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

Haohong Lin[1], Wenhao Ding[1], Zuxin Liu[1], Yaru Niu[1], Jiacheng Zhu[1], Yuming Niu[2] and Ding Zhao[1]

***Abstract*—In the domain of autonomous driving, the offline Reinforcement Learning (RL) approaches exhibit notable efficacy in addressing sequential decision-making problems from offline datasets. However, maintaining safety in diverse safety-critical scenarios remains a significant challenge due to long-tailed and unforeseen scenarios absent from offline datasets. In this paper, we introduce the saFety-aware strUctured Scenario representatION (FUSION), a pioneering representation learning method in offline RL to facilitate the learning of a generalizable end-to-end driving policy by leveraging structured scenario information. FUSION capitalizes on the causal relationships between the decomposed reward, cost, state, and action space, constructing a framework for structured sequential reasoning in dynamic traffic environments. We conduct extensive evaluations in two typical real-world settings of the distribution shift in autonomous vehicles, demonstrating the good balance between safety cost and utility reward compared to the current state-of-the-art safe RL and IL baselines. Empirical evidence in various 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 offline safe RL algorithm. Our code implementation is available on the project website.**

***Index Terms*—Intelligent Transportation Systems, Representation Learning, Reinforcement Learning**

I. Introduction

LEARNING from Demonstration (LfD) techniques have achieved huge success in autonomous driving [1]–[3] by improving the representation quality in an end-to-end framework. Among all the solutions categorized as LfD, Offline Reinforcement Learning has shown its superiority in many other robotic tasks, including locomotion and manipulation [4]. However, in the context of autonomous driving, the safety and generalizability of learning-based policies in various safety-critical scenarios remain elusive [5]–[7]. The distribution shift between offline training samples and online testing environments makes it harder to deploy the learning algorithms to the online environments safely. Prior studies [8], [9] illustrate that even minor domain shifts in road structures or surrounding vehicles can result in catastrophic outcomes due to the high-stakes nature of autonomous driving.

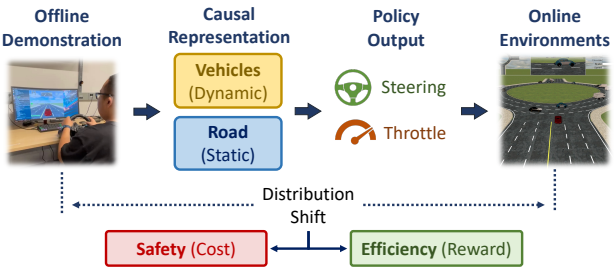

Fig. 1. Diagram depicting offline-to-online generalization via a modular reasoning framework. The agent learns a causal abstraction from offline demonstration trajectories and then applies it to different environmental components during online implementation. The distribution shift between offline datasets and online environment can lead to unsatisfying safety or efficiency in driving performance. This abstracted representation enables learning agile agents for unseen scenarios in a zero-shot manner while enhancing safety and efficiency.

Although existing research has successfully applied end-to-end learning-based algorithms to racing cars [10]–[12], urban driving scenarios remain complex for existing learning-based agents. Complexity arises from the fact that urban settings require structural reasoning in context-rich and safety-critical situations [13]. For instance, *humans* can effortlessly adapt their driving behaviors based on static contexts such as roadblocks or dynamic contexts such as surrounding traffic, often making intuitive judgments, as illustrated in Figure. 1. Although such causal abstraction is straightforward to humans with high reasoning capabilities, end-to-end approaches, such as vanilla deep RL methods, usually fail due to the distribution shift in various driving scenarios neglecting the underlying structures of the scenarios and usually resulting in being over-conservative or over-aggressive. As a consequence, two pivotal challenges emerge under such distribution shifts: (i) ensuring safety performance unseen driving contexts and (ii) striking a balance between safety and driving efficiency.

Recent LfD advances in autonomous driving improve the trustworthiness of learned policies through representation learning in offline RL or IL, including the object-centric representation [6], safety-enhanced scene representation [7], [14], multi-modal sensory representation [15], domain-invariant state representation [16], agile action abstraciton [2], and hierarchical action representation [17]. However, a recurring limitation of these representation learning works is the assumption of access to perfect expert demonstrations, which may not be accessible in diverse urban scenarios.

To mitigate the reliance on perfect expert demonstrations, multiple offline RL [18], [19] and safe RL [20], [21] approaches

Manuscript received: October 29, 2023; Revised February 4, 2024; Accepted February 26, 2024.

This paper was recommended for publication by Editor Jens Kober upon evaluation of the Associate Editor and Reviewers' comments.

[1]Haohong Lin, Wenhao Ding, Zuxin Liu, Yaru Niu and Ding Zhao are all with the Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA. {haohongl, wenhaod, zuxinl, yarun, jzhu4}@andrew.cmu.edu, dingzhao@cmu.edu,

[2]Yuming Niu is with the Ford Motor Company, Dearborn, MI 48126 USA. {yniu4}@ford.com,

Digital Object Identifier (DOI): see top of this page.

have been proposed. These methodologies harbor the potential to equilibrate the RL agents' priorities between safety and efficiency, especially when learning from non-expert demonstrations. Encouragingly, some studies [12], [22] manage to surpass expert policies during online deployment by using these batch RL methods, which are based on improved real-world data. Although these works overcome the limitation of perfect expert demonstration, they mostly assume that online environments will mirror the dynamics of those from which offline trajectories were collected. In reality, the scarcity and lack of diversity of high-quality expert data always exist and lead to significant distribution mismatch between training and deployment. This is particularly apparent in autonomous driving, where static (e.g., road layouts) and dynamic (e.g., traffic flow) contexts differ markedly across locales. How to achieve generalizability in unseen scenarios remains an open research question.

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

- We introduce a safety-aware offline reinforcement learning framework that aims to improve generalizability under distribution shifts during the online deployment stage.
- We develop a self-supervised causal representation learning paradigm to regularize the scenario representation, encouraging a better balance between the safety and efficiency of the learned policies.
- 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 the existing state-of-the-art approaches in offline safe RL [23] and IL-based methods [15]–[17].

II. Related Works

Safety-aware Decision Making from Offline Data. To bring up safety awareness of autonomous vehicles, the most recent works formulate the safe decision-making problem as constrained optimization [7], [24], [25]. However, there have been several different roadmaps for solving this problem. For the IL-based approach, [15], [26] propose implicit safe constraints in IL via uncertainty quantification and Bayesian abstraction from expert data. These approaches depend on their *safety* on the small discrepancy between the learned trajectory and the 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 information on the reward, cost, and cost threshold along the trajectories [23], [27]. To fully extract temporal information from offline trajectories, recent works turn offline RL into a sequential modeling problem using the power of transformers [21], [28]–[30]. Most of these works ignore the inherent structures of MDP in either the spatial or temporal domain, which limits the generalizability of the policy.

State Abstraction for Decision Making. To improve the performance of decision-making agents with some extra structural information, some recent work has focused on deriving state abstraction for generalizable decision-making using representation learning tricks. In the IL realm, [16] proposes Invariant Causal Imitation Learning (ICIL) to deal with the distribution shift with domain-invariant causal features. Based on uncertainty quantification, [15], [26], [31] propose an ensemble representation that leverages multi-modal sensor inputs to improve 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 [32], hierarchical skill decomposition [17] and deep bisimulation metric learning [33], [34]. In autonomous driving applications, state and action space are usually factorizable, [35], [36] 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. Prior to this work, however, the intersection of state abstraction with offline Safe RL is unexplored, which is crucial to advance the learning-based methods in the autonomous driving domain.

III. Problem Formulation

As stated in Section I, this work essentially aims to tackle a generalizable safe RL problem under distribution shifts in an offline setting. To better model such distribution shifts, we follow the definition of contextual MDP in [37] 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 $\big(\Omega, \mathcal{M}(\omega)\big)$, where $\mathcal{M}$ is a function that maps any contexts $\Omega \in \Omega$ to a constrained MDP $\mathcal{M}(\omega) = \big(\mathcal{S}, \mathcal{A}, P_\omega, r, c, s_0, \gamma\big)$.*

Here $P_\omega : \mathcal{S} \times \mathcal{A} \times \Omega \to \mathcal{S}$ 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 additional intuitive performance preference for driving agents. In addition, 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 cumulative reward while satisfying the safety constraints on cumulative expected cost under a certain target context ω. In formal terms, our problem can be defined as the following constrained optimization problem $\max_\pi \; J_r(\pi, \omega) \;\; s.t. \; J_c(\pi, \omega) \leq \kappa_c$, 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: $J_c(\pi, \omega) \leq c, \forall \pi \in \Pi, \omega \in \Omega$, i.e. imposing constraint satisfaction under varying behavior policies π_β and environment contexts ω. Meanwhile, we assume

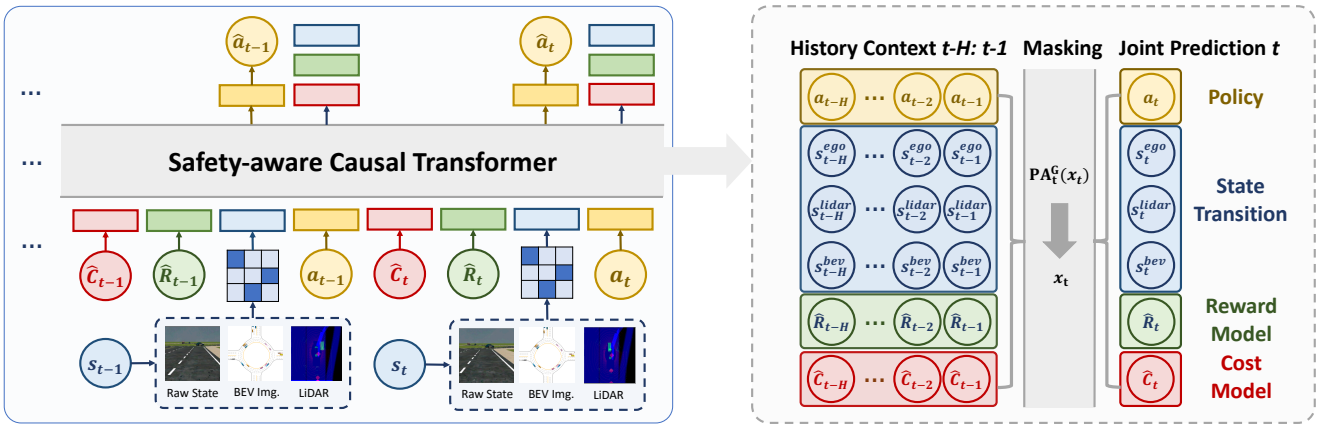

Fig. 2. Overview of Safety-aware structural Scenario Representation Framework. The diagram on the left shows a safety-aware decision transformer that conducts sequential decision-making based on the temporal contexts. The right diagram shows the general form of the graphical model in the CEWM and Policy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

that the preference for the reward function r and the cost function c remain unchanged in 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 three events: (i) collision with others, (ii) out-of-road, and (iii) overspeeding. Collision and out-of-road costs are binary indicators that are non-zero only when the corresponding event happens, and overspeeding costs are a continuous cost that occurs once the vehicle exceeds 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 formulation in this paper is to learn a safe policy with good generalizability at the deployment stage, under distribution shifts that occur: (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}. This difference also indicates the difference in MDP $\mathcal{M}(\omega_1) \neq \mathcal{M}(\omega_2)$. More specifically, we define the distribution shift in *transition dynamics* T_ω (e.g., the density of traffic) as follows: $p(\cdot|s, a; \omega_{train}) \neq p(\cdot|s, a; \omega_{test})$.

IV. 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).

A. Causal Ensemble World Model Learning

In autonomous driving problems, the entire state space can be decomposed into several disjoint subspaces [15], 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.*

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 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 [38],*

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

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

For general offline RL problems, SCM aims to jointly parameterize the world model and the policy model between different nodes in the state, action, reward, and safety cost. To parameterize the functions f in this SCM, we use a Safety-aware Causal Transformer, as shown in Figure 2. For example, the child node s_j is determined by its parent tokens $\mathbf{PA}_t^{\mathcal{G}}(s_j)$ 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 spatio-temporal domain, which

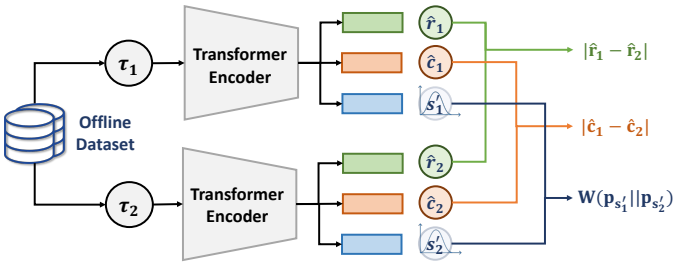

Fig. 3. Safety-aware bisimulation metrics with the distribution distance in transition dynamics, rewards, and safety cost.

is essentially captured by the attention weights, as we will discuss later in Figure 6 of the experiment parts. In addition to 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 unnecessary dependencies between the descendent nodes to indirect ancestors or non-parent nodes. This 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:

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

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.

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

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

B. 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. Inspired by the DBC algorithm for off-policy RL in [33], 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 [33], [39] 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 determined not only by the stepwise 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 [40] of safety-aware bisimulation relationship, denoted as Safety-aware Bisimulation Metrics as is shown in Figure 3.

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

Finally, at inference time, we take advantage of the prediction of values in online inference time, as shown in Algorithm 2. By taking the minimum cost-to-go preference and cost prediction, and the maximum reward-to-go preference and reward prediction at each step, we aim to improve the safety and efficiency of FUSION conditioned on the input human preference during the online deployment stage.

V. Experiments

In this section, we first go through the environments and evaluation protocols that we use based on the MetaDrive simulator [41]. 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.

Algorithm 1: Safety-aware CBL

Data: Offline (mixed) trajectories, cost limit C
Result: State encoder ϕ of policy π
for $k = 0, \cdots, N-1$ **do**
  Sample minibatch: $\mathcal{B} \leftarrow \text{Sample}(\mathcal{D}_{\pi_\beta})$;
  Construct transition pairs: $(s_1, a_1, s_1') \leftarrow \mathcal{B}$;
  Permute samples: $(s_2, a_2, s_2') \leftarrow \text{permute}(\mathcal{B})$;
  Compute bisimulation distance: With (5) ;
  Update encoder: $\phi_{k+1} \leftarrow \phi_k - \nabla_\phi \mathcal{L}_{\text{bisim}}$ with (6);

Algorithm 2: 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. 1;
/* 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: $s_t, r_t, c_t = \text{env.step}(a_{t-1})$;
  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: $a_t \leftarrow \pi_{\theta,\phi}(\mathbf{o})$;

A. Experiment Setup

a) Evalation Environment: We evaluate our algorithm on MetaDrive [41], 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, which is an 84×84×5 multi-channel image that captures the road contexts and the recent trajectories of the ego and surrounding vehicles.

We collect the offline dataset by IDM polices [42] 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 offline dataset consists of 2,000 trajectories with over 400,000 timesteps under 6 different road configurations.

We evaluate the following quantitative metrics to demonstrate the effectiveness of FUSION:

- 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 5). As introduced in (2), the safety violation costs are due to 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 end if any one of the risk 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.

b) 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 entering uncertain and unsafe regions. This kind of baseline includes Safe Behavior Cloning (**Safe BC** [2]) that only uses safe trajectories to train the agent, Invariant Causal Imitation Learning (**ICIL** [16]) that derives invariant state abstraction to learn generalizable policies by the model ensemble, like **GSA** [17] and **BNN** [15], which both use hierarchical state abstraction in generalizable decision making.

On the other hand, 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 [19] and BCQ [18], respectively. Constrained Penalized Q-Learning (**CPQ** [20]) aims to learn safe policy by penalizing the cost from the offline dataset. All Offline Safe RL baselines set an episodic cost constraint threshold $\kappa_c = 1$. Based on the

TABLE I
EVALUATION PERFORMANCE IN BOTH POLICY MISMATCH AND DYNAMICS MISMATCH SETTINGS. EACH OF THE BASELINE RESULTS IS EVALUATED UNDER 5 RANDOM SEEDS. **BOLD** MEANS THE BEST.

Mismatch	Metrics	Safe BC	ICIL	BNN	GSA	BEAR-Lag	BCQ-Lag	CPQ	FUSION
Policy	Reward (↑)	106.28±7.49	122.66±4.85	118.61±3.09	89.94±6.84	109.62±3.91	111.36±5.26	9.01±0.87	**139.95±4.24**
	Cost (↓)	12.79±0.70	11.07±1.11	4.46±0.41	13.18±1.26	4.46±0.29	0.89±0.08	1.05±0.18	**0.52±0.06**
	Succ. Rate (↑)	0.47±0.10	0.76±0.05	0.74±0.11	0.34±0.08	0.72±0.06	0.79±0.08	0.00±0.00	**0.90±0.03**
Dynamics	Reward (↑)	81.07±3.80	88.21±5.30	113.35±5.68	102.40±6.44	113.38±5.25	**122.72±7.64**	7.47±0.59	117.40±4.30
	Cost (↓)	9.44±0.55	7.29±0.72	19.16±0.55	11.88±0.98	7.86±0.66	6.22±0.76	**0.71±0.09**	**0.90±0.14**
	Succ. Rate (↑)	0.12±0.06	0.32±0.05	0.59±0.06	0.03±0.02	0.32±0.05	0.39±0.08	0.00±0.00	**0.82±0.04**

design of the safety cost introduced in Section V-A, when the episodic cost is lower than 1, it means that no critical violence, including collision and out-of-lane, occurred in this episode.

B. 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 baselines?
- **(RQ3)** Can FUSION consistently outperform other baselines and expert policies in 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 with the baselines aforementioned in both policy mismatch and dynamics mismatch settings. The results in Table I demonstrate the advantages of FUSION compared to baselines in both the safety cost and driving reward performance. (i) In the policy mismatch setting where the agent must 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 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 training environments, the performance gap between our methods and other baselines even enlarged, for example, we can see that 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 more than 30%.

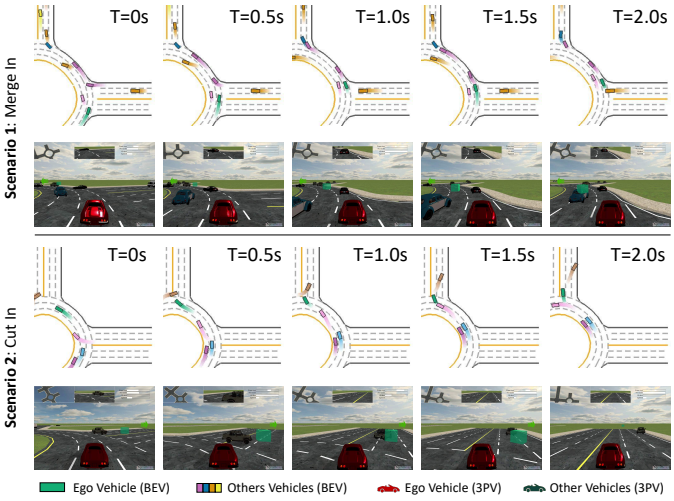

Fig. 4. The figure shows both birds-eye-view (BEV) and third-person-view (3PV) images of two case studies in roundabouts. The first case is a merge-in behavior from normal traffic, and the ego vehicles controlled by FUSION will decelerate reasonably to keep the distance from the front vehicle. The second case is an adversarial driver trying to cut in from the wrong side of the roundabout exit, FUSION manages to yield to it safely.

For **RQ3**, we take a deeper look at the exact driving performance by case studies in Figure 4. We also provide comparisons of safety metrics in different road contexts as a radar plot in Figure 5. The larger the pentagon is, the better overall safety performance it has. We calculate the safety metrics by the episode-wise frequency of five different safety behavior categories, including (i) **AR**: arrival rate in all episodes; (ii) **NS**: not speeding in the episode, which counts the time step ratio in which 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 agent can drive reasonably under complex contexts, especially in the hardest Roundabout environment.

For **RQ4**, we provide additional ablation studies in Table II. We compare FUSION with three of its variants: (i) **FUSION-Short**, which uses a shorter context in the safety-aware

TABLE II
ABLATION STUDIES ON FUSION'S VARIANTS TO SHOW THE CONTRIBUTION OF EACH MODULE. **BOLD** MEANS THE BEST.

Mismatch	Metrics	FUSION Short	FUSION w/o CEWM	FUSION w/o CBL	FUSION	Expert Policy
Policy	Reward (↑)	100.86±3.40	94.24±6.50	104.54±4.04	**139.95±4.24**	131.32±29.60
	Cost (↓)	0.77±0.09	**0.67±0.11**	3.46±0.21	**0.52±0.06**	16.02±9.46
	Succ. Rate (↑)	0.34±0.07	0.41±0.06	0.58±0.09	**0.90±0.03**	0.81±0.15
Dynamics	Reward (↑)	98.63±2.36	81.70±3.82	90.34±4.28	**117.40±4.30**	129.71±28.84
	Cost (↓)	0.79±0.06	**0.60±0.04**	5.60±0.32	**0.90±0.14**	17.58± 9.71
	Succ. Rate (↑)	0.34±0.04	0.24±0.04	0.08±0.01	**0.82±0.04**	0.72±0.20

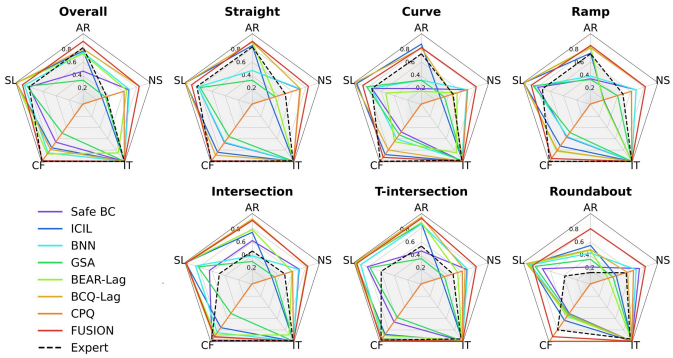

Fig. 5. The figure shows the comparison of FUSION on different road configurations with baselines. The larger lidar plot on each coordinate stands for the safer performance in each safety metric. (**AR**: Arrival, **NS**: Not speeding, **IT**: In-time, **CF**: Collision-free, **SL**: Stay in-lane.)

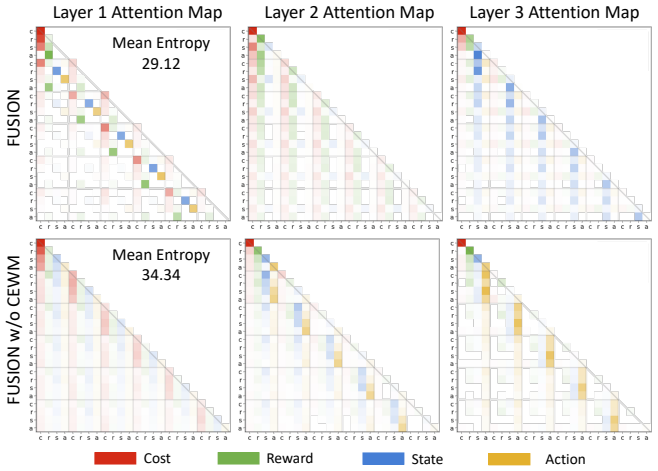

Fig. 6. Visualization of average attention matrix over 30 trajectories. We compare different layers of the attention map of two models: FUSION and FUSION w/o CEWM. We compare the mean entropy through all three attention layers in one head of our Transformer encoder. The result shows that FUSION has a lower entropy than the ablation variant, which means its attention map is more sparse compared to the baselines without causal representation.

transformer to model the whole sequence; (ii) **FUSION w/o CEWM**, which does not consider the learning of the causal ensemble world model, and only uses the behavior cloning term as supervised signals; (iii) **FUSION w/o CBL**, which neglects safety-aware bisimulation learning. The result confirms that the FUSION benefits from all its design, including the spatio-temporal information from CEWM and additional safety awareness in the transformer model via CBL.

Furthermore, we visualize the normalized attention map of FUSION's safety-aware causal transformer in Figure 6. 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 quite 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** and **reward-to-go** to the future state and action tokens, trying to balance both for the decision-making process in a long horizon. (iii) FUSION captures world dynamics and policy by attending previous **states** to the future value prediction and action. 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. On the contrary, as shown in the second row of Figure 6, FUSION without CEWM has a higher average entropy among all the layers, indicating that it does not capture the above sparsity and interpretability. The reason is that the variant without CEWM ignores sequential awareness which has more informative training signals during the offline training stage.

VI. 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 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 several strong baselines of LfID and causal abstraction in diverse autonomous driving scenarios. We also conduct extensive case analysis to analyze the benefits of different modules that we design in FUSION and show a comprehensive and interpretable evaluation of FUSION. One potential limitation is that all the experiments are conducted in the portable MetaDrive simulator instead of more high-fidelity simulators like CARLA. Meanwhile, in the FUSION pipeline, CBL relies on a good estimation of transition dynamics, which in general requires good coverage and diversity in offline samples. In the future, it would be interesting to extend FUSION's framework to other autonomous vehicle platforms and tackle more challenging scenarios in multi-agent RL settings.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support from the National Science Foundation under grants CNS-2047454 and gift funding from Ford Motor Company.

REFERENCES

[1] J. Chen, B. Yuan, and M. Tomizuka, "Model-free deep reinforcement learning for urban autonomous driving," in *2019 IEEE intelligent transportation systems conference (ITSC)*. IEEE, 2019, pp. 2765–2771.

[2] Y. Pan, C.-A. Cheng, K. Saigol, K. Lee, X. Yan, E. A. Theodorou, and B. Boots, "Imitation learning for agile autonomous driving," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 286–302, 2020.

[3] J. Chen, S. E. Li, and M. Tomizuka, "Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5068–5078, 2021.

[4] R. Rafailov, T. Yu, A. Rajeswaran, and C. Finn, "Offline reinforcement learning from images with latent space models," in *Learning for Dynamics and Control*. PMLR, 2021, pp. 1154–1168.

[5] W. Ding, H. Lin, B. Li, and D. Zhao, "Causalaf: causal autoregressive flow for safety-critical driving scenario generation," in *Conference on Robot Learning*. PMLR, 2023, pp. 812–823.

[6] K. Renz, K. Chitta, O.-B. Mercea, A. Koepke, Z. Akata, and A. Geiger, "Plant: Explainable planning transformers via object-level representations," *arXiv preprint arXiv:2210.14222*, 2022.

[7] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in *Conference on Robot Learning*. PMLR, 2023, pp. 726–737.

[8] W. Ding, C. Xu, M. Arief, H. Lin, B. Li, and D. Zhao, "A survey on safety-critical driving scenario generation—a methodological perspective," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[9] M. Xu, Z. Liu, P. Huang, W. Ding, Z. Cen, B. Li, and D. Zhao, "Trustworthy reinforcement learning against intrinsic vulnerabilities: Robustness, safety, and generalizability," *arXiv preprint arXiv:2209.08025*, 2022.

[10] F. Fuchs, Y. Song, E. Kaufmann, D. Scaramuzza, and P. Dürr, "Super-human performance in gran turismo sport using deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4257–4264, 2021.

[11] P. R. Wurman, S. Barrett, K. Kawamoto, J. MacGlashan, K. Subramanian, T. J. Walsh, R. Capobianco, A. Devlic, F. Eckert, F. Fuchs *et al.*, "Outracing champion gran turismo drivers with deep reinforcement learning," *Nature*, vol. 602, no. 7896, pp. 223–228, 2022.

[12] D. Shah, K. Stachowicz, A. Bhorkar, I. Kostrikov, and S. 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] A. Mohan, A. Zhang, and M. Lindauer, "Structure in reinforcement learning: A survey and open problems," *arXiv preprint arXiv:2306.16021*, 2023.

[14] X. Jia, P. Wu, L. Chen, J. Xie, C. He, J. Yan, and H. 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*, 2023, pp. 21 983–21 994.

[15] K. Lee, Z. Wang, B. Vlahov, H. Brar, and E. A. Theodorou, "Ensemble bayesian decision making with redundant deep perceptual control policies," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 2019, pp. 831–837.

[16] I. Bica, D. Jarrett, and M. van der Schaar, "Invariant causal imitation learning for generalizable policies," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3952–3964, 2021.

[17] R. Akrour, F. Veiga, J. Peters, and G. Neumann, "Regularizing reinforcement learning with state abstraction," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 534–539.

[18] A. Kumar, J. Fu, M. Soh, G. Tucker, and S. Levine, "Stabilizing off-policy q-learning via bootstrapping error reduction," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *International conference on machine learning*. PMLR, 2019, pp. 2052–2062.

[20] H. Xu, X. Zhan, and X. Zhu, "Constraints penalized q-learning for safe offline reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 8753–8760.

[21] Z. Liu, Z. Guo, Y. Yao, Z. Cen, W. Yu, T. Zhang, and D. Zhao, "Constrained decision transformer for offline safe reinforcement learning," *arXiv preprint arXiv:2302.07351*, 2023.

[22] X. Fang, Q. Zhang, Y. Gao, and D. Zhao, "Offline reinforcement learning for autonomous driving with real world driving data," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 3417–3422.

[23] Z. Liu, Z. Guo, H. Lin, Y. Yao, J. Zhu, Z. Cen, H. Hu, W. Yu, T. Zhang, J. Tan *et al.*, "Datasets and benchmarks for offline safe reinforcement learning," *arXiv preprint arXiv:2306.09303*, 2023.

[24] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International conference on machine learning*. PMLR, 2017, pp. 22–31.

[25] Z. Liu, Z. Cen, V. Isenbaev, W. Liu, S. Wu, B. Li, and D. Zhao, "Constrained variational policy optimization for safe reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 13 644–13 668.

[26] K. Menda, K. Driggs-Campbell, and M. J. Kochenderfer, "Ensembledagger: A bayesian approach to safe imitation learning," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5041–5048.

[27] H. Le, C. Voloshin, and Y. Yue, "Batch policy learning under constraints," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3703–3712.

[28] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," *Advances in neural information processing systems*, vol. 34, pp. 15 084–15 097, 2021.

[29] H. Liu, Z. Huang, X. Mo, and C. Lv, "Augmenting reinforcement learning with transformer-based scene representation learning for decision-making of autonomous driving," *arXiv preprint arXiv:2208.12263*, 2022.

[30] M. Janner, Q. Li, and S. Levine, "Offline reinforcement learning as one big sequence modeling problem," *Advances in neural information processing systems*, vol. 34, pp. 1273–1286, 2021.

[31] A. Loquercio, M. Segu, and D. Scaramuzza, "A general framework for uncertainty estimation in deep learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3153–3160, 2020.

[32] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1134–1141.

[33] A. Zhang, R. McAllister, R. Calandra, Y. Gal, and S. Levine, "Learning invariant representations for reinforcement learning without reconstruction," *arXiv preprint arXiv:2006.10742*, 2020.

[34] R. Dadashi, S. Rezaeifar, N. Vieillard, L. Hussenot, O. Pietquin, and M. Geist, "Offline reinforcement learning with pseudometric learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2307–2318.

[35] W. Ding, H. Lin, B. Li, and D. Zhao, "Generalizing goal-conditioned reinforcement learning with variational causal reasoning," in *Advances in Neural Information Processing Systems*, 2022.

[36] W. Ding, L. Shi, Y. Chi, and D. Zhao, "Seeing is not believing: Robust reinforcement learning against spurious correlation," *arXiv preprint arXiv:2307.07907*, 2023.

[37] B. Chen, Z. Liu, J. Zhu, M. Xu, W. Ding, L. Li, and D. Zhao, "Context-aware safe reinforcement learning for non-stationary environments," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 10 689–10 695.

[38] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[39] L. Li, T. J. Walsh, and M. L. Littman, "Towards a unified theory of state abstraction for mdps." in *AI&M*, 2006.

[40] N. Ferns, P. Panangaden, and D. Precup, "Metrics for finite markov decision processes." in *UAI*, vol. 4, 2004, pp. 162–169.

[41] Q. Li, Z. Peng, L. Feng, Q. Zhang, Z. Xue, and B. Zhou, "Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[42] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model mobil for car-following models," vol. 1999, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2007, pp. 86–94.