RENEW: Towards Learning World Models and Repairing Model Exploitation from Preferences

Logan M. Bhamidipaty[1*], Mykel J. Kochenderfer[2], Subramanian Ramamoorthy[1] [1]University of Edinburgh, [2]Stanford University

*l.m.bhamidipaty@sms.ed.ac.uk

Motivation

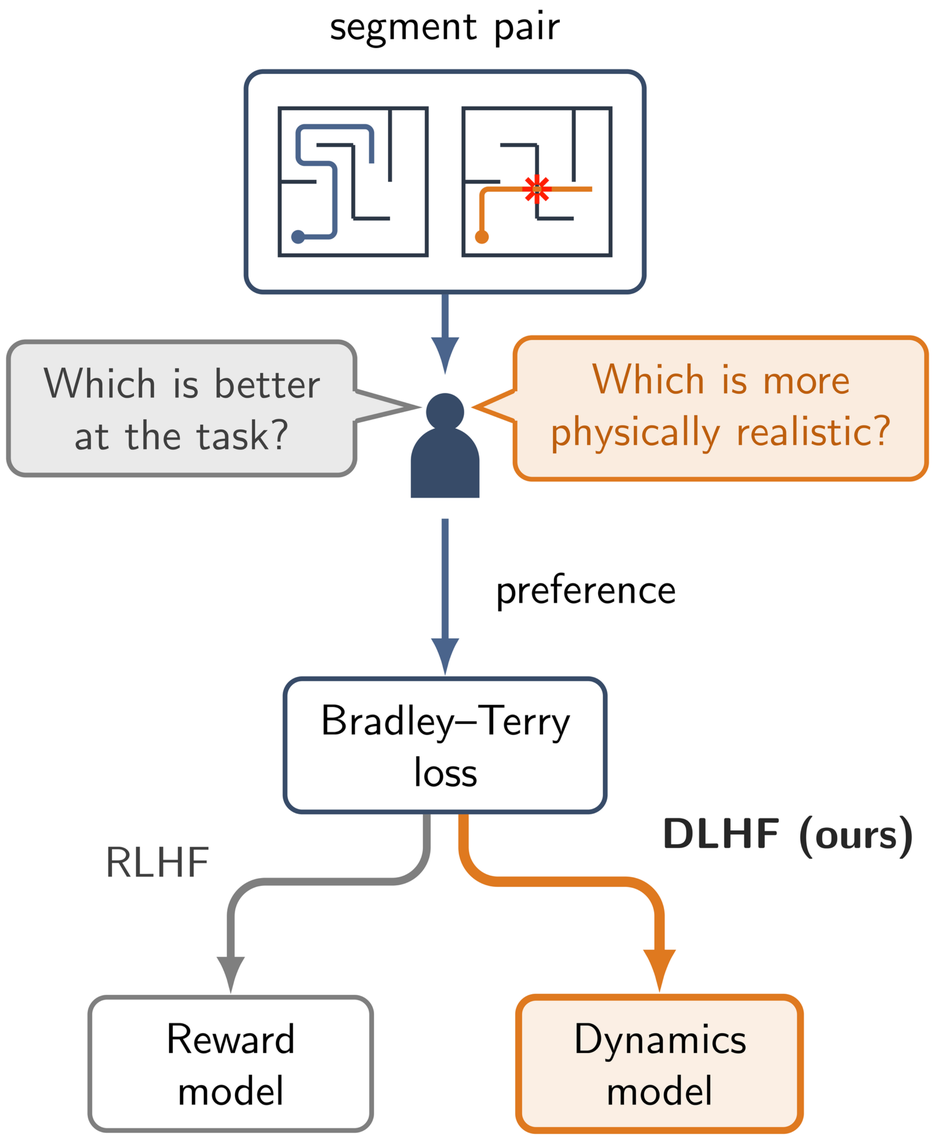
Offline model-based RL suffers **distribution shift** and **model exploitation.** Two typical fixes:

1. Collect more demos → **$$$, unsafe, or finite**
2. Limit OOD exploration → **Poor generalization**

Can we improve 🌍 without more demos / pessimism?

Learning 🌍 from Human Feedback

🔑 **Insight:** *Humans can intuit simple physics!*



RL from Human Feedback (RLHF)

$$L_{DLHF} = \sum \mathrm{CE}\left(y, \mathrm{logistic}\left(\boldsymbol{R_\phi}(\sigma^0) - \boldsymbol{R_\phi}(\sigma^1)\right)\right)$$

Dynamics Learning from Human Feedback (DLHF)

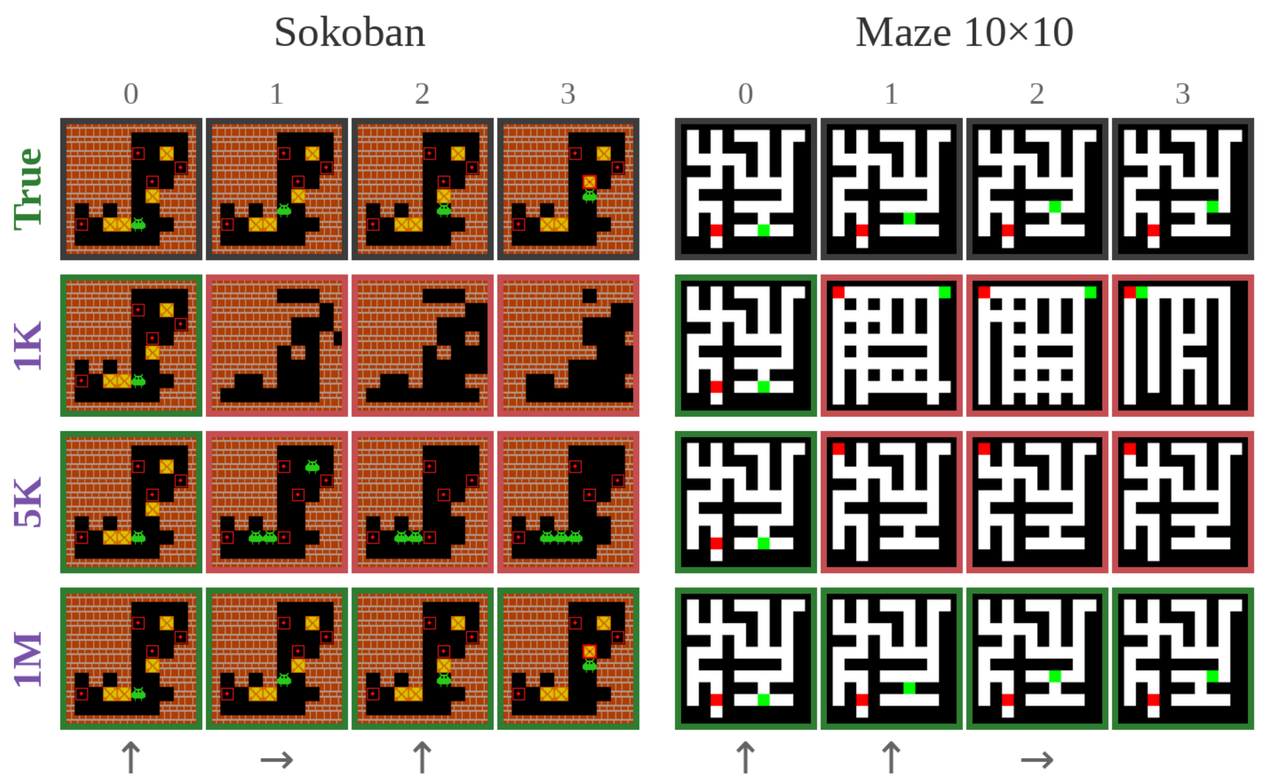
$$L_{DLHF} = \sum \mathrm{CE}\left(y, \mathrm{logistic}\left(\boldsymbol{\ell_\theta}(\sigma^0) - \boldsymbol{\ell_\phi}(\sigma^1)\right)\right)$$

Algorithm 1 RENEW

Require: Pretrained world model $\hat{T}_\theta$, preference budget N, iterations I

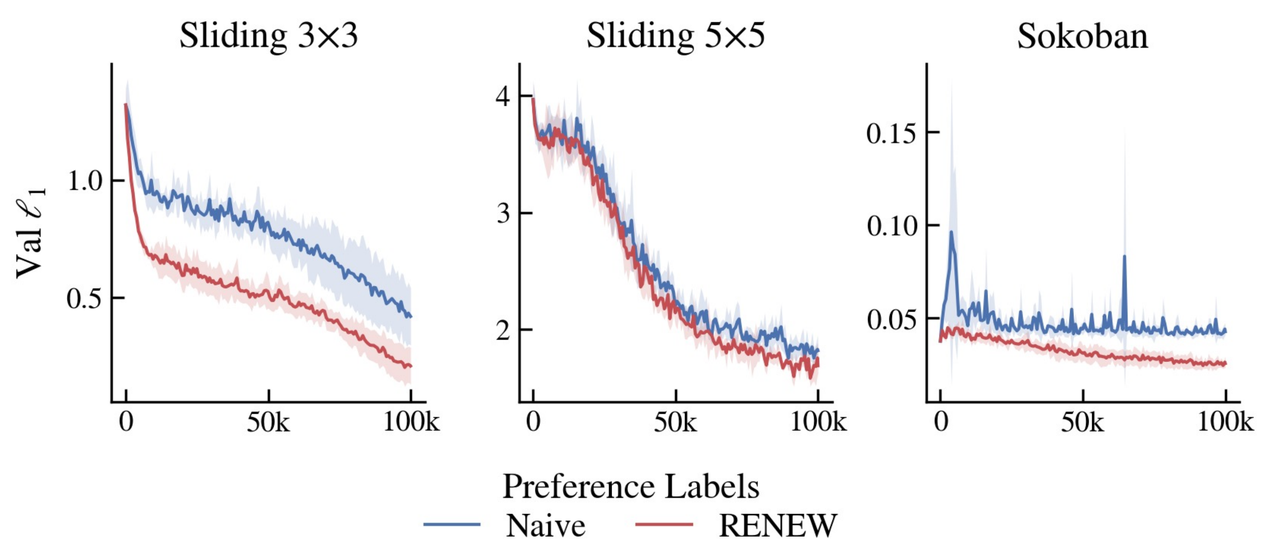
1: **for** $i = 1, \ldots, I$ **do**
2: Compute uncertainty u under $\hat{T}_\theta$
3: Sample segment pairs (σ^0, σ^1) from $\hat{T}_\theta \propto u$
4: Elicit N/I preferences $\mathcal{D}_\succ = \{(\sigma_j^0, \sigma_j^1, y_j)\}_{j=1}^{N/I}$
5: Update $\theta \leftarrow \arg\min_\theta \mathcal{L}_{\mathrm{DLHF}}(\theta; \mathcal{D}_\succ)$
6: **end for**
7: **return** $\hat{T}_\theta$

Q1: Can DLHF learn discrete 🌍?



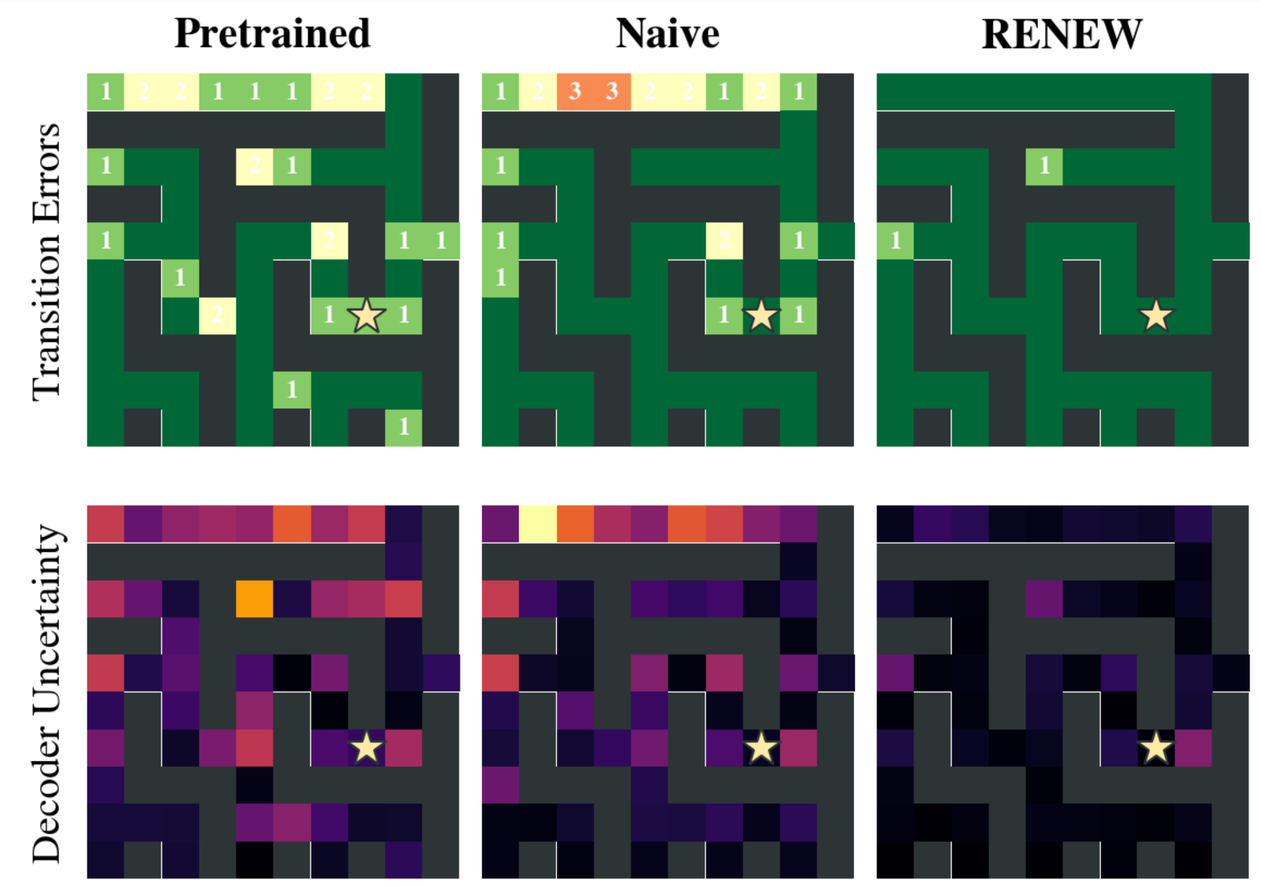
A1: *Yes, but naïve DLHF is very **sample inefficient.***

Q2: Can we boost DLHF sample efficiency?



A2: *Yes, **RENEW is more efficient** than naïve DLHF.*

Q3: Can RENEW target model exploitation?



A3: Yes,* *RENEW reduces uncertainty & forgetting.*

This work was supported by a UKRI Turing AI World Leading Researcher Fellowship on AI for Person-Centred and Teachable Autonomy (grant EP/Z534833/1)