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

Haohong Lin[1], **Wenhao Ding**[1], **Zuxin Liu**[1], **Yaru Niu**[1], **Jiacheng Zhu**[1]
Yuming Niu[2], **Ding Zhao**[1]
{haohongl, wenhaod, zuxinl, yarun, jzhu4}@andrew.cmu.edu,
yniu4@ford.com, dingzhao@cmu.edu,
Carnegie Mellon University [*], Ford Motor Company [†]

Abstract

The Learning from Demonstration (LfD) paradigm has exhibited notable efficacy in addressing sequential decision-making problems within the domain of autonomous driving. However, consistently achieving safety in varying traffic contexts, especially in safety-critical scenarios, poses a significant challenge due to the long-tailed and unforeseen scenarios absent from offline datasets. In this paper, we introduce the saFety-aware strUctured Scenario representatION (FUSION), a pioneering methodology conceived to facilitate the learning of an adaptive end-to-end driving policy by leveraging structured scenario information. FUSION capitalizes on the causal relationships between decomposed reward, cost, state, and action space, constructing a framework for structured sequential reasoning under dynamic traffic environments. We conduct rigorous evaluations in two typical settings of distribution shift for autonomous vehicles, demonstrating the good balance between safety cost and utility reward of FUSION compared to contemporary state-of-the-art safety-aware LfD baselines. Empirical evidence under diverse driving scenarios attests that FUSION significantly enhances the safety and generalizability of autonomous driving agents, even in the face of challenging and unseen environments. Furthermore, our ablation studies reveal noticeable improvements in the integration of causal representation into the safe offline RL problem.

1 Introduction

Learning from Demonstration (LfD) techniques, such as Imitation Learning (IL) and offline Reinforcement Learning (RL) [1, 2, 3, 4], have revolutionized end-to-end frameworks in autonomous vehicles. Nonetheless, the safety and generalizability of learning-based driving policies across diverse scenarios remain elusive [5, 6, 7]. These challenges become even more pronounced in intricate contexts involving complex vehicle-to-road and vehicle-to-vehicle interaction. Prior studies [8, 9] illustrate that minor domain shifts in road structures or surrounding vehicles can result in catastrophic outcomes, given the high-stakes nature of autonomous driving.

While existing research has successfully applied end-to-end learning-based algorithms to racing cars [10, 11, 12], urban driving scenarios remain a puzzle. The complexity arises from the fact that urban settings demand robust structural reasoning from context-rich, safety-critical situations [13]. For instance, *humans* can effortlessly adapt their driving behaviors based on static contexts like roadblocks or dynamic contexts such as surrounding traffic, often making intuitive judgments, as illustrated in Figure. 1. Although such abstraction is straightforward to humans with high reasoning

[*][1]Carnegie Mellon University, Pittsburgh, PA 15213 USA
[†][2]Ford Motor Company, Dearborn, MI 48126 USA

Symposium on Machine Learning for Autonomous Driving (ML4AD 2023).

capabilities, end-to-end approaches like vanilla deep RL methods usually fail due to the distribution shift under diverse driving scenarios, leading to a consequence of staying either over-conservative or over-aggressive. As a consequence, two pivotal challenges emerge under such distribution shifts: (i) striking a balance between safety and driving efficiency, and (ii) ensuring safety performance in unseen driving contexts.

Recent LfD advancements in autonomous driving have strived for safety improvements through various means, including agile actions [1, 3], object-centric world models [5, 6, 14, 15], safety-enhanced scene representation [7, 16], and structure-aware representation of multi-modal sensory inputs [17]. Moreover, techniques like domain-invariant IL [18] and hierarchical IL [19] further bolster the generation of a safe and universally applicable causal representation. However, a recurring limitation is the presupposition of access to perfect expert demonstrations, which are often unattainable in intricate urban scenarios.

To circumvent the reliance on impeccable expert demonstrations, researchers are pivoting towards RL-based techniques, encompassing offline RL [20, 21] and safe RL [22, 23]. These methodologies harbor the potential to equilibrate RL agents' priorities between safety and efficiency, especially when learning from non-expert demonstrations. Encouragingly, some studies [12, 24] even manage to surpass expert policies during online deployment by utilizing these batch RL methods, which draw from enhanced real-world data. However, a prevailing assumption – often misguided – is that online environments will mirror the dynamics of those from which offline trajectories were sourced. In reality, the scarcity and lack of diversity in available data, exacerbated when prioritizing safety, curtails the generalizability of offline RL. This is particularly apparent in autonomous driving, where both static (e.g., road layouts) and dynamic (e.g., traffic flow) contexts differ markedly across locales. As a result, achieving generalizability in unseen environments remains challenging.

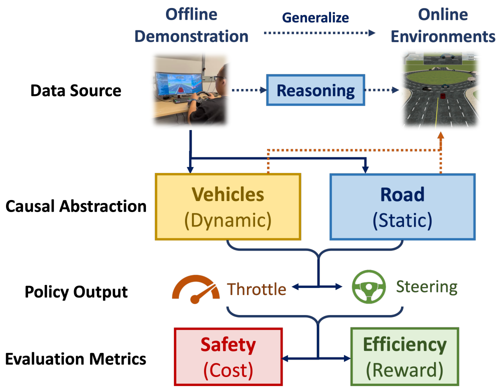

Figure 1: Diagram depicting offline-to-online generalization via a modular reasoning framework. The agent learns a causal abstraction from offline demonstration trajectories and then applies it to different environmental components during online implementation. This abstracted representation enables learning agile agents for unseen scenarios in a zero-shot manner while enhancing safety and efficiency.

In this study, we introduce sa**F**ety-aware str**U**ctural **S**cenario representat**ION** (FUSION), which aims to improve the generalizability of safety performance of self-driving cars in unseen scenarios. More concretely, our contributions are summarized as follows:

- We introduce a safety-aware offline reinforcement learning framework that successfully balances the trade-off between efficiency and safety, termed the Causal Ensemble World Model (CEWM).
- We develop a Causal Bisimulation Learning (CBL) paradigm that regularizes the state representation in a compact way, enabling better generalizability towards OOD state inputs during the online deployment stage.
- We provide comprehensive evaluations on the offline dataset collected from the human beings and Intelligent Driver's Model (IDM), showing the advantage of FUSION over existing baselines in offline safe IL and RL.

2 Problem Formulation

As stated in the Introduction section, this work aims to tackle a generalizable safe RL problem under some distribution shifts in different task domains. To better model such distribution shift, we follow the definition of contextual MDP in [25] to define the Constrained Contextual Markov Decision Process, or C^2-MDP, to model this generalizable safe RL problem as follows:

Definition 1. *Constrained Contextual Markov Decision Process (C^2-MDP) is a Contextual MDP with a tuple $\big(\mathcal{P}, \mathcal{M}(\omega)\big)$, where $\mathcal{M}$ is a function that maps any contexts $\omega \in \mathcal{P}$ to a constrained MDP $\mathcal{M}(\omega) = \big(\mathcal{S}, \mathcal{A}, T_\omega, r, c, s_0, \gamma\big)$,* where $T_\omega : \mathcal{S} \times \mathcal{A} \times \mathcal{P} \to \mathcal{S}$ is the context-specific transition function, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the cost function, s_0 is the initial state, and γ is the discount factor. C^2-MDP defines the safety cost as an intuitive additional performance preference for the driving agents. Also, it includes different MDPs according to different contexts ω. This additional context aims to model the phenomena that the traffic environment varies across different contexts (e.g. road types or traffic densities) in the autonomous driving scenarios.

Following the above definition, we introduce our problem formulation and then give a sketch of our proposed learning pipeline for generalizable safe RL problems in autonomous driving. Based on the Definition 1, the Constrained Contextual MDP aims to maximize the cumulative reward while satisfying the safety constraints over the cumulative expected cost under a certain target context ω. In formal terms, our problem can be defined as the following constrained optimization problem $\max_\pi \ J_r(\pi, \omega) \quad s.t. \ J_c(\pi, \omega) \le \kappa_c$, where we define the reward objective $J_r(\pi, \omega) \triangleq \mathbb{E}_{\omega,\pi} \sum_{t=1}^{T} r(s_t, a_t)$ and similarly the cost objective, $J_c(\pi, \omega) = \mathbb{E}_{\omega,\pi} \sum_{t=1}^{T} c(s_t, a_t)$.

To achieve generalizable safety, we aim to optimize a policy that satisfies safety constraints: $J_c(\pi, \omega) \le c, \forall \pi \in \Pi, \omega \in \Omega$, i.e. imposing constraint satisfaction under varying behavior policies π_β and environment contexts ω. Meanwhile, we assume that the preference of the reward function r and the cost function c remain unchanged across different contexts.

In our autonomous driving problem, the reward is composed of a forwarding reward in the longitude direction, a continuous reward for the vehicle speed, and an additional sparse reward once the vehicles reach the goal or other terminal states:

$$\begin{aligned} r_t &= w_1^r r_{\text{forward}} + w_2^r r_{\text{speed}} + w_3^r r_{\text{term}} \\ &= w_1^r (d_t - d_{t-1}) + w_2^r v_t + w_3^r \mathbb{I}(s_t = g) \end{aligned} \tag{1}$$

In our urban driving task: the safety cost comes from (i) collision with others, (ii) out-of-road conditions, and (iii) over-speeding. The collision and out-of-road costs are binary indicators that are 1 only when they occurs, and the over-speeding cost is a continuous cost that occurs once the vehicle goes over a certain speed limit v_{limit}.

$$\begin{aligned} c_t = & w_1^c c_{\text{collision}} + w_2^c c_{\text{out road}} + w_3^c c_{\text{overspeed}} \\ =& w_1^c \mathbb{I}(s \in s_{\text{collision}}) + w_2^c \mathbb{I}(s \notin s_{\text{road}}) + w_3^c \max(0, v_t - v_{\text{limit}}) \end{aligned} \tag{2}$$

The core problem in this paper is to learn a safe policy with good generalizability under some distribution shifts. In autonomous driving scenarios, we wish the agents could generalize (i) between offline data collected from mixed-quality policies and online environments, i.e. $\pi_\beta \neq \pi^*$, and (ii) between varying contexts of C^2-MDP, i.e. training environments ω_{train} for data collection are different from online testing environments ω_{test}. Such difference also indicates the difference in the MDP $\mathcal{M}(\omega_1) \neq \mathcal{M}(\omega_2)$. More specifically, we define the distribution shift in *transition dynamics* (e.g. the density of the traffic) as follows: $p(\cdot|s, a; \omega_{train}) \neq p(\cdot|s, a; \omega_{test})$.

3 Methodology

In this section, we zoom in on more details about our proposed FUSION with two important modules: (i) Causal Ensemble World Model (CEWM), and (ii) safety-aware Causal Bisimulation Learning (CBL).

3.1 Causal Ensemble World Model Learning

In autonomous driving problems, the entire state space can be decomposed into several disjoint subspaces [17], including the (estimated) ego navigation state, lidar observation, and visual observation, e.g. the birds-eye-view observation that serve as input to FUSION in Figure 2.

Definition 2 (Factorizable State Space). *The factorizable state space in MDP indicates a disjoint state space decomposition, where $S = S_1 \cup S_2 \cup \cdots \cup S_N$, and N indicates how many disjoint state components we have in a certain problem.*

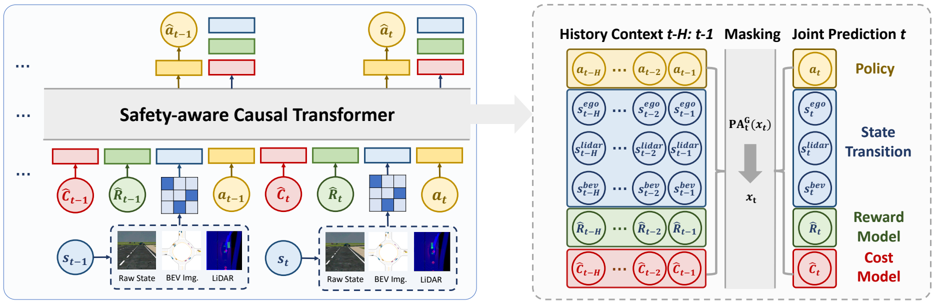
$\hat{a}_{t-1}$
$\hat{a}_t$
...
Safety-aware Causal Transformer
$\hat{C}_{t-1}$
$\hat{R}_{t-1}$
a_{t-1}
$\hat{C}_t$
$\hat{R}_t$
a_t
s_{t-1}
s_t
Raw State
BEV Img.
LiDAR
History Context t-H: t-1
Masking
Joint Prediction t
a_{t-H} ... a_{t-2} a_{t-1}
s^{ego}_{t-H} ... s^{ego}_{t-2} s^{ego}_{t-1}
s^{lidar}_{t-H} ... s^{lidar}_{t-2} s^{lidar}_{t-1}
s^{bev}_{t-H} ... s^{bev}_{t-2} s^{bev}_{t-1}
$\hat{R}_{t-H}$ ... $\hat{R}_{t-2}$ $\hat{R}_{t-1}$
$\hat{C}_{t-H}$ ... $\hat{C}_{t-2}$ $\hat{C}_{t-1}$
$\mathrm{PA}^{\mathcal{G}}_t(x_t)$
x_t
a_t
s^{ego}_t
s^{lidar}_t
s^{bev}_t
$\hat{R}_t$
$\hat{C}_t$
Policy
State Transition
Reward Model
Cost Model

Figure 2: Overview of Safety-aware structural Scenario Representation Framework. The left diagram shows a safety-aware decision transformer that conducts sequential decision-making based on the temporal contexts. The right diagram shows the general form of the graphical model in the CEWM and Policy Learning modules in FUSION, where the connection between different timesteps will be determined by the attention weights in the causal transformer. The nodes at a later timestep depend on their parental nodes in the previous timesteps.

To help the FUSION framework gain better awareness of the structure of the state and action space, we propose the CEWM based on multi-modal observations, as is defined The factorized state space Definition 2, along with the reward, cost, and action variables, form the nodes in this world model. To better describe the structural dependency between them, we further design the CEWM according to the following definition of Structured Causal Model (SCM).

Definition 3. *An SCM $(\mathcal{S}, \mathcal{E})$ consists of a set of variables $\mathcal{S}$, along with d functions [26],*

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

where $\mathbf{PA}^{\mathcal{G}}_j \subset \{s_1, \dots, s_d\} \backslash \{s_j\}$ are called parents of s_j in the Directed Acyclic Graph (DAG) $\mathcal{G}$, and $\mathcal{E} = \{\epsilon_1, \dots, \epsilon_d\}$ follows a joint distribution over the noise variables, which are required to be jointly independent.

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 $\theta$ with CEWM by (4);
    Update Encoder $\phi$ 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: $a_t \leftarrow \pi_{\theta,\phi}(\mathbf{o})$ ;
```

For general offline RL problems, SCM aims to jointly parameterize the world model as well as the policy model between different nodes in the state, action, reward, and safety cost. In order to parameterize the functions f in this SCM, we use a Safety-aware Causal Transformer, as is shown in Figure 2. For instance, the child node s_j is determined by its parent tokens $\mathbf{PA}^{\mathcal{G}}_t(s_j)$ in the previous tokens $\tau_{t-H:t}$, and the exogenous noise variable ϵ_j, which are aggregated by a variable-specific function f_j empowered by the attention mechanism of Transformer. The edges between different nodes represent their causal dependency in the spatiotemporal domain, which is essentially captured by the attention weights, as we will discuss later in Figure 5 of the experiment parts. Besides capturing the cause-and-effect relationship between the reward, cost, and factorizable state space, the SCM also enjoys a great property in that the child nodes (e.g. the state and reward/cost in subsequent timesteps) are only dependent on their parent nodes (in the state or action space in the previous timesteps), while removing the unnecessary dependencies between the descendent nodes to indirect ancestors or non-parent nodes. Such property improves both generalizability and efficiency for an autoregressive inference during the online deployment.

Based on this property, we derive the CEWM under the SCM, which can then be decomposed into the following disjoint components, including the reward-to-go model, cost-to-go model, the factorized state-action transition dynamics, and the policy optimization, as is shown below:

$$
\begin{aligned}
&p(\tau_t|\tau_{t-H:t}) = p(a_t, s_t, R_t, C_t|a_{t-1}, s_{t-1} \cdots R_{t-H}, C_{t-H}) \\
&= \underbrace{p\Big(r_t|\mathbf{PA}_t^G(r_t)\Big)}_{\text{Reward-to-go}} \underbrace{p\Big(c_t|\mathbf{PA}_t^G(c_t)\Big)}_{\text{Cost-to-go}} \underbrace{p\Big(a_{t+1}|\mathbf{PA}_t^G(a_{t+1})\Big)}_{\text{Policy Optimization}} \underbrace{\prod_{i \in \dim(S)} p\Big(s_{t+1}^i|\mathbf{PA}_t^G(s_{t+1}^i)\Big)}_{\text{Factorized Dynamics}}
\end{aligned} \tag{3}
$$

Therefore, we exert an auxiliary task of trajectory optimization in the optimization process of safety-aware decision transformer to estimate the three components in (3), i.e.

$$
\begin{aligned}
\mathcal{L}_{\text{traj}} &= -\log p(\tau_{t+1}|\tau_{t-H:t}) = -\log p(R_t|\mathbf{PA}_t^G(R_t)) \\
&-\log p(C_t|\mathbf{PA}_t^G(C_t)) - \log p(a_{t+1}|\mathbf{PA}_t^G(a_{t+1})) - \sum_{i \in \dim(S)} \log p(s_{t+1}^i|\mathbf{PA}_t^G(s_{t+1}^i)) \\
&= \mathcal{L}_{\text{rtg}} + \mathcal{L}_{\text{ctg}} + \mathcal{L}_{\text{act}} + \mathcal{L}_{\text{dyn}}
\end{aligned} \tag{4}
$$

This trajectory optimization objective benefits our safety-aware DT with better structural awareness of the trajectory level between the state, action, reward-to-go, and cost-to-go. The design of this safety-aware DT model manages to parameterize CEWM that we propose in (3), as the latter token is generated conditioned on the previous tokens in an auto-regressive way.

3.2 Safety-aware Bisimulation Learning

Though CEWM provides an *explicit* structure to model the causality, learning such a model from offline datasets is non-trivial. The reason is that demonstrations in the mixed-quality dataset have diverse levels of safety due to spurious correlations between actions and states. To avoid getting misled by such spurious correlation, we introduce an additional self-supervised regularization term in an *implicit* way, namely Causal Bisimulation Learning, or CBL.

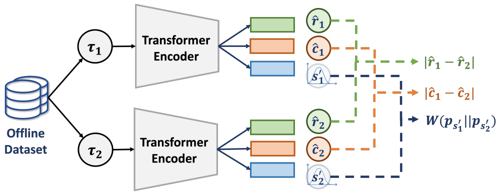

Figure 3: Safety-aware bisimulation metrics with the distribution distance in transition dynamics, rewards, and safety cost.

Inspired by the DBC algorithm for off-policy RL in [27], we further regularize the FUSION model with safety-aware Bisimulation Learning in our offline RL setting. We first extend the traditional bisimulation relationships for MDP in [27, 28] with an extra safety term:

Definition 4 (Safety-aware Bisimulation Relation)**.** *A safety-aware bisimulation relation* $\mathcal{U} \subset \mathcal{S} \times \mathcal{S}$ *is a binary relation which satisfies,* $\forall (s_1, s_2) \in \mathcal{U}$*:* $\forall a \in \mathcal{A}, r(s_1, a) = r(s_2, a)$*,* $\forall a \in \mathcal{A}, c(s_1, a) = c(s_2, a)$*,* $\forall a \in \mathcal{A}, s' \in \mathcal{S}, p(s'|s_1, a) = p(s'|s_2, a)$*.*

Intuitively, in the Constrained MDP setting, the bisimilarity between two states is not only determined by the step-wise reward and transition dynamics but also by their similarity in the step-wise cost. In practice, the reward, cost, and transition dynamics could hardly match exactly for two different states, therefore, we propose a smooth alternative [29] of safety-aware bisimulation relationship, denoted as Safety-aware Bisimulation Metrics as is shown in Figure 3:

```
Algorithm 2: Safety-aware CBL
  Data: Offline (mixed) trajectories from
        controller/human, cost limit $C$
  Result: Policy $\pi$'s state encoder $\phi$
  for $k = 0, \cdots, N-1$ do
      Sample minibatch: $\mathcal{B} \leftarrow$ Sample($\mathcal{D}_{\pi_\beta}$);
      Construct transition pairs: $(s_1, a_1, s'_1) \leftarrow \mathcal{B}$ ;
      Permute samples: $(s_2, a_2, s'_2) \leftarrow$ permute($\mathcal{B}$) ;
      Compute bisimulation distance: With (5) ;
      Update encoder: $\phi_{k+1} \leftarrow \phi_k - \nabla_\phi \mathcal{L}_{\text{bisim}}$
        with (6);
```

Definition 5 (Safety-aware Bisimulation Metrics)**.** *The bisimulation metric* $d^\pi : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^+$ *is a mapping from the joint state space to a non-negative scalar. According to* *the definition of the safety-aware bisimulation relationship, the distance is defined as:*

$$
\begin{aligned}
d^{\pi}(s_1, s_2) = \mathbb{E}_{\substack{a_1 \sim \pi(\cdot|s_1),\\ a_2 \sim \pi(\cdot|s_2)}} \Big[&|r(s_1, a_1) - r(s_2, a_2)| \\
&+ \lambda |c(s_1, a_1) - c(s_2, a_2)| + \gamma W_2(\hat{p}(\cdot|s_1, a_1), \hat{p}(\cdot|s_2, a_2)) \Big],
\end{aligned}
\tag{5}
$$

The Lagrangian multiplier λ aims to balance the safety term, and $W_2(\cdot,\cdot)$ is the 2-Wasserstein distance measuring the similarity between two transition dynamics distribution. We use the following learning objectives to align the state representation with the bisimulation metrics in the latent space:

$$
\mathcal{L}_{\text{bisim}} = \mathbb{E}_{s_1, s_2 \sim p_{\pi_\beta}} \Big(\|\phi(s_1) - \phi_{sg}(s_2)\|_1 - d^{\pi}(s_1, s_2) \Big)^2, \tag{6}
$$

where ϕ_{sg} means stop gradient of state encoder ϕ.

In the inference time, we greedily exploit the value prediction in the online inference time, as is shown in Algorithm 1. Notably, we take the minimum cost-to-go preference and cost prediction, and the maximum reward-to-go preference and reward prediction at each step. This strategy aims to improve the safety and efficiency of FUSION given the preference in the online deployment stage.

4 Experiments

In this section, we first go through the environments and evaluation protocols that we use based on the MetaDrive simulator [30]. Next, we conduct experiments and ablation studies to answer four research questions, aiming to demonstrate how well our proposed methods could learn a safe and generalizable policy based on the offline driver's data. The evaluation results illustrate the effectiveness of the FUSION model.

4.1 Experiment Setup

Evalation Environment We evaluate our algorithm on MetaDrive [30], a light-weighted, realistic, and diverse autonomous driving simulator, which can specifically test the generalizability of learned agents on unseen driving environments with its capability to generate an unlimited number of scenes with various road networks and traffic flows.

The observation of the agents consists of the following components: (i) the ego states and navigation information, which contains the estimation of the ego vehicle's relative position with respect to the closest waypoint for navigation; (ii) the LiDAR observation with 240 laser bins; (iii) the Birds-eye-view (BEV) observation of the ego vehicle and surrounding vehicles, which is an 84×84×5 multi-channel image that describes the road contexts and the past trajectories of surrounding vehicles.

We collect the offline dataset by IDM polices [31] with diverse levels and styles of aggressiveness of the ego and surrounding drivers. We manually set different acceleration and deceleration rates to adjust the aggressiveness level in the IDM policy. The total offline dataset consists of 2,000 trajectories with more than 400,000 timesteps under six different road contexts.

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	**139.95±4.24**	**0.52±0.06**	**0.90±0.03**	**117.40±4.30**	**0.90±0.14**	**0.82±0.04**

Table 1: Evaluation Performance in both policy mismatch and dynamics mismatch settings. Evaluation metrics include (i) utility reward, the higher the better; (ii) safety cost, the lower the better; and (iii) success rate, the higher the better. Each of the baseline results is evaluated under 5 random seeds. **Bold** means the best.

Our evaluation protocol includes the following three metrics: The **Utility Reward** metric evaluates the efficacy and efficiency of autonomous vehicles to finish the task, which is a weighted combination of the cumulative driving distance, driving speed, and waypoint arrival, as is introduced in (1). The **Safety Cost** metric evaluates the overall safety level of autonomous vehicles, which comes from three safety-critical scenarios in autonomous driving, including collision, out-of-lane, and over-speed, as is defined in (2). The speed limit v_{limit} is set to be 40 *kph*. The **Success Rate** metric indicates the ratio of episodes in which the agent successfully reaches the destination within a maximum number of timesteps.

We test our methods in six different types of road configurations (see Figure 4). As introduced in (2), the safety violation costs result from three sources: (i) collision, (ii) out-of-lane, and (iii) over-speed. The cost for collision and out-of-lane is 1 at each occurrence, and the over-speed cost $c_{\text{speed}} = \max\{0, 0.02(v - v_{\text{limit}})\}$. An episode will terminate if any one of the risky scenarios (i) (ii) happens, or the overall timestep is greater than a preset decision horizon of 1,000. When the agent reaches the destination without any collision or getting off the road, it will be counted as a success.

We compare our proposed methods and baselines in the following two settings. **Policy Mismatch** stands for the case where the offline dataset is sampled from the non-perfect expert policy, and the agents need to tackle the generalization challenge from mixed-quality and potentially unsafe offline data towards the deployment in the online environment. **Dynamics mismatch** stands for the case where the agent needs to tackle another generalization challenge from the training environments (where the offline data is collected) with sparser traffic flows, towards the testing environments where the traffic flows are 1.5× denser than the training.

Baselines We illustrate our results by comparing FUSION against two types of baselines: (i) safe imitation learning and (ii) offline safe reinforcement learning. Specifically, the implementation of these baselines aims to solve the multi-modal sensory inputs in the sequential decision-making problems of autonomous driving.

IL-based methods select safe trajectories or conduct uncertainty quantification to avoid getting into uncertain and unsafe regions. This kind of baselines includes Safe Behavior Cloning (**Safe BC** [3]) that only uses safe trajectories to train the agent, Invariant Causal Imitation Learning (**ICIL** [18]) that derives invariant state abstraction to learn generalizable policies by the model ensemble, like **GSA** [19] and **BNN** [17], which both use hierarchical state abstraction in generalizable decision making.

On the other hand, the offline Safe RL baselines generally solve a constrained optimization problem of C^2-MDP by adding Lagrangian terms in the policy evaluation step. Two of them are BEAR Lagrangian (**BEAR-Lag**) and BCQ Lagrangian (**BCQ-Lag**), which are safety-aware variants of Offline RL algorithms BEAR [21] and BCQ [20], respectively. Constrained Penalized Q-Learning (**CPQ** [22]) aims to learn safe policy by penalizing the cost from the offline dataset. All the Offline Safe RL baselines set episodic cost constraint threshold $\kappa_c = 1$. Based on the design of the safety cost introduced in Section 4.1, when the episodic cost is lower than 1, it means no critical violence, including collision and out-of-lane, occurred in this episode.

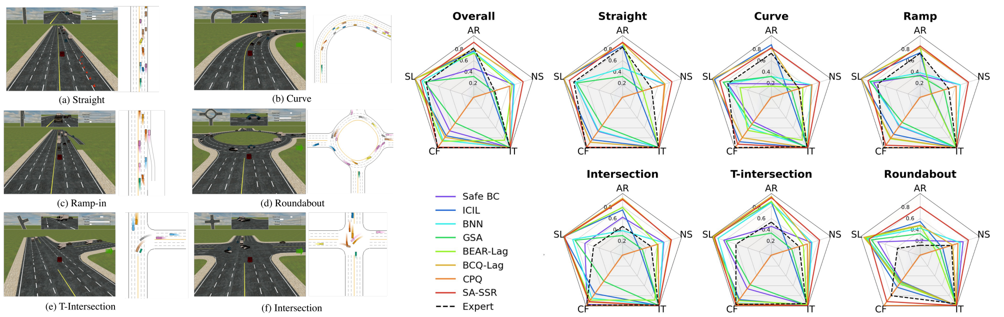
(a) Straight
(b) Curve
(c) Ramp-in
(d) Roundabout
(e) T-Intersection
(f) Intersection

Figure 4: The left figure shows the diverse scenarios of multiple road configurations in MetaDrive. The right figure shows the analysis of the results under each of them. We compare the performance of FUSION on different types of roads with Safe IL and offline RL baselines, as well as the expert policy for offline RL. The larger lidar plot on each coordinate stands for the safer performance in each safety metric. (**AR**: Arrival, **NS**: Not speeding, **IT**: In-time, **CF**: Collision-free, **SL**: Stay in-lane.)

4.2 Results and Analysis

We design experiments and corresponding ablation studies to answer the following important research questions:

- **(RQ1)** How does FUSION perform with non-perfect offline data with diverse behavior policies from IDM and humans, compared with Safe Offline IL and RL baselines?
- **(RQ2)** How does FUSION perform under unseen dynamics that the offline dataset does not cover, compared with all the baselines?
- **(RQ3)** Can FUSION consistently outperform other baselines and expert policies under diverse road contexts?
- **(RQ4)** Do sequential modeling and causal representation learning benefit FUSION in capturing spatio-temporal dynamics contexts?

For **RQ1** and **RQ2**, we compare FUSION against the aforementioned baselines in both policy mismatch and dynamics mismatch settings. The results in Table 1 clearly demonstrate the advantages of FUSION against baselines in both safety cost and driving reward performance. (i) In the policy mismatch setting where the agent needs to overcome the suboptimality of the offline data, FUSION performs better in the reward (driving efficiency), cost (safety performance), and success rate. Notice that all the Safe IL baselines failed to learn a low-cost driving policy because these IL-based methods do not have explicit cost or reward feedback, and only fitting on those safe state and action transition pairs are insufficient to satisfy the safety requirements due to the imperfection of the offline demonstrations. Meanwhile, the Safe RL baselines seem to perform better, as they explicitly constrain the learned target policy with a preset cost threshold. The actor-critic framework that alternates between policy improvement and policy evaluation could implicitly guide the target policy to avoid some low-reward or high-cost behaviors. However, CPQ seems to be overly conservative in that it fails to balance efficiency and safety, thus always procrastinating near the starting zone to avoid getting a large cost penalty. On the other hand, ICIL, BNN, BEAR-Lag, and BCQ-Lag seem to have high success rates in policy mismatch settings, yet FUSION could still outperform them by a large margin (over 10%). (ii) In the dynamics mismatch case where the online testing environments have significantly different traffic dynamics and different types of roadblocks from the training environments, the performance gap between our methods and other baselines even enlarged, for example, we can see the success rate of Bear-Lag and BCQ-Lag drops by 40%, and the evaluation cost of BCQ-Lag also violates the cost constraints. In contrast, although FUSION has a slightly lower reward than what it has in policy mismatch, the cost is still below the set threshold 1, and the success rate is also significantly higher than other baselines by over 30%.

For **RQ3**, we take a deeper look at the exact driving performance in different road contexts in Figure 4. We provide a failure mode analysis in the following radar plot (Figure 4), the larger the pentagon is, the better overall safety performance it has. The plot provides an episode-wise frequency of five different safety behavior categories, including (i) **AR**: arrival rate among all episodes; (ii) **NS**: not speeding in the episode, which counts the ratio of timestep where the agent exceeds a speed limit of 40 *kph* on the urban local roads; (iii) **IT**: in-time (complete the route within the time limit of 1,000 steps per episode); (iv) **CF**: collision-free in a single episode; (v) **SL**: stay in-lane without violating the lane constraints. The result shows that our proposed FUSION (red) can consistently outperform the expert policies (grey with shadows) and other baselines, especially in the hardest Roundabout environment.

For **RQ4**, we provide additional ablation studies in Table 2. We compare FUSION with three of its variants: (i) **FUSION-Short**, which uses shorter context in the safety-aware transformer to model the whole sequence; (ii) **FUSION w/o CEWM**, which does not consider the causal ensemble world model learning, and only uses the behavior cloning term as supervised signals; (iii) **FUSION w/o CBL**, which neglects the safety-aware bisimulation learning. The result confirms that the FUSION benefits from all its design, including the spatiotemporal information from CEWM and additional safety awareness in the transformer model via CBL.

Furthermore, we visualize the learned attention map of FUSION's safety-aware causal transformer in Figure 5. The x-axis represents the source (previous) nodes and the y-axis represents the target (future) nodes. The attention map is a low-triangular matrix because only the tokens of previous timesteps affect the tokens in the future.

We find that FUSION has a clear hierarchy in the attention map: (i) the attention map of the first layer is actually very sparse, as FUSION only attends tokens from previous *one* timestep, which essentially models the whole decision-making process in a Markovian manner. (ii) FUSION attends the preference tokens that include cost-to-go (red) and reward-to-go (green) to the future state and action tokens, trying to balance both of them for the decision-making process in a long horizon. (iii) FUSION captures world dynamics and policy by attending previous states (blue) to the future value prediction and action nodes. Such semantically meaningful interpretation as well as the heterogeneity of attention weights on different layers indicate that FUSION benefits from CEWM by hierarchically capturing structural information reflected in the attention maps. In contrast, as shown in the second row of Figure 5, FUSION without CEWM does not capture the above sparsity and interpretability, yet the second and third layers tend to resemble each other. The reason is that the variant without CEWM lacks sequential awareness with more informative training signals during the offline training stage.

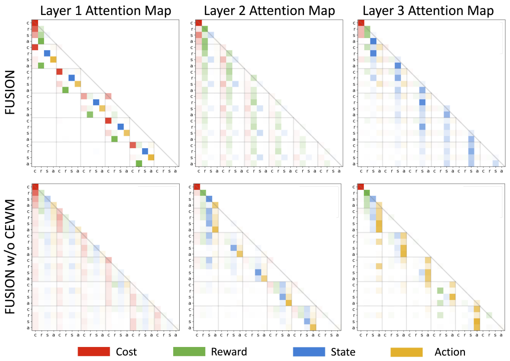

Figure 5: Visualization of Attention matrix on the same trajectories. We compare different layers of the attention map of two models: FUSION and FUSION without CEWM. More visualization over diverse trajectories is on our website.

Variants	**Policy Mismatch**			**Dynamics Mismatch**		
	Reward (↑)	Cost (↓)	Succ. Rate (↑)	Reward (↑)	Cost (↓)	Succ. Rate (↑)
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
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
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	131.32±29.60	16.02±9.46	0.81±0.15	129.71±28.84	17.58± 9.71	0.72±0.20

Table 2: Ablation studies on FUSION's variants. **Bold** means the best.

5 Conclusions

In this paper, we propose FUSION, a trustworthy autonomous driving system with a causality-empowered safe reinforcement learning algorithm in an offline setting. We first design a safety-aware causal transformer termed CEWM to model the causal relationship between the state space, reward value, and cost value at different timesteps. Then we regularize the learned representation in CEWM with a CBL to enforce their compactness via safety-aware bisimulation in an implicit way, then greedily infer the action during online deployment. Exhaustive empirical results show that our method consistently outperforms offline demonstration and several strong baselines in safe IL or offline safe RL under diverse urban autonomous driving scenarios. We also conduct extensive analysis to analyze the benefits of different modules that we design in FUSION and show a comprehensive and interpretable evaluation of FUSION against its variants or other baselines. One potential limitation is that all the experiments are conducted in the MetaDrive simulator since it is more portable than CARLA or other autonomous driving simulators. It would be interesting to extend FUSION's framework to other autonomous vehicle simulators with higher fidelity, as well as the multi-agent RL settings in the near future.

References

[1] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[2] Jianyu Chen, Bodi Yuan, and Masayoshi Tomizuka. Model-free deep reinforcement learning for urban autonomous driving. In *2019 IEEE intelligent transportation systems conference (ITSC)*, pages 2765–2771. IEEE, 2019.

[3] Yunpeng Pan, Ching-An Cheng, Kamil Saigol, Keuntaek Lee, Xinyan Yan, Evangelos A Theodorou, and Byron Boots. Imitation learning for agile autonomous driving. *The International Journal of Robotics Research*, 39(2-3):286–302, 2020.

[4] Jianyu Chen, Shengbo Eben Li, and Masayoshi Tomizuka. Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5068–5078, 2021.

[5] Wenhao Ding, Haohong Lin, Bo Li, and Ding Zhao. Causalaf: causal autoregressive flow for safety-critical driving scenario generation. In *Conference on Robot Learning*, pages 812–823. PMLR, 2023.

[6] Katrin Renz, Kashyap Chitta, Otniel-Bogdan Mercea, A Koepke, Zeynep Akata, and Andreas Geiger. Plant: Explainable planning transformers via object-level representations. *arXiv preprint arXiv:2210.14222*, 2022.

[7] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning*, pages 726–737. PMLR, 2023.

[8] Wenhao Ding, Chejian Xu, Mansur Arief, Haohong Lin, Bo Li, and Ding Zhao. A survey on safety-critical driving scenario generation—a methodological perspective. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[9] Mengdi Xu, Zuxin Liu, Peide Huang, Wenhao Ding, Zhepeng Cen, Bo Li, and Ding Zhao. Trustworthy reinforcement learning against intrinsic vulnerabilities: Robustness, safety, and generalizability. *arXiv preprint arXiv:2209.08025*, 2022.

[10] Florian Fuchs, Yunlong Song, Elia Kaufmann, Davide Scaramuzza, and Peter Dürr. Superhuman performance in gran turismo sport using deep reinforcement learning. *IEEE Robotics and Automation Letters*, 6(3):4257–4264, 2021.

[11] Peter R Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, et al. Outracing champion gran turismo drivers with deep reinforcement learning. *Nature*, 602(7896):223–228, 2022.

[12] Dhruv Shah, Kyle Stachowicz, Arjun Bhorkar, Ilya Kostrikov, and Sergey Levine. Fastrlap: A system for learning high-speed driving via deep rl and autonomous practicing. In *ICRA2023 Workshop on Pretraining for Robotics (PT4R)*, 2023.

[13] Aditya Mohan, Amy Zhang, and Marius Lindauer. Structure in reinforcement learning: A survey and open problems. *arXiv preprint arXiv:2306.16021*, 2023.

[14] Wenhao Ding, Haohong Lin, Bo Li, and Ding Zhao. Generalizing goal-conditioned reinforcement learning with variational causal reasoning. In *Advances in Neural Information Processing Systems*, 2022.

[15] Wenhao Ding, Laixi Shi, Yuejie Chi, and Ding Zhao. Seeing is not believing: Robust reinforcement learning against spurious correlation. *arXiv preprint arXiv:2307.07907*, 2023.

[16] Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21983–21994, 2023.

[17] Keuntaek Lee, Ziyi Wang, Bogdan Vlahov, Harleen Brar, and Evangelos A Theodorou. Ensemble bayesian decision making with redundant deep perceptual control policies. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 831–837. IEEE, 2019.

[18] Ioana Bica, Daniel Jarrett, and Mihaela van der Schaar. Invariant causal imitation learning for generalizable policies. *Advances in Neural Information Processing Systems*, 34:3952–3964, 2021.

[19] Riad Akrour, Filipe Veiga, Jan Peters, and Gerhard Neumann. Regularizing reinforcement learning with state abstraction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 534–539. IEEE, 2018.

[20] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.

[22] Haoran Xu, Xianyuan Zhan, and Xiangyu Zhu. Constraints penalized q-learning for safe offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8753–8760, 2022.

[23] Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constrained decision transformer for offline safe reinforcement learning. *arXiv preprint arXiv:2302.07351*, 2023.

[24] Xing Fang, Qichao Zhang, Yinfeng Gao, and Dongbin Zhao. Offline reinforcement learning for autonomous driving with real world driving data. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 3417–3422. IEEE, 2022.

[25] Baiming Chen, Zuxin Liu, Jiacheng Zhu, Mengdi Xu, Wenhao Ding, Liang Li, and Ding Zhao. Context-aware safe reinforcement learning for non-stationary environments. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10689–10695. IEEE, 2021.

[26] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[27] Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

[28] Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. In *AI&M*, 2006.

[29] Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *UAI*, volume 4, pages 162–169, 2004.

[30] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[31] Arne Kesting, Martin Treiber, and Dirk Helbing. General lane-changing model mobil for car-following models. volume 1999, pages 86–94. SAGE Publications Sage CA: Los Angeles, CA, 2007.

[32] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.

[33] Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Steven Wu, Bo Li, and Ding Zhao. Constrained variational policy optimization for safe reinforcement learning. In *International Conference on Machine Learning*, pages 13644–13668. PMLR, 2022.

[34] Kunal Menda, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Ensembledagger: A bayesian approach to safe imitation learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5041–5048. IEEE, 2019.

[35] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

[36] Zuxin Liu, Zijian Guo, Haohong Lin, Yihang Yao, Jiacheng Zhu, Zhepeng Cen, Hanjiang Hu, Wenhao Yu, Tingnan Zhang, Jie Tan, et al. Datasets and benchmarks for offline safe reinforcement learning. *arXiv preprint arXiv:2306.09303*, 2023.

[37] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

[38] Haochen Liu, Zhiyu Huang, Xiaoyu Mo, and Chen Lv. Augmenting reinforcement learning with transformer-based scene representation learning for decision-making of autonomous driving. *arXiv preprint arXiv:2208.12263*, 2022.

[39] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286, 2021.

[40] Antonio Loquercio, Mattia Segu, and Davide Scaramuzza. A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, 5(2):3153–3160, 2020.

[41] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018.

[42] Robert Dadashi, Shideh Rezaeifar, Nino Vieillard, Léonard Hussenot, Olivier Pietquin, and Matthieu Geist. Offline reinforcement learning with pseudometric learning. In *International Conference on Machine Learning*, pages 2307–2318. PMLR, 2021.

A Additional Related Works

Safety-aware Decision Making from Offline Data. To bring up safety awareness of autonomous vehicles, most of the recent works formulate the safe decision-making problem with constrained optimization [7, 32, 33]. Still, there have been several different roadmaps for solving this problem. For the IL-based approach, [17, 34] propose implicit safe constraints in IL via uncertainty quantification and Bayesian abstraction from the expert data. These approaches depend their *safety* on the small discrepancy between the learned trajectory and expert trajectory. More explicitly, InterFuser [7] proposes a safe controller that utilizes interpretable intermediate features to directly constrain the controller output within a safety set. On the other hand, offline Reinforcement Learning (RL) agents manage to balance safety and efficiency with additional reward, cost, and cost threshold information along the trajectories [35, 36]. To fully extract the temporal information in offline trajectories, recent works turn offline RL into a sequential modeling problem by utilizing the power of transformers [23, 37, 38, 39]. However, most of these works ignore the inherent structures of MDP in either spatial or temporal domain, which limits the policy's generalizability.

State Abstraction for Decision Making. To improve the performance of decision-making agents with some extra structural information, some recent works have focused on deriving state abstraction for generalizable decision-making using representation learning tricks. In the IL realm, [18] proposes Invariant Causal Imitation Learning (ICIL) to deal with the distribution shift with domain-invariant causal features. Based on uncertainty quantification, [17, 34, 40] propose ensemble representation that leverages multi-modal sensor inputs to improve the generalizability for self-driving agents. PlanT [6] proposes a learnable planner module based on object-centric representations. The RL field has seen developments in state abstraction through self-supervised learning methods, including time contrastive learning [41], hierarchical skill decomposition [19] and deep bisimulation metric learning [27, 42]. In autonomous driving applications, state and action space are usually factorizable, [14, 15] propose to train RL agents under the guidance of causal graphs to improve generalizability by discovering the latent structure in the world or policy model. The intersection of state abstraction with offline Safe RL is unexplored, yet crucial, as enhanced LfD can advance real-world autonomous driving.