SLOPE: Optimistic Potential Landscape Shaping for Model-based Reinforcement Learning

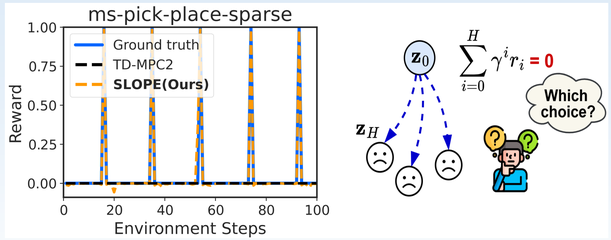
Yao-Hui Li*, Zeyu Wang*, Xin Li†, Wei Pang, Yingfang Yuan, Zhenkun Chen, Boya Zhang, Riashat Islam, Alex Lamb, Yonggang Zhang

*Equal contribution, †Corresponding author

NEURAL INFORMATION PROCESSING SYSTEMS

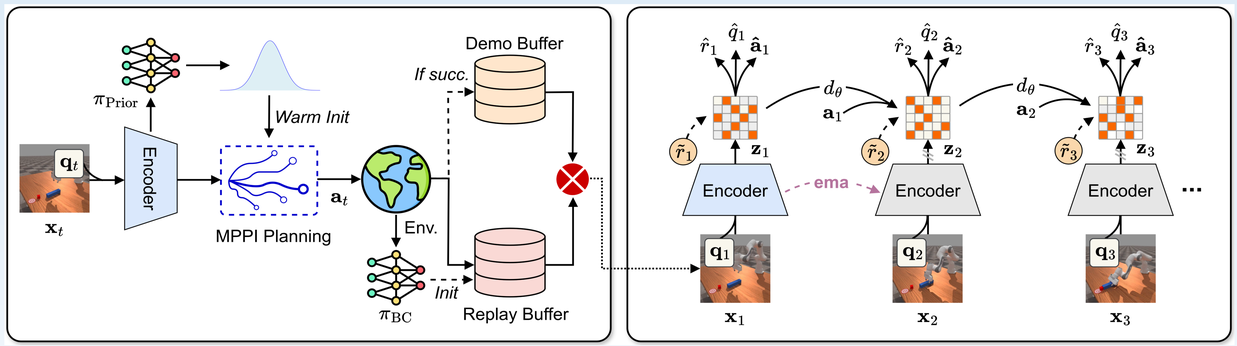
Motivation

- **Reward learning failure under extreme data imbalanceControllability.**
- **Ineffectiveness of ground-truth rewards for planning.**



*How can we construct a **dense**, **informative guiding signal** for MBRL in **sparse reward** settings that provides reliable directional feedback for trajectory planning, while remaining aligned with **the true task objective**?*

Model-Based Planning via Potential Landscape



- **Potential-Based Planning:** Transform sparse rewards into dense guidance for planning.
- **Optimistic Landscape Learning**: Amplify rare success signals to encourage exploration.
- **Stable and Efficient Training**: stable and efficient convergence.

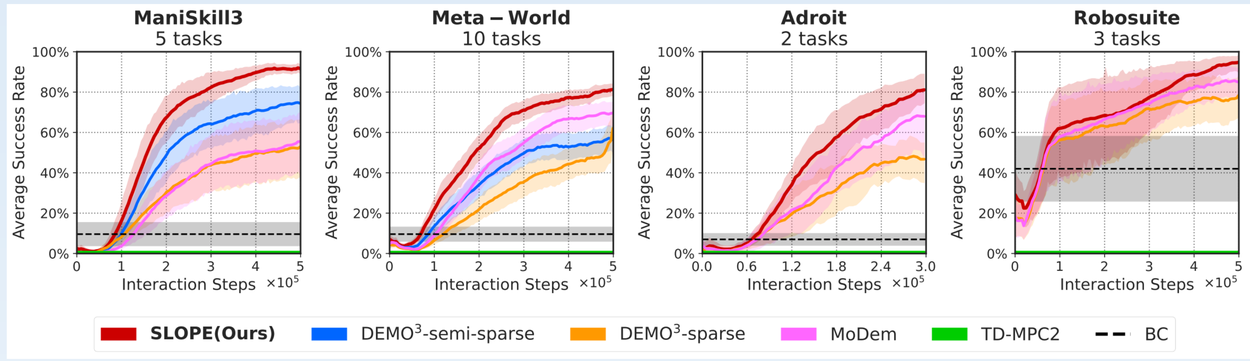
$$\tilde{r}(\mathbf{s},\mathbf{a}) = r(\mathbf{s},\mathbf{a}) + \gamma \mathbb{E}_{\mathbf{s}' \sim \mathcal{P}}[\Phi(\mathbf{s}')] - \Phi(\mathbf{s})]$$

$$\Phi(\mathbf{s}) = \eta \max_{\mathbf{a}} Q_{\bar{\theta}}(\mathbf{s},\mathbf{a})]$$

$$\mathcal{L}_{QCE}(\theta) = \mathbb{E}_{\mathcal{D}}[w(\delta) \cdot CE(\Psi(y), Q_\theta(\mathbf{s},\mathbf{a}))]]$$

$$w(\delta) = \tau \cdot \mathbb{1}_{\{\delta<0\}} + (1-\tau) \cdot \mathbb{1}_{\{\delta \geq 0\}}]$$

Experiments



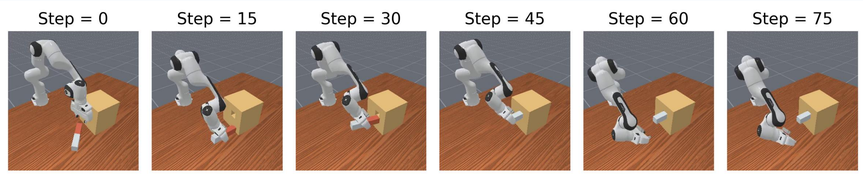
Real-robot Deployment

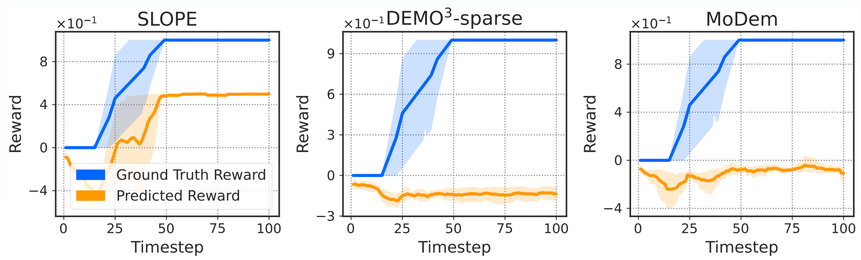


Table 2: Comparison of methods on real-world tasks at $15k$ environment steps ($\approx$ 1 hour of real-world training). Results are averaged over 20 evaluation trials. Success Rate and Intervention Rate are reported in percentage (%). The best results are highlighted in **colorbox**.

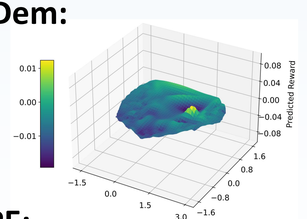
Method	Press Button		Push Cube		Grasp Cube	
	Success (↑)	Intervention (↓)	Success (↑)	Intervention (↓)	Success (↑)	Intervention (↓)
SAC	45.0	16.8	0.0	46.5	0.0	45.3
MoDem	65.0	10.0	15.0	18.4	35.0	24.0
DEMO^3	**100.0**	11.1	15.0	19.1	20.0	23.2
SLOPE (Ours)	**100.0**	**7.3**	**65.0**	**18.2**	**65.0**	**21.3**

*SLOPE's predicted reward exhibits **a clear upward trend** that **aligns with the temporal progression of task completion**.*





MoDem:



SLOPE:

