Safety-aware Causal Representation for Trustworthy Reinforcement Learning in Autonomous Vehicles

Haohong Lin, Wenhao Ding, Zuxin Liu, Yaru Niu, Jiacheng Zhu, Yuming Niu, and Ding Zhao

Carnegie Mellon University, Ford Motor Company

Motivation

Autonomous driving systems…

- Desires safety & generalizability
- Lacks structural awareness of the world

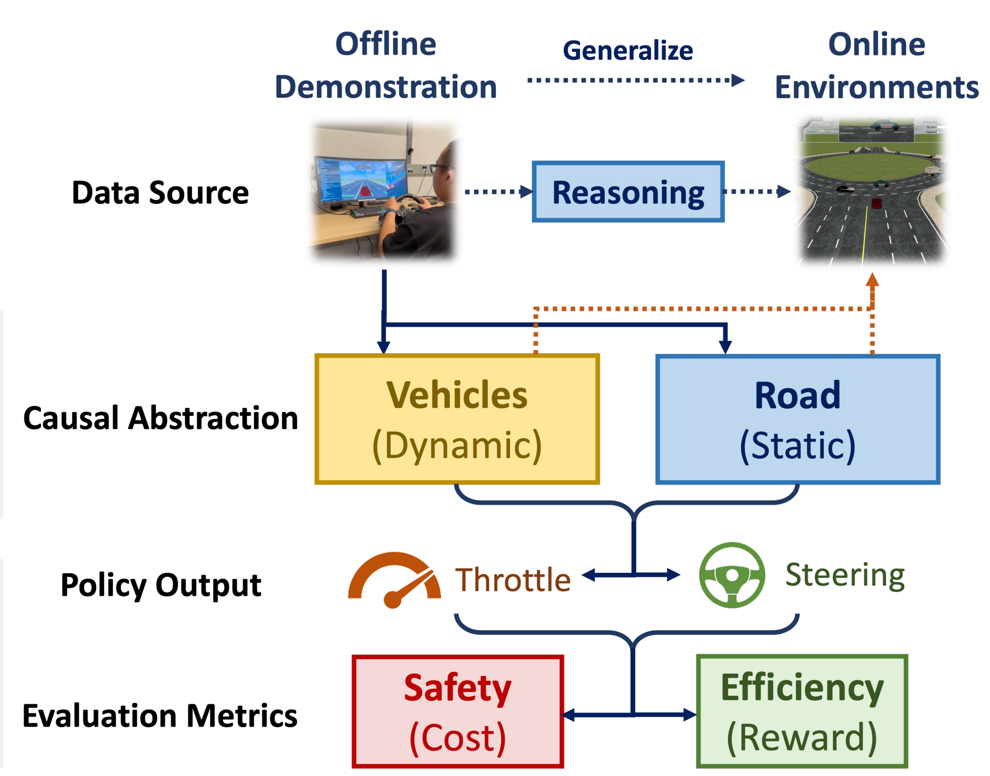
Existing approach along the pipeline…

- End-to-end solutions that are scalable?
- Balance safety and efficiency?

Simulator: *CARLA, MetaDrive*
[Dosovitskiy et al., CoRL 17'], [Li et al., TPAMI 22']
Dataset: *Waymo, Argoverse*
[Sun et al., CVPR 20'], [Chang et al., CVPR 19']

Explicit: *CDL, GRADER*
[Wang et.al. ICML 22'], [Ding et al., NeurIPS 22']
Implicit: *DBC, Denoised MDP.*
[Zhang et al., ICLR 21'], [Wang et al., ICML 22']

Explicit Constraints: *InterFuser*
[Shao et.al. CoRL 22']
Value-based: *SaFormer, CPQ*
[Xu et al., AAAI 22'] [Zhang et al., ICLR 23']



Problem Formulation

Constrained optimization:

$$\max_{\pi} \; J_r(\pi,\omega) \qquad r = w_1^r r_{forward} + w_2^r r_{speed} + w_3^r r_{term}$$

$$s.t. \;\; J_c(\pi,\omega) \le \kappa_c \qquad c = w_1^c c_{collide} + w_2^c c_{out\_road} + w_3^c c_{speed}$$

Structured Causal Model

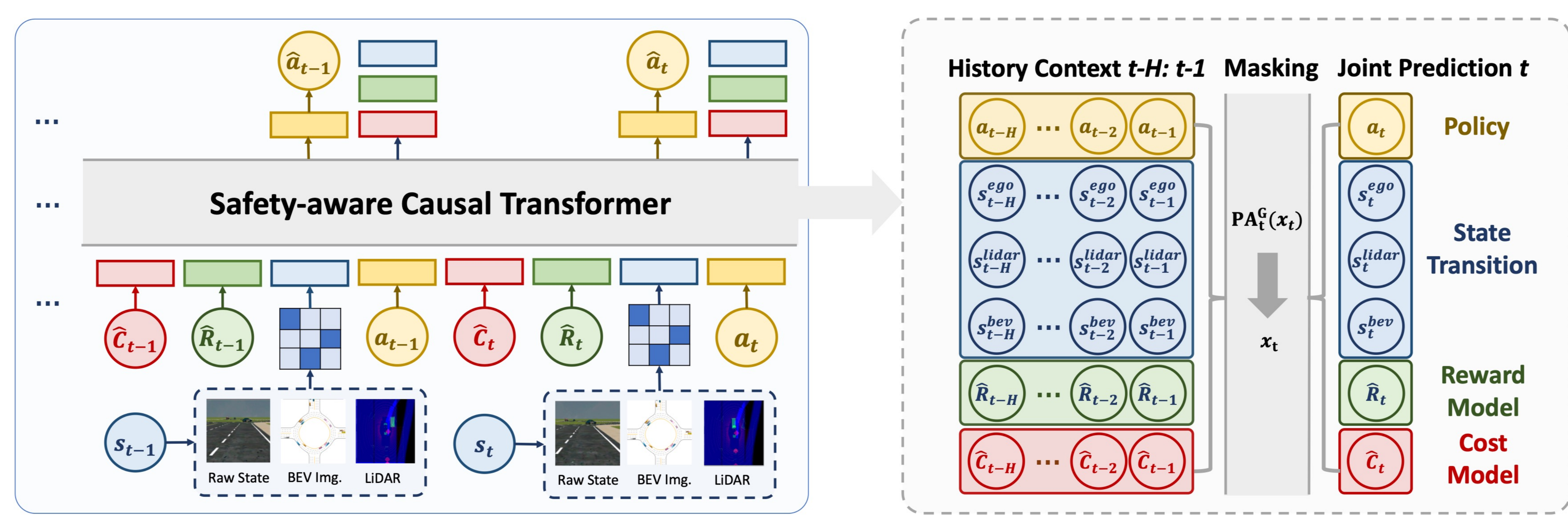
$$s_j := f_j(\mathbf{PA}^{\mathcal{G}}(s_j), \epsilon_j), \quad j \in [d],$$

Set of variables | Aggregation function | Parental Nodes in Causal Graph G | Exogenous Noise Variables

Methodology

FUSION: saFety-aware strUctural Scenario representatION

Step I: Causal Ensemble World Model



$$p(\tau_{t+1}|\tau_{t-K:t}) = p(a_{t+1}, s_{t+1}, R_{t+1}, C_{t+1}|a_t, s_t, R_t, C_t, \cdots)$$
$$= \underbrace{p\Big(r_t|\mathbf{PA}^G(r_t)\Big)}_{\text{Reward-to-go}} \underbrace{p\Big(c_t|\mathbf{PA}^G(c_t)\Big)}_{\text{Cost-to-go}}$$
$$\underbrace{\prod_{i\in\dim(S)} p(s^i_{t+1}|\mathbf{PA}^G(s^i_{t+1}))}_{\text{Factorized Dynamics}}$$
$$\underbrace{p(a_{t+1}|\mathbf{PA}^G(a_{t+1}))}_{\text{Policy Optimization}}$$

$$\mathcal{L}_{traj} = \log p(\tau_{t+1}|\tau_{t-K:t})$$
$$= \log p(r_t|\mathbf{PA}^G(r_t)) + \log p(c_t|\mathbf{PA}^G(c_t))$$
$$+ \sum_{i\in\dim(S)} \log p(s^i_{t+1}|\mathbf{PA}^G(s^i_{t+1}))$$
$$+ \log p(a_{t+1}|\mathbf{PA}^G(a_{t+1}))$$
$$= \underbrace{\mathcal{L}_{rtg}}_{\text{Reward Critic}} + \underbrace{\mathcal{L}_{ctg}}_{\text{Cost Critic}} + \underbrace{\mathcal{L}_{dyn}}_{\text{Transition Dynamics}} + \underbrace{\mathcal{L}_{act}}_{\text{Policy Optimization}}$$

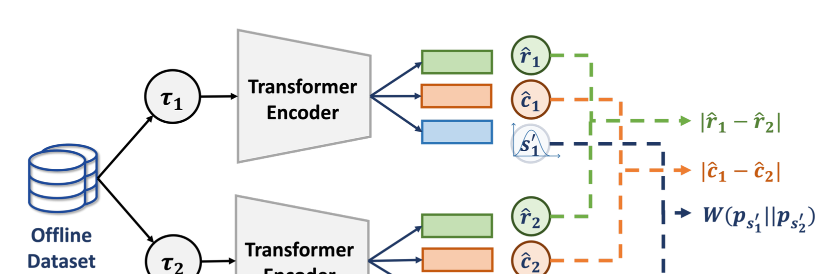
Step II: Causal Bisimulation Learning

Definition: Safety-aware Bisimulation Relationship

- $\forall a \in \mathcal{A}, r(s_1, a) = r(s_2, a)$
- $\forall a \in \mathcal{A}, c(s_1, a) = c(s_2, a)$
- $\forall a \in \mathcal{A}, s' \in \mathcal{S}, p(s'|s_1, a) = p(s'|s_2, a)$

Definition: Safety-aware Bisimulation Metrics

$$d^\pi(s_1, s_2) = \mathbb{E}_{\substack{a_1\sim\pi(\cdot|s_1),\\ a_2\sim\pi(\cdot|s_2)}} \Big[|r(s_1,a_1) - r(s_2,a_2)|$$
$$+ \lambda|c(s_1,a_1) - c(s_2,a_2)| + \gamma W_2(\hat{p}(\cdot|s_1,a_1), \hat{p}(\cdot|s_2,a_2))\Big]$$



Algorithm 1: Training and Inference of FUSION

Data: Context length H, Reward target R_0, Cost limit C_0
Result: Policy $\pi_{\theta,\phi}$
`/* Offline Training */`
for $k = 0, \cdots, N-1$ **do**
  Update Transformer θ with CEWM by (4);
  Update Encoder ϕ with CBL by Alg. 2;
`/* Online Inference with context H */`
$s_0 \leftarrow$ env.reset();
$\mathbf{o} \leftarrow \{C_0, R_0, s_0\}$;
$a_0 \leftarrow \pi_{\theta,\phi}(\mathbf{o})$;
for $t = 1, \cdots, T-1$ **do**
  Rollout: $s_t, r_t, c_t =$ env.step(a_{t-1});
  Predict reward value: $\hat{R}(s_t, a_t) \leftarrow \phi^r(s_t)$;
  Predict cost value: $\hat{C}(a_t, s_t) \leftarrow \phi^c(s_t)$;
  Update rtg token:
   $R_t \leftarrow \max\{\hat{R}(s_t, a_t), R_{t-1} - r_t\}$;
  Update ctg token:
   $C_t \leftarrow \min\{\hat{C}(s_t, a_t), C_{t-1} - c_t\}$;
  Update context: $\mathbf{o} \leftarrow \{\{a_{t-1}, C_t, R_t, s_t\}\}_{t-H:t}$;
  Predict action: $a_t \leftarrow \pi_{\theta,\phi}(\mathbf{o})$;

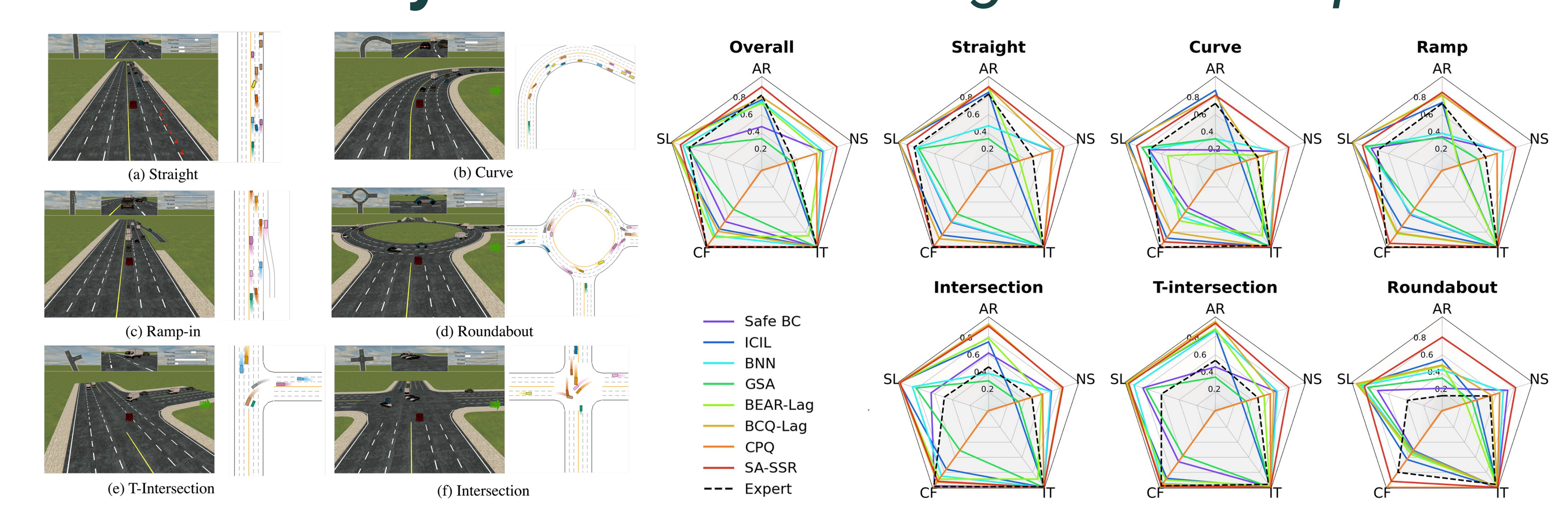
Experiments and Analysis

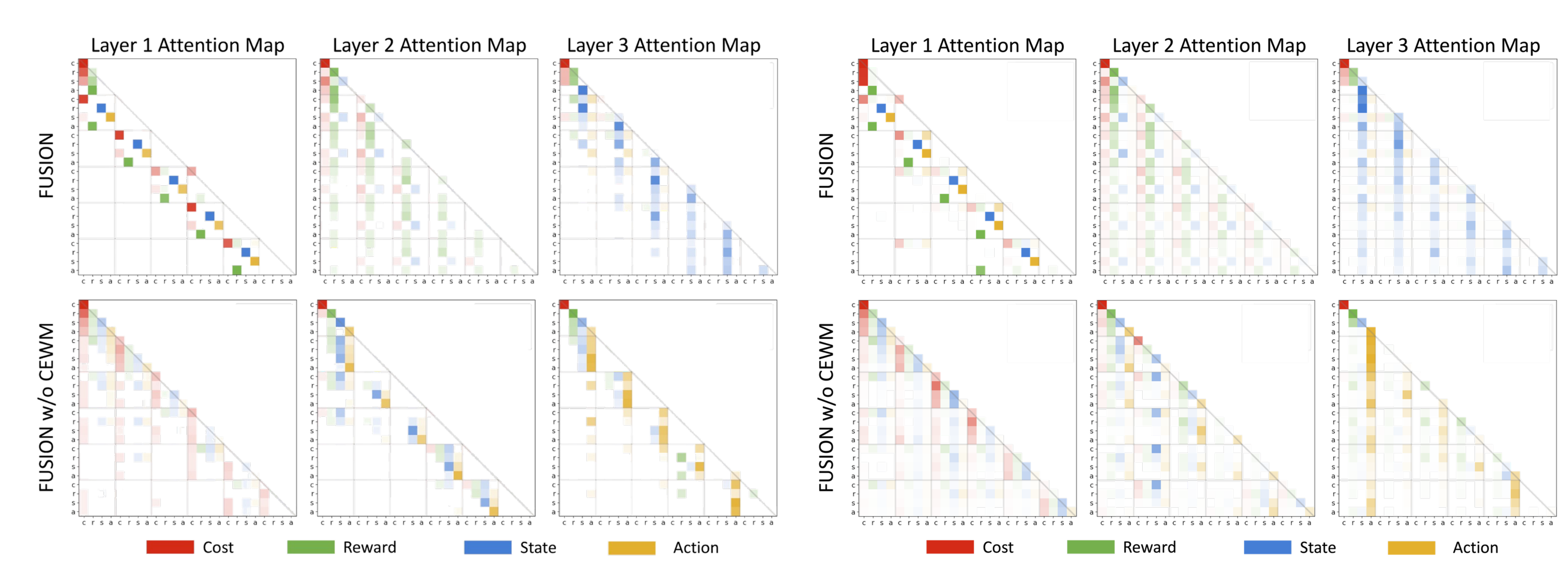
Evaluation Settings ($\kappa_c = 1$):

- Policy Mismatch (imperfect demonstration)
- Dynamics Mismatch (dense traffic)

Method	Policy Mismatch			Dynamics Mismatch		
	Reward (↑)	Cost (↓)	Succ. Rate (↑)	Reward (↑)	Cost (↓)	Succ. Rate (↑)
Safe BC	106.28±7.49	12.79±0.70	0.47±0.10	81.07±3.80	9.44±0.55	0.12±0.06
ICIL	122.66±4.85	11.07±1.11	0.76±0.05	88.21±5.30	7.29±0.72	0.32±0.05
BNN	118.61±3.09	4.46±0.41	0.74±0.11	113.35±5.68	19.16±0.55	0.59±0.06
GSA	89.94±6.84	13.18±1.26	0.34±0.08	102.40±6.44	11.88±0.98	0.03±0.02
BEAR-Lag	109.62±3.91	4.46±0.29	0.72±0.06	113.38±5.25	7.86±0.66	0.32±0.05
BCQ-Lag	111.36±5.26	**0.89**±0.08	0.79±0.08	122.72±7.64	6.22±0.76	0.39±0.08
CPQ	9.01±0.87	1.05±0.18	0.00±0.00	7.47±0.59	0.71±0.09	0.00±0.00
FUSION (Ours)	**139.95**±4.24	**0.52**±0.06	**0.90**±0.03	**117.40**±4.30	**0.90**±0.14	**0.82**±0.04
FUSION-Short	100.86±3.40	0.77±0.09	0.34±0.07	98.63±2.36	0.79±0.06	0.34±0.04
FUSION w/o CEWM	94.24±6.50	**0.67**±0.11	0.41±0.06	81.70±3.82	**0.60**±0.04	0.24±0.04
FUSION w/o CBL	104.54±4.04	3.46±0.21	0.58±0.09	90.34±4.28	5.60±0.32	0.08±0.01
FUSION	**139.95**±4.24	**0.52**±0.06	**0.90**±0.03	**117.40**±4.30	**0.90**±0.14	**0.82**±0.04
Expert Policy	131.32±29.60	16.02±9.46	0.81±0.15	129.71±28.84	17.58± 9.71	0.72±0.20

Result Analysis: *Diverse Config. / Attn. Map*





Take-aways

- CEWM transforms the offline RL as a sequence modeling problem, while adding more sequential awareness accounts for better results.
- CBL empowers the structural dynamics by enforcing extra sparsity.
- Comprehensive empirical evaluations with safety-aware LfD baselines