FOUNDER: Grounding Foundation Models in World Models for Open-Ended Embodied Decision Making

Yucen Wang[1 2], **Rui Yu**[1 2], **Shenghua Wan**[1 2], **Le Gan**[1 2], **De-Chuan Zhan**[1 2]

[1]School of Artificial Intelligence, Nanjing University, China
[2] National Key Laboratory for Novel Software Technology, Nanjing University, China

Introduction

- The Vision: Truly General-Purpose Robots
  - Open-ended multi-modal embodied task interpretation and solving
  - Instruct agents with natural language or videos
  - Adapt to out-of-distribution tasks seamlessly
  - Learn without needing hand-crafted rewards for every single task

- Using Foundation Models (FMs) to interpretate the multi-modal tasks?
  - FMs are **not grounded** into the embodied domain and the physical world

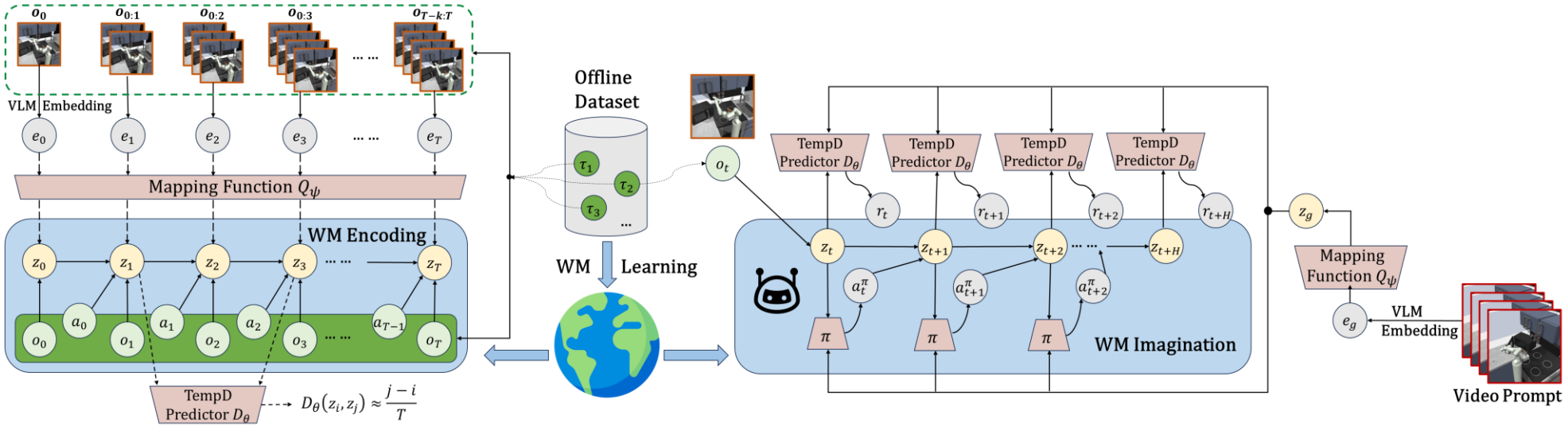
Introduction

- The Gap: High-Level "Brains" vs. Low-Level "Physics Engines"
  - Foundation Models - The "Scholar"
    - **Strengths**: Rich world knowledge, understands complex text/video prompts
    - **Weakness**: **Not grounded** in the physical world. Doesn't know "how to act".
  - World Models - The "Artisan"
    - **Strengths**: Models physical dynamics, efficient for control via imagination.
    - **Weakness**: Cannot understand open-ended tasks. Requires per-task reward engineering.

- **Fundamental Question:** How can we bridge the gap between high-level semantic understanding and low-level physical control?

Method Overview

- Our Solution: Building a Bridge between FMs and WMs with FOUNDER
  - We ground FM task representations into actionable goal states within the WM.
  - This enables open-ended task solving in a reward-free, model-based manner.
  - **Task Input (Text/Video) → FM → Task Representation**
  - **Task Representation → Mapping Function → Goal State in WM**
  - **Agent → [WM Imagination + Temporal-Distance-Based Rewards] → Goal-Conditioned Policy**



Problem Setting

- The agent is given
  - A Reward-free trajectory offline dataset, consisting of observations (images) and actions, pre-collected from the target embodied environment
  - A Vision Language Model (VLM)
- The agent cannot interact with the environment or obtain the ground-truth reward

- Offline + Visual + Reward-Free + Multi-Task

Method

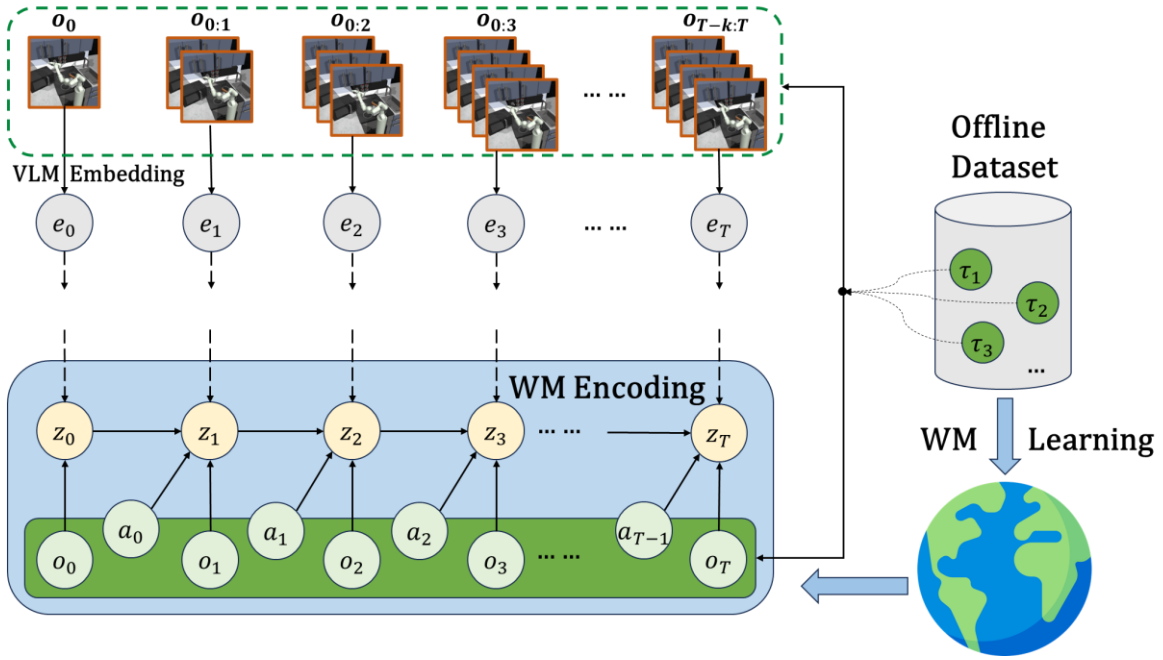
- Phase 1: Pretraining the WM and the Mapping Function
  - We first learn a Dreamer-V3 style WM trained on the offline dataset
  - Then we learn a mapping function: $\hat{z} \sim Q(\cdot \mid e)$ by aligning the embeddings of VLM and WM on the offline dataset, using an auto-encoder structure

$$e_t = VLM(o_{t-k:t}), t = k, \cdots, T$$
$$z_t \sim WM(\cdot \mid o_{\leq t}, a_{<t}), t = k, \cdots, T$$

$$\min_{Q_\psi, P_\psi} \sum_{t=1}^{T} \mathbb{D}_{\mathrm{KL}}[Q_\psi(\cdot \mid e_t) \parallel WM(\cdot \mid o_{\leq t}, a_{<t})] + \mathbb{E}_{\hat{z}_t \sim Q_\psi(\cdot \mid e_t)}[-\ln P_\psi(e_t \mid \hat{z}_t)]$$

  - External VLM embeddings e_g can be mapped into corresponding WM goal states: $z_g \sim Q_\psi(\cdot \mid e_g)$

The mapping function translates the abstract "what" of a task (VLM representation e) into a concrete "where" in the world model's state space (the inferred corresponding WM state $\hat{z}$)

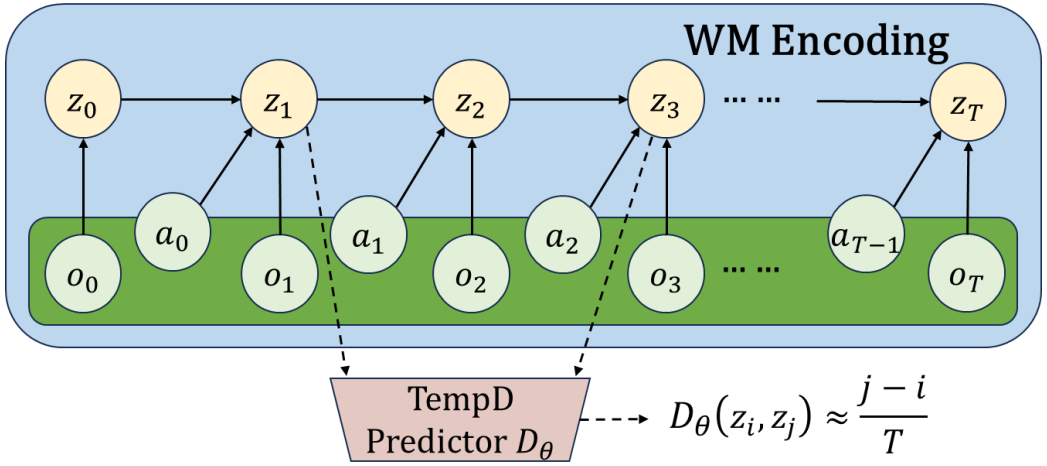


Method

- Phase 2: Behavior Learning
  - Once we have a goal state in the WM, all we need is to specify the reward function used to guide policy learning towards the goal.
  - We propose using temporal distance as the reward signal.
  - We learn a temporal distance prediction model within the WM that predicts how many steps it takes to get from one state to another. This is also done in pretraining.

$$\min_{D_\theta} MSE(D_\theta(z_t, z_{t+c}), \frac{c}{T})$$

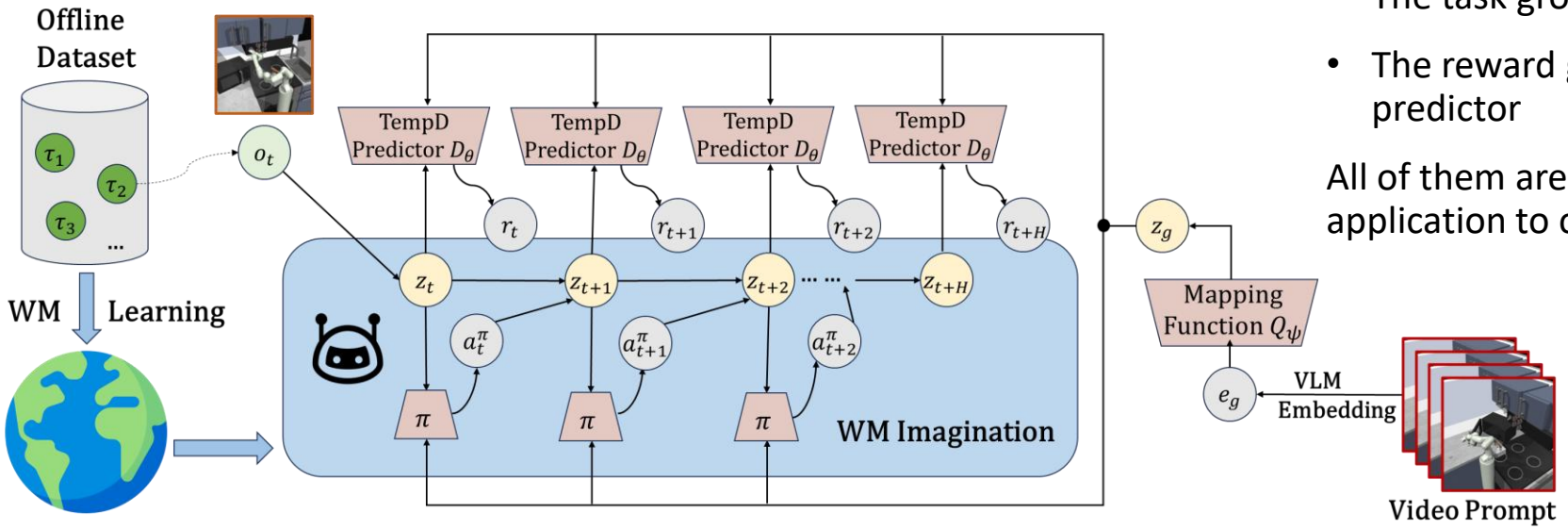
$$\min_{D_\theta} MSE(D_\theta(z^i, z^j), 1)$$



Method

- Phase 2: Behavior Learning
  - Goal-Conditioned policy learning in the WM through imagination
  - using the predicted temporal distance between current state and the goal state as reward

$$r_D(z_t, z_g) = -D_\theta(z_t, z_g)$$



Action model: $a_t \sim \pi(a_t \mid z_t, z_g)$

Value model: $v(z_t, z_g) \approx \mathbb{E}_{\pi,\mathrm{WM}}\left[\sum_{k=t}^{H} \gamma^{k-t} r_D(z_t, z_g)\right]$

The behavior learning stage integrates:

- the high-level knowledge in the VLM
- The environment modeling capability of the WM
- The task grounding ability of the mapping function
- The reward generation of the temporal distance predictor

All of them are task-agnostic, allowing for their potential application to open-ended, multimodal downstream tasks

Experiments: Language Tasks

Superior Performance: Consistently outperforms prior methods on multi-task visual control benchmarks.

Table 1. Normalized test performance of FOUNDER and baselines on DMC and Kitchen benchmarks. Mean scores (higher is better) with standard deviation are recorded across 6 seeds for each task.

Task	GenRL	WM-CLIP	GenRL-TempD	FOUNDER w/o TempD	FOUNDER
Cheetah Stand	0.93 ± 0.03	0.93 ± 0.03	0.42 ± 0.04	0.91 ± 0.01	**1.02 ± 0.01**
Cheetah Run	0.68 ± 0.06	0.51 ± 0.04	0.37 ± 0.06	0.21 ± 0.01	**0.81 ± 0.02**
Cheetah Flip	-0.04 ± 0.01	-0.11 ± 0.11	0.06 ± 0.05	-0.26 ± 0.01	**0.97 ± 0.02**
Walker Stand	0.81 ± 0.16	**1.01 ± 0.02**	0.92 ± 0.06	**1.02 ± 0.01**	**1.01 ± 0.02**
Walker Walk	**0.95 ± 0.02**	**0.95 ± 0.03**	0.42 ± 0.10	0.19 ± 0.02	**0.94 ± 0.04**
Walker Run	**0.81 ± 0.02**	0.69 ± 0.05	0.68 ± 0.03	0.21 ± 0.01	0.78 ± 0.04
Walker Flip	0.48 ± 0.04	**0.59 ± 0.04**	0.50 ± 0.04	0.28 ± 0.02	0.47 ± 0.03
Stickman Stand	0.60 ± 0.11	0.41 ± 0.06	0.49 ± 0.05	0.53 ± 0.04	**0.91 ± 0.04**
Stickman Walk	0.83 ± 0.03	0.69 ± 0.13	0.84 ± 0.07	0.26 ± 0.03	**0.91 ± 0.03**
Stickman Run	0.38 ± 0.03	0.37 ± 0.04	0.38 ± 0.03	0.17 ± 0.00	**0.48 ± 0.02**
Stickman Flip	0.29 ± 0.05	**0.62 ± 0.03**	0.38 ± 0.04	0.25 ± 0.03	0.41 ± 0.03
Quadruped Stand	0.95 ± 0.06	0.84 ± 0.20	**0.97 ± 0.04**	**0.99 ± 0.02**	**0.98 ± 0.01**
Quadruped Walk	0.73 ± 0.19	0.64 ± 0.21	0.60 ± 0.17	0.51 ± 0.02	**0.90 ± 0.05**
Quadruped Run	0.72 ± 0.21	0.58 ± 0.13	0.51 ± 0.09	0.52 ± 0.01	**0.94 ± 0.03**
Kitchen Light	0.00 ± 0.00	0.35 ± 0.48	0.92 ± 0.28	**1.00 ± 0.00**	0.97 ± 0.18
Kitchen Slide	0.62 ± 0.49	**1.00 ± 0.00**	**1.00 ± 0.00**	0.97 ± 0.18	**1.00 ± 0.00**
Kitchen Microwave	**1.00 ± 0.00**	0.63 ± 0.48	**1.00 ± 0.00**	0.98 ± 0.13	**1.00 ± 0.00**
Kitchen Burner	0.35 ± 0.48	0.10 ± 0.30	0.63 ± 0.48	**1.00 ± 0.00**	0.60 ± 0.49
Kitchen Kettle	**0.35 ± 0.48**	0.07 ± 0.25	0.05 ± 0.22	0.05 ± 0.22	**0.33 ± 0.47**
Overall	0.60	0.57	0.59	0.52	**0.81**

Experiments: Cross-domain Video Tasks

Deep Semantic Understanding:
Effectively captures task semantics beyond simple visual matching, especially in scenarios with domain gaps.

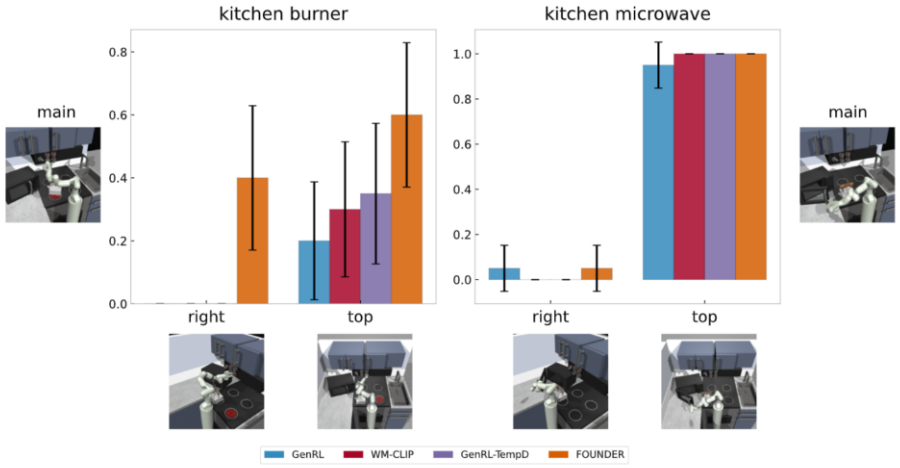


Figure 3. Performance of FOUNDER and baselines over 4 seeds on two tasks in Kitchen. The agent's observations is captured from the Main viewpoint, while task video prompts are provided from the Right or Top viewpoints, yielding 4 cross-view tasks.

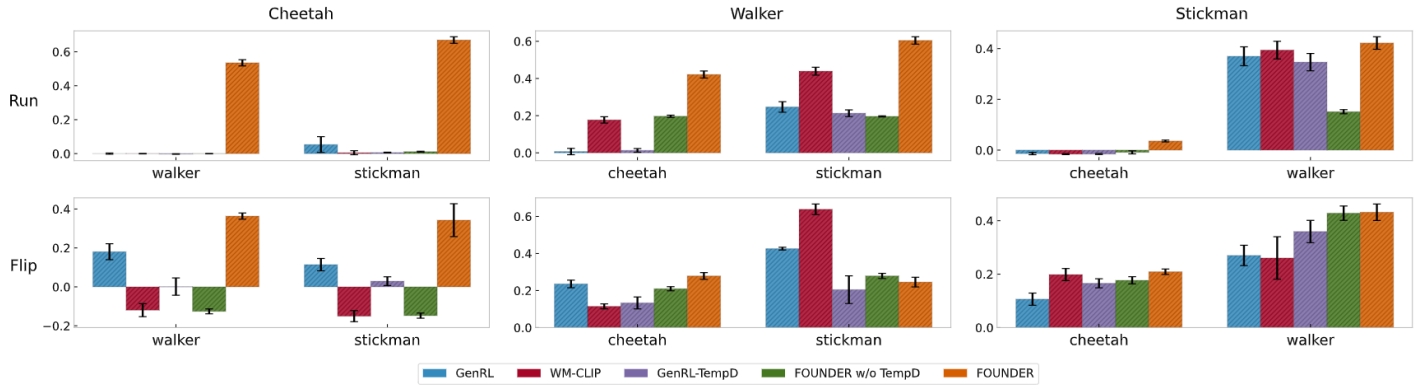


Figure 2. Normalized evaluation performance on cross-embodiment tasks built upon DMC. Each row corresponds to one of the *Run* or *Flip* tasks, while each column represents the domain in which the agent is evaluated. Each subplot presents the results of respectively using videos from the remaining two domains as task prompts. This yields 6 domain combinations: (Cheetah, Walker), (Cheetah, Stickman), (Walker, Cheetah), (Walker, Stickman), (Stickman, Cheetah), and (Stickman, Walker), each evaluated on *Run* or *Flip*, totaling 12 evaluation tasks. Mean scores with standard deviation are recorded across 4 seeds.

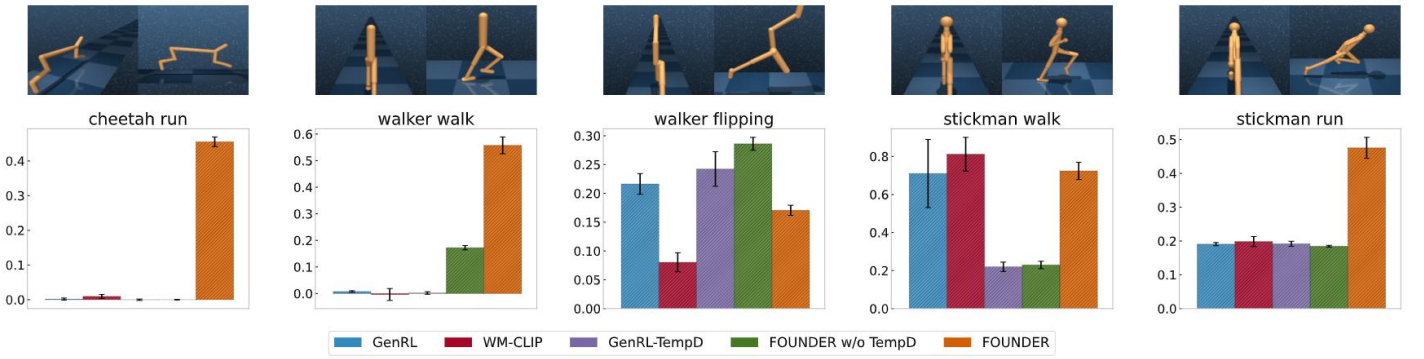


Figure 4. Performance evaluation of FOUNDER and baselines over 4 seeds across 5 cross-viewpoint video tasks in DMC. Visualizations of video prompts indicating the task semantics and the agent's observation in the corresponding environment is presented at first row.

Experiments: Reward Evaluation

Consistent Reward:
FOUNDER's learned reward shows strong correlation with ground-truth rewards.

Table 2. Evaluation of the consistency between learned pseudo rewards and ground-truth rewards, averaged on 7 tasks in DMC. The results for each task are shown in Appendix D.4.

Methods	Corr↑	Regret↓	Precision↑	Recall↑	F1↑
GenRL	0.12	0.37	0.47	0.44	0.44
WM-CLIP	0.40	0.26	0.61	**0.69**	**0.63**
GenRL-TempD	0.05	0.75	0.46	0.46	0.40
FOUNDER w/o TempD	-0.02	0.90	0.16	0.15	0.15
FOUNDER	**0.54**	**0.07**	**1.0**	0.47	0.59

For policy learning based on our assigned pseudo-rewards, avoiding the misclassification of low-reward behavior as high-reward is far more critical than identifying all high-reward behaviors, as mistakenly favoring low-reward behaviors can lead to reward hacking and undesirable outcomes. In this context, precision outweighs recall.

Experiments: Minecraft

Superior Performance:
FOUNDER also outperforms baselines on this harder open-ended environment with more challenging task instructions and complex observations

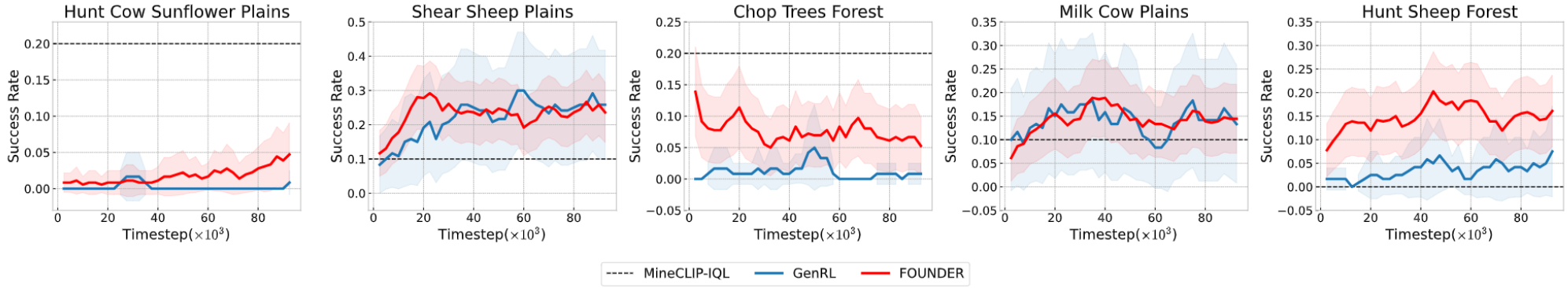


Figure 5. Performance of FOUNDER and baselines over 3 seeds across 5 tasks in Minecraft during behavior learning. Each task is specified in a text prompt. The solid curves and the shaded region indicate the average episodic success rates and the 95% confidence intervals across different runs. We apply a moving average to smooth the curves.

THANK YOU!

Project Website

Contact: wangyc@lamda.nju.edu.cn