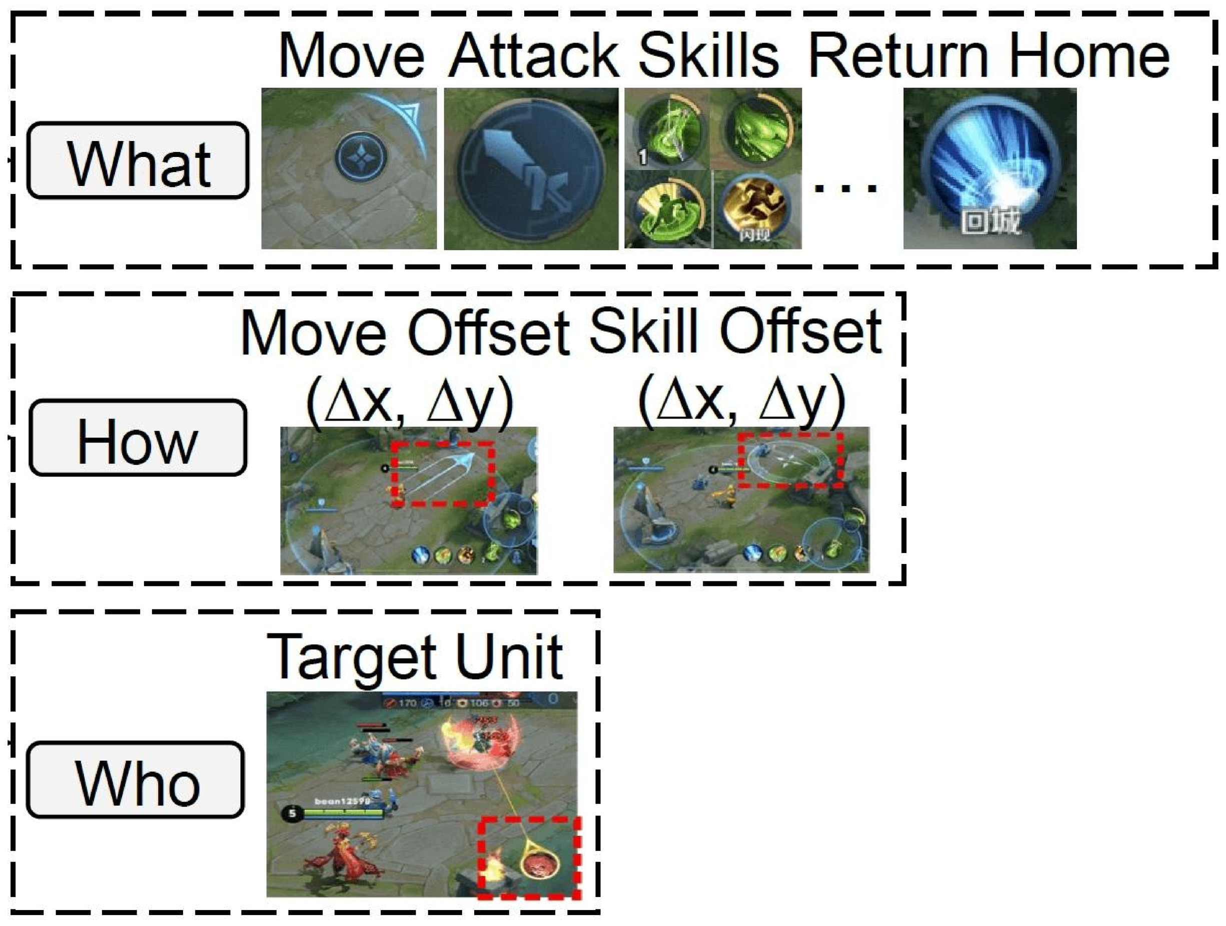
Mult-stage Phasic Policy Gradient for HoK 1v1

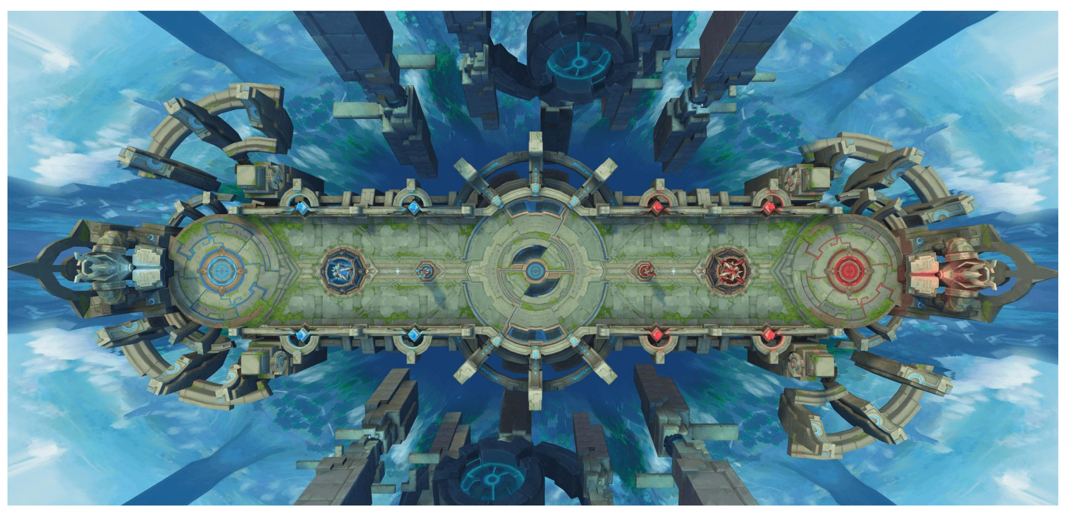
Yixing Chen, Songlin Jiang, Qihao Luo
ACM Honor Class, Shanghai Jiao Tong University

Background

HoK Game Agent



1v1 Solo Match



Task

In HOK 1v1 solo match, an agent needs to control one hero to fight the enemy hero, obtain money and exp through units, and destroy enemy tower.
The agent must master details like basic control action, skill combos while learning to gain advantage in the whole game and achieve victory.

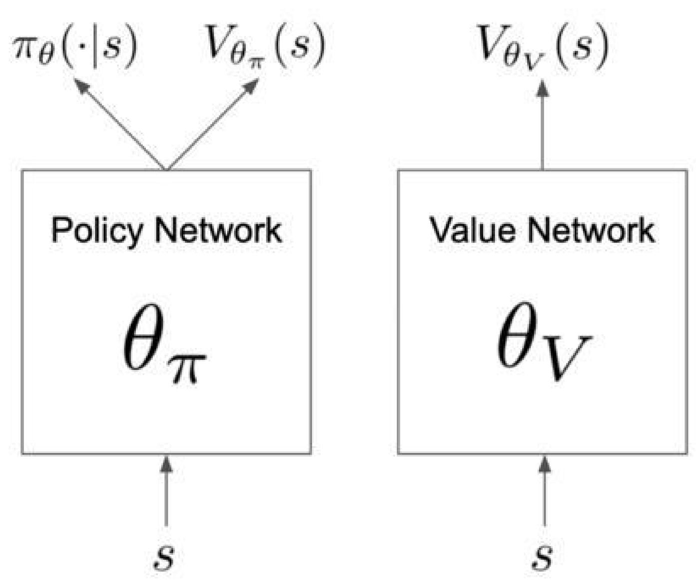
Challenges

High Action Dimension: agent must handle various kinds of actions that concern positions and targets.
Complex Skill Combinations and Environment: agent must learn to use different heroes by mastering their skill sets and deal with the enemy.
Long Range Optimization: agent must learn how to gain and maintain advantage or to reverse enemy's advantage and seek a comeback.

Our Approach

Phasic Policy Gradient (PPG)

During the policy phase, we adopt the dual-clip PPO objective to update the policy. After a fixed number of policy phase updates, we switch to the auxiliary phase.
In the auxiliary phase, the model is updated using data cached during the policy phase. Two losses are optimized: (1) a mean squared error (MSE) loss between the predicted value and a stored target value to improve value function accuracy, and (2) a Kullback-Leibler (KL) divergence loss to maintain policy consistency.



$$\mathcal{L}_{\text{value}} = \frac{1}{2}\|V_\theta(s_t) - V_{\text{target}}\|^2$$

$$\mathcal{L}_{\text{KL}} = \text{KL}\left[\pi_{\theta_{\text{old}}}(\cdot \mid s_t) \parallel \pi_\theta(\cdot \mid s_t)\right]$$

$$\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{value}} + \beta \cdot \mathcal{L}_{\text{KL}}$$

Advantage Normalization

$$\hat{A} = \frac{A - \mu_A}{\sigma_A + \epsilon}$$

Reward and Curriculum Design

Intuition: design a suitable set of rewards and curriculum that makes agent training stable and achieves better performance.

Reward	Weight	Time Scaling Factor	Meaning
hp_point	5.0	1.001	the health point of the hero
tower_hp_point	10.0	1.002	the health point of the turret
money	0.007	0.9998	money gained
exp	0.007	0.9998	experience points gained
ep_rate	0.75	1.0	the rate of mana
death	-1.2	1.0015	being killed
kill	-0.6	0.995	killing an enemy hero
last_hit	0.6	1.001	striking the last hit to an enemy unit
forward	0.05	1.001	moving forward
ult_hit	5.0	1.0015	hitting an enemy hero with ultimate skill

Experiments

Setup

Method: 1v1 solo match.
Metrics: to measure the overall performance of agent in a 1v1 solo match in terms of win rate and K/D.

- K/D: The ratio of agent's kill and death averaged over all games.

Compared Models

- Baseline 1: Trained for 3h, 10000 steps.
- Baseline 2: Trained for 5h, 20000 steps.
- Baseline 3: Trained for 8h, 30000 steps.

Results

Model	Win Rate (%)	K/D
Baseline-1	90	3.62
Baseline-2	68	2.16
Baseline-3	12	0.75

PPG + Adv Norm	Multi-stage Reward	Win Rate (%)	K/D
✓	✓	90	3.62
✓	✗	68	3.64
✗	✓	36	0.85
✗	✗	8	0.44