Prompt-Informed Reinforcement Learning for Visual Coverage Path Planning

Venkat Margapuri
Department of Computing Sciences
Villanova University, USA

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.

INTRODUCTION

Problem:

- Visual coverage path planning (VCPP) for UAVs requires balancing coverage, energy, and redundancy.
- Traditional RL methods depend on manually tuned, static reward functions that lack semantic adaptability.

Gap:

- Reward hacking and environment-specific rewards hinder generalization.
- LLMs (e.g., GPT-3.5) possess contextual reasoning but lack environmental grounding.

Core Idea: PIRL combines the adaptability of LLMs with the optimization strength of RL by using LLMs as dynamic reward shapers rather than policy generators.

KEY CONTRIBUTIONS

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.

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

METHODOLOGY

1. Task Setup:
The Visual Coverage Path Planning (VCPP) task is modeled as a Markov Decision Process M = (S, A, P, R, γ).
Each state $s_t = (p_t, \theta_t, b_t)$ includes the UAV's 3D position (x, y, z), camera configuration (tilt, pan, zoom), and normalized battery level (0–1).
Actions include movement (x±, y±, z±) and camera control (tilt±, pan±, zoom±) with fixed step sizes of 5°, 15°, and 0.1 respectively.

2. PPO Learning:
A Proximal Policy Optimization (PPO) agent learns a policy $\pi\theta(a|s)$ that maximizes expected return $J(\theta) = E[\Sigma \gamma^t r_t]$.
The clipped PPO objective ensures stable updates and prevents large policy shifts. This curiosity-driven policy balances exploration, coverage, and energy efficiency during 3D flight.

3. Reward Composition:
The total reward $r_{t+1} = \Sigma \lambda_i f_i(s_{t+1})$ combines several terms:

- Coverage gain (ΔC_t) for new visual area
- Curiosity incentive for exploring unseen regions
- Battery and collision penalties
- Camera variation encouragement
- LLM-based auxiliary reward fLLM for semantic alignment
- Each λ_i weight adjusts the importance of these behaviors.

4. Prompt-Adaptive Reward Engine (PARE):
PARE integrates GPT-3.5 feedback into the reward loop:

1. A structured prompt $\pi(s_t)$ describes the UAV state, environment, and task goal.
2. GPT-3.5 returns a recommended next state (position and camera configuration).
3. PARE computes an auxiliary reward that measures alignment between the agent's actions and the LLM's suggestions.

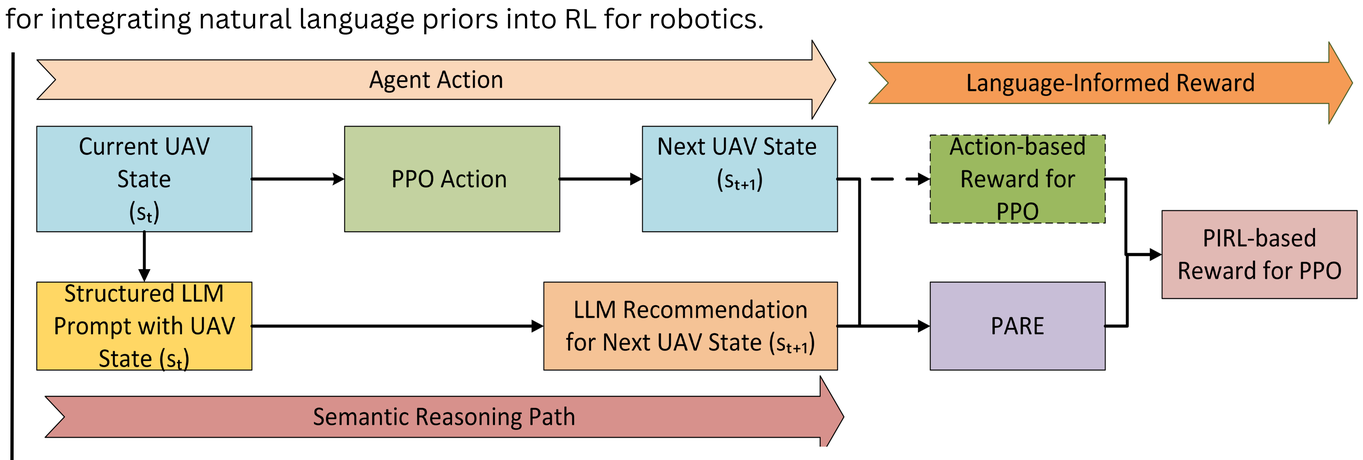
5. Alignment Rewards:
Movement alignment combines direction and position similarity:

f_move = α·DirAlign + (1−α)·PosAlign,

encouraging the UAV to move in directions semantically consistent with LLM feedback.
Camera alignment penalizes deviation from the suggested pan-tilt-zoom configuration using

f_cam = −||θ_agent − θ_LLM||.

The total auxiliary reward is **f_LLM = λ_cam·f_cam + λ_move·f_move**.



PIRL Architecture for Visual Coverage Path Planning

EXPERIMENT

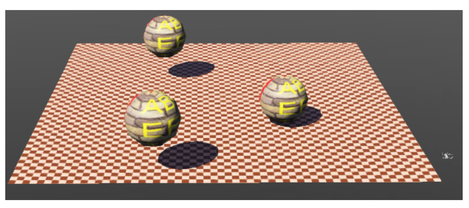
Agent Training:
The UAV is trained in a custom 3D environment (15×15×3) built in OpenAI Gym with a pan-tilt-zoom camera. Each episode begins with 2–5 random spherical obstacles, and wind disturbances along the X and Y axes simulate real-world conditions.
Training uses a curiosity-driven PPO algorithm with Exploratory Weight Random Initialization (EWRI), which randomizes reward weights each episode to encourage broad exploration and improve generalization.
Key reward factors include coverage gain, redundancy and collision penalties, curiosity, camera motion, battery efficiency, and LLM alignment.
At each timestep, the UAV state (position, battery, coverage, camera) is converted into a structured prompt for GPT-3.5, which returns a recommended next position and camera configuration. These suggestions are not enforced directly but integrated into the reward via the Prompt-Adaptive Reward Engine (PARE).

EVALUATION METRICS

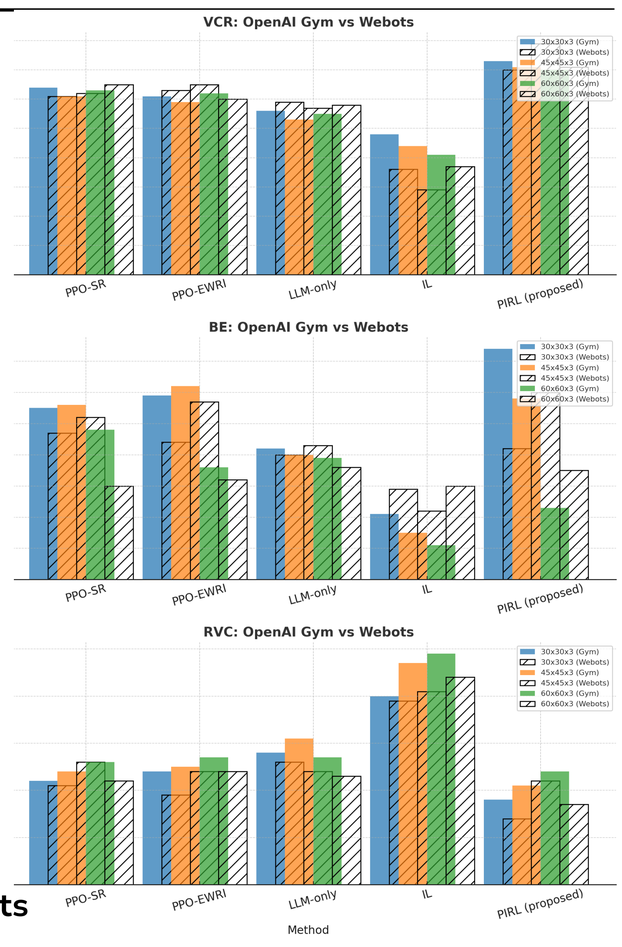
Visual Coverage Ratio (VCR): Fraction of unique ground cells viewed by the UAV — higher is better.
Battery Efficiency (BE): Coverage achieved per unit battery consumed.
Redundant Views per Covered Cell (RVC): Average number of duplicate observations — lower is better.

RESULTS

The policy is tested zero-shot in both OpenAI Gym and Webots, evaluating robustness under realistic physics and noise. Environments of size 30×30×3, 45×45×3, and 60×60×3 are used, with scaling factors to match episode length and energy limits.
Across 100 episodes per setup, PIRL consistently outperforms all baselines:
In Gym, PIRL achieved up to 14 % higher VCR, 25 % higher battery efficiency, and 18 % lower redundancy compared to PPO baselines.
In Webots, PIRL maintained strong sim-to-real transfer, achieving 27 % higher coverage and the lowest redundancy across all grid sizes.
These results confirm that LLM-guided reward shaping enables smarter exploration, improved energy efficiency, and better generalization to unseen 3D environments.



Environment with three obstacles in Webots



DISCUSSION

The proposed PIRL framework was compared against four key baselines: PPO-SR (static reward), PPO-EWRI (stochastic reward initialization), LLM-only (no RL training), and imitation learning with TabNet. These comparisons highlight distinct trade-offs in exploration and adaptability. PPO-SR performs consistently but lacks semantic flexibility, while PPO-EWRI improves early exploration through random reward weighting yet remains task-specific. The LLM-only agent demonstrates contextual reasoning but lacks stability without reinforcement feedback. TabNet captures imitation patterns but fails to generalize beyond demonstrations. In contrast, PIRL unifies the semantic reasoning of LLMs with the optimization strength of PPO, enabling adaptive, context-aware exploration that balances coverage, efficiency, and generalization across unseen 3D environments.

FUTURE WORK AND CONCLUSION

This work presented PIRL, which uses large language models to dynamically shape PPO rewards for UAV visual coverage in 3D environments. PIRL outperforms multiple baselines in both OpenAI Gym and Webots, achieving higher coverage and lower redundancy. Future efforts will focus on real-time visual signal fusion and confidence-based LLM guidance to enhance robustness and real-world adaptability.