Diffusion-Reinforcement Learning Hierarchical Motion Planning in Adversarial Multi-agent Games (Supplementary)

Zixuan Wu*, Sean Ye*, Manisha Natarajan* and Matthew C. Gombolay*

I. Environment Details

A. Terrain

The terrain has varying visibility levels, with dense forests in Prisoner Escape and wave height regions in Narco Interdiction that hinder the agents' detection abilities. The two domains are different in size, team composition and configuration. The Narco Traffic Interdiction domain is approximately three times as large as Prison Escape and with one more pursuit agent. The details of the environment is summarized at Table I-II where 1 grid represents 0.021 km:

TABLE I
PRISONER ESCAPE DOMAIN PARAMETERS

Parameter Name	Notation	Value
Size	s_p	2428×2428 (grid)
Start Region	l_{p0}	$2078 \sim 2428$ (grid)
Candidate Hideout Num	n_{ph}	20

TABLE II
NARCO INTERDICTION DOMAIN PARAMETERS

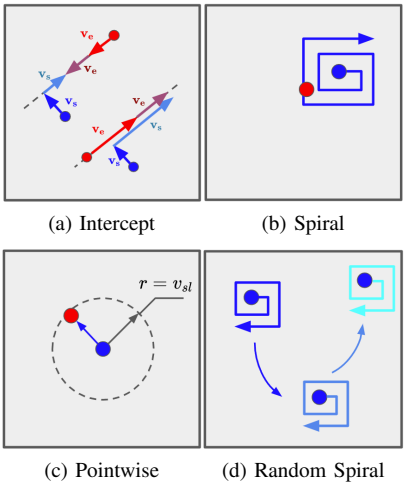
Parameter Name	Notation	Value
Size	s_n	7884×3538 (grid)
Start Region	l_{n0}	$0 \sim 350$ (grid)
Candidate Hideout Num	n_{nh}	6

B. Pursuit Policy

If the evader is detected nearby and within reaching distance, the search agent will directly move to the detected location using a pointwise policy (Figure 1c). Otherwise, the search agent will intercept by moving to the interception point perpendicular to the opponent agent's velocity and then towards the opponent's estimated location (Figure 1a). If the evader is not detected, the search agent will spiral around the estimated evader location for a set time (Figure 1b). If the evader is still not found, the policy is reset, and the search agent starts a spiral search at a random location (Figure 1d).

C. Team Composition and Parameters

We use two domains to test our diffusion-RL hierarchy: prisoner escape and narco interdiction domains. The pursuit team composition are different in the two domains. We have one helicopter and one search party in the prisoner escape domain and one plane and two search boats in the narco interdiction domain. The speed limits of each agent on the two domains are summarized at Table III:



(a) Intercept (b) Spiral

(c) Pointwise (d) Random Spiral

Fig. 1. Pursuit policies used in the search agent heuristics.

TABLE III
SPEED LIMIT (SL) OF EACH AGENT

Parameter Name	Parameter Notation	Value (km/h, grid/step)
Prisoner SL	s_{pri}	18.9, 15
Search Party SL	s_{sp}	25.2, 20
Helicopter SL	s_{heli}	160.0, 127
Smuggler SL	s_{sm}	37.8, 30
Search Boat SL	s_{sb}	50.4, 40
Plane SL	s_{pla}	315, 250

II. Training Details

In this section, we will briefly introduce the diffusion and RL training details.

A. Diffusion and RL

We will introduce the hyperparameters used in diffusion and RL training at the two different domains respectively. We train our diffusion model from RRT* waypoints dataset with 10000 downsampled paths and validate with 1000 paths each of which includes 10 waypoints. We also set a small weight decay coefficient as 5×10^{-4} to prevent over-fitting.

We will list the RL hyperparameters. Both DDPG and SAC are off-policy actor-critic RL methods and we set the policy and critic learning rate as 0.003. They are all using target critic with soft updating tricks to stablize the training process and we set the coefficient as 0.01. Since SAC

*All authors are associated with the Institute of Robotics and Intelligent Machines (IRIM), Georgia Institute of Technology, Atlanta, GA 30308, USA.

Correspondance Author: Zixuan Wu zwu380@gatech.edu

TABLE IV

DIFFUSION HYPERPARAMETERS (PRISONER ESCAPE)

Parameter Name	Notation	Value
waypoint number	N_w	10
learning rate	l_d	2e-5
weight decay	wd	5e-4
epoch number	N_d	100

TABLE V

DIFFUSION HYPERPARAMETERS (NARCO INTERDICTION)

Parameter Name	Notation	Value
waypoint number	N_w	10
learning rate	l_d	2e-6
weight decay	wd	5e-4
epoch number	N_d	150

has a self-adjusted entropy regularization term, the entropy learning rate, entropy target and regularization coefficient are set as 0.003, 3, 0.2 respectively.

TABLE VI

RL HYPERPARAMETERS (PRISONER ESCAPE)

Parameter Name	Notation	Value
critic lr.	l_c	0.003
policy lr.	l_p	0.003
delay coef.	s_c	0.01
entropy lr.	l_e	0.003
entropy target	e_{trgt}	3
entropy regularization	α	0.2
waypoint reach threshold	d_w	10
discount factor	γ	0.97
detection penalty	r_{adv}	-1
distance penalty coeff.	c_d	$1 \sim 0.05$
distance penalty	r_d	$-2.5 * c_d * d$
waypoint reach reward	r_g	13

TABLE VII

RL HYPERPARAMETERS (NARCO INTERDICTION)

Parameter Name	Notation	Value
critic lr.	l_c	0.003
policy lr.	l_p	0.003
delay coef.	s_c	0.01
entropy lr.	l_e	0.003
entropy target	e_{trgt}	3
entropy regularization	α	0.2
waypoint reach threshold	d_w	30
discount factor	γ	0.97
detection penalty coeff.	r_{adv}	$1 \sim 0.2$
distance penalty	r_d	$-5 * c_d * d$
waypoint reach reward	r_g	20

We summarize the hyperparameters on our two domains in Table IV-VII. We save the diffusion model checkpoint every 2 epochs and select the one that has the best performance in our validation set to help our RL. The detection penalty coefficient c_d is linearly decreasing and d is the distance to the next waypoint.

B. Costmap Construction

We summarize the parameters used in costmap construction in Table VIII-IX:

TABLE VIII

COSTMAP CONSTRUCTION (PRISONER ESCAPE)

Parameter Name	Notation	Value
Distance Threshold	ϵ	20 (grid)
Gaussian STD	σ	24 (grid)
Sample Num	N_s	30

TABLE IX

COSTMAP CONSTRUCTION (NARCO INTERDICTION)

Parameter Name	Notation	Value
Distance Threshold	ϵ	30 (grid)
Gaussian STD	σ	24 (grid)
Sample Num	N_s	30

III. TESTING DETAILS

We test our algorithms in both prisoner escape domain and narco interdiction domains. To have repeatable and persuasive benchmarking, we test on 100 trajectories with the fixed random seeds different from the training. We set the random seed as $1 \times 10^6 + 1 \sim 1 \times 10^6 + 101$ for 100 testing episodes and collect data accordingly. Here we also show a boxplot to intuitively illustrate the results we get in Figure 2.

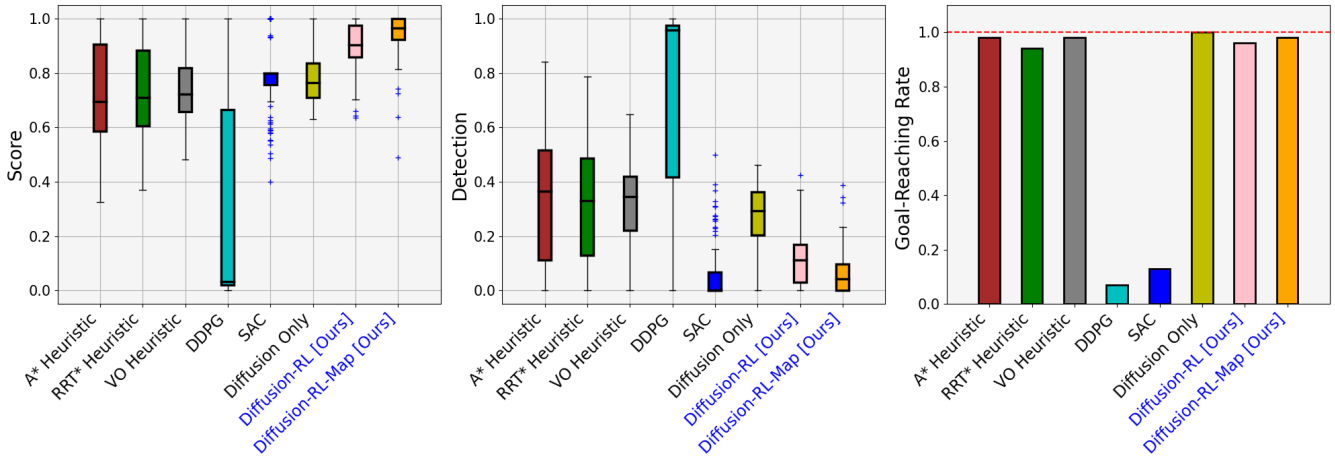


Fig. 2. Comparison of Score, Detection, and Goal-Reaching rate of our methods versus heuristics and learning based approaches.