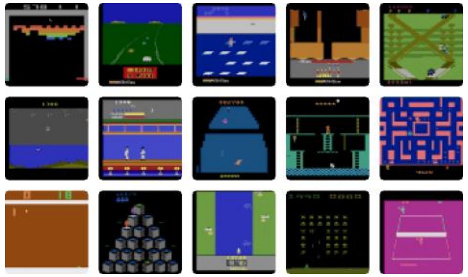
Grounded Answers for Multi-agent Decision-making Problem through Generative World Model

Zeyang Liu · Xinrui Yang · Shiguang Sun ·
Long Qian · Lipeng Wan · Xingyu Chen* · Xuguang Lan*

National Key Laboratory of Human-Machine Hybrid Augmented Intelligence
National Engineering Research Center for Visual Information and Application
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
*: Corresponding authors

Multi-agent reinforcement learning is an important way to solve the optimized decision-making of complex intelligent systems.

Go



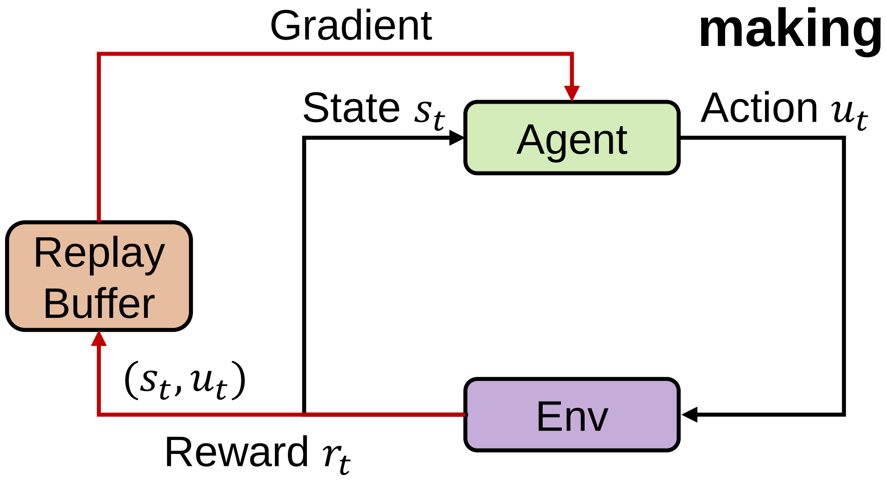
Video Game

Smart Logistics

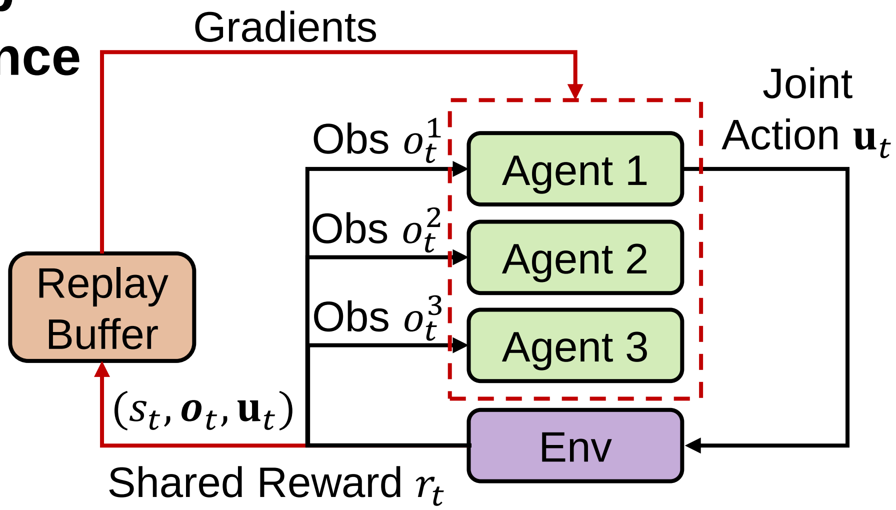
Traffic Control

Decision-making

Group Intelligence



Single-agent Reinforcement Learning



Multi-agent Reinforcement Learning

The current offline policy generation methods use pessimistic estimation and conditional sequence generation. However, they cannot find the correct answers through trial and error like humans.

Image from the scene

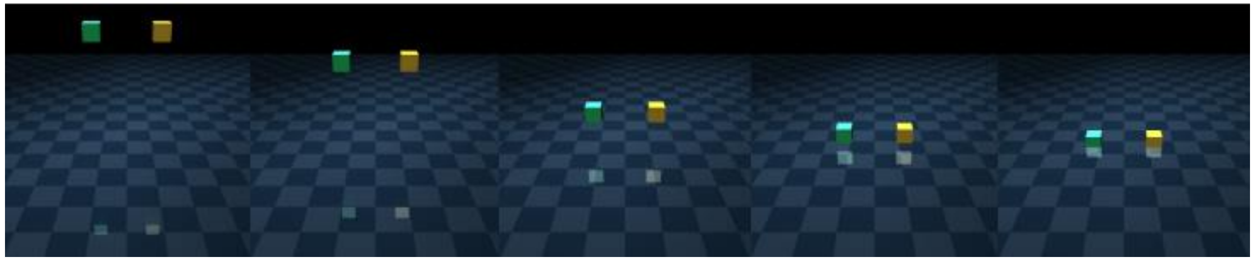
Question

Let's say we're controlling five marines on the left. What strategy should we use to defeat the six enemy marines on the right?

GPT-4 Output

1. Analyze the Enemy Composition
2. Positioning and Formation
3. Focus Fire
4. Utilize Abilities Strategically
5. Adapt to Enemy Movements
6. Retreat and Heal

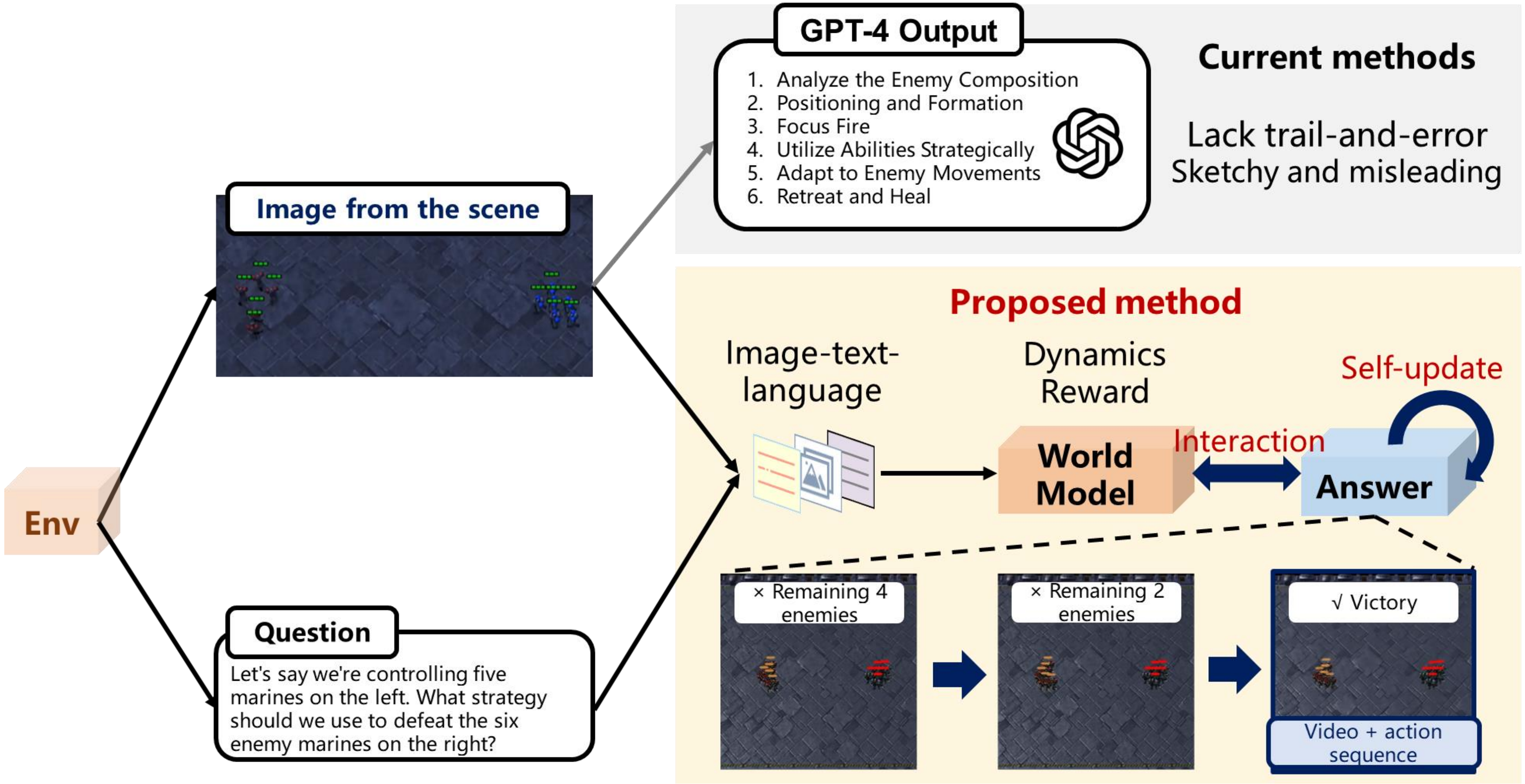
Sketchy and misleading



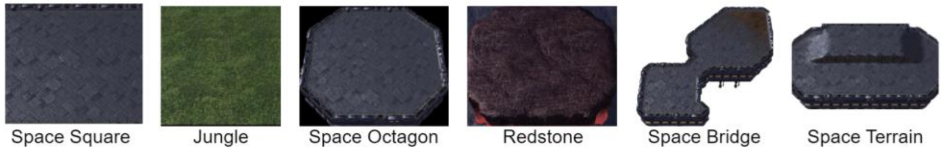
Limited to addressing issues related to physical facts and cannot handle decision-making problems

Only for single-agent tasks with manually crafted reward functions, making them incapable of handling multi-agent coordination tasks

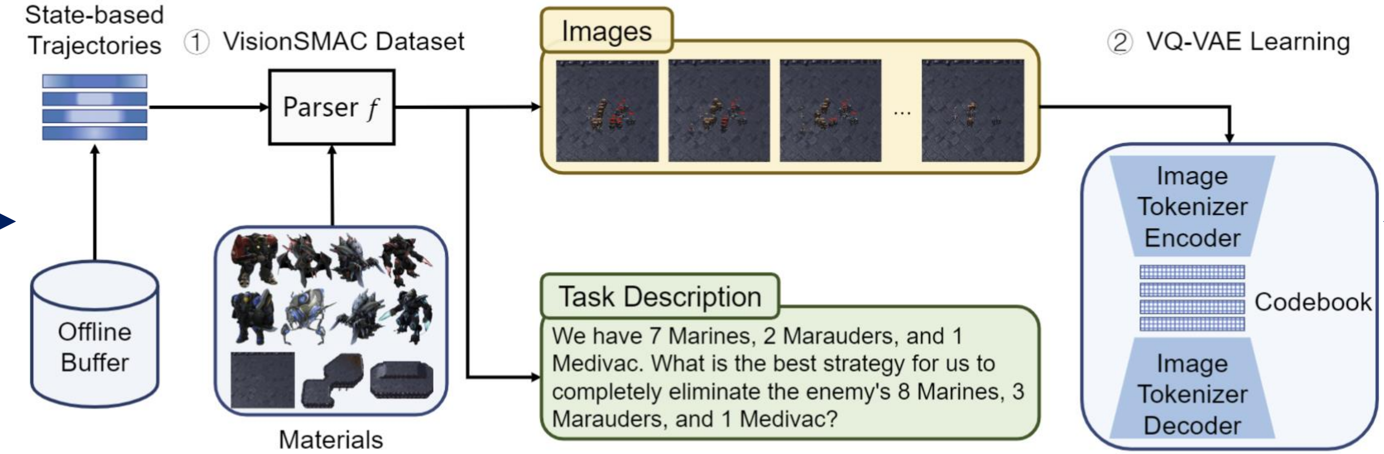
To enhance response quality for decision-making problems, we can integrate multi-agent reinforcement learning with offline policy learning in world models



VisionSMAC: we convert the state into images and languages through a parser f, decoupled from StarCraft, making it easy to create new contents

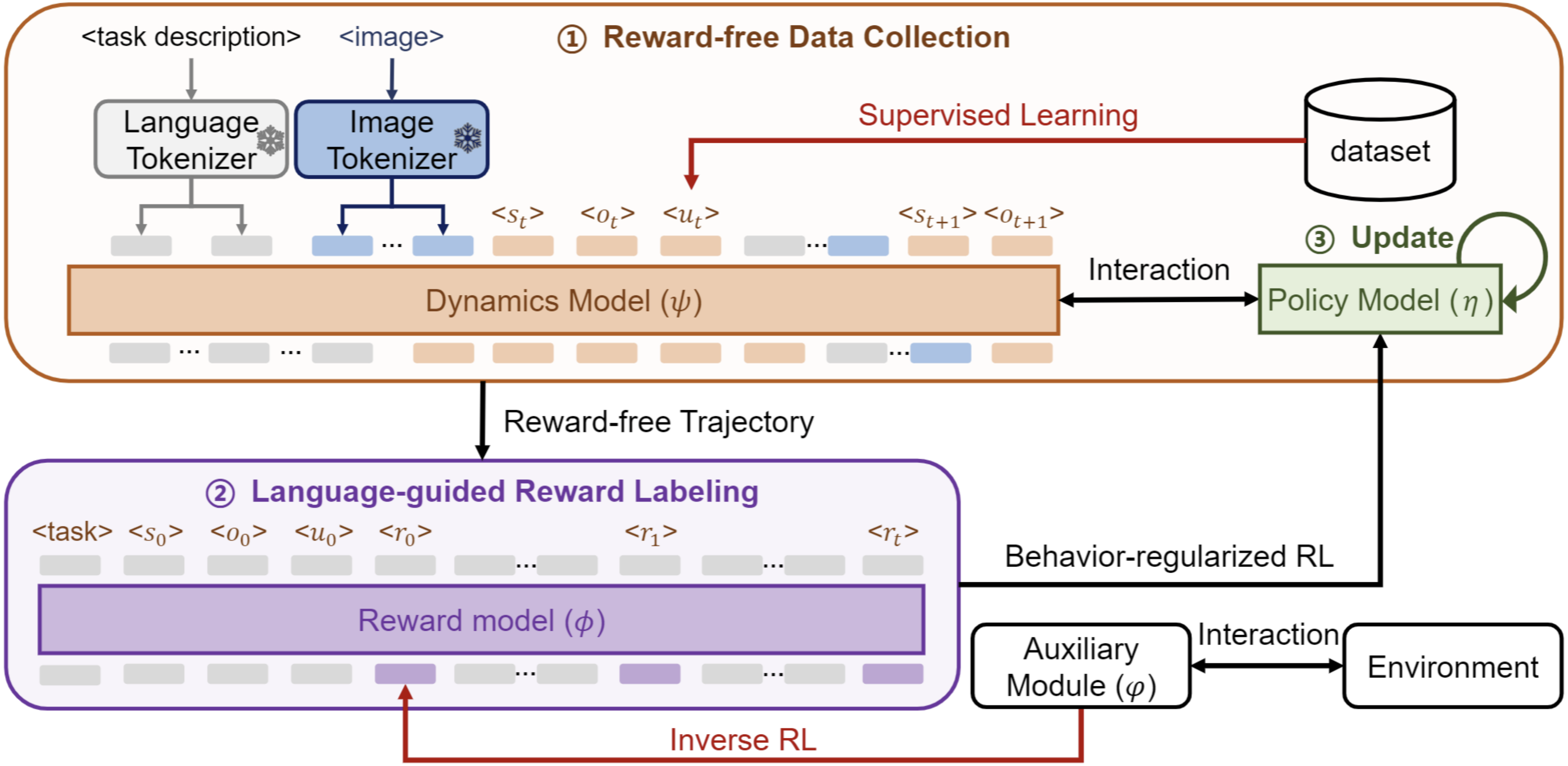


Map Name	Return Distribution	Map Name	Return Distribution
3s5z	19.43 ± 1.86	5m_vs_6m	19.83 ± 2.16
1c3s5z	19.66 ± 1.25	6h_vs_8z	18.84 ± 2.09
10m_vs_11m	19.75 ± 1.03	3s5z_vs_3s6z	19.76 ± 1.26
2c_vs_64zg	19.98 ± 0.71	corridor	19.69 ± 1.48
3s_vs_5z	19.88 ± 1.40	MMM2	19.63 ± 2.07



Interactive Simulator: (1) Image Tokenize, (2) Dynamics Model, and (3) Reward Model.

Inference: Learning Policy in the Simulator



Dynamics Model (causal transformer)

$$x = \{\dots, z^L, z_t^I, s_t, o_t^1, \dots, o_t^n, u_t^1, \dots, u_t^n, z_{t+1}^L, z_{t+1}^I, s_{t+1}, \dots\}$$

$$\Delta s_{t+1} = s_{t+1} - s_t$$

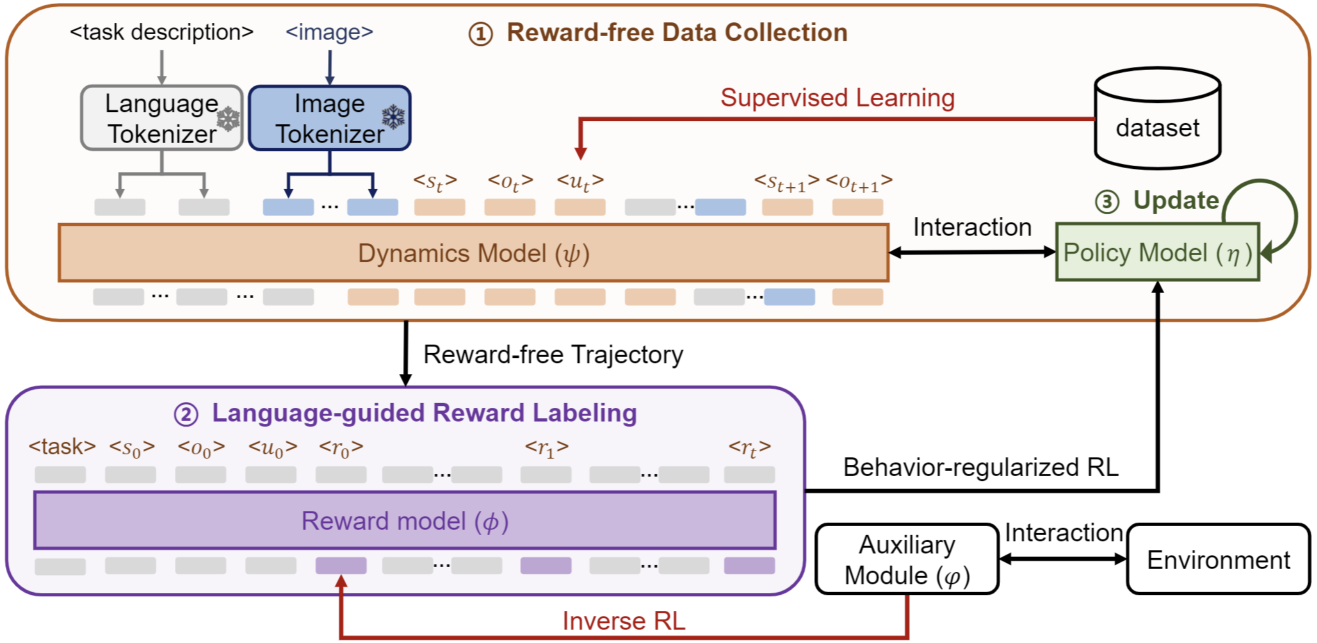
Residual Dynamics

Reward Model (bidirectional transformer)

$$\tilde{x} = \{\dots, s_t, z^L, \widehat{r_t^i}, \dots, \widehat{r_t^n}, \hat{r}_t, s_{t+1}, \dots\}$$

$$\nabla_\phi \mathcal{L} = E_{\tau\sim\pi^\theta}\left[\sum_i^n \sum_t \gamma^t \nabla_\phi \widehat{r_t^i}(\tau, u_i^t;\phi)\right] - E_{\tau\sim D}\left[\sum_i^n \sum_t \left(\gamma^t \nabla_\phi \widehat{r_t^i}(\tau, u_i^t;\phi) - |\widehat{r_t^i}(\tau;\phi)|_2^2\right)\right]$$

Reward Constraint



Learning Policy in the Simulator

$$\max_{\bar{\pi}} E\left[\sum_{t=0}^{\infty}\gamma\left(\sum_{i=1}^{n}\left(\widehat{r_t^i}(\tau;\phi)-\alpha\cdot\log\left(\frac{\bar{\pi}_i\left(u_t^i|o_t^i;\eta\right)}{q_s\left(u_t^i|x_{<u_t^i};\psi\right)}\right)\right)\right)\right]$$

Behavior Regularization

① Generate reward-free trajectory by interacting with the dynamics model
② Generate reward for each state-action pair using the reward model
③ Update agents using any off-policy MARL algorithm

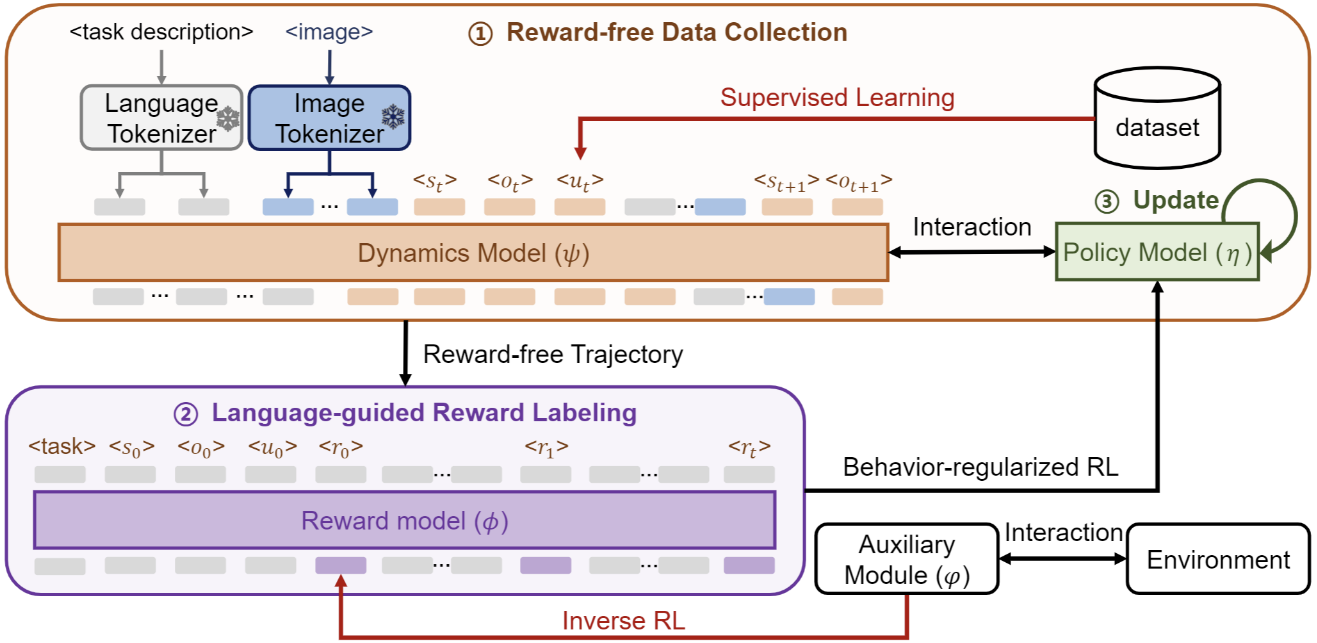


Table 1: Test win rates (%) and standard deviations compared with reward-free imitation learning methods.

Map Name	BC	MA-AIRL	MADT	MAPT	MA-TREX	LBI
1c3s5z	16.44± 1.35	7.88± 2.49	61.35± 7.26	74.77± 5.15	64.76± 11.62	94.59± 3.41
10m_vs_11m	26.19± 4.42	41.69± 7.12	82.76± 4.41	66.85± 9.28	48.78± 11.28	90.45± 6.99
2c_vs_64zg	17.37± 10.12	24.75± 10.83	61.90± 5.74	58.28± 7.84	22.45± 7.74	71.44± 8.83
3s_vs_5z	0.00± 0.00	0.05± 0.03	80.90± 0.45	72.33± 3.93	55.38± 18.03	92.82± 6.25
5m_vs_6m	13.78± 2.15	11.59± 6.75	79.78± 4.98	56.01± 3.17	50.01± 14.87	87.98± 5.10
6h_vs_8z	9.28± 5.06	16.47± 8.08	30.94± 25.54	37.16± 6.27	28.38± 5.31	66.61± 4.57
3s5z_vs_3s6z	0.00± 0.00	0.00± 0.00	27.44± 9.49	34.90± 6.84	36.16± 3.68	83.34± 4.27
corridor	0.00± 0.00	0.76± 0.15	69.85± 1.54	45.91± 15.47	30.59± 9.86	87.45 ± 2.94
MMM2	0.00± 0.00	0.00± 0.00	54.34± 12.83	19.21± 5.59	21.52± 6.58	95.96± 4.65

Table 2: Test return and standard deviations compared with offline reinforcement learning methods.

Map Name	BCQ-MA	CQL-MA	ICQ	OMAR	OMIGA	LBI
5m_vs_6m	9.13± 0.21	10.15± 0.15	9.47± 0.45	8.76± 0.52	10.38± 0.50	18.96± 0.56
2c_vs_64zg	18.86± 0.35	19.20± 1.25	18.47± 0.25	17.10± 0.94	19.25± 0.38	20.45± 0.25
6h_vs_8z	11.91± 0.44	9.95± 0.32	11.55± 0.15	9.74± 0.28	12.74± 0.21	18.97± 0.28
corridor	16.42± 1.55	6.64± 0.90	16.74± 1.78	8.15± 0.89	17.10± 1.33	19.50± 0.73

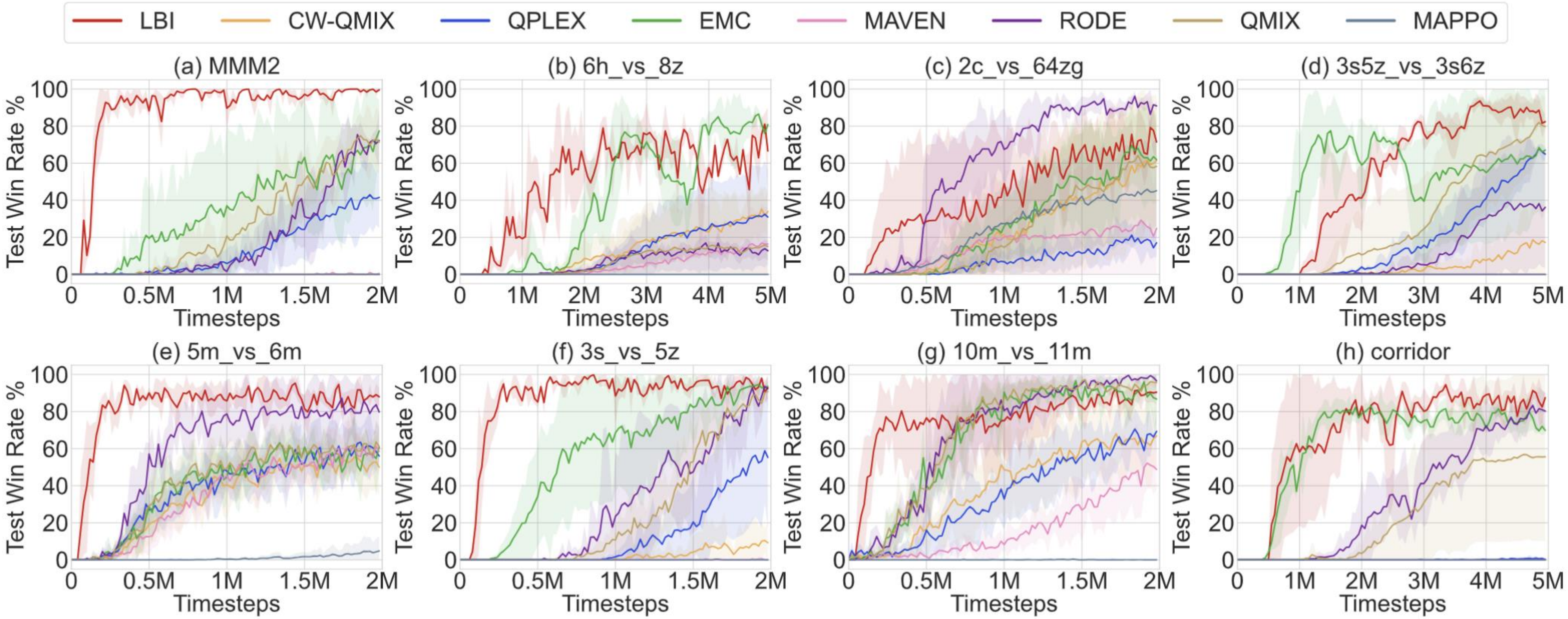


Table 3: Test win rates (%) and standard deviations on unseen tasks.

Unseen Task	MADT	MA-TREX	LBI	Unseen Task	MADT	MA-TREX	LBI
1c3s	16.21± 5.38	23.53 ± 8.83	56.47± 5.63	1c2s7z	6.16± 3.09	5.69±3.81	28.26± 6.41
6m	49.28± 4.06	37.12±2.59	97.85± 2.15	6m_vs_7m	73.45± 7.22	32.88±4.47	81.07± 5.17
1c_vs_32zg	2.08± 1.51	11.41±3.41	58.33± 6.44	3s4z	90.21± 1.82	79.71±3.56	87.55± 1.76
3s2z_vs_2s3z	0.00± 0.00	9.16±5.62	18.22± 2.46	3s5z_vs_3s7z	10.21± 3.66	15.88±4.34	22.08± 7.63
1c3s6z	16.41± 6.44	58.09±3.41	65.38± 5.12	9m_vs_11m	76.44± 4.17	70.91±6.95	75.05± 2.16

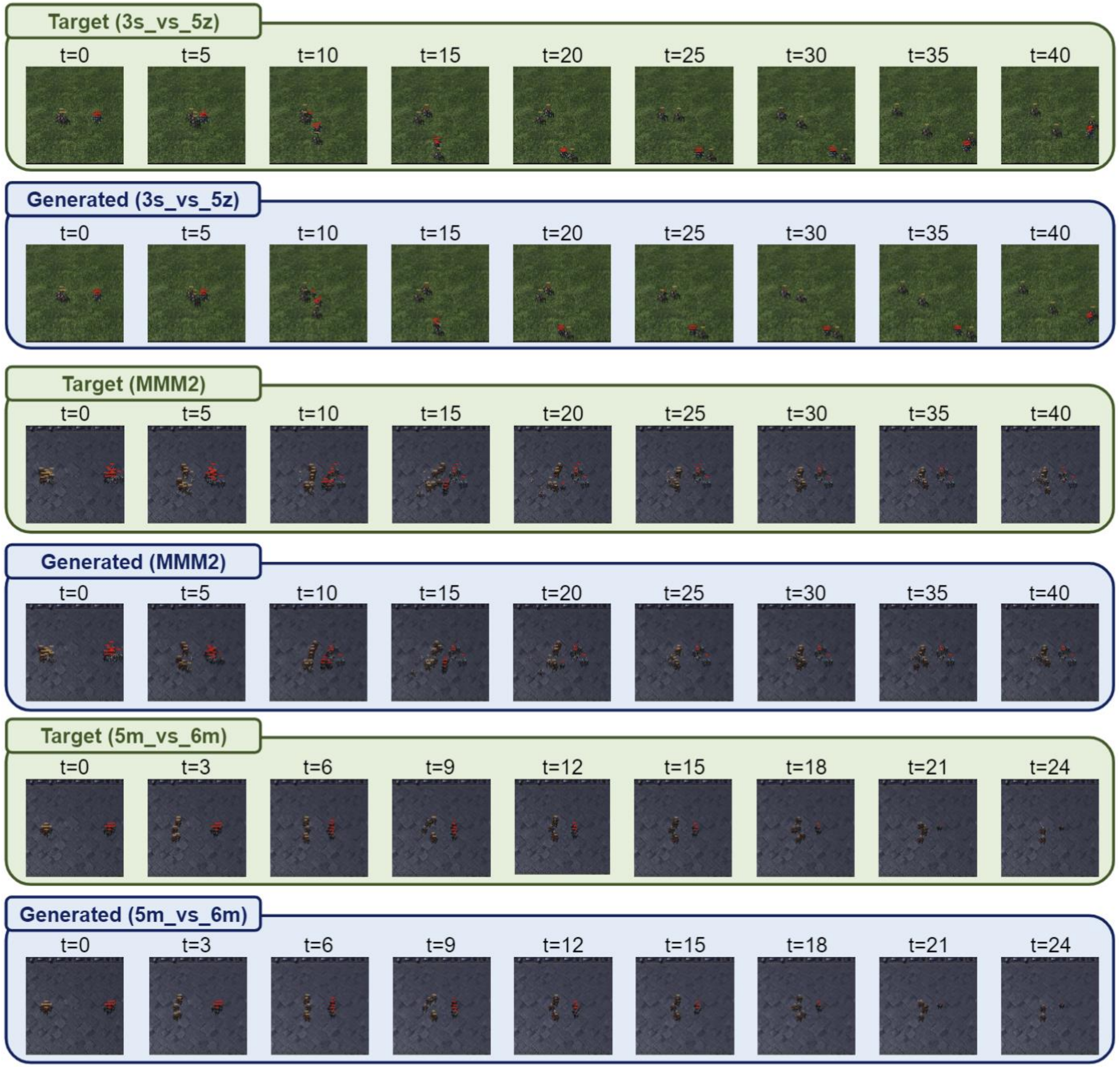
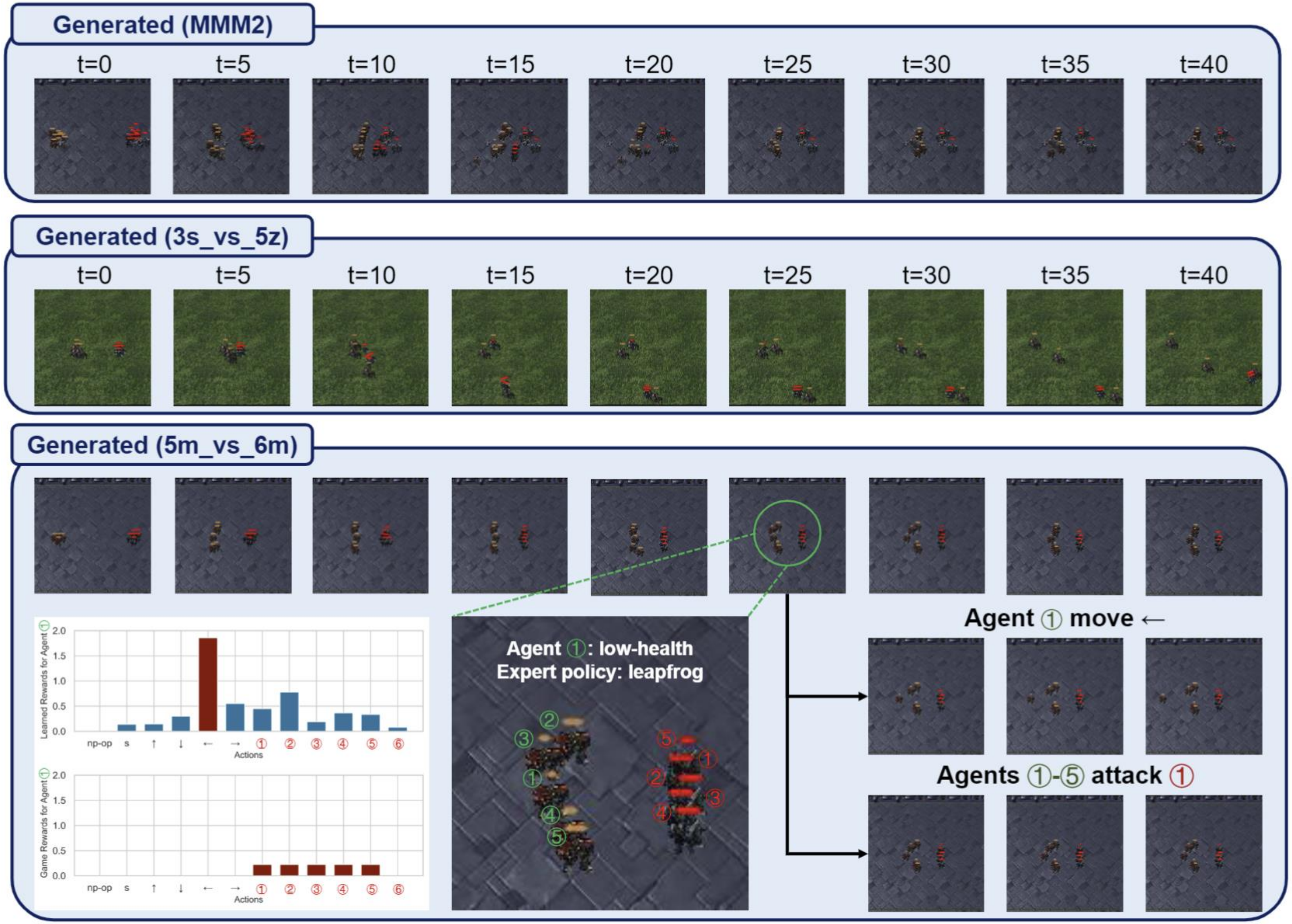


Table 4: The ablation results for the dynamics model without residual term (wo-RT), image reference (wo-IR), and using ground-truth image (GTI) as the reference for state prediction.

Algorithm	Prediction error	Return (all)
LBI	0.016 ± 0.023	18.91 ± 1.33
LBI-GTI	0.014 ± 0.018	18.98 ± 0.89
LBI-wo-RT	0.434 ± 0.351	14.25 ± 1.84
LBI-wo-IR	0.029 ± 0.041	18.63 ± 1.01
LBI-wo-RT&IR	0.744 ± 1.164	12.13 ± 2.33

Table 5: The ablation results for the reward model without reward constraint (wo-RC), behavior regularization (wo-BR), and using ground-truth rewards (w-GTR) provided by the SMAC benchmark.

Algorithm	Return (training)	Return (unseen)
LBI	19.47 ± 0.77	18.54 ± 1.49
LBI-GTR	16.68 ± 1.55	14.07 ± 2.79
LBI-wo-RC	17.85 ± 0.59	14.75 ± 1.67
LBI-wo-BR	18.82 ± 1.28	17.46 ± 2.01
LBI-wo-RC&BR	12.35 ± 2.38	9.83 ± 1.46


Generated (MMM2)
t=0
t=5
t=10
t=15
t=20
t=25
t=30
t=35
t=40
Generated (3s_vs_5z)
t=0
t=5
t=10
t=15
t=20
t=25
t=30
t=35
t=40
Generated (5m_vs_6m)
Learned Rewards for Agent ①
Game Rewards for Agent ①
Actions
np-op
s
↑
↓
←
→
①
②
③
④
⑤
⑥
Agent ①: low-health
Expert policy: leapfrog
Agent ① move ←
Agents ①-⑤ attack ①