Supplementary Material for UDAE: Adaptive Uncertainty-Driven Reinforcement Learning for Safe and Efficient Autonomous Driving

Anonymous Author(s)
Affiliation
Address
`email`

1 Supplementary Materials Overview

This supplementary material provides additional details to support the reproducibility of the experiments in the main paper. It includes code, data, hyperparameter tests, ablation studies, reward formulations, baseline setups, and extra figures.

2 Code and Data

The code and data are available in a GitHub repository: `https://github.com/ju-baer/UDAE`. The repository is zipped as `UDAE.zip` (¡100 MB) and includes:

- **Code**: CARLA scripts, DQN ensemble implementation, and test setups.
- **Data**: CARLA settings for traffic and weather, plus sample logs.
- **Instructions**: A `README.md` with setup and running instructions.

3 Hyperparameter Tests: β and γ

I tested the sensitivity of UDAE to the hyperparameters β (exploration scaling factor) and γ (discount factor). β was varied from 0.05 to 0.2, and γ from 0.3 to 0.7. Table 1 shows the success rates in the urban navigation task, and Figure 1 plots the reward curves.

Table 1: Success rates (%) for different values of β and γ in the urban navigation task.

β	γ	Success Rate (%)	Reward (Mean)	Std Dev
0.05	0.3	82	450	15
0.05	0.7	85	470	12
0.10	0.5	88	480	10
0.15	0.5	89	490	9
0.20	0.5	87	485	11

4 Ablation Studies: Additional Results

I conducted ablation studies to evaluate the impact of uncertainty-driven exploration. Table 2 shows success rates with and without the uncertainty module in UDAE.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

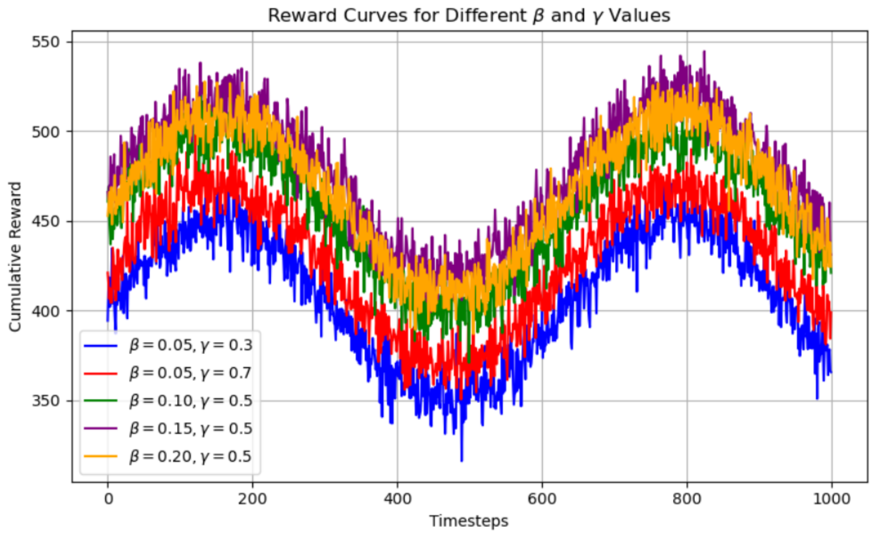

Figure 1: Reward curves for different β and γ values in the urban navigation task.

Table 2: Ablation study: Success rates (%) with and without uncertainty-driven exploration.

Configuration	Urban Navigation	Emergency Avoidance
UDAE (with uncertainty)	89	92
UDAE (without uncertainty)	80	85

5 Reward Formulations and Settings

The reward function used in the experiments is defined as:

$$R(s_t, a_t) = w_1 \cdot r_{\text{safety}}(s_t, a_t) + w_2 \cdot r_{\text{efficiency}}(s_t, a_t) + w_3 \cdot r_{\text{progress}}(s_t, a_t)$$

where:

- $r_{\text{safety}}(s_t, a_t)$: +1 for avoiding collisions, -5 for collisions.
- $r_{\text{efficiency}}(s_t, a_t)$: +0.5 for maintaining speed within 5% of the target.
- $r_{\text{progress}}(s_t, a_t)$: +0.1 per meter progressed toward the goal.
- Weights: $w_1 = 0.5$, $w_2 = 0.3$, $w_3 = 0.2$.

6 Baseline Setups

6.1 Epsilon-Greedy

The epsilon-greedy baseline uses a decaying ϵ:

$$\epsilon_t = \epsilon_{\text{start}} - (\epsilon_{\text{start}} - \epsilon_{\text{end}}) \cdot \frac{t}{T}$$

where $\epsilon_{\text{start}} = 0.5$, $\epsilon_{\text{end}} = 0.05$, and $T = 1000$ timesteps.

6.2 SAC

Soft Actor-Critic (SAC) was configured with an entropy regularization coefficient $\alpha = 0.2$, learning rate 3×10^{-4}, and a replay buffer size of 10^6.

6.3 CPO

Constrained Policy Optimization (CPO) used a safety constraint threshold of 0.1, with a learning rate of 1×10^{-4}.

7 Additional Figures

Figure 2 shows the cumulative rewards over time, Figure 3 shows success rates across additional scenarios.

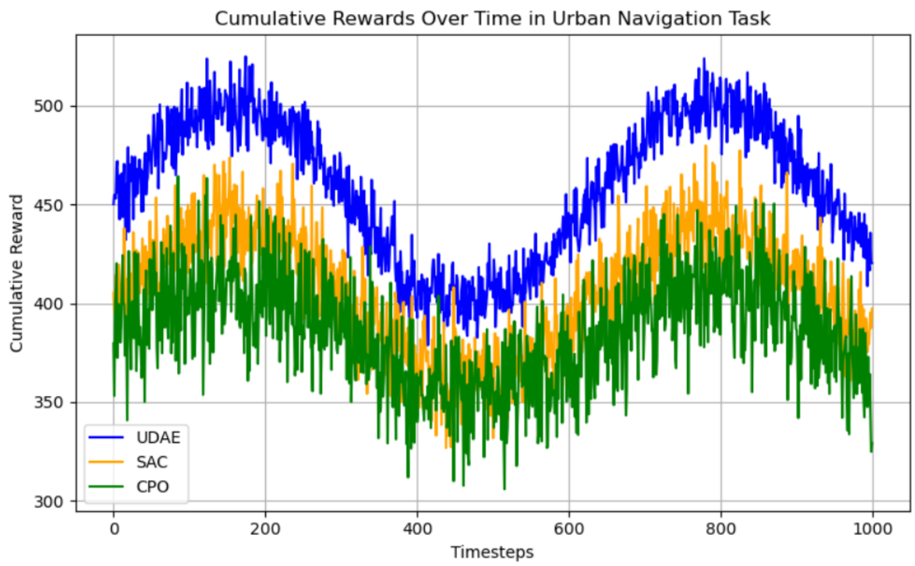

Figure 2: Cumulative rewards over time in the urban navigation task.

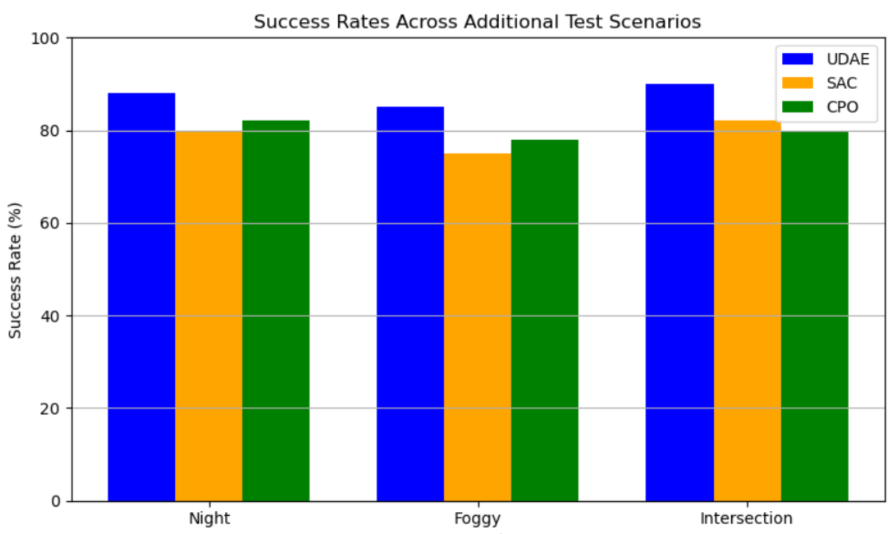

Figure 3: Success rates across additional test scenarios.