Code Paper

VVC-Gym: A Fixed-Wing UAV Reinforcement Learning Environment for Multi-Goal Long-Horizon Problems

Xudong Gong[1,2] Dawei Feng[1,2] Kele Xu[1,2] Weijia Wang[3]
Zhangjun Sun[3] Xing Zhou[4] Si Zheng[5] Bo Ding[1,2] Huaimin Wang[1,2]

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, Hunan, China
[2]State Key Laboratory of Complex & Critical Software Environment, Changsha, Hunan, China
[3]Flight Automatic Control Research Institute, AVIC, Xian, Shaanxi, China
[4]College of Intelligence Science and Technology, National University of Defense Technology,Changsha, Hunan, China
[5]Qiyuan Lab, Beijing, China

Outline

1. Motivation
2. VVC-Gym
3. Experiments
4. Discussion

Motivation

Multi-Goal Problems

A UAV must be capable of achieving not only the left-side goal but also the right-side goal

Long-Horizon Problems

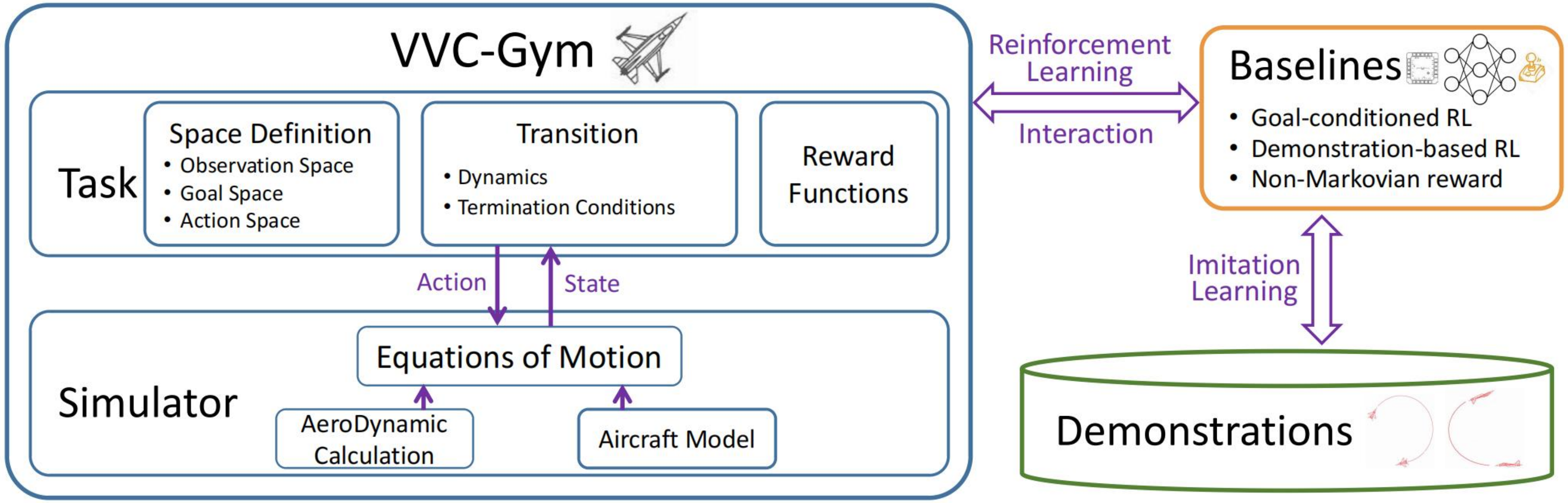
When completing an ascending turn, it is necessary to perform a horizontal turn first, then accelerate in a straight line, and finally climb in altitude (Long interaction sequence)

Challenge

- The **spatial complexity** of exploration: introduces the additional goal space that requires to explore
- The **temporal complexity** of exploration: the learning signal decreases exponentially with the horizon

Motivation

- Existing work predominantly focuses on the design of algorithms, neglecting the importance of **environment design** and the potential benefits that **demonstrations** can provide during training.
- To **facilitate study on multi-goal long-horizon problems**, we:
  - Provide the GCRL community with the first RL environment on realistic fixed-wing UAV's velocity vector control (VVC) task, **VVC-Gym**.
  - Conduct **ablation studies** on the environment design of VVC-Gym.
  - Equip VVC-Gym with multi-quality **demonstration** datasets.
  - Provide **baselines** on VVC-Gym and corresponding demonstrations



Outline

1. Motivation
2. VVC-Gym
3. Experiments
4. Discussion

RL Environment

- Problem Formulation

  Manipulating the UAV's velocity vector to match a desired velocity vector.

- Transition

  - 7 termination conditions are employed to avoid collecting ineffective samples

    - **R**each **T**arget Termination (RT)
    - **T**imeout termination (T)
    - **C**rash Termination (C)
    - **C**ontinuously **M**ove **A**way Termination (CMA)
    - **C**ontinuously **R**oll Termination (CR)
    - **E**xtreme **S**tate Termination (ES)
    - **N**egative **O**verload and **B**ig **R**oll Angle Termination (NOBR)

  - a general distance-based goal-conditioned reward is designed to facilitate effective learning

$$r_{g,t} = \begin{cases} 0, & \text{if triggers RT} \\ r_{penalty}, & \text{if triggers any of CMA, CR, C, ES, or NOBR} \\ -\left(\frac{\|\zeta(s_t)-g\|}{\sigma}\right)^b, & \text{else} \end{cases}$$

Demonstrations

- Demonstration Generating Method

1. Generating seed demonstrations with a PID controller
2. Augmenting demonstrations based on symmetry
3. Generating more and high-quality demonstrations through the IRPO[1] algorithm

- Demonstration quantity and quality

Demonstration	Number of trajectories	Goal space coverage (%)	Average length of trajectories	Number of transitions	Accuracy		
					$error_v$	$error_\mu$	$error_\chi$
$\mathcal{D}_E^0$	10184	20.08	282.01±149.98	2872051	6.56±3.25	0.36±0.35	0.53±0.45
$\overline{\mathcal{D}_E^0}$	10264	20.24	281.83±149.48	2892731	6.56±3.25	0.36±0.36	0.53±0.45
$\mathcal{D}_E^1$	24924	49.15	124.64±53.07	3106516	4.12±3.45	0.59±0.32	0.57±0.41
$\overline{\mathcal{D}_E^1}$	27021	53.28	119.64±47.55	3232896	4.47±3.49	0.58±0.32	0.60±0.44
$\mathcal{D}_E^2$	33114	65.29	117.65±46.24	3895791	4.83±3.45	0.57±0.33	0.66±0.54
$\overline{\mathcal{D}_E^2}$	34952	68.92	115.76±45.65	4045887	5.16±3.47	0.56±0.33	0.68±0.60
$\mathcal{D}_3$	38654	76.22	116.59±46.81	4506827	5.24±3.41	0.60±0.34	0.71±0.69
$\overline{\mathcal{D}_E^3}$	39835	78.55	116.56±47.62	4643048	5.29±3.38	0.60±0.35	0.74±0.75

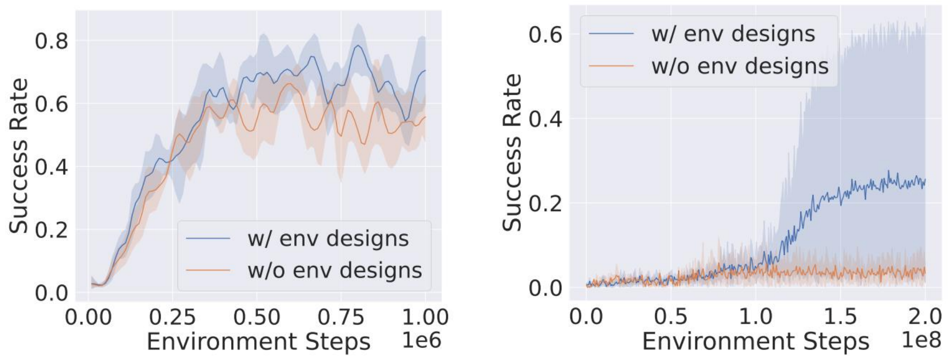
1. Xudong G, Dawei F, Xu K, et al. Iterative regularized policy optimization with imperfect demonstrations[C]//Forty-first International Conference on Machine Learning. 2024.

Outline

1. Motivation
2. VVC-Gym
3. Experiments
4. Discussion

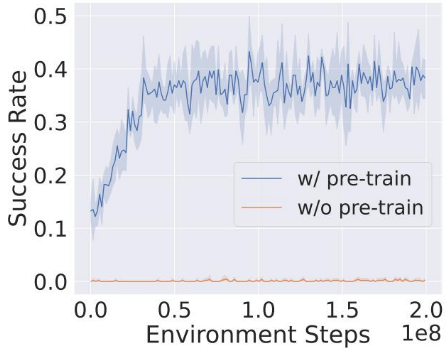
Main Results

- Evaluating the effectiveness of termination conditions and reward function
  - The termination conditions and the dense reward effectively facilitate more efficient training for GCRL algorithms



(a) SAC+HER (b) PPO

- Evaluating the benefits of demonstrations for GCRL training
  - Demonstrations facilities more efficient GCRL training



Experiments

Baselines

(a) Baselines on GCRL methods

RL type	Algorithm	Success rate
Off-policy	SAC	1.08±0.48
	HER	8.32±1.86
On-policy	PPO	0.04±0.03
	GCBC + PPO	38.31±1.62

- Both RL and GCRL algorithms perform poorly on VVC-Gym, suggesting **VVC-Gym poses a challenging multi-goal long-horizon task**

(b) Baselines on Curriculum methods

Curriculum	Success rate
None	38.31±1.62
RIG	49.03±1.54
DISCERN	49.36±1.91
MEGA	48.62±2.35

- self-curriculum methods can enhance learning effectiveness, indicating that **VVC-Gym is suitable for studying self-curriculum in GCRL**

(c) Baselines on demonstration-based methods

Demos	GCBC	GCBC + PPO
$\mathcal{D}_E^0$	17.08±0.57	38.31±1.62
$\mathcal{D}_E^1$	36.54±1.97	53.83±0.80
$\mathcal{D}_E^2$	41.79±0.44	68.47±1.20
$\mathcal{D}_E^3$	42.77±1.35	71.68±2.86

- both GCBC and GCBC+PPO exhibit improved policy performance as the quantity and quality of the demonstrations increase, suggesting that **VVC-Gym and the demonstrations are well-suited for studying demonstration-based RL**

Outline

1. Motivation
2. VVC-Gym
3. Experiments
4. Discussion

Discussion

Our Contributions

- We propose **VVC-Gym**, a fixed-wing UAV environment suited for researching multi-goal long-horizon problems.
- We equip VVC-Gym with multi-quality **demonstration datasets**.
- We provide **baselines** for GCRL, demonstration-based RL algorithms on VVC-Gym and its demonstrations.

Future Work

- Construct tasks with longer control sequences, including BFMs such as Slow Roll and Knife Edge
- Establish baselines for automatic sub-goal generation methods
- Explore methods for collecting low-cost demonstrations for velocity vector control tasks from human play data

Thanks for watching!

- Code is available at:
  - https://github.com/GongXudong/fly-craft
  - https://github.com/GongXudong/fly-craft-examples

Fly-Craft

Fly-Craft-Examples

- Happy to answer any questions by email:

gongxudong_cs@aliyun.com davyfeng.c@qq.com