ICML 2024
Forty-first International Conference on Machine Learning

ICML
International Conference
On Machine Learning

HarmonyDream: Task Harmonization Inside World Models

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

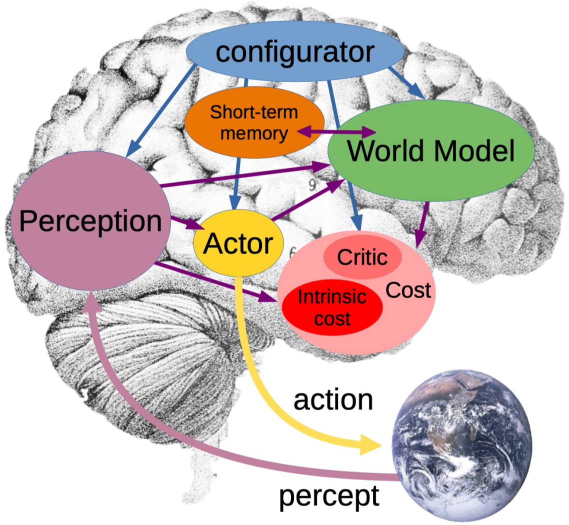
Haoyu Ma [*1] **Jialong Wu** [*1] **Ningya Feng** [1] **Chenjun Xiao** [2] **Dong Li** [2] **Jianye Hao** [2 3] **Jianmin Wang** [1] **Mingsheng Long** [1]

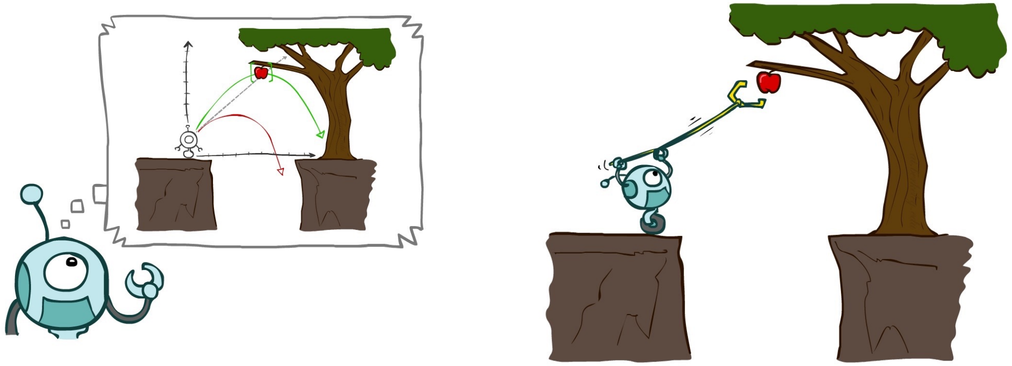
*Equal contribution [1]School of Software, BNRist, Tsinghua University.
[2]Huawei Noah's Ark Lab. [3]College of Intelligence and Computing, Tianjin University.

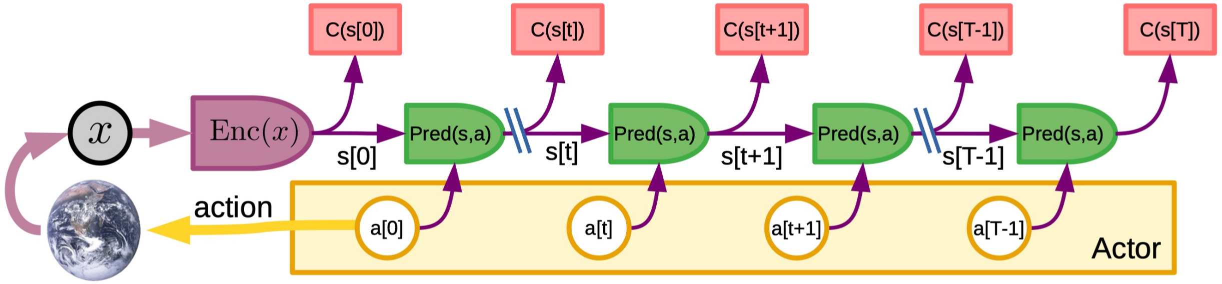
Tsinghua University

HUAWEI

World Models







World Models:

Internal models of how the world works

Model-based Agents:

Act through an optimization procedure (planning) running the world model.

Yann LeCun. A path towards autonomous machine intelligence. 2022.
Dan Klein and Pieter Abbeel. Introduction to Artificial Intelligence.

Video Generation Models as World Simulators?

OpenAI **Sora!**

Abandon generative models!

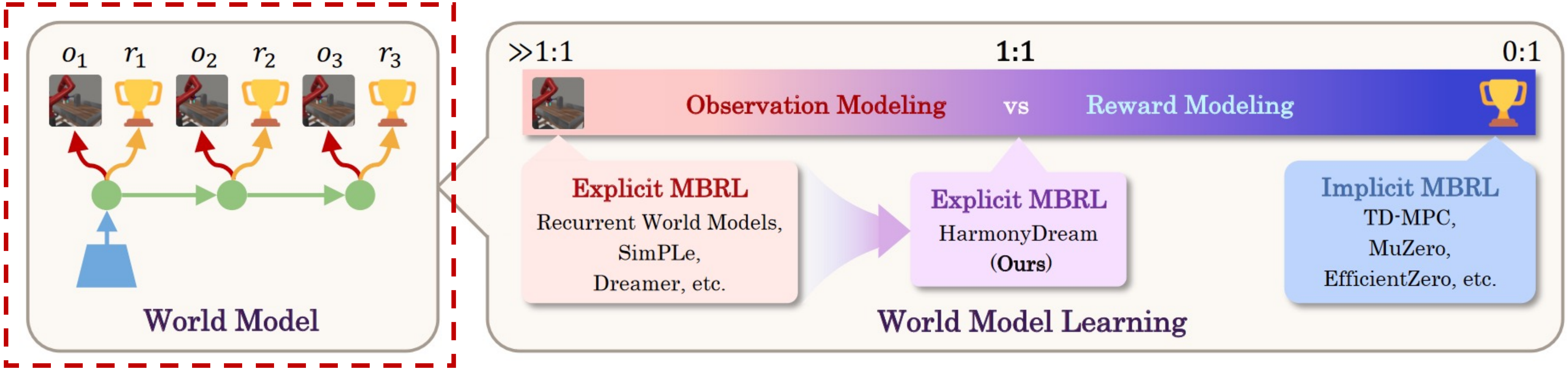
"Modeling the world for action by generating pixel is as wasteful and doomed to failure..."

"It's much more desirable to generate abstract representations of those continuations that eliminate details in the scene that are irrelevant to any action we might want to take."

Pixel-Driven vs. **Objective-Driven**

OpenAI. https://openai.com/research/video-generation-models-as-world-simulators
Yann LeCun. https://twitter.com/ylecun/status/1758740106955952191

A Multi-task View of World Models



Two key tasks in world models:

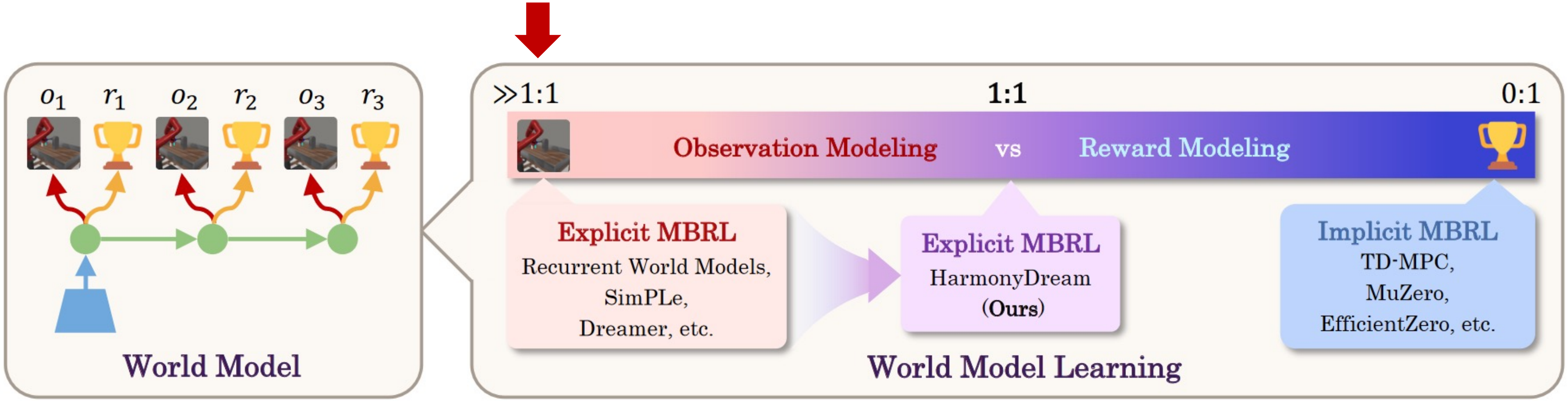
- **Observation Modeling:** how the environment transits and is observed

$$p\left(o_{t+1:T} \mid o_{1:t}, a_{1:T}\right)$$

- **Reward Modeling:** how the task has been progressed

$$p\left(r_{t+1:T} \mid o_{1:t}, a_{1:T}\right)$$

A Multi-task View of World Models

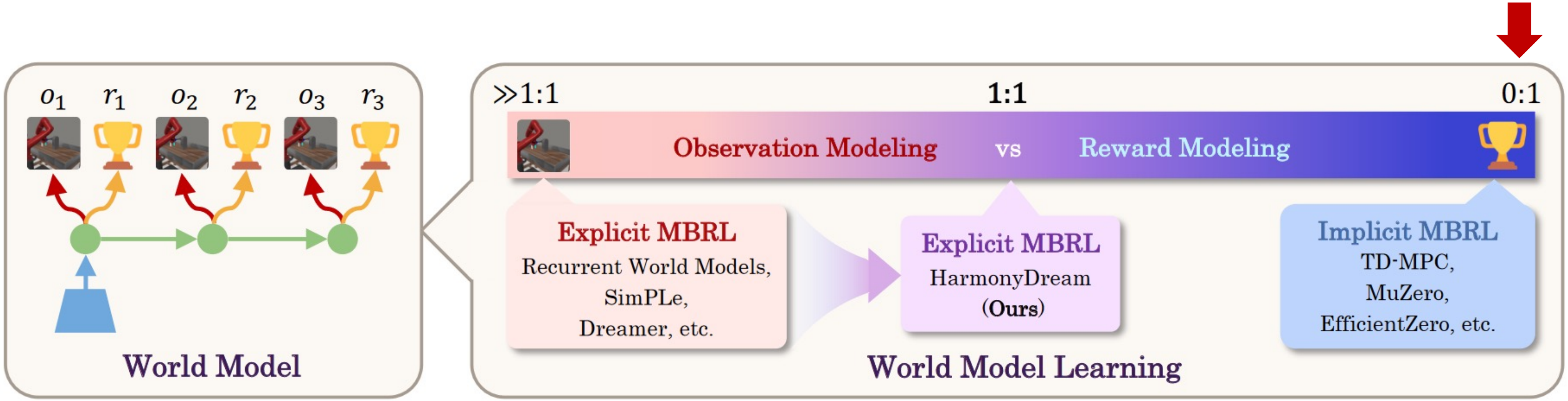


Unifying MBRL in concept (1/2): **Explicit MBRL**

- Learns an exact duplicate of the environment
- Typically dominated by **observation modeling**
- Limited by environment complexity (irrelevant details!) and model capacity

Thomas M. Moerland, Model-based reinforcement learning: A survey, 2023

A Multi-task View of World Models



Unifying MBRL in concept (2/2): **Implicit MBRL**

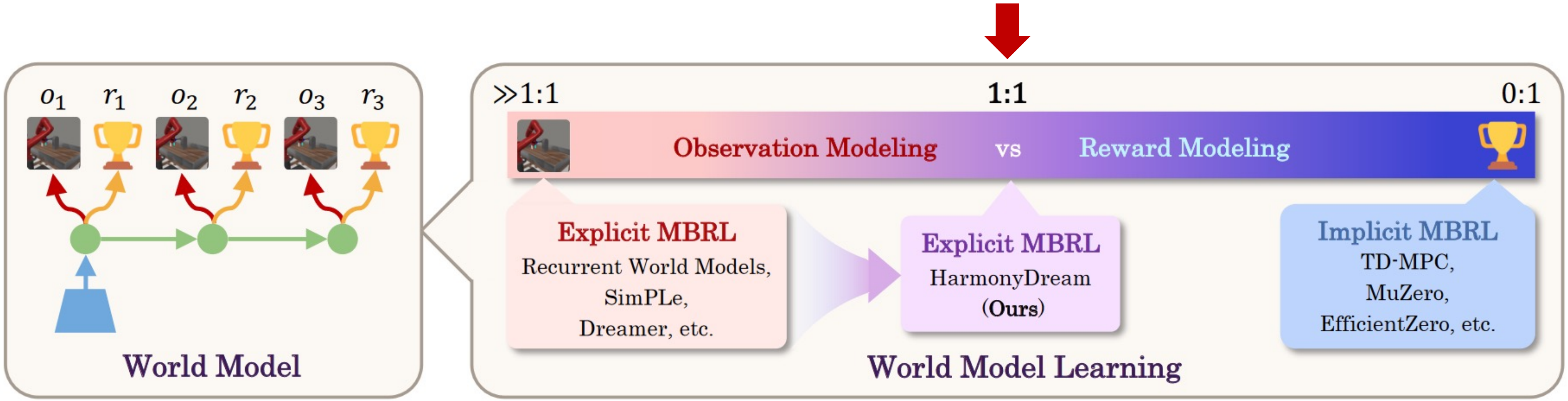
- Learns task-centric world models
- Relies solely on **reward modeling**
- Limited by sparse learning signals

Value equivalence principle:
Predicted rewards of the world model match that of the real environment.

Thomas M. Moerland, Model-based reinforcement learning: A survey, 2023

Schrittwieser, Julian, et al. Mastering atari, go, chess and shogi by planning with a learned model. Nature 588 (2020): 604-609.

Our Contributions



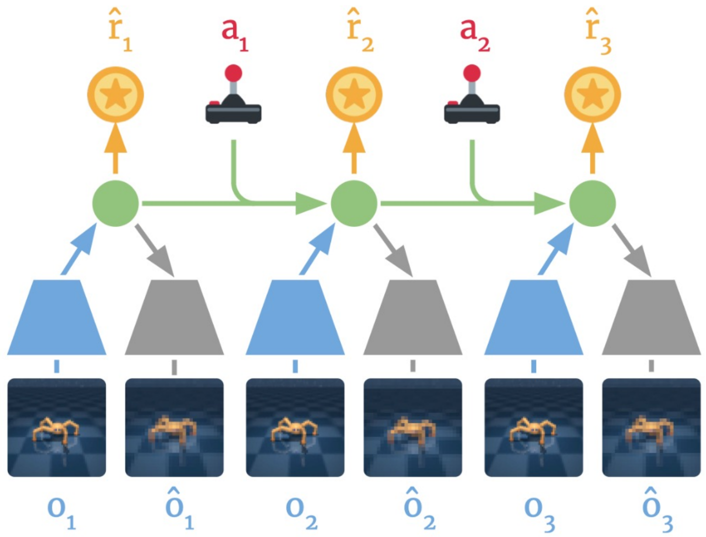
1. Systematically identify the multi-task essence of world models and analyze the deficiencies by task domination.
2. HarmonyDream, a world model learning approach to mitigate the domination of either task.
3. Extensive experiments on visual robotic tasks and video game benchmarks.

✓ Three findings

✓ One simple yet effective method

✓ Eight Domains

Dreamer: An Instantiation of Explicit World Models



Representation model: $z_t \sim q_\theta(z_t \mid z_{t-1}, a_{t-1}, o_t)$

Transition model: $\hat{z}_t \sim p_\theta(\hat{z}_t \mid z_{t-1}, a_{t-1})$

Observation model: $\hat{o}_t \sim p_\theta(\hat{o}_t \mid z_t)$

Reward model: $\hat{r}_t \sim p_\theta(\hat{r}_t \mid z_t)$

Model Learning with Sequential Variational Inference

$$\mathcal{L}(\theta) \doteq \mathbb{E}_{q_\theta(z_{1:T} \mid a_{1:T}, o_{1:T})}\Big[\sum_{t=1}^{T}\Big(\underbrace{-\ln p_\theta(o_t \mid z_t)}_{\text{Observation loss}} \underbrace{-\ln p_\theta(r_t \mid z_t)}_{\text{Reward loss}} + \beta_z \underbrace{\mathrm{KL}\big[q_\theta(z_t \mid z_{t-1}, a_{t-1}, o_t) \,\|\, p_\theta(\hat{z}_t \mid z_{t-1}, a_{t-1})\big]}_{\text{Dynamics loss between prior and posterior}}\Big)\Big].$$

Behavior Learning: Purely on imaginary latent trajectories

Hafner, Danijar, et al. Dream to control: Learning behaviors by latent imagination. ICLR 2020.
Hafner, Danijar, et al. Mastering atari with discrete world models. ICLR 2021.

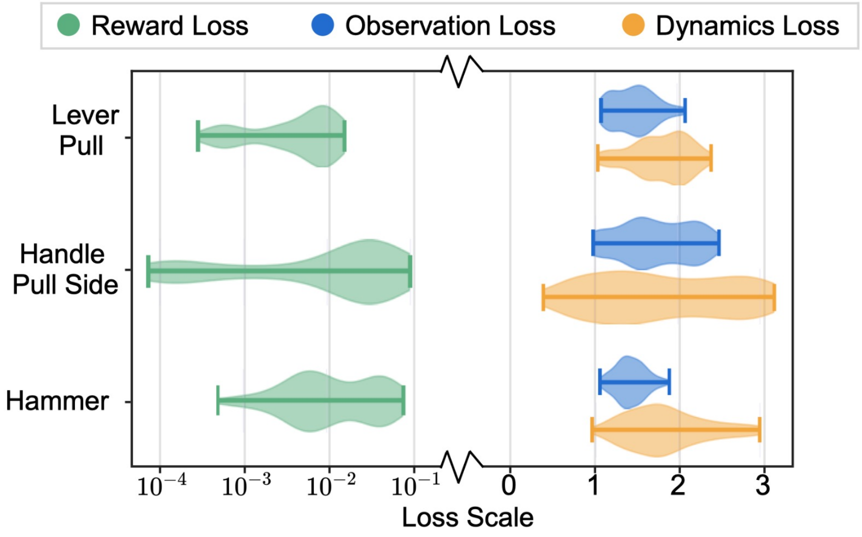
Dive into World Model Learning

Observation loss: $\mathcal{L}_o(\theta) = -\log p_\theta(o_t \mid z_t) = -\sum_{h,w,c} \log p_\theta\left(o_t^{(h,w,c)} \mid z_t\right)$

It aggregates H×W×C dimensions

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

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



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

Typical but suboptimal practice:

Approximately equal weights

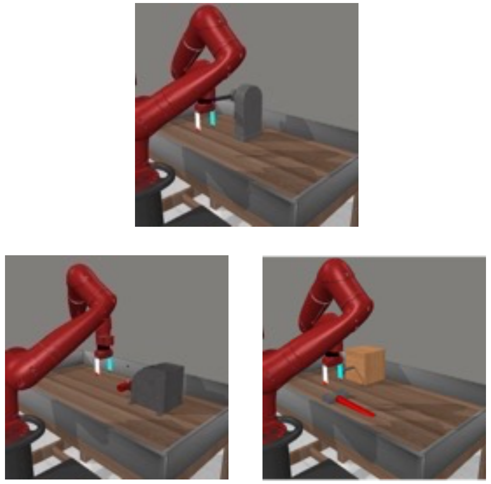
$$w_o = w_r = w_d = 1.0$$

Imbalanced nature of world model learning

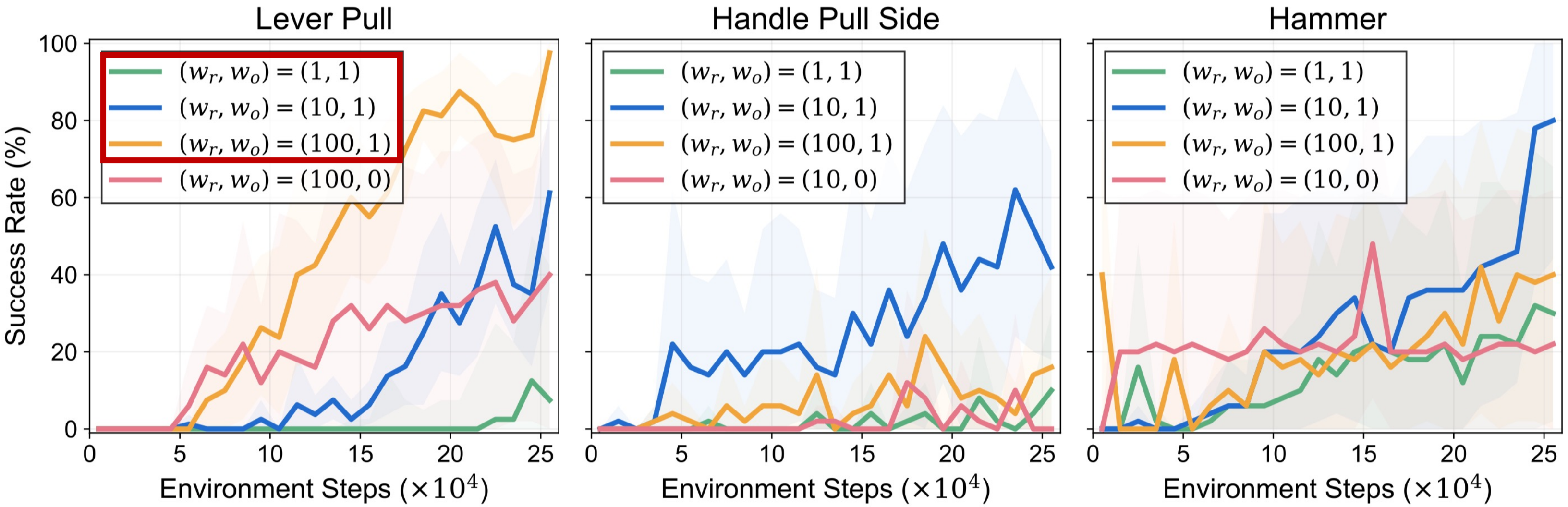
Potential benefits of multi-task learning yet properly exploited!

Task Weighting is Crucial

Dramatically boosted sample efficiency!



Testbed:
Three manipulation tasks from Meta-world

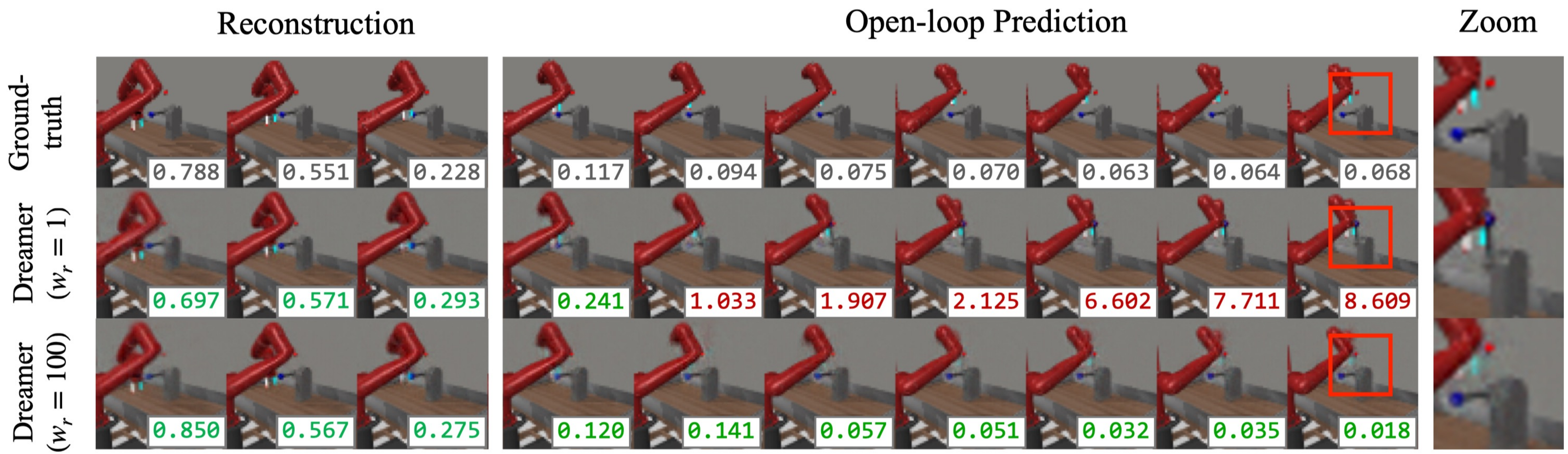


$$\mathcal{L}(\theta) = w_o \mathcal{L}_o(\theta) + w_r \mathcal{L}_r(\theta) + w_d \mathcal{L}_d(\theta)$$

(↑)

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.

Observation Modeling Learns Spurious Correlations



Finding 2. Observation modeling as a dominating task can result in world models establishing spurious correlations without realizing incorrect reward predictions.

Observation Modeling Learns Spurious Correlations

Hallucinations!

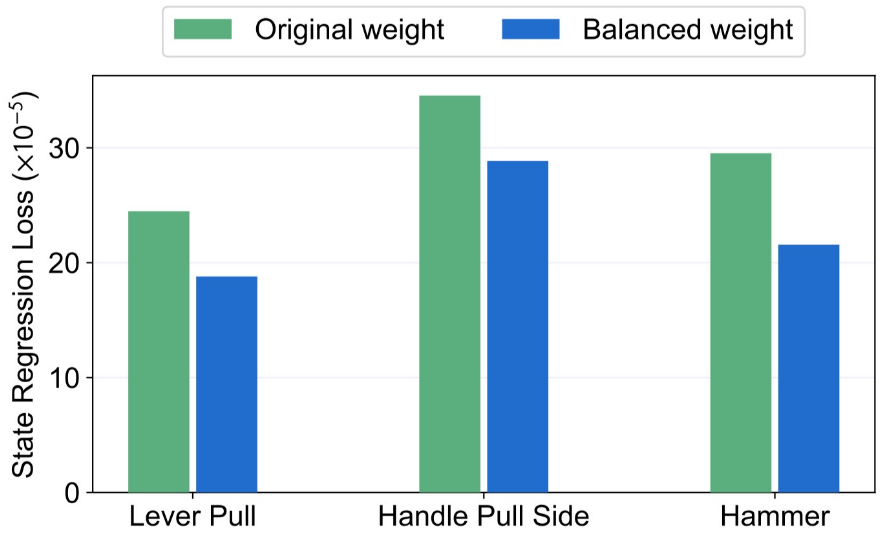
How to mitigate this?

Emphasizing task-relevant information

Finding 2. Observation modeling as a dominating task can result in world models establishing spurious correlations without realizing incorrect reward predictions.

Observation Modeling Learns Spurious Correlations

Properly balancing the reward loss learns task-centric representations capable of better predicting ground truth states



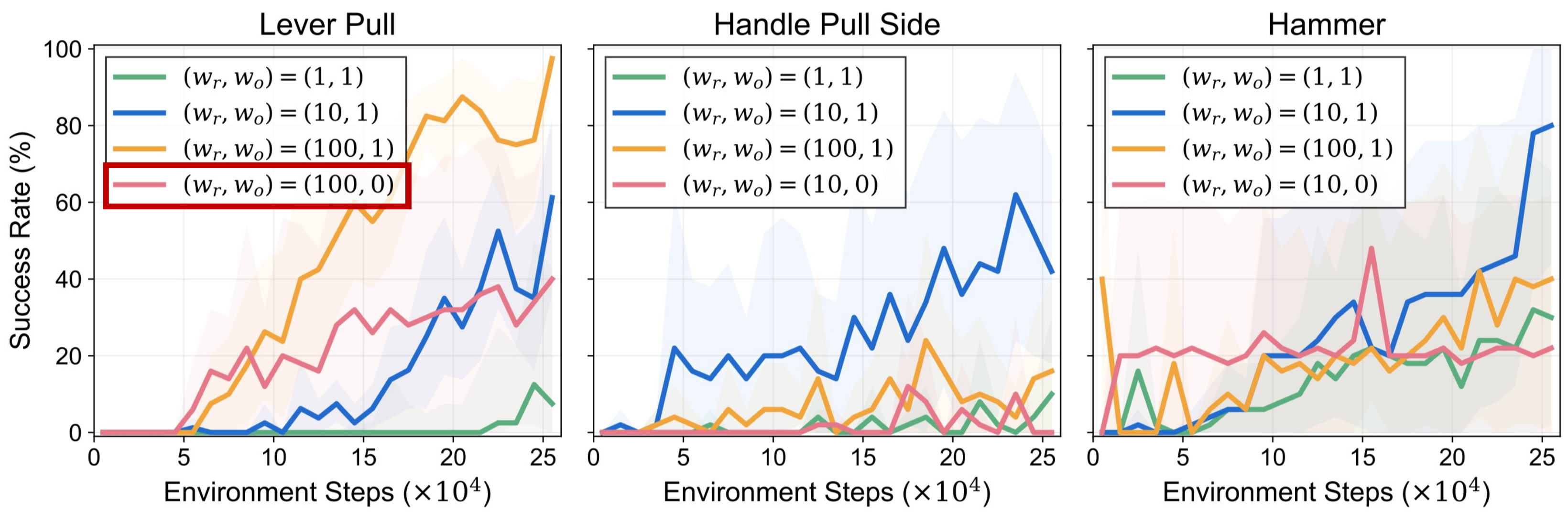
Hallucinations!

How to mitigate this?

Emphasizing task-relevant information

Finding 2. Observation modeling as a dominating task can result in world models establishing spurious correlations without realizing incorrect reward predictions.

Reward Modeling Alone is Not Enough



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

$(= \mathbf{0})$

Limited capability of representation learning…

Finding 3. Learning signal of world models from rewards alone without observations is inadequate for sample-efficient model-based learning.

HarmonyDream

Harmonious interaction between the two world model tasks

Facilitates representation learning

Observation Modeling ⇄ **Reward Modeling**

Enhance task-centric representations

Our principle: Losses scaled to the same constant

A straightforward but suboptimal approach

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

$$w_i = \text{sg}\left(\frac{1}{\mathcal{L}_i}\right), i \in \{o, r, d\}$$

✗ Fluctuate throughout training

✗ Sensitive to outlier values

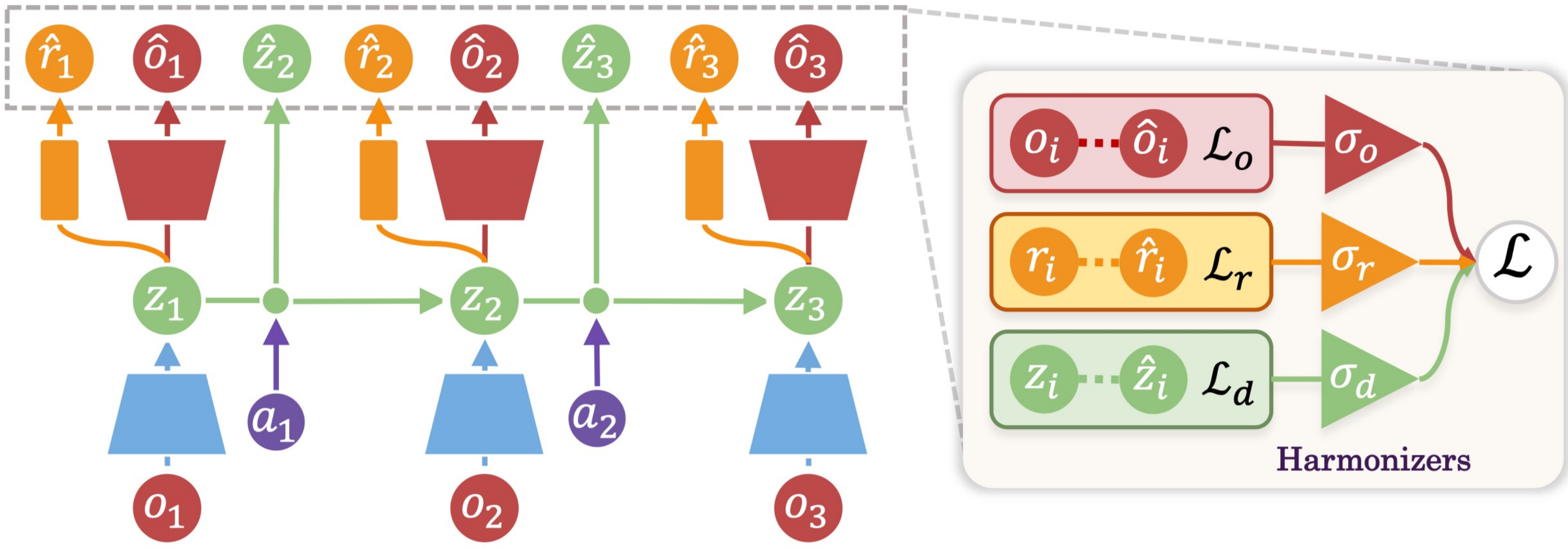
A Variational Approach and Its Rectification

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

$$\sigma^* = \mathbb{E}[\mathcal{L}]$$
$$\mathbb{E}[\mathcal{L}/\sigma^*] = 1$$

A "global" reciprocal of the loss scale

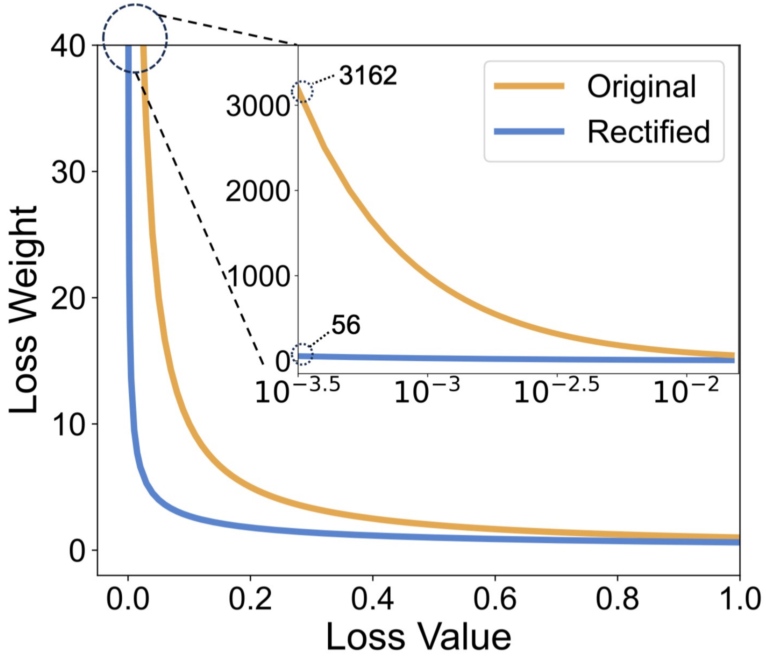
Dynamically but smoothly



A Variational Approach and Its Rectification

Extremely large coefficient hurts training stability

$$1/\sigma \approx \mathcal{L}^{-1} \gg 1$$



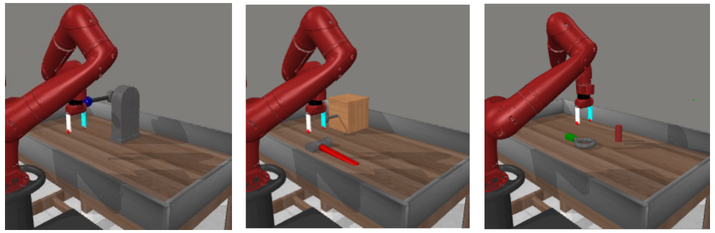
$$\mathcal{L}\left(\theta, \sigma_o, \sigma_r, \sigma_d\right) = \sum_{i \in \{o,r,d\}} \hat{\mathcal{H}}\left(\mathcal{L}_i(\theta), \sigma_i\right)$$

$$= \sum_{i \in \{o,r,d\}} \frac{1}{\sigma_i}\mathcal{L}_i(\theta) + \log\left(1 + \sigma_i\right)$$

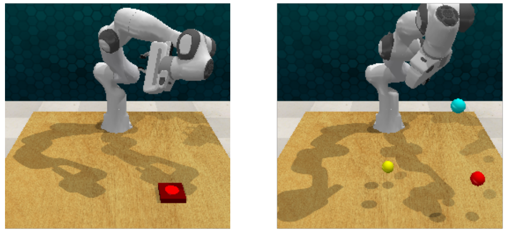
$$\mathbb{E}\left[\mathcal{L}/\sigma^*\right] = \frac{2}{1 + \sqrt{1 + 4/\mathbb{E}[\mathcal{L}]}} < 1$$

Prevent extremely large loss weights

Experiments: Extensive Benchmarks and Tasks

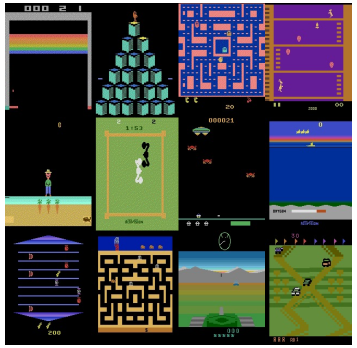


Meta-World
Yu et al. CoRL 2020



RLBench
James et al. IEEE RA-L 2020

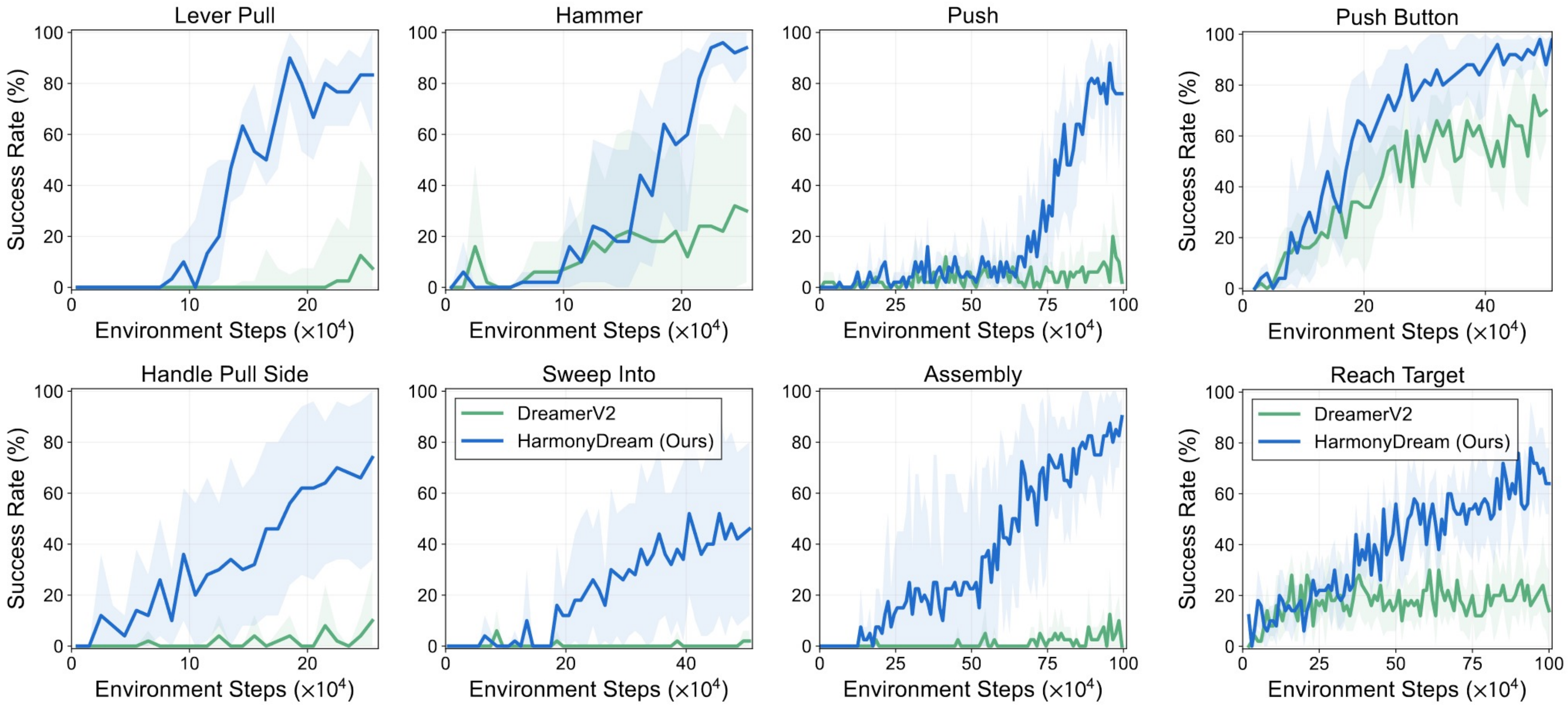
Distracted DMC Variants
Tassa et al. 2018; Zhang et al. 2018



Atari100K
Kaiser et al. ICLR 2020

Minecraft
Fan et al. NerulPS 2022

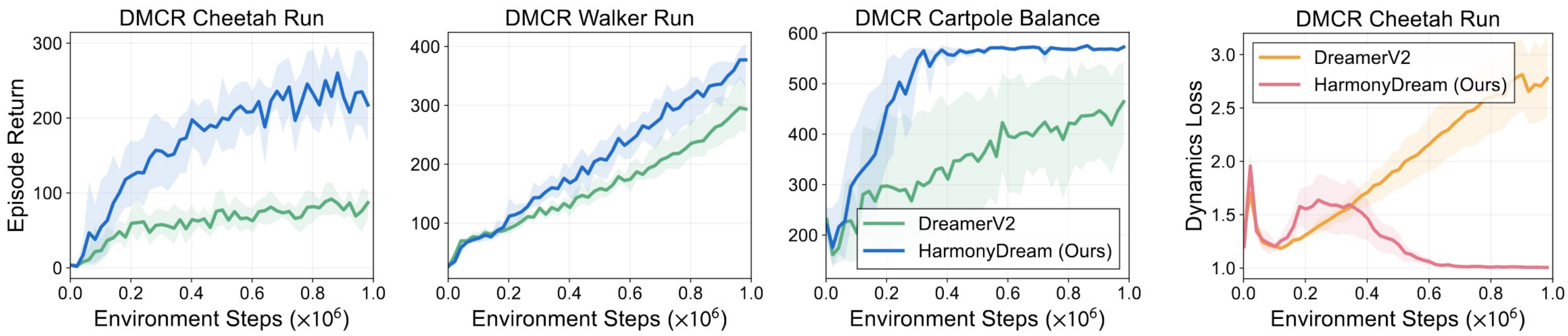
Main Results: Meta-world & RLBench



(a) Meta-world (b) RLBench

By simply adding harmonizers, HarmonyDream demonstrates superior performance in terms of both sample efficiency and final success rate

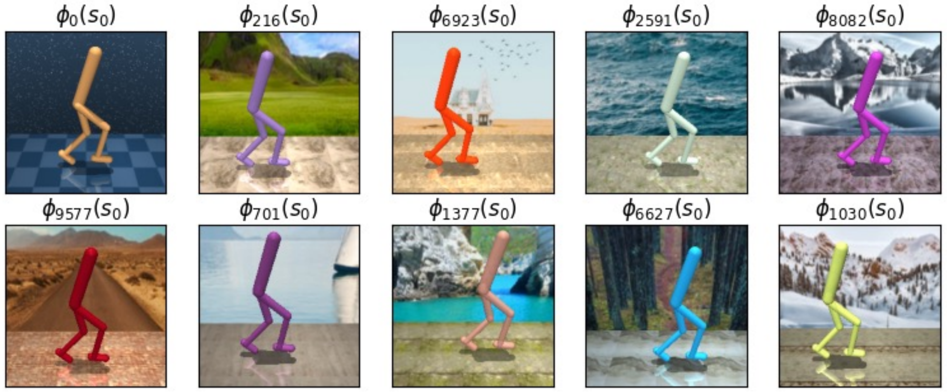
Main Results: DMC Remastered



(a) Learning curves

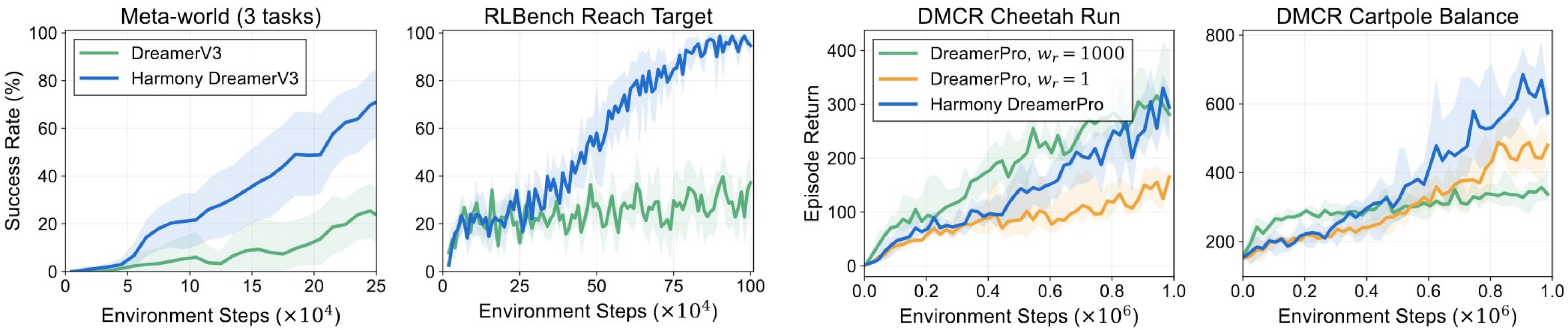
(b) Dynamics loss

On visual generalization benchmark, HarmonyDream bypasses distractors in observations and can learn task-centric transitions more easily.



Visual generalization benchmark: Seven visual factors randomly initialized on each episode

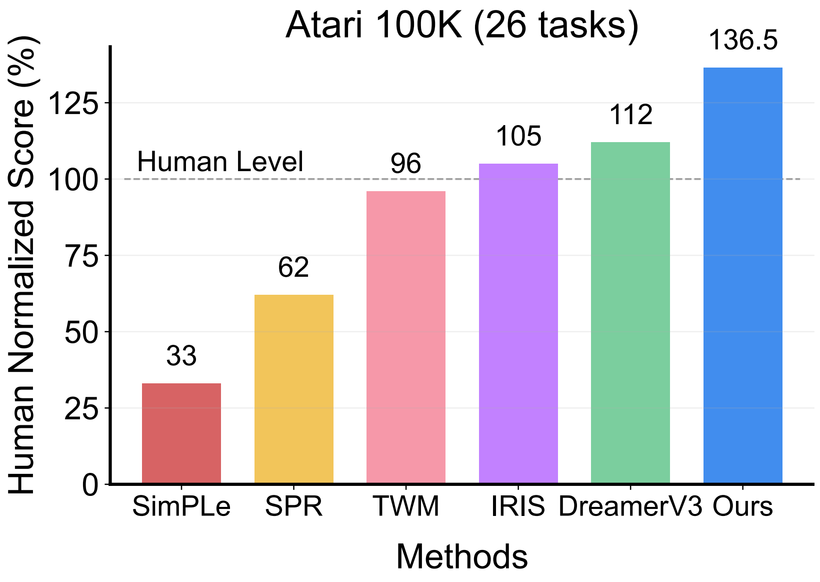
Generality to Base Model-based RL Methods

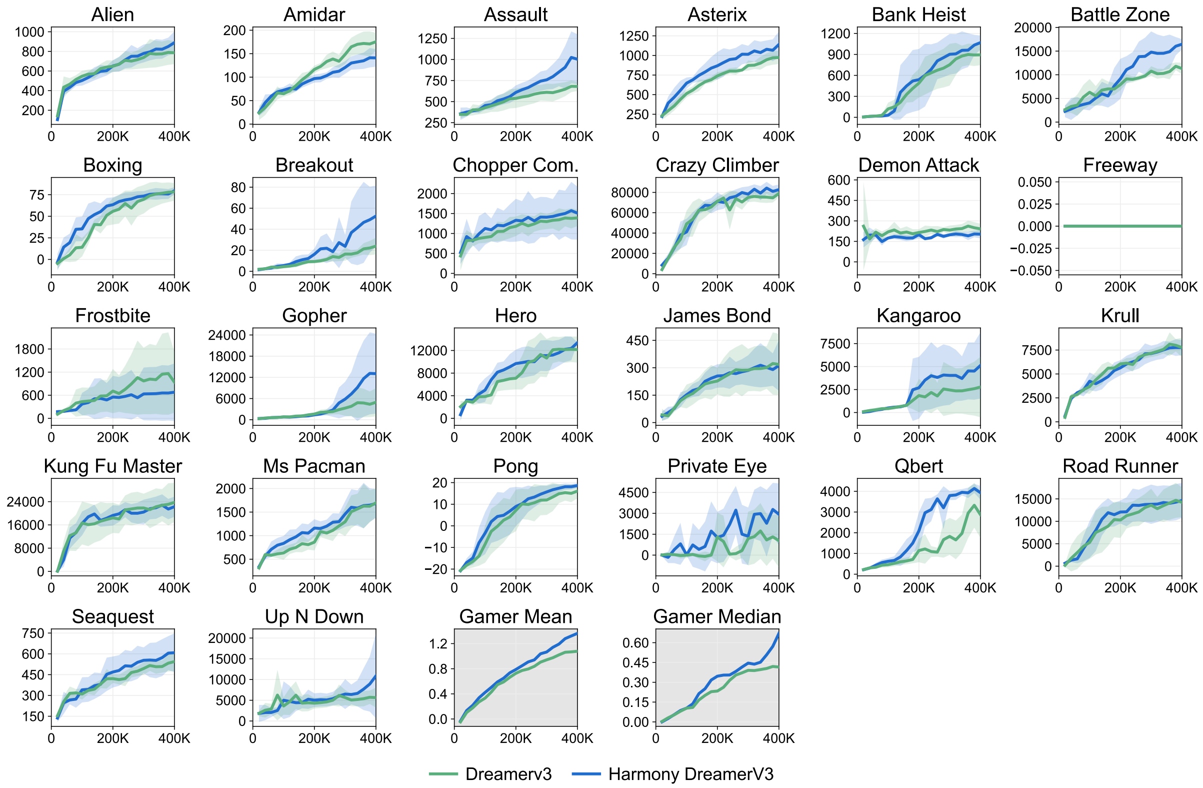


HarmonyDream exhibits excellent generality to DreamerV3, significantly boosting sample efficiency.

Although DreamerPro also leverages a high reward coeff ($w_r = 1000$), HarmonyDream still performs better on average.

Harmony DreamerV3 on Atari100K

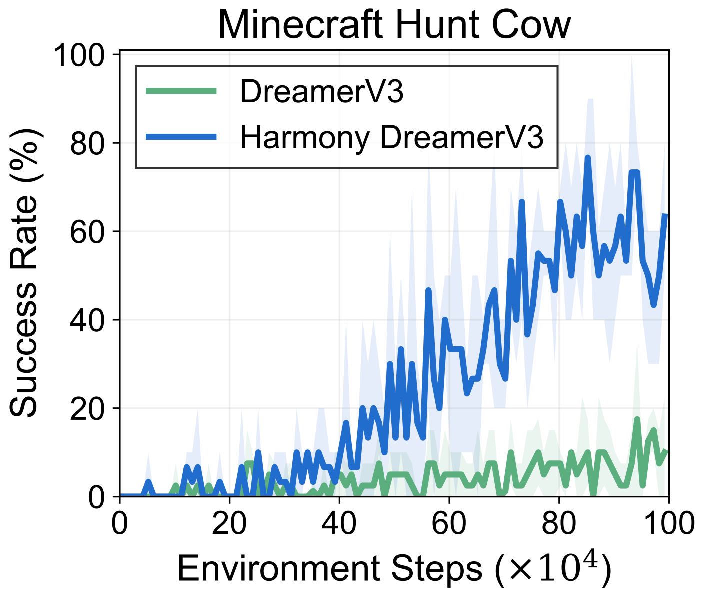




Harmony DreamerV3 significantly improves DreamerV3's performance, setting a new state of the art.

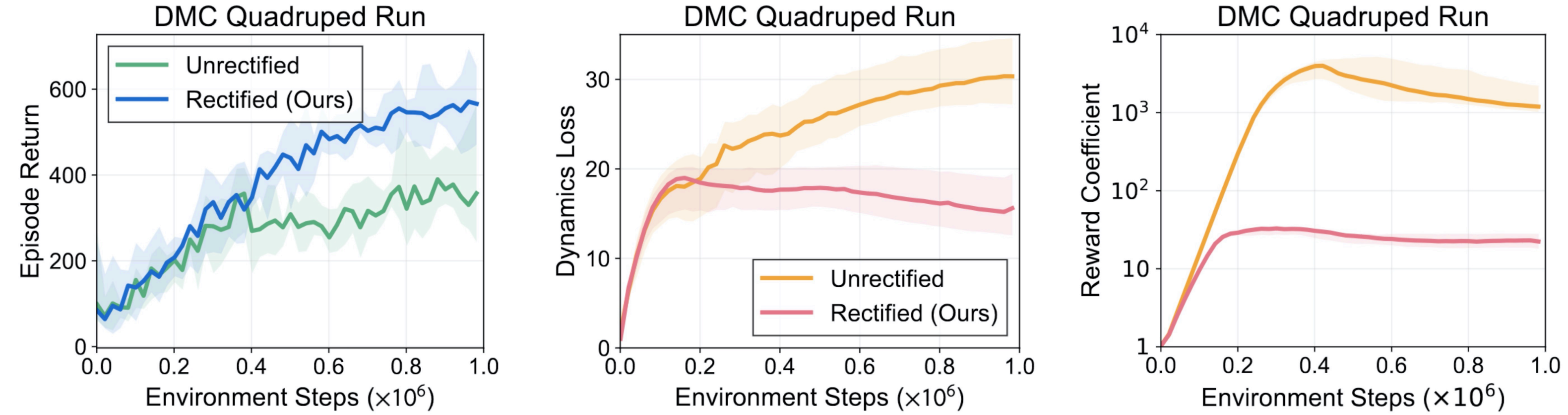
Either matching or surpassing DreamerV3 in 23/26 tested environments.

Harmony DreamerV3 on Minecraft

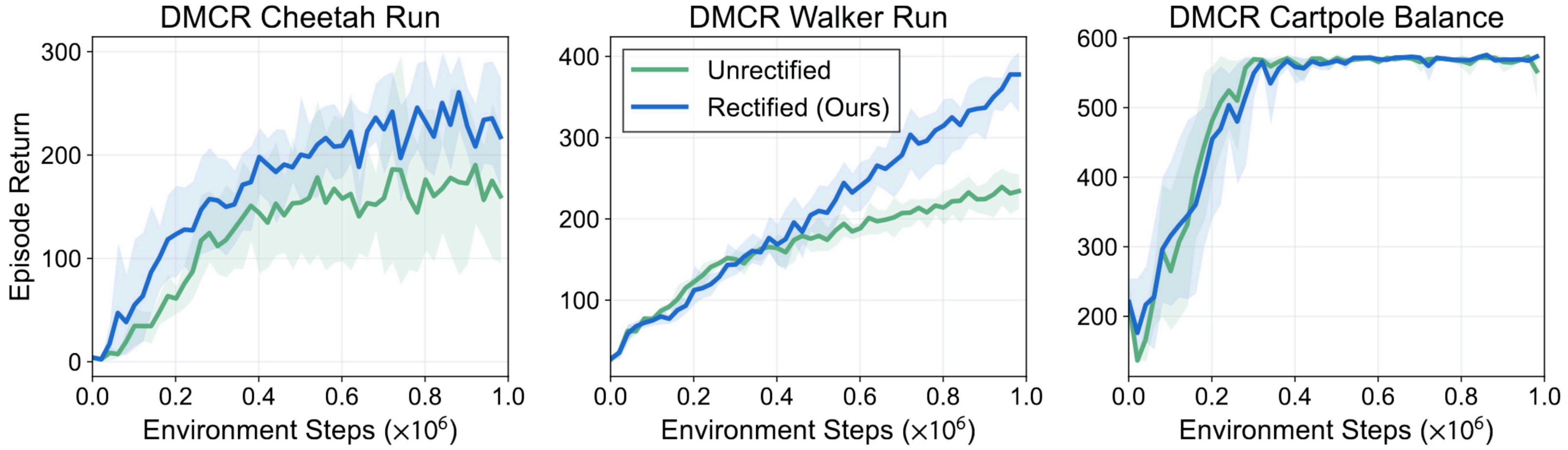


Harmony DreamerV3 successfully learns a basic skill *Hunt Cow* within 1M interactions, while DreamerV3 fails.

Ablation on Rectified Harmonious Loss

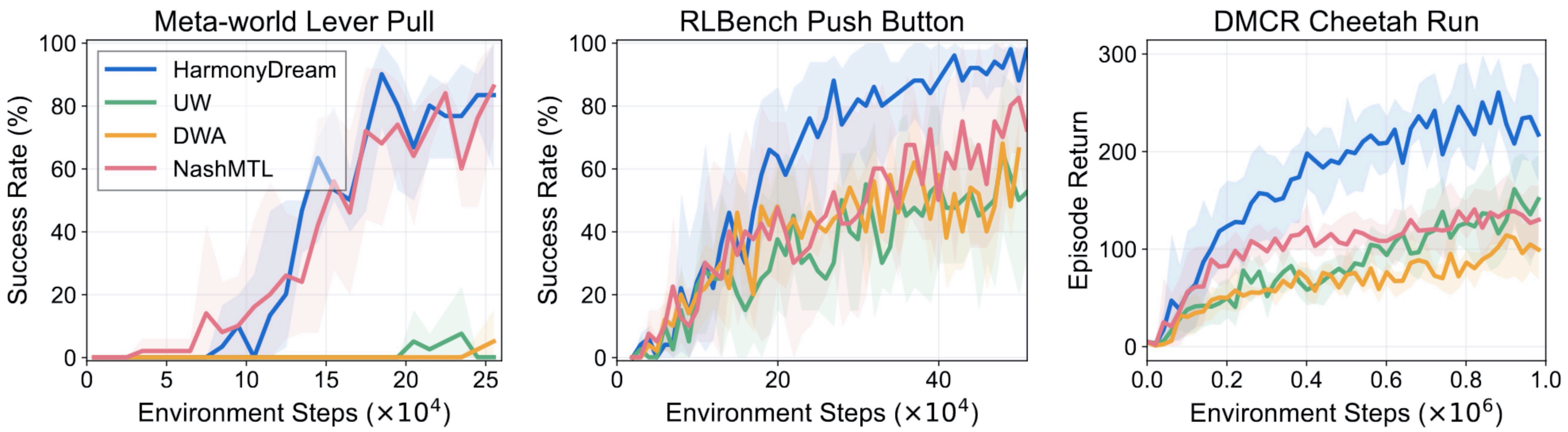


(a) Learning curves. (b) Dynamics loss. (c) Reward coefficient.



Using a regularization term of $\log(1+\sigma_i)$ instead of $\log \sigma_i$ is essential to maintaining a proper balance between tasks.

Comparison to Multi-task Learning Methods



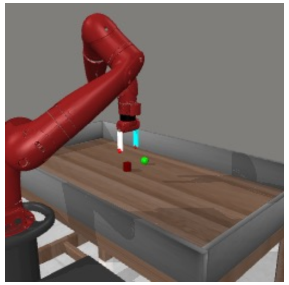
Takeaways:

1. In world model learning, the data in the replay buffer is growing and non-stationary. Learning statistics may not accurately measure learning progress.
2. Loss coefficients in world model learning needs to be properly rectified. Extreme loss weights usually leads to inferior performance.
3. HarmonyDream's improvement mainly attributes to balancing two modeling tasks, instead of solely tuning the dynamics loss.

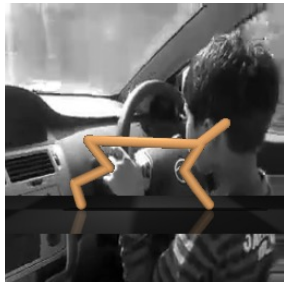
Applicability of HarmonyDream

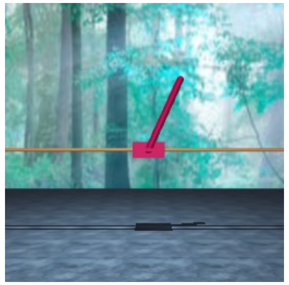
Typical realistic scenarios:

- ✓ **Fine-grained task-relevant observations**: Robotics manipulation tasks and video games require accurately modeling interactions with **small objects**.



- ✓ **Highly varied task-irrelevant observations**: **Redundant visual components** can easily distract visual agents if task-relevant information is not emphasized correctly.





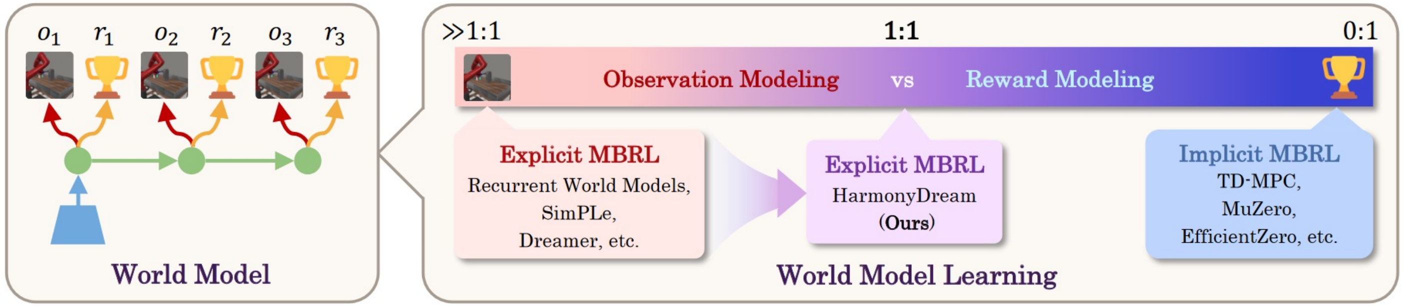
- ✓ **Hybrid of both**: More difficult **open-world** tasks (e.g., Minecraft) can encounter both, including small target entities and abundant visual details.

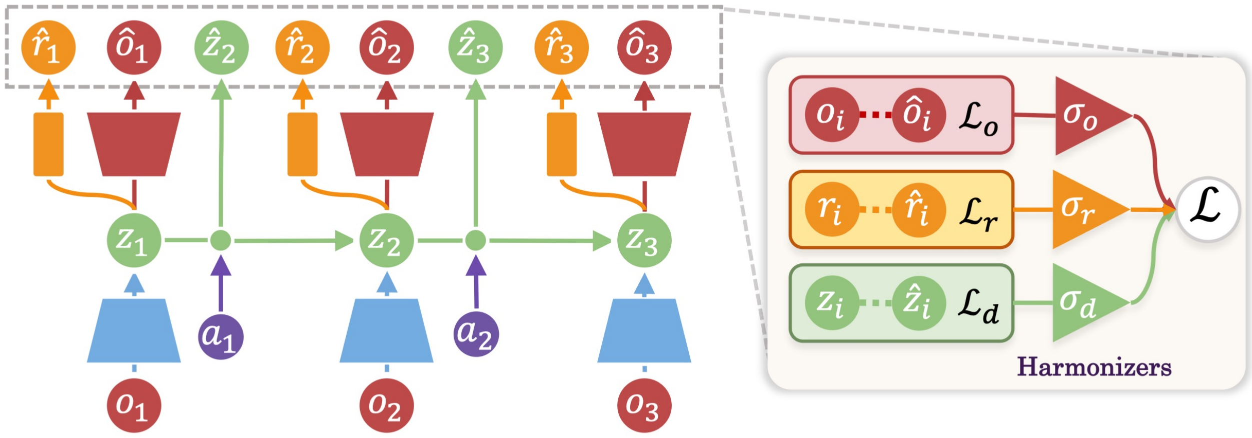
Summary

A multi-task view of world models

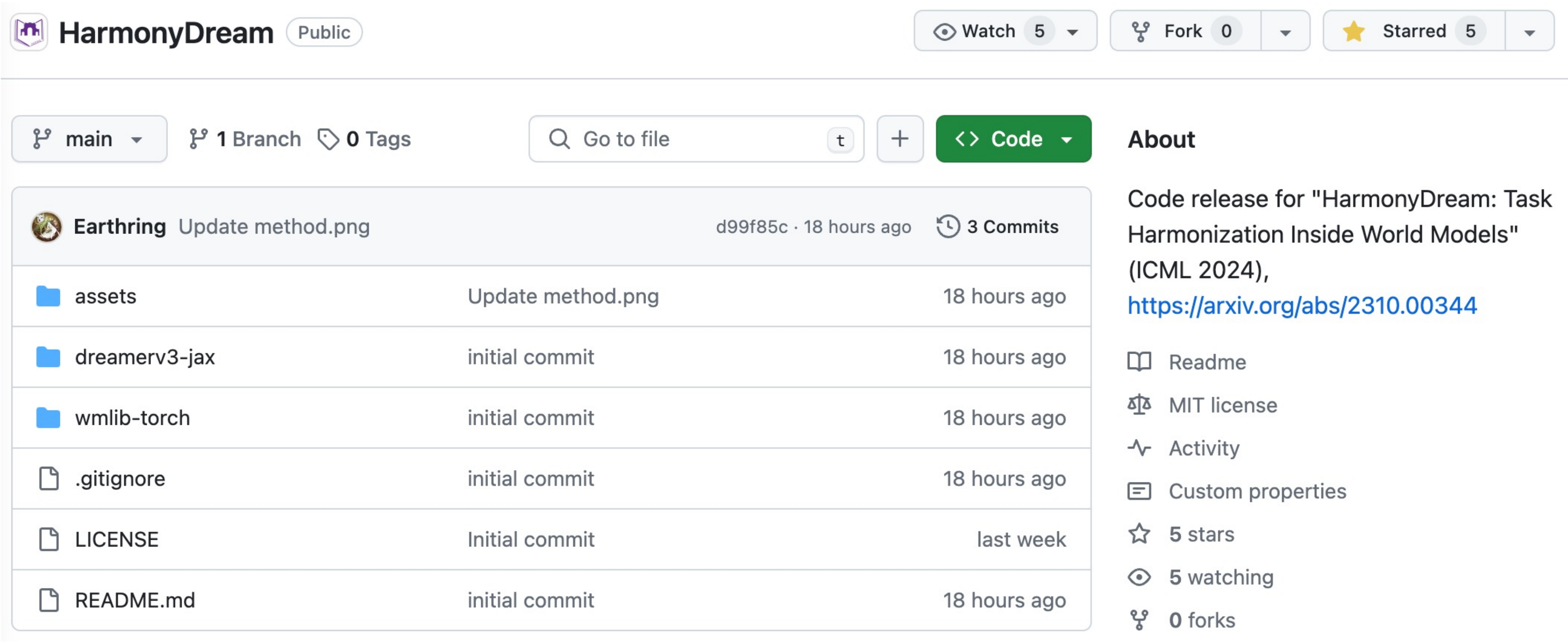
↓ mitigate task domination

A simple yet effective world model learning approach





Open Source



https://github.com/thuml/HarmonyDream

Unified implementations of DreamerV2 and DreamerV3 in PyTorch
with plug-and-play HarmonyDream

Thank You!

Contact:

mhy22@mails.tsinghua.edu.cn

wujialong0229@gmail.com

Researcher who tried HarmonyDream:

"It was super easy to implement";

"It works very smoothly"

Machine Learning Group, School of Software, Tsinghua University

http://ise.thss.tsinghua.edu.cn/~mlong/

清華大學 Tsinghua University