A. Implementation Details

A.1. PPO Modeling & Hyperparameters

One of the most challenging aspects of implementing a reinforcement learning agent is hyperparameter tuning, since these choices strongly influence stability, convergence speed, and the final behavior of the learned policy. In our work, we adopt the Proximal Policy Optimization (PPO) implementation from Stable-Baselines3 [33] and treat the hyperparameters as a carefully tuned configuration for reliable training in the wildfire digital twin.

Table 4 summarizes the main PPO hyperparameters used for all experiments, including optimization settings, rollout configuration, and training horizon.

B. Extended Quantitative Results

This appendix reports the full quantitative results for all six models across the five evaluation tasks described in Section 4. Each table summarizes performance for a single task in terms of total return, wildfire visibility, detection latency, distance traveled, and runtime. Together, these results highlight how VLM-guided reward shaping, dual-camera sensing, and directional guidance contribute to improved wildfire localization, tracking, and robustness.

B.1. Task 1: Wildfire in Initial Field of View

In Task 1, wildfire is already visible in the initial top-down frame, so all models can detect and track the fire without significant exploration. As expected, all six variants achieve 100% time-in-FOV and near-identical time-to-detection. Differences therefore arise mainly in total reward and flight efficiency. The VLM-guided final model attains the highest total reward while maintaining 100% coverage, indicating slightly more efficient control and reward optimization, whereas the VLM-only model travels substantially less distance due to the absence of a base exploration reward.

B.2. Task 2: Wildfire Near UAV (Out of Initial FOV)

Task 2 requires short-range search: the UAV starts roughly 100 m from the fire with no flames in the initial top-down view. Here, model differences become more pronounced. The VLM-guided final model achieves the highest total reward, highest time-in-FOV, and shortest time-to-detection, indicating more efficient exploration and transition from search to tracking. The VLM-guided model without segmentation performs similarly, while the baseline PPO and VLM-only variants show lower FOV coverage and much slower detection.

B.3. Task 3: Wildfire at Long Range

Task 3 is the most challenging setting: the UAV starts roughly 1 km from the fire, with no flames in the top-down view and only a small smoke cue in the angled view. Under this long-range, sparse-reward regime, only the two dual-camera VLM-guided models succeed in detecting the wildfire within the 4000-timestep limit. All four other models fail to bring the fire into the top-down FOV, yielding zero FOV time and undefined detection latency. The final model achieves the highest total reward, non-zero FOV coverage, and a finite time-to-detection, highlighting the importance of VLM-based semantic guidance for distant search.

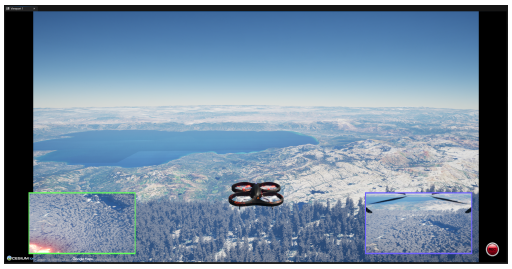


Figure 5. UAV deployed with wildfire in the initial FOV of its top-down camera (Wildfire in FOV).

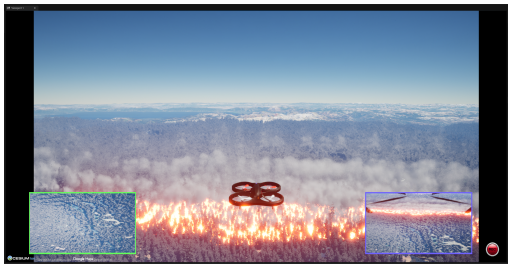


Figure 6. UAV deployed approximately 100 m away from the nearest wildfire instance (Wildfire near UAV).

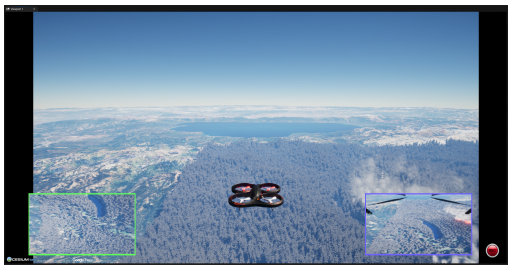


Figure 7. UAV deployed approximately 1 km away from the nearest wildfire instance (Wildfire in distance).

B.4. Task 4: Robustness to Wind

Task 4 reuses the Task 2 spatial configuration (wildfire $\approx$ 100 m away, initially outside the top-down FOV) but adds a 5 m/s crosswind. All models experience some degradation relative to Task 2, yet the VLM-guided final model again attains the highest total reward, high FOV coverage, and low detection latency. The close performance between the fi-

Hyperparameter	Name	Value	Description
Learning Rate	learning_rate	3×10^{-4}	Adam optimizer learning rate
Steps per Batch	n_steps	2000	Environment steps per policy update
Minibatch Size	batch_size	400	Minibatch size for gradient updates
Number of Epochs	n_epochs	20	PPO epochs per update
Discount Factor γ	gamma	0.99	Return discount factor
Loss Function	loss_func	ClipPPOLoss	Clipped PPO surrogate loss
Value Estimator	value_estimator	GAE	Generalized advantage estimation
GAE λ	gae_lambda	0.95	Smoothing parameter for GAE
Clip Epsilon ϵ	clip_range	0.20	PPO clipping range
Entropy Coefficient	ent_coef	0.01	Entropy regularization coefficient
Value Function Coeff.	vf_coef	0.6	Value function loss coefficient
Max Gradient Norm	max_grad_norm	0.5	Gradient clipping norm
Steps per Episode	max_steps	4000	Maximum timesteps per episode
Evaluation Frequency	eval_freq	20000	Evaluation callback frequency
Total Timesteps	total_timesteps	200000	Total training timesteps

Table 4. PPO hyperparameters used for training the wildfire-monitoring UAV policy with Stable-Baselines3 [33].

Model	R_{total}	FOV%	TTD (s)	(steps)	Dist (m)	Runtime (s)
Base PPO	5112.54	**100.0%**	**6.12**	**1.0**	4119.91	**4420.2**
VLM-only rew. shaping	3936.43	**100.0%**	6.33	**1.0**	2028.45	4510.1
VLM-int., top-down only	5118.06	**100.0%**	6.38	**1.0**	**6089.79**	4708.9
VLM-int., angled only	4612.58	**100.0%**	6.42	**1.0**	6076.34	4710.8
VLM-guided, unsegmented	5129.60	**100.0%**	6.33	**1.0**	6078.10	4705.0
VLM-guided final model	**5146.59**	**100.0%**	6.32	**1.0**	6073.74	4707.1

Table 5. Task 1 results: wildfire initially in the UAV's top-down FOV. All models achieve 100% visibility; differences arise mainly in total reward and path characteristics.

Model	R_{total}	FOV%	TTD (s)	(steps)	Dist (m)	Runtime (s)
Base PPO	4278.30	84.8%	669.23	596	3890.03	**4491.5**
VLM-only rew. shaping	3520.37	89.0%	494.21	432	2048.45	4576.2
VLM-int., top-down only	4674.29	91.3%	332.52	280	5218.33	4697.3
VLM-int., angled only	3910.21	88.2%	375.28	290	5436.07	4706.7
VLM-guided, unsegmented	5001.84	96.5%	182.69	116	**5721.53**	4708.1
VLM-guided final model	**5018.22**	**97.1%**	**180.05**	**113**	5710.59	4716.0

Table 6. Task 2 results: wildfire near the UAV (≈ 100 m) but initially outside the top-down FOV. The VLM-guided models detect the fire substantially faster and achieve higher FOV coverage.

nal model and the VLM-guided variant without segmentation suggests that dual-view VLM integration is the primary driver of robustness, with directional guidance providing an additional refinement.

B.5. Task 5: Robustness to Low Lighting

Task 5 again mirrors the Task 2 geometry but under low-light conditions to emulate nighttime or low-visibility operation. The VLM-guided final model achieves the highest total reward, highest FOV coverage, and fastest detection, indicating that VLM-based semantic cues help compensate for reduced RGB contrast. Notably, all VLM-integrated models outperform the baseline PPO in both time-in-FOV and time-to-detection, suggesting that wildfire-specific semantics (flame glow, plume structure) remain informative even when the overall scene is dark.

B.6. Qualitative Reward and Trajectory Trends

Beyond scalar metrics, these results show consistent qualitative patterns across Tasks 1–3. Reward curves exhibit three phases: an initial exploration phase, a stabilization phase as the agent converges toward the frontline, and a monitoring phase with relatively steady or increasing reward once the fire remains in view. Smoothed reward curves (50-step moving average) highlight this structure by reducing short-term variance.

Trajectory and altitude plots further confirm the effect of the shaped base reward: the agent typically ascends rapidly early in the episode, then slows its climb and emphasizes horizontal coverage as it begins tracking the fireline. Velocity traces show high variability during early exploration, followed by more stable speeds once the agent

Model	R_{total}	FOV%	TTD (s)	(steps)	Dist (m)	Runtime (s)
Base PPO	486.39	0.0%	None	None	3513.58	**4238.4**
VLM-only rew. shaping	314.82	0.0%	None	None	1596.23	4378.5
VLM-int., top-down only	629.13	0.0%	None	None	1902.76	4511.7
VLM-int., angled only	637.54	0.0%	None	None	2562.39	4496.5
VLM-guided, unsegmented	2953.47	42.7%	2684.19	2210	4484.91	4640.1
VLM-guided final model	**3612.46**	**45.1%**	**2492.54**	**2118**	**4875.80**	4656.4

Table 7. Task 3 results: wildfire at long range ($\approx$ 1 km). Only the two dual-camera VLM-guided models successfully detect and track the distant fire within 4000 timesteps.

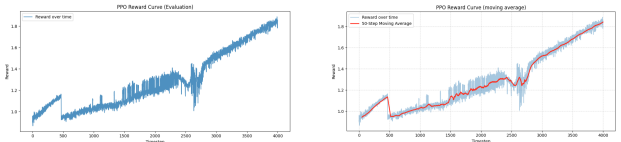
Model	R_{total}	FOV%	TTD (s)	(steps)	Dist (m)	Runtime (s)
Base PPO	4291.64	83.5%	676.90	602	4236.58	**4486.1**
VLM-only rew. shaping	3538.61	86.4%	492.76	429	2416.52	4572.3
VLM-int., top-down only	4690.39	88.9%	358.29	292	5505.19	4689.4
VLM-int., angled only	3891.02	86.5%	388.17	294	5617.43	4691.3
VLM-guided, unsegmented	5012.85	93.7%	187.34	120	**6030.84**	4706.9
VLM-guided final model	**5034.62**	**96.9%**	**183.61**	**115**	5925.30	4711.2

Table 8. Task 4 results: wildfire near the UAV with added cross-wind ($\approx$ 5 m/s). The VLM-guided models remain robust under plume-induced drift and environmental disturbance.

Model	R_{total}	FOV%	TTD (s)	(steps)	Dist (m)	Runtime (s)
Base PPO	4291.65	85.0%	601.84	595	3891.20	**4015.6**
VLM-only rew. shaping	3751.94	89.0%	443.75	429	2052.96	4056.0
VLM-int., top-down only	4859.16	91.5%	290.30	282	5216.71	4091.3
VLM-int., angled only	3992.37	88.8%	292.43	289	5432.15	4073.2
VLM-guided, unsegmented	5015.07	97.0%	117.90	112	**5710.34**	4102.6
VLM-guided final model	**5039.43**	**97.4%**	**112.96**	**108**	5686.39	4084.5

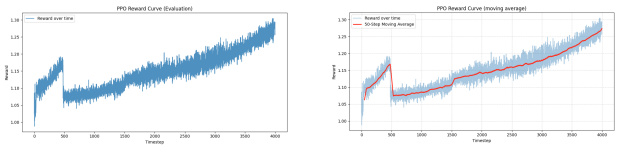
Table 9. Task 5 results: wildfire near the UAV under low-light conditions. The VLM-guided models maintain high visibility and rapid detection despite reduced RGB contrast.

locks onto the firefront. These extended results collectively show that the final VLM-guided model not only achieves higher quantitative scores but also exhibits more structured, interpretable flight behavior across diverse wildfire scenarios.



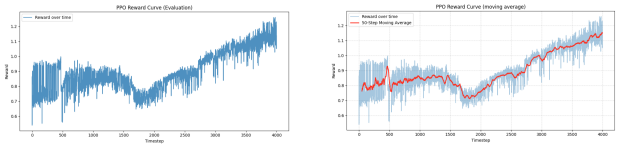
(a) Raw reward (b) 50-step moving average

Figure 8. Task 1 reward curves (wildfire in initial FOV).



(a) Raw reward (b) 50-step moving average

Figure 9. Task 2 reward curves (wildfire near UAV, outside initial FOV).



(a) Raw reward (b) 50-step moving average

Figure 10. Task 3 reward curves (wildfire at $\sim$1 km distance).

C. Reward Curves Across Tasks

We also report full reward trajectories for the final VLM-guided model over the three core tasks (wildfire in FOV, wildfire near UAV, wildfire in distance). For each task we plot both the raw per-timestep reward and a 50-step moving average to visualize the underlying trend.

Across all three tasks, the smoothed curves show a consistent pattern: an initial high-variance exploration phase, followed by a stabilization phase as the agent converges toward the firefront, and a monitoring phase with relatively steady reward once the wildfire remains in view.