Prompt-Informed Reinforcement Learning for Visual Coverage Path Planning

Venkat Margapuri
vmargapu@villanova.edu

Department of Computing Sciences
Villanova University
Villanova, PA

Abstract

Visual coverage path planning with unmanned aerial vehicles (UAVs) requires agents to strategically coordinate UAV motion and camera control to maximize coverage, minimize redundancy, and maintain battery efficiency. Traditional reinforcement learning (RL) methods rely on environment-specific reward formulations that lack semantic adaptability. This study proposes Prompt-Informed Reinforcement Learning (PIRL), a novel approach that integrates the zero-shot reasoning ability and in-context learning capability of large language models with curiosity-driven RL. PIRL leverages semantic feedback from an LLM, GPT-3.5, to dynamically shape the reward function of the Proximal Policy Optimization (PPO) RL policy guiding the agent in position and camera adjustments for optimal visual coverage. The PIRL agent is trained using OpenAI Gym and evaluated in various environments. Furthermore, the sim-to-real-like ability and zero-shot generalization of the agent are tested by operating the agent in Webots simulator which introduces realistic physical dynamics. Results show that PIRL outperforms multiple learning-based baselines such as PPO with static rewards, PPO with exploratory weight initialization, imitation learning, and an LLM-only controller. Across different environments, PIRL outperforms the best-performing baseline by achieving up to 14% higher visual coverage in OpenAI Gym and 27% higher in Webots, up to 25% higher battery efficiency, and up to 18% lower redundancy, depending on the environment. The results highlight the effectiveness of LLM-guided reward shaping in complex spatial exploration tasks and suggest a promising direction for integrating natural language priors into RL for robotics.

1 Introduction

In recent times, large language models (LLMs) have advanced significantly in their ability to comprehend and generate human-like text. Notable models such as Generative Pretrained Transformer (GPT) [14], Bidirectional Encoder Representations from Transformers (BERT) [8], and Text-to-Text Transfer Transformers (T5) [19] have exhibited remarkable capabilities in tasks ranging from natural language understanding to sophisticated decision-making process. LLMs are widely adopted in different domains including code generation [27], customer chatbots [23], and reinforcement learning (RL) [11, 28] systems for their context-aware decision-making ability. Simultaneously, unmanned aerial vehicles (UAVs) have been widely adopted across different sectors such as agriculture [26], delivery services [2], and military [10]. UAVs are immensely useful for visual coverage path planning (VCPP) tasks where the goal is to cover large swathes of land for tasks such as field surveying, terrain mapping, and crop monitoring. These applications require sophisticated decision-making algorithms to optimize navigation and minimize energy consumption while ensuring comprehensive visual coverage.

The use of RL to train autonomous agents enables them to navigate complex environments by making data-driven decisions based on real-time environmental feedback. However, RL paradigms rely heavily on the reward function, whose curation is often non-trivial as it involves balancing competing objectives such as motion, energy usage, and task relevance. Suboptimal reward design can lead to reward hacking [18], where agents exploit spurious strategies that maximize rewards without accomplishing intended goals such as favoring rapid movement over comprehensive environmental exploration in VCPP tasks. Among RL approaches, Proximal Policy Optimization (PPO) [22] is prevalent due to its training stability and sample efficiency. However, PPO optimizes against an environment-driven reward formulation which encodes task objectives implicitly but lacks semantic adaptability, potentially limiting the policy's ability to generalize across tasks or environments. Hence, the design of a reward function that accounts for dynamic task constraints remains a bottleneck.

LLMs such as OpenAI GPT-3.5 possess zero-shot reasoning [15] and in-context learning abilities [9], generating task-relevant responses from structured prompts without requiring finetuning. While LLMs can produce semantically aligned actions, they lack environmental grounding leading to suboptimal performance. This work proposes Prompt-Informed Reinforcement Learning (PIRL), a hybrid framework that uses LLMs not as policy generators but as reward shapers to dynamically guide RL agents. PIRL combines the adaptability and generalization of LLMs with the optimization strength of RL, enabling agents to learn flexible behaviors aligned with high-level semantic intent. Unlike PPO, PIRL yields task-aware generalization through adaptive reward shaping, and unlike LLM-based planners, PIRL learns from feedback grounded in real-time environmental interaction. By shaping the reward rather than prescribing actions, PIRL enables better alignment with task goals while preserving policy flexibility and improvement without requiring explicit finetuning.

This research contributes to the growing body of work in the field of embodied vision using simulated active agents. By adjusting position and camera parameters, the agent (UAV) trained using PIRL performs VCPP by focusing on maximizing visual coverage while minimizing battery usage and redundant scanning. The key contributions of the work are:

1. The use of an LLM, OpenAI GPT-3.5, coupled with PPO RL technique to train a UAV for VCPP in 3D virtual environments using OpenAI Gym and Webots.

2. Introduction of PIRL, a novel reward shaping paradigm in which context-aware LLMs dynamically guide the reward function using structured zero-shot prompting, mitigating risks such as reward hacking.

3. Leveraging the zero-shot reasoning ability of LLMs to provide high-level guidance for RL without expert demonstrations enabling scalable policy learning with minimal supervision.

2 Related Work

RL for Visual Coverage Path Planning: Recent advancements in RL have significantly enhanced frameworks for VCPP by integrating camera aware and curiosity driven strategies.

Multiple studies [21, 29, 32] have integrated vision sensors for agricultural coverage path planning use cases such as fruit picking and terrain mapping. Another study [5] introduced a unified framework leveraging value-based RL techniques for zero-shot coverage path planning, emphasizing adaptability across varying map sizes and configurations. In the realm of curiosity driven approaches, [7, 30] integrated a curiosity mechanism to enhance the performance of PPO method in a realistic 3D virtual environment. Furthermore, [20] proposed a convergence-based approach and a guidance-based approach using a domain network to explore unexplored areas in a 3D environment while minimizing redundancy. The aforementioned research studies serve as the motivation behind the use of the curiosity driven PPO approach with structured zero-shot prompting to train the UAV for VCPP in 3D environments.

Using Language for Reward Shaping: The use of LLMs to map natural language instructions to rewards has been explored by prior research [12, 16, 24, 31]. [13] used LLMs for grounding high-level tasks expressed in natural language to a chosen set of actionable steps, demonstrating the idea of decomposing high-level tasks to mid-level plans. [16] integrated GPT-3 to train objective-aligned RL agents using few-shot prompting for the Ultimatum game and zero-shot prompting for Matrix games. The work by [31] demonstrated how LLMs can be used to interpret natural language instructions and generate corresponding reward functions using a reward translator. Furthermore, the rewards were converted to code to control robots. This study draws inspiration from the reward translator [31] in its implementation of Prompt-Adaptive Reward Engine (PARE), described in section 3.

Imitation Learning: Imitation learning helps RL agents learn from expert demonstrations of the task in a supervised learning setting. This technique relies on the availability of large labeled datasets that demonstrate the multitude of scenarios associated with learning a task. Imitation learning has been widely explored in related domains such as autonomous driving [4, 25], and path planning [6, 17]. This study uses imitation learning as one of the baseline approaches to compare the proposed PIRL approach.

3 Methodology

The VCPP task is modeled as a Markov Decision Process (MDP) given by the tuple $\mathcal{M} = (S, A, P, R, \gamma)$, where:

- $S \subseteq \mathbb{Z}^3 \times \Theta \times \mathbb{R}_{[0,1]}$ is the hybrid state space, where each state $s_t \in S$ includes the UAV's 3D position $\mathbf{p}_t = (x_t, y_t, z_t) \in \mathbb{Z}^3$, camera configuration $\theta_t = (\phi_t, \psi_t, \zeta_t) \in \Theta$ encoding tilt, pan, and zoom respectively, where $\Theta := \{(\phi, \psi, \zeta) \in \mathbb{Z} \times \mathbb{Z} \times \mathbb{Q} \mid 0 \le \phi \le 90,\ \phi \equiv 0 \pmod 5;\ -90 \le \psi \le 90, \psi \equiv 0 \pmod{15}, 0.5 \le \zeta \le 2.0, \zeta = 0.1k,\ k \in \mathbb{Z}\}$, and (normalized) battery level $b_t \in \mathbb{R}_{[0,1]}$, where $\mathbb{R}_{[0,1]} := \{b \in \mathbb{R} \mid 0 \le b \le 1\}$. The rationale for the design choice is provided in section A.1 of the Appendix.

- $A_{\text{move}} = (x^+, x^-, y^+, y^-, z^+, z^-)$ and $A_{\text{cam}} = (\text{tilt}^+, \text{tilt}^-, \text{pan}^+, \text{pan}^-, \text{zoom}^+, \text{zoom}^-)$ denote the ordered unit movement and camera control (fixed step sizes of 5°(tilt), 15°(pan), and 0.1(zoom)) actions, respectively. The discrete action space is defined as the ordered tuple $A = A_{\text{move}} \,||\, A_{\text{cam}}$, where $||$ denotes tuple concatenation.

- $P : S \times A \times S \to [0,1]$ models the probability $P(s_{t+1} \mid s_t, a_t)$ of transitioning to s_{t+1} after executing a_t in s_t, accounting for deterministic actions and stochastic perturbations like winds.

- $R : S \times A \rightarrow \mathbb{R}$ is the reward function composed of multiple components designed to encourage exploratory and efficient behavior. The total reward received at time $t+1$, resulting from taking action a_t in state s_t at time t, is defined as: $r_{t+1} = \lambda_c \cdot \Delta C_t + \lambda_r \cdot \mathbb{1}_{\text{redundant}}(s_{t+1}) + \lambda_b \cdot f_{\text{battery}}(b_{t+1}) + \lambda_{\text{cam}} \cdot f_{\text{cam}}(\theta_{t+1}) + \lambda_{\text{cur}} \cdot f_{\text{curiosity}}(s_{t+1}) + \lambda_{\text{collision}} \cdot \mathbb{1}_{\text{collision}}(s_{t+1}) + \lambda_{\text{idle}} \cdot \mathbb{1}_{\text{idle}}(s_{t+1}) + \lambda_{\text{LLM}} \cdot f_{\text{LLM}}(s_{t+1})$ where
  - ΔC_t: change in visual coverage between timesteps t and $t+1$ computed on the ground plane $z = 0$ to prioritize surface-level exploration.
  - $\mathbb{1}_{\text{redundant}}(s_{t+1}), \mathbb{1}_{\text{collision}}(s_{t+1}), \mathbb{1}_{\text{idle}}(s_{t+1})$: binary indicators for redundant views, collisions, and idle behavior respectively in s_{t+1}.
  - $f_{\text{battery}}(b_{t+1})$: energy consumption penalty for battery usage observed at $t+1$.
  - $f_{\text{cam}}(\theta_{t+1})$: incentive to encourage useful variations to pan, tilt, and zoom camera parameters in s_{t+1}.
  - $f_{\text{curiosity}}(s_{t+1})$: reward for the exploration of under-visited areas in s_{t+1}.
  - $f_{\text{LLM}}(s_{t+1})$: auxiliary reward component from PARE that evaluates deviation in position and camera parameters from LLM-based action recommendation at time $t+1$.
  - The λ coefficients are tunable non-negative scalars $(\mathbb{R}_{\geq 0})$ to balance multiple behavioral objectives.
- $\gamma \in [0,1]$ is the discount factor that indicates how much future rewards are valued.

This work emphasizes visual cell coverage derived from the camera's configuration unlike traditional path planning which rewards cell visitation within the environment. At each step of exploration, the UAV observes a subset of the 3D environment using a forward-facing view cone defined by the camera's configuration. At time t, $\mathbf{p}_t = (x_t, y_t, z_t)$ and $\theta_t = (\phi_t, \psi_t, \zeta_t)$ describe the UAV's 3D position and camera parameters for (tilt, pan, zoom) respectively. Let $\text{ViewCone} : \mathbb{Z}^2 \times \Theta \rightarrow 2^{\mathbb{Z}^3}$ be the function that maps a given UAV position and camera configuration to a set of observed cells in the 3D environment. At time t, the set of observed cells is defined as $\mathcal{V}_t = \left\{(i,j,k) \in \mathbb{Z}^3 \,\middle|\, (i,j,k) \in \text{ViewCone}(\mathbf{p}_t, \theta_t)\right\}$. These observations are used to update the visual coverage map $\mathcal{C}^{(t)}_{i,j,k} \in \{0,1\}$ at time $t+1$ as:

$$\mathcal{C}^{(t+1)}_{i,j,k} = \begin{cases} 1 & \text{if}(i,j,k) \in \mathcal{V}_t \\ \mathcal{C}^{t}_{i,j,k} & \text{otherwise} \end{cases}.$$

The agent policy $\pi_\theta(a|s)$ is learned using PPO which maximizes the expected discounted return $J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$ where r_t is the reward at time t and $\gamma \in [0,1]$. PPO ensures that the policy updates are stable using the clipped surrogate objective which clips the policy change at each optimization step. It is given as:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}\big(r_t(\theta), 1-\varepsilon, 1+\varepsilon\big)\hat{A}_t\right)\right],$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ and $\hat{A}_t$ is the Generalized Advantage Estimation (GAE).

Unlike prior research [31] where LLM generates task-specific reward functions, this work proposes prompt-informed reinforcement learning (PIRL), a technique that supplements the curiosity-driven action-based reward signal of the PPO RL policy with zero-shot reasoning feedback from an LLM. As illustrated in Figure 1, PIRL dynamically shapes the reward signal at each step of learning based on an LLM's semantic guidance, introducing an adaptive feedback mechanism beyond conventional RL reward formulations.

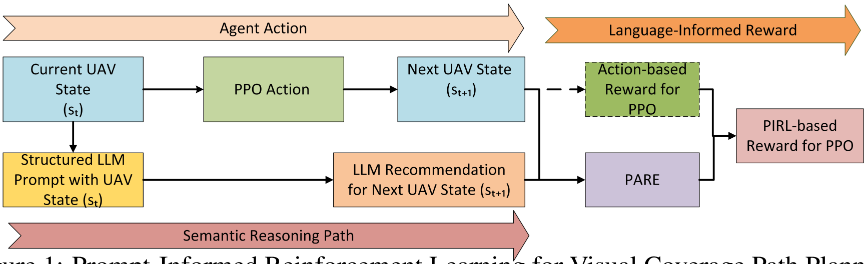

Figure 1: Prompt-Informed Reinforcement Learning for Visual Coverage Path Planning

A structured text prompt $\pi(s_t)$, comprising the concatenation of the following three components is generated using a deterministic parser: 1) task description π_{task}, apprising the LLM of the VCPP task, 2) Environment summary π_{env}, including the UAV's 3D position, camera parameters, battery state, obstacles, and coverage, 3) Request template π_{query}, requesting the LLM for recommendations on the UAV's position and camera parameters. Therefore, $\pi(s_t) = \pi_{\text{task}} || \pi_{\text{env}} || \pi_{\text{query}}$ where $||$ denotes string concatenation. Given the structured text prompt $\pi(s_t)$ at time t, the LLM acts as a mapping function $\text{LLM} : \mathcal{L} \rightarrow \mathbb{Z}^3 \times \Theta$ generating the recommendation $\mathbf{a}_{t+1}^{\text{LLM}} = (\mathbf{p}_{t+1}^{\text{LLM}}, \theta_{t+1}^{\text{LLM}})$ where $\mathbf{p}_{t+1}^{\text{LLM}} = (x_{t+1}^{\text{LLM}}, y_{t+1}^{\text{LLM}}, z_{t+1}^{\text{LLM}}) \in \mathbb{Z}^3$, and $\theta_{t+1}^{\text{LLM}} = (\phi_{t+1}^{\text{LLM}}, \psi_{t+1}^{\text{LLM}}, \zeta_{t+1}^{\text{LLM}}) \in \Theta$ encode the recommendations for the UAV's 3D position and camera parameters (tilt, pan, zoom) respectively, for time $t+1$.

The proposed PARE module integrates LLM recommendations into the RL policy by computing an auxiliary reward that penalizes the deviation in direction, position, and camera alignment between the agent's actions and the LLM's recommendations. The composite movement alignment reward combines both directional and positional alignment. At time $t+1$, let $\Delta\mathbf{p}_{t+1} = \mathbf{p}_{t+1} - \mathbf{p}_t$ and $\Delta\mathbf{p}_{t+1}^{LLM} = \mathbf{p}_{t+1}^{LLM} - \mathbf{p}_t$ be the agent and LLM-recommended movement vectors respectively. The directional alignment, which quantifies the alignment between the agent's trajectory and LLM's recommended direction (independent of distance), is computed as their cosine similarity defined as: $\text{DirAlign}(s_{t+1}) = \frac{\Delta\mathbf{p}_{t+1} \cdot \Delta\mathbf{p}_{t+1}^{\text{LLM}}}{\|\Delta\mathbf{p}_{t+1}\|_2 \|\Delta\mathbf{p}_{t+1}^{\text{LLM}}\|_2}$ where $\|\cdot\|_2$ denotes Euclidean (L2) norm. The positional alignment is computed using inverse normalized Euclidean distance, given as: $\text{PosAlign}(s_{t+1}) = 1 - \frac{\|\mathbf{p}_{t+1} - \mathbf{p}_{t+1}^{\text{LLM}}\|_2}{d_{\max}}$, where $d_{\max}$ is the maximum possible distance in the environment. The composite movement alignment reward is given by: $f_{\text{move}}(s_{t+1}) = \alpha \cdot \text{DirAlign}(s_{t+1}) + (1-\alpha) \cdot \text{PosAlign}(s_{t+1})$, where $\alpha \in [0,1]$ balances the two alignment components. Likewise at time $t+1$, the camera alignment penalty is computed as the deviation between the agent camera parameters θ_{t+1} and LLM-recommended camera parameters $\theta_{t+1}^{\text{LLM}}$. It is defined as: $f_{\text{cam}}(s_{t+1}) = -\|\theta_{t+1} - \theta_{t+1}^{\text{LLM}}\|_1$ where $\|\cdot\|_1$ is the Taxicab (L1) norm. The overall reward augmentation by PARE is given as: $f_{\text{LLM}}(s_{t+1}) = \lambda_{\text{cam}} \cdot f_{\text{cam}}(s_{t+1}) + \lambda_{\text{move}} \cdot f_{\text{move}}(s_{t+1})$ where $\lambda_{\text{cam}}, \lambda_{\text{move}} \in \mathbb{R}_{\geq 0}$.
Further rationale for the composite movement alignment reward and treating camera alignment as a constraint while rewarding movement alignment for compliance is provided in section A.2 of the Appendix. This enables the agent to leverage the LLM as a context-aware, soft-task planner dynamically guiding the reward signal for VCPP.

4 Experiment

4.1 Research Questions

This study is motivated by the following research questions (RQ):

RQ1: Can a pretrained LLM effectively inform and enhance the reward shaping of a PPO-based RL policy in 3D visual coverage environments?

RQ2: How does prompt informed reward shaping compare with other learning-based baselines in terms of exploration efficiency, battery utilization, and redundancy minimization?

RQ3: Can language-guided reward signals improve generalization across previously unseen 3D environments?

4.2 Agent Training

The VCPP task is simulated in a custom 3D environment of size 15x15x3 built using the OpenAI Gym framework [3]. The UAV is equipped with a logical pan-tilt-zoom (PTZ) camera whose parameters are trained to maximize visual coverage while minimizing redundant observations and battery consumption. At the start of each episode, spherical obstacles, between two and five, are introduced into the environment dynamically to ensure that UAV learns to avoid obstacles as it explores the environment. Furthermore, at each timestep, stochastic wind disturbances are applied to the environment along the X and Y axes to simulate real-world conditions. The agent is trained using the curiosity-driven PPO approach whose reward function is initialized using exploratory weight random initialization (EWRI). EWRI encourages early exploration by introducing stochasticity into the reward function (RF), drawing reward weights (RW) from a distribution at the beginning of each training episode (shown in Table 1), thereby promoting robustness and generalization. The chosen ranges ensure that the relative priorities among RF coefficients are preserved and positive shaping terms such as coverage, curiosity, and camera movement have a significant impact unless strongly offset by violations like collisions or redundancy. The LLM alignment (λ_{LLM}) weight range is deliberately set lower than the upper bound of coverage gain to ensure that maximizing visual coverage remains the most important goal.

At each timestep of agent training, the UAV state which includes 3D UAV position, battery state, coverage, and camera parameters, is curated into a structured zero-shot prompt and passed as input to OpenAI's GPT-3.5 LLM. In return, the LLM recommends the future state values for the UAV position along the Cartesian axes, and camera parameters such as pan, tilt, and zoom. The recommendations made by the LLM are not enforced directly, but incorporated into the reward function using PARE. The rationale behind choosing OpenAI GPT-3.5, and curation of the structured zero-shot prompt is provided in sections A.3 and A.4 of the Appendix respectively.

Table 1: EWRI Ranges for RF Coefficients

RW	Range	Description
λ_c	$\mathcal{U}(0.5,\ 1.5)$	coverage gain
λ_r	$\mathcal{U}(-1.0,\ -0.2)$	redundancy penalty
λ_b	$\mathcal{U}(-0.5,\ -0.1)$	battery efficiency
λ_{cam}	$\mathcal{U}(0.2,\ 0.6)$	PTZ parameter usage
λ_{cur}	$\mathcal{U}(0.2,\ 0.5)$	curiosity incentive
$\lambda_{\text{collision}}$	$\mathcal{U}(-1.5,\ -0.8)$	collision penalty
λ_{idle}	$\mathcal{U}(-0.8,\ -0.3)$	idle/inactivity penalty
λ_{LLM}	$\mathcal{U}(0.2,\ 0.6)$	LLM alignment

4.3 Evaluation Metrics

The trained PPO agent is evaluated using the following core metrics which reflect task success and operational efficiency:

4.3.1 Visual Coverage Ratio (VCR)

VCR is defined as the fraction of unique ground cells (cells at $z = 0$) in the 3D environment that are seen by the UAV's camera view cone as the camera configuration is altered during an episode. It is given by $\text{VCR} = \frac{|\mathcal{C}_{\text{covered}}|}{|\mathcal{C}_{\text{total}}|}$ where $\mathcal{C}_{\text{covered}} \subseteq \mathcal{C}_{\text{total}}$; $\mathcal{C}_{\text{covered}}$ and $\mathcal{C}_{\text{total}}$ are the set of uniquely observed ground cells and total ground cells in the environment respectively, for a given episode. The higher the VCR, the better the UAV's coverage of the environment.

4.3.2 Battery Efficiency (BE)

BE is defined as the VCR per unit battery consumed, and is given by $\text{BE} = \frac{\text{VCR}}{2 - \frac{b_t}{b_i}}$, where b_i and b_t represent the initial and final battery level for the episode. The higher the BE, the better the UAV's battery usage.

4.3.3 Redundant Views per Covered Cell (RVC)

RVC measures spatial redundancy in ground-level observations, and is computed as the average number of redundant views per uniquely covered ground cell. It is given by $\text{RVC} = \frac{|\mathcal{C}_{\text{redundant}}|}{|\mathcal{C}_{\text{covered}}|}$, where $\mathcal{C}_{\text{redundant}}$ is the set of ground cells observed more than once during the episode. The lower the RVC, the better the efficiency in coverage with minimal redundancy.

4.4 Results

The trained PPO agent is inferred in two different simulation environments, OpenAI Gym, and Webots. While OpenAI Gym offers a grid-based, logic-level simulation, Webots allows for physics-based, sensor-level robotic simulation. Therefore, evaluating the model in Webots allows for sim-to-real-like evaluation by determining whether the learned policy can tolerate an imperfect and realistic 3D environment. The agent is evaluated zero-shot in test environments of varying sizes including 30x30x3, 45x45x3, and 60x60x3 with the number of steps and battery strength scaled using the scaling factor $\alpha = \frac{x_{\text{test}} \cdot y_{\text{test}} \cdot z_{\text{test}}}{x_{\text{train}} \cdot y_{\text{train}} \cdot z_{\text{train}}}$. The OpenAI Gym environment is emulated in Webots (Figure 2) by creating a logical mirror of the UAV simulation state. The trained agent is simulated on a DJI Mavic 2 Pro fitted with a controllable gimbal-mounted camera in the Webots simulator. Although Webots does not interpret image streams natively for RL, it allows position-based coverage approximation consistent with the OpenAI Gym implementation. Zero-shot inference is performed in the Webots simulator with the agent not undergoing any training in a Webots environment. The evaluation metrics described in section 4.3 are logged across 100 independent episodes for each environment wherein obstacles and UAV start position are

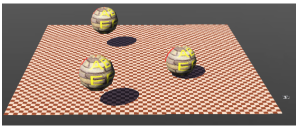

Figure 2: Environment with Three Obstacles and UAV in Webots

varied for each episode, essentially inferring on 100 varied environments of the same size to determine the performance of the trained agent. To ensure consistency, the obstacle locations and start positions in different environments are preserved as the proposed PIRL agent is compared with baseline agents.

The inference results of the proposed PIRL approach are compared with multiple baseline approaches. A curiosity-driven PPO baseline with static reward function and no LLM guidance (PPO-SR), and a PPO policy initialized with EWRI without LLM guidance (PPO-EWRI), LLM-only without PPO (LLM-only), and imitation learning (IL) using TabNet [1] are used to compare with the proposed approach. The mean of each of the evaluation metrics from running inference on 100 episodes in OpenAI Gym and Webots is as shown in Tables 2 and 3, respectively. A graph comparing the results, and further insight into the setup of the baseline approaches is provided in section A.5 of the Appendix.

Table 2: Mean of Results in OpenAI Gym

Method	30x30x3			45x45x3			60x60x3		
	VCR	BE	RVC	VCR	BE	RVC	VCR	BE	RVC
PPO-SR	0.64	0.55	0.22	0.61	0.56	0.24	0.63	**0.48**	0.26
PPO-EWRI	0.61	0.59	0.24	0.59	**0.62**	0.25	0.62	0.36	0.27
LLM-only	0.56	0.42	0.28	0.53	0.40	0.31	0.55	0.39	0.27
IL	0.48	0.21	0.40	0.44	0.15	0.47	0.41	0.11	0.49
PIRL	**0.73**	**0.74**	**0.18**	**0.71**	0.58	**0.21**	**0.70**	0.23	**0.24**

Table 3: Mean of Results in Webots

Method	30x30x3			45x45x3			60x60x3		
	VCR	BE	RVC	VCR	BE	RVC	VCR	BE	RVC
PPO-SR	0.61	**0.47**	0.21	0.62	0.52	0.26	0.65	0.30	0.22
PPO-EWRI	0.63	0.44	0.19	0.65	0.57	0.24	0.60	0.32	0.24
LLM-only	0.59	0.40	0.26	0.57	0.43	0.24	0.58	**0.36**	0.23
IL	0.36	0.29	0.39	0.29	0.22	0.41	0.37	0.30	0.44
PIRL	**0.70**	0.42	**0.14**	**0.79**	**0.60**	**0.22**	**0.71**	0.35	**0.17**

4.5 Discussion

RQ1: Can a pretrained LLM effectively inform and enhance the reward shaping of a PPO-based RL policy in 3D visual coverage environments?

The results from OpenAI Gym and Webots environments demonstrate that the LLM-guided PIRL significantly enhances the visual coverage capabilities of the PPO agent. In OpenAI Gym, on the 30x30x3 environment, PIRL achieves a VCR of 0.73 compared to 0.64 (PPO-SR), 0.61 (PPO-EWRI), 0.56 (LLM-only), and 0.48 (IL), exhibiting relative improvements ranging from 14% (over PPO-SR) to 52% (over IL). Similarly, PIRL records VCR of 0.71 and 0.70 in larger environments of sizes 45x45x3 and 60x60x3 respectively, outperforming the best baseline (PPO-SR) by 16% and 11% respectively. Furthermore, a similar trend is observed in the Webots simulator which introduces realistic physical dynamics and environmental perturbations such as wind. In the 30x30x3 environment, PIRL scores a VCR of 0.70 surpassing PPO-EWRI (0.63), PPO-SR (0.61), LLM-only (0.59), and IL (0.36). It translates to performance gains ranging from 11% (over PPO-EWRI) to 94% (over IL). Likewise, PIRL achieves a VCR of 0.79 and 0.71 on 45x45x3 and 60x60x3 environments respectively, maintaining margins of 27% and 9% over the best baselines of PPO-SR and PPO-EWRI respectively. The results affirm the feasibility of LLM-informed reward shaping to improve the visual coverage ability of a PPO-based RL policy in 3D environments.

RQ2: How does prompt informed reward shaping compare with other learning-based baselines in terms of exploration efficiency, battery utilization, and redundancy minimization?

The proposed PIRL approach improves the exploration efficiency of the PPO agent by guiding behavior that minimizes redundancy and promotes diverse viewpoint adjustments. This is evidenced by the best RVC values achieved by PIRL across both OpenAI Gym and Webots simulators. In OpenAI Gym, PIRL achieves RVC of 0.18, 0.21, and 0.24 in the 30x30x3, 45x45x3, and 60x60x3 environments respectively. They are lower (better) than those achieved by the best baseline, PPO-SR, which achieves RVC scores of 0.22, 0.24, and 0.26 on the same environments translating to relative improvements of 18%, 4%, and 8% respectively in redundancy. Similarly, in Webots, PIRL achieves the best RVC scores of 0.14, 0.22, and 0.17 in the 30x30x3, 45x45x3, and 60x60x3 environments, outperforming all the baseline approaches.

BE results offer more nuanced trends compared to RVC. In OpenAI Gym, PIRL attains a BE of 0.74 in the 30x30x3 environment, 25% higher than the nearest baseline (PPO-EWRI - 0.59) but exhibits lower BE in larger environments. Specifically, PIRL's BE drops to 0.58 in the 45x45x3 environment (lower than PPO-EWRI's 0.62) and 0.23 in the 60x60x3 environment, trailing PPO-SR (0.48), PPO-EWRI (0.36), and LLM-only (0.39). Despite this drop, PIRL continues to achieve the highest VCR which indicates that PIRL favors maximizing energy utilization for diverse observations rather than minimizing consumption, a learned preference shaped by dynamic reward signals. In Webots, PIRL maintains a competitive or superior BE profile, achieving 0.60 on the 45x45x3 grid, the highest among all approaches. Though it is slightly outperformed in the 30x30x3 and 60x60x3 environment, PIRL's superior RVC and VCR values across all grid sizes compensate for the minor BE trade-off. Therefore, PIRL encourages the PPO agent to explore meaningfully capturing more unique observations with a reasonable trade-off between coverage and energy expenditure.

RQ3: Can language-guided reward signals improve generalization across previously unseen 3D environments?

As part of the experiment, all the agents are initially trained in a 15x15x3 environment using OpenAI Gym. The agents are inferred in environments of sizes 30x30x3, 45x45x3, and 60x60x3 in Webots to test the zero-shot ability of the agents in a realistic physics-based simulator. The robustness of the PIRL agent in Webots, despite being trained exclusively in OpenAI Gym, demonstrates its strong generalization capacity to zero-shot, sim-to-real-like environments. PIRL achieves high VCR and low RVC values across all three grid sizes in Webots without any additional fine-tuning. For instance, PIRL improves VCR over IL by 94% on the 30×30×3 grid, and maintains strong lead margins of 27% and 9% on the 45×45×3 and 60×60×3 grids, respectively, over the best baselines. Furthermore, PIRL achieves the lowest RVC scores in all Webots settings, and while its BE is not the highest in all environments, it reaches a peak of 0.60 BE on the 45×45×3 grid, outperforming all baselines in that configuration. These results validate that the language-guided feedback used for reward modulation in PIRL allows the agent to learn generalizable exploration behaviors that remain effective even in high-fidelity, unstructured environments with physical disturbances.

4.6 Limitations

Although the proposed PIRL approach demonstrates robust performance across diverse environments, the approach comes with a few limitations. They are:

- PIRL relies on access to a pretrained LLM with zero-shot reasoning and in-context learning capabilities to generate semantic guidance during training. This dependency may introduce computational overhead and latency, which could be limiting for resource-constrained settings or cost-sensitive applications.
- The current formulation models the state space discretely for both position and camera parameters, which simplifies the learning problem but may limit the granularity of control and realism. Extending PIRL to continuous or hybrid state-action spaces would enable more precise control and facilitate real-world deployment.
- While sim-to-real-like evaluation using Webots introduces physical realism, real-world

deployment would require further considerations such as sensor noise, localization error, and computational constraints onboard the UAV platform.

5 Conclusion and Future Work

This study presents PIRL that leverages LLMs for dynamic reward shaping of a PPO policy aimed at visual coverage of 3D environments. Through robust experimental setup and extensive evaluation in OpenAI Gym and Webots, this study demonstrates the superior performance of PIRL for VCPP by comparing with multiple learning-based baseline approaches. While the proposed approach highlights the potential of LLM-guided reward shaping, future work will focus on enhancing PIRL by integrating visual signal fusion where real-time camera streams are semantically interpreted to adaptively influence reward dynamics. Furthermore, incorporating LLM self-evaluation mechanisms such as confidence scores can help modulate the influence of LLM guidance during training in different environments. These research directions will be actively pursued to increase the robustness, and real-world applicability of PIRL-based systems.

Acknowledgement

This work is supported by the Pollinator Health: Research and Application no. 2024-67014-42301 from the U.S. Department of Agriculture's National Institute of Food and Agriculture.

References

[1] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.

[2] Francesco Betti Sorbelli. Uav-based delivery systems: A systematic review, current trends, and research challenges. *Journal on Autonomous Transportation Systems*, 1(3): 1–40, 2024.

[3] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[4] Eli Bronstein, Mark Palatucci, Dominik Notz, Brandyn White, Alex Kuefler, Yiren Lu, Supratik Paul, Payam Nikdel, Paul Mougin, Hongge Chen, et al. Hierarchical model-based imitation learning for planning in autonomous driving. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8652–8659. IEEE, 2022.

[5] José Pedro Carvalho and A Pedro Aguiar. Deep reinforcement learning for zero-shot coverage path planning with mobile robots. *IEEE/CAA Journal of Automatica Sinica*, 2025.

[6] Lin Chen, Yaonan Wang, Zhiqiang Miao, Yang Mo, Mingtao Feng, Zhen Zhou, and Hesheng Wang. Transformer-based imitative reinforcement learning for multirobot path planning. *IEEE Transactions on Industrial Informatics*, 19(10):10233–10243, 2023.

[7] Rupayan Das, Angshuman Khan, and Gunjan Paul. A proximal policy optimization with curiosity algorithm for virtual drone navigation. *Engineering Research Express*, 6 (1):015057, 2024.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[10] Michael Gargalakos. The role of unmanned aerial vehicles in military communications: application scenarios, current trends, and beyond. *The Journal of Defense Modeling and Simulation*, 21(3):313–321, 2024.

[11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[12] Hengyuan Hu and Dorsa Sadigh. Language instructed reinforcement learning for human-ai coordination. In *International Conference on Machine Learning*, pages 13584–13598. PMLR, 2023.

[13] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR, 2022.

[14] Louis Kirsch, James Harrison, Jascha Sohl-Dickstein, and Luke Metz. General-purpose in-context learning by meta-learning transformers. *arXiv preprint arXiv:2212.04458*, 2022.

[15] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[16] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. Reward design with language models. *arXiv preprint arXiv:2303.00001*, 2023.

[17] Sha Luo, Hamidreza Kasaei, and Lambert Schomaker. Self-imitation learning by planning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4823–4829. IEEE, 2021.

[18] Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. Inform: Mitigating reward hacking in rlhf via information-theoretic reward modeling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435.

[20] Jeremy Roghair, Amir Niaraki, Kyungtae Ko, and Ali Jannesari. A vision based deep reinforcement learning algorithm for uav obstacle avoidance. In *Proceedings of SAI Intelligent Systems Conference*, pages 115–128. Springer, 2021.

[21] Andres J Sanchez-Fernandez, Luis F Romero, Gerardo Bandera, and Siham Tabik. Vpp: visibility-based path planning heuristic for monitoring large regions of complex terrain using a uav onboard camera. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:944–955, 2021.

[22] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[23] Farooq Shareef. Enhancing conversational ai with llms for customer support automation. In *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, pages 239–244. IEEE, 2024.

[24] Pratyusha Sharma, Balakumar Sundaralingam, Valts Blukis, Chris Paxton, Tucker Hermans, Antonio Torralba, Jacob Andreas, and Dieter Fox. Correcting robot plans with natural language feedback. *arXiv preprint arXiv:2204.05186*, 2022.

[25] Siyu Teng, Long Chen, Yunfeng Ai, Yuanye Zhou, Zhe Xuanyuan, and Xuemin Hu. Hierarchical interpretable imitation learning for end-to-end autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 8(1):673–683, 2022.

[26] Francesco Toscano, Costanza Fiorentino, Nicola Capece, Ugo Erra, Danilo Travascia, Antonio Scopa, Marios Drosos, and Paola D'Antonio. Unmanned aerial vehicle for precision agriculture: A review. *IEEE access*, 2024.

[27] Jianxun Wang and Yixiang Chen. A review on code generation with llms: Application and evaluation. In *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, pages 284–289. IEEE, 2023.

[28] Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy. Reinforcement learning enhanced llms: A survey. *arXiv preprint arXiv:2412.10400*, 2024.

[29] Yinchu Wang, Zhi He, Dandan Cao, Li Ma, Kai Li, Liangsheng Jia, and Yongjie Cui. Coverage path planning for kiwifruit picking robots based on deep reinforcement learning. *Computers and Electronics in Agriculture*, 205:107593, 2023.

[30] Huilin Yin, Yinjia Lin, Jun Yan, Qian Meng, Karin Festl, Lukas Schichler, and Daniel Watzenig. Agv path planning using curiosity-driven deep reinforcement learning. In *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, pages 1–6. IEEE, 2023.

[31] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, et al. Language to rewards for robotic skill synthesis. *arXiv preprint arXiv:2306.08647*, 2023.

[32] Xiangyu Zeng, Tobias Zaenker, and Maren Bennewitz. Deep reinforcement learning for next-best-view planning in agricultural applications. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2323–2329. IEEE, 2022.