
$\mathrm{RMM}^{\mathrm{MLOO}}$	**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)

Reward	RMM ↑	Collision (%) ↓	Offroad (%) ↓	ADE (m) ↓	minADE (m) ↓
SMART-tiny [34] (ref. model)	0.7769	5.67	15.14	2.59	**1.30**
$\mathrm{RMM}^{\mathrm{MLOO}}$	**0.7818**	4.53	14.71	2.55	1.31
Collision-offroad-ADE	0.7788	4.93	14.73	**2.39**	1.32
Collision-offroad	0.7769	**4.51**	**13.95**	2.62	1.36

RQ3: Is there an effective approach to **condition rollouts on specific goals**?

(Goal rep., Goal criterion)	Passing Miss Rate ↓	Kinematic ↑	Interactive ↑	Map-Based ↑	RMM
Goal-Free (RLFTSim)	16.631 (9.8e-2)	**0.4924** (5.0e-4)	**0.8070** (4.1e-4)	**0.9182** (5.0e-4)	**0.7830** (3.3e-4)
(Concatenation, Soft)	10.473 (6.5e-2)	0.4794 (4.9e-4)	0.8045 (4.1e-4)	0.9134 (6.0e-4)	0.7776 (3.3e-4)
(Concatenation, Hard)	14.978 (9.1e-2)	0.4791 (4.9e-4)	0.8045 (4.1e-4)	0.9129 (6.1e-4)	0.7774 (3.3e-4)
(Indication, Soft)	**9.180** (5.9e-2)	0.4887 (4.9e-4)	**0.8068** (4.1e-4)	0.9175 (5.7e-4)	0.7819 (3.2e-4)
(Indication, Hard)	13.393 (8.2e-2)	0.4916 (5.1e-4)	**0.8068** (4.2e-4)	0.9179 (5.8e-4)	0.7827 (3.4e-4)

Takeaways

Closed-loop RL beats open-loop imitation
Imitation alone cannot fix error accumulation in closed-loop traffic simulation.

MLOO: a dense, low-variance reward
First RL use of the WOSAC realism meta-metric, with $O(1/N^2T)$ variance — quadratic reduction vs. RLOO.

State-of-the-art realism on WOSAC
RMM 0.7867 — outperforms CAT-K and prior fine-tuning baselines; model-agnostic across SMART and TrafficBots.

Controllability without losing realism
Goal-conditioned fine-tuning with Hindsight Experience Replay distills steerable behaviors from a single seed scenario.

Questions?

Ehsan's LinkedIn

Hunter's LinkedIn

Project page