Model-Based Policy Adaptation for Closed-Loop End-to-End Autonomous Driving

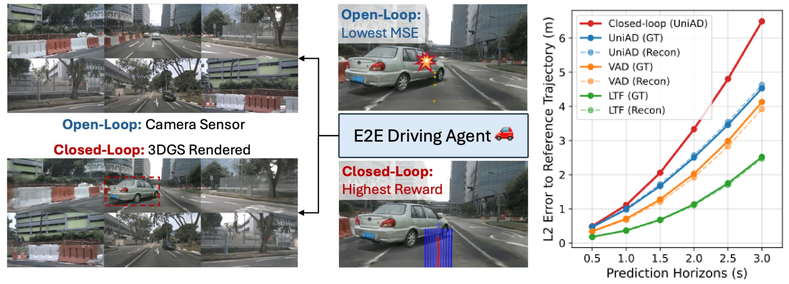
Haohong Lin[1], Yunzhi Zhang[2], Wenhao Ding[3], Jiajun Wu[2], Ding Zhao[1]
[1]CMU, [2]Stanford, [3]NVIDIA

Motivation

Problem: Existing driving foundation models fall short in closed-loop evaluation, especially in the long-tailed cases.

Goal: Close the **open-** and **closed-loop** performance gap.

Approach: Use 3DGS-synthetic data to learn a policy adapter and a value model for closed-loop planning.



Cause of Performance Gap

For closed-loop E2E driving, policy π^*, there would be:

(1) Objective mismatch: covariate shift under closed-loop deployment.

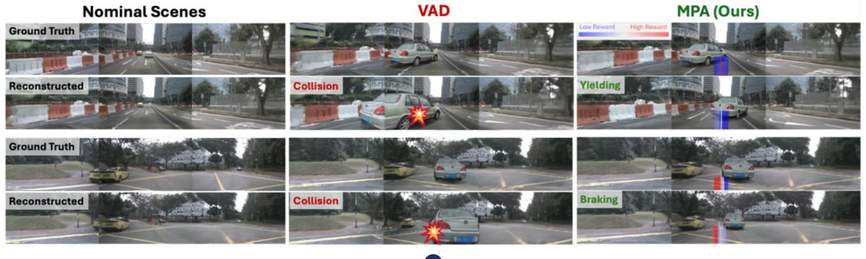
(2) Observation mismatch: between the simulated and real observations.

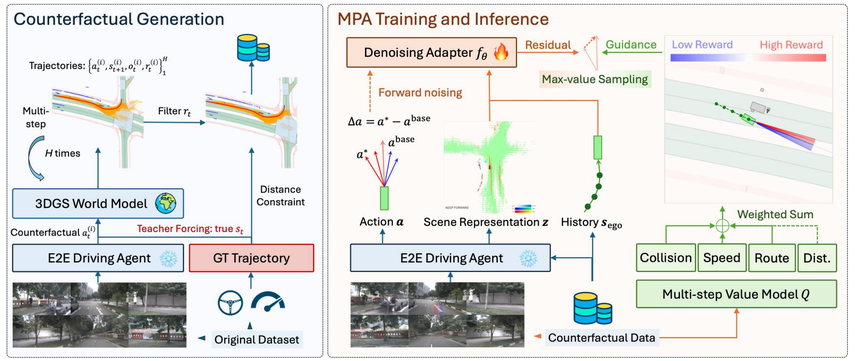
$$\text{Open-loop}: \ \hat{\pi}^* = \arg\min_{\pi} \sum_{t=1}^{T} \mathbb{E}_{(s_t,a_t)\sim\pi_{\text{ref}}} \|a_t - \pi(s_t)\|_2^2,$$

$$\text{Closed-loop}: \pi^* = \arg\max_{\pi} \sum_{t=1}^{T} \mathbb{E}_{s_t\sim P(s_{t-1},a_{t-1}),\ a_{t-1}\sim\pi(o_{t-1},s_{t-1}),\ o_{t-1}\sim P_{\text{obs}}(s_{t-1})} [r(s_t,a_t)],$$

$$\text{Simulation}: \pi^* = \arg\max_{\pi} \sum_{t=1}^{T} \mathbb{E}_{s_t\sim P(s_{t-1},a_{t-1}),\ a_{t-1}\sim\pi(o_{t-1},s_{t-1}),\ o_{t-1}\sim \hat{P}_{\text{obs}}(s_{t-1})} [r(s_t,a_t)].$$

Model-Based Policy Adaptation





What is MPA? Diffusion adapter + multi-principled Q head

How to train MPA? With **counterfactual** rollouts by 3DGS.

$$\mathbb{E}_{\Delta a^{(0)},k,\epsilon} \min_i \left\| f_\theta(\Delta a^{(k)}, k, z, s_{\text{ego}}, a^{\text{base}})[i] - \Delta a^{(0)} \right\|_2^2, \quad Q(o_t, s_t, a_t; T) = \sum_{t=1}^{T} \gamma^t r(s, a_t)$$

$$\text{where } \Delta a^{(k)} = \sqrt{\bar{\alpha}_k}\Delta a^{(0)} + \sqrt{1-\bar{\alpha}_k}\epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \quad Q = \textstyle\sum_{i\in\{\text{collision, dist, ...}\}} w_i \times Q_i$$

How to use MPA? **Inference-time Scaling** with Q-values

$$\Delta\hat{a}^* = \arg\max_{\Delta a \in a^{\text{adapt}}} Q(o_t, s_{\text{ego}}, a^{\text{base}} + \Delta a; T), \quad \hat{a}^* = a^{\text{base}} + \Delta\hat{a}^*.$$

Experiments and Analysis

Closed-Loop Evaluation on the nuScenes dataset and HUGSIM simulator

290 Training Scenarios, 70 In-domain, 70 unseen, and 10 safety-critical scenarios

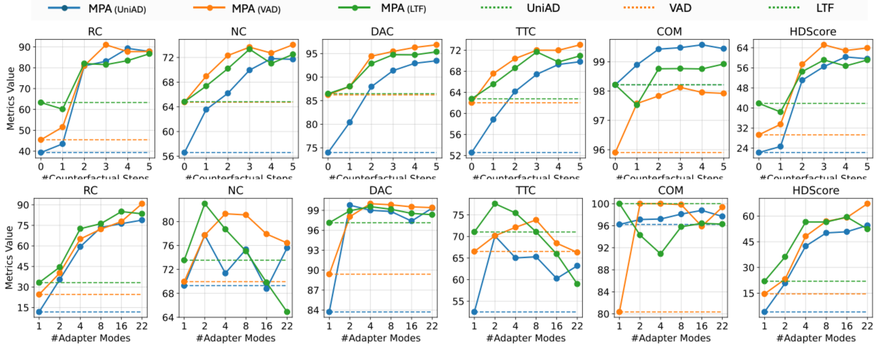
NAVSIM-like Metrics: $\text{HDScore} = \text{RC} \times \frac{1}{T}\sum_{t=0}^{T}\Big\{ \prod_{m\in\{\text{NC, DAC}\}} \text{score}_m \times \frac{\sum_{m\in\{\text{TTC, COM}\}} \text{weight}_m \times \text{score}_m}{\sum_{m\in\{\text{TTC,COM}\}} \text{weight}_m} \Big\}_t.$

Model	Ego Status	Camera	Curation	RC	NC	DAC	TTC	COM	HDScore
UniAD	✓	✓	✗	39.4	56.9	75.1	52.1	**98.7**	19.4
VAD	✓	✓	✗	50.1	68.4	87.2	66.1	90.2	31.9
LTF	✓	✓	✗	65.2	71.3	92.1	67.6	98.4	46.7
AD-MLP	✓	✗	✓	13.4	**80.2**	86.2	**79.4**	90.1	6.5
BC-Safe	✓	✓	✓	57.0	59.8	87.9	55.2	89.4	33.6
Diffusion	✓	✓	✓	71.8	67.4	88.1	64.5	91.5	45.1
MPA (UniAD)	✓	✓	✓	93.6	76.4	92.8	72.8	91.8	66.4
MPA (VAD)	✓	✓	✓	**94.9**	75.4	**93.6**	72.5	92.8	**67.0**
MPA (LTF)	✓	✓	✓	93.1	70.8	90.9	67.9	94.9	60.0

	Unseen Nominal Scenes						Safety-Critical Scenes					
Model	RC	NC	DAC	TTC	COM	HDScore	RC	NC	DAC	TTC	COM	HDScore
UniAD	39.3	56.6	74.0	52.6	**98.2**	22.2	11.4	76.2	82.1	57.8	95.9	4.5
VAD	45.4	64.8	86.2	62.0	95.9	29.3	25.4	77.0	88.3	73.2	88.4	16.0
LTF	63.3	64.8	86.5	62.8	**98.2**	41.9	35.1	80.9	96.8	78.1	**100.0**	24.2
AD-MLP	7.6	**71.6**	82.2	**69.8**	92.3	3.3	4.9	**93.5**	96.2	**93.4**	85.9	4.3
BC-Safe	59.2	59.8	81.2	56.3	95.9	34.6	20.2	80.1	91.7	67.3	86.7	13.5
Diffusion	57.9	62.1	83.5	58.3	96.2	35.1	20.9	84.3	92.3	72.4	86.3	13.1
MPA (UniAD)	**93.7**	69.5	92.9	66.6	97.6	60.9	95.1	76.8	98.9	74.2	97.7	70.4
MPA (VAD)	90.9	71.0	**94.4**	68.8	97.7	**61.2**	**96.6**	79.8	**99.0**	77.3	97.7	**74.7**
MPA (LTF)	91.8	68.3	91.0	66.5	96.9	57.0	87.3	72.0	94.0	66.9	97.8	56.3

Observation: MPA manages to outperform both the pretrained baselines and baselines solely trained on the counterfactual dataset.
The closed-loop performance of MPA generalizes well in the OOD and safety-critical scenes.

Performance w.r.t. *#Counterfactual Step*(T) and *#Sample Modes*



Short Takeaways: Using counterfactual data to learn policy adapter and Q-value critic leads to better closed-loop performance in E2E autonomous driving!