Safety-Aware Human-in-the-Loop Reinforcement Learning With Shared Control for Autonomous Driving

Wenhui Huang, *Graduate Student Member, IEEE*, Haochen Liu, *Graduate Student Member, IEEE*, Zhiyu Huang, *Graduate Student Member, IEEE*, and Chen Lv, *Senior Member, IEEE*

***Abstract*— The learning from intervention (LfI) approach has been proven effective in improving the performance of RL algorithms; nevertheless, existing methodologies in this domain tend to operate under the assumption that human guidance is invariably devoid of risk, thereby possibly leading to oscillations or even divergence in RL training as a result of improper demonstrations. In this paper, we propose a safety-aware human-in-the-loop reinforcement learning (SafeHIL-RL) approach to bridge the abovementioned gap. We first present a safety assessment module based on the artificial potential field (APF) model that incorporates dynamic information of the environment under the Frenet coordinate system, which we call the Frenet-based dynamic potential field (FDPF), for evaluating the real-time safety throughout the intervention process. Subsequently, we propose a curriculum guidance mechanism inspired by the pedagogical principle of whole-to-part patterns in human education. The curriculum guidance facilitates the RL agent's early acquisition of comprehensive global information through continual guidance while also allowing for fine-tuning local behavior through intermittent human guidance through a human-AI shared control strategy. Consequently, our approach enables a safe, robust, and efficient reinforcement learning process independent of the quality of guidance human participants provide. The proposed method is validated in two highway autonomous driving scenarios under highly dynamic traffic flows (https://github.com/OscarHuangWind/Safe-Human-in-the-Loop-RL). The experiments' results confirm the superiority and generalization capability of our approach when compared to other state-of-the-art (SOTA) baselines, as well as the effectiveness of the curriculum guidance.**

***Index Terms*— Autonomous driving, human-in-the-loop reinforcement learning, safe reinforcement learning, curriculum guidance, shared control.**

I. Introduction

AUTONOMOUS vehicle (AV) is a complicated and holistic intelligent system that consists of various ingredients such as perception, decision-making, and planning [1], [2], [3]. Among these components, intelligent decision-making serves as a "brain" of the AVs to decide where to go and how to go, playing a significant role in achieving fully autonomous driving techniques [4]. Over the past years, scholars have dedicated their efforts to developing such decision-making strategies through various domain knowledge, and in general, these approaches are categorized into two families: rule-based and learning-based methods [5]. In the early stage, the contributions were intensively located on the former one, which employs finite-state machine (FSM) [6] or online optimization approaches [7], [8] due to their attractive level of interpretability. However, the rule-based method suffers from a significant computational efficiency limitation since the AVs are required to drive under a complex and highly dynamic environment, and it is costly to pre-design the exhausted definitions, as well as the constraints for the different driving scenarios [9]. Alternatively, the learning-based approach has gained growing attention recently due to its superior capability of handling high-dimensional information. Reinforcement learning (RL) [10] is one of the most widely utilized learning-based approaches for decision-making and control problems in the autonomous driving field [11], [12], [13]. However, the intrinsic learning mechanism, namely trial-with-error, leads to a significant safety concern that hinders the practical application of the RL-based methods [14].

To address the safety issue mentioned above, a subfield within the RL community entitled Safe RL is proposed [15], [16]. Among various categorized groups within the Safe RL framework, a promising method is leveraging human intelligence in the learning loop of the RL algorithms, namely learning from intervention (LfI) [17]. The objective of the LfI approach is to prevent catastrophic actions from occurring and replace them with real-time human intervention, guiding the RL agent toward a safe and efficient exploration strategy. Recently, several related works have been proposed and successfully applied to the autonomous navigation [18] and autonomous driving [19] domain. These approaches do not explicitly perform safety assessment since they assume that human beings have the dominant knowledge of the task, and thus, the human demonstration is consistently safe or at least better than that of the RL agents. In reality, however, there is no theoretical guarantee that each human participant has qualified driving proficiency, and it is equally possible that even the experts could perform degraded or catastrophic operations due to various reasons, e.g., degraded attention and

Manuscript received 14 July 2023; revised 25 February 2024 and 19 June 2024; accepted 24 June 2024. Date of publication 11 July 2024; date of current version 1 November 2024. This work was supported in part by the Agency for Science, Technology and Research (A*STAR), Singapore, under the MTC Individual Research under Grant M22K2c0079, in part by the ANR-NRF Joint under Grant NRF2021-NRF-ANR003 HM Science, and in part by the Ministry of Education (MOE), Singapore, under the Tier 2 under Grant MOE-T2EP50222-0002. The Associate Editor for this article was S. Ahn. *(Corresponding author: Chen Lv.)*

The authors are with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wenhui001@e.ntu.edu.sg; haochen002@e.ntu.edu.sg; zhiyu001@e.ntu.edu.sg; lyuchen@ntu.edu.sg).

Digital Object Identifier 10.1109/TITS.2024.3420959

intrinsic uncertainty. Therefore, performing a real-time safety assessment for the human demonstration and not entirely relying on human intervention is crucial to the subsequent learning process, especially for the intermediate stage in which the RL agents have learned and possessed a certain level of intelligence.

In light of this, we propose a novel safety-aware human-in-the-loop RL (SafeHIL-RL) framework to bridge the abovementioned gap. Specifically, we construct a safety module by employing artificial potential field (APF) theory and dynamic information of the environment under the Frenet coordination system, namely the Frenet-based dynamic potential field (FDPF), to perform real-time safety assessment for the human demonstration whenever the intervention occurs. Subsequently, we design a dynamic control authority allocation mechanism to realize the shared autonomy between human participants and RL agents based on the potential hazard level provided by the safety module. Furthermore, we present a curriculum guidance mechanism, which is inspired by the pedagogical principle of whole-to-part patterns in human education, within the proposed framework to facilitate the RL agent's early acquisition of comprehensive global information through continual guidance while also allowing for fine-tuning local behavior through intermittent human guidance through a human-AI shared autonomy technique. It is important to highlight that our work does not focus on providing absolute safety assurance for the entire training process since the failure experiences are valuable for RL learning as well. Instead, SafeHIL-RL aims to present a novel safety-aware LfI approach to prevent policy oscillations or even divergence phenomena caused by inappropriate human demonstrations during the intervention, thereby enhancing learning efficiency and robustness. The proposed method is thoroughly evaluated in the highway autonomous driving scenario with stochastic traffic flow. The experiment and ablation study results confirm the advance of the proposed approach compared to other state-of-the-art (SOTA) methods in data efficiency and driving performance. The main contributions of this paper can be summarized as follows:

1) A novel safety-aware human-in-the-loop RL method within the Safe RL framework is realized by performing the safety assessment over real-time human intervention to enhance the sampling efficiency and quality.
2) A safety module that integrates the APF model and dynamic information of the environment within the Frenet coordination system (FDPF) is presented to evaluate the real-time risk during the human intervention period. Moreover, a dynamic control authority allocation mechanism is designed to achieve the human-AI shared autonomy approach in the training loop of the RL algorithms.
3) We propose a novel intervention mechanism called curriculum guidance, which draws inspiration from the pedagogical principle of whole-to-part patterns observed in human education. This mechanism aims to encourage the early acquisition of comprehensive global information by RL agents, followed by the fine-tuning of local behavior in subsequent stages, ultimately guiding the RL learning process towards smoothness and robustness.
4) We instantiated a concrete human-in-the-loop RL algorithm for the proposed approach and thoroughly evaluated it in the highway autonomous driving scenario under stochastic traffic flow, providing an efficient and reliable end-to-end autonomous driving solution.

II. Related Works

In this section, we present a review of the RL-based autonomous driving policy and LfI approach.

A. Reinforcement Learning for Autonomous Driving

Reinforcement learning (RL) has shown great potential in handling decision-making and control problems in complicated environments due to its superior representation capability. In general, RL algorithms can be categorized into three main groups, the value-based approaches such as deep Q-network (DQN) [20] and its variants [21], [22], [23], [24], policy-based methods like trust region policy optimization (TRPO) [25] and proximal policy optimization (PPO) [26], and actor-critic-based framework such as deep deterministic policy gradient (DDPG) [27], twin delayed deep deterministic policy gradient (TD3) [28], and soft actor-critic (SAC) [29]. Recently, these algorithms have been intensively employed in the autonomous driving field. For instance, [30] successfully realizes the end-to-end decision-making for autonomous vehicles in the simulation through the DQN algorithm with the gray-scaled image only. In [31] and [32], the DDPG algorithm is utilized to learn autonomous driving strategy by considering traffic rules. Similarly, [33] advances the same algorithm to real-world driving under driver supervision. Subsequently, [34] employs the SAC algorithm to realize the high-speed autonomous drifting function, while [35] addresses the end-to-end control tasks through the SAC algorithm in an urban environment. In [36], a sim-to-real application of autonomous navigation is presented by means of the same algorithm, with the augmented scene encoder called Goal-enabled Transformer (GoT). Moreover, the TD3 algorithm is employed to handle the unprotected left-turn and congestion scenarios for autonomous vehicles in [37].

B. Learning From Intervention

Despite the various successful applications, the RL algorithms notoriously suffer from poor data efficiency stemming from their intrinsic learning mechanism: trial-with-error. A promising solution for addressing the abovementioned issue is leveraging human intelligence in the training loop of the RL agents. In light of this, [38] proposes a learning from intervention (LfI) approach titled human intervention RL (HIRL) to lead the intelligent agent with a safe exploration strategy by means of avoiding catastrophic actions under human supervision. In the subsequent years, scholars have dedicated themselves to enhancing the LfI approach from various domains. For example, [39] proposes a novel approach entitled expert intervention learning (EIL) to introduce human

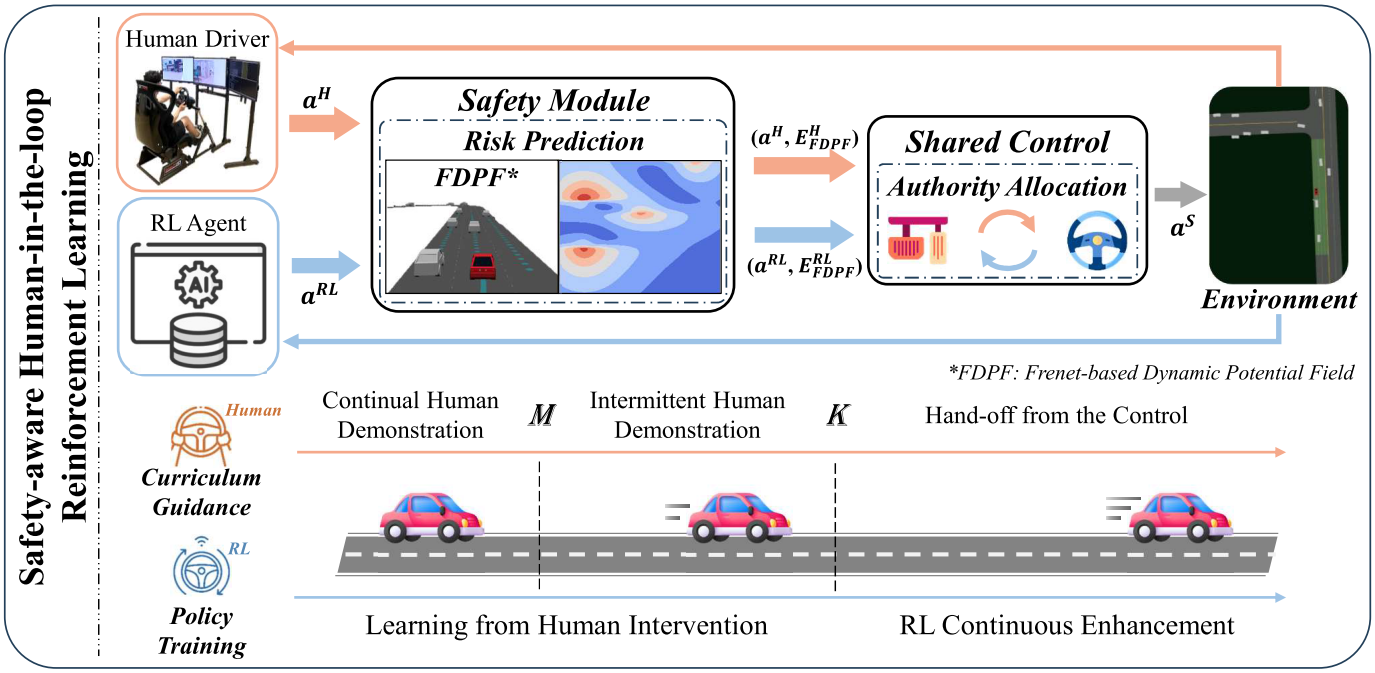


Fig. 1. Overall framework of the proposed approach.

guidance into the value function as a constraint to achieve a safe autonomous navigation strategy. Instead of augmenting the loss function of the value network, [18] presents an intervention-aided RL (IARL) method that adds an imitation term in the actor network of the PPO algorithm and evaluates the obstacle-avoiding performance in a real UAV platform. Similarly, in [40], a method called human-guidance-based deep RL (Hug-DRL) is developed to address the end-to-end autonomous driving control tasks. Alternatively, [41] learns the probability of disengagement from the human intervention rather than the deterministic actions, realizing a safe navigation strategy in a sidewalk environment. Acknowledging that consistent monitoring would fatigue the human participants and hence degrade the quality of the demonstration, another modality of human intelligence is explored in [42]. The authors propose a novel expert-guided policy optimization (EGPO) method that employs an expert policy rather than a human participant to actively intervene throughout the entire training loop of the RL agents, enhancing the training and test-time safety.

III. Methodology

This section elaborates on the primary components of the proposed approach shown in Fig. 1, starting from the conventional RL to HIL-RL, safety assessment, control authority allocation, and curriculum guidance settings.

A. Reinforcement Learning

Considering autonomous driving as an end-to-end control task, we employ the RL algorithm within actor-critic (AC) framework in this paper to address the continuous action space problem. The RL algorithm formulates such a control task as a standard Markov decision process (MDP). More specifically, at an arbitrary time step t, the RL agent executes a concrete action a_t based on the perceived state s_t and triggers the state transition $s_t \rightarrow s_{t+1}$ with providing an immediate reward r_t. Inheriting the dynamic programming (DP) property, the RL algorithm employs action value function $Q(s_t, a_t)$ and state function $V(s_t)$ to represent the cumulative discounted future payoffs $\sum_t^T \gamma^t \cdot r_t$ within the finite time horizon T, where $\gamma \in (0, 1]$ is discounting factor. The action value function and value function can be expressed through Bellman equation, computed as follows:

$$\begin{aligned} V^\pi(s_t) &= \mathop{\mathbb{E}}_{a_t \sim \pi(\cdot|s_t)} [Q^\pi(s_t, a_t)] \\ Q^\pi(s_t, a_t) &= r_t + \gamma \cdot \mathop{\mathbb{E}}_{s_{t+1} \sim \rho} [V^\pi(s_{t+1})] \\ &= r_t + \gamma \cdot \mathop{\mathbb{E}}_{(s_{t+1}, a_{t+1}) \sim \rho_\pi} [Q^\pi(s_{t+1}, a_{t+1})] \end{aligned} \tag{1}$$

where ρ and π represents the transition probability and the behavior policy. Therefore, the optimal policy can be obtained through maximizing the overall future payoffs:

$$\pi^* = \mathop{argmax}_{\pi} \mathop{\mathbb{E}}_{s_t \sim \mathcal{D}, a_t \sim \pi} [Q^\pi(s_t, a_t)] \tag{2}$$

where $\mathcal{D}$ is the replay buffer in which stores the transitions. In addition, to prevent the policy from trapping in the local optima, it usually introduces the entropy of the policy into the objective, denoted as:

$$\begin{aligned} \mathcal{L}(\phi) &= \mathop{\mathbb{E}}_{s_t \sim \mathcal{D}, a_t \sim \pi_\phi} [Q^\pi(s_t, a_t) + \alpha \cdot \mathcal{H}(\pi_\phi(a_t|s_t))] \\ &= \mathop{\mathbb{E}}_{s_t \sim \mathcal{D}} \Big[\mathop{\mathbb{E}}_{a_t \sim \pi_\phi} [Q^\pi(s_t, a_t) - \alpha \cdot log(\pi_\phi(a_t|s_t))] \Big] \end{aligned} \tag{3}$$

where ϕ, $\mathcal{H}$ and α indicates the network parameters, entropy of policy and temperature parameter. Equation 3 is usually employed as the objective function of the actor network. Regarding the critic network, the parameters are updated

by minimizing the mean squared Bellman error (MSBE), denoted as:

$$\mathcal{L}(\theta_i) = \mathop{\mathbb{E}}_{\substack{(s_t,a_t,r_t,s_{t+1}) \\ \sim \mathcal{D}}} [Q^{\pi}_{\theta_i}(s_t, a_t) - (r_t + \min_{j=1,2} Q^{\pi}_{\theta_j}(s_{t+1}, a_{t+1}))] \tag{4}$$

where θ and i denotes the network parameters and the index of double critic networks.

B. Human-in-the-Loop Reinforcement Learning

Introducing the human guidance into the training loop of the RL algorithms is a promising method to mitigate the poor efficiency issue that originates from the intrinsic learning mechanism of the RL. In this work, we employ the human participants to consistently monitor the RL agent and provide human guidance through Logitech G29 set whenever it is necessary. Hence, the shared control command a_t^s is a blend of the actions between human and RL agent when the intervention is triggered, denoted as:

$$a_t^s = \left(\mathbf{1} - \Lambda_t\right) \cdot a_t^{RL} + \Lambda_t \cdot a_t^{\mathbf{H}} \tag{5}$$

where Λ indicates the control authority of the human and the computation details will be discussed in Section III-C. It should be mentioned that a_t^s is equal to a_t^{RL} (we will omit superscript RL and use a_t to represent the action of RL agent in the rest of the paper for the convenience) when $\Lambda_t = 0$. Once such a shared control occurs, the human action, intervention discriminator $\mathcal{I}$, and state transitions are stored in the replay buffer, denoted as follows:

$$\mathcal{D} \leftarrow (s_t, a_t^s, r_t, s_{t+1}, a_t^{\mathbf{H}}, \Lambda_t, \mathcal{I}) \tag{6}$$

Besides introducing human intervention, the HIL-RL aims to effectively learn from the experiences consisting of the RL exploration and human demonstration data. Thus, we reformulate the Eq. 4 as follows:

$$\mathcal{L}(\theta_i) = \mathop{\mathbb{E}}_{\substack{(s_t,a_t^s,r_t,s_{t+1}) \\ \sim \mathcal{D}}} [Q^{\pi}_{\theta_i}(s_t, a_t^s) - (r_t + \min_{j=1,2} Q^{\pi}_{\theta_j}(s_{t+1}, a_{t+1}))] \tag{7}$$

As for the actor network, we add a human guidance related term to the original objective function. Therefore, the actor network aims to mimic the human demonstration in addition to maximizing the action value function, shown as:

$$\mathcal{L}(\phi) = \mathop{\mathbb{E}}_{\substack{(s_t,a_t^{\mathbf{H}},\Lambda_t,\mathcal{I})\sim\mathcal{D} \\ a_t\sim\pi_\phi}} [Q^{\pi}(s_t, a_t) + \alpha \cdot \mathcal{H}(\pi_\phi(a_t|s_t)) - \mathcal{I} \cdot \omega_t \cdot ||a_t - a_t^{\mathbf{H}}||^2] \tag{8}$$

where ω_t is the weight for the human guidance term which represents the confidence that RL agent holds about the human, and we employ the control authority Λ_t for this value in this paper. Additionally, the mean square error (MSE) in Eq. 8 does not apply any restrictions for the human guidance term which means it can be alternated by any other loss function such as Kullback-Leibler (KL) divergence or negative log-likelihood (NLL).

C. Safety Assessment and Control Authority Allocation

Though human participants can react to a complex environment rapidly, consistent monitoring would fatigue human beings and subsequently affect the guidance quality. In light of this, we propose a novel artificial potential field (APF)-based safety module and design a control authority allocation mechanism to adaptively transfer the control authority based on real-time safety assessment instead of letting the human participant completely dominate the vehicle control whenever the intervention occurs.

Recently, the APF method has been widely utilized in the autonomous driving field for safety assessment [43], [44]. Inspired by [45] and [46], we construct the APF function with two ingredients: kinematic potential field (KPF) intensity $E_{KPF}(x, y)$ and road potential field (RPF) intensity $E_{RPF}(x, y)$, where (x, y) represents for the Cartesian coordinate. In addition, we advance the APF function to the Frenet frame [47] and consider the dynamic states of the environment into the field modeling to more precisely estimate the driving safety and term it Frenet-based dynamic potential field (FDPF), formulated as follows:

$$E_{FDPF}(s, l) = \sum_i \sum_j E_{KPF_i}(s, l) + E_{RPF_j}(s, l) \tag{9}$$

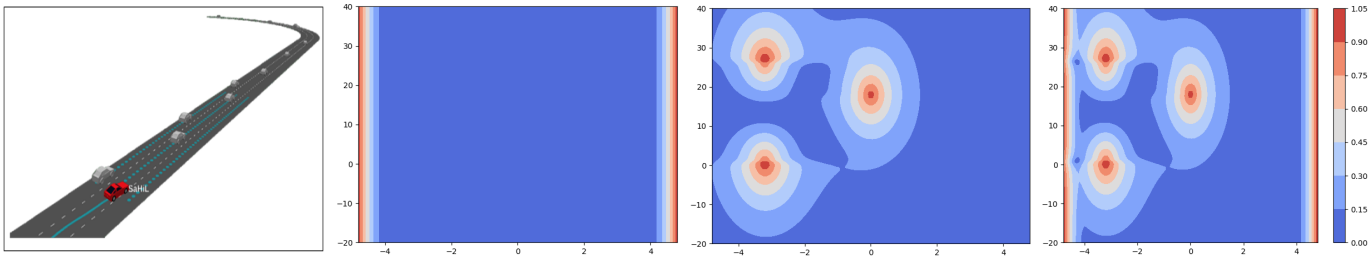
where i represents the index of the other road users such as surrounding vehicles, $j \in \{left, right\}$ denotes for the index of the road boundary, and (s, l) indicates the coordination under the Frenet frame. Recognizing that the closer to the other vehicles the more dangerous would the situation be, the intensity of KPF is defined as:

$$E_{KPF_i}(s, l) = \omega_{KPF} \cdot e^{-\beta \cdot \sqrt{\alpha_i^s (s_{ego} - s_i)^2 + \alpha_i^l (l_{ego} - l_i)^2}} \cdot \frac{d_i}{|d_i|}$$
$$\alpha_i^{k=s,l} = \frac{\lambda^k}{\gamma^k \cdot \lambda^k + max(v_{ego} - v_i, 0)} \tag{10}$$

where indices β, γ, λ, and d_i denote decay rate, scale factor, shape factor, and position vector starting from the surrounding vehicles to the ego vehicle under the Frenet frame. As for the RPF, we consider the safety regarding the road boundary, which means closer to the shoulder, the higher the potential risk is and computed as follows:

$$E_{RPF_j}(s, l) = \omega_{RPF} \cdot (max(0, Dist^l_{ego} - Dist^l_{marginal}))^2 \cdot \frac{d_j}{|d_j|} \tag{11}$$

where $Dist^l_{ego}$ represent for lateral distance between ego vehicle and road boundary, while $Dist^l_{marginal}$ denote the lateral marginal distance from the road boundary. Moreover, $d_j = l_{ego} - l_j$ is the orientation vector points from the ego vehicle to the left or right boundary. Figure 2 demonstrates an example of the constructed FDPF. More specifically, a general highway scenario is shown in Fig. 2(a), followed by the heatmap of the RPF and KPF in Fig. 2(b) and 2(c). Finally, Fig 2(d) illustrates the overall constructed FDPF that integrates RPF and KPF mentioned above. We can observe from these figures that the landscape of the potential field for each road



(a) Highway scene. (b) RPF. (c) KPF. (d) FDPF.

Fig. 2. An example of the FDPF modeling. a) Highway: A general scenario in the Highway environment. b) RPF: High risk near the shoulder. c) KPF: High risk near other road users. d) FDPF: Overall potential field.

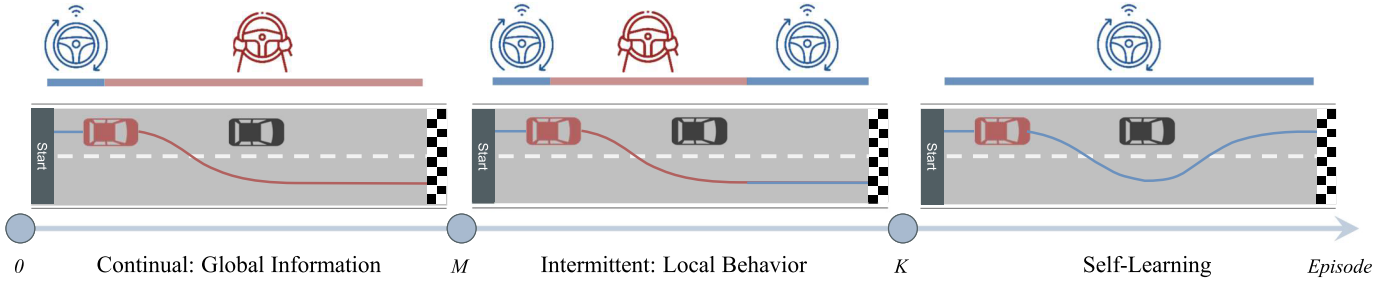


Fig. 3. Learning from curriculum guidance. The steering wheel and trajectory in blue color are controlled and performed by RL agents, while those in red color are operated by human participants. During the continual guidance phase, the participants are allowed to control the vehicle until one episode completes once they intervene, aiming at motivating the SafeHIL-RL to acquire global information at the early stage. On the contrary, the intermittent guidance allows the human participants only for the necessary duration within one episode, fine-tuning the local behavior based on the learned policy. In the self-learning phase, the human participants are asked to completely hand off from the vehicle control and motivate continuous enhancement of the driving policy through self-exploration.

user is different from the other since the real-time speed information is considered during the modeling.

To conduct risk prediction, we first individually forecast the trajectories of the ego vehicle for the human and RL agent over a time horizon based on the intelligent driver model (IDM). Recognizing that a large time horizon would result in the over-shoot of the trajectory prediction, which may negatively impact the accuracy of the safety assessment, we set the prediction horizon as 0.5 seconds in this work. It should be mentioned that such a trajectory prediction is simple but sufficient for implementing our approach. While a SOTA trajectory prediction approach may be preferred, it falls beyond the scope of our work. Once the safety assessment results are calculated through Eq. 9, namely the potential hazard levels of the human ($E^{\mathbf{H}}_{FPDF}$) and the RL agent (E^{RL}_{FDPF}), we can dynamically allocate the control authority. In order to keep patience for the low risk and be sensitive as the hazard grows, we first scale the field strength and apply the exponential function over it to compute the control authority, shown as:

$$\Lambda_t = \frac{e^{(\eta \cdot E^{RL}_{FDPF}(s,l))}}{e^{(\eta \cdot E^{\mathbf{H}}_{FDPF}(s,l))} + e^{(\eta \cdot E^{RL}_{FDPF}(s,l))}} \tag{12}$$

Finally, the output action is the blend of the Human and RL agent, as shown in Eq. 5.

D. Curriculum Guidance

In this study, we propose a novel intervention mechanism called curriculum guidance, shown in Fig. 3, to achieve a smooth and robust human-in-the-loop learning process. Specifically, during the training phase, human participants have the power to control the vehicle with a Logitech G29 driving set manually. This process begins with continuous guidance for the initial M episodes and transitions to intermittent guidance for the subsequent periods, continuing through to the Kth episode. In the phase of continuous guidance (spanning from 0 to M episodes), participants have the power to take over vehicle control for the duration of an entire episode once intervention occurs. This strategy is designed to encourage the SafeHIL-RL system to gather comprehensive information early on, facilitating its ability to complete the driving task successfully.

Conversely, during the intermittent guidance phase (from M to K episodes), human participants can only offer guidance for critical moments within an episode. This approach helps refine specific behaviors, e.g., decelerating and executing smooth lane changes, adjusting them based on the policy learned thus far. In the self-learning phase (after K episodes), the human participants are asked to completely hand off the vehicle control and motivate the RL agent to enhance the driving policy through self-exploration continuously. It is important to emphasize that insufficient intervention is unlikely to lead to favorable outcomes in the learning process, while excessive intervention can lead to overfitting to the human demonstration. With these considerations in mind, we restrict human intervention to a maximum of three instances within every ten episodes and confine the involvement of human participants to the first 400 episodes (K=400), which accounts for half of the entire training duration. Furthermore, the transition from continuous to intermittent guidance is empirically set at the 200th episode (M=200), where it is observed that the RL agent achieves a level of competency sufficient to complete the whole trajectory, though its performance may not be optimal.

It is important to distinguish curriculum guidance from curriculum learning. Curriculum learning is a training approach that introduces a sequence of progressively challenging tasks to facilitate policy learning and tackle the ultimate complex problems effectively [48]. The core idea of curriculum learning revolves around the arrangement of tasks in a manner that allows the policy to learn varying levels of complexity from the algorithm's perspective. In contrast, the proposed curriculum guidance does not revolve around task or environment complexity since the characteristics of the interactive road users are entirely stochastic (traffic flow, speed, and cooperation willingness) throughout the whole learning process. Instead, curriculum guidance centers on researching and enhancing the LfI approaches from the perspective of human experts within the safe RL framework.

By clarifying the distinction between curriculum guidance and curriculum learning, we can better comprehend the specific focus and objectives of each approach in enhancing the reinforcement learning process. While curriculum learning emphasizes the sequencing of tasks to aid policy learning, which is from the algorithm point of view, curriculum guidance delves into researching how humans can guide the learning process effectively, focusing on the learning patterns from the human perspective. We provide the formal definition of the curriculum guidance in the following.

Definition 1: (**Curriculum Guidance**) Suppose the RL agent learns optimal policy via E interaction episodes with an environment, and let $\mathcal{I}$ be a set of the guidance from the human intervention. Then, the curriculum guidance consists of a sequence of the state-action pairs over two types of interventions, that is, continual guidance $\mathcal{I}_c = \{< s_t, a_t^{\mathbf{H}} >, \cdots, < s_T, a_T^{\mathbf{H}} >\}$ within M episodes and intermittent guidance $\mathcal{I}_i = \{< s_t, a_t^{\mathbf{H}} >, \cdots, < s_\tau, a_\tau^{\mathbf{H}} >\}$ within K episodes, where T and τ denote trajectory and intervention terminal time step, M and K subject to $0 < M < K < E$.

Intuitively, such a guidance mechanism is in line with human pedagogy. According to the modes of learning mentioned in [49], the authors mention that the whole-to-part pattern offers an exciting result, encouraging students to prioritize the construction of meaning rather than simply decoding text. For instance, when students acquire new knowledge in school, teachers often acquaint the child with the whole framework first, and then the different portions of it may be analyzed and studied intensively. Therefore, following the above-mentioned principle, the curriculum guidance aims to motivate the SafeHIL-RL to acquire global information over a complete trajectory at the early stage through continual guidance and fine-tuning the local behavior via intermittent human guidance in the subsequent learning process, ultimately guiding toward the smooth and robust RL learning process.

Integrating all the elements above, we present the detailed implementation of our algorithm in Algorithm 1.

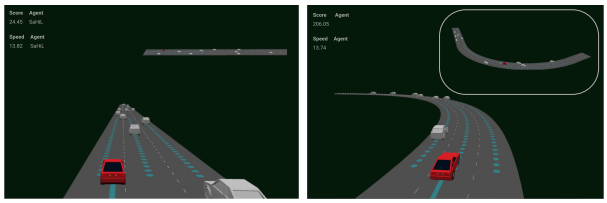
Algorithm 1 Safety-Aware Human-in-the-Loop Reinforcement Learning (SafeHIL-RL)

Initialize network parameters: ϕ, θ;
Initialize temperature parameter: α;
Determine batch size N and initialize replay buffer $\mathcal{D} \leftarrow \varnothing$;
Determine the episode threshold of curriculum guidance M (continual) and K (intermittent);
for *episode=1 to* E_1 **do**
  Initialize the environment: $s_t \sim Env$;
  while *not done* **do**
    Observe the state s_t;
    if *episode* $< M$ **then**
      if *Intervened (continual guidance)* **then**
        Human Takeover: $a_t^s = a_t^{\mathbf{H}}$
      end
    else if $M \leq$ *episode* $< K$ **then**
      if *Intervened (intermittent guidance)* **then**
        Authority Λ_t allocation based on FDPF;
        Human-AI Shared Control: $a_t^s \leftarrow$ Eq. 5;
      end
    else
      Sample an action: $a_t^s = a_t^{RL} \leftarrow \pi_\phi(a_t|s_t)$;
    end
    Interact with the environment $r_t, s_{t+1} \sim Env$;
    Store the tuple:
      $\mathcal{D} \leftarrow \mathcal{D} \cup (s_t, a_t^s, r_t, s_{t+1}, a_t^{\mathbf{H}}, \Lambda_t, \mathcal{I})$;
    if *time to update* **then**
      Sample a batch (N) of the data;
      Calculate critic loss $\mathcal{L}(\theta)$ based on Eq. 7;
      Update critic network parameter θ;
      Calculate actor loss $\mathcal{L}(\phi)$ based on Eq. 8;
      Update actor network parameter ϕ;
    end
  end
end

IV. Experiments

A. Autonomous Driving Scenarios

We evaluate the performance of the proposed approach in solving highway autonomous driving issues under dense traffic flows. In this study, two highway scenarios are designed, as shown in Fig. 4. Figure 4(a) demonstrates the first-person and top-down bird's eye view of the road users and environment where the SafeHIL-RL agent trains its policy. In Fig. 4(b), an additional unobserved highway environment that consists of straight and curve roads is introduced to validate the generalization capability of the post-trained driving policy. In our case, evaluating the performance under curve structure is significant since we would like to confirm that the policy trained by our approach is not over-fitted to the human demonstration under the straight road but learns the causal inference from states to actions.



(a) Scenario I. (b) Scenario II.

Fig. 4. Highway environment in SMARTS platform; a) Straight road. b) Straight road with the curve.

To replicate authentic highway scenarios, we utilize domain randomization to introduce randomness in road users' initial positions, speed distribution, and driving behaviors using Gaussian modeling. Furthermore, these road users have the ability to actively change lanes with varying degrees of willingness to cooperate. By combining all these elements, we regularly generate high-density traffic flows to ensure nearby vehicles are always available for interaction.

B. MDP Formulation

In this section, we present the MDP formulation of our approach.

1) State Space: The 100 × 100 bird-eye view (BEV) top-down images are taken as the image state for the RL agent. The three most recent frames are stacked to enable the temporal perception, and thereby, the dimension of the image state is 100 × 100 × 3. Moreover, we employ the previous actions as the conditional state, encoding it to the latent features and feed to the subsequent networks.

2) Action Space: The final output of the RL model consists of three elements: pedal, braking, and steering. The pedal and braking values are restricted within the range of 0 to 1, while steering ranges from −1 to 1, representing the left and right directions, respectively.

3) Reward Function: Our reward function comprises both heuristic and sparse rewards, denoted as:

$$r(s_t, a_t) = r_h + r_s \tag{13}$$

The heuristic term in Eq. 13, as a continuous and dense reward, should consider driving efficiency and performance. Therefore, we shape the heuristic reward as follows:

$$r_h = r_{speed} + r_{shoulder} + r_{lane} \tag{14}$$

where $r_{speed} = \omega_{speed} \cdot v_{speed}$ encourages the ego vehicle to move forward and $r_{shoulder}$ is a small punishment whenever the ego vehicle drive on the shoulder. Moreover, r_{lane} is a cost function that motivates the ego vehicle to keep the lane center an drive along the road direction, calculated as:

$$r_{lane} = -\omega_{offset} \cdot (\Delta l)^2 - \omega_{heading} \cdot (\Delta\varphi)^2 \tag{15}$$

where Δl and $\Delta\varphi$ represent for the lateral offset and heading error.

The sparse reward is mainly related to the driving task and safety, shaped as follows:

$$r_s = r_{goal} + r_{collision} + r_{offroad} \tag{16}$$

where r_{goal}, $r_{collision}$, and $r_{offroad}$ are reward for arriving goal position, punishment for collision and drive off the road, denoted as:

$$\begin{aligned} r_{goal} &= \begin{cases} 3.0, & if \quad goal\ reached \\ 0, & otherwise \end{cases} \\ r_{collision} &= \begin{cases} -7.0, & if \quad crashed \\ 0, & otherwise \end{cases} \\ r_{offroad} &= \begin{cases} -7.0, & if \quad offroad \\ 0, & otherwise \end{cases} \end{aligned} \tag{17}$$

TABLE I
PARAMETER SETTINGS FOR SAFEHIL-RL

Parameter	Value
minibatch size	32
reply buffer size	50000
learning start episode	10
discount factor	0.99
action repeat	1
actor network update frequency	1
critic network update frequency	1
target network update parameter	0.005
RMSProp learning rate for actor network	0.001
RMSProp learning rate for critic network	0.001
Adam learning rate for temperature parameter	0.0005

C. Baseline Algorithms

In this paper, we employ several SOTA LfI and RL algorithms to thoroughly validate the proposed approach. These baselines are:

1) **HIRL**: Human intervention reinforcement learning (HIRL) [17] is the first work attempt to formalize online human intervention into the RL framework. The action from the RL agent would be replaced by the one from human policy when human intervention is triggered, avoiding the catastrophic actions that occur during the training.
2) **PHIL-RL**: The SOTA LfI method known as Prioritized human-in-the-hoop RL (PHIL-RL) [37] aims to address the end-to-end autonomous driving task by incorporating human intelligence into the training process. In contrast to our proposed approach, the PHIL-RL algorithm employs an takeover strategy when intervention occurs, granting complete control authority to the human. The PHIL-RL utilizes the prioritized replay technique to augment performance, which is not the primary focus of this paper; therefore, we omit the prioritization and concentrate our attention solely on the learning mechanism associated with human intervention.
3) **SAC**: Soft actor-critic (SAC) is a SOTA off-policy RL algorithm that obtains the optimal policy by balancing the overall future payoff and Shannon entropy. In our experiment, we employ the optimal dual version that automatically adjusts the entropy of the behavior policy proposed in [50].

D. Configurations

All algorithms are trained on a computer equipped with an Intel Core i7-10700 CPU, 64 GB of RAM, and an NVIDIA GTX 1660 SUPER graphics card. The training process encompasses 800 episodes, with each episode comprising a maximum of 300 steps. An episode concludes when the target position is reached, a collision transpires, or the ego vehicle exhausts its allotted number of steps. A detailed parameter settings are reported in Table I.

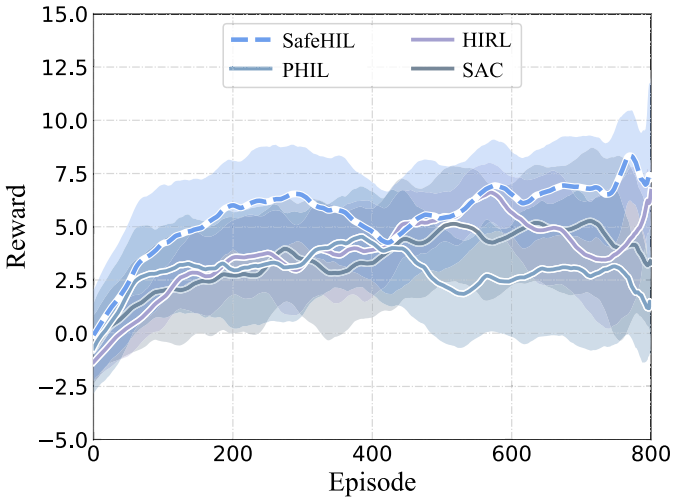


Fig. 5. Reward curve of SafeHIL-RL and baselines. The cornflower-blue dotted line and solid lines represent the average rewards of our approach and baselines per episode, while the shaded areas depict the variances over five runs.

V. Results and Discussion

This section intensively evaluates and presents the training and testing performance of the proposed approach against the SOTA baselines mentioned above, followed by a detailed analysis of the effectiveness of the FDPF and curriculum guidance through an ablation study.

A. Training Performance

Figure 5 illustrates the learning curves of SafeHIL-RL and other baselines. We run each algorithm with five different random seeds to measure statistics and evaluate the performance. The average rewards per episode are represented by the cornflower-blue dotted line for SafeHIL-RL and solid lines with various colors for the baselines. In addition, the shaded areas indicate the variances over the five runs, representing the robustness of each approach. The provided figure reveals that our approach exhibits the fastest convergence rate and achieves the highest level of rewards when compared to other SOTA baselines. The pure RL algorithm, the SAC, exhibits the slowest learning process, indicating its limited capability in effectively addressing the complex task of end-to-end autonomous driving control in such a challenging environment. Furthermore, HIRL demonstrates improvement compared to the SAC algorithm in terms of convergence speed, thanks to the integration of human guidance. Conversely, although the reward curve of the PHIL-RL initially displays rapid growth during the early stages where human intervention is allowed, the subsequent learning process without human guidance experiences a significant degradation in performance, which is deemed unacceptable within the LfI framework. One potential explanation is that the PHIL-RL approach operates under the assumption that human guidance is consistently reliable during the intervention, leading to an adversarial impact on the actor network when inadequate demonstrations or incorrect guidance is provided. In contrast, our approach effectively addresses this issue by evaluating real-time safety measures, enabling the filtration of improper guidance. As a result, our method ensures a resilient learning process regardless of the guidance quality.

TABLE II
Efficiency Improvement (%) of SafeHIL-RL Compared With Baselines

	$Epoc\ (Base \mid R^*(Base))$	$Epoc\ (SafeHIL \mid R^*(Base))$	Improved Percentage(%)
PHIL [37]	382	92	71.52
HIRL [17]	800	573	28.38
SAC [50]	712	137	80.76

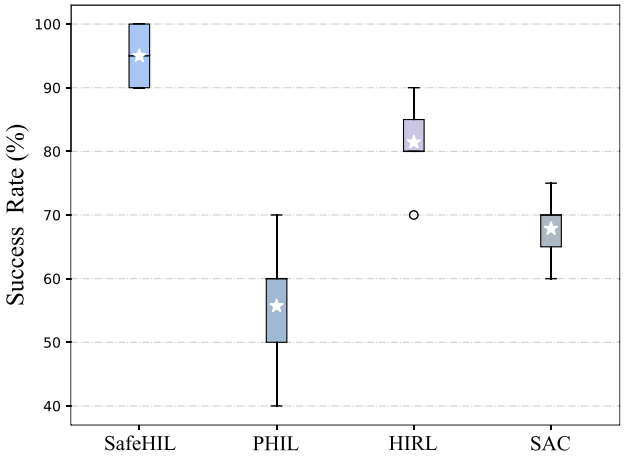


Fig. 6. Boxplot of success Rate. The black-solid line and "star" located at the box body denote the median and average.

In addition to evaluating the robustness of our training method, we aim to quantitatively assess the superior data efficiency of our approach by calculating the magnitude of average efficiency improvement of SafeHIL-RL compared to baseline methods. This average efficiency improvement metric is computed using the following formula:

$$\frac{Epoc(Base|R^*(Base)) - Epoc(SafeHIL|R^*(Base))}{Epoc(Base|R^*(Base))} \quad (18)$$

where $Epoc(Base|R^*(Base)$ represents the number of the episode the baseline needs to achieve its best reward and $Epoc(SafeHIL|R^*(Base))$ denotes the number of the episode our approach requires to reach the baseline's best reward. The comparison between our method and the baseline approaches is detailed in Table II. It is evident that SafeHIL-RL significantly enhances data efficiency over other LfI and DRL baselines. For example, our method surpasses the PHIL-RL strategy by as much as 71.52% and requires 80.76% less data than the SAC algorithm on average, exhibiting the most superior data efficiency among all approaches.

B. Testing Performance

Besides the learning curves, post-trained performance is also a significant metric to evaluate the superiority of the different approaches. To this end, we test the success rate for all the trained policies under scenario I (Fig. 4(a)) with ten additional unobserved seeds and run for twenty episodes for each. The success rate is defined as the number of episodes the ego vehicle drives without collision or off-road, divided by the total episodes run for each seed. The final result is demonstrated in Fig. 6. The boxplot clearly depicts that the overall performance

TABLE III
TESTING PERFORMANCE OF SAFEHIL-RL AND BASELINES

Performance	Method	Dynamic Metric			Static Metric
		Average Speed (m/s)	Average Distance (m)	Average Lateral Offset (m)	Success Rate (%)
Upper Bound	SafeHIL-RL	10.60 ± 2.61	**200.84 ± 0.30**	**0.22 ± 0.15**	**100**
	PHIL-RL[37]	9.82 ± 1.73	159.87 ± 67.27	0.39 ± 0.05	70
	HIRL[17]	9.36 ± 1.20	188.41 ± 35.99	**0.23 ± 0.04**	90
	SAC[50]	**13.24 ± 0.83**	157.67 ± 69.15	0.73 ± 0.09	75
Lower Bound	SafeHIL-RL	9.48 ± 2.20	**196.55 ± 9.15**	0.31 ± 0.15	**90**
	PHIL-RL[37]	10.34 ± 3.22	109.48 ± 70.22	0.54 ± 0.35	40
	HIRL[17]	8.94 ± 1.37	167.12 ± 67.50	**0.27 ± 0.14**	70
	SAC[50]	**13.33 ± 0.33**	146.65 ± 67.10	0.79 ± 0.12	60

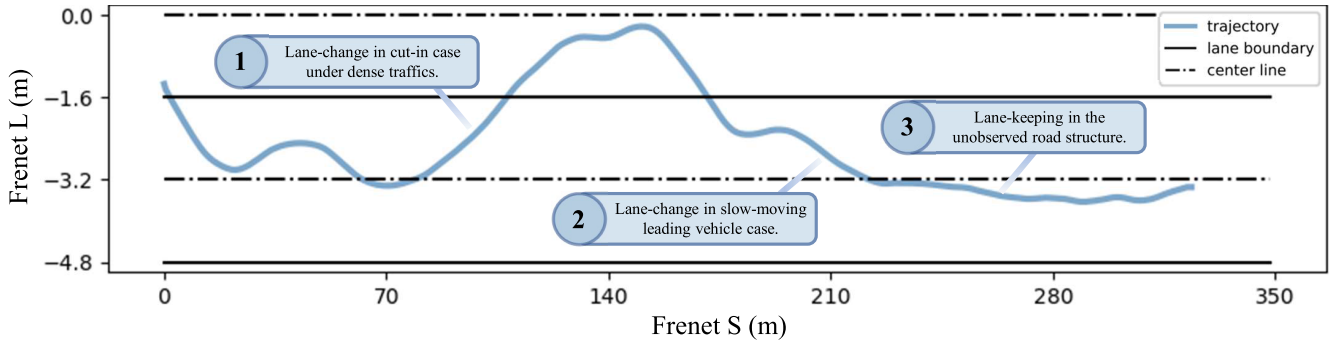
of our approach completely dominates that of other SOTA baselines in terms of average success rate and corresponding variance, demonstrating superiority and robustness in addressing end-to-end autonomous driving tasks. Furthermore, we presented the statistical analysis (refer to Table III) of the performance of different algorithms across multiple trials. This analysis allows us to quantitatively evaluate the upper and lower limits of driving performance. Notably, the SafeHIL-RL agent drives the longest distance given the fixed time steps in any case, and most importantly, the performance deviation between the upper and lower bounds is obviously less than that of other baselines. Although the SAC agent exhibited a higher average driving speed compared to the other agents, this was achieved at the expense of a lower success rate, which is entirely undesirable from the safe autonomous driving perspective. As for the keeping lane performance along the lane center, the HIRL agent demonstrated a slightly lower average lateral offset compared to our approach. This can be attributed to the conservative nature of the policy trained by the HIRL method. The HIRL policy tends to slow down and maintain lane position whenever encountering slow-moving leading vehicles. On the contrary, the policy trained by our approach prefers to perform lane-changing behavior to actively get rid of the surrounding vehicles unless the dynamically congested traffic (where all lanes are occupied by slow-moving vehicles) necessitates slowing down. This is also a key reason why the average driving distance of the SafeHIL-RL agent is significantly longer than that of HIRL. These results collectively demonstrate that our proposed approach leads the reinforcement learning (RL) agent to converge towards a more desirable, robust, and safer driving policy compared to SOTA methods.

To confirm the generalization ability and verify whether our approach is overfitted to the human demonstration, we further evaluate the performance of the proposed approach in the unobserved environment, which consists of the 200m straight road and 120m curve, shown in Fig. 4(b). If the performance presented in the Table. III is due to the overfitting of the human guidance, the new road structure, which requires a consistent steering operation unlike the straight road, would significantly challenge the performance of the SafeHIL-RL. In light of this, we tested the performance of the SafeHIL-RL with extra episodes and recorded the ego vehicle's trajectory information under the Frenet coordinate system. Figure 7 illustrates one example of the recorded data over an entire trajectory. The cornflower blue demonstrates the trajectory of the ego vehicle, while the solid and dashed line in black color represents the lane boundary and center line, respectively. From the figure, we can see that the trained policy under the proposed approach successfully completes the episode without any collision by performing lane-keep and lane-change operations. Especially we highlight three moments along the trajectory for better demonstration. Figure 7(b) illustrates a dense traffic flow scenario where one of the surrounding vehicles cuts into the ego lane (the lane where the ego vehicle locates). Motivated by the speed reward and human guidance, the ego vehicle executes the safe lane-change decision, rather than the slow-down operation, to actively get rid of the potential traffic congestion case that may significantly impact driving efficiency. A similar condition on the junction that connects the straight and curve is shown in Fig. 7(c) but with sparser surrounding traffic. In Fig. 7(d), the ego vehicle demonstrates the lane-keep performance on the curve. We surprisingly find that the trained policy under the SafeHIL-RL approach still can drive along the center line of the ego lane even though such a road structure is never seen during the training process, indicating the excellent generalization ability of the proposed method.

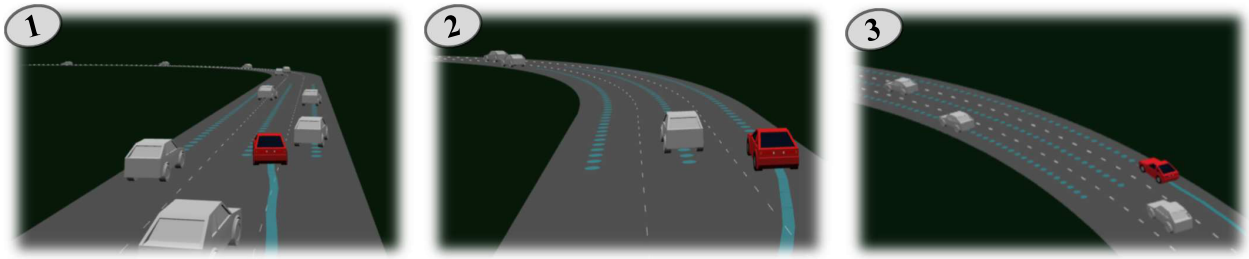
C. Frenet-Based Dynamic Potential Field Visualization

In this section, we would like to visualize and highlight the dynamic effect of the proposed FDPF model. One of the key benefits of establishing the safety assessment model based on the Frenet coordination system is that it is straightforward to measure and visualize the potential risk along a curve in the space, no matter the curvature of the road, since the Frenet coordination system uses the arc length of the curve as the reference axis. Moreover, we incorporate the dynamic information of the environment into the safety field modeling, as shown in Eq. 10, to extend or shrink the safety boundary depending on the situation.

We now provide an example to visualize the dynamic effect of the FDPF in Fig. 8. Figure 8(a) illustrates three moments of the highway driving scene captured from one trajectory. From the illustration, we can observe that the ego vehicle encounters

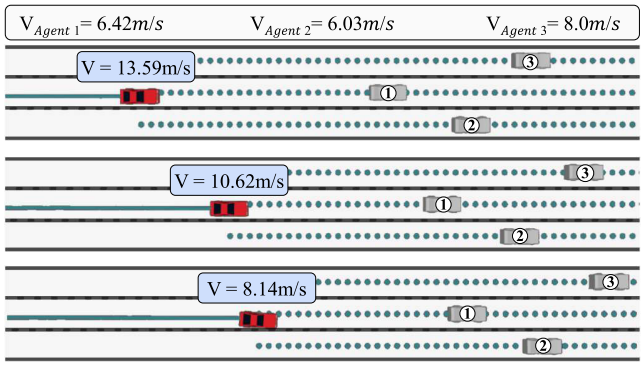


(a) Recorded trajectory of the ego vehicle.

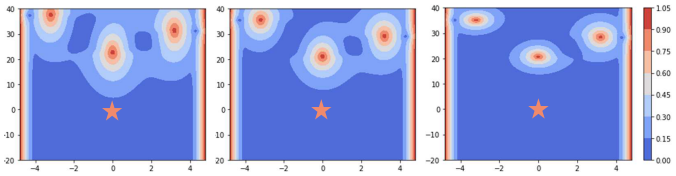


(b) Moment I. (c) Moment II. (d) Moment III.

Fig. 7. An example of the recorded entire trajectory performed by the policy trained through SafeHIL-RL method; a) Trajectory of the ego vehicle performs under Frenet coordination system; b) Moment I: Lane-change in cut-in case with dense surrounding traffics; c) Moment II: Lane-change in slow-moving leading vehicle case; d) Moment III: Lane-keeping in the unobserved road structure.



(a) Three moments captured from a trajectory.



(b) Moment I. (c) Moment II. (d) Moment III

Fig. 8. Visualization of the FDPF modeling. a) Three moments captured from a trajectory under the dynamic traffic congestion scene; b) Safety field at moment I; c) Safety field at moment II; d) Safety field at moment III.

a dynamic congested scenario. In this scene, a slow-moving leading vehicle with a speed of only 6.42 m/s is positioned ahead of the ego lane. Additionally, the other two lanes are occupied by vehicles traveling at similar speeds.

The constructed safety field at the first moment is visualized in Fig. 8(b). In this representation, the star denotes the position of the ego vehicle, and the transition from blue to red indicates an increasing potential hazard. In this case, it is evident that the ego vehicle is positioned near the safety boundary with a current driving velocity of 13.59 m/s, which is more than twice the speed of the leading vehicle. However, as the ego vehicle reduces its velocity, the safety boundary of the surrounding road users begins to shrink (Fig. 8(c)). At the third moment (Fig. 8(d)), the situation is deemed significantly safer compared to the first moment. Therefore, despite the relative distance being similar, any acceleration actions at the first moment would be considered potentially dangerous, whereas it is undoubtedly allowed for the third moment, highlighting the reasonable logic and flexibility in the proposed FDPF model.

D. Effect of Curriculum Guidance

Despite the fact that we have confirmed the superiority of curriculum guidance against pure continual guidance by comparing the performance with HIRL and PHIL-RL, there could still be uncertainty regarding whether the SafeHIL-RL's outstanding performance is a result of intermittent guidance rather than curriculum guidance. Therefore, we conduct an additional ablation study regarding the curriculum versus intermittent guidance. More specifically, we keep all the settings, including the hyperparameter, network structures, intervention times, and safety assessment, but we omit the continual guidance mode during training. Thus, human participants are only allowed to intervene intermittently during necessary periods instead of throughout the entire trajectory.

Figure 9 shows the training effort over three seeds. It is clear that the performance of the curriculum guidance completely dominates that of the intermittent guidance, especially at early stage, demonstrating the superiority of the proposed human guidance mechanism. This phenomenon is not surprising because pure continual guidance can lead to overfitting in the long run, while pure intermittent guidance may lack global information, resulting in a biased future payoff estimation.

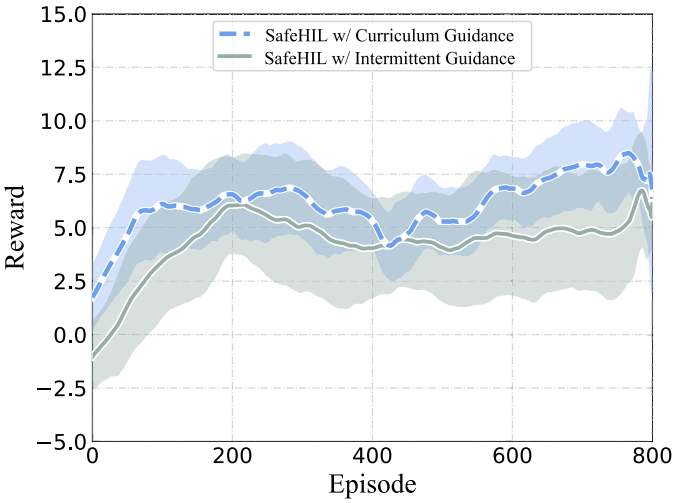


Fig. 9. Training curve of SafeHIL-RL with curriculum guidance versus intermittent guidance.

In contrast, incorporating curriculum guidance allows for the extraction of benefits from both approaches, effectively addressing the abovementioned issues and leading to a more robust and efficient learning process.

VI. Conclusion

In this paper, a safety-aware human-in-the-loop reinforcement learning (SafeHIL-RL) approach and the concept of curriculum guidance are proposed within the safe RL framework. In order to accomplish this, we present a safety assessment module based on the artificial potential field (APF) model that incorporates dynamic information of the environment under the Frenet coordinate system, which we call the Frenet-based dynamic potential field (FDPF), for evaluating the real-time safety throughout the human-guided learning process. We then introduce a curriculum guidance mechanism inspired by the whole-to-part pattern in pedagogy. This mechanism allows the RL agent to acquire global information early on through continuous guidance and fine-tune local behavior through intermittent human guidance with human-AI shared autonomy. As a result, the proposed approach enables a safe, robust, and efficient reinforcement learning process, regardless of the quality of guidance provided by human participants. The proposed method is validated in two complex highway autonomous driving scenarios, and the experiment results confirm the superiority and generality, as well as the effectiveness of the curriculum guidance, compared with other SOTA baselines. The universal framework and methodology proposed in this work have significant potential for real-world applications in the future, providing a safe and promising learning process for realizing the continuous improvement of learning-based policies after real-world deployment.

References

[1] L. Le Mero, D. Yi, M. Dianati, and A. Mouzakitis, "A survey on imitation learning techniques for end-to-end autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14128–14147, Sep. 2022.

[2] L. Yang, H. Yang, H. Wei, Z. Hu, and C. Lv, "Video-based driver drowsiness detection with optimised utilization of key facial features," *IEEE Trans. Intell. Transp. Syst*, pp. 1–13, Jan. 2024, doi: 10.1109/TITS.2023.3346054. [Online]. Available: https://ieeexplore-ieee-org.remotexs.ntu.edu.sg/document/10382460

[3] Z. Shan et al., "Safe and efficient trajectory planning considering longitudinal and lateral coupled limits for autonomous vehicles," *IEEE Trans. Veh. Technol*, pp. 1–6, Feb. 2024, doi: 10.1109/TVT.2024.3367125. [Online]. Available: https://ieeexplore-ieee-org.remotexs.ntu.edu.sg/document/10439656

[4] L. Chen et al., "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 2, pp. 1046–1056, Feb. 2023.

[5] Q. Liu, X. Li, S. Yuan, and Z. Li, "Decision-making technology for autonomous vehicles: Learning-based methods, applications and future outlook," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 30–37.

[6] S. Hwang, K. Lee, H. Jeon, and D. Kum, "Autonomous vehicle cut-in algorithm for lane-merging scenarios via policy-based reinforcement learning nested within finite-state machine," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17594–17606, Oct. 2022.

[7] W. Zhan, C. Liu, C.-Y. Chan, and M. Tomizuka, "A non-conservatively defensive strategy for urban autonomous driving," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 459–464.

[8] H. Wei, Y. Wang, J. Chen, and H. Zhang, "Resilient predictive control of constrained connected and automated vehicles under malicious attacks," in *Proc. IEEE 6th Int. Conf. Ind. Cyber-Physical Syst. (ICPS)*, May 2023, pp. 1–6.

[9] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58443–58469, 2020.

[10] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[11] H. Liu, Z. Huang, X. Mo, and C. Lv, "Augmenting reinforcement learning with transformer-based scene representation learning for decision-making of autonomous driving," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 3, pp. 4405–4421, Mar. 2024.

[12] X. He, W. Huang, and C. Lv, "Toward trustworthy decision-making for autonomous vehicles: A robust reinforcement learning approach with safety guarantees," *Engineering*, vol. 33, pp. 77–89, Feb. 2024.

[13] C. Zhang, W. Huang, X. Zhou, C. Lv, and C. Sun, "Expert-demonstration-augmented reinforcement learning for lane-change-aware eco-driving traversing consecutive traffic lights," *Energy*, vol. 286, Jan. 2024, Art. no. 129472.

[14] K. Yuan et al., "Evolutionary decision-making and planning for autonomous driving based on safe and rational exploration and exploitation," *Engineering*, vol. 33, pp. 108–120, Feb. 2024.

[15] J. Garcia and F. Fernández, "A comprehensive survey on safe reinforcement learning," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1437–1480, 2015.

[16] X. He et al., "Fear-neuro-inspired reinforcement learning for safe autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 267–279, Jan. 2024.

[17] W. Saunders, G. Sastry, A. Stuhlmüller, and O. Evans, "Trial without error: Towards safe reinforcement learning via human intervention," in *Proc. 17th Int. Conf. Auto. Agents MultiAgent Syst.*, 2018, pp. 2067–2069.

[18] F. Wang et al., "Intervention aided reinforcement learning for safe and practical policy optimization in navigation," in *Proc. Conf. Robot Learn.*, 2018, pp. 410–421.

[19] Q. Li, Z. Peng, and B. Zhou, "Efficient learning of safe driving policy via human-AI copilot optimization," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–19.

[20] V. Mnih, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[21] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell.*, 2016, vol. 30, no. 1, pp. 1–7.

[22] M. Fortunato et al., "Noisy networks for exploration," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–7.

[23] M. Hessel et al., "Rainbow: Combining improvements in deep reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–8.

[24] W. Huang, C. Zhang, J. Wu, X. He, J. Zhang, and C. Lv, "Sampling efficient deep reinforcement learning through preference-guided stochastic exploration," *IEEE Trans. Neural Netw. Learn. Syst*, pp. 1–12, Oct. 2023, doi: 10.1109/TNNLS.2023.3317628. [Online]. Available: https://ieeexplore-ieee-org.remotexs.ntu.edu.sg/document/10269149

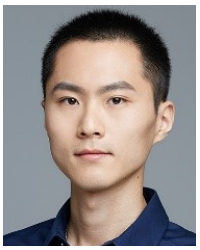
[25] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 1889–1897.

[26] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.
[27] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.
[28] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.
[29] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.
[30] A. Yu, R. Palefsky-Smith, and R. Bedi, "Deep reinforcement learning for simulated autonomous vehicle control," Standford Univ., Stanford, CA, USA, 2016. [Online]. Available: https://cs231n.stanford.edu/reports/2016/pdfs/112_Report.pdf
[31] W. Huang, F. Braghin, and S. Arrigoni, "Autonomous vehicle driving via deep deterministic policy gradient," in *Proc. Conf. Adv. Vehicle Technol., 16th Int. Conf. Design Educ.*, vol. 59216, Aug. 2019, Art. no. V003T01A017.
[32] W. Huang, F. Braghin, and Z. Wang, "Learning to drive via apprenticeship learning and deep reinforcement learning," in *Proc. IEEE 31st Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2019, pp. 1536–1540.
[33] A. Kendall et al., "Learning to drive in a day," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 8248–8254.
[34] P. Cai, X. Mei, L. Tai, Y. Sun, and M. Liu, "High-speed autonomous drifting with deep reinforcement learning," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1247–1254, Apr. 2020.
[35] J. Chen, S. E. Li, and M. Tomizuka, "Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5068–5078, Jun. 2022.
[36] W. Huang, Y. Zhou, X. He, and C. Lv, "Goal-guided transformer-enabled reinforcement learning for efficient autonomous navigation," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 1832–1845, Feb. 2024.
[37] J. Wu, Z. Huang, W. Huang, and C. Lv, "Prioritized experience-based reinforcement learning with human guidance for autonomous driving," *IEEE Trans. Neural Netw. Learn. Syst*, vol. 35, no. 1, pp. 855–869, Jan. 2024, doi: 10.1109/TNNLS.2022.3177685.
[38] W. Saunders, G. Sastry, A. Stuhlmueller, and O. Evans, "Trial without error: Towards safe reinforcement learning via human intervention," 2017, *arXiv:1707.05173*.
[39] J. Spencer et al., "Learning from interventions: Human–robot interaction as both explicit and implicit feedback," in *Proc. 16th Robot., Sci. Syst.* Cambridge, MA, USA: MIT Press, 2020.
[40] J. Wu, Z. Huang, Z. Hu, and C. Lv, "Toward human-in-the-loop AI: Enhancing deep reinforcement learning via real-time human guidance for autonomous driving," *Engineering*, vol. 21, pp. 75–91, Feb. 2023.
[41] G. Kahn, P. Abbeel, and S. Levine, "LaND: Learning to navigate from disengagements," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1872–1879, Apr. 2021.
[42] Z. Peng, Q. Li, C. Liu, and B. Zhou, "Safe driving via expert guided policy optimization," in *Proc. Conf. Robot Learn.*, 2022, pp. 1554–1563.
[43] L. Li, J. Gan, X. Ji, X. Qu, and B. Ran, "Dynamic driving risk potential field model under the connected and automated vehicles environment and its application in car-following modeling," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 122–141, Jan. 2020.
[44] W. Huang, Y. Zhou, J. Li, and C. Lv, "Potential hazard-aware adaptive shared control for human–robot cooperative driving in unstructured environment," in *Proc. 17th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Dec. 2022, pp. 405–410.
[45] J. Wang, J. Wu, and Y. Li, "The driving safety field based on driver-vehicle–road interactions," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2203–2214, Aug. 2015.
[46] H. Wang, Y. Huang, A. Khajepour, Y. Zhang, Y. Rasekhipour, and D. Cao, "Crash mitigation in motion planning for autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 9, pp. 3313–3323, Sep. 2019.
[47] M. Werling, J. Ziegler, S. Kammel, and S. Thrun, "Optimal trajectory generation for dynamic street scenarios in a Frenet frame," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 987–993.
[48] Y. Bengio, J. Louradour, and R. Collobert, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, Aug. 2009, pp. 41–48.
[49] H. Dombey and M. Moustafa, *Whole to Part Phonics: How Children Learn to Read and Spell*. Portsmouth, NH, USA: ERIC, 1998.
[50] T. Haarnoja et al., "Soft actor-critic algorithms and applications," 2018, *arXiv:1812.05905*.

Wenhui Huang (Graduate Student Member, IEEE) received the B.S. degree in vehicle engineering from Wuhan University of Technology, Wuhan, China, and the M.Sc. degree in mechanical engineering from the Polytechnic University of Milan, Milan, Italy, in 2018. He is currently pursuing the Ph.D. degree with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore. He was a Research Engineer with HoloMatic Technology (Beijing) Company Ltd., from 2019 to 2020. His research interests include autonomous driving, decision-making, foundation model, and lifelong reinforcement learning.

Haochen Liu (Graduate Student Member, IEEE) received the B.E. degree from the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China, in 2021. He is currently pursuing the Ph.D. degree with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore. His current research interests include deep learning-enabled motion prediction and decision-making.

Zhiyu Huang (Graduate Student Member, IEEE) received the B.E. degree from the School of Automobile Engineering, Chongqing University, Chongqing, China, in 2019, and the Ph.D. degree from the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore, in 2024. His current research interests include machine learning-based methods for perception, decision-making, testing, and simulation in autonomous driving, including deep reinforcement learning, behavior prediction, generative modeling, traffic simulation, and data-driven motion planning.



Chen Lv (Senior Member, IEEE) received the Ph.D. degree from the Department of Automotive Engineering, Tsinghua University, China, in January 2016. He was a joint Ph.D. Researcher with UC Berkeley, USA, from 2014 to 2015, and a Research Fellow with Cranfield University, U.K., from 2016 to 2018. He has been with NTU and founded the Automated Driving and Human–Machine System (AutoMan) Research Laboratory, since June 2018. He is currently a Nanyang Assistant Professor with the School of Mechanical and Aerospace Engineering and the Cluster Director of the Future Mobility Solutions, Nanyang Technological University, Singapore. His research interests include intelligent vehicles, automated driving, and human–machine systems.