清華大學车辆与运载学院
School of Vehicle and Mobility, Tsinghua University
Li Auto

NeurIPS 2026

CRAFT: Counterfactual-to-Interactive Reinforcement Fine-Tuning for Driving Policies

Keyu Chen[1,2] Nanfei Ye[2] Yida Wang[2] Wenchao Sun[1] Danqi Zhao[1] Hao Cheng[1] Sifa Zheng[1]

[1] School of Vehicle and Mobility, Tsinghua University [2] Li Auto Inc

CRAFT Project

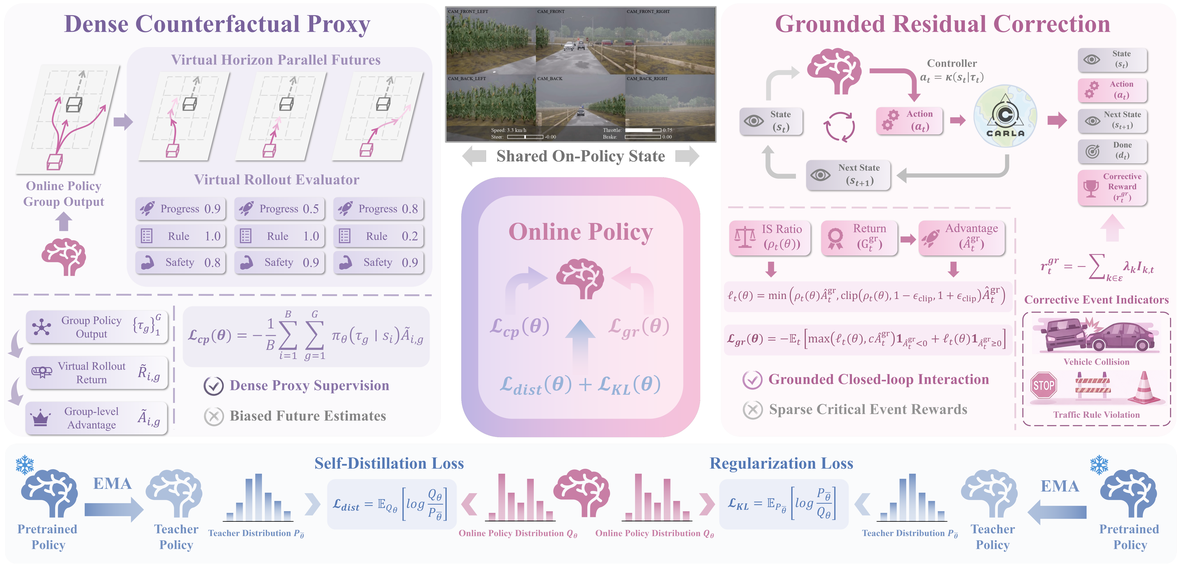
Paper · Code · Driving videos

currychen77.github.io/CRAFT

NEURAL INFORMATION PROCESSING SYSTEMS

01 Method: dense proxy with grounded correction

Driving policies trained by imitation can drift during closed-loop deployment.
Counterfactual scores are dense but biased; interaction feedback is grounded but sparse.



CRAFT combines counterfactual supervision, grounded correction, and EMA self-distillation.

Proxy–residual policy gradient

$$\nabla_\theta J_{\text{real}}(\pi_\theta) = \mathbb{E}\Big[\nabla_\theta \log \pi_\theta(\tau \mid s)\, \Phi_{\pi_\theta}(s,\tau)\Big] + \mathbb{E}\Big[\nabla_\theta \log \pi_\theta(\tau \mid s)\, \Delta_{\pi_\theta}(s,\tau)\Big],$$

Φ denotes the dense counterfactual proxy; Δ is the residual between real and proxy advantages.
Both terms share on-policy states; grounded event rewards approximate the residual.

02 Closed-loop results on Bench2Drive

Matched fine-tuning settings. † Vocabulary-based variant of SparseDriveV2.

Driving Policy	Method	Closed-loop Metrics		Multi-Ability (%)					
		DS ↑	SR (%) ↑	Merging ↑	Overtaking ↑	Emergency Brake ↑	Give Way ↑	Traffic Sign ↑	Mean ↑
HiP-AD [46] *(Hierarchical Planning)*	Pre-trained	73.46	53.64	**46.30**	53.30	63.30	**50.00**	46.30	51.80
	REINFORCE++	74.82	55.00	36.20	**64.40**	71.70	**50.00**	44.20	53.30
	PPO	73.58	53.18	41.20	55.60	68.30	30.00	43.20	47.70
	GRPO	75.27	55.91	42.50	60.00	70.00	40.00	**48.40**	52.20
	CRAFT (Ours)	**75.79**	**56.36**	40.00	**64.40**	**73.30**	**50.00**	**48.40**	**55.20**
MindDrive [24] *(Vision Language Action)*	Pre-trained	62.60	40.45	26.30	51.10	50.00	40.00	32.60	40.00
	REINFORCE++	63.68	41.36	32.50	46.70	50.00	**50.00**	33.70	42.60
	PPO	64.61	42.73	30.00	48.90	55.00	**50.00**	35.80	43.90
	GRPO	65.12	43.18	27.50	**57.80**	53.30	**50.00**	34.70	44.70
	CRAFT (Ours)	**67.45**	**46.36**	**37.50**	51.10	**56.70**	**50.00**	**37.90**	**46.60**
SparseDriveV2† [22] *(Vocabulary Scoring)*	Pre-trained	66.43	45.45	**43.80**	40.00	56.70	**50.00**	32.60	44.60
	REINFORCE++	66.60	46.36	41.20	**53.30**	55.00	**50.00**	26.30	45.20
	PPO	67.44	49.09	**43.80**	**53.30**	61.70	**50.00**	33.70	48.50
	GRPO	68.40	48.18	41.20	48.90	61.70	**50.00**	36.80	47.70
	CRAFT (Ours)	**71.10**	**50.45**	40.00	**53.30**	**68.30**	**50.00**	**37.90**	**49.90**

Best DS / SR across all three policies; up to +5.91 pp SR vs. pre-trained.

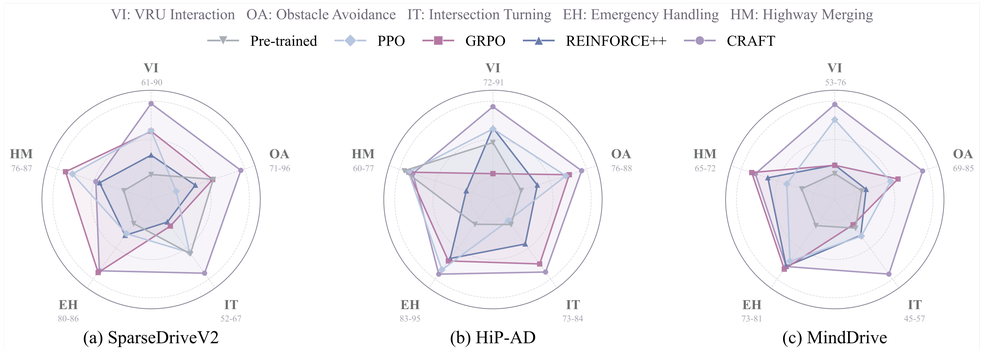


Fig. 3 Fine-grained ability profiles across three policy architectures.

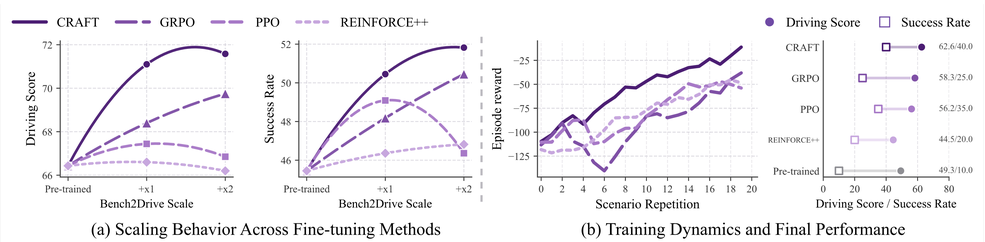


Fig. 4 Scaling behavior and closed-loop reward dynamics.

03 Component ablation

ID	SD	CP	GR	Closed-loop Metrics			
				DS ↑	RC ↑	IS ↑	SR ↑
1	✗	✗	✗	66.43	77.53	0.80	45.45
2	✓	✗	✗	66.49	76.65	0.81	46.36
3	✓	✓	✗	68.40	78.28	0.82	48.18
4	✓	✓	✓	**71.10**	**80.01**	**0.84**	**50.45**

SD: self-distillation · CP: proxy · GR: residual

04 Cross-dataset generalization

Train	Test	Closed-loop Metrics			
		Δ DS ↑	Δ RC ↑	Δ IS ↑	Δ SR ↑
Bench2Drive	Bench2Drive	**+4.67**	**+2.48**	**+0.04**	**+5.00**
Bench2Drive	Longest6 v2 [25]	+2.18	+1.12	+0.06	+0.00
LEAD‡	LEAD‡ [45]	**+2.14**	**+2.60**	**+0.01**	**+1.89**
LEAD‡	Bench2Drive	+0.58	-0.34	+0.01	+0.91

Gains over pre-trained policies; ‡ Town12 subset of LEAD.

Dense counterfactual supervision + grounded interaction correction + stable adaptation