Beyond the Map: Learning to Navigate Unseen Urban Dynamics Using Diffusion-Guided Deep Reinforcement Learning

Monu Nagar , Debasis Das
Indian Institute of Technology Jodhpur, India
{nagar.3, debasis}@iitj.ac.in

Abstract

Vision-based motion planning is a crucial task in Autonomous Driving (AD). Recent advancements in urban AD show that integrating Imitation Learning (IL) with Deep Reinforcement Learning (DRL) improves decision-making to be more like humans. However, IL methods depend on expert demonstrations to learn the optimal policy. The main drawback of this approach is the assumption that expert demonstrations are always optimal, which is not always true in real-world settings. This creates challenges in adapting to diverse weather conditions and dynamic traffic scenarios, often resulting in higher collision rates and increased risks to pedestrian safety. To address these challenges, we propose a **D**iffusion-**G**uided **D**eep **R**einforcement **L**earning (DGDRL) framework that integrates a diffusion model with a Soft Actor-Critic DRL method to effectively mitigate environmental uncertainties and enable self-learning beyond the training maps for new tasks. This framework follows a novel modified partially observable Markov decision process (mPOMDP) to choose optimal action from original and diffusion-generated observations, ensuring that the policy behavior remains consistent with the current action. We use the CARLA NoCrash benchmark to train and evaluate the proposed framework. The method is validated in diverse urban environments (e.g., empty, regular, and dense) across multiple towns. Additionally, we compare our model against state-of-the-art techniques to ensure robustness and generalizability to new environments. The project page and code are available at link.[1]

1 Introduction

In recent years, Autonomous Driving (AD) has gained significant research attention due to advancements in Imitation Learning (IL) and Deep Reinforcement Learning (DRL) techniques. AD systems require comprehensive understanding of the driving environment to ensure safety and effective generalization for deployment in real-world settings. IL uses supervised learning to make a driving policy based on expert datasets. [Shafiullah *et al.*, 2022; Wang *et al.*, 2023; Jain and Unhelkar, 2024]. The goal of IL is to train an agent to replicate the behavior of human experts as closely as possible. However, IL algorithms have the following limitations: 1) The driving policy is inherently restricted to the expert's performance and it is impractical to gather expert data for every conceivable driving situation, and 2) IL algorithms often experience a distribution mismatch between training data and real-world scenarios, as they rarely encounter failure situations during training. This lack of exposure prevents the system from learning how to manage such situations effectively.

To address the limitations of IL, researchers have shifted their focus to DRL techniques to develop driving policies through direct interaction with the environment. These techniques utilize reward mechanisms to evaluate optimal actions in various environmental states, such as dense traffic and dynamic weather conditions (e.g., rain, fog, and cloudy) [Ahmed *et al.*, 2021; Buddareddygari *et al.*, 2022; Wu *et al.*, 2022; Zhang *et al.*, 2023b; Coelho *et al.*, 2023; Chowdhury *et al.*, 2024]. The main goal of DRL techniques is to maximize expected cumulative rewards. Since the agent learns by interacting with the environment, DRL alleviates distribution mismatch issues and is not constrained by the performance of any expert. However, DRL suffers from low sample efficiency and requires significantly more data than IL to achieve convergence due to the extensive exploration needed during training to understand and navigate the environment completely. In advanced methodologies, researchers are focusing on integrating data augmentation, IL, and DRL techniques to develop robust and efficient DRL systems capable of effectively managing the complexities of real-world driving environments [Yuan *et al.*, 2022b; Liu *et al.*, 2024; Wang *et al.*, 2024b; Hansen and Wang, 2021]. Despite recent progress, the widespread deployment of Autonomous Vehicles (AVs) on public roads remains a challenge. The main difficulties stem from dense urban traffic environments characterized by numerous dynamic entities, such as cars, bicycles, and pedestrians, as well as complex road layouts and intricate user interactions. Existing solutions sometimes make incorrect decisions in these complex and dynamic settings, potentially leading to serious accidents or traffic violations.

[1]Project Page: https://autovisionproject.github.io/project/

$O_1\,(t = 1s)$ $O_2\,(t = 3s)$ $O_3\,(t = 5s)$ $O_4\,(t = 7s)$

Figure 1: Sequential observations at different intervals show minimal changes in subsequent scenes and nearly identical actions between closely spaced frames. This indicates that minor variations in observations do not significantly influence vehicle behavior. Such stability is crucial, as it provides a robust foundation for developing a generalized solution capable of effectively addressing environmental uncertainties.

A survey report by the AAA Newsroom highlights the significant challenge of building public trust in autonomous vehicles, particularly due to recent incidents that have heightened safety concerns among the public. The statistics indicate potential risks for the future of the autonomous industry. Therefore, this research aims to develop a generalized solution for handling environmental uncertainties in AD systems, primarily focusing on enhancing their performance in dynamic environments. We utilize the CARLA simulation environment to emulate real-world conditions such as varied towns, weather, and highway scenarios. To advance the performance of AVs, we propose the following contributions:

- We propose the Diffusion-Guided Deep Reinforcement Learning method to enhance generalization in the occurrence of unseen events and reduce dependency on imitation learning by integrating a diffusion model with the Soft Actor-Critic (SAC) method.
- We introduce a novel modified Partially Observable Markov Decision Process (mPOMDP) to optimize the behavior of the SAC policy for managing uncertainty in the dynamic environment. This design enhances the policy network's ability to make optimal decisions in closely related states.
- We conducted extensive testing to evaluate the effectiveness of the proposed method in various CARLA environments. Our approach outperformed the benchmarks in NoCrash scenarios (Empty, Regular, Dense) in both the Towns (Town01 and Town02), resulting in significant improvements in pedestrian safety. The pedestrian impact was reduced to 0.01/km, and the task completion success rate increased to 95%.

2 Related Work

This section presents a thorough review of existing literature on IL and explores methods to improve generalization through DRL techniques.

2.1 Imitation Learning in Driving

The effectiveness of IL has been widely acknowledged in AD, enabling the development of driving policies applicable in simulated and real-world urban driving scenarios. Learning from All Vehicles (LAV) [Chen and Krähenbühl, 2022] enhances sample efficiency by incorporating behaviors from all vehicles in the scene. Similarly, Learning by Cheating (LBC) [Chen *et al.*, 2020a] and Roach [Zhang *et al.*, 2021] adopt a RL coach, trained on-policy with privileged information, to guide the learning process. The CaT (Coaching a Teachable student) framework facilitates efficient knowledge transfer from a privileged teacher to a sensorimotor student, utilizing ResNet architectures for effective learning [Zhang *et al.*, 2023a]. The Behavior-Aware Trajectory (BAT) prediction model [Liao *et al.*, 2024b] employs LSTM and attention-based techniques to manage prediction uncertainties and understand interactions. It mainly addresses issues with non-continuous behavior labeling. The RLfOLD [Coelho *et al.*, 2024] integrates IL with DRL by leveraging online demonstrations to bridge the distribution gap between demonstration and training environments. Furthermore, researchers have also focused on integrating diffusion models with IL. The Diffusion Q [Wang *et al.*, 2022] conceptualizes the diffusion model as a policy, incorporating regularization and maximizing the action value function to optimize action selection. Diffusion-BC [Pearce *et al.*, 2023] improves traditional behavior cloning by employing the diffusion model as a policy, addressing expressiveness limitations. Additionally, CDSTraj [Liao *et al.*, 2024a] combines the Diffusion Module and the Spatio-Temporal Interaction Module to predict future traffic scenarios, enabling safe and efficient navigation in dynamic environments. DiffAIL [Wang *et al.*, 2024a] enhances adversarial IL by integrating the diffusion model, strengthening the discriminator's ability to capture and represent complex distributions effectively. However, despite the advancements in IL approaches, a significant challenge persists in addressing the distribution gap between demonstration datasets, and few works have explored the effectiveness of IL models for generalization. This gap remains a key challenge to the effective deployment of AD systems in real-world scenarios and highlights the need for further research.

2.2 Enhancing Generalization using DRL

Researchers have explored various approaches to enhance generalization capabilities in visually diverse environments. Data Augmentation (DA) and Domain Randomization (DR) have proven two effective methods for this purpose. [Hansen and Wang, 2021] employed a BYOL-like [Grill *et al.*, 2020] architecture to separate augmentation from policy learning, enhancing generalization. To manage high variance during DA, DrQ [Yarats *et al.*, 2020] updated the temporal difference

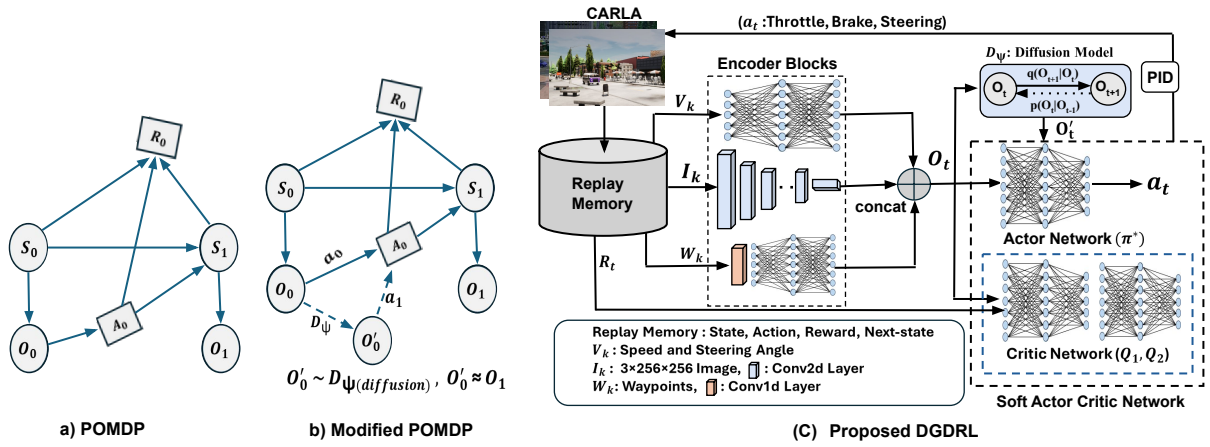

Figure 2: Comparison of (a) the standard POMDP, (b) the mPOMDP, and (c) the proposed DGDRL framework, highlighting key architectural enhancements. The DGDRL model addresses uncertainty in stochastic environments by leveraging three encoder blocks to extract features O_t. These features are processed by a diffusion model and the Actor (Policy) network to generate perturbed features O'_t and optimal actions a_t. The framework minimizes the distribution mismatch between O_t and O'_t to jointly optimize the Actor and Critic networks.

loss using augmented Q-values and target Q-values. SVEA [Hansen *et al.*, 2021] identified two pitfalls causing instability in DrQ and proposed augmenting Q-values only during training. [Yuan *et al.*, 2022a] proposed a task-aware data augmentation method using the Lipschitz constant to maintain training stability. Additionally, [Hafner *et al.*, 2023] worked on improving sample efficiency for learning from pixel-based observations. [Ma *et al.*, 2024] proposed Random PadResize (Rand PR) and Cycling Augmentation (CycAug) to enhance the efficacy of DA by improving spatial diversity and maintaining data distribution consistency. The most popular approach in domain randomization involves training a DRL agent on numerous source domains to develop a generalized policy, as demonstrated by Andrychowicz et al. [Andrychowicz *et al.*, 2020]. This method helps the agent to ignore irrelevant variations and focus on common features. [Lee *et al.*, 2024] introduced a spatial domain adaptation algorithm that enhances images through random transformations such as flipping, cropping, and rotating. These techniques use deterministic transformations, which sometimes fail to capture the continuous nature of state transitions and unexpected events in a dynamic environment. The proposed method as depicted in Figure 2c, learns smooth interpolation between states and allows the model to develop more robust representations of the states.

3 Methodology

In this section, we introduce the modified Partially Observable Markov Decision Process (mPOMDP) method for improving generalization. Subsequently, we detail the integration of the diffusion model with the Soft Actor-Critic Deep Reinforcement Learning framework.

3.1 Problem Definition

Autonomous vehicle (AV) frequently encounters challenges in interpreting scenes from images due to the dynamic nature of the environment (e.g., rapid movements of vehicles and pedestrians) and its inherent uncertainties. Consequently, the sequential decision-making process for AVs can be modeled as a mPOMDP, characterized by the tuple (S, O, A, r, P, γ). Here, $S = \{s_1, s_2, \ldots, s_{N_S}\}$ represents the finite set of possible states of the vehicle and its environment, $O = \{o_1, o_2, \ldots, o_{N_O}\}$ denotes the finite set of possible observations derived from sensor data, where o_2 is derived from the diffusion generated latent representation of o_1 and the initial original latent representation (o_1), and $A = \{a_1, a_2, \ldots, a_{N_A}\}$ is the finite set of actions the vehicle can take. The reward function $r : S \times A \to \mathbb{R}$ assigns rewards based on actions taken in specific states, such that $r(s, a) \in [0, R_{\max}]$, where $R_{\max} \in \mathbb{R}^+$ is the maximum possible reward. The state transition function $P : S \times A \times S \to [0, 1]$ gives the probability $P(s_{t+1} \mid s_t, a_t)$ of transitioning from state s_t to state s_{t+1} after taking action a_t. The observation function $O : S \times O \to [0, 1]$ defines the probability $O(o_t \mid s_t)$ of making observation o_t given the current state s_t. Finally, $\gamma \in [0, 1]$ is the discount factor that places higher importance on immediate rewards over future rewards. Our main objective is to determine an optimal policy π^* that maximizes the expected cumulative reward over time. Mathematically, this can be defined as:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{a_t \sim \pi(\cdot|s_t), s_t \sim P} \left[\sum_{t=1}^{T} \gamma^t r(s_t, a_t) \right] \quad (1)$$

where the actions a_t are sampled from the policy $\pi(\cdot \mid s_t)$ and the states s_t evolve according to the transition probabilities $P(s_{t+1} \mid s_t, a_t)$, starting from an initial state $s_0 \in S$. The optimal policy π^* specifies the best action a_t to take in each state s_t to maximize the expected discounted sum of rewards over a time horizon T.

Use case: Making complex sequential decisions in a non-deterministic, stochastic environment is a key challenge in autonomous driving (AD). Traditional methods often approach this task using Markov Decision Processes (MDPs) or POMDPs. These methods typically assume that the next state is fully determined by the current state and the action taken, as shown in Figure 2a. However, this assumption

can be inadequate in AD, where sudden or significant state transitions may lead to unsafe conditions, such as collisions. Real-world driving dynamics require state transitions to be smooth and constrained to ensure safety and stability, as represented in Figure 1. To address these challenges, the proposed mPOMDP framework allows the vehicle to make optimal decisions by analyzing original and diffusion-generated observations as illustrated in Figure 2b. These generated observations enable the model to learn a more robust, generalized policy by exposing it to uncertainty. The primary objective of integrating the diffusion model within the mPOMDP is to create anticipated future states with controlled uncertainty.

3.2 Encoder block

The AV decision-making process involves interpreting observations o_t and selecting actions a_t to maximize cumulative rewards over time. Each observation o_t at timestep t is defined as: $o_t = \{(\mathbf{I}_{t-k}, \mathbf{W}_{t-k}, \mathbf{V}_{t-k})\}_{k=0}^{1}$, where $\mathbf{I}_{t-k} \in \mathbb{R}^{3\times256\times256}$ represents the image from the vehicle's camera, $\mathbf{W}_{t-k} \in \mathbb{R}^{N\times2}$ denotes the 2D coordinates of $N = 10$ future waypoints, and $\mathbf{V}_{t-k} \in \mathbb{R}^2$ includes the vehicle's speed and steering angle. The agent computes an action $a_t = (\text{throttle}, \text{brake}, \text{steering})$, where throttle, brake $\in [0, 1]$ and steering $\in [-1, 1]$, based on these observations. The environment provides a reward r_t and the next observation o_{t+1}. The throttle and brake values are determined by a Proportional-Integral-Derivative (PID) controller that aligns with the predicted target speed from the policy network. To handle partial observability, we transform observations into latent representations. We used a similar setup for the encoder configurations based on the approach defined in RLfOLD [Coelho *et al.*, 2024]. The image encoder f_{img} processes the image sequence to generate latent features $\mathbf{i}_t$. This encoder uses convolutional layers enhanced with Adaptive Local Signal Mixing (A-LIX) and data augmentations such as color jittering and Gaussian blur to improve feature extraction and reduce overfitting. The waypoint encoder f_{wp} applies 1D convolutions to the waypoint coordinates $\mathbf{W}_t$, producing a latent representation. The vehicle measurements encoder f_{veh} uses a Multi-Layer Perceptron (MLP) to process the concatenated vehicle data. The overall latent state $\mathbf{o}_t$ is formed by concatenating these individual latent features: $o_t = (I_h, W_h, V_h)_t$, where $o_t \in \mathbb{R}^{d_{\text{total}}}$, with d_{total} representing the combined dimensionality of the latent features from the image, waypoint, and vehicle measurement encoders. These latent representation $\mathbf{o}_t$ encapsulates the critical information from the environment and vehicle state, allowing the policy network to take optimal actions.

3.3 Diffusion for Latent Space Uncertainty

Diffusion models are generative methods that iteratively transform Gaussian noise into a desired target distribution, potentially incorporating contextual information as needed. We modified the output layer of the DiffAIL [Wang *et al.*, 2024a] model to process encoded latent representations and generate representations that are closely aligned with some uncertainty. For simplicity, define $o_t = (I_{hi}, W_{hi}, V_{hi})_t$, where t denotes the time step in the diffusion process, and i indicates the agent environment steps at which a latent representation is used for action in the trajectory. The forward diffusion process is parameter-free and is given by:

$$q(o_t \mid o_{t-1}) = \mathcal{N}\left(o_t \mid \sqrt{1-\gamma_t} \cdot o_{t-1}, \gamma_t I\right) \tag{2}$$

where, γ_t represents the variance at time step t. The reverse process of the parameterized diffusion model is defined as:

$$p_\psi(o_{0:T} \mid o_t) = \mathcal{N}\left(o_{t-1} \mid \mu_\psi(o_t, t), \Sigma_\psi(o_t, t)\right). \tag{3}$$

The covariance matrix is fixed as $\Sigma_\psi(o_t, t) = \sigma_t^2 I = \gamma_t I$, making it non-trainable and represented by a predefined schedule. According to the forward process, o_t can be derived from o_{t-1} at any time step t. The mean of the reverse process is expressed as:

$$\mu_\psi(o_t, t) = \frac{1}{\sqrt{1-\gamma_t}}\left(o_t - \frac{\gamma_t}{\sqrt{1-\delta_t}}\epsilon_\psi(o_t, t)\right) \tag{4}$$

where, $\delta_t = \prod_{s=1}^{t}(1-\gamma_s)$ and $\epsilon_\psi(o_t, t)$ represents the predicted noise at time step t. The diffusion model over state-action pairs is formulated to predict the noise at each time step in the reverse diffusion process:

$$L(\psi) = \mathbb{E}_{o'_t \sim o_t}\left[\left\|\epsilon - \epsilon_\psi\left(\sqrt{\delta_t} \cdot o_t + \sqrt{1-\delta_t} \cdot \epsilon, t\right)\right\|^2\right] \tag{5}$$

Equation (5), is used to train the diffusion model to align with o_t. We incorporate the predicted noise into the latent features to enhance the policy network's ability to generalize across different scenarios. The term $(\sqrt{\delta_t} \cdot o_t + \sqrt{1-\delta_t} \cdot \epsilon)$ represents the noisy latent state generated by the forward pass, referred to as o_t^{pred}. In the reverse pass, the diffusion model learns the added noise, denoted as ϵ_{pred}. Therefore, we can obtain o'_t by subtracting the learned noise ϵ_{pred} from o_t^{pred}. Once the diffusion model is end-to-end trained, o'_t will converge to o_t. Therefore, o'_t can be mathematically represented as:

$$\mathbf{o'_t} = \frac{\mathbf{o_t^{\text{pred}}} - \sqrt{1-\delta_\mathbf{t}} \cdot \epsilon_{\text{pred}}}{\sqrt{\delta_\mathbf{t}}} \tag{6}$$

where, o'_t represents the adjusted latent feature, δ_t is the parameter of the diffusion process, and ϵ_{pred} is the predicted noise. Eq. (6), then used to guide the policy network in selecting new actions based on this input, effectively reducing the distribution gap between the actions taken by o_t and o'_t. This process is detailed in the following subsection.

3.4 Diffusion-Guided Soft Actor-Critic Method

Soft Actor-Critic [Haarnoja *et al.*, 2018] method is a model-free, off-policy actor-critic DRL algorithm designed to optimize a maximum-entropy objective with a discount factor γ. It learns two Q-functions Q_{θ_1} and Q_{θ_2}, a stochastic policy π_ϕ, and a temperature parameter α. Our approach bypasses the use of a value network and instead trains two Q-networks by optimizing them based on the one-step soft Bellman residual:

$$L_{\theta_k, i, w, v} = \mathbb{E}_{o_t, a_t, o_{t+1} \sim D}\left[\left(Q_{\theta_k}(o_t, a_t) - \hat{y}\right)^2\right] \tag{7}$$

Table 1: Comparison of success rates (%) on the CARLA NoCrash benchmark using current state-of-the-art methods. The evaluation covers two towns, T01 (Town01) and T02 (Town02), and two sets of weather conditions: W01 (ClearNoon, WetNoon, HardRainNoon, ClearSunset) and W02 (SoftRainSunset, WetSunset).

Weather	Town	Task	CILRS	LBC	CADRE	GRIAD	WOR	RLfOLD	OURS
Train (W01)	Train (T01)	Empty	97.23	89.11	95.32	97.15	97.43	**100**	**100**
		Regular	83.54	87.34	92.21	**98.43**	97.32	94.12	96.43
		Dense	42.65	75.36	80.04	94.54	92.46	90.42	**95.52**
Train (W01)	Test (T02)	Empty	66.44	86.56	92.21	94.32	94.54	**100**	**100**
		Regular	49.65	79.34	78.23	93.54	89.34	92.34	**98.77**
		Dense	23.23	53.65	61.32	78.22	74.76	82.26	**89.97**
Test (W02)	Train (T01)	Empty	96.46	60.72	94.32	83.97	90.23	96.23	**97.90**
		Regular	77.65	60.67	86.23	87.34	90.65	88.02	**95.25**
		Dense	39.56	54.46	76.33	83.74	84.23	85.42	**92.23**
Test (W02)	Test (T02)	Empty	66.56	36.32	78.54	69.12	78.31	98.47	**100**
		Regular	56.35	36.76	72.76	63.23	82.76	86.99	**92.32**
		Dense	24.43	12.67	52.23	52.56	66.34	68.62	**82.12**
Average			60.31	61.08	81.14	83.01	86.53	90.07	**95.04**

$$\hat{y} = r_t + \gamma \left[\min_{k=1,2} \bar{Q}_{\theta_k}(o_{t+1}, \tilde{a}_{t+1}) - \alpha \log \pi_\phi(\tilde{a}_{t+1}|o_{t+1}) \right] \tag{8}$$

The policy loss function for the SAC method is enhanced by integrating diffusion-based latent features from Eq. (6), aiming to increase the expected return and foster exploration via entropy regularization. Therefore, the policy loss function can be defined as:

$$\begin{aligned} L_\phi^\pi = -\mathbb{E}_{o_t \sim D} \Bigg[& \mathbb{E}_{\tilde{a}_t \sim \pi_\phi(\cdot|o_t)} \Bigg(\min_{k=1,2} Q_{\theta_k}(o_t, \tilde{a}_t) \\ & - \alpha \log \pi_\phi(\tilde{a}_t|o_t) \Bigg) \Bigg] \\ & + \boldsymbol{\lambda} \mathbb{E}_{\boldsymbol{o_t' \sim \text{Diffusion}(o_t)}} \left[\mathbf{D}_{\text{KL}} \left(\boldsymbol{\pi_\phi(\cdot|o_t) \| \pi_\phi(\cdot|o_t')} \right) \right] \end{aligned} \tag{9}$$

The policy loss function L_ϕ^π in (9), combines the expected return, which drives the policy network to maximize rewards using KL divergence, weighted by λ, to regularize the policy's distribution. For Gaussian policies, the KL divergence between the two distributions can be modeled as $\pi_\phi(\cdot|o_t) \sim \mathcal{N}(\mu_{o_t}, \sigma_{o_t}^2)$ and $\pi_\phi(\cdot|o_t') \sim \mathcal{N}(\mu_{o_t'}, \sigma_{o_t'}^2)$. The higher KL divergence value helps the agent stay consistent with previously learned policies, particularly in new or uncertain states. As the diffusion model continues to learn over time, the gradient of the KL divergence with respect to the policy parameters ($\frac{\partial D_{\text{KL}}}{\partial \phi}$), progressively decrease to zero. Therefore, we can draw the following observations: $(\mu_{o_t} - \mu_{o_t'}) \to 0$ and $\sigma_{o_t} \approx \sigma_{o_t'}$. Hence, we can formulate:

$$\lim_{t \to \text{last episode}} D_{\text{KL}} \left(\pi_\phi(\cdot \mid o_t) \parallel \pi_\phi(\cdot \mid o_t') \right) \approx 0 \tag{10}$$

Thus, the gradient of the policy loss with respect to ϕ simplifies, as the KL divergence term no longer majorly contributes to the gradient. Hence, the policy optimization reduces to:

$$\frac{\partial L_\phi^\pi}{\partial \phi} = \frac{\partial L_{\text{SAC}}}{\partial \phi} + \frac{\partial D_{\text{KL}}}{\partial \phi} \approx \frac{\partial L_{\text{SAC}}}{\partial \phi} \tag{11}$$

Reducing KL divergence weakens the penalty on policy differences between states, allowing the policy to focus more on optimizing actions based on Q-value differences rather than maintaining consistency. This smooth transition from high to low regularization reflects the policy's shift from broad exploration to specialization, initially exploring and generalizing across states and later focusing on exploiting optimal actions as confidence grows. The importance of integrating the diffusion model with SAC method is explicitly delineated in **supplementary file: Section 1**.

4 Experiments and Result Discussion

We trained and evaluated our model using an NVIDIA L40S GPU with 46GB of RAM in the *endless-full-nocrash* Carla environment. The training process took 6 days and 19 hours to complete 10^6 environment steps. Evaluations were conducted every 20k steps, with episode returns averaged over 10 episodes in each assessment. We utilized the reward function mechanism described in [Zhang *et al.*, 2021]. In this study, we utilized CARLA [Dosovitskiy *et al.*, 2017] version 0.9.10, an open-source simulator designed for AD research. CARLA features high-fidelity maps with both static objects, such as buildings, traffic signs, and dynamic objects. In vehicle avoidance scenarios, the ego vehicle must stop when a stationary vehicle appears 20 meters ahead and proceed once cleared. The explicit details of the hyperparameter configurations used during model training are provided in the **supplementary file**.

4.1 Evaluation Metrics and NoCrash Benchmark

To evaluate the generalization ability of the proposed model, we utilized the NoCrash benchmark [Codevilla *et al.*, 2019a]. Which provides three traffic conditions of varying difficulty: empty (no dynamic objects), regular (moderate numbers of pedestrians and vehicles), and dense (high volume of pedestrians and vehicles). Additionally, it defines six weather conditions and features 25 routes in Town01 (characterized by one-lane roads and T-junctions) for training, while Town02, a scaled-down version with distinct textures, is used for testing. The evaluation focused on the success rate, indicating the proportion of routes completed without collisions.

4.2 Model Comparison

We compare the proposed method with recent state-of-the-art (SOTA) approaches, including CILRS [Codevilla *et al.*,

Figure 3: Generalization to unseen tasks: Despite being trained in Town01 (two-lane setup), the model autonomously learns and successfully executes smooth lane-switching tasks on a four-lane road in a completely new town setup without any collisions. This underscores the model's capability to learn and master tasks beyond its training scope through self-learning. (Refer Supplementary File: **Lane-switching.mp4**)

Figure 4: Analysis of the model's generalization capability to ensure safety measures during unexpected events in the Town 2 dense setup under test weather conditions. (Refer supplementary file: **Safety01.mp4, Safety02.mp4**)

2019b], LBC [Chen *et al.*, 2020b], WOR [Chen *et al.*, 2021], GRIAD [Chekroun *et al.*, 2023], CADRE [Zhao *et al.*, 2022], and RLFOLD [Coelho *et al.*, 2024]. Table 1, presents a comprehensive comparison of success rates on the NoCrash benchmark across these methods. LBC, CADRE, and RLFOLD utilize single-camera approaches, while GRIAD and WOR use three cameras as input. Most of these methods rely on IL in offline or online demonstration mode. A major drawback of IL is its limited generalization to unseen scenarios. IL methods mainly rely on the quality of demonstration data, which can result in unsafe or suboptimal decisions when faced with rare or unpredictable events in dense traffic scenarios. The proposed method efficiently handled this problem and choose the optimal decision close to the previous decision instead of following predefined actions. The DGDRL demonstrates remarkable efficacy across diverse traffic conditions. In the Train(W01)-Train(T01) scenario, DGDRL achieves a success rate of 95.52% in the dense task, outperforming other methods. However, it performs relatively lower in the Regular task, where GRIAD achieves a higher success rate of 98.43%. In the Test(W02)-Test(T02) scenario, where other methods struggle to handle dense traffic conditions effectively, DGDRL continues to demonstrate outstanding performance. In dense traffic scenarios, it achieves an 82.12% success rate, representing a substantial 24% improvement over RLFOLD. Similarly, in regular task conditions, DGDRL attains a 92% success rate, surpassing RLFOLD by 6%. Notably, the method maintains an impressive average success rate of 95.04%, underscoring its robust performance and superiority compared to existing state-of-the-art approaches.

Handling Unseen Tasks and Unexpected Events: Scene understanding is a critical task for AVs operating in highly dynamic environments, where even a minor deviation in perception and interpretation of surroundings can lead to significant safety challenges for pedestrians. Figures 4 and 3 present experimental results evaluating the generalization capability of the proposed method during unexpected events and its ability to handle unseen tasks. The results show that the vehicle successfully detects pedestrians and executes appropriate safety protocols by reducing speed to a complete stop, ensuring safe road crossing for pedestrians with zero collision incidents. The vehicle then resumes its planned trajectory once the road is clear. Furthermore, we critically analyzed the captured testing videos and observed that the model, despite not being explicitly trained for lane-switching, executes smooth lane-switching tasks. These findings demonstrate the vehicle's reliability in handling unexpected events and highlight its potential for real-world AV applications. Additionally, we investigated the root cause of the 5% gap in the success rate from the perfect score. Our analysis revealed that the behavior of the vehicle is inconsistent when encountering traffic lights, as shown in Figure 5. The agent stops when it detects

Table 2: The ablation study presents task success rates and safety metrics under varying values of λ, highlighting the performance of the regular task under test conditions. Results are based on four runs (20 episodes each), presented as $mean \pm std$.

Metric	RL baseline	(λ=0.01)	(λ=0.1)	(λ=0.5)
Success Rate (%, ↑)	86 ± 2	85 ± 4	92 ± 2	89 ± 3
Trajectory Completion Rate (%, ↑)	94 ± 2	95 ± 1	98 ± 1	96 ± 2
Pedestrian Impact (#/Km, ↓)	0.10 ± 0.03	0.08 ± 0.04	**0.01 ± 0.02**	0.05 ± 0.03
Vehicle Collision (#/Km, ↓)	0.32 ± 0.11	0.30 ± 0.10	0.16 ± 0.06	0.22 ± 0.08
Layout Collision (#/Km, ↓)	0.15 ± 0.04	0.21 ± 0.03	0.09 ± 0.04	0.11 ± 0.05
Agent Obstruction (#/Km, ↓)	0.21 ± 0.13	0.18 ± 0.11	0.14 ± 0.08	0.16 ± 0.10

Figure 5: Model behavior in handling traffic lights in the CARLA NoCrash Regular setup under new weather conditions, observed at different timestamps ($t_1 \rightarrow t_4$) in a continuous sequence. (Refer supplementary file: **Success.mp4, Failure.mp4**).

another stopped vehicle in the lane. However, the agent sometimes passes through a red light when the road is clear. This behavior highlights the limitations of the proposed method and requires further improvements.

5 Ablation Study

In the ablation study, we evaluate the impact of varying the diffusion regularization parameter (λ), given in Eq. (9), on task completion and safety metrics. As shown in Table 2, increasing λ leads to a substantial improvement in the task success rate. Specifically, the success rate rises from 86% for the RL baseline to 92% when λ is set to 0.1. This improvement is due to the additional regularization term in the loss function, which helps the policy network generate more robust actions to state perturbations. We observed a significant reduction in pedestrian impacts and vehicle collisions, with values of 0.01 and 0.14 per km, respectively, when λ was set to 0.1. This indicates that the model generalizes effectively to unseen events, such as sudden pedestrian movements. However, when λ exceeded 0.1, the diffusion loss became dominant over the policy loss, causing the model to deviate from optimal sequential decisions. In real-world driving scenarios, state transitions are continuous in nature. For instance, when moving from sunny to rainy weather, the change occurs gradually: sunny change to cloudy, followed by light rain and heavy rain. Considering these factors, we utilized linear noise scheduling, which produced state-of-the-art results. We also tested exponential noise scheduling, which led to aggressive exploration and caused learning instability. We set the timesteps (t) to 1000 and observed that reducing the t to 600 made the model less generalized while increasing it to 1200 resulted in no further improvement in results. Additionally, we conducted a series of testing experiments using Town03, Town04, and Town05 to evaluate the generalization capability of the proposed model to unseen dynamic environments, as detailed in the **supplementary file**.

6 Conclusion

This work introduces the Diffusion-Guided Deep Reinforcement Learning (DGDRL) framework, which integrates a Diffusion model with a Soft Actor-Critic DRL approach to enhance autonomous driving performance in complex urban environments. The framework follows a novel modified Partially Observable Markov Decision Process (mPOMDP) principle. It effectively addresses the limitations of traditional Imitation Learning by incorporating a diffusion process, which improves the model's ability to handle environmental uncertainties and unforeseen events. Empirical results demonstrate DGDRL's significant superiority over existing state-of-the-art methods, achieving an average success rate of 95.04% across diverse towns and weather conditions. The ablation study reveals that increasing the diffusion regularization parameter λ to 0.1 shows substantial improvements in task completion rates and reduces collision rates, underscoring the critical role

of the diffusion model in developing robust and generalized autonomous driving policies. In the future, we plan to focus on enhancing DGDRL's capabilities to manage complex tasks, such as handling traffic lights and tunnels, along with exploring its application in other domains.

References

[Ahmed *et al.*, 2021] Marwa Ahmed, Ahmed Abobakr, Chee Peng Lim, and Saeid Nahavandi. Policy-based reinforcement learning for training autonomous driving agents in urban areas with affordance learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):12562–12571, 2021.

[Andrychowicz *et al.*, 2020] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

[Buddareddygari *et al.*, 2022] Prasanth Buddareddygari, Travis Zhang, Yezhou Yang, and Yi Ren. Targeted attack on deep rl-based autonomous driving with learned visual patterns. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10571–10577. IEEE, 2022.

[Chekroun *et al.*, 2023] Raphael Chekroun, Marin Toromanoff, Sascha Hornauer, and Fabien Moutarde. Gri: General reinforced imitation and its application to vision-based autonomous driving. *Robotics*, 12(5):127, 2023.

[Chen and Krähenbühl, 2022] Dian Chen and Philipp Krähenbühl. Learning from all vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17222–17231, 2022.

[Chen *et al.*, 2020a] Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Conference on Robot Learning*, pages 66–75. PMLR, 2020.

[Chen *et al.*, 2020b] Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 66–75. PMLR, 30 Oct–01 Nov 2020.

[Chen *et al.*, 2021] Dian Chen, Vladlen Koltun, and Philipp Krähenbühl. Learning to drive from a world on rails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15590–15599, 2021.

[Chowdhury *et al.*, 2024] Jayabrata Chowdhury, Venkataramanan Shivaraman, Suresh Sundaram, and PB Sujit. Graph-based prediction and planning policy network (gp3net) for scalable self-driving in dynamic environments using deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11606–11614, 2024.

[Codevilla *et al.*, 2019a] Felipe Codevilla, Eder Santana, Antonio Lopez, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9328–9337, 2019.

[Codevilla *et al.*, 2019b] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9329–9338, 2019.

[Coelho *et al.*, 2023] Daniel Coelho, Miguel Oliveira, and Vitor Santos. Rlad: Reinforcement learning from pixels for autonomous driving in urban environments. *IEEE Transactions on Automation Science and Engineering*, 2023.

[Coelho *et al.*, 2024] Daniel Coelho, Miguel Oliveira, and Vitor Santos. Rlfold: Reinforcement learning from online demonstrations in urban autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11660–11668, 2024.

[Dosovitskiy *et al.*, 2017] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[Grill *et al.*, 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[Hafner *et al.*, 2023] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy P. Lillicrap. Mastering diverse domains through world models. *CoRR*, abs/2301.04104, 2023.

[Hansen and Wang, 2021] Nicklas Hansen and Xiaolong Wang. Generalization in reinforcement learning by soft data augmentation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13611–13617. IEEE, 2021.

[Hansen *et al.*, 2021] Nicklas Hansen, Hao Su, and Xiaolong Wang. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. *Advances in neural information processing systems*, 34:3680–3693, 2021.

[Jain and Unhelkar, 2024] Abhinav Jain and Vaibhav Unhelkar. Go-dice: Goal-conditioned option-aware offline imitation learning via stationary distribution correction estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 12763–12772, 2024.

[Lee *et al.*, 2024] Kyunghyun Lee, Ukcheol Shin, and Byeong-Uk Lee. Learning to control camera exposure via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2975–2983, 2024.

[Liao *et al.*, 2024a] Haicheng Liao, Xuelin Li, Yongkang Li, Hanlin Kong, Chengyue Wang, Bonan Wang, Yanchen Guan, KaHou Tam, and Zhenning Li. Cdstraj: characterized diffusion and spatial-temporal interaction network for trajectory prediction in autonomous driving. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 7331–7339, 2024.

[Liao *et al.*, 2024b] Haicheng Liao, Zhenning Li, Huanming Shen, Wenxuan Zeng, Dongping Liao, Guofa Li, and Chengzhong Xu. Bat: Behavior-aware human-like trajectory prediction for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10332–10340, 2024.

[Liu *et al.*, 2024] Haochen Liu, Zhiyu Huang, Xiaoyu Mo, and Chen Lv. Augmenting reinforcement learning with transformer-based scene representation learning for decision-making of autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2024.

[Ma *et al.*, 2024] Guozheng Ma, Linrui Zhang, Haoyu Wang, Lu Li, Zilin Wang, Zhen Wang, Li Shen, Xueqian Wang, and Dacheng Tao. Learning better with less: effective augmentation for sample-efficient visual reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[Pearce *et al.*, 2023] Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, et al. Imitating human behaviour with diffusion models. *arXiv preprint arXiv:2301.10677*, 2023.

[Shafiullah *et al.*, 2022] Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *Advances in neural information processing systems*, 35:22955–22968, 2022.

[Wang *et al.*, 2022] Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193*, 2022.

[Wang *et al.*, 2023] Huandong Wang, Changzheng Gao, Yuchen Wu, Depeng Jin, Lina Yao, and Yong Li. Pategail: a privacy-preserving mobility trajectory generator with imitation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14539–14547, 2023.

[Wang *et al.*, 2024a] Bingzheng Wang, Guoqiang Wu, Teng Pang, Yan Zhang, and Yilong Yin. Diffail: Diffusion adversarial imitation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15447–15455, 2024.

[Wang *et al.*, 2024b] Shuo Wang, Zhihao Wu, Xiaobo Hu, Jinwen Wang, Youfang Lin, and Kai Lv. What effects the generalization in visual reinforcement learning: policy consistency with truncated return prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 5590–5598, 2024.

[Wu *et al.*, 2022] Jingda Wu, Zhiyu Huang, and Chen Lv. Uncertainty-aware model-based reinforcement learning: Methodology and application in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 8(1):194–203, 2022.

[Yarats *et al.*, 2020] Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International conference on learning representations*, 2020.

[Yuan *et al.*, 2022a] Zhecheng Yuan, Guozheng Ma, Yao Mu, Bo Xia, Bo Yuan, Xueqian Wang, Ping Luo, and Huazhe Xu. Don't touch what matters: Task-aware lipschitz data augmentation for visual reinforcement learning. *arXiv preprint arXiv:2202.09982*, 2022.

[Yuan *et al.*, 2022b] Zhecheng Yuan, Zhengrong Xue, Bo Yuan, Xueqian Wang, Yi Wu, Yang Gao, and Huazhe Xu. Pre-trained image encoder for generalizable visual reinforcement learning. *Advances in Neural Information Processing Systems*, 35:13022–13037, 2022.

[Zhang *et al.*, 2021] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. End-to-end urban driving by imitating a reinforcement learning coach. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15222–15232, 2021.

[Zhang *et al.*, 2023a] Jimuyang Zhang, Zanming Huang, and Eshed Ohn-Bar. Coaching a teachable student. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7805–7815, 2023.

[Zhang *et al.*, 2023b] Linrui Zhang, Qin Zhang, Li Shen, Bo Yuan, Xueqian Wang, and Dacheng Tao. Evaluating model-free reinforcement learning toward safety-critical tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15313–15321, 2023.

[Zhao *et al.*, 2022] Yinuo Zhao, Kun Wu, Zhiyuan Xu, Zhengping Che, Qi Lu, Jian Tang, and Chi Harold Liu. Cadre: A cascade deep reinforcement learning framework for vision-based autonomous urban driving. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3481–3489, 2022.