2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Uncertainty-aware Reinforcement Learning for Autonomous Driving with Multimodal Digital Driver Guidance

Wenhui Huang[1], Zitong Shan[2], Shanhe Lou[1] and Chen Lv[1]

***Abstract*— While existing Learning from intervention (LfI) methods within the human-in-the-loop reinforcement learning (HiL-RL) paradigm mainly operate on the assumption that human policies are homogeneous and deterministic with low variance, natural human driving behaviors are multimodal with intrinsic uncertainties, and hence, accommodating diverse human capabilities is significant for its practical applications. This work proposes an enhanced LfI approach for learning the optimal RL policy by leveraging multimodal human behaviors in the setting of N-driver concurrent interventions. Specifically, we first learn the N number of human digital drivers from the multi-human demonstration dataset, wherein each driver possesses its own policy distribution. Then, the post-trained drivers will be kept in the training loop of the RL algorithms, providing diverse driving guidance whenever the intervention is required. Additionally, to better utilize the provided guidance, we augment the RL regarding the fundamental architecture and optimization objectives to facilitate the proposed uncertainty-aware reinforcement learning (UnaRL) algorithm. The proposed approach, which won 2^{nd} place in the Alibaba Future Car Innovation Challenge 2022, is solidly compared in two challenging autonomous driving scenarios against state-of-the-art (SOTA) LfI baselines, and results of both simulation and real-world experiment confirm the superiority of our method in terms of learning robustness and driving performance. Videos and source code are provided.** [1]

I. Introduction

Deep reinforcement learning (DRL) algorithms have demonstrated various degrees of success across numerous fields [1]–[3]. In recent years, researchers have made noteworthy contributions to the use of DRL algorithms for decision-making and control tasks, employing techniques such as deep Q-network (DQN) [4], twin delayed deep deterministic policy gradient (TD3) [5], and soft actor-critic (SAC) [6] to tackle both discrete and continuous action problems. Despite their impressive performance, DRL algorithms are notorious for their poor data efficiency, primarily due to the exploitation and exploration dilemma inherent in their learning mechanism. This issue is amplified when DRL algorithms are applied to complex and diverse tasks, such as learning reliable autonomous driving policies in dense traffic, as they often require hundreds or thousands of environmental interactions to converge [7].

Recently, the exploration of a novel paradigm referred to as learning from intervention (LfI), as situated within the context of human-in-the-loop reinforcement learning (HiL-RL) [8], has gained emerging attention in the RL community. LfI aims to block any catastrophic actions before they happen and override them with online human guidance, leading the RL agent with a safe and efficient exploration strategy. Existing works have made attempts to achieve such an approach by utilizing reward shaping or objective augmentation techniques for online human guidance and successfully demonstrated enhanced data efficiency in the Atari game [9], autonomous navigation [10], [11], and autonomous driving domain [12]–[16]. However, the LfI method remains in a nascent stage of development, as the previous works are founded on impractical assumptions, evident in two ways: Firstly, in the confidence that human policy is deterministic without uncertainties, and secondly, in the belief that human behaviors are homogeneous without diversity.

In the learning-based autonomous driving field [17], [18], especially with the aid of human intelligence, uncertainty can be categorized into two types: policy and model uncertainty [7], which are also entitled aleatoric and epistemic uncertainty [19], [20]. Recently, there has been a growing interest in incorporating uncertainty estimation into DRL and DIL approaches to address challenging issues in autonomous driving. For instance, extensive quantified uncertainty estimation metrics are presented in [21] for decision-making and control tasks in E2E autonomous driving. [22] proposes an uncertainty-aware imitation learning approach to construct a safe action selection mechanism by considering aleatoric uncertainty, validating the effectiveness through the simulation in the Carla platform [23]. Alternatively, [24] applies the epistemic uncertainty to the model-based RL algorithm for realizing uncertainty-aware E2E autonomous driving. Nevertheless, to the best of our knowledge, there has yet to be an existing work that formalizes uncertainty estimation into the LfI paradigm, enabling RL agents to accommodate multimodal and distinct human behaviors.

To bridge the abovementioned research gap, we present a novel learning from multimodal guidance (LfMG) approach for optimal policy learning by leveraging multimodal human behaviors in the setting of N-driver concurrent interventions. In particular, we learn the N-human digital drivers through the demonstration dataset gathered from multiple human drivers, encompassing distinct genders, age groups,

This work was supported in part by the A*STAR, Singapore, under AME YIRG (No.A2084c0156), the MTC IRG (M22K2c0079), the ANR-NRF Joint Grant (No.NRF2021-NRF-ANR003 HM Science), and the MOE Tier 2 Grant (MOE-T2EP50222-0002).

[1]W. Huang, S. Lou, and C. Lv are with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore, 639798. (E-mail: wenhui001@e.ntu.edu.sg, shanhe.lou@ntu.edu.sg, lyuchen@ntu.edu.sg)

[2]Z. Shan is with the State Key Laboratory of Automotive Simulation and Control, Jilin University, Changchun 130022, China. (E-mail: shanzt20@mails.jlu. edu.cn)

Corresponding author: Chen Lv. (E-mail: lyuchen@ntu.edu.sg)

[1]https://github.com/OscarHuangWind/Learning-from-Intervention.

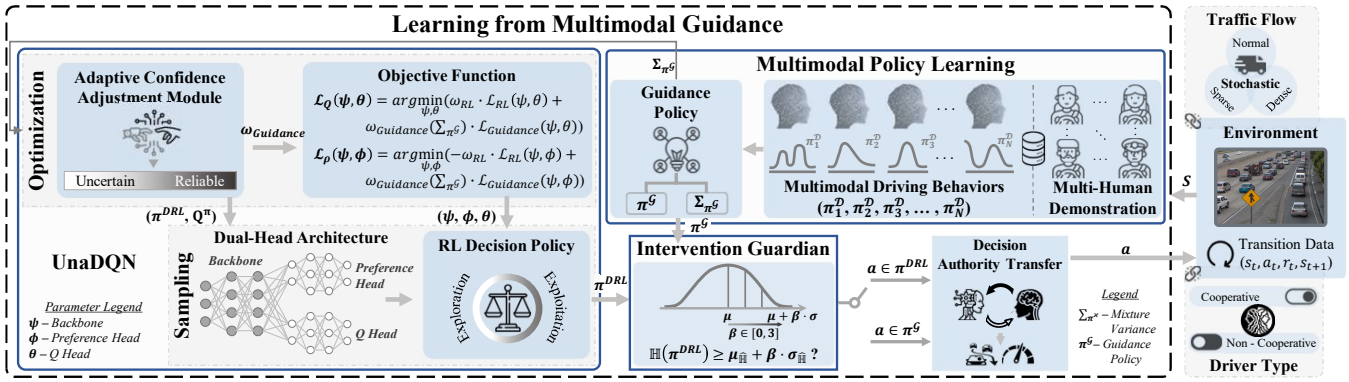


Fig. 1. Universal scheme of the proposed LfMG approach. LfMG consists of three main ingredients: multimodal policy learning, UnaRL, and intervention guardian.

and driving proficiency, and keep post-trained drivers in the training loop of the RL to provide multimodal guidance. Then, to enhance the utilization of the guidance, we design an intervention guardian and an augmented dual-head architecture deep reinforcement learning architecture for making decisions that align with human behavior preferences. Furthermore, we present an adaptive confidence adjustment module that optimizes the objectives based on the estimated confidence (uncertainties) over human demonstrations. This module adaptively assigns rational weights to objective functions during instances of N-driver interventions, realizing a dynamic learning process for the RL algorithm. Lumping all the above elements completes our proposed uncertainty-aware reinforcement learning (UnaRL) algorithm. We have solidly compared the proposed approach against the SOTA LfI baselines in the decision-making of two challenging autonomous driving scenarios and conducted real-world experiments to evaluate its real-time driving performance. The main contributions of this work are: 1) We propose a novel learning from multimodal guidance (LfMG) approach to break assumptions of homogeneous, deterministic human behaviors in the HiL-RL framework and bridge the research gap by accommodating diverse and multimodal human behaviors. 2) We design a concrete intervention guardian and an uncertainty-aware reinforcement learning (UnaRL) algorithm featuring a dual-head network architecture and adaptive confidence adjustment module to enhance the utilization of multimodal guidance, leading to efficient and robust policy optimization even in the face of uncertainty.

II. Method

The universal paradigm of the proposed LfMG approach is shown in Fig. 1. It consists of three essential ingredients: multimodal policy learning, uncertainty-aware reinforcement learning (UnaRL), and intervention guardian. In the following, we elaborate on details for each module.

A. Preliminaries

A typical decision-making problem in RL can be formulated as a standard Markov decision process (MDP) represented by a tuple $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$, where $\mathcal{S}$ is a set of states that the RL agent possibly visits, $\mathcal{A}$ denotes action space, $\mathcal{P}$ represents the transition of the environment, $\mathcal{R}$ is the reward function estimating the overall future return, and $\gamma \in (0,1]$ is a discount factor. At each time step t, the RL agent percepts the state $s_t \in \mathcal{S}$ and executes an action $a_t \in \mathcal{A}$, receiving an immediate reward $r_t = \mathcal{R}(s_t, a_t) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, as well as next state $s_{t+1} \in \mathcal{S}$ based on the transition probability $\mathcal{P}(s_{t+1}|s_t, a_t) : \mathcal{S} \times \mathcal{A} \to [0, 1]$. The policy $\pi(\cdot|s_t) : \mathcal{S} \to \mathcal{A}$ represents a probability distribution that the candidate action follows based on the current state s_t. Inherited from the dynamic programming (DP) property [25], the goal of the RL algorithm is to find a policy that maximizes the discounted overall future return, e.g., the state value $V^\pi(s_t)$ or state-action value $Q^\pi(s_t, a_t)$.

B. Multimodal Policy Learning

Human behavior is diverse and has a great deal of heterogeneity, which stems from the inherent uncertainty and limited cognition of human beings [26]. For instance, when driving on a highway and encountering a slow-moving vehicle ahead, a driver may reasonably decide to turn left or right [20]. Therefore, accommodating multimodal human policies and learning robust driving behaviors under uncertainties are substantial for the LfMG approach. It is crucial to highlight that the term 'multi-modality' used in this work refers to the diversity and heterogeneity of driving policies of reinforcement learning [27].

1) Policy Uncertainty. To investigate the inherent uncertainty in human behaviors, policy distribution, as opposed to deterministic decision command, is learned in this work. For classification tasks, e.g., decision-making problems in E2E autonomous driving, the uncertainty information can be captured by predictive entropy [21]. Notably, even though human policy is widely known to be non-deterministic, the data gathered during the demonstration are still represented in a deterministic format, i.e., $\pi^{\mathcal{H}}$ is a one-hot vector in decision-making tasks. Consequently, minimizing the forward Kullback-Leibler (KL) divergence (D_{KL}) [28] would cause the policy model to overfit the deterministic data, as it is equivalent to minimizing the cross-entropy between the learned policy and the one-hot vector. To avoid this issue,

we can instead minimize Reverse Kullback-Leibler (RKL) divergence [29], formulated as:

$$\varphi^* = \underset{\varphi}{argmin}\ D_{RKL}(\pi_\varphi^{\mathcal{D}} \mid\mid \pi^{\mathcal{H}}) \tag{1}$$

where $\pi_\varphi^{\mathcal{D}}$, $\pi^{\mathcal{H}}$, and φ denote the policy of the digital driver, human driver, and policy parameters. The Eq. 1 can be further expanded as:

$$\begin{aligned} \underset{\varphi}{min}\ D_{RKL}(\pi_\varphi^{\mathcal{D}} \mid\mid \pi^{\mathcal{H}}) &= \underset{\varphi}{min}\ -\sum \pi_\varphi^{\mathcal{D}} log(\frac{\pi^{\mathcal{H}}}{\pi_\varphi^{\mathcal{D}}}) \\ &\equiv \underset{\varphi}{min}\big(-\mathbb{E}_{\pi_\varphi^{\mathcal{D}}}[log\pi^{\mathcal{H}}]\big) - \underset{\varphi}{max}\ \mathbb{H}(\pi_\varphi^{\mathcal{D}}) \end{aligned} \tag{2}$$

where $\mathbb{H}$ and $\mathbb{E}$ represent Shannon entropy and expectation operation respectively. The first term in Eq. 2, identified as cross-entropy, elucidates that the learned policy $\pi_\varphi^{\mathcal{D}}$ can effectively pursue the primary mode of human policy $\pi^{\mathcal{H}}$. This is because the expectation is taken for the learned policy, demanding that samples generated by it exhibit a high likelihood under the human policy (mode-seeking behavior [30]). On the contrary, the second term concurrently motivates $\pi_\varphi^{\mathcal{D}}$ to maximize its Shannon entropy, thereby preventing the learned policy from converging to a sharp unimodal distribution. Therefore, minimizing the RKL divergence ensures the modeling of the policy uncertainty through optimization.

2) N-human Digital Drivers with Model Uncertainty. Even though we have considered policy uncertainty, one human digital driver can only represent one driving pattern and thus lack diversity and heterogeneity. Therefore, learning multiple human digital drivers to generate diverse and multimodal decision policies is crucial for practical applications of the LfMG method. Such diversity and multi-modality, originating from distinct cognitive abilities for interpreting the scenes, are referred to as model uncertainty or epistemic uncertainty. Estimating model uncertainty typically involves two methods: Monte-Carlo (MC)-dropout [31] and deep ensemble [32], with the latter employed in this study. Specifically, we split the collected dataset into ten groups and employed N networks to learn from these human demonstrations. In order to guarantee behavior diversity, we let each of the human digital driver models learn through randomly selected eight among ten groups with different initialization of the parameters. The demonstration data from selected groups were then divided into training and validation datasets, with a ratio of eight to two, and the learning process terminates either when the maximum iteration limit is reached or when the validation loss turns to increase. Subsequently, the ultimate guidance policy and the uncertainty can be obtained by integrating all the N-driver policy patterns in the form of a mixture mean and variance:

$$\begin{aligned} \pi^{\mathcal{G}}(\cdot|s) &= \frac{1}{N}\sum_{i=1}^{N} \pi_{\varphi_i}^{\mathcal{D}}(\cdot|s) \\ \Sigma_{\pi^{\mathcal{G}}}(s) &= \frac{1}{N}\sum_{i=1}^{N}\big(\Sigma_{\pi_{\varphi_i}^{\mathcal{D}}}(s)\big) + \frac{1}{N}\sum_{i=1}^{N}\big(\pi_{\varphi_i}^{\mathcal{D}}(\cdot|s)\big)^2 - \big(\pi^{\mathcal{G}}(\cdot|s)\big)^2 \end{aligned} \tag{3}$$

where i indicates i-th digital driver, the ultimate guidance policy $\pi^{\mathcal{G}}(\cdot|s) \in \mathbb{R}^{\mathbf{1}\times dim(\mathcal{A})}$ is constrained by $\sum_{j=1}^{dim(\mathcal{A})} \pi^{\mathcal{G}}(a_j|s) = 1$, and $\Sigma_{\pi^{\mathcal{G}}}(s)$ is a diagonal covariance matrix with $dim(\mathcal{A}) \times dim(\mathcal{A})$ dimension. Subsequently, the LfMG draws decision command from the guidance policy $\pi^{\mathcal{G}}(\cdot|s)$ during the intervention process and employs the mixture variance $\Sigma_{\pi^{\mathcal{G}}}(s)$ to compute adaptive weights in the adaptive confidence adjustment module for the RL optimization process.

C. Uncertainty-aware Reinforcement Learning

In this paper, we present a novel uncertainty-aware reinforcement learning (UnaRL) algorithm, namely UnaDQN, aimed at maximizing the effective use of multimodal guidance. It is worth noting that the "Una" paradigm does not apply any constraints to RL algorithms and thus can be flexibly adapted to off-policy actor-critic RL algorithms as well.

The UnaDQN is an enhanced algorithm based on our previous work PGDQN, which has confirmed its superior performance on OpenAI Gym benchmarks [33]. Fig. 2 demonstrates the overall architecture of the UnaDQN algorithm. We keep the dual architecture design of PGDQN for our algorithm and title each branch as preference head and Q head, sharing a common vision transformer (ViT) encoder as the backbone to learn the latent representation from diverse knowledge in multitask learning fashion [34]. Both heads consist of two fully connected layers with layer normalization [35] operation in between, and the output of the preference head is policy distribution while that of the Q head is state-action value Q.

Given a batch of $M = M_{DRL} + M_{\mathcal{G}}$ samples, the objective function of the Q head is minimizing the Bellman loss and imitation loss, denoted as follows:

$$\begin{aligned} \mathcal{L}_Q(\theta) = &\frac{1}{M_{DRL}}\sum_{i=1}^{M_{DRL}} \|r_t + \gamma \underset{a}{max} Q(s_{t+1}, a_{t+1}^{DRL}|\theta') - Q(s_t, a_t^{DRL}|\theta)\|^2 \\ &+ \frac{1}{M_{\mathcal{G}}}\sum_{j=1}^{M_{\mathcal{G}}} \|r_t + \gamma \underset{a}{max} Q(s_{t+1}, a_{t+1}^{DRL}|\theta') - Q(s_t, a_t^{\mathcal{G}}|\theta)\|^2 \\ &+ \frac{1}{M_{\mathcal{G}}}\sum_{j=1}^{M_{\mathcal{G}}} \boldsymbol{\omega}_j \cdot \|softmax\big(Q(s_t, \pi^{DRL}(\cdot|s_t)|\theta)\big) - \mathbb{1}_{a^{\mathcal{G}}}\|^2 \end{aligned} \tag{4}$$

where $Q(s_t, \pi^{DRL}(\cdot|s_t)|\theta) \in \mathbb{R}^{\mathbf{1}\times dim(\mathcal{A})}$ indicates a vector of Q values for all the actions and $\mathbb{1}_{a^{\mathcal{G}}} \in \mathbb{R}^{\mathbf{1}\times dim(\mathcal{A})}$ represents the one-hot encoding over the guidance action $a^{\mathcal{G}}$. In addition, $\boldsymbol{\omega}_j$ is the weight for j-th transition data, denoting the confidence held by the DRL agent regarding the guidance. Most importantly, this weight is adaptive to the uncertainty of the provided guidance, i.e., the mixture variance $\Sigma_{\pi^{\mathcal{G}}}(s)$, and dynamically computed and assigned by the adaptive confidence adjustment module for the parameter optimization.

As for the preference head, we first define the sampling policy of the UnaDQN algorithm. In this study, we employ the preference-guided ϵ-greedy policy [33] for UnaDQN instead of vanilla ϵ-greedy as the former is a generalized

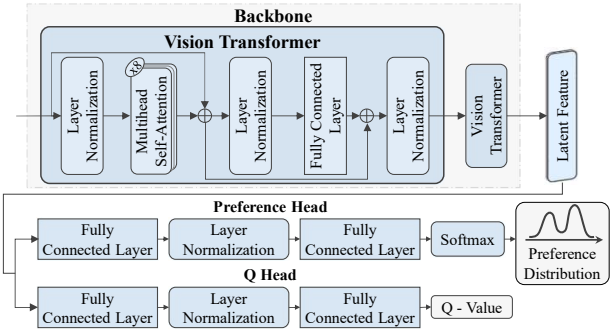


Fig. 2. The overall architecture of the UnaDQN. The input state is fed into the backbone network, which consists of two blocks of the ViT, and encoded to latent features; then, the latent features are delivered to the preference head and Q head to generate the preference distribution and estimated future payoffs.

form of the latter, possessing superior exploration efficiency. Therefore, given the state s_t at an arbitrary timestep t, the sampling policy is as follows:

$$\pi^{DRL}(a_t^{DRL}|s_t) = \begin{cases} 1-\epsilon+\epsilon\cdot\rho_\phi(a_t^{DRL}|s_t), & if\ a=a^* \\ \epsilon\cdot\rho_\phi(a_t^{DRL}|s_t), & a\neq a^* \end{cases} \tag{5}$$

where $a^* = argmax\ Q(a_t^{DRL}, s_t)$ denotes greedy action, ϕ is the parameters of the preference head and $\rho_\phi(a_t^{DRL}|s_t)$ indicates the preference probability of selected action a_t^{DRL} which is inferred by preference head. As shown in Eq. 5, the decision commands are sampled from preference distribution $\rho_\phi(\cdot|s_t)$ within the probability of ϵ or assigned as greedy actions otherwise. To mitigate the overestimation issue, we employ on-policy learning to train the preference distribution, and it is optimized by maximizing entropy augmented advantage function while simultaneously minimizing the cross-entropy between preference distribution $\rho(\cdot|s_t)$ and guidance policy $\pi^{\mathcal{G}}(\cdot|s_t)$, computed as follows:

$$\begin{aligned}\mathcal{L}_\rho(\phi) = &\mathop{\mathbb{E}}_{(s_t,a_t^{\mathcal{G}})\sim\mathcal{P}_{\pi^{\mathcal{G}}}}\left[A^\rho(s_t,a_t^{\mathcal{G}})\cdot\rho_\phi(a_t^{\mathcal{G}}|s_t)+\boldsymbol{\omega_t}\cdot\mathbb{H}(\rho(\cdot|s_t))\right] \\ &-\mathop{\mathbb{E}}_{(s_t,a_t^{\mathcal{G}})\sim\mathcal{P}_{\pi^{\mathcal{G}}}}\left[\boldsymbol{\omega_t}\cdot\mathbb{H}(\rho(\cdot|s_t),\ \pi^{\mathcal{G}}(\cdot|s_t))\right] \\ &=\mathop{\mathbb{E}}_{(s_t,a_t^{\mathcal{G}})\sim\mathcal{P}_{\pi^{\mathcal{G}}}}\left[A^\rho(s_t,a_t^{\mathcal{G}})\cdot\rho_\phi(a_t^{\mathcal{G}}|s_t)-\boldsymbol{\omega_t}\cdot D_{KL}\Big(\rho(\cdot|s_t)||\pi^{\mathcal{G}}(\cdot|s_t)\Big)\right]\end{aligned} \tag{6}$$

Unlike the existing works, the LfMG approach does not apply any restriction for the human drivers as it can effectively lean the robust driving policy through diverse and multimodal human behaviors. This advantage can be attributed to the adaptive weight adjustment module of UnaDQN through estimating the confidence of the real-time guidance based on the mixture variance. Inspired from [36], we first employ a dynamic normalization technique to normalize the mixture variances within the finite moving horizons k, computed as follows:

$$\begin{aligned}\boldsymbol{\Sigma} &= \{\Sigma_{\pi^{\mathcal{G}}}(s_{t-k+1}), \Sigma_{\pi^{\mathcal{G}}}(s_{t-k+2}), ..., \Sigma_{\pi^{\mathcal{G}}}(s_{t-1})\} \\ \overline{\Sigma}_{\pi^{\mathcal{G}}}(s_t) &= \frac{\Sigma_{\pi^{\mathcal{G}}}(s_t) - min\boldsymbol{\Sigma}}{max\boldsymbol{\Sigma} - min\boldsymbol{\Sigma}}\end{aligned} \tag{7}$$

(a) Left Turn.

(b) Ramp Merge.

(c) Driving simulator set.

Fig. 3. Testing scenarios and driving simulator set.

where $\overline{\Sigma}_{\pi^{\mathcal{G}}}(s_t) \in [0,1]$ is the normalized mixture variance. Then, we map it to the weight (confidence) of the guidance-related term in the exponential form:

$$\boldsymbol{\omega_t} = e^{-a\cdot\overline{\Sigma}_{\pi^{\mathcal{G}}}(s_t)+b} \tag{8}$$

where a and b are the scale and bias constant. The Eq. 8 demonstrates that when the mixture variance is high, the UnaDQN agent has low confidence in the provided guidance, resulting in a small weight value. Back to Eq. 6, we can observe that both Shannon entropy and cross-entropy term will be large when the weight is small, encouraging the DRL agent to explore policy by itself and learn less from the guidance policy. On the contrary, if the DRL agent has high confidence in the provided guidance, i.e., assigns a high value to the weight, it will focus more on imitating the guidance policy rather than exploring diverse decision options.

D. Intervention Guardian

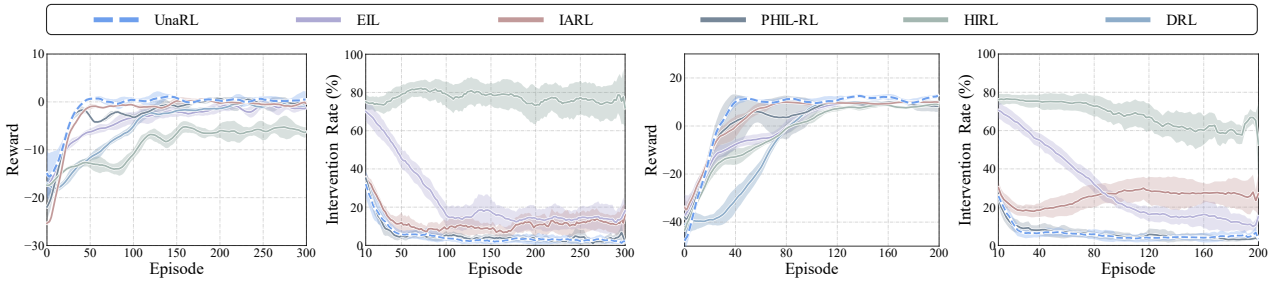
In order to motivate the RL agent to learn driving policies simultaneously from free exploration and multimodal guidance, we introduce an intervention guardian system. This mechanism allows the guidance policy to intervene in vehicle control in a situation where the policy uncertainty of the RL agent is higher than a certain threshold. Drawing inspiration from the HG-DAgger [37], we learn a dynamic threshold from N-driver interventions instead of selecting a fixed empirical number. In particular, we capture the entropy of guidance policy at each time step, subsequently deriving the moving average and moving standard deviation across finite moving horizons τ:

$$\begin{aligned}\hat{\mathbb{H}} &= \{\mathbb{H}(\pi^{\mathcal{G}}(\cdot|s_{t-\tau+1})), \mathbb{H}(\pi^{\mathcal{G}}(\cdot|s_{t-\tau+2})), ..., \mathbb{H}(\pi^{\mathcal{G}}(\cdot|s_{t-1}))\} \\ \mu_{\hat{\mathbb{H}}} &= mean(\hat{\mathbb{H}}), \quad \sigma_{\hat{\mathbb{H}}} = std(\hat{\mathbb{H}})\end{aligned} \tag{9}$$

Then we compute the dynamic threshold as follows:

$$\delta = \mu_{\hat{\mathbb{H}}} + (-\frac{3.0}{\mathbb{H}(U)}\cdot\mathbb{H}(\rho(\cdot|s_t)) + 3.0)\cdot\sigma_{\hat{\mathbb{H}}} \tag{10}$$

where U denotes uniform distribution. From the above equation, we can observe that the threshold anneals to the moving average at the initial stages when the RL agent considers each decision an equal possibility. However, as the reinforcement learning agent refines its decision policy through training, the intervention threshold becomes progressively more stringent, culminating in an eventual convergence to the value corresponding with the 3σ rule of thumb.



(a) Left Turn. (b) Left Turn. (c) Ramp Merge. (d) Ramp Merge.

Fig. 4. The reward and intervention rate curves for LfMG and baselines.

III. Experiments and Results

A. Experimental Setup

Although the efficacy of the proposed algorithm has been confirmed through its achievement of the second position in the Alibaba Future Car Innovation Challenge 2022 [2], this paper aims to extend the comparison of our method against SOTA LfI baselines by applying it to two additional challenging autonomous driving scenarios in SMARTS [38]: namely, the ramp merge and unprotected left turn (shown in Fig. 3(a) and Fig. 3(b)), employed in the 2022 NeurIPS Driving SMARTS Competition [3]. The primary objective in both scenarios is to effectively interact with various types (cooperative and non-cooperative) of social agents, execute a successful merge into the junction or intersection, and navigate safely to the outer lane of the main road under highly stochastic traffic flow. In order to accommodate multi-modal human behaviors, we hired twenty human participants with distinct genders, age groups, and driving proficiency to perform the demonstrations in these scenarios through driving monitors and the Logitec G29 set, as shown in Fig. 3(c). We obtained a total of 180 trajectories from human demonstrations in the form of state-action pairs. During the policy learning process, each digital driver can access a random 80% of the whole data, and the selected dataset is further divided by 80% and 20% for policy training and validation.

As for the RL training, the state space consists of three consecutive top-down bird-eye-view (BEV) RGB images with the dimension of 120 $\times$ 120 $\times$ 3, denoted as $s_t \in \mathbb{R}^{\mathbf{120}\times\mathbf{120}\times\mathbf{9}}$. As for the action space, our model takes charge of making high-level decisions, encompassing *Keep Lane*, *Slow Down*, *Change Lane Right*, *Change Lane Left*, and adopts the speed and lane-following controllers for the subsequent command execution. Last but not least, our reward function contains both heuristic and sparse rewards related to ego speed, action consistency, intervention cost, on-shoulder driving, goal-reaching, and off-road/crash penalty.

To thoroughly confirm the superior performance of the proposed approach, we employ SOTA LfI and DRL algorithms as our baselines, and they are HIRL [9], IARL [10], EIL [13], PHIL-RL [15], and dueling double DQN (D3QN) [39]. We allow digital drivers to intervene only in even-numbered episodes to prevent data distribution shifts or overfitting issues related to guidance policy. In odd-numbered iterations, the DRL agent is liberated from intervention and trained in standard RL fashion.

TABLE I
Efficiency Improvement (%) of UnaRL compared with Baselines.

	$Epoc\ (Base \mid R^*(Base))$					Improved Percentage(%)				
Scenario	HIRL	IARL	EIL	PHIL-RL	D3QN	>HIRL	>IARL	>EIL	>PHIL-RL	>D3QN
Left Turn	161	184	253	235	276	80.75	71.20	**81.82**	76.60	**82.25**
Ramp Merge	161	83	136	125	121	**73.91**	48.19	68.38	65.60	**64.46**

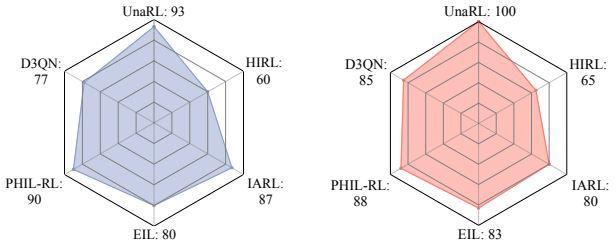
B. Benchmark Comparison: Qualitative Results

The learning curves of UnaRL and all the baseline methods are given in Fig. 4. We train each algorithm with five random seeds to measure statistical performance. The average rewards per episode are represented by the cornflower-blue dotted line for UnaRL 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. As shown in Fig. 4(a) and 4(c), the UnaRL method exhibits the fastest convergence speed with the lowest variance compared to other LfI and DRL baselines, demonstrating superior data efficiency and robust characteristics. The worst performance in the unprotected left turn scenario originated from the HIRL, while that of the ramp merge scenario belongs to the D3QN algorithm due to the notoriously known poor data efficiency issue.

Furthermore, an analysis of the intervention rate reveals notable insights, as depicted in Fig. 4(b) and Fig. 4(d). It is evident that the intervention rates of UnaRL and PHIL-RL asymptotically reduce to nearly zero towards the conclusion of the learning process. This trend signifies that both approaches have effectively learned optimal policies, enabling the ego vehicle to drive safely with minimal interventions. A similar trend can be observed in the EIL method. On the contrary, the intervention rate for HIRL consistently maintains a relatively high level throughout the entire learning phase in both scenarios, demonstrating the worst performance among

[2] https://tianchi.aliyun.com/competition/entrance/531996/rankingList

[3] https://codalab.lisn.upsaclay.fr/competitions/6618



(a) Left Turn. (b) Ramp Merge.

Fig. 5. Average success rate (%) of UnaRL compared with baselines.

all the algorithms. Interestingly, the intervention curve of IARL experiences an initial rapid decline but subsequently exhibits an upward trend in the later stages of learning, ultimately converging into an unsatisfactory suboptimal policy.

C. Benchmark Comparison: Quantitative Results

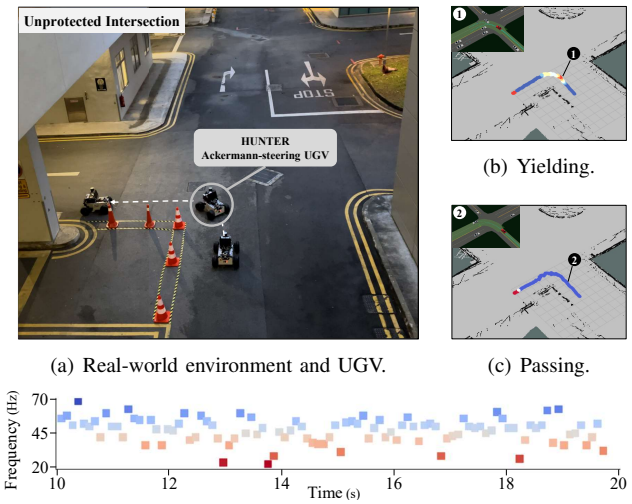
To quantitatively measure the superiority of our approach, we compute the amplitude of efficiency improvement of UnaRL against baselines according to the following metric:

$$\frac{Epoc(Base|R^*(Base)) - Epoc(UnaRL|R^*(Base))}{Epoc(Base|R^*(Base))} \quad (11)$$

where $Epoc(Base|R^*(Base)$ represents the number of the episode the baseline needs to achieve its best reward and $Epoc(UnaRL|R^*(Base))$ denotes the number of the episode our approach requires to reach the baseline's best reward. Furthermore, we meticulously select the optimal model for both our algorithm and the baseline methods, conducting success rate assessments with three additional, previously unobserved seeds, each comprising ten episodes. The comprehensive comparative results encompassing various statistical metrics are presented in Table I and Fig. 5. It is clear that UnaRL demonstrates the efficiency enhancement by a large margin against LfI and DRL baselines in terms of data efficiency. For instance, our approach outperforms the EIL method by up to 81.82% and consumes 73.91% less data than the HIRL on average, exhibiting the fastest learning speed among all approaches. As for the testing performance, the success rate of UnaRL (93% and 100% on average over three unseen seeds) decisively outperforms those of the baselines, demonstrating superior and robust characteristics of the proposed approach.

D. Real-world Experiment

To validate the feasibility and real-time performance, especially under limited onboard computational power, we also deployed the post-trained algorithm on the hardware platform and conducted real-world experiments. As Fig. 6(a) shows, the experiment is carried out in a real-world unprotected intersection scene at Nanyang Technological University with an Ackermann-steering unmanned grounded vehicle (UGV) called HUNTER. The HUNTER is equipped with an edge computing platform NVIDIA Jetson Xavier 16GB, an inertial measurement unit (IMU), and a 16-line LiDAR. To comply with legal requirements, we did not experiment with the natural traffic flow. Instead, we employ the stochastic traffic



(a) Real-world environment and UGV. (b) Yielding. (c) Passing. (d) Real-time on-board decision frequency.

Fig. 6. Real-world experiment settings and results.

flow from the SUMO and deliver the corresponding BEV images through the ROS. Then, the deployed algorithm must generate real-time decision inference with onboard computation and send it to the UGV chassis via CAN communication to realize motion control.

Fig. 6(b) and 6(c) present a visual representation of qualitative assessments regarding the driving performance of two distinct selected cases, as observed against ground-truth trajectories. Notably, these trajectories are color-coded by the velocity profile executed by the HUNTER, transitioning from blue (indicating maximum allowable speed) to red (signifying zero speed). The result clearly demonstrates that our algorithm can perform yielding behavior when encountering an oncoming right-of-way vehicle and effectively execute passing maneuvers with non-conservative driving decisions when interacting with other cooperative drivers. Moreover, Fig. 6(d) demonstrates a satisfactory real-time decision capability (with the least frequency of 22 Hz), underscoring the substantial potential of our transformer-enabled RL-based algorithm regarding practical applications.

IV. Conclusion

This paper proposes a novel learning from multimodal guidance (LfMG) approach to consider the multi-modality and intrinsic uncertainty of human behaviors in the context of the HiL-RL framework. More specifically, we learn N-human digital drivers from the multi-human demonstration and let them provide multimodal driving guidance during the RL learning process. To accommodate diverse human behaviors and learn robust driving policies under uncertainties, we present a concrete intervention guardian and an uncertainty-aware RL (UnaRL) algorithm equipped with dual-head architecture and an adaptive confidence adjustment module. Comprehensive simulation and real-world experiments conducted in two challenging autonomous driving scenarios confirm the superiority of our approach compared with SOTA baselines and reveal the substantial potential in real-world applications.

REFERENCES

[1] A. Kendall, J. Hawke, D. Janz, P. Mazur, D. Reda, J.-M. Allen, V.-D. Lam, A. Bewley, and A. Shah, "Learning to drive in a day," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8248–8254.

[2] X. He, W. Huang, and C. Lv, "Toward trustworthy decision-making for autonomous vehicles: A robust reinforcement learning approach with safety guarantees," *Engineering*, 2023.

[3] W. Huang, Y. Zhou, X. He, and C. Lv, "Goal-guided transformer-enabled reinforcement learning for efficient autonomous navigation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 2, pp. 1832–1845, 2024.

[4] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[5] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.

[6] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.

[7] Z. Huang, J. Wu, and C. Lv, "Efficient deep reinforcement learning with imitative expert priors for autonomous driving," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[8] D. Abel, J. Salvatier, A. Stuhlmüller, and O. Evans, "Agent-agnostic human-in-the-loop reinforcement learning," *arXiv preprint arXiv:1701.04079*, 2017.

[9] W. Saunders, G. Sastry, A. Stuhlmüller, and O. Evans, "Trial without error: Towards safe reinforcement learning via human intervention," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018, pp. 2067–2069.

[10] F. Wang, B. Zhou, K. Chen, T. Fan, X. Zhang, J. Li, H. Tian, and J. Pan, "Intervention aided reinforcement learning for safe and practical policy optimization in navigation," in *Conference on Robot Learning*. PMLR, 2018, pp. 410–421.

[11] G. Kahn, P. Abbeel, and S. Levine, "Land: Learning to navigate from disengagements," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1872–1879, 2021.

[12] W. Huang, F. Braghin, and Z. Wang, "Learning to drive via apprenticeship learning and deep reinforcement learning," in *2019 ieee 31st international conference on tools with artificial intelligence (ictai)*. IEEE, 2019, pp. 1536–1540.

[13] J. Spencer, S. Choudhury, M. Barnes, M. Schmittle, M. Chiang, P. Ramadge, and S. Srinivasa, "Learning from interventions: Human-robot interaction as both explicit and implicit feedback," in *16th Robotics: Science and Systems, RSS 2020*. MIT Press Journals, 2020.

[14] Q. Li, Z. Peng, and B. Zhou, "Efficient learning of safe driving policy via human-ai copilot optimization," in *International Conference on Learning Representations*, 2021.

[15] J. Wu, Z. Huang, W. Huang, and C. Lv, "Prioritized experience-based reinforcement learning with human guidance for autonomous driving," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[16] Z. Peng, Q. Li, C. Liu, and B. Zhou, "Safe driving via expert guided policy optimization," in *Conference on Robot Learning*. PMLR, 2022, pp. 1554–1563.

[17] W. Huang, F. Braghin, and S. Arrigoni, "Autonomous vehicle driving via deep deterministic policy gradient," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 59216. American Society of Mechanical Engineers, 2019, p. V003T01A017.

[18] X. He, J. Wu, Z. Huang, Z. Hu, J. Wang, A. Sangiovanni-Vincentelli, and C. Lv, "Fear-neuro-inspired reinforcement learning for safe autonomous driving," *IEEE transactions on pattern analysis and machine intelligence*, 2023.

[19] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[20] S. Choi, K. Lee, S. Lim, and S. Oh, "Uncertainty-aware learning from demonstration using mixture density networks with sampling-free variance modeling," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6915–6922.

[21] R. Michelmore, M. Kwiatkowska, and Y. Gal, "Evaluating uncertainty quantification in end-to-end autonomous driving control," *arXiv preprint arXiv:1811.06817*, 2018.

[22] L. Tai, P. Yun, Y. Chen, C. Liu, H. Ye, and M. Liu, "Visual-based autonomous driving deployment from a stochastic and uncertainty-aware perspective," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 2622–2628.

[23] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[24] J. Wu, Z. Huang, and C. Lv, "Uncertainty-aware model-based reinforcement learning: methodology and application in autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2022.

[25] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[26] Z. Zhu and H. Zhao, "A survey of deep rl and il for autonomous driving policy learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14 043–14 065, 2021.

[27] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *International conference on machine learning*. PMLR, 2017, pp. 1352–1361.

[28] S. K. S. Ghasemipour, S. Gu, and R. Zemel, "Understanding the relation between maximum-entropy inverse reinforcement learning and behaviour cloning," 2019. [Online]. Available: https://openreview.net/forum?id=rkeXrIIt_4

[29] A. Malinin and M. Gales, "Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[30] L. Ke, S. Choudhury, M. Barnes, W. Sun, G. Lee, and S. Srinivasa, "Imitation learning as f-divergence minimization," in *Algorithmic Foundations of Robotics XIV: Proceedings of the Fourteenth Workshop on the Algorithmic Foundations of Robotics 14*. Springer, 2021, pp. 313–329.

[31] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *International conference on machine learning*. PMLR, 2017, pp. 1183–1192.

[32] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[33] W. Huang, C. Zhang, J. Wu, X. He, J. Zhang, and C. Lv, "Sampling efficient deep reinforcement learning through preference-guided stochastic exploration," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[34] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," *Advances in neural information processing systems*, vol. 31, 2018.

[35] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[36] J. Yan, R. Wan, X. Zhang, W. Zhang, Y. Wei, and J. Sun, "Towards stabilizing batch statistics in backward propagation of batch normalization," in *International Conference on Learning Representations*.

[37] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer, "Hg-dagger: Interactive imitation learning with human experts," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8077–8083.

[38] M. Zhou, J. Luo, J. Villella, Y. Yang, D. Rusu, J. Miao, W. Zhang, M. Alban, I. Fadakar, Z. Chen *et al.*, "Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving," *arXiv preprint arXiv:2010.09776*, 2020.

[39] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *International conference on machine learning*. PMLR, 2016, pp. 1995–2003.