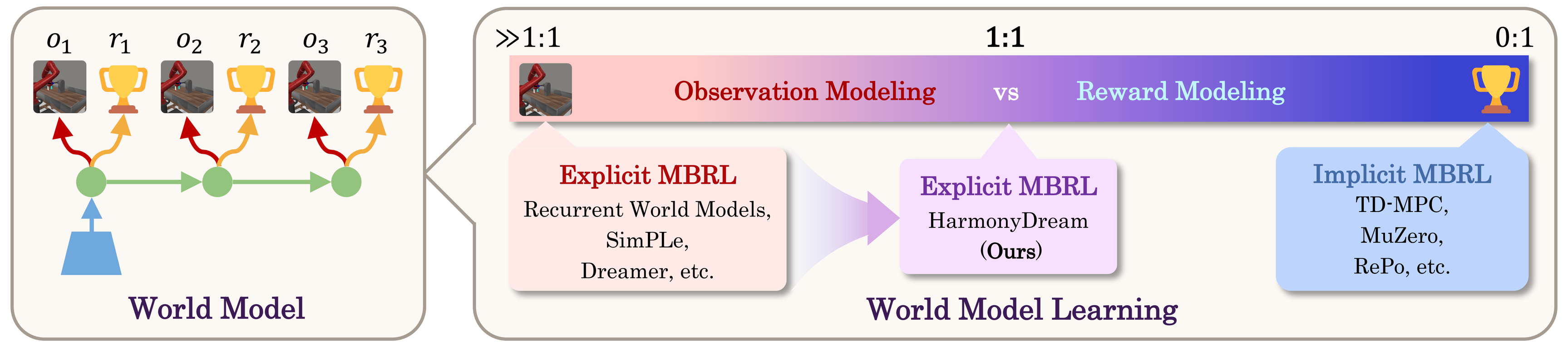
HarmonyDream: Task Harmonization Inside World Models

Haoyu Ma*, Jialong Wu*, Ningya Feng, Chenjun Xiao, Dong Li, Jianye Hao, Jianmin Wang, Mingsheng Long

ICML International Conference On Machine Learning

Introduction

- **World models**: Internal models of how the world works
- **Two key tasks** in world model learning:
  - Observation Modeling: how the environment transits and is observed.
  - Reward Modeling: how the task has been progressed.
  - Provide a unified multi-task view of MBRL.



- **Explicit MBRL**: Learns an exact duplicate of the environment.
  - Typically dominated by observation modeling.
  - Limited by environment complexity (irrelevant details) and model capacity.
- **Implicit MBRL**: Learns only task-centric world models.
  - Relies solely on reward modeling to achieve value equivalence.
  - Limited by sparse learning signals from a single scalar reward.

Research Problem

How do model-based RL methods properly exploit the intrinsic multi-task benefits within world model learning?

- **Contributions**:
  - A systematic analysis of deficiencies brought by task domination.
  - HarmonyDream, a simple but effective method to mitigate domination.
  - Significant improvement of sample efficiency on various domains

Overview of World Model Learning

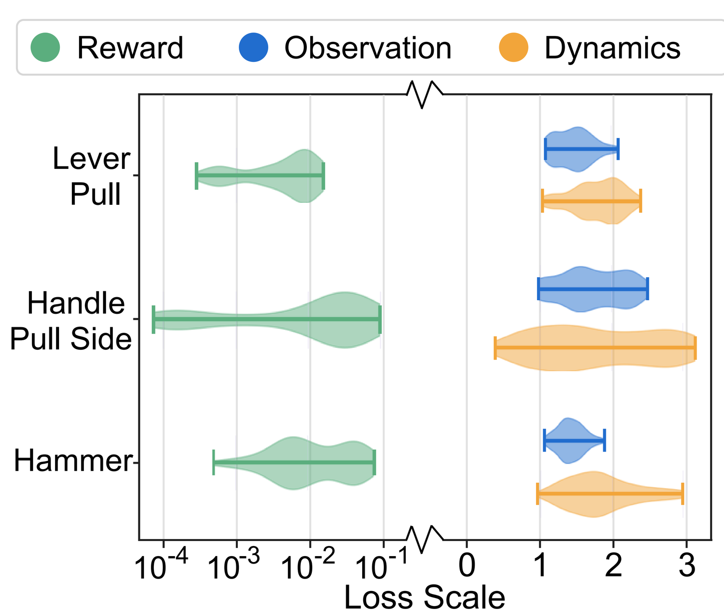
- **Optimization objectives**:

Observation: $\mathcal{L}_o(\theta) = -\log p_\theta(o_t \mid z_t)$

Reward: $\mathcal{L}_r(\theta) = -\log p_\theta(r_t \mid z_t)$

Dynamics: $\mathcal{L}_d(\theta) = \mathrm{KL}[q_\theta(z_t \mid z_{t-1}, a_{t-1}, o_t) \,\|\, p_\theta(\hat{z}_t \mid z_{t-1}, a_{t-1})]$.

$$\mathcal{L}(\theta) = w_o\mathcal{L}_o(\theta) + w_r\mathcal{L}_r(\theta) + w_d\mathcal{L}_d(\theta).$$



- **Dimension difference**: The observation loss aggregates $\mathrm{H} \times \mathrm{W} \times \mathrm{C}$ dimensions, while reward is only a scalar.
- **Typical practice**: Approximately equal weights $w_o = w_r = w_d = 1$, overlooking the imbalanced nature of world model learning.

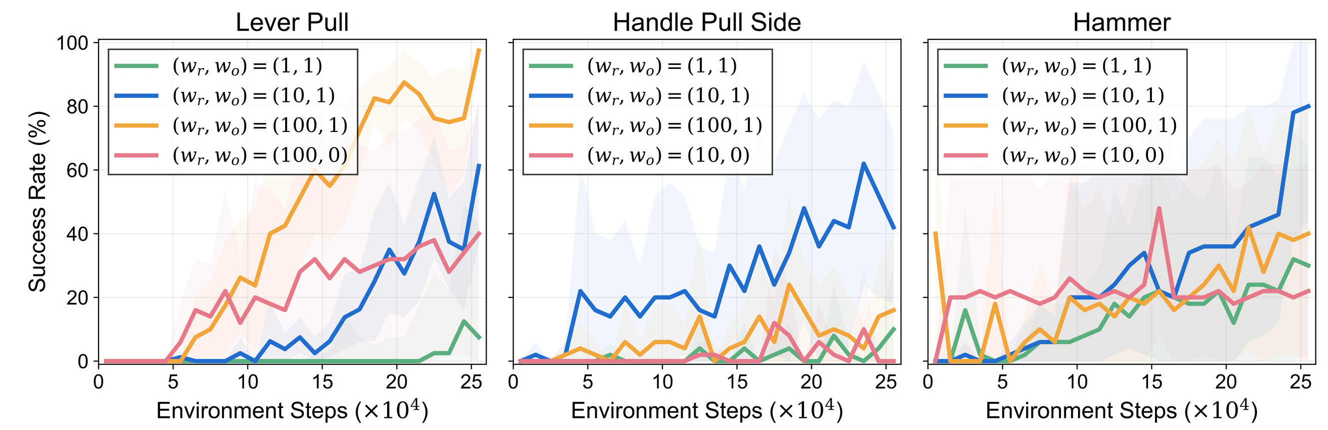
Our Insight

There exists potential benefits of multi-task learning yet to be exploited.

Dive Into World Model Learning

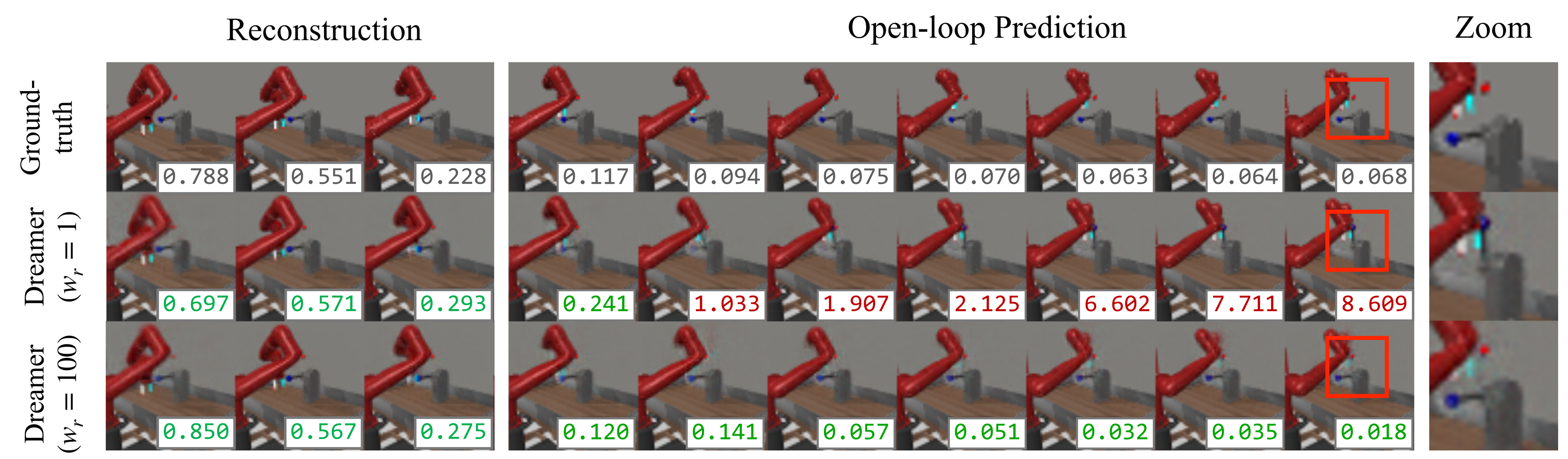
Finding 1

Leveraging the reward loss by adjusting its coefficient in world model learning has a great impact on the sample efficiency of model-based agents.



Finding 2

Domination of observation modeling can result in world models establishing spurious correlations without realizing incorrect reward predictions.



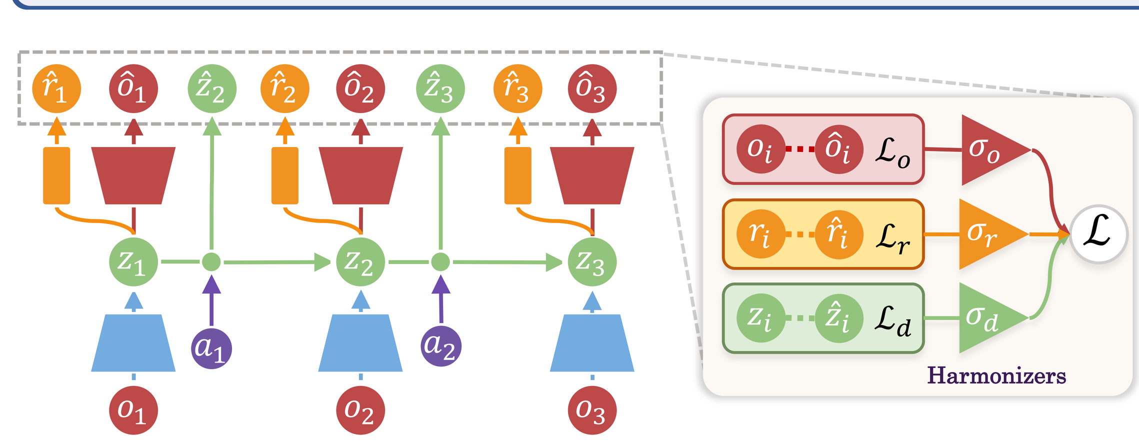
Finding 3

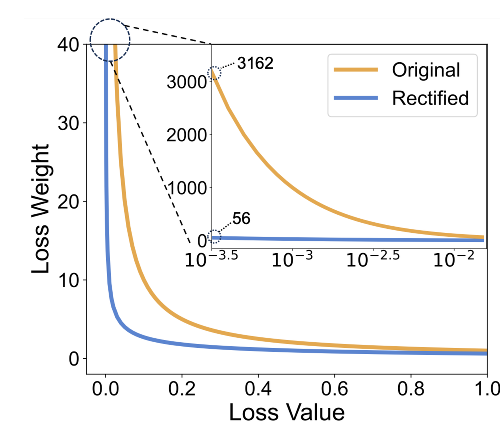
Learning signal of world models from rewards alone without observations is inadequate for sample-efficient model-based learning.

HarmonyDream

Principle

Scale losses to the same constant to harmonize interactions between tasks.



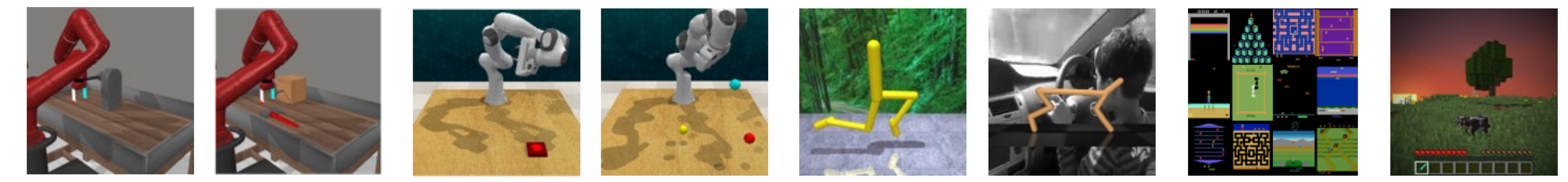


- **Harmonious loss**:

$$\mathcal{L}(\theta, \sigma_o, \sigma_r, \sigma_d) = \sum_{i \in \{o,r,d\}} \frac{1}{\sigma_i}\mathcal{L}_i(\theta) + \log(1 + \sigma_i).$$

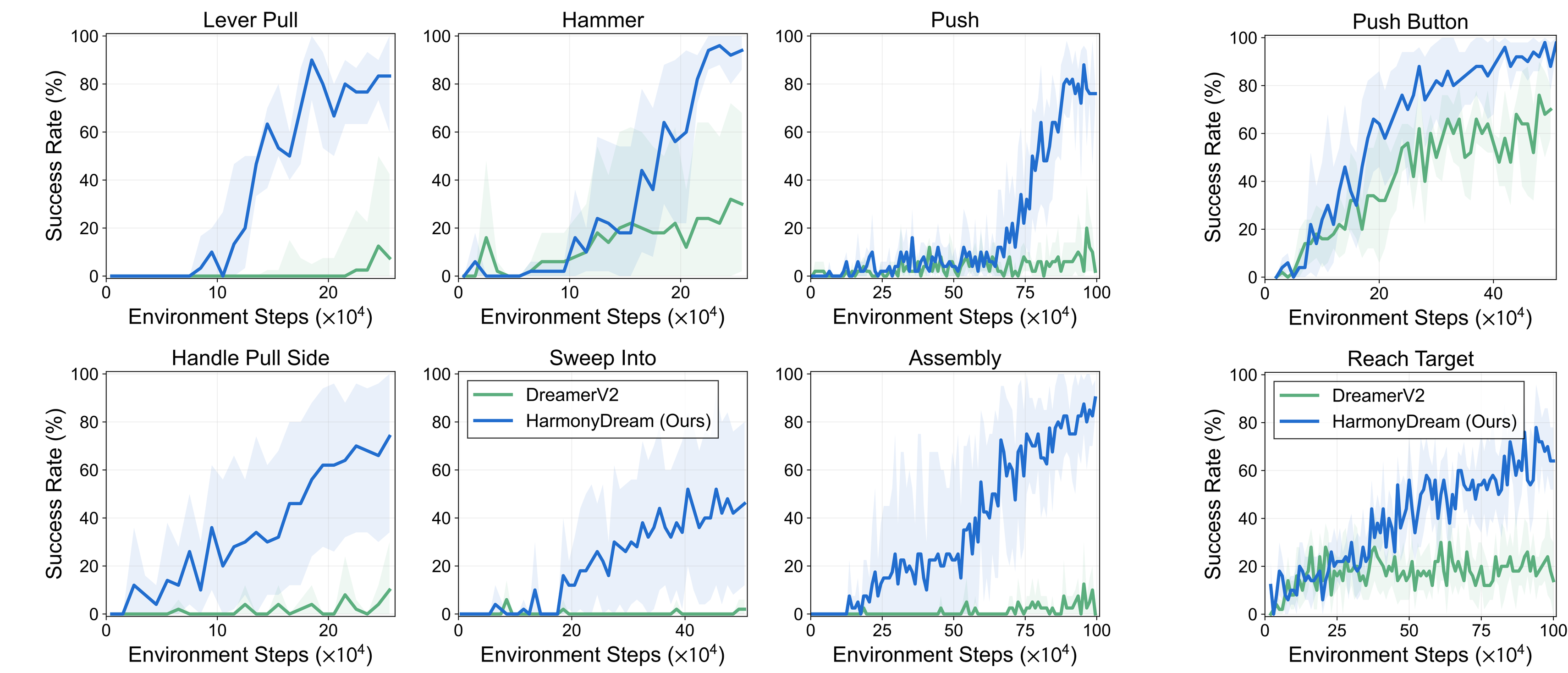
  - Variational approach: Dynamically adjusting σ towards $\mathbb{E}[\mathcal{L}/\sigma^*] = 1$.
  - Additional rectification: Prevent extremely large loss weights.

Experiment Results

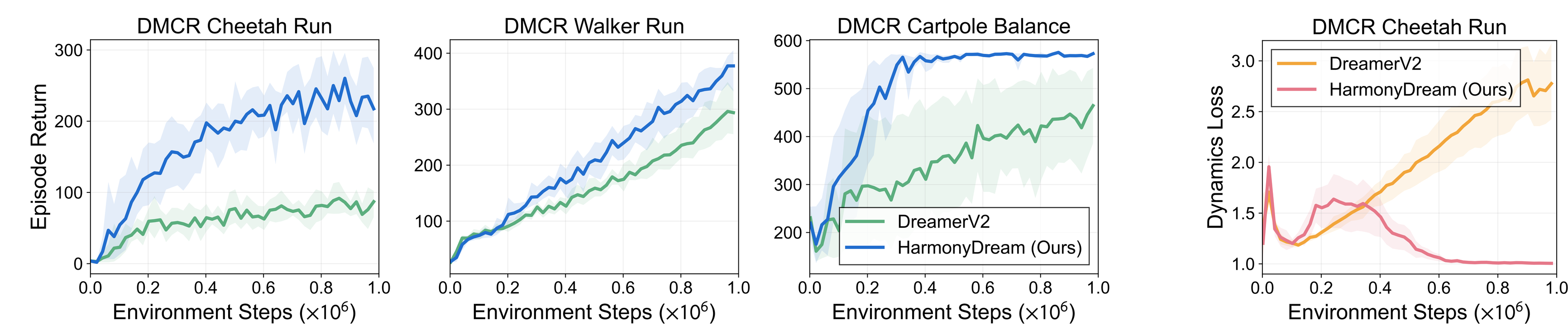


Meta-world RLBench Distracted DMC variants Atari Minecraft

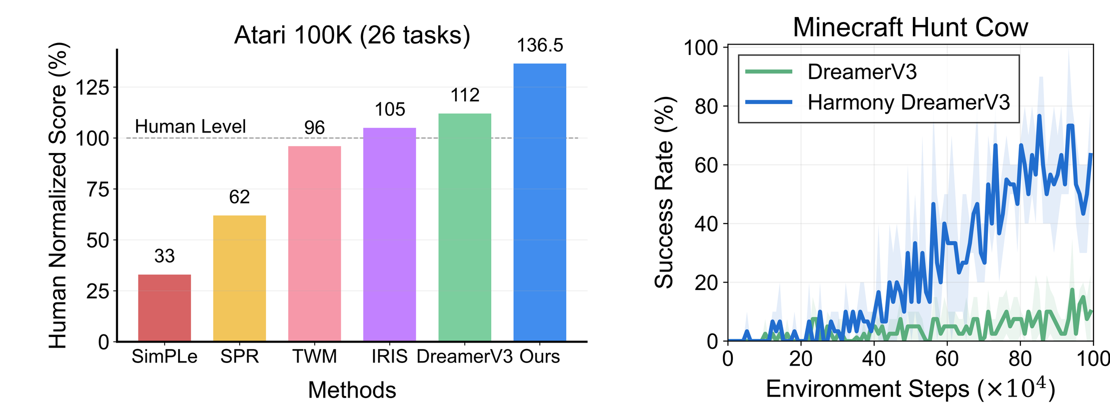
- **Meta-world & RLBench**: Simply adding harmonizers, HarmonyDream shows superior performance in terms of both sample efficiency and final success rate.



- **Distracted control**: HarmonyDream bypasses distractors in observations and can learn task-centric transitions more easily.



- **Video games**: HarmonyDream further unleashes the potentials of DreamerV3, setting a new state of the art on Atari 100k, and greatly improving on Minecraft.



Applicability

- Fine-grained task-relevant observations: Robotic manipulation tasks and video games require accurately modeling interactions with small objects.
- Highly varied task-irrelevant observations: Redundant visual components can easily distract agents if task-relevant information is not emphasized correctly.
- Hybrid of both: More difficult open-world tasks (e.g., Minecraft) can encounter both, including small target entities and abundant visual details.

School of Software - BNRist - Tsinghua University - China

Code: https://github.com/thuml/HarmonyDream

Mail: mhy22@mails.tsinghua.edu.cn / wujialong0229@gmail.com