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-IV where 1 grid represents 0.021 km:

TABLE I

Prisoner Escape Domain Parameters (prisoner has a global view of searching team)

Parameter Name	Notation	Value
Size	s_p	2428×2428 (grid)
Start Region	l_{p0}	$2078 \sim 2428$ (grid)
Hideout Num	n_{ph}	20
Visibility Model	$\alpha, \beta_{cam}, \beta_{heli}, \beta_{sp}, \eta$	$12, 0.5, 0.75, 0.75, 1$

TABLE II

Prisoner Escape Domain Parameters (prisoner has only local view of searching team)

Parameter Name	Notation	Value
Size	s_p	2428×2428 (grid)
Start Region	l_{p0}	$2078 \sim 2428$ (grid)
Hideout Num	n_{ph}	20
Visibility Model	$\alpha, \beta_{cam}, \beta_{heli}, \beta_{sp}, \eta$	$12, 0.5, 0.75, 0.75, 0.5$

TABLE III

Narco Interdiction Domain Parameters (smuggler has a global view of searching team)

Parameter Name	Notation	Value
Size	s_n	7884×3538 (grid)
Start Region	l_{n0}	$0 \sim 350$ (grid)
Hideout Num	n_{nh}	6
Visibility Model	$\alpha, \beta_{cam}, \beta_{pla}, \beta_{sb}, \eta$	$12, 2.0, 1.5, 1.5, 1$

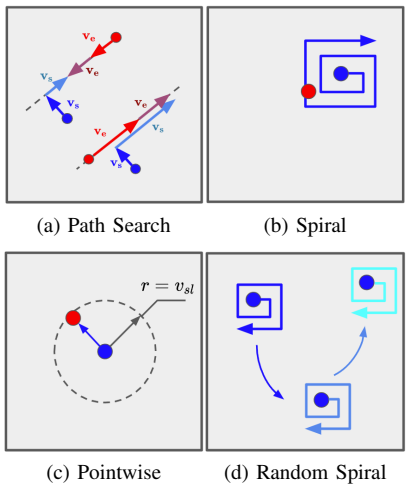
B. Pursuit Policy

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 search the predicted evading path 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).

TABLE IV

Narco Interdiction Domain Parameters (smuggler has only local view of searching team)

Parameter Name	Notation	Value
Size	s_n	7884×3538 (grid)
Start Region	l_{n0}	$0 \sim 350$ (grid)
Hideout Num	n_{nh}	6
Visibility Model	$\alpha, \beta_{cam}, \beta_{pla}, \beta_{sb}, \eta$	$12, 2.0, 1.5, 1.5, 1$



(a) Path Search (b) Spiral

(c) Pointwise (d) Random Spiral

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

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 V:

II. Training Details

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

*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 V

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

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.

TABLE VI

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 VII

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

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 smoothing coefficient as 0.01. Since SAC 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 VIII

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 IX

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 VI-IX. 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

In our implementation, ϵ and σ represents adjustable parameters that defines the risk range. A larger ϵ means a larger risk range the evader claims. A larger sigma implies a greater uncertainty regarding the risk locations, necessitating a broader area to be designated as risky. We summarize the parameters used in costmap construction in Table X-XI:

TABLE X

COSTMAP CONSTRUCTION (PRISONER ESCAPE)

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

TABLE XI

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 boxplots to intuitively illustrate the results we get for two domains in Figure 2-3.

In addition, we also benchmark the scenarios where the evader has only local view to the searching team. The results are shown on Table XII. Our method still outperforms baselines by at least 18.2% and 20.67% in prisoner escape and narco interdiction domain respectively.

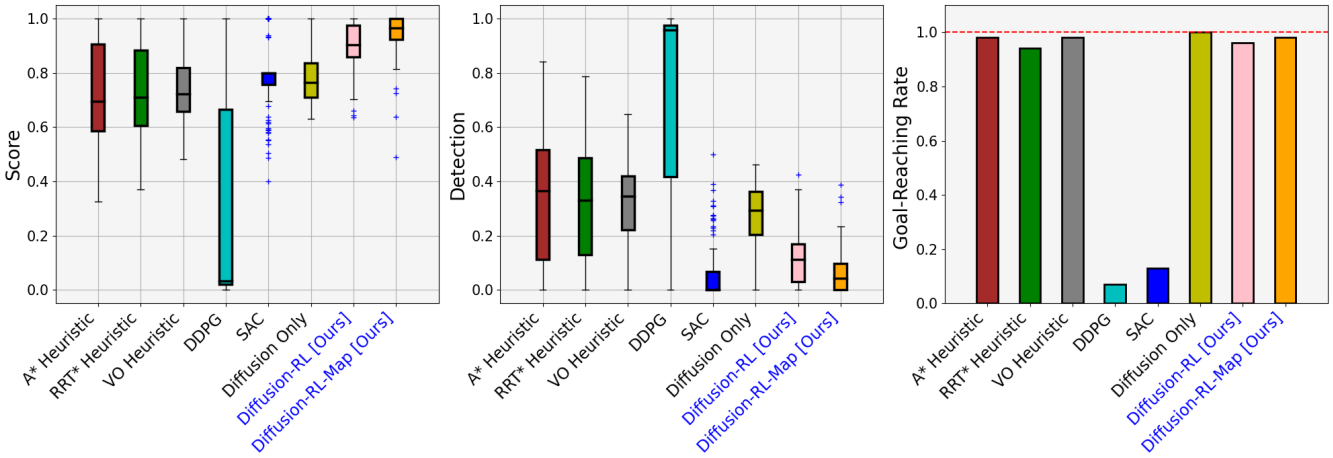


Fig. 2. Comparison of Score, Detection, and Goal-Reaching rate in the prisoner escape domain.

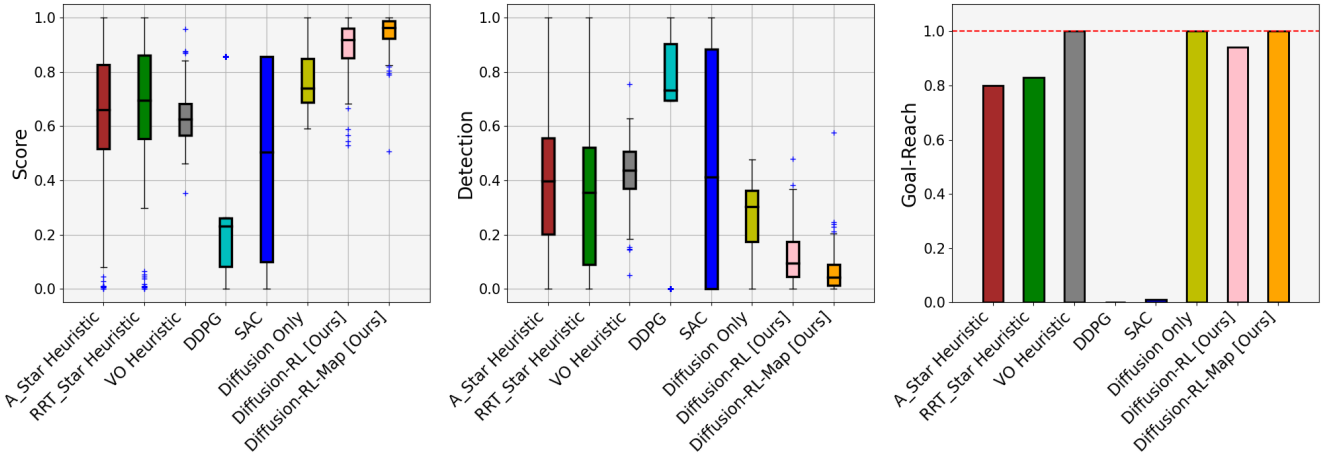


Fig. 3. Comparison of Score, Detection, and Goal-Reaching rate in the narco interdiction domain.

TABLE XII

DIFFUSION-RL BENCHMARKS [EVADER HAS ONLY LOCAL VIEW OF SEARCHING TEAM] (MEAN ± STD)

Domain	Prisoner Escape			Narco Interdiction		
	Score ↑	Detection ↓	Goal-Reach. ↑	Score ↑	Detection ↓	Goal-Reach. ↑
Non-Learning Approaches						
A-Star Heuristic	0.801±0.117	0.248±0.146	1.000±0.000	0.525±0.290	0.514±0.306	0.760±0.427
RRT-Star Heuristic	0.766±0.112	0.292±0.140	1.000±0.000	0.587±0.263	0.455±0.295	0.840±0.367
VO-Heuristic	0.731±0.121	0.334±0.154	0.990±0.099	0.636±0.095	0.425±0.111	1.000±0.000
Learning Approaches						
DDPG	0.231±0.337	0.729±0.400	0.070±0.255	0.260±0.356	0.697±0.416	0.000±0.000
SAC	0.731±0.123	0.098±0.147	0.050±0.218	0.506±0.402	0.410±0.469	0.000±0.000
Diffusion Only	0.789±0.103	0.264±0.128	1.000±0.000	0.779±0.129	0.258±0.150	1.000±0.000
Our Approaches						
Diffusion-RL [ours]	0.896±0.120	0.106±0.102	0.900±0.300	0.891±0.114	0.112±0.109	0.910±0.286
Diffusion-RL-Map [ours]	**0.947±0.086**	0.054±0.072	0.950±0.218	**0.940±0.075**	0.062±0.072	0.950±0.218