Towards General World Models:

Pre-training, Multi-Modality, and Scalable Architecture

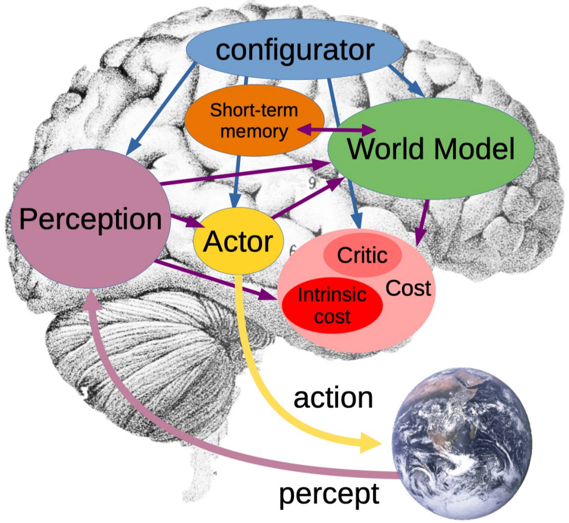
Mingsheng Long

School of Software, Tsinghua University

July 2024

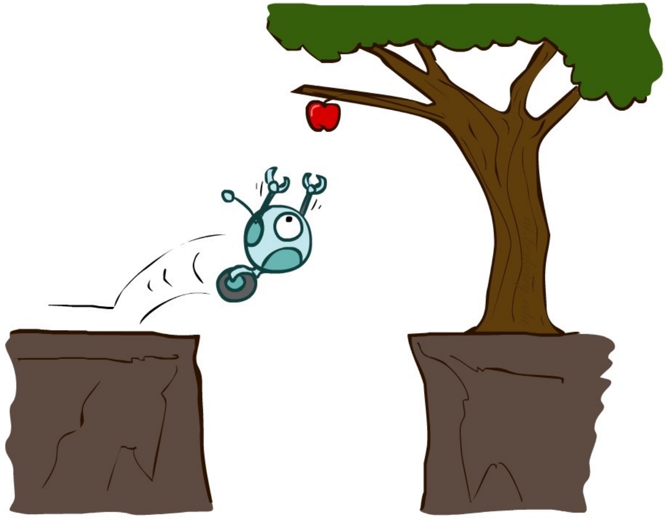
清華大學
Tsinghua University

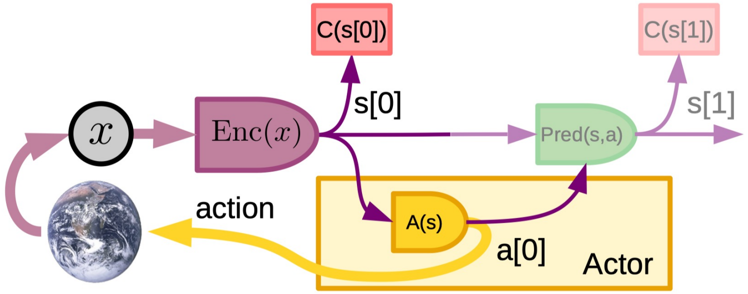
World Models: From System-1 to System-2



World Models:

internal models of how the world works



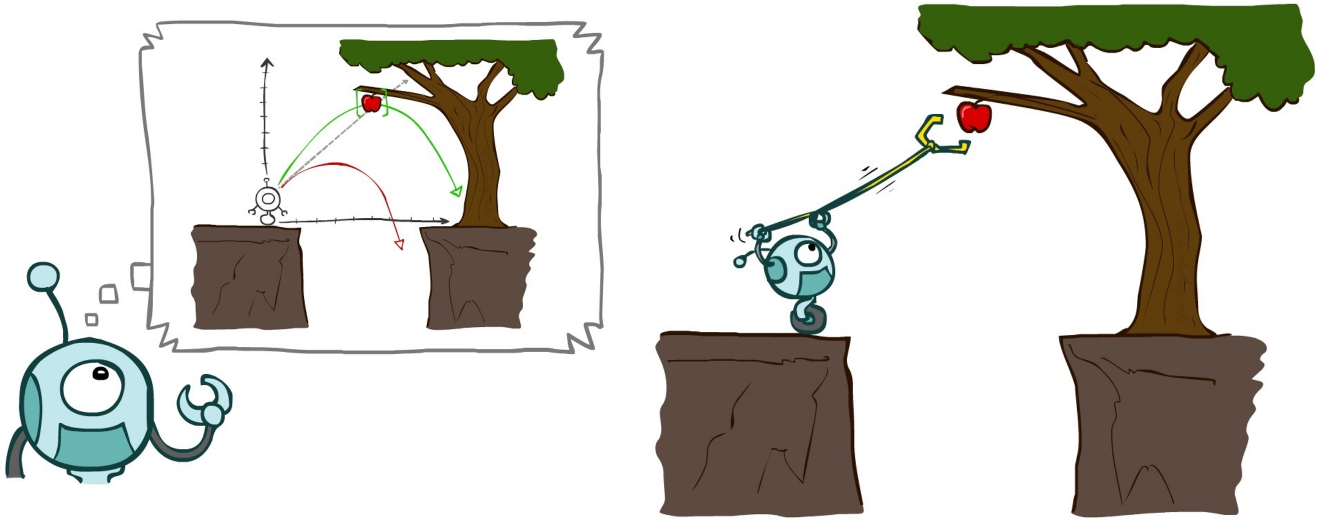


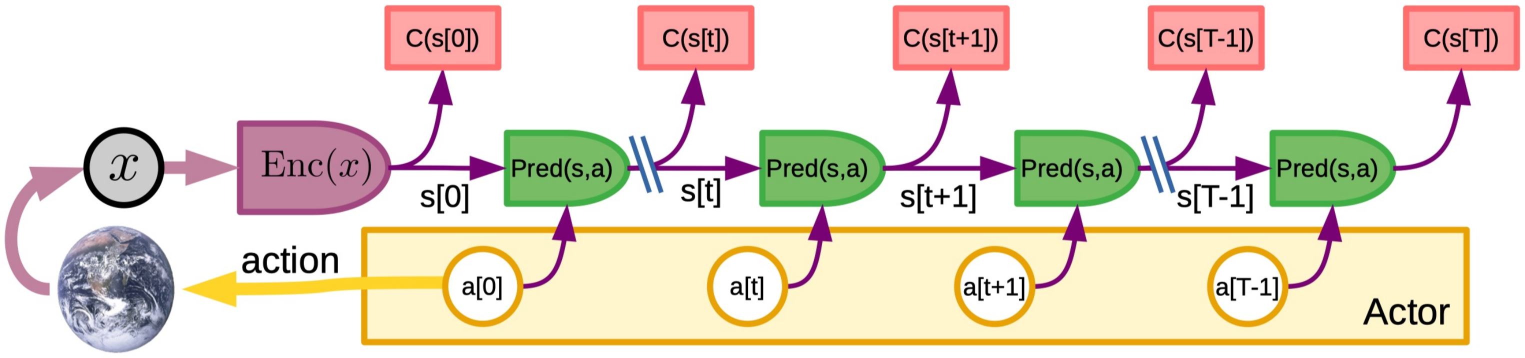
System-1 Agent (Reflex):

Not utilize the world model nor the cost.

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

World Models: From System-1 to System-2





System-2 Agent (Planning):

Act through an optimization procedure running the world model.

Amortized Inference:

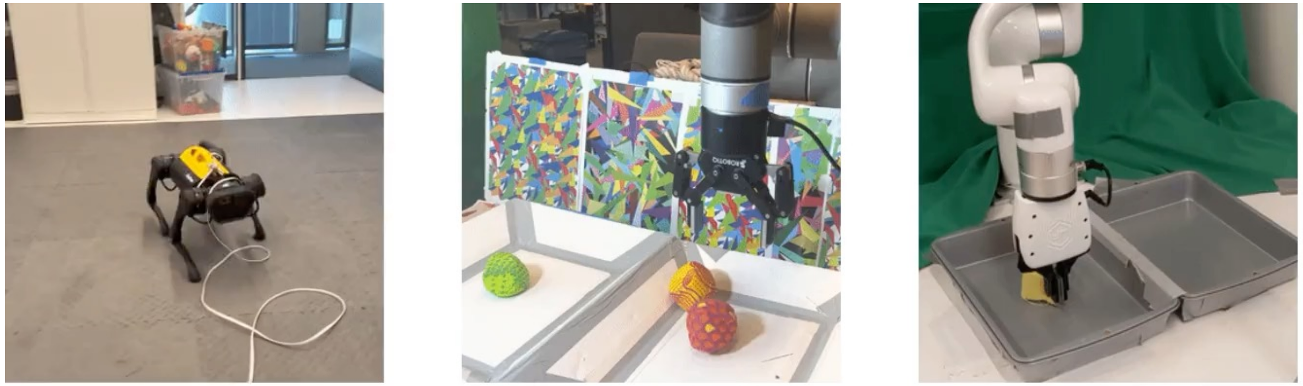
A policy module mimicking the optimal actions

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

World Models: Applications

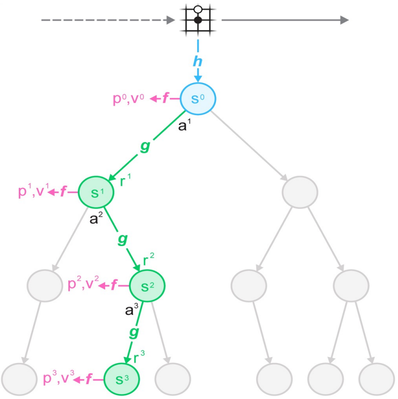
Autonomous Driving

Alex Kendall. CVPR 2023 E2EAD Workshop.



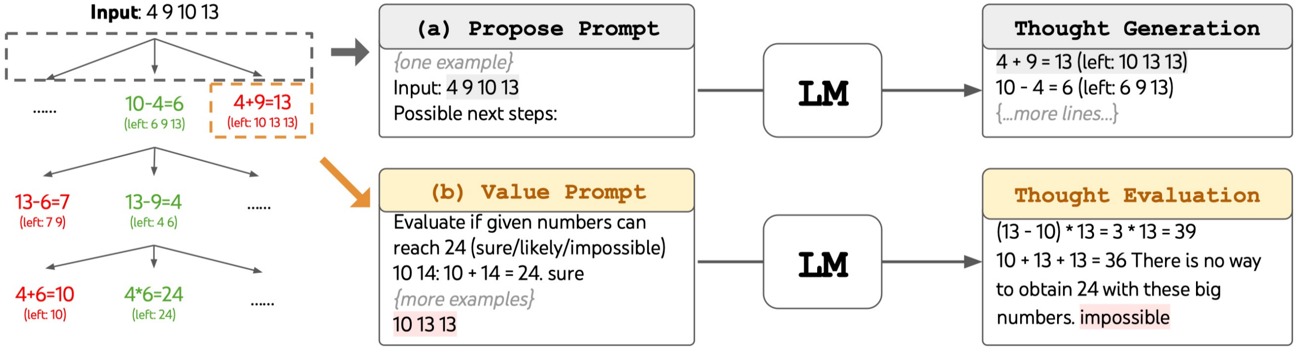
Robotics

Wu, Philipp, et al. CoRL 2022.



Games

Schrittwieser, Julian, et al. Nature 588 (2020).

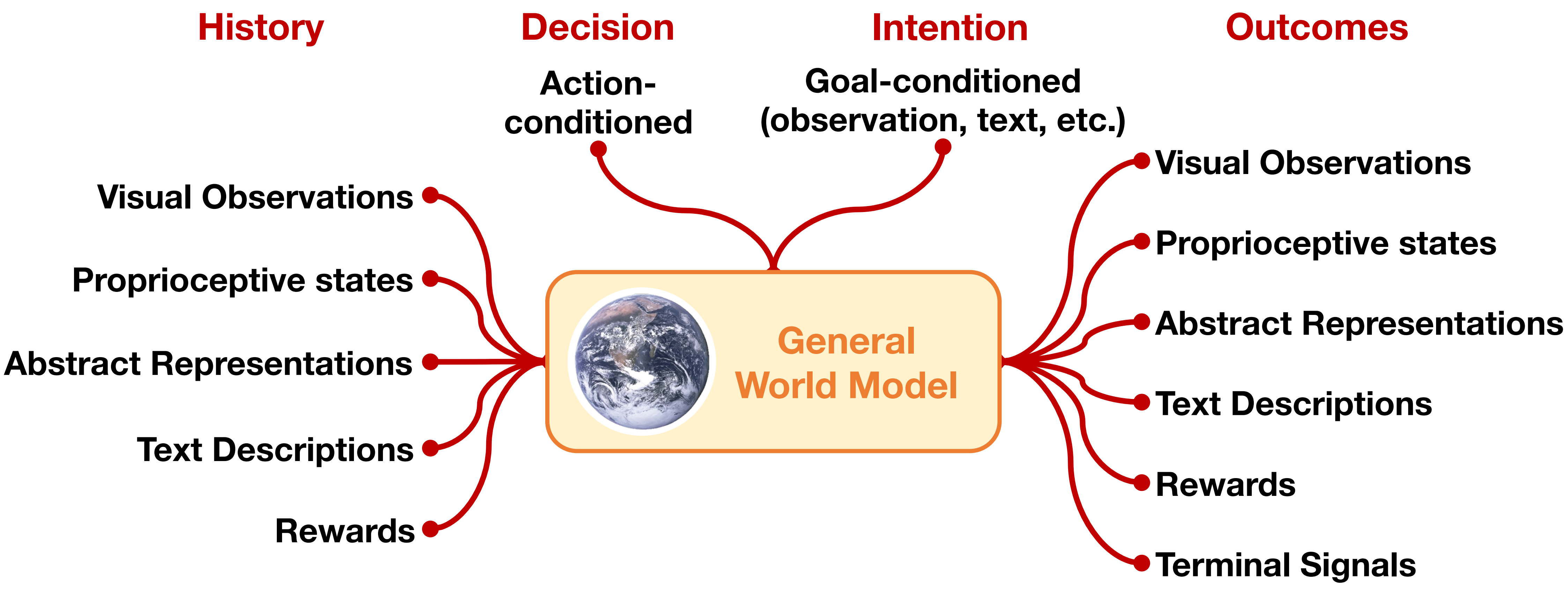


Large Language Models

Yao, Shunyu, et al. arXiv 2023.

General World Models

Any-to-Any Prediction with Any Conditions



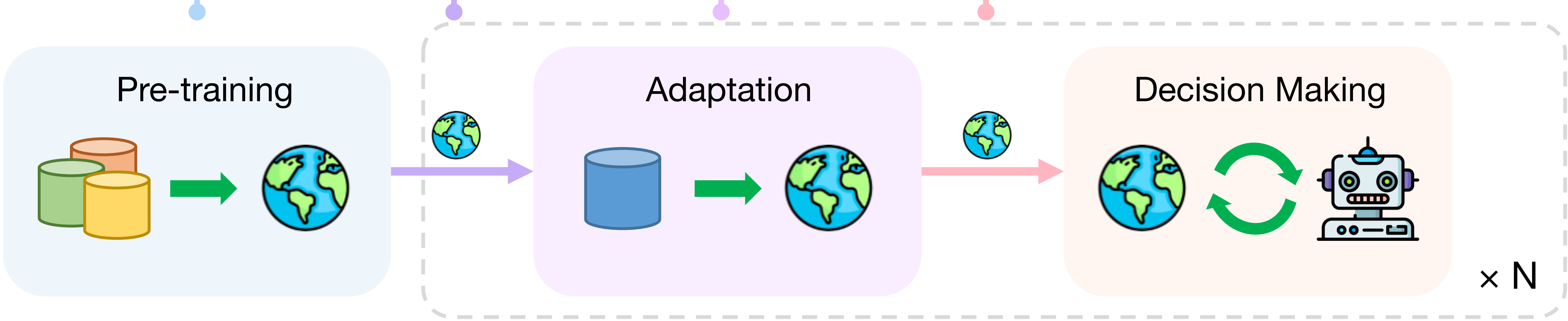
Life Cycle of A General World Model

Challenge 1: Scalable architecture for pre-training ⟵ **iVideoGPT**

Challenge 2: Knowledge transfer from pre-training ⟵ **ContextWM**

Challenge 3: multimodal world model learning ⟵ **HarmonyDream**

Common practice: Model-based planning or RL



NeurIPS | 2023
Thirty-seventh Conference on Neural Information Processing Systems

NEURAL INFORMATION PROCESSING SYSTEMS

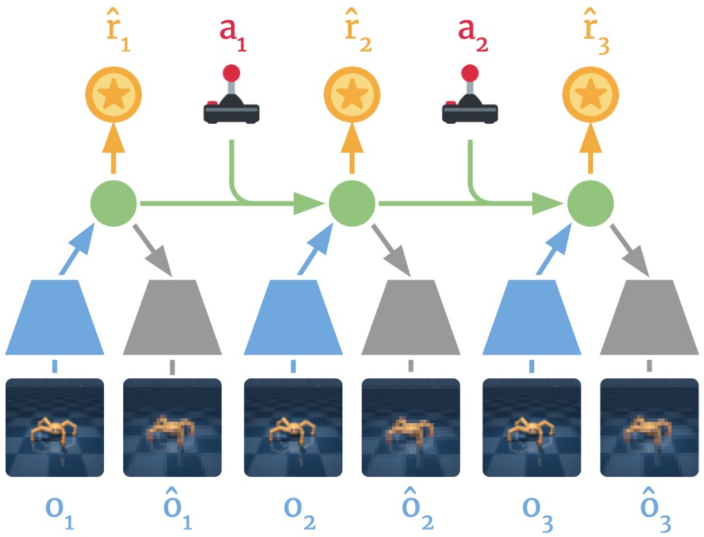
Pre-training Contextualized World Models with In-the-wild Videos for Reinforcement Learning

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

Jialong Wu,* Haoyu Ma,* Chaoyi Deng, Mingsheng Long✉
School of Software, BNRist, Tsinghua University, China
wujialong0229@gmail.com, {mhy22,dengcy23}@mails.tsinghua.edu.cn
mingsheng@tsinghua.edu.cn

清華大學
Tsinghua University

Dreamer: An Instantiation of World Models



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

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

Image decoder: $\hat{o}_t \sim p_\theta(\hat{o}_t \mid z_t)$

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

Model Learning with Sequential Variational Inference

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

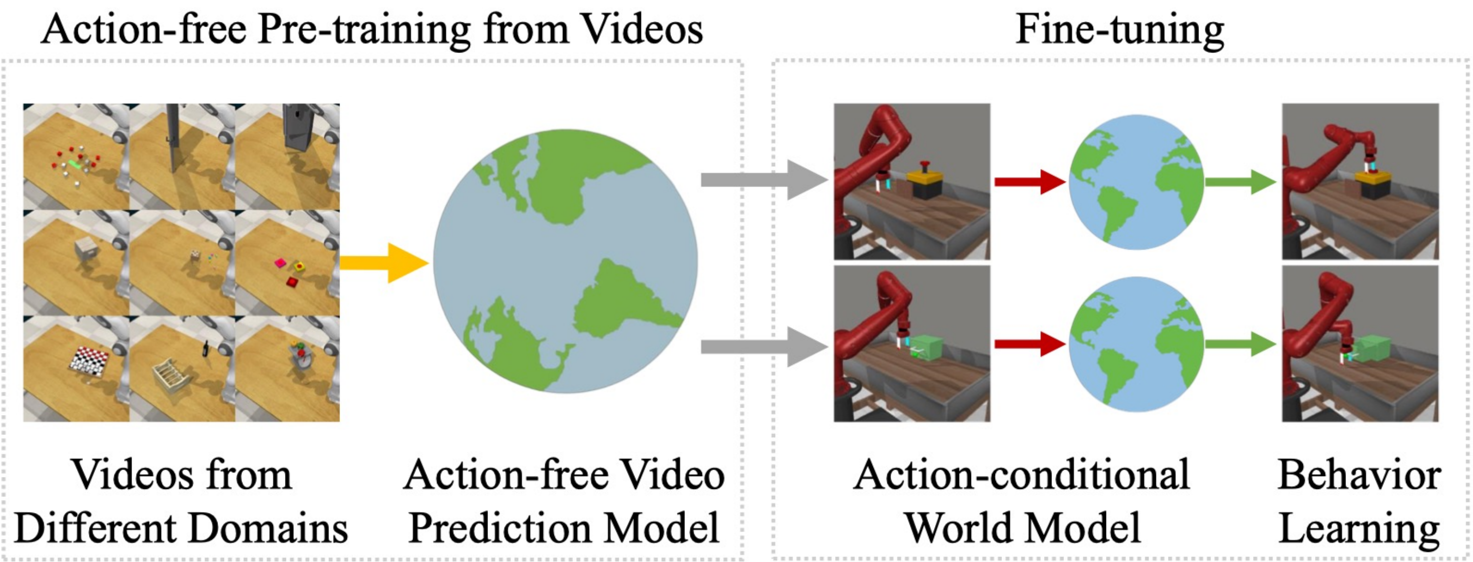
Behavior Learning: Purely on imaginary latent trajectories

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.

APV: Action-free Pre-training from Videos

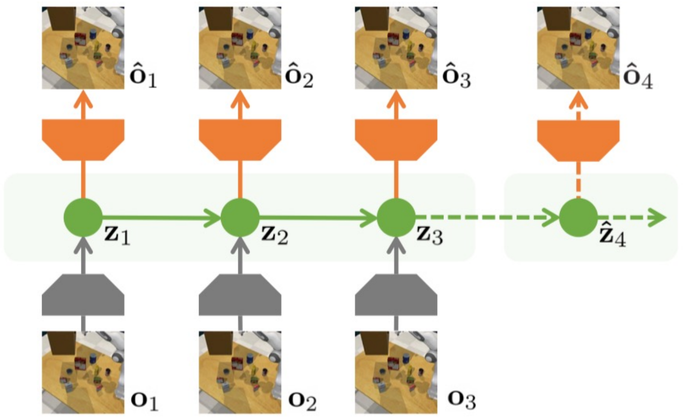
How to represent and acquire prior knowledge for RL?

Learning representations useful for understanding the dynamics via generative pretraining on videos



Seo, Younggyo, et al. Reinforcement learning with action-free pre-training from videos. ICML 2022.

APV: Action-free Pre-training from Videos



Stacked Latent Prediction Model

Action-free

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

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

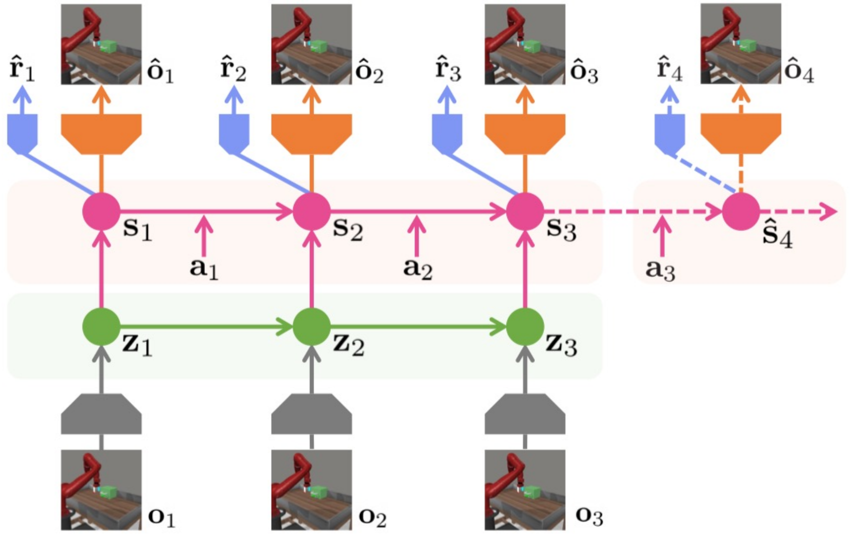
Image decoder: $p_\theta(\hat{o}_t \mid s_t)$

Action-conditional

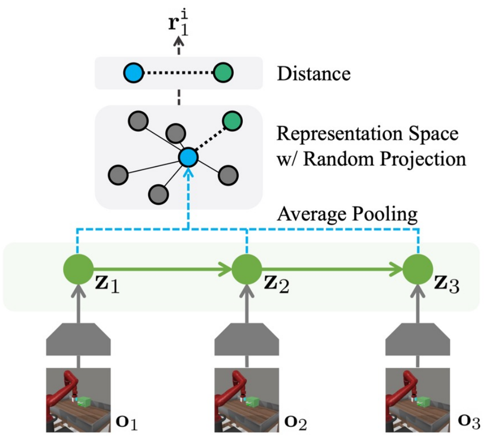
Representation: $q_\phi(s_t \mid s_{t-1}, a_{t-1}, z_t)$

Transition: $p_\phi(\hat{s}_t \mid s_{t-1}, a_{t-1})$

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



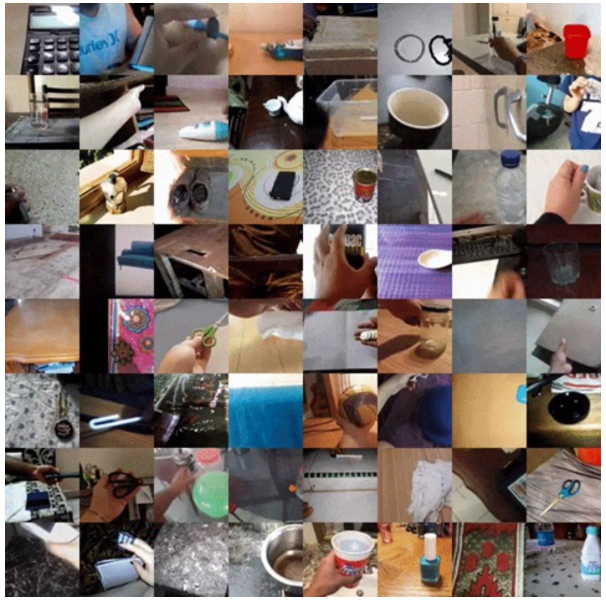
1. Pre-train an action-free latent video prediction model
2. Stack an action-conditional model when fine-tuned for MBRL
3. Video-based intrinsic bonus for better exploration



Seo, Younggyo, et al. Reinforcement learning with action-free pre-training from videos. ICML 2022.

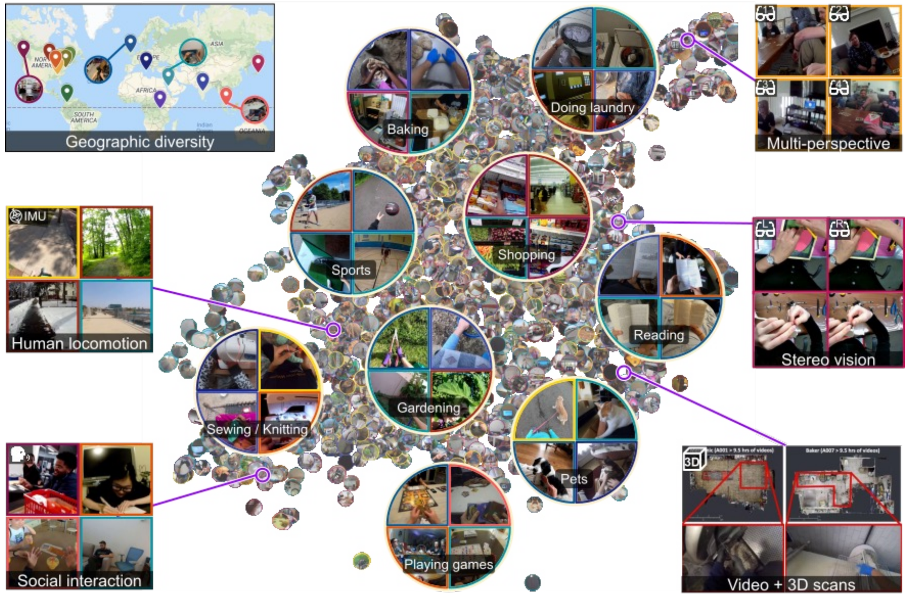
Our Work: Towards a General World Model

General world knowledge for a variety of downstream tasks
from abundant in-the-wild videos on the Internet



Something-Something V2

Goyal et al. ICCV 2017



Ego4D

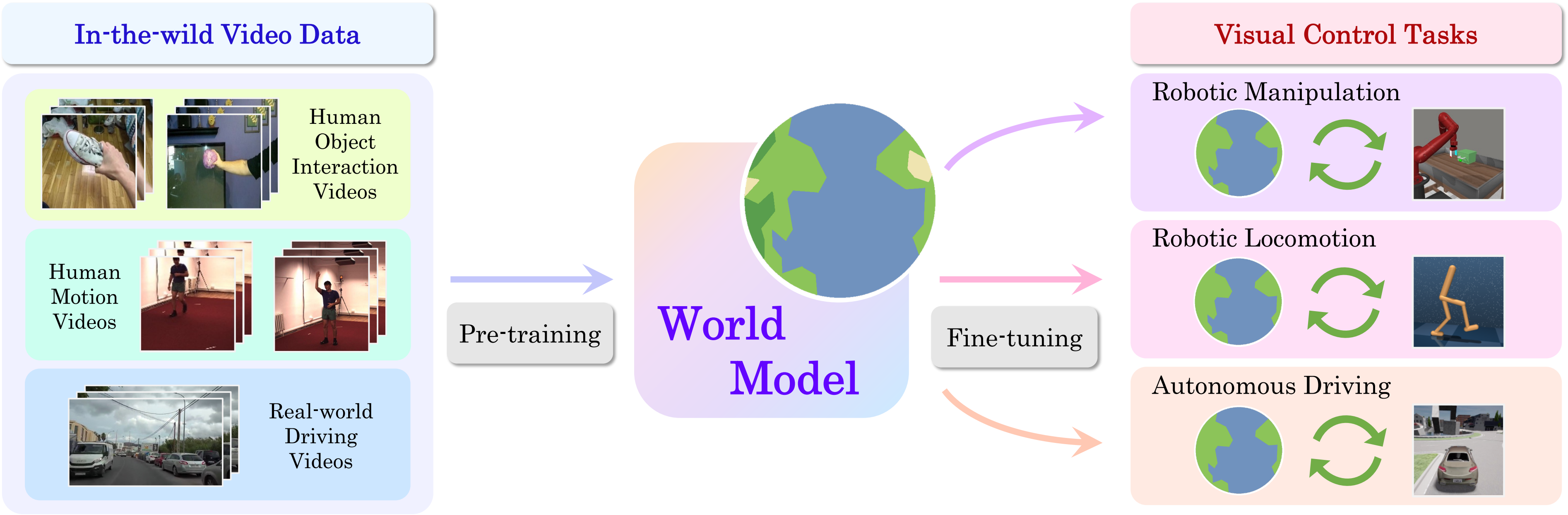
Grauman et al., Facebook AI. CVPR 2022

- ✓ Task-agnostic
- ✓ Widely available
- ✓ Broad Knowledge

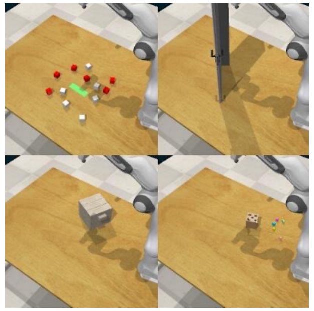
IPV: In-the-wild Pre-training from Videos

Towards a general world model:

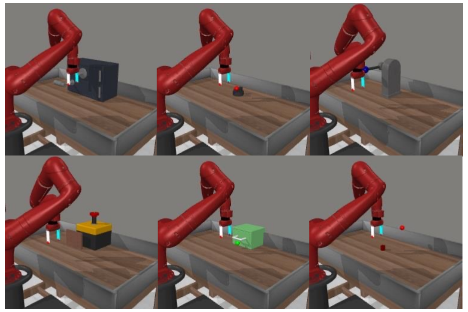
- How to overcome the visual complexity and diversity?
- What is the shared knowledge transferable from in-the-wild video domain to visual control tasks?



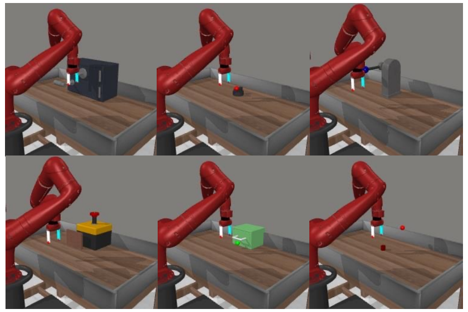
Failure of Plain World Models on In-the-wild Videos

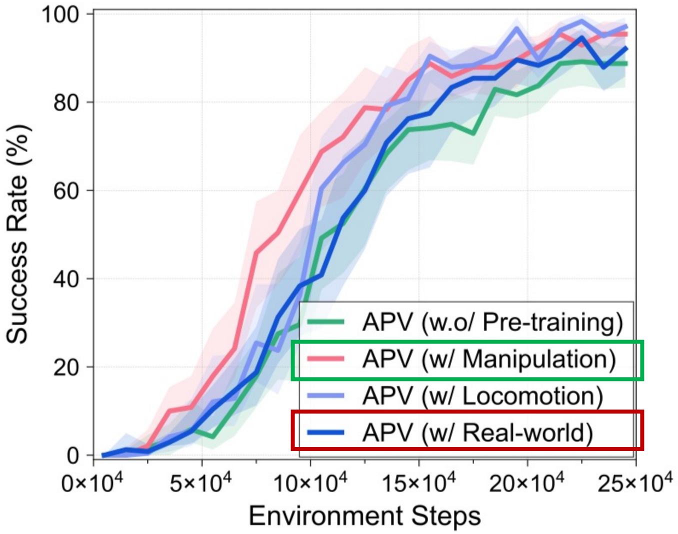


Positive Transfer



No Transfer





Why pre-training fails?

Seo et al.: Video prediction model suffers from severe underfitting

Wasting model capacity on modeling low-level **contextual** information!

Seo, Younggyo, et al. Reinforcement learning with action-free pre-training from videos. ICML 2022.

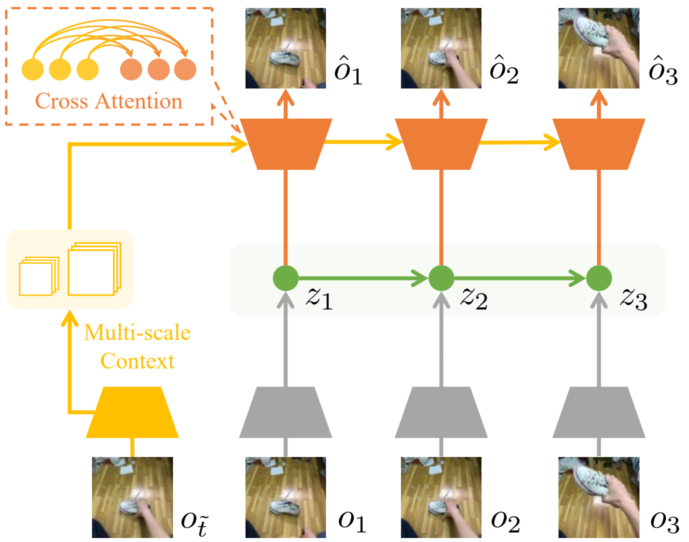
Contextualized World Models (ContextWM)

Overview:

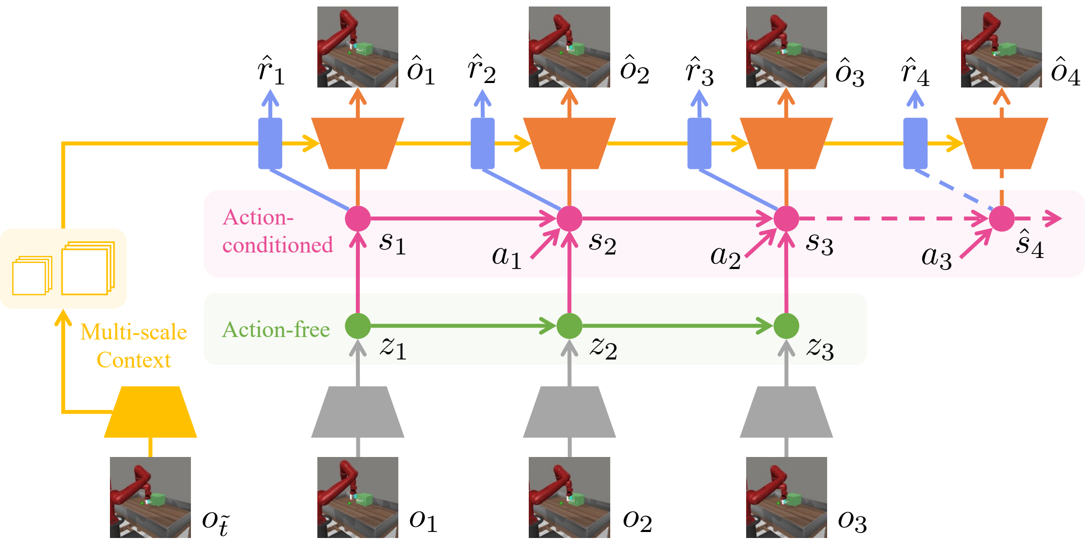
ContextWM empowers the **image decoder** by incorporating a **context encoder** that operates in parallel with the **latent dynamics model**

✓ Less inductive bias

✓ Diverse datasets & tasks



Step 1. Pre-training with in-the-wild videos by action-free video prediction



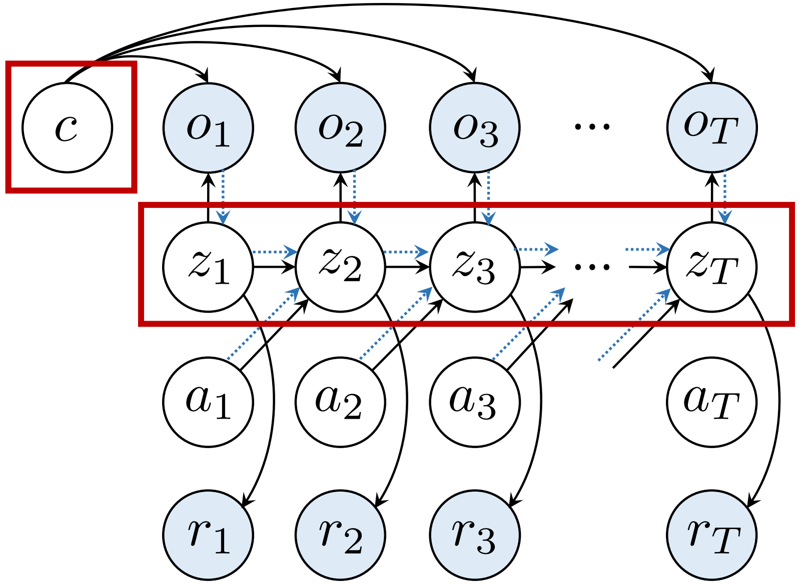
Step 2. Fine-tuning on downstream visual control tasks with MBRL

Contextualized Latent Dynamics Models

Our insight:

Even across distinct scenes (**contexts**), the environment **dynamics** and physics share a similar structure.

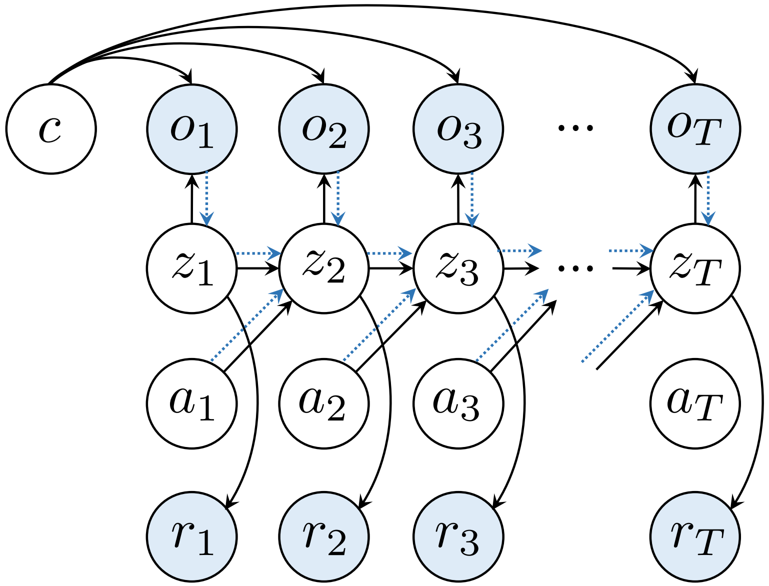
Distinct
contexts



Similar
dynamics

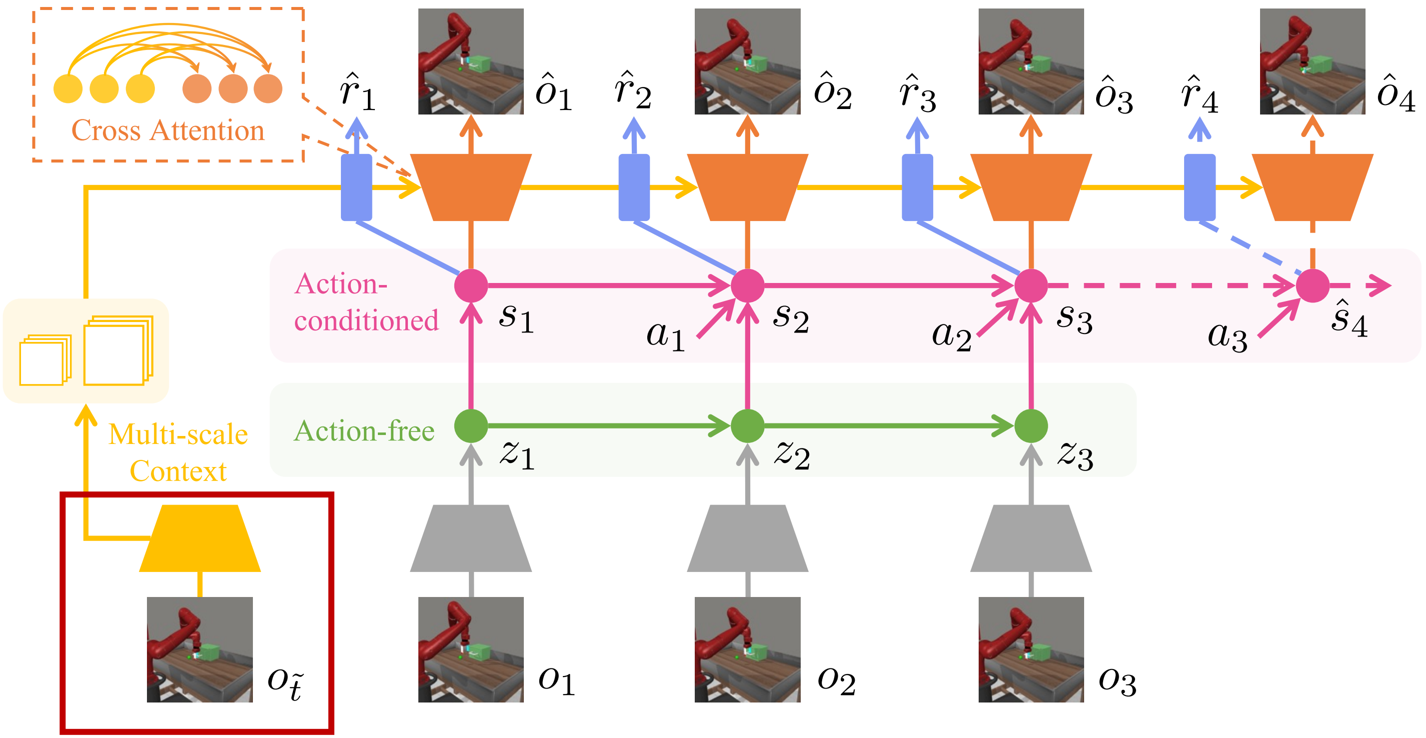
Contextualized Latent Dynamics Models

$$\mathcal{L}(\theta) \doteq \underbrace{\mathbb{E}_{q_\theta(z_{1:T}|a_{1:T},o_{1:T})}}_{\text{context-unaware latent inference}} \Big[\sum_{t=1}^{T} \Big(\underbrace{-\ln p_\theta(o_t \mid z_t, c)}_{\text{contextualized image loss}} - \ln p_\theta(r_t \mid z_t) + \beta_z \, \mathrm{KL}\left[q_\theta(z_t \mid z_{t-1}, a_{t-1}, o_t) \,\|\, p_\theta(\hat{z}_t \mid z_{t-1}, a_{t-1})\right] \Big) \Big]$$



- Learn with ELBO of conditional $\ln p_\theta(o_{1:T}, r_{1:T} \mid a_{1:T}, c)$ without the need to model the context distribution
- **Contextualized** image decoders with rich information beyond the expressiveness of latent variables
- Latent **dynamics** inference concentrates on essential temporal variations

Contextualized World Models: An Implementation



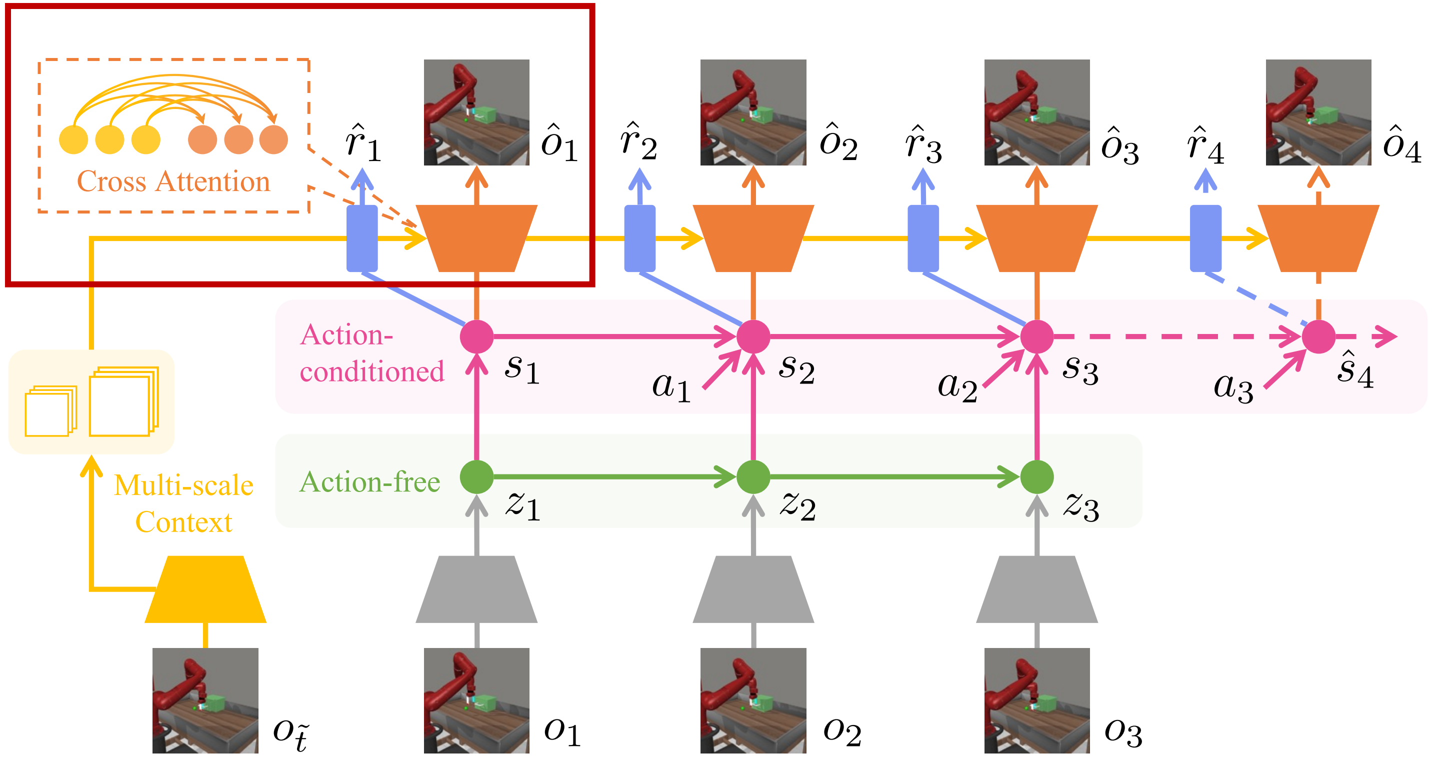
Context formulation:

A random single frame from the trajectory segment

$$c \doteq o_{\tilde{t}}, \ \tilde{t} \sim \text{Uniform}\{T\}$$

By random selection, the context encoder learns to be robust to temporal variations

Contextualized World Models: An Implementation



Multi-scale cross-attention:

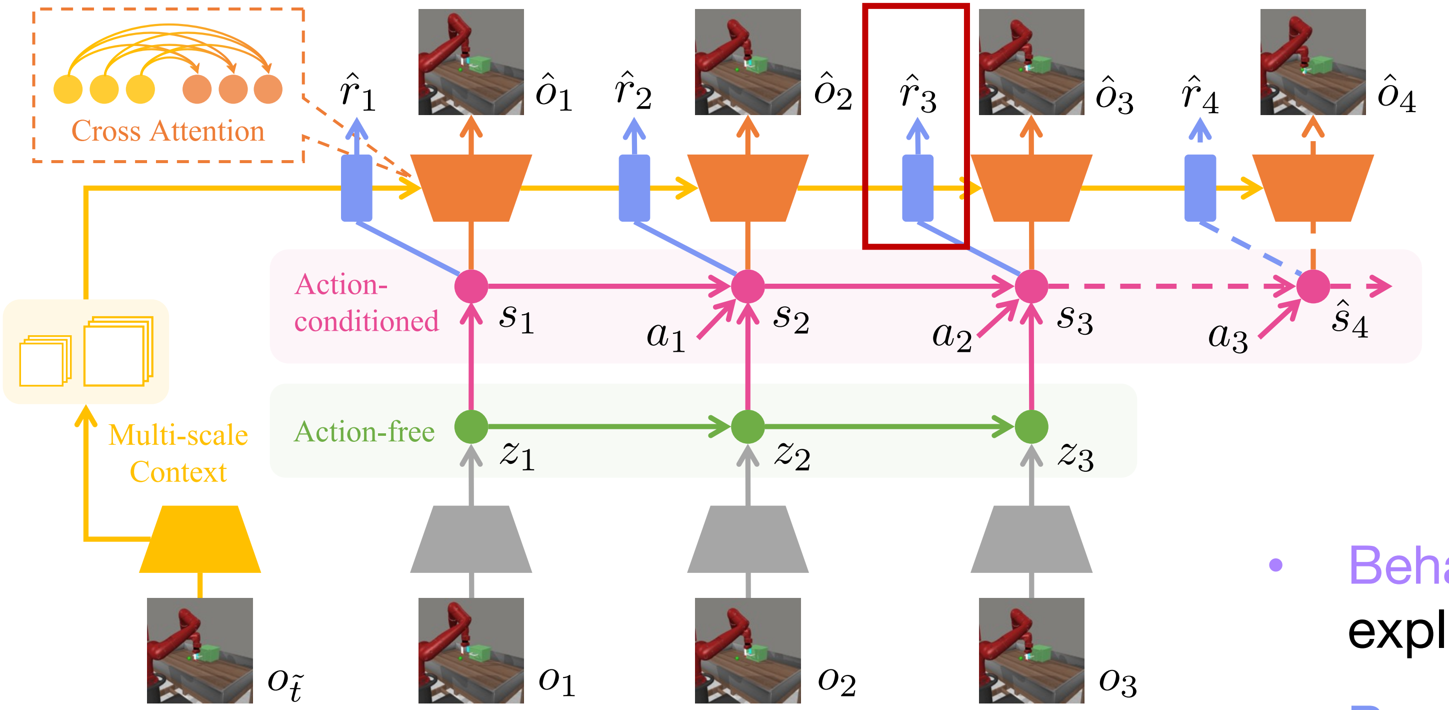
1. U-Net-style multi-scale feature shortcuts
2. Instead of naive concatenation forcing a spatial alignment, adaptive cross-attention mechanism is utilized

$$\text{Flatten: } Q = \operatorname{Reshape}(X) \in \mathbb{R}^{hw \times c}, \;\; K = V = \operatorname{Reshape}(Z) \in \mathbb{R}^{hw \times c}$$

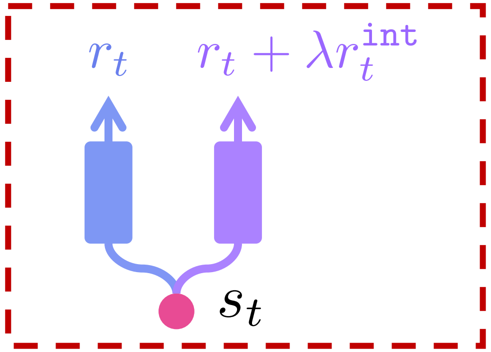
$$\text{Cross-Attention: } R = \operatorname{Attention}\left(QW^Q, KW^K, VW^V\right) \in \mathbb{R}^{hw \times c}$$

$$\text{Residual-Connection: } X = \operatorname{ReLU}\left(X + \operatorname{BatchNorm}\left(\operatorname{Reshape}(R)\right)\right) \in \mathbb{R}^{c \times h \times w}.$$

Contextualized World Models: An Implementation



Dual reward predictors:



- Behavioral reward predictor: exploratory reward for behavior learning
- Representative reward predictor: pure task reward for task-relevant representation learning

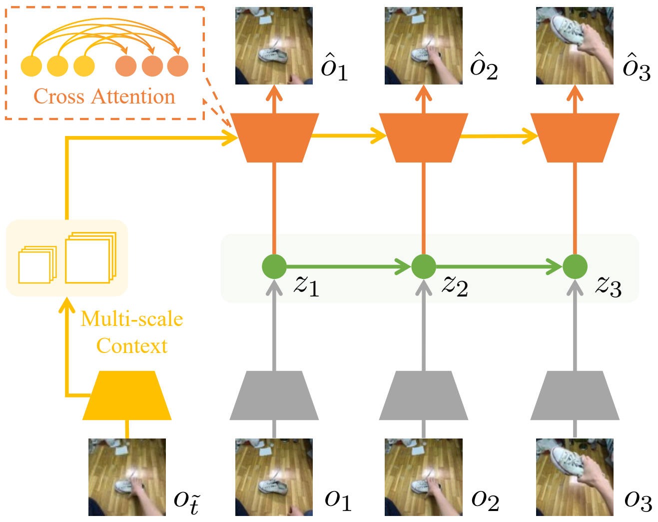
Contextualized World Models: An Implementation

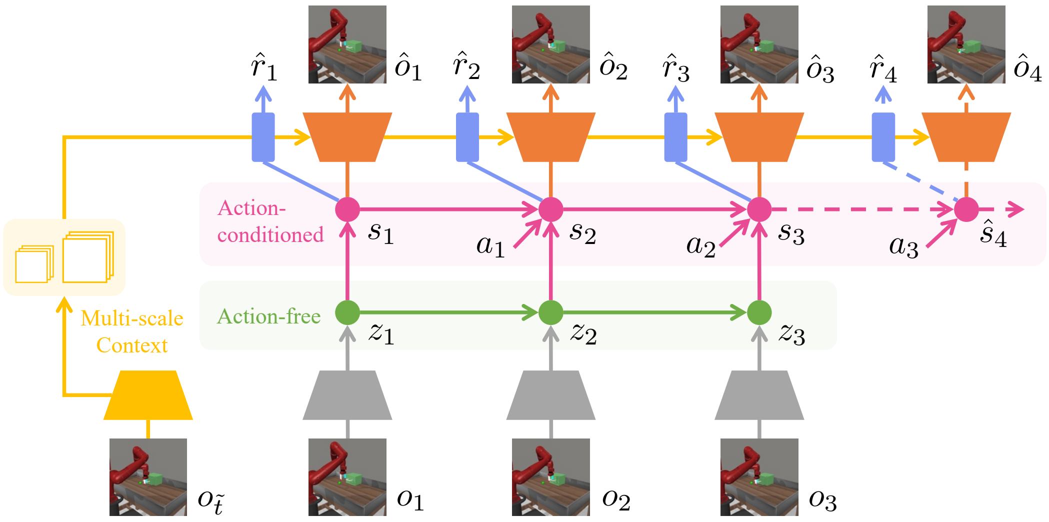
Overall objective:

$$\mathcal{L}^{\texttt{CWM}}(\phi,\varphi,\theta) \doteq \underbrace{\mathbb{E}_{q_\phi(s_{1:T}\,|\,a_{1:T},z_{1:T}),\,q_\theta(z_{1:T}\,|\,o_{1:T})}}_{\text{context-unware latent inference}} \Big[\sum_{t=1}^{T} \Big(\underbrace{-\ln p_\theta(o_t\,|\,s_t,c)}_{\text{contextualized image loss}}$$

$$\underbrace{-\ln p_\phi(r_t+\lambda r_t^{\texttt{int}}\,|\,s_t)}_{\text{behavioral reward loss}} \; \underbrace{-\beta_r \ln p_\varphi(r_t\,|\,s_t)}_{\text{representative reward loss}} \; \underbrace{+\beta_z\,\mathrm{KL}\left[q_\theta(z_t\,|\,z_{t-1},o_t)\,\|\,p_\theta(\hat{z}_t\,|\,z_{t-1})\right]}_{\text{action-free KL loss}}$$

$$\underbrace{+\beta_s\,\mathrm{KL}\left[q_\phi(s_t\,|\,s_{t-1},a_{t-1},z_t)\,\|\,p_\phi(\hat{s}_t\,|\,s_{t-1},a_{t-1})\right]}_{\text{action-conditional KL loss}} \Big) \Big].$$





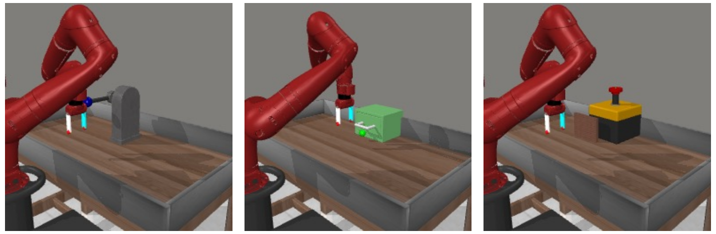
Experiments: Diverse Datasets & Tasks

Something-Something V2
Goyal et al. ICCV 2017

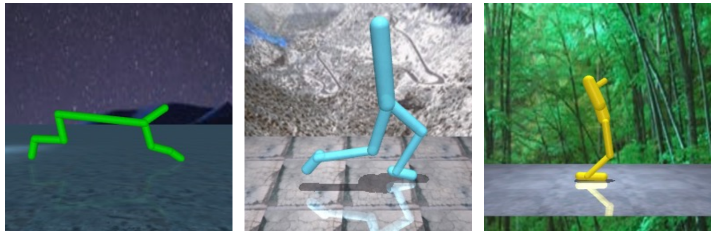
Human3.6M
Ionescu et al. TPAMI 2014

YouTube Driving
Zhang et al. ECCV 2022

Transfer



Meta-World
Yu et al. CoRL 2020

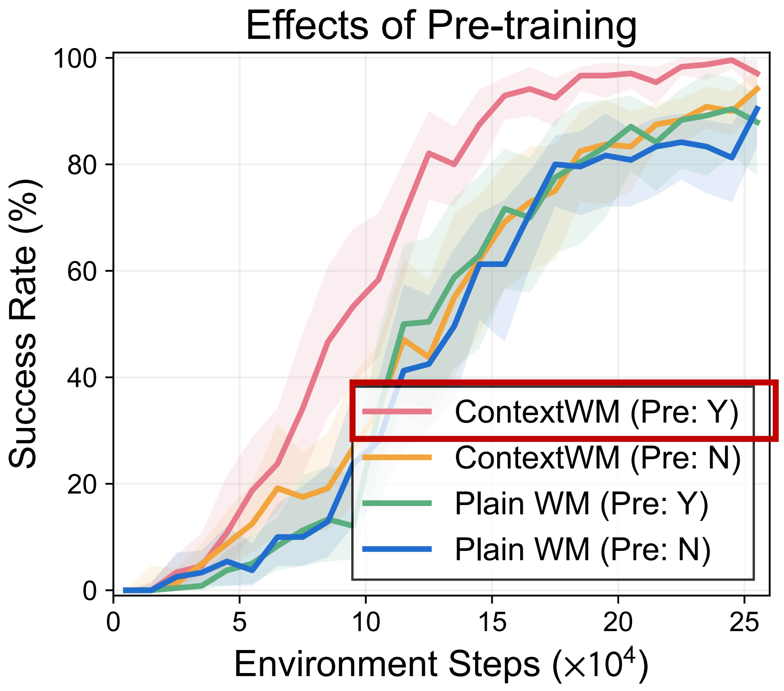


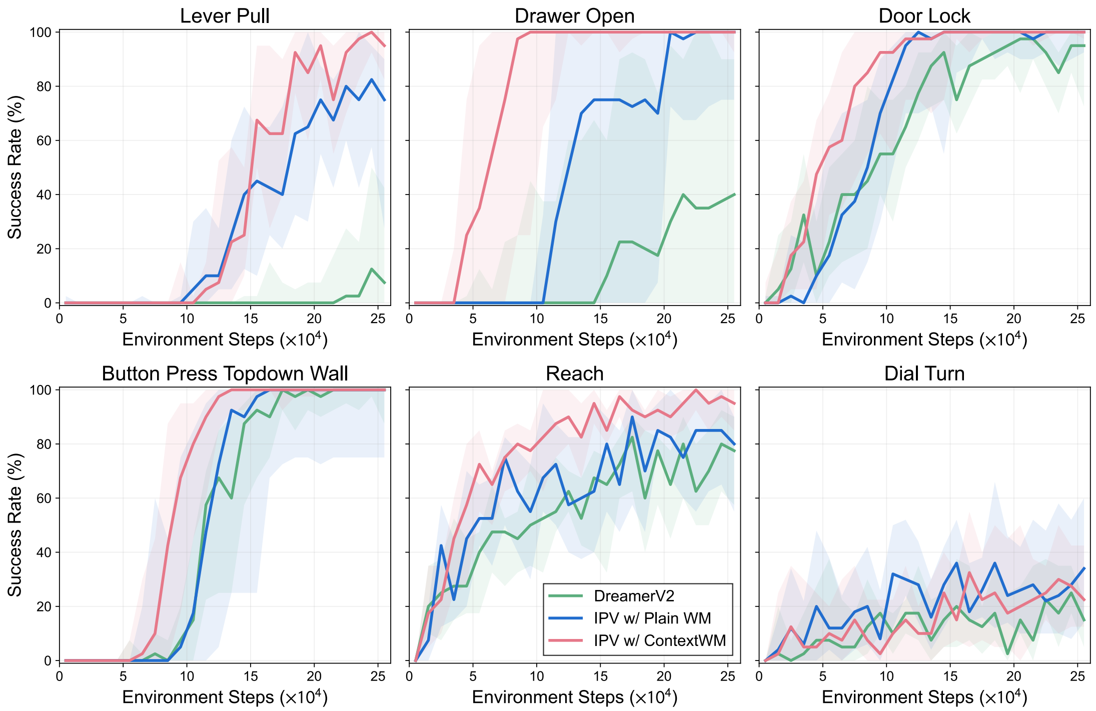
DMC Remastered
Grigsby et al. 2020

CARLA
Dosovitskiy et al. CoRL 2017

Main Results: Meta-world

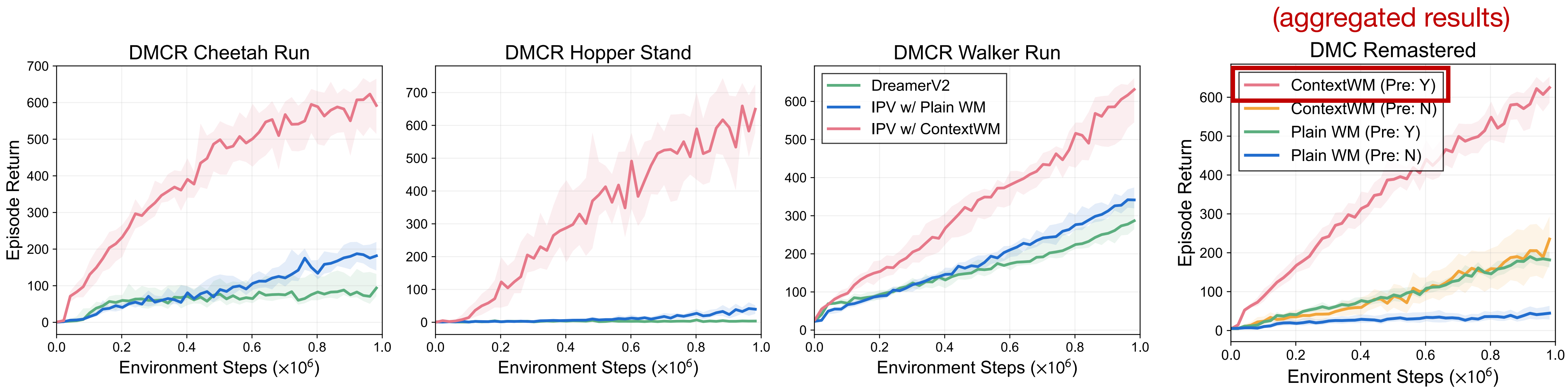
(aggregated results)



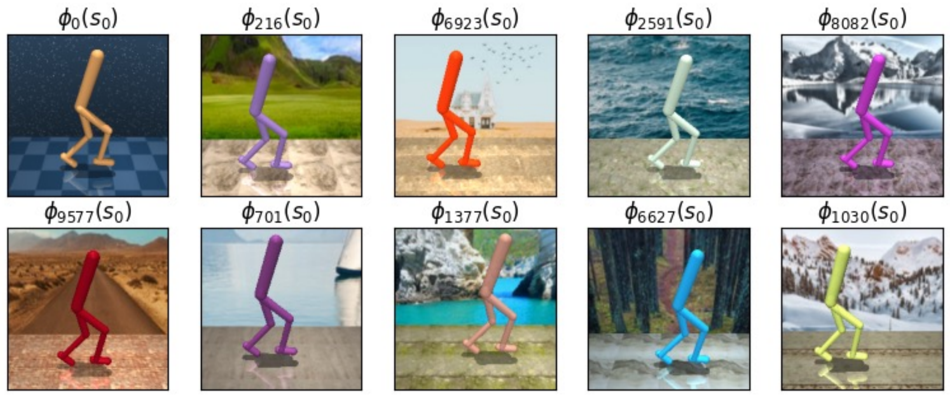


On six Meta-world tasks, ContextWM achieves significant positive transfer (from SSv2) in terms of sample efficiency, while a plain WM fails.

Main Results: DMC Remastered

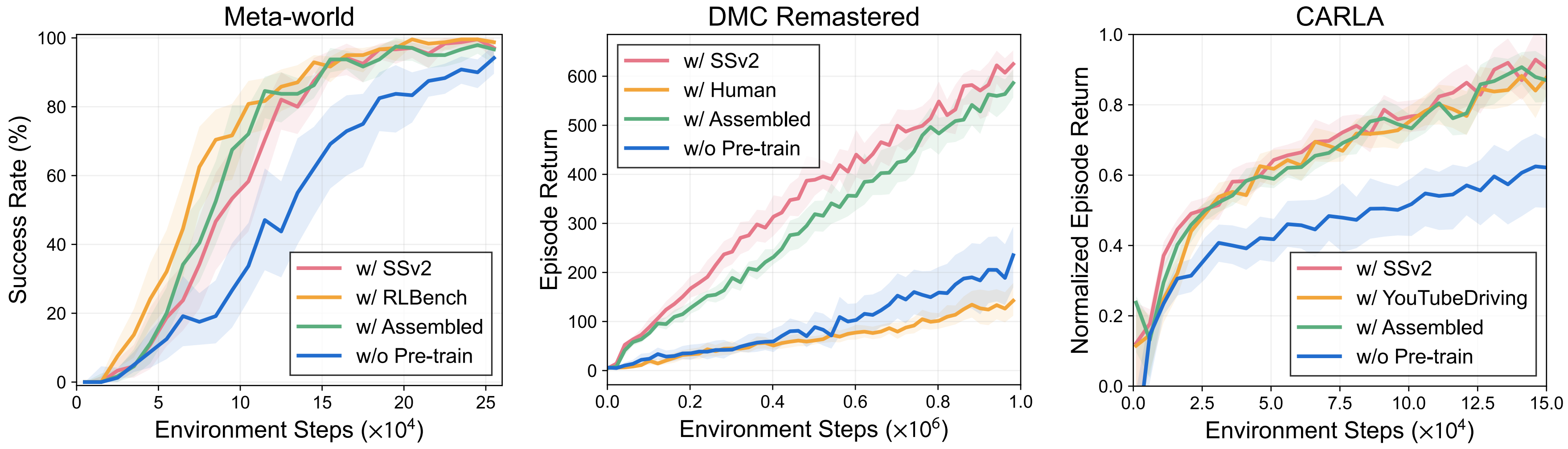


On visual generalization benchmark, pre-training from in-the-wild videos (SSv2) incurs significant performance boost, which is further unleashed by ContextWM.



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

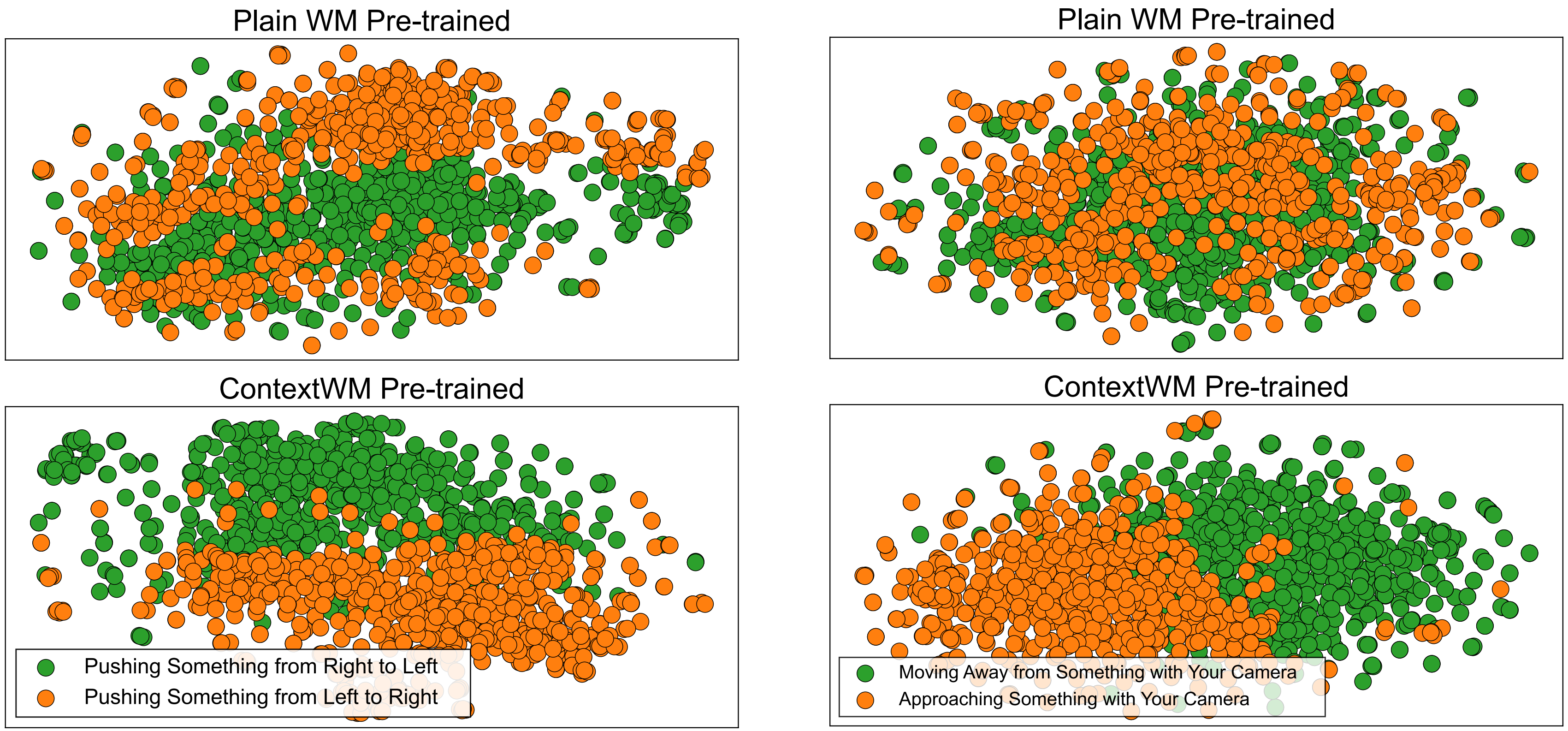
Effects of Pre-training Dataset Domain



Takeaways:

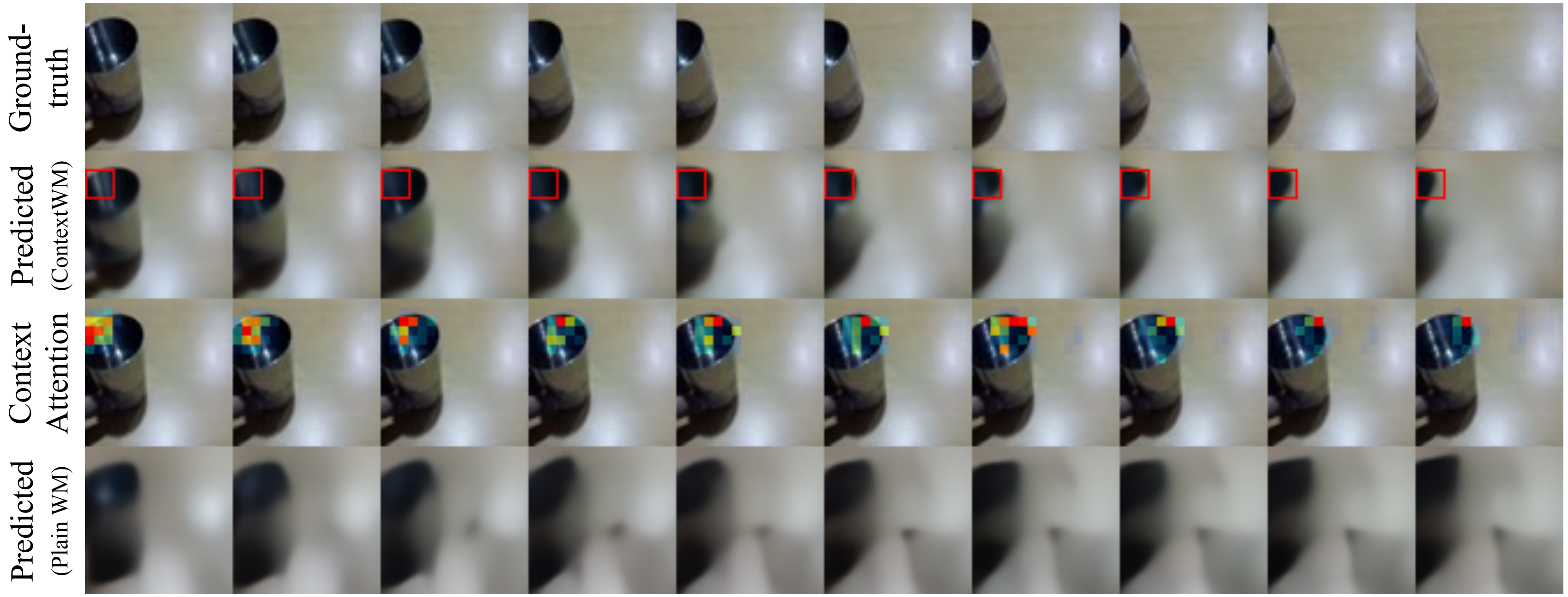
1. Human-object interaction data (SSv2) are generally beneficial.
2. A more similar domain (e.g. RLBench) is more useful, but more diverse datasets can serve as promising scalable alternatives.
3. Pre-training data lack of diversity (Human3.6M) can even be harmful.

Qualitative Evaluation: Video Representations



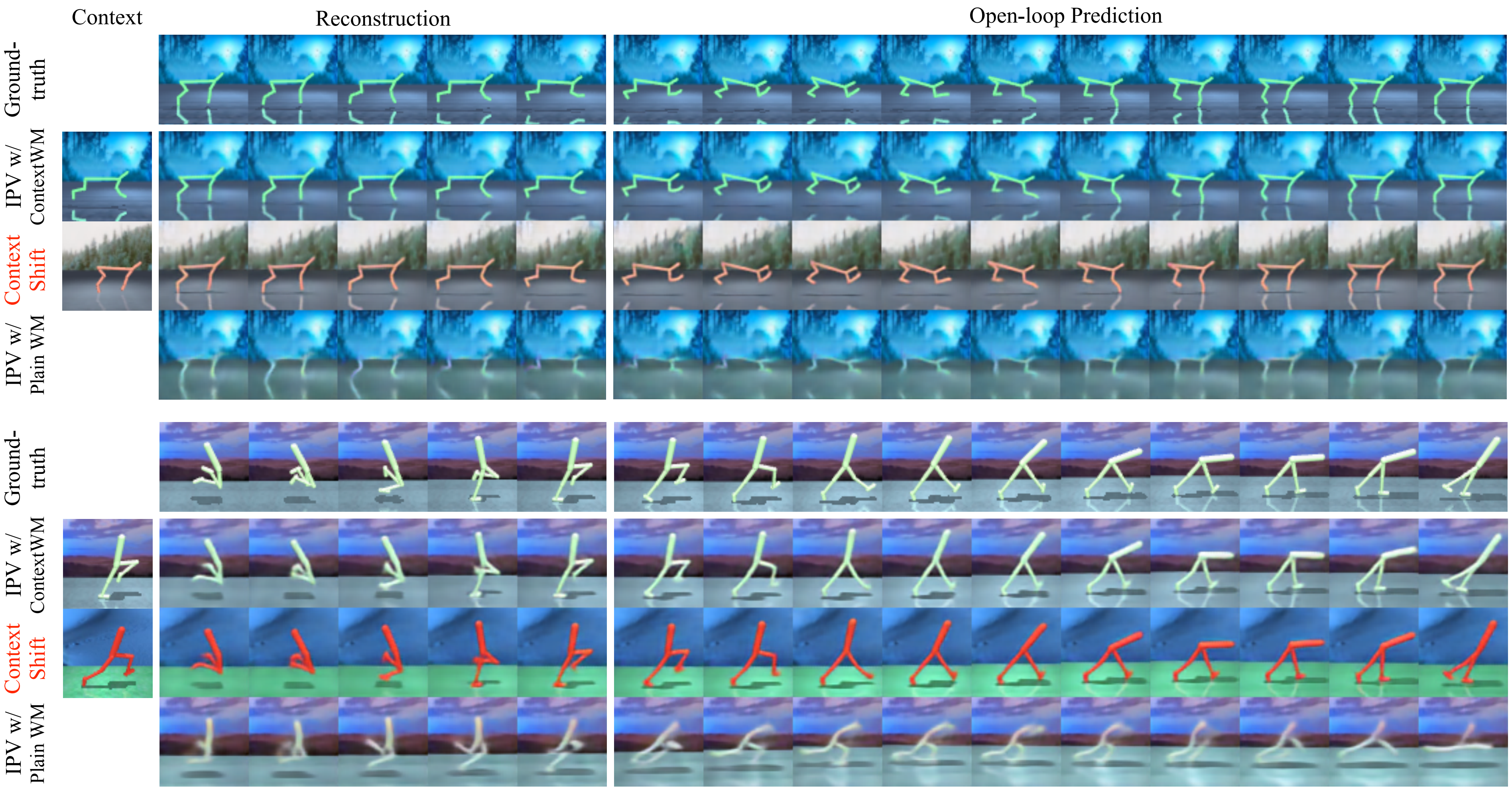
ContextWM learns representations well distributed according to different directions of motion, while not utilizing any labels of the videos in pre-training

Qualitative Evaluation: Video Prediction



1. Predictions from ContextWM well capture the shape and motion of the water cup.
2. Cross-attentions from different frames successfully attend to varying spatial positions of the context frame.

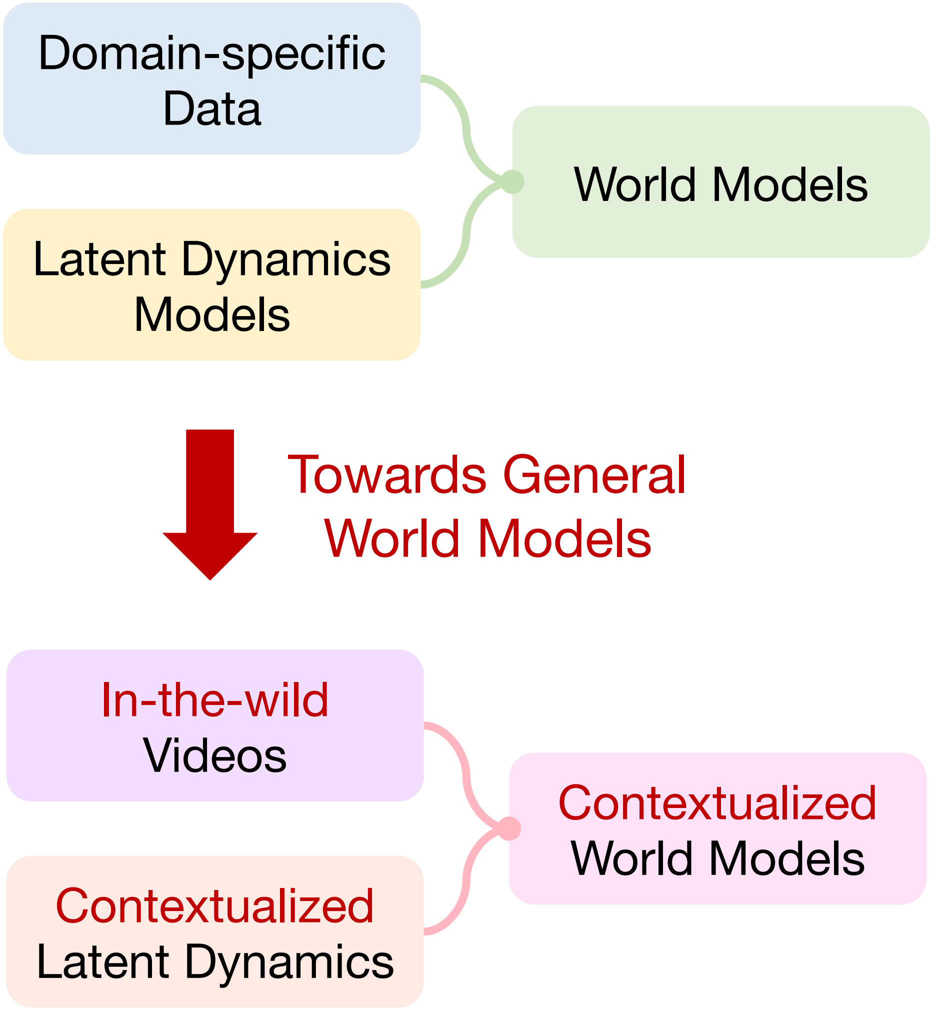
Qualitative Evaluation: Compositional Decoding



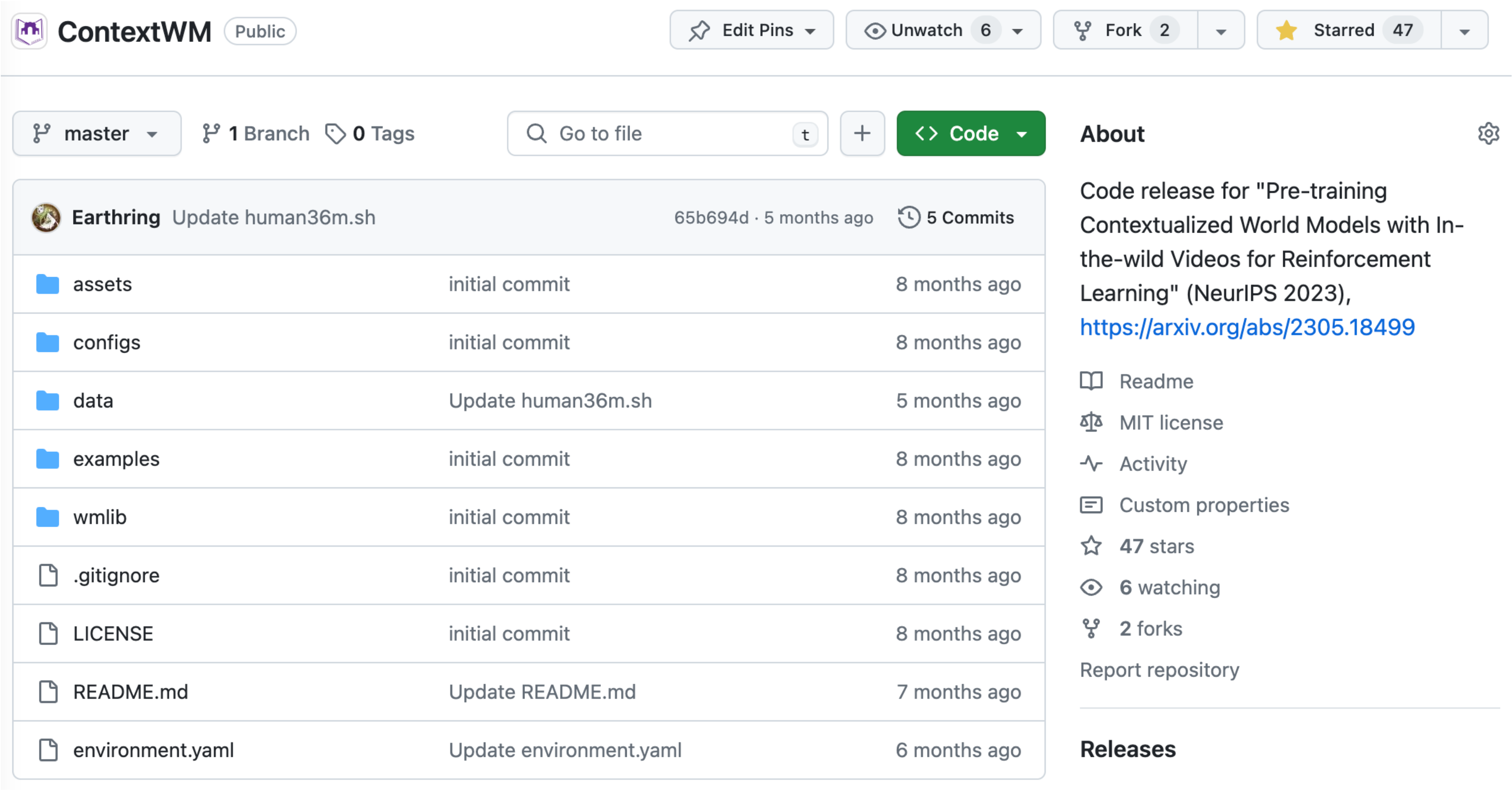
Excellent compositionality to combine new contexts with the original dynamics by **disentangled representations**

Summary

- Introduces Contextualized World Models (ContextWM)
- Applies it to the paradigm of In-the-wild Pre-training from Videos (IPV)
- Followed by fine-tuning on downstream tasks to boost learning efficiency of MBRL



Open Source



https://github.com/thuml/ContextWM

Unified implementations of DreamerV2, APV, ContextWM in PyTorch

ICML 2024
Forty-first 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.

清華大學 Tsinghua University

HUAWEI

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..."

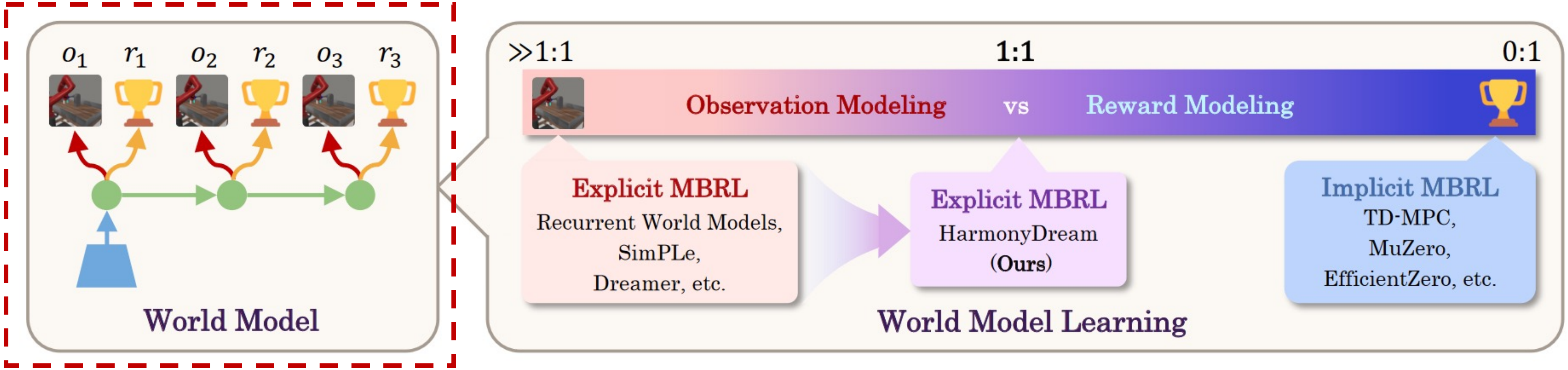
"It's much more desirable to generate abstract representations of those continuations that eliminate details in the scene that are irrelevant to any action we might want to take."

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:

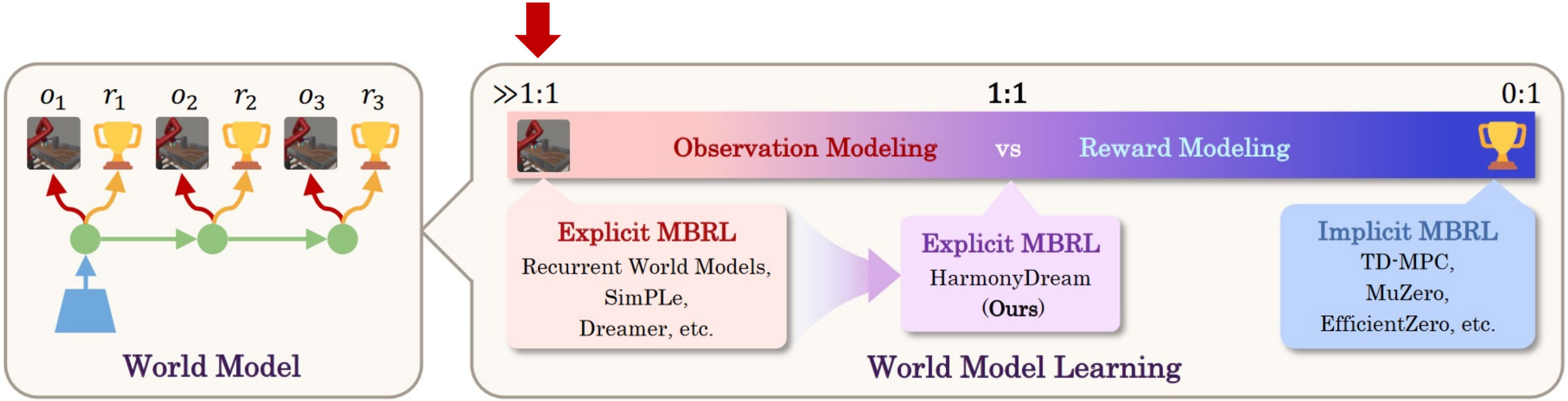
- **Observation Modeling:** how the environment transits and is observed

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

- **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

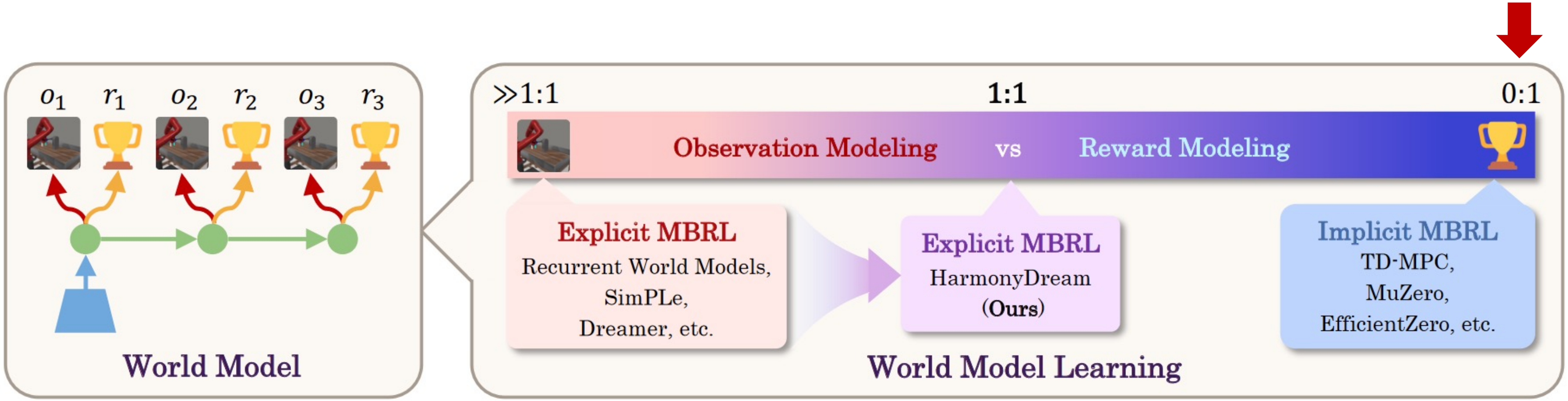


Unifying MBRL in concept (1/2): **Explicit MBRL**

- 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



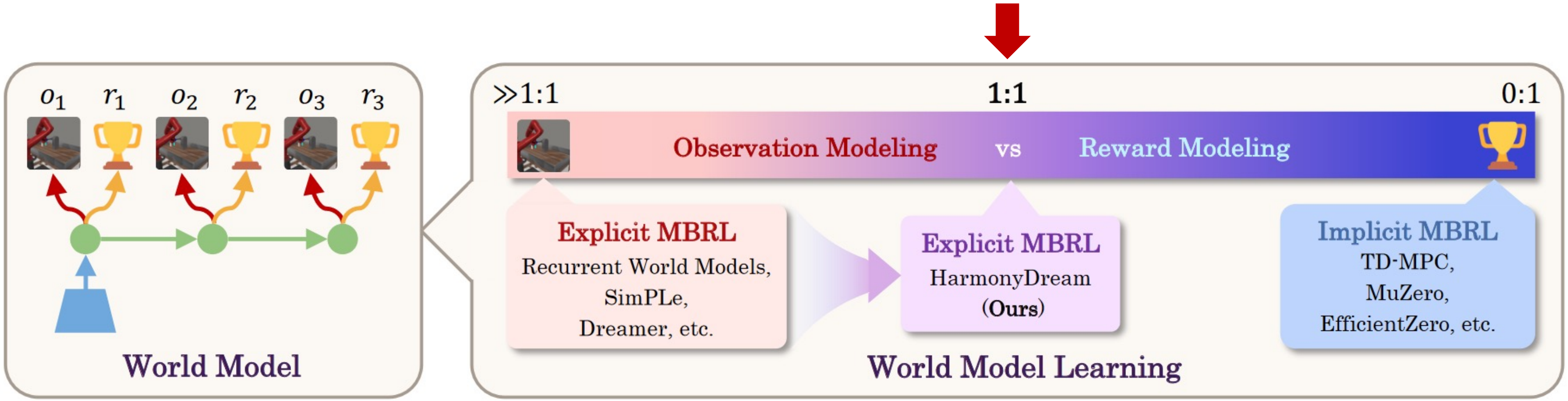
Unifying MBRL in concept (2/2): **Implicit MBRL**

- 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 Work



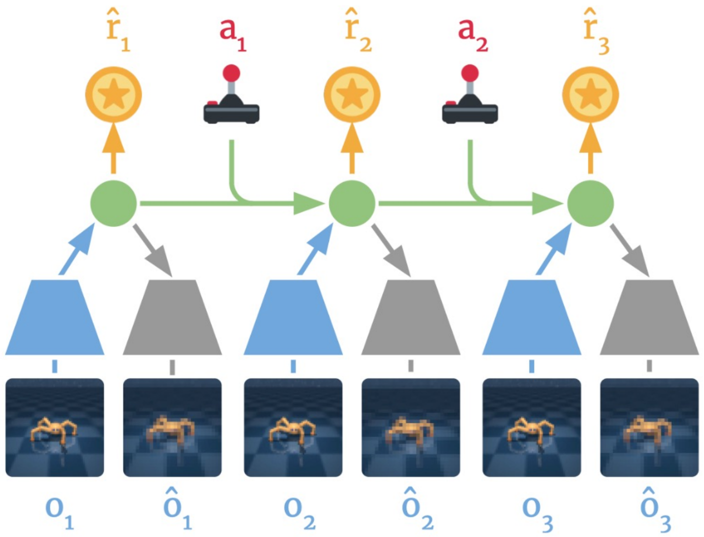
1. Systematically identify the multi-task essence of world models and analyze the deficiencies by task domination.
2. HarmonyDream, a world model learning approach to mitigate the domination of either task.
3. Extensive experiments on visual robotic tasks and video game benchmarks.

✓ Three findings

✓ One simple yet effective method

✓ Eight Domains

Recap: World Model Learning in Dreamer



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

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

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

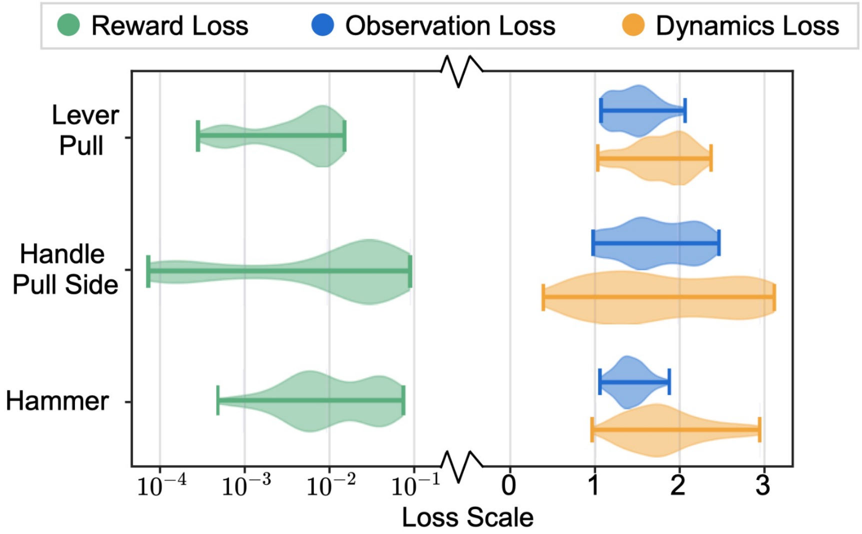
Observation loss: $\mathcal{L}_o(\theta) = -\log p_\theta(o_t \mid z_t) = -\sum_{h,w,c} \log p_\theta\left(o_t^{(h,w,c)} \mid z_t\right)$

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

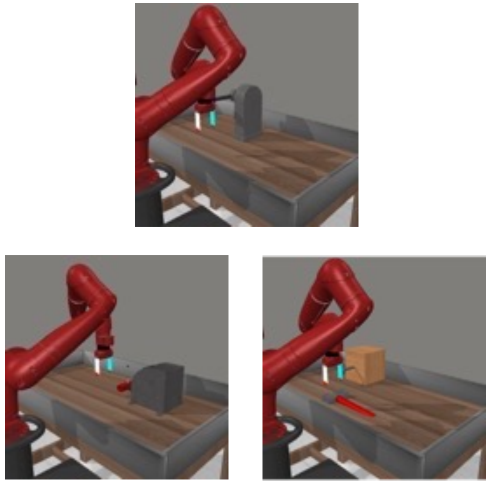
$$w_o = w_r = w_d = 1.0$$

Imbalanced nature of world model learning

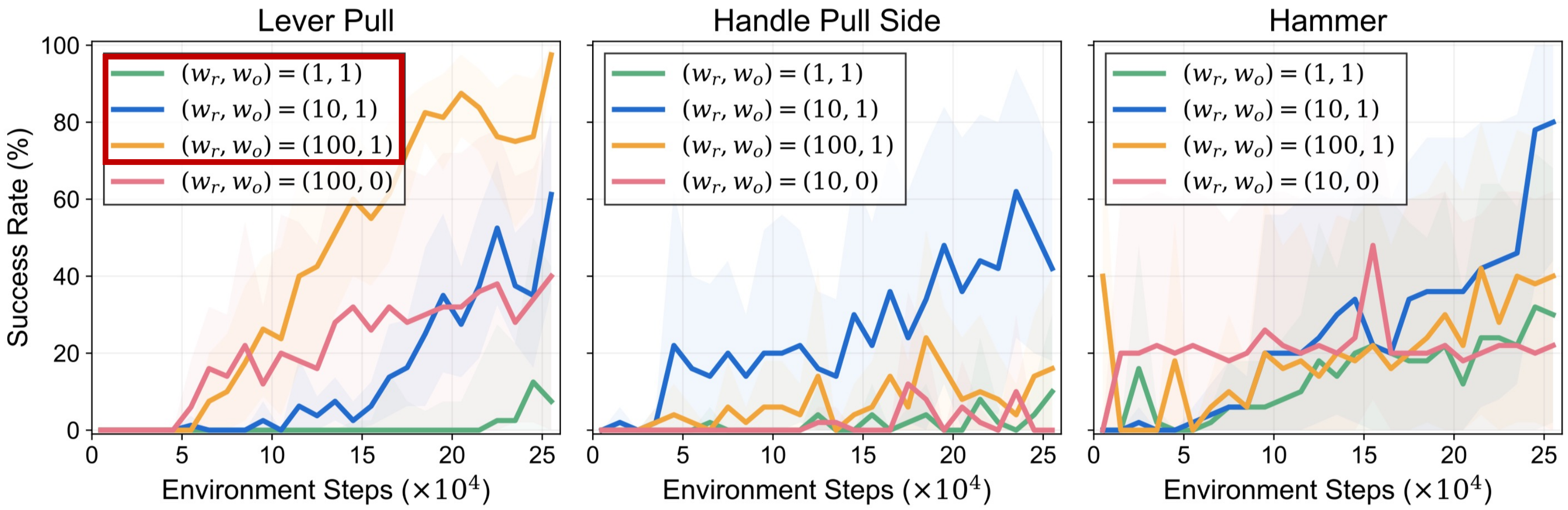
Potential benefits of multi-task learning yet properly exploited!

Task Weighting is Crucial

Dramatically boosted sample efficiency!



Testbed:
Three manipulation tasks from Meta-world

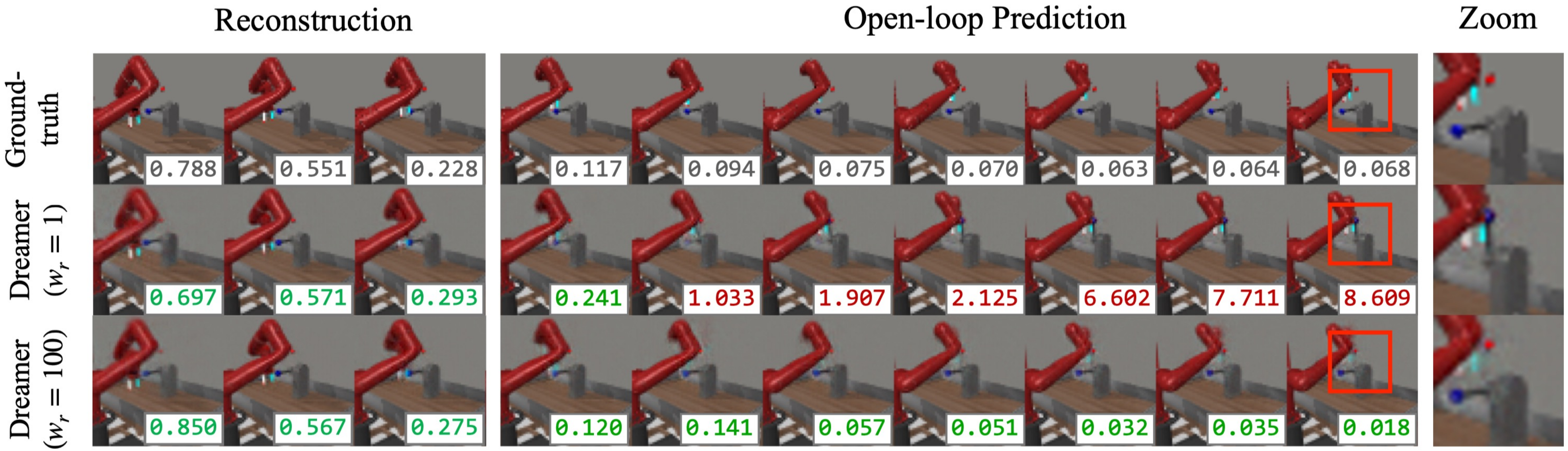


$$\mathcal{L}(\theta) = w_o\mathcal{L}_o(\theta) + w_r\mathcal{L}_r(\theta) + w_d\mathcal{L}_d(\theta)$$

$(\uparrow)$

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

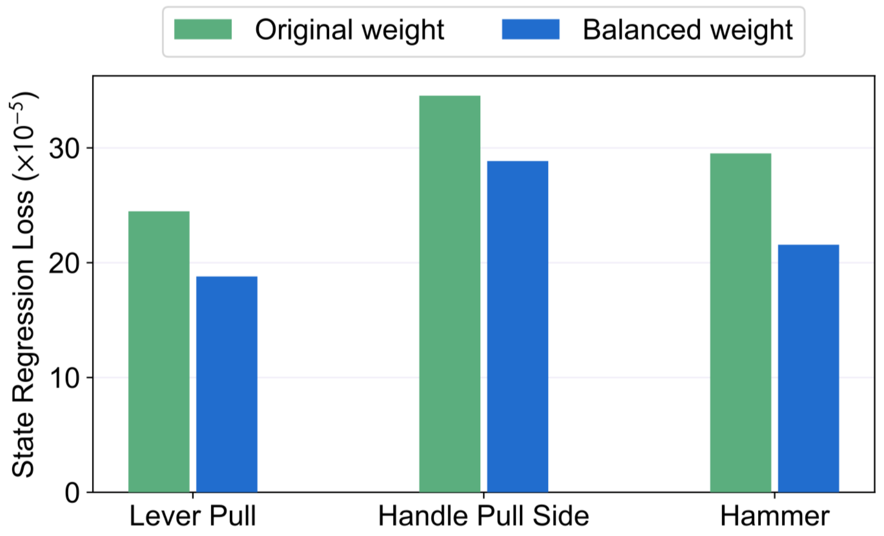
Hallucinations!

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



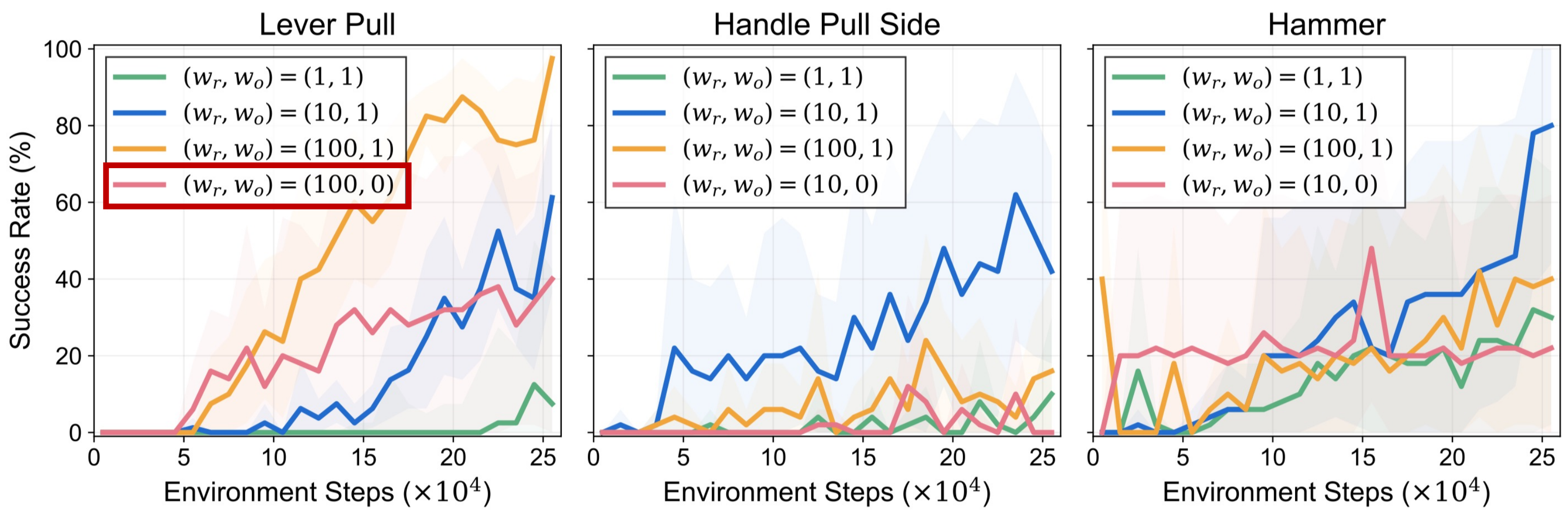
Hallucinations!

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

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

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

✗ Fluctuate throughout training

✗ Sensitive to outlier values

A Variational Approach and Its Rectification

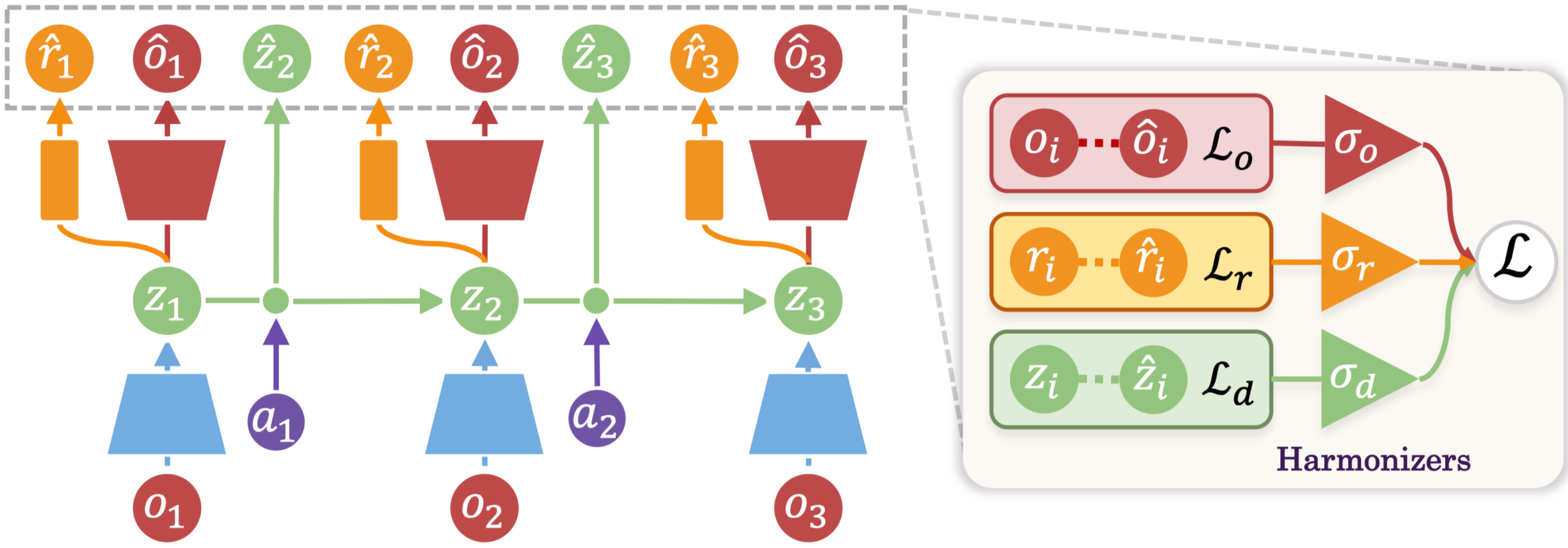
$$\mathcal{L}\left(\theta, \sigma_o, \sigma_r, \sigma_d\right) = \sum_{i \in \{o,r,d\}} \mathcal{H}\left(\mathcal{L}_i(\theta), \sigma_i\right)$$

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

$$\boxed{\begin{aligned} \sigma^* &= \mathbb{E}[\mathcal{L}] \\ \mathbb{E}\left[\mathcal{L}/\sigma^*\right] &= 1 \end{aligned}}$$

A "global" reciprocal of the loss scale

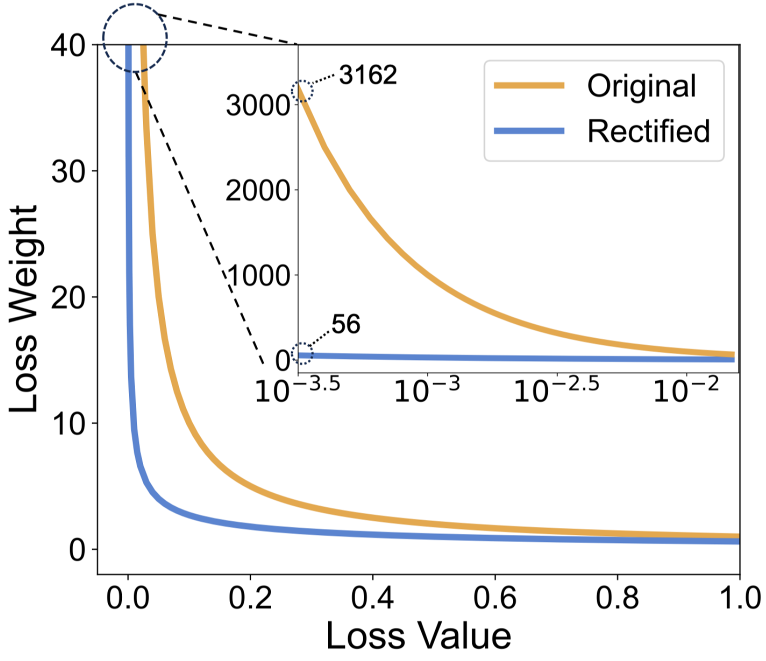
Dynamically but smoothly



A Variational Approach and Its Rectification

Extremely large coefficient hurts training stability

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



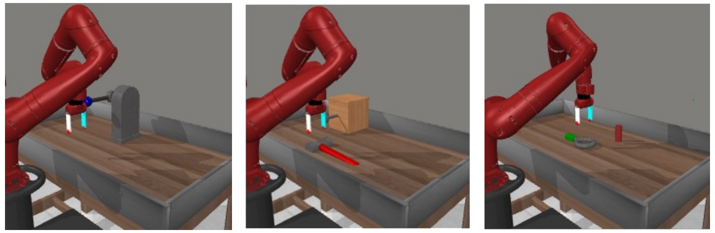
$$\mathcal{L}\left(\theta, \sigma_o, \sigma_r, \sigma_d\right) = \sum_{i \in \{o,r,d\}} \hat{\mathcal{H}}\left(\mathcal{L}_i(\theta), \sigma_i\right)$$

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

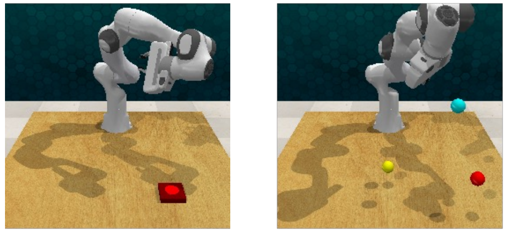
$$\boxed{\mathbb{E}\left[\mathcal{L}/\sigma^*\right] = \frac{2}{1+\sqrt{1+4/\mathbb{E}[\mathcal{L}]}} < 1}$$

Prevent extremely large loss weights

Experiments: Extensive Benchmarks and Tasks

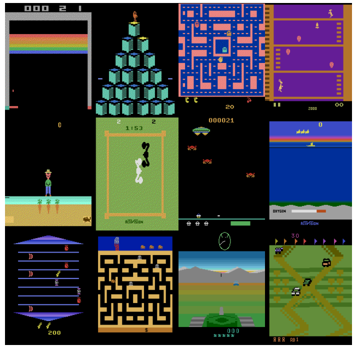


Meta-World
Yu et al. CoRL 2020



RLBench
James et al. IEEE RA-L 2020

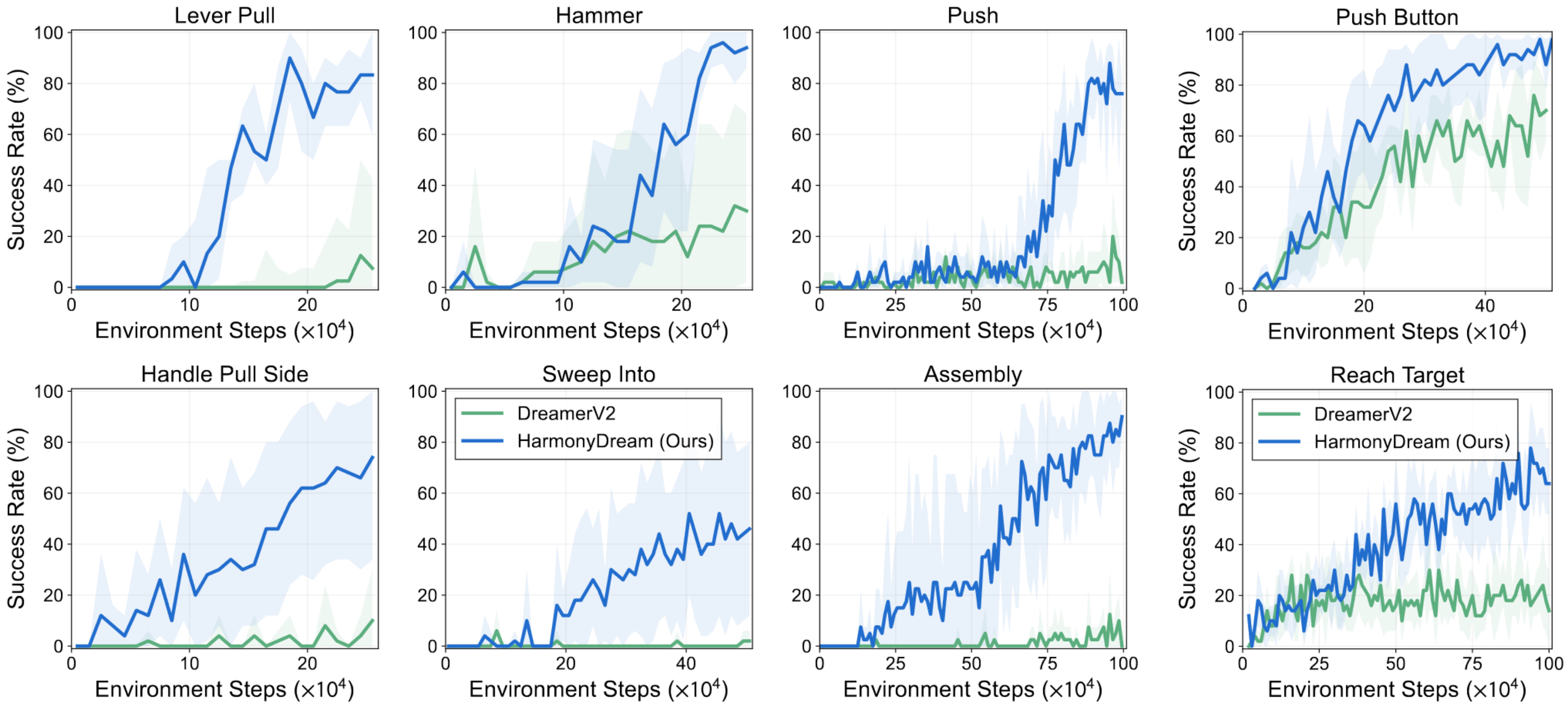
Distracted DMC Variants
Tassa et al. 2018; Zhang et al. 2018



Atari100K
Kaiser et al. ICLR 2020

Minecraft
Fan et al. NerulPS 2022

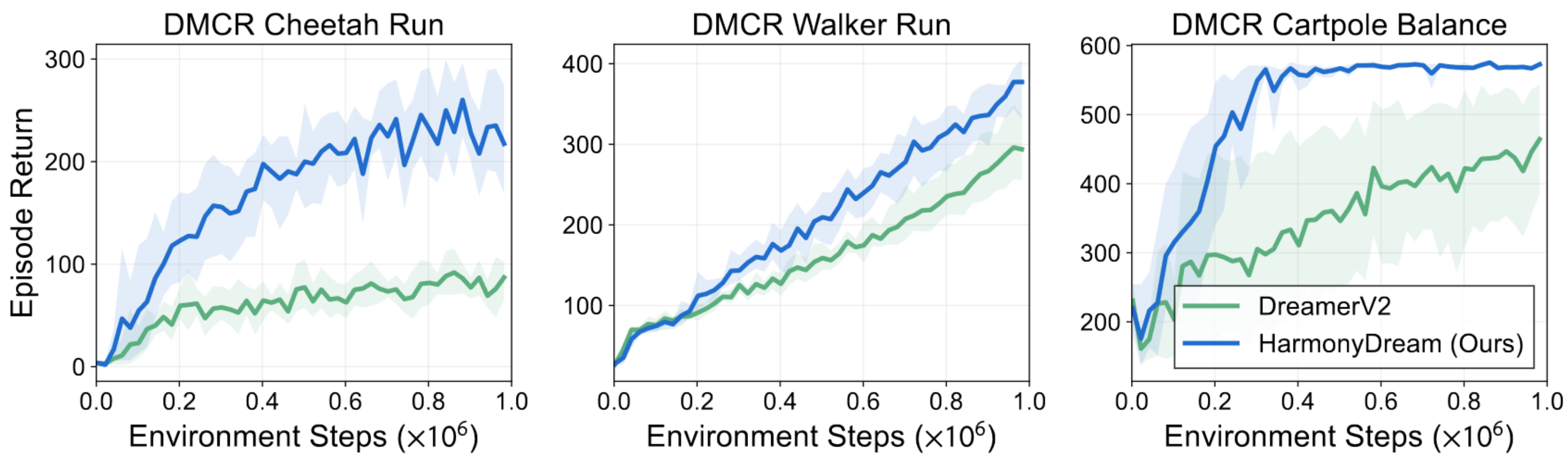
Main Results: Meta-world & RLBench



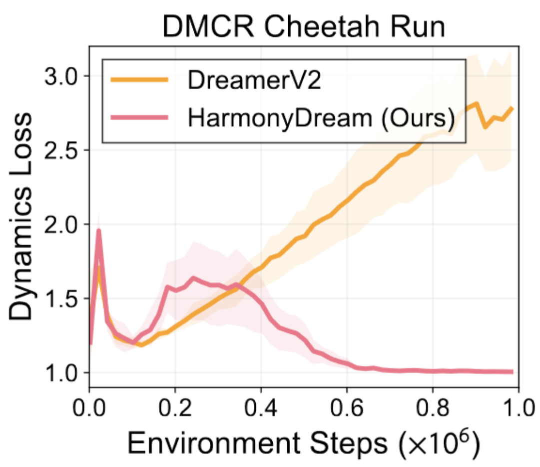
(a) Meta-world (b) RLBench

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

Main Results: DMC Remastered

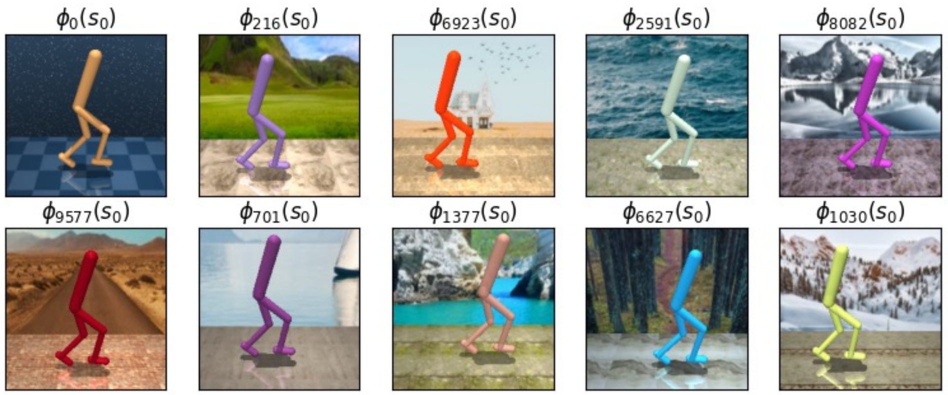


(a) Learning curves



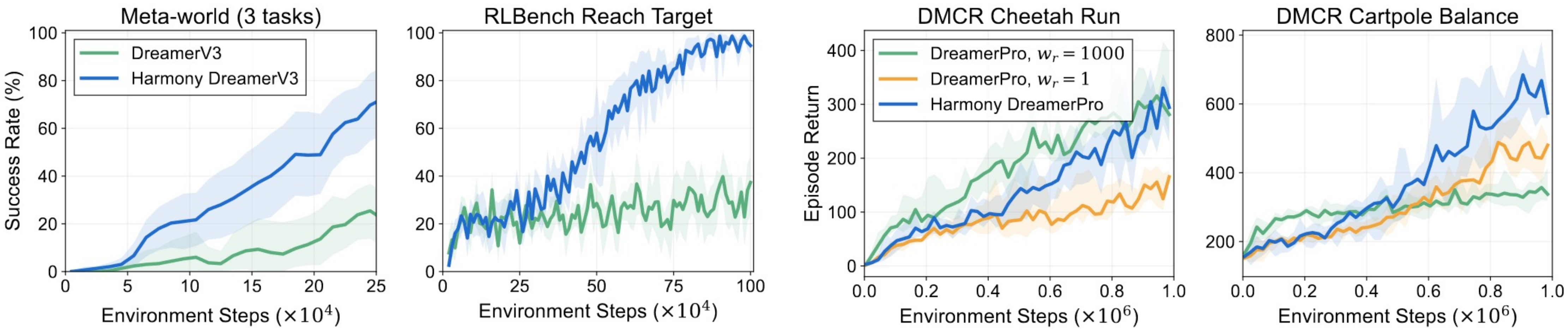
(b) Dynamics loss

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

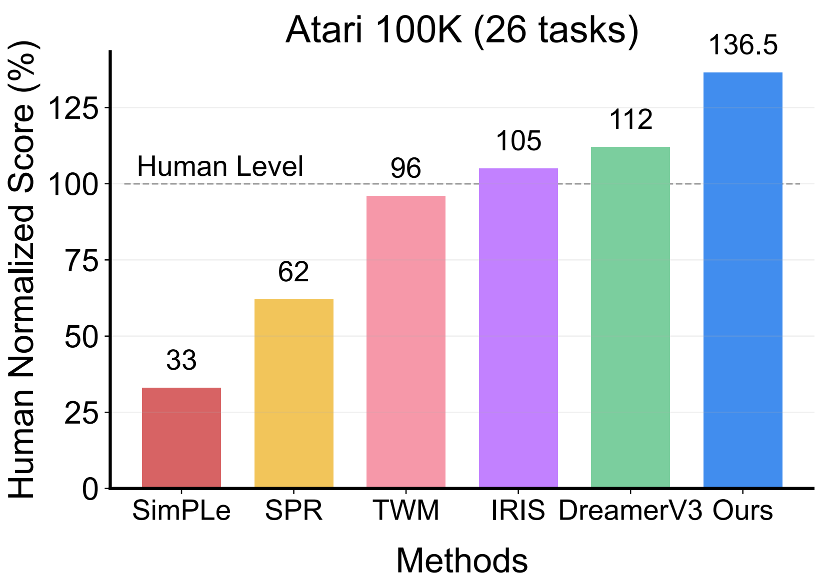
Generality to Base Model-based RL Methods

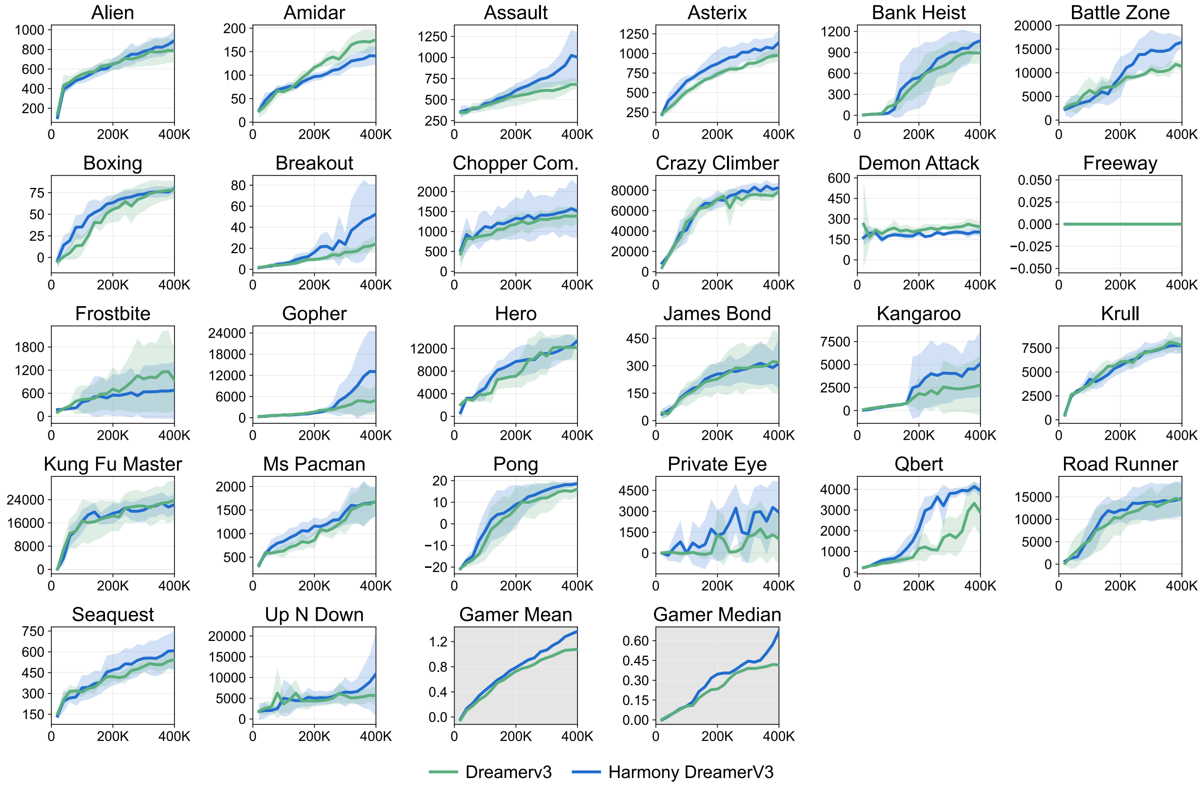


HarmonyDream exhibits excellent generality to DreamerV3, significantly boosting sample efficiency.

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

Harmony DreamerV3 on Atari100K

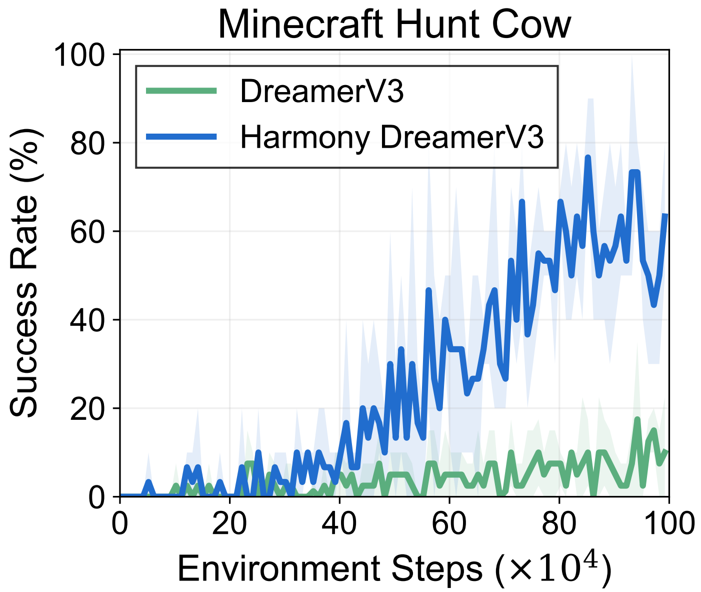




Harmony DreamerV3 significantly improves DreamerV3's performance, setting a new state of the art.

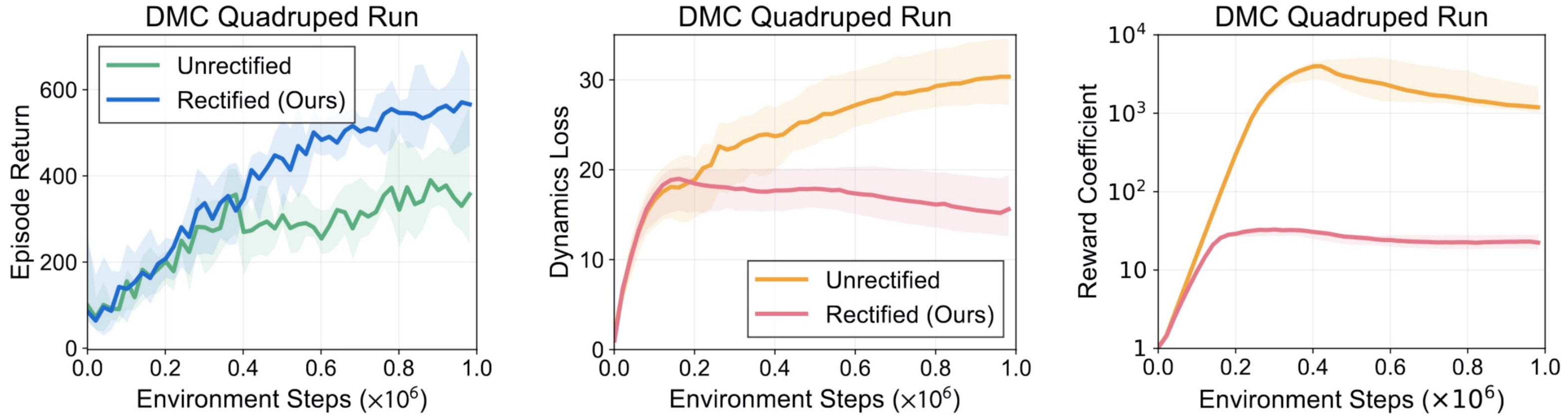
Either matching or surpassing DreamerV3 in 23/26 tested environments.

Harmony DreamerV3 on Minecraft

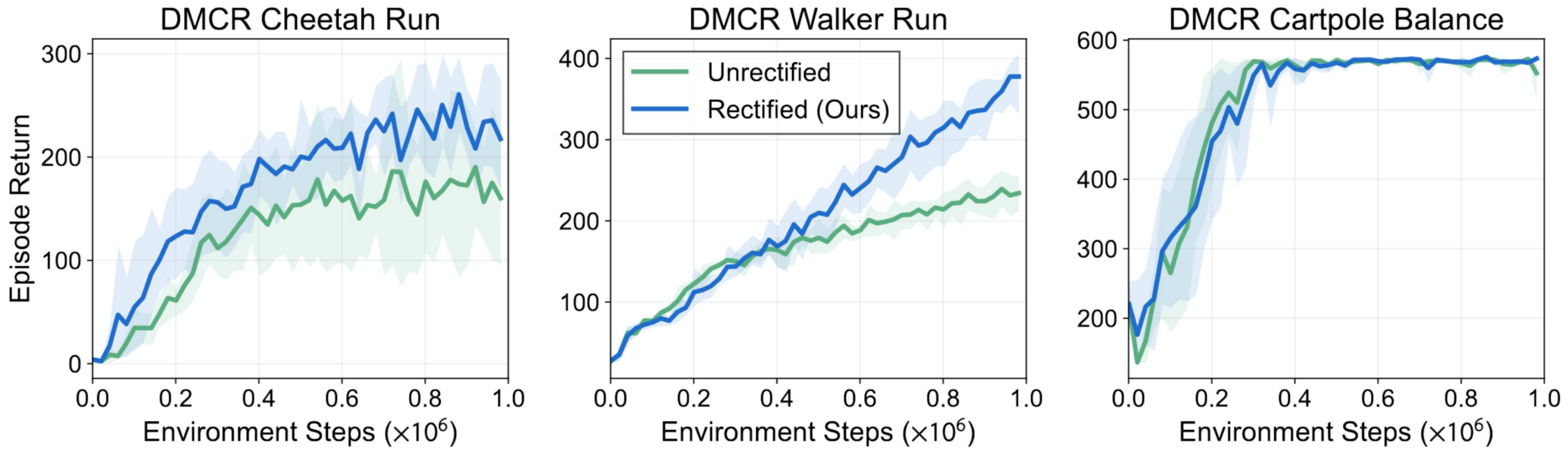


Harmony DreamerV3 successfully learns a basic skill *Hunt Cow* within 1M interactions, while DreamerV3 fails.

Ablation on Rectified Harmonious Loss

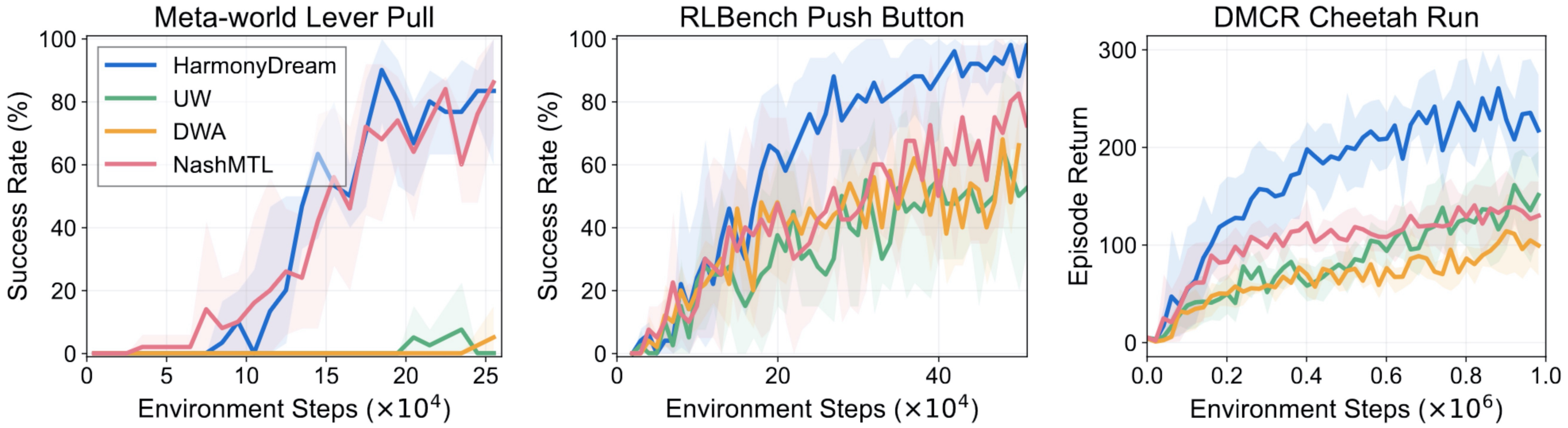


(a) Learning curves. (b) Dynamics loss. (c) Reward coefficient.



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



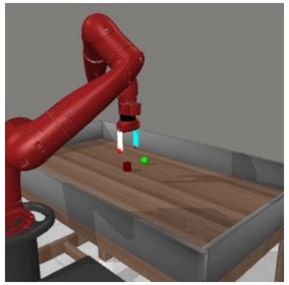
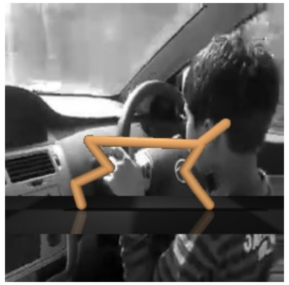
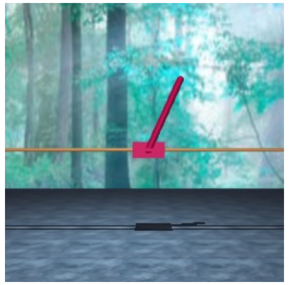
Takeaways:

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.

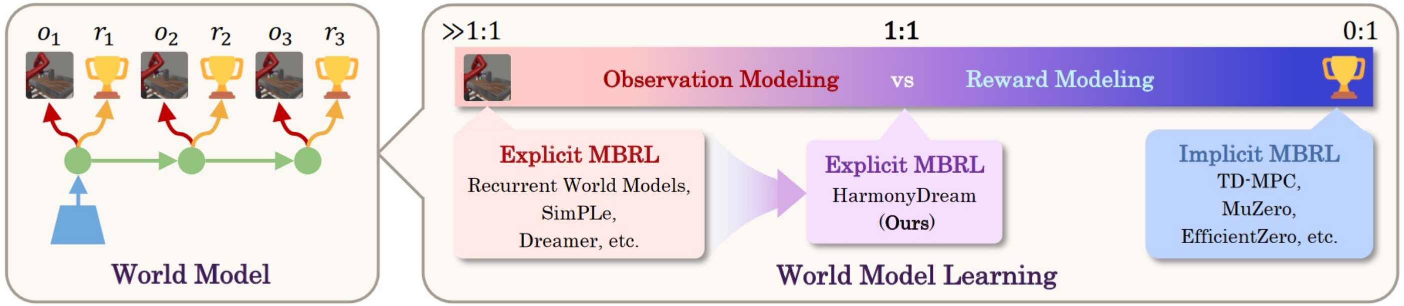





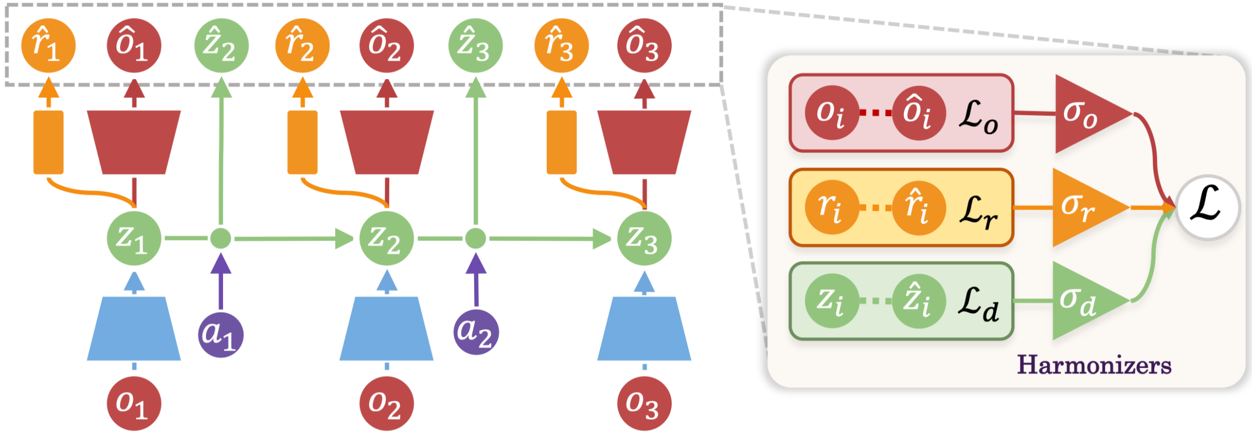

Summary

A multi-task view of world models

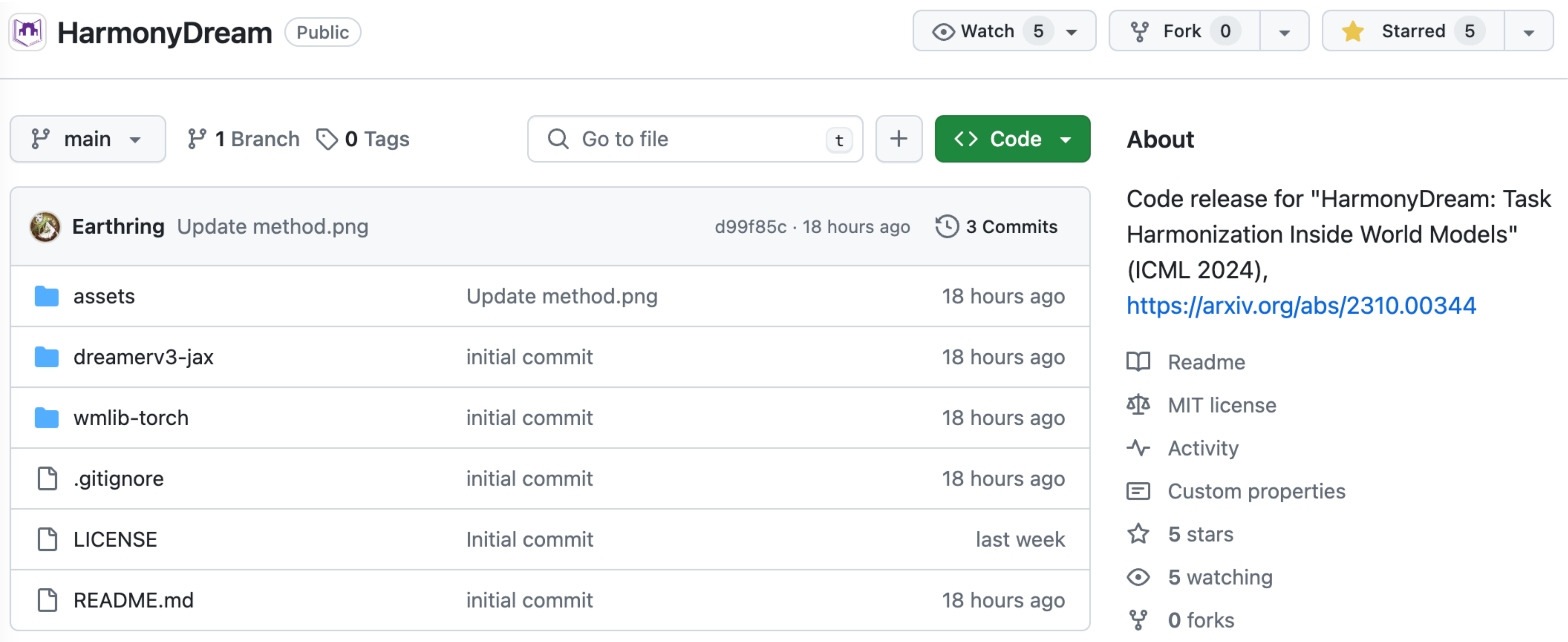
mitigate task domination

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

iVideoGPT: Interactive VideoGPTs are Scalable World Models

https://thuml.github.io/iVideoGPT

Jialong Wu[1]*, Shaofeng Yin[1,2]*, Ningya Feng[1], Xu He[3], Dong Li[3], Jianye Hao[3,4], Mingsheng Long[1]✉

[1]School of Software, BNRist, Tsinghua University, [2]Zhili College, Tsinghua University
[3]Huawei Noah's Ark Lab, [4]College of Intelligence and Computing, Tianjin University

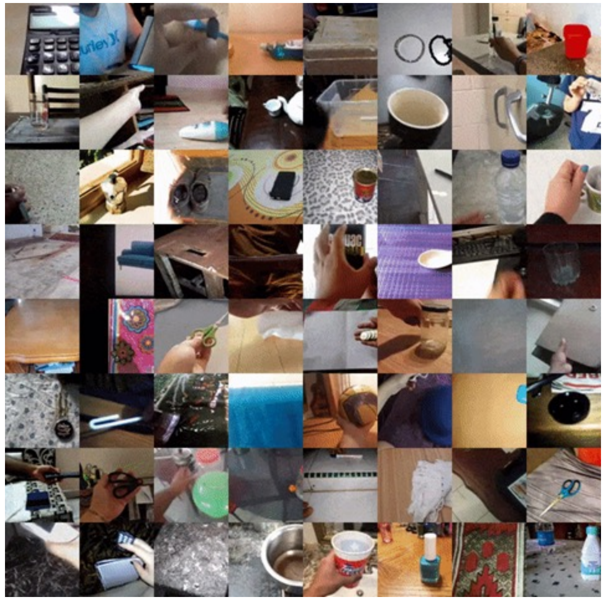
wujialong0229@gmail.com, ysf22@mails.tsinghua.edu.cn, mingsheng@tsinghua.edu.cn

Tsinghua University

HUAWEI

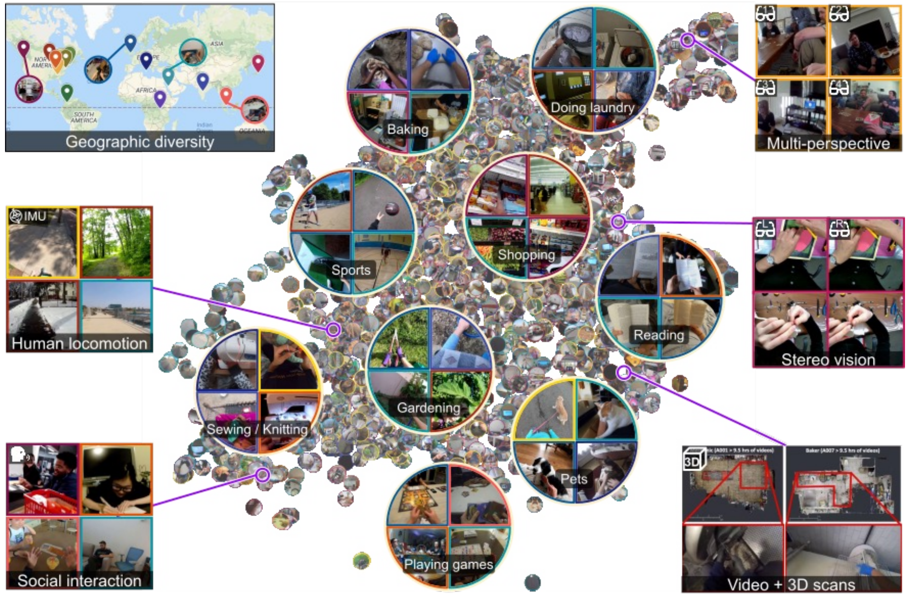
Recap: Towards a General World Model

General world knowledge for a variety of downstream tasks
from abundant in-the-wild videos on the Internet



Something-Something V2

Goyal et al. ICCV 2017

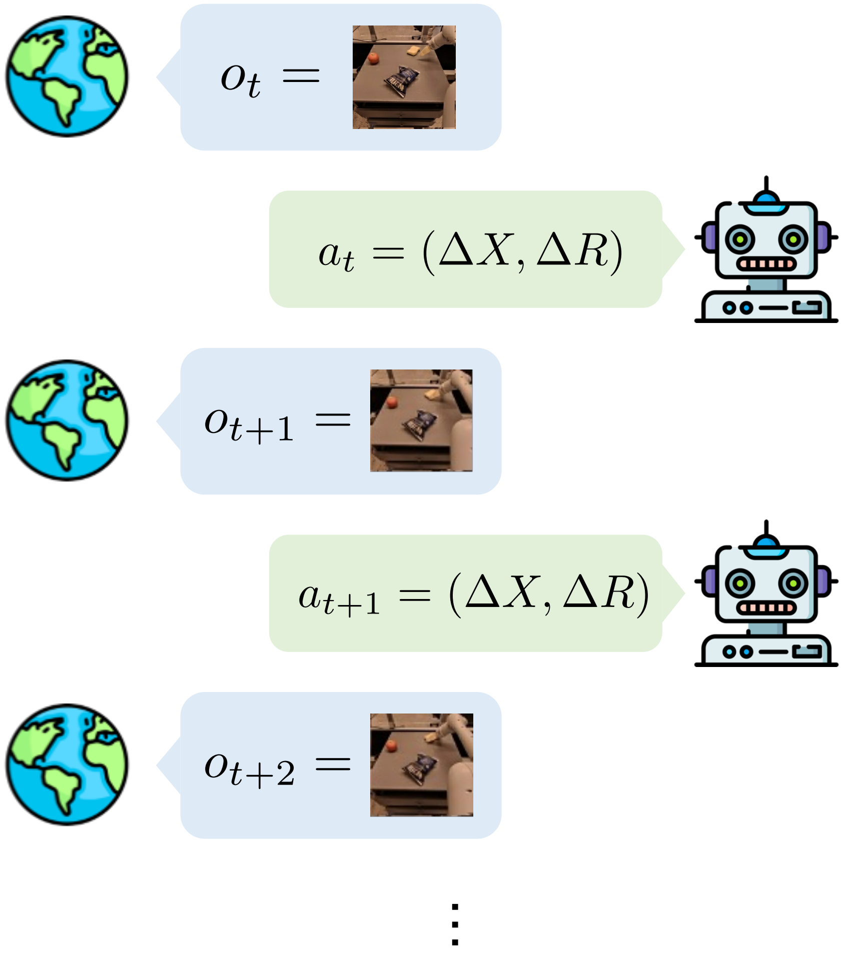


Ego4D

Grauman et al., Facebook AI. CVPR 2022

- ✓ Task-agnostic
- ✓ Widely available
- ✓ Broad Knowledge

World Model as Interactive Video Prediction



A process of making decisions and imagine outcomes:

$$p(o_{T_0+1:T}, a_{T_0:T-1} \mid o_{1:T_0})$$

$$= \underbrace{p(a_{T_0:T-1}|o_{1:t})}_{\text{Agent}}\underbrace{p(o_{T_0+1:T}|o_{1:T_0}, a_{T_0:T-1})}_{\text{World model}}$$

Non- (Low-) interactive

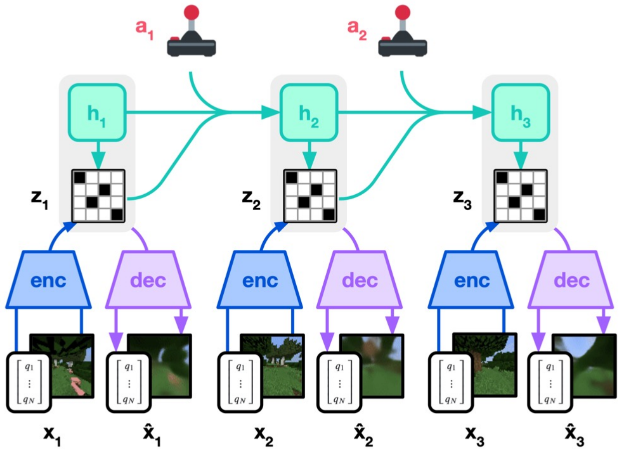
$$= \prod_{t=T_0}^{T-1} \underbrace{p(a_t|o_{1:t})}_{\text{Agent}}\underbrace{p(o_{t+1}|o_{1:t}, a_{T_0:t})}_{\text{World model}}$$

Interactive

A problem with fundamental connection to video prediction/generation models, referred to as interactive video prediction

Recurrent World Models Have Limited Scalability

DreamerV3: Naturally allows step-by-step transitions but with limited capability

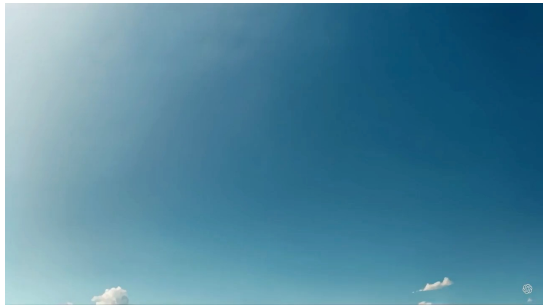


A case study on Minecraft

Ground truth

Prediction (DreamerV3-L)

Sora: Internet-scale video generative models can synthesize realistic long videos



High-fidelity Minecraft simulation:

Video Generative Models Have Limited Interactivity

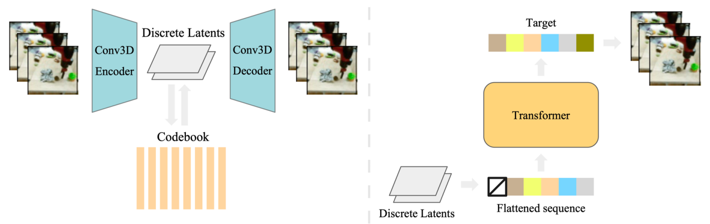
Typically design **non-causal temporal modules**

⬇

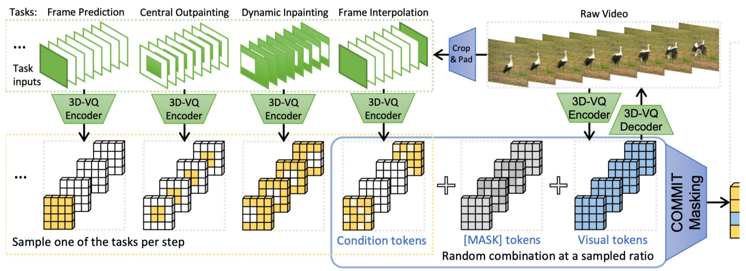
Provide only **trajectory-level interactivity**

- Allow text/action conditions only at the beginning of the video
- Lacking the ability for intervention during simulations
- Typically produce videos of a fixed length

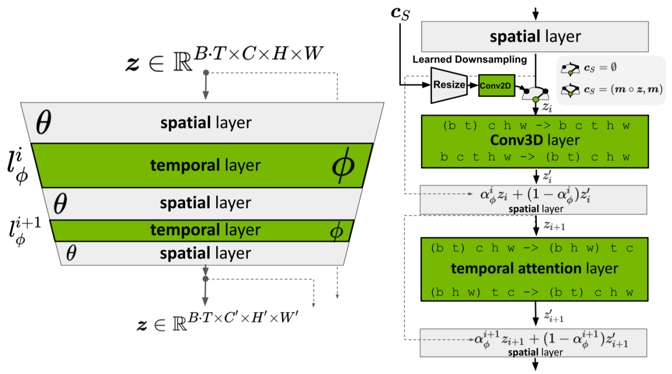
Our work: achieve **step-level interactivity**



Autoregressive model: VideoGPT



Masked model: MAGVIT

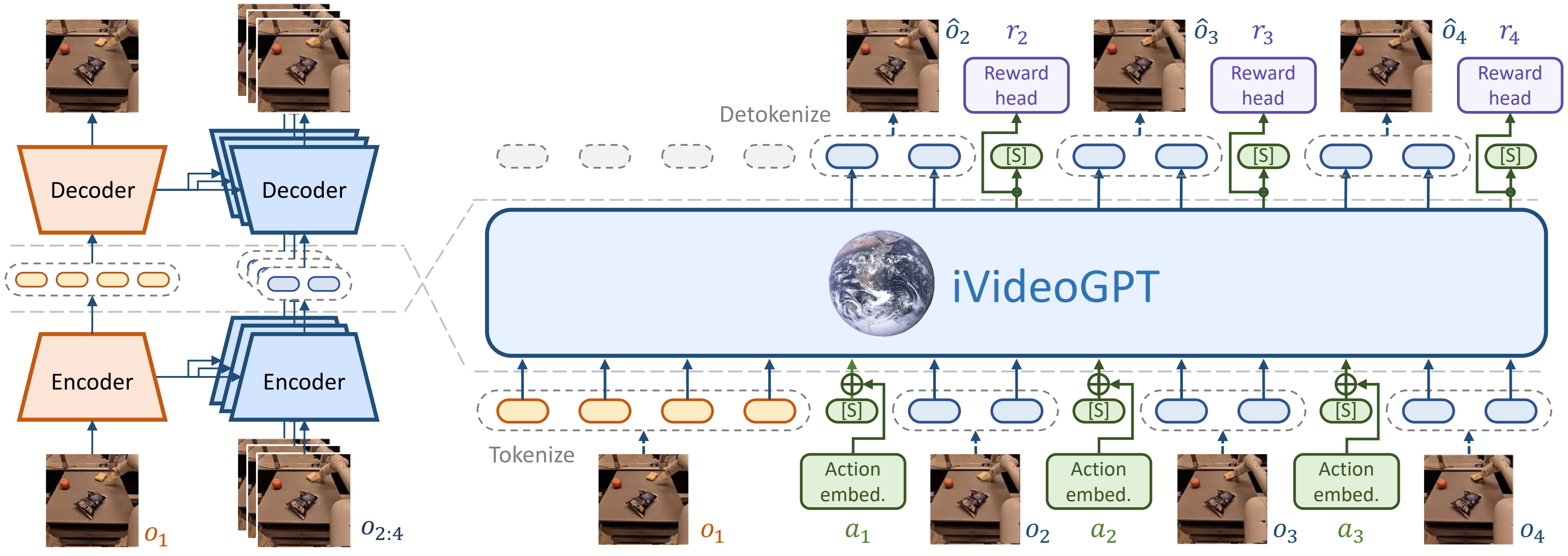


Diffusion model: Stable Video Diffusion

iVideoGPT: Interactive VideoGPT

Overview:

iVideoGPT integrates multimodal signals—visual observations (via **compressive tokenization**), actions, and rewards—into a sequence of tokens, and providing interactive experience via next-token prediction of an **autoregressive transformer**.



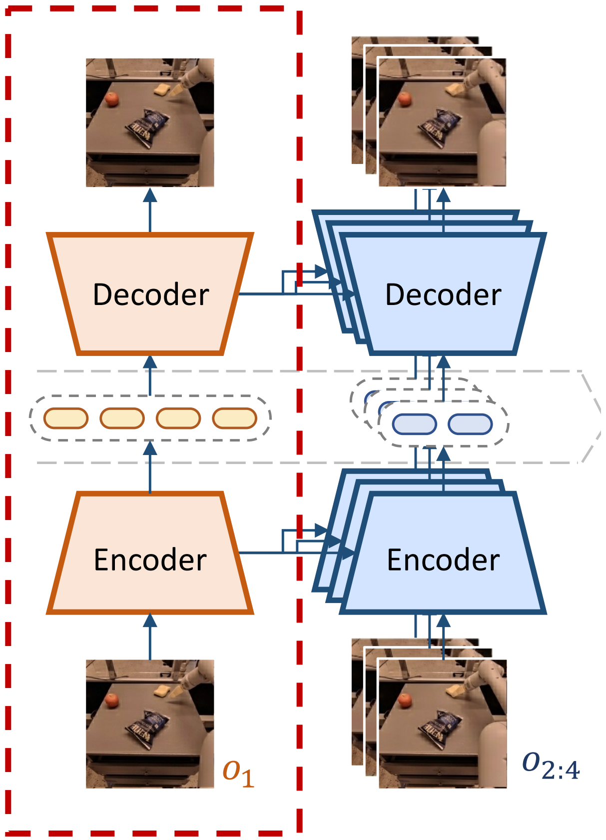
Compressive tokenization

Interactive prediction with Transformers

✓ **Scalability**

✓ **Interactivity**

Compressive Tokenization



($T_0 = 1$ for simplicity)

Transformers particularly shine when operating over sequences of discrete tokens → Commonly used visual tokenizer: **VQGAN**

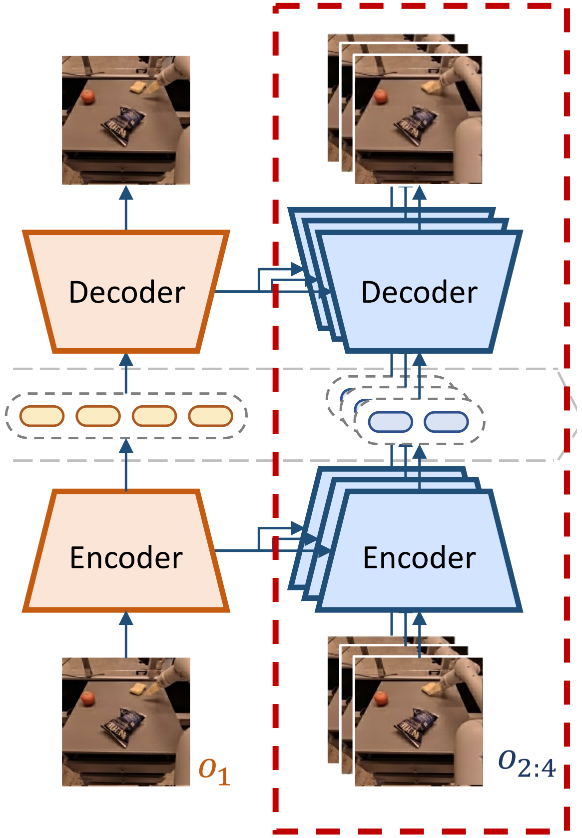
Context frames independently tokenized:

- Rich in contextual information
- Discretized into N tokens each frame:

$$z_t^{(1:N)} = E_c\left(o_t\right), \hat{o}_t = D_c\left(z_t\right) \text{ for } t = 1, \ldots, T_0$$

- To tokenize future frames as well? Low efficiency!

Compressive Tokenization



($T_0 = 1$ for simplicity)

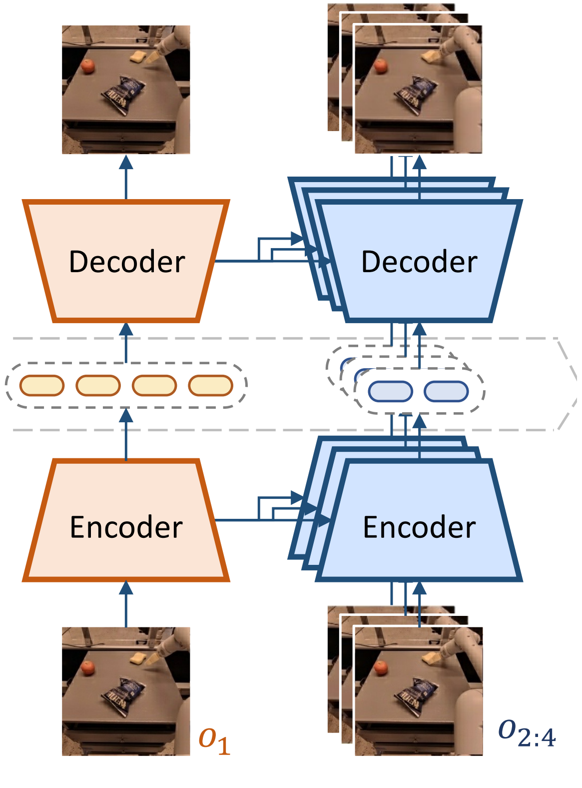
Future frames conditionally tokenized:

- Temporal redundancy between context and future frames
- Discretized into $n \ll N$ tokens each frame through **conditional VQGAN**:

$$z_t^{(1:n)} = E_p\left(o_t \mid \underbrace{o_{1:T_0}}_{\text{conditional encoder}}\right), \hat{o}_t = D_p\left(z_t \mid \underbrace{o_{1:T_0}}_{\text{conditional decoder}}\right) \qquad \text{for } t = T_0+1, \ldots, T$$

- Conditioning mechanism using cross-attention between multi-scale feature maps (the same as in **ContextWM**)

Compressive Tokenization



($T_0 = 1$ for simplicity)

Overall objective:

$$\mathcal{L}_{\text{tokenizer}} = \sum_{t=1}^{T_0} \underbrace{\mathcal{L}_{\text{VQGAN}}\left(o_t; E_c(\cdot), D_c(\cdot)\right)}_{\text{context frames}} + \sum_{t=T_0+1}^{T} \underbrace{\mathcal{L}_{\text{VQGAN}}\left(o_t; E_p\left(\cdot \mid o_{1:T_0}\right), D_p\left(\cdot \mid o_{1:T_0}\right)\right)}_{\text{future frames}}$$

Benefits:

✓ Shorter token sequence, faster rollouts for model-based planning and reinforcement learning

✓ Maintain temporal consistency of the context much easier and focus on modeling essential dynamics information

Interactive Prediction with Transformers

A sequence of tokens:

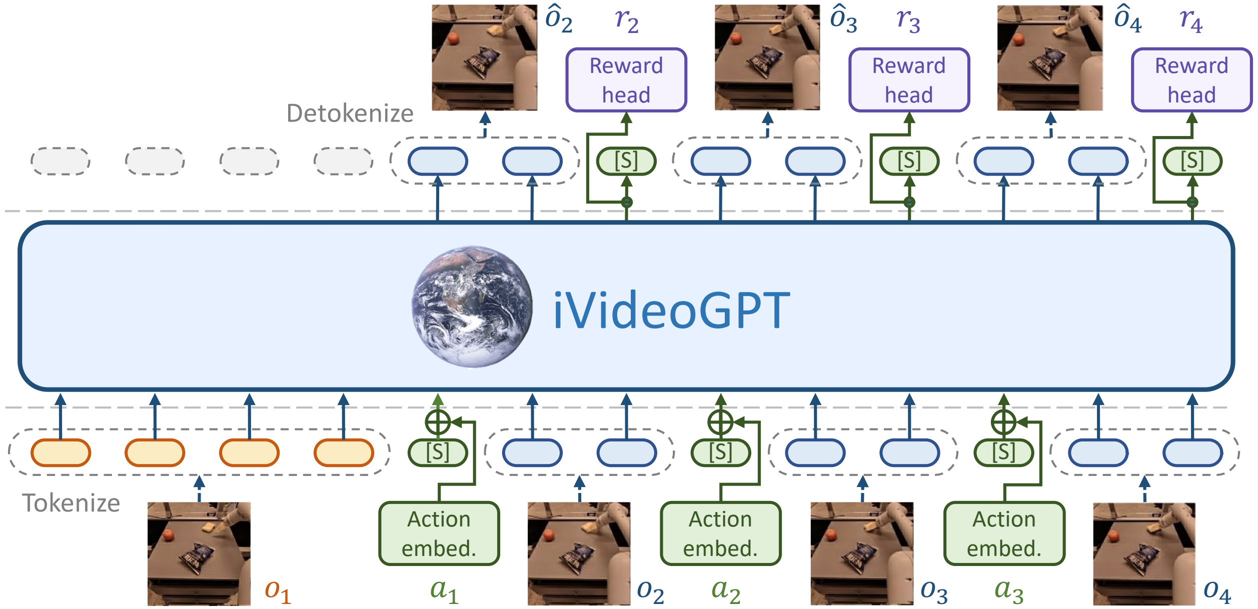
Delineate frame boundaries and facilitate optional action and reward integration

$$x = \Big(\underbrace{z_1^{(1)}, \ldots, z_1^{(N)}}_{\text{context frame}}, \texttt{[S]}, z_2^{(1)}, \ldots, z_2^{(N)}, \ldots, \underbrace{\texttt{[S]}}_{\text{slot token}}, \underbrace{z_{T_0+1}^{(1)}, \ldots, z_{T_0+1}^{(n)}}_{\text{future frame}}, \ldots\Big)$$

Total length $L = (N+1)T_0 + (n+1)(T - T_0) - 1$ grows linearly with frame numbers but at a much smaller rate ($n \ll N$)

GPT-2 size,
LLaMA architecture:

Embrace the latest innovations for LLM architecture



Pre-Training and Fine-Tuning

Action-free video prediction:

Not trained to generate context frames, focusing on dynamics information

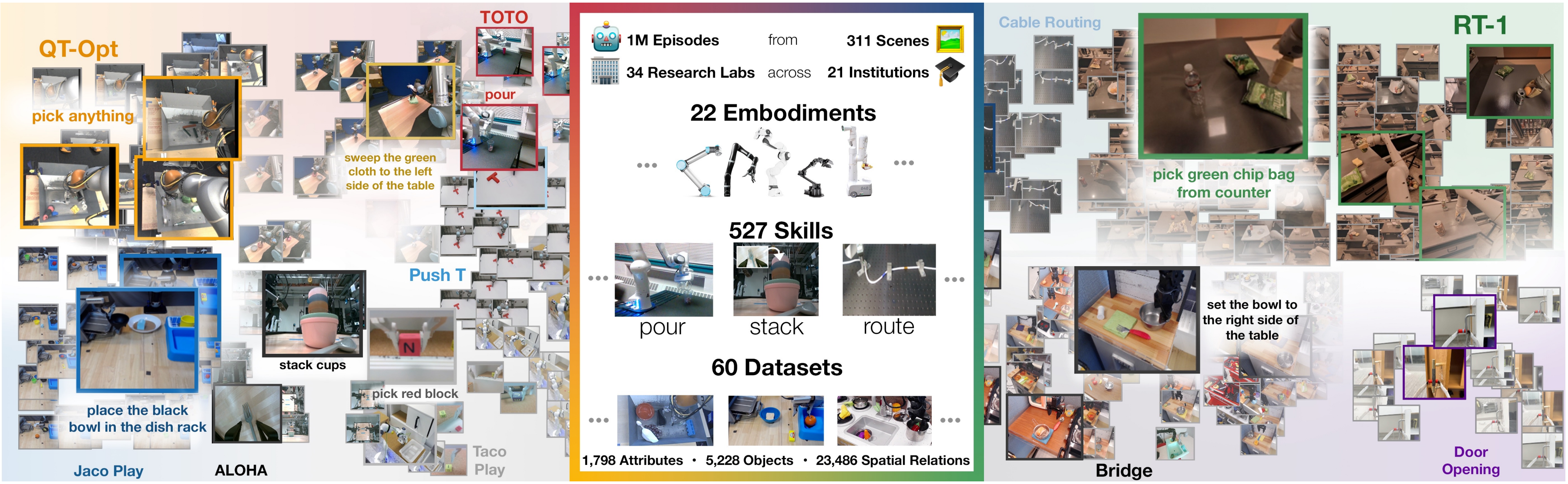
$$\mathcal{L}_{\text{pre-train}} = -\sum_{i=(N+1)T_0+1}^{L} \log p\left(x_i \mid x_{<i}\right)$$

First token index of predicted frames

Flexibly incorporate extra modalities:

- **Action conditioning**: linear projection and adding to the slot token embeddings
- **Reward prediction**: linear head to the last token's hidden state of each observation; mean-squared error (MSE) loss

Pre-Training Data



Open X-Embodiment

Padalkar et al. 2023

Something-Something V2

Goyal et al. ICCV 2017

Total 1.5 million trajectories:

- Select 35 datasets from OXE, in addition to SSv2, by excluding mobile robots, excessive repetition, and low image resolutions
- Filter out overlaps with downstream test data
- Sampling weights based on sizes and diversity
- Varied frame step sizes, based on control frequency

Video Prediction

Per-frame tokenization suffers from temporal inconsistency and flicker artifacts

BAIR [20]	FVD↓	PSNR↑	SSIM↑	LPIPS↓
action-free & 64×64 resolution				
VideoGPT [97]	103.3	-	-	-
MaskViT [26]	93.7	-	-	-
FitVid [3]	93.6	-	-	-
MCVD [89]	89.5	16.9	78.0	-
MAGVIT [100]	**62.0**	19.3	78.7	12.3
iVideoGPT (ours)	75.0±0.20	**20.4**±0.01	**82.3**±0.05	**9.5**±0.01
action-conditioned & 64×64 resolution				
MaskViT [26]	70.5	-	-	-
iVideoGPT (ours)	**60.8**±0.08	**24.5**±0.01	**90.2**±0.03	**5.0**±0.01

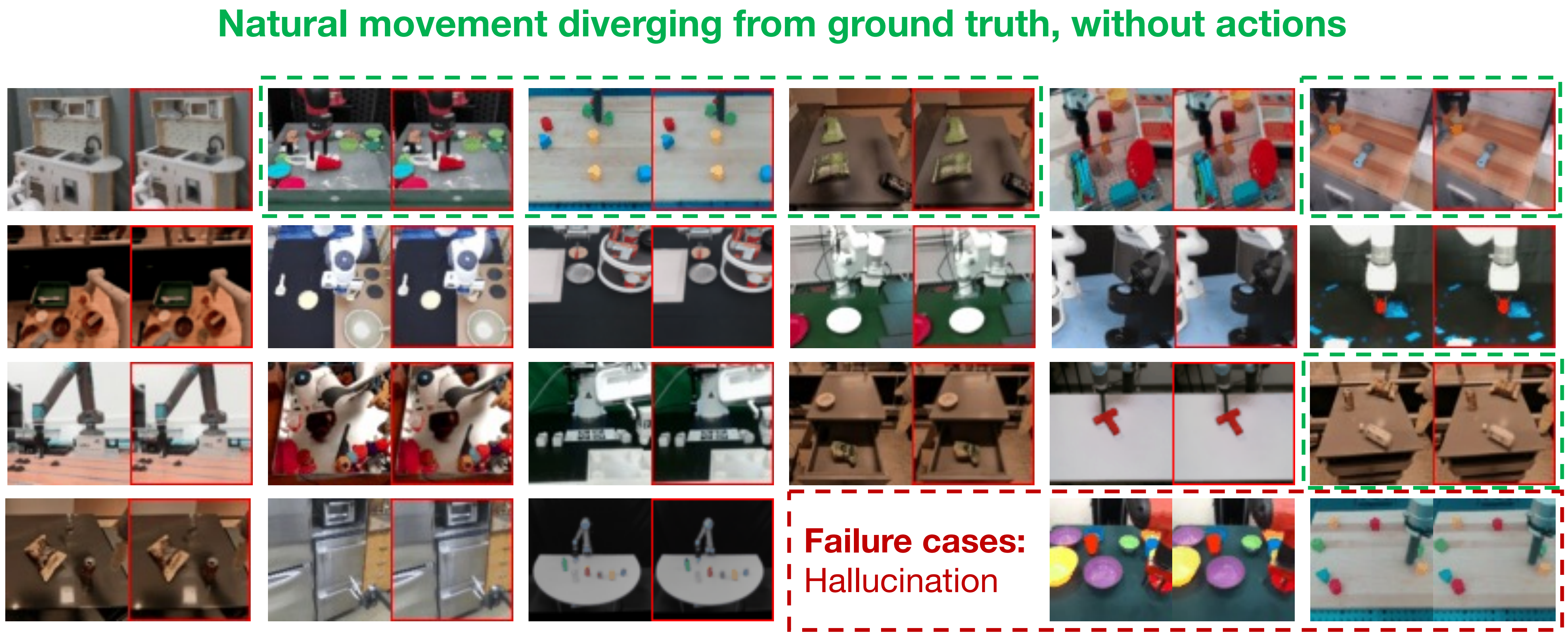
RoboNet [15]	FVD↓	PSNR↑	SSIM↑	LPIPS↓
action-conditioned & 64×64 resolution				
MaskViT [26]	133.5	23.2	80.5	4.2
SVG [87]	123.2	23.9	87.8	6.0
GHVAE [94]	95.2	24.7	89.1	3.6
FitVid [3]	**62.5**	**28.2**	89.3	**2.4**
iVideoGPT (ours)	63.2±0.01	27.8±0.01	**90.6**±0.02	4.9±0.00
action-conditioned & 256×256 resolution				
MaskViT [26]	211.7	20.4	67.1	17.0
iVideoGPT (ours)	**197.9**±0.66	**23.8**±0.00	**80.8**±0.01	**14.7**±0.01

Initially pre-trained action-free, flexibly allows for action-conditioning

Primary experiments at 64×64, easily extended to high resolution 256×256

iVideoGPT provides competitive performance compared to state-of-the-art methods, MAGVIT for BAIR and FitVid for RoboNet

Video Samples: Open X-Embodiment (Action-free)



Left: ground truth, right: prediction.
Red border: context frames, green border: predicted frames.

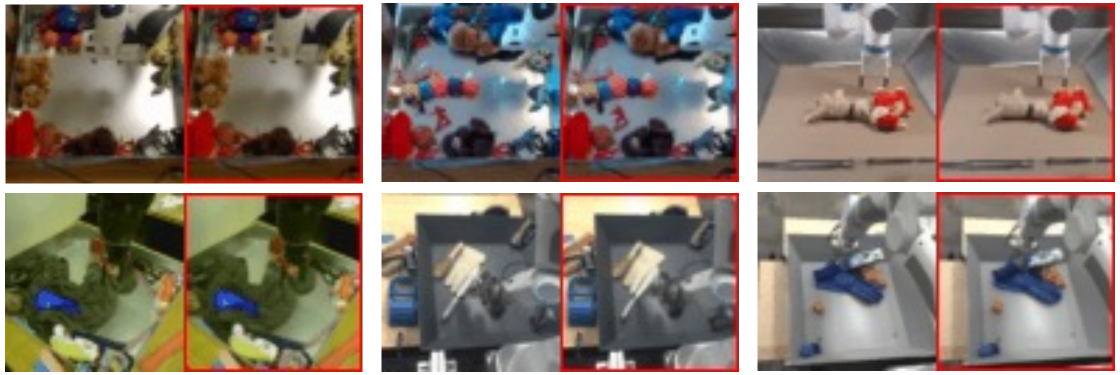
Video Samples: BAIR Robot Pushing & RoboNet

BAIR Robot Pushing Ebert et al. CoRL 2017

Action-free

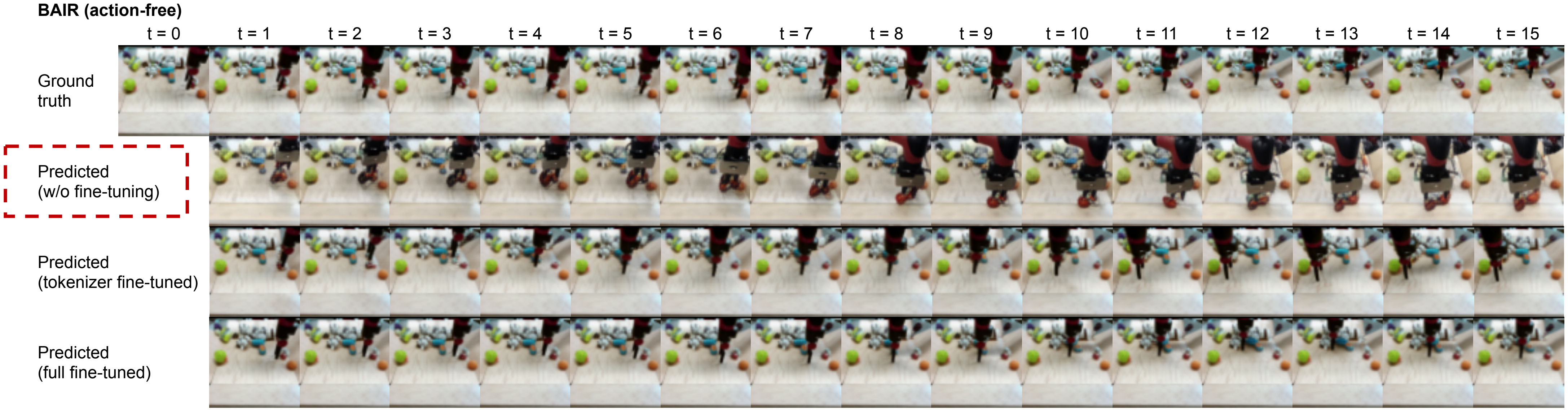
Action-conditioned

RoboNet (Action-conditioned) Dasari et al. CoRL 2019



High Resolution: 256×256

Zero-shot Prediction & Tokenization Adaptation



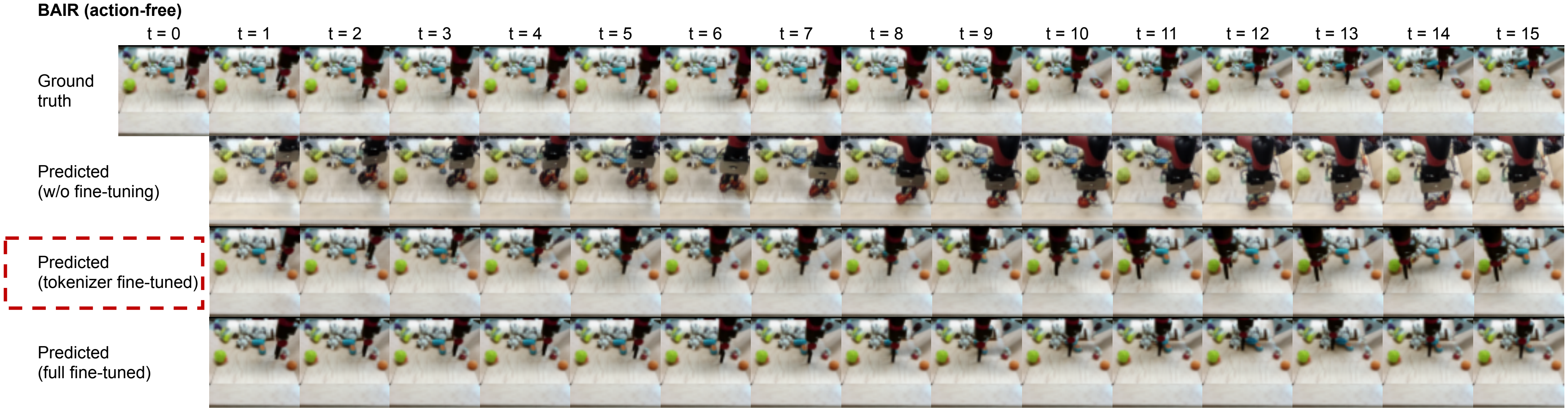
Zero-shot prediction:

Interestingly, without any fine-tuning, iVideoGPT can predict natural movements of a robot gripper—albeit another one originally from our pre-training dataset.

× Insufficient diversity of pre-training data

√ Effectively separates context and motions

Zero-shot Prediction & Tokenization Adaptation

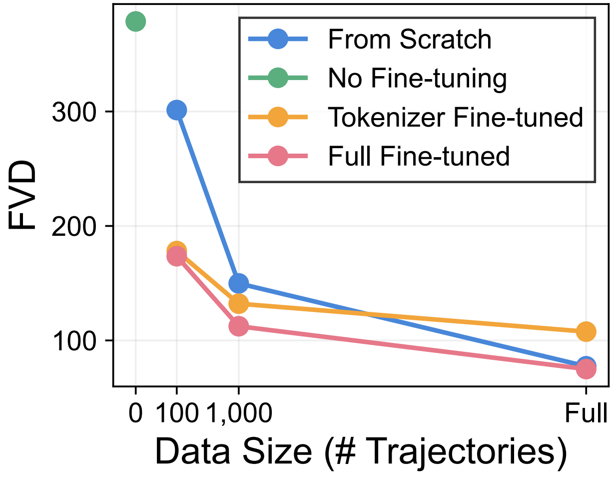


Tokenization adaptation:

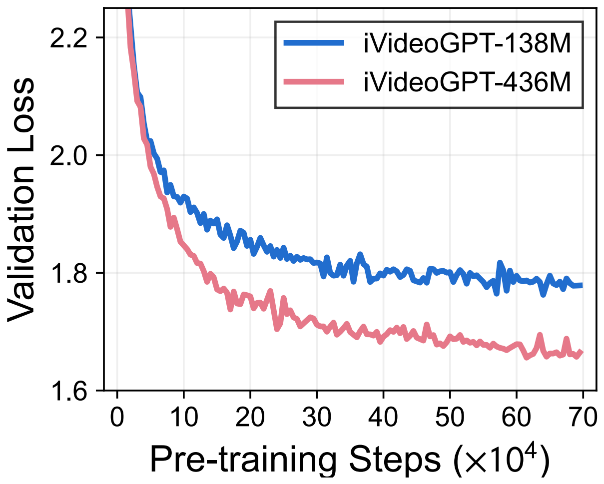
After adapting tokenizer, the transformer that is not fine-tuned itself successfully transfers the pre-trained knowledge and predicts movements for the new robot type, providing a similar perceptual quality as the fully fine-tuned model

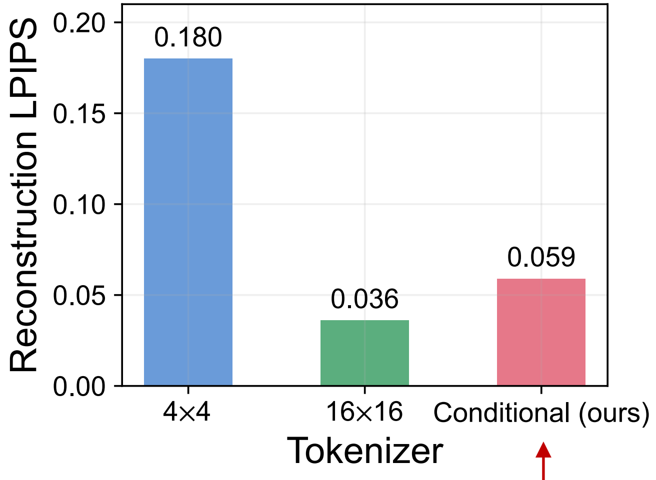
√ Lightweight alignment while keeping the transformer intact

Model Analysis



138M: 12 layers, 768 hidden dim
436M: 24 layers, 1024 hidden dim





Context frames: 16 x 16 tokens
Future frames: 4 x 4 tokens

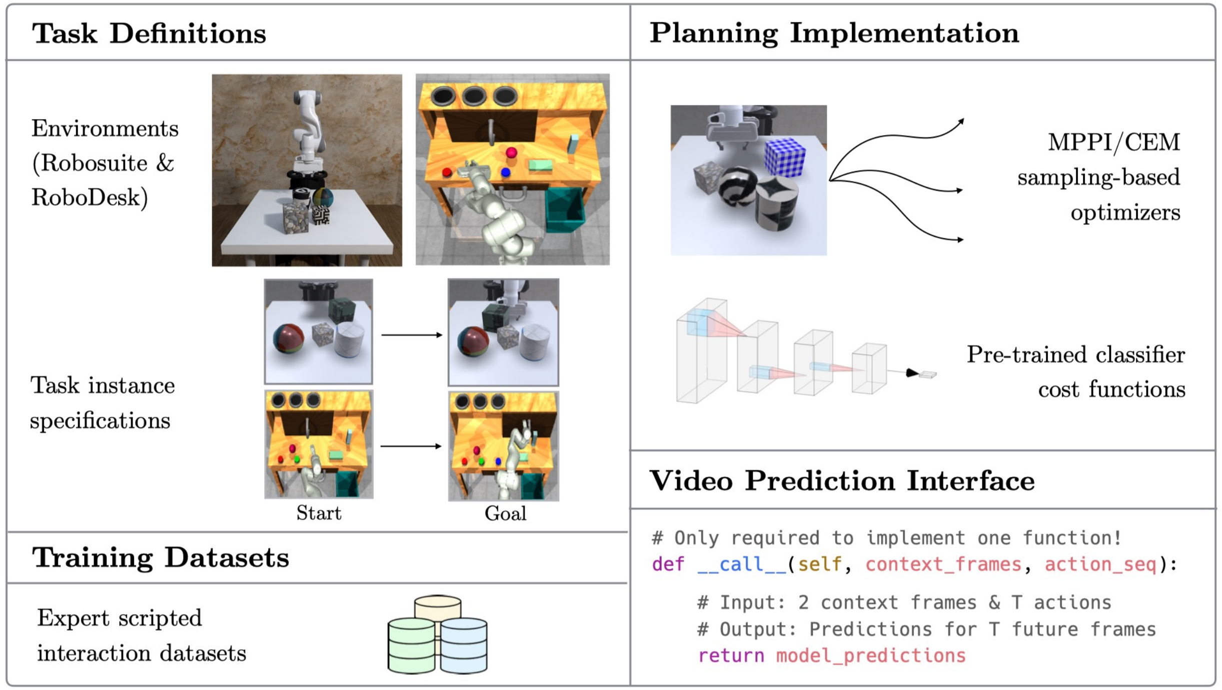
Takeaways:

1. Pre-training offers minimal benefits with full downstream data available, yet the advantages become significant under data scarcity.
2. Larger model sizes and increased computation can build more powerful iVideoGPTs
3. The proposed conditional tokenization slightly compromises reconstruction but significantly reduces the number of an autoregressive transformer's forward passes by 16×.

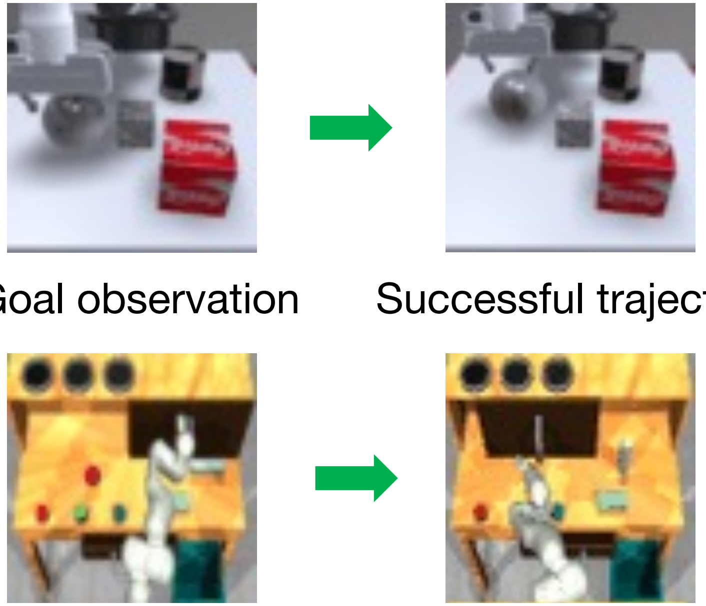
Visual Planning

Excellent perceptual metrics do not always correlate with effective control performance

VP2: A control-centric benchmark for video prediction



Model-predictive control

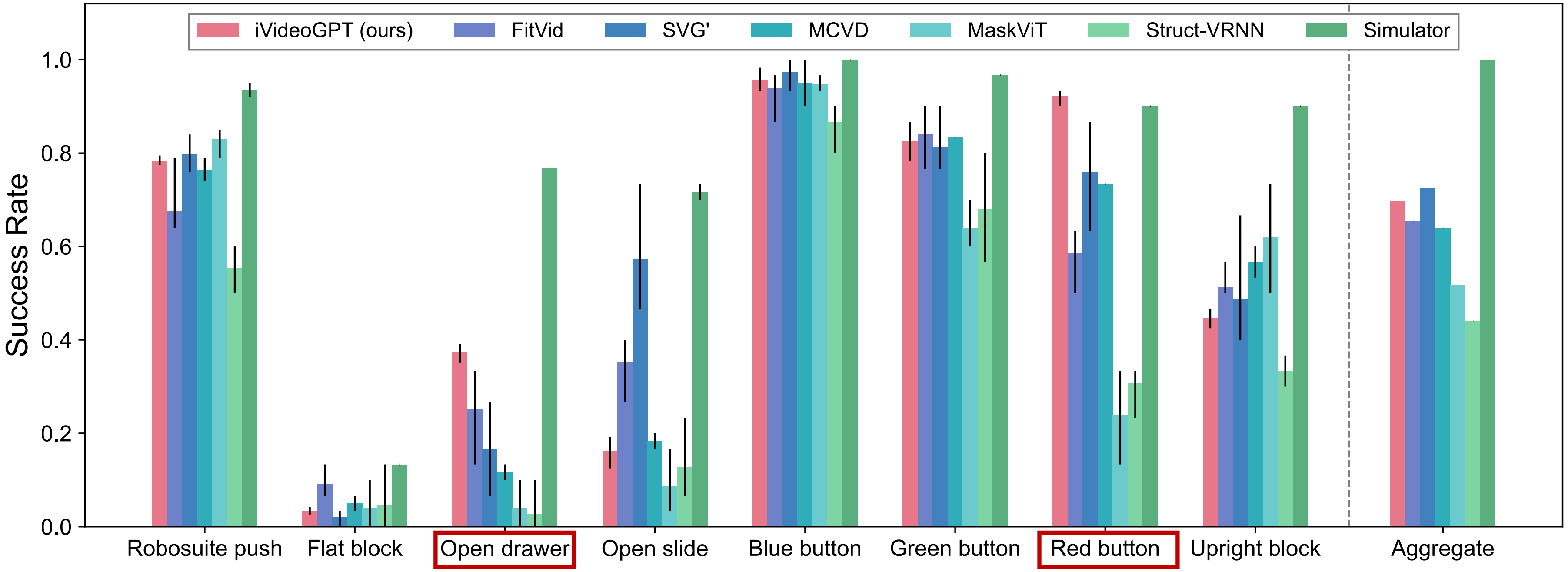


Goal observation | Successful trajectory

Goal observation | Successful trajectory

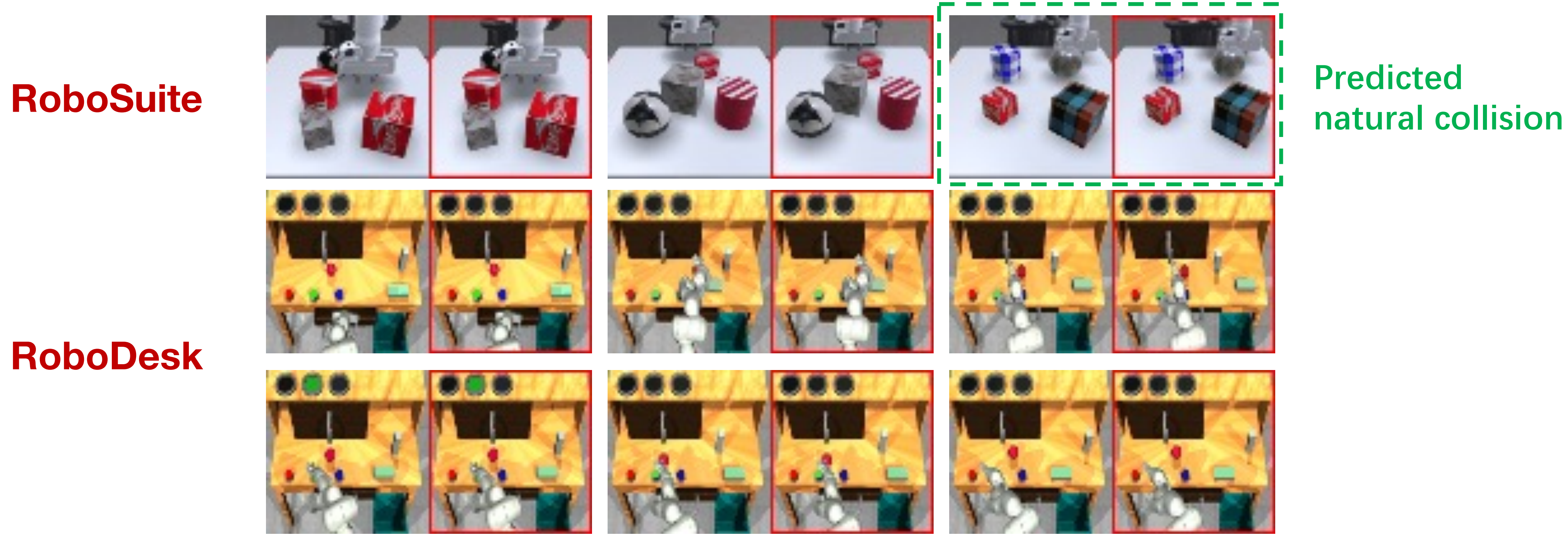
Tian, Stephen, et al. A Control-Centric Benchmark for Video Prediction. ICLR 2023.

Visual Planning: VP2



iVideoGPT outperforms all baselines in two RoboDesk tasks with a large margin and achieves comparable average performance to the strongest model.

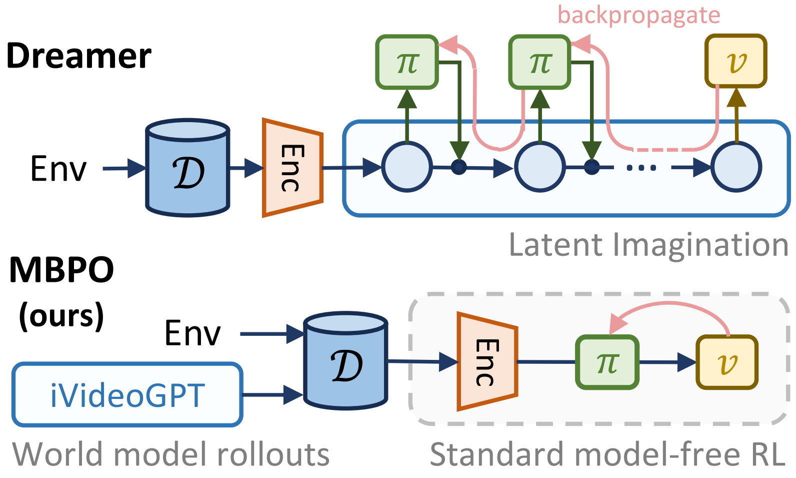
Video Samples: VP2



Visual Model-based RL

Model-based RL with iVideoGPT:

- **Adapted from MBPO**: Augments the replay buffer with synthetic rollouts into replay buffer to train a standard actor-critic RL algorithm (DrQ-v2)
- **Eliminate latent imagination**: Decoupling model and policy learning can substantially simplify the design space, facilitating real-world applications.



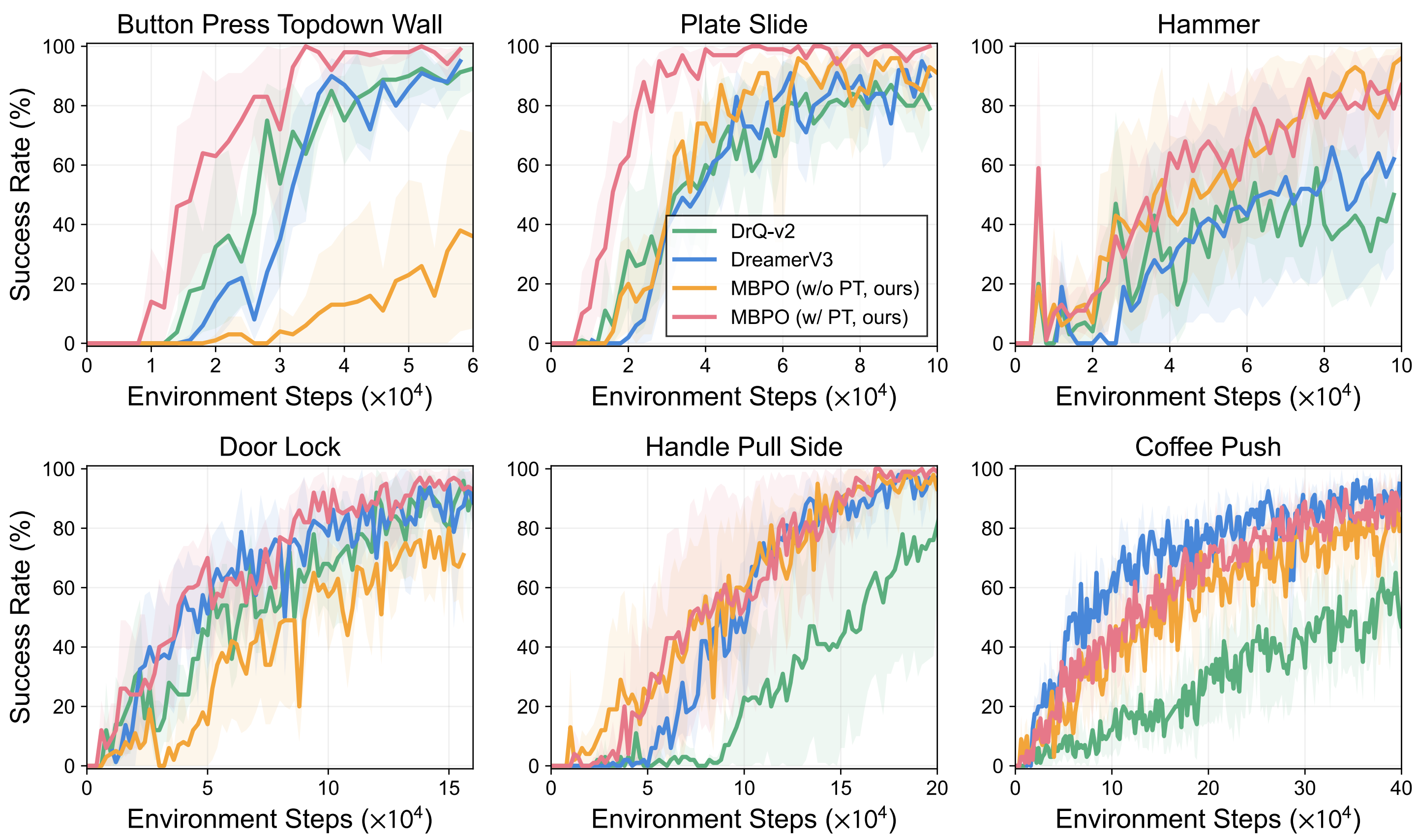
Algorithm 1 Model-Based Policy Optimization (MBPO), adapted from [40]

Initialize actor-critic π_ϕ, v_ψ, world model p_θ
Initialize real replay buffer $\mathcal{D}_{\text{real}}$ with random policy
Initially train model p_θ on $\mathcal{D}_{\text{real}}$
Initialize imagined replay buffer $\mathcal{D}_{\text{imag}}$ with random rollouts using p_θ
for N steps **do**
    `// Training`
    if model update step **then**
        Update world model p_θ on a mini-batch from $\mathcal{D}_{\text{real}}$
    end if
    Update actor-critic π_ϕ, v_ψ with model-free objectives on a mini-batch from $\mathcal{D}_{\text{imag}} \cup \mathcal{D}_{\text{real}}$
    `// Data collection`
    if model rollout step **then**
        Sample a mini-batch of o_t uniformly from $\mathcal{D}_{\text{real}}$
        Perform k-step model rollout starting from o_t using policy π_ϕ; add to $\mathcal{D}_{\text{imag}}$
    end if
    Take action in environment according to π_ϕ; add to $\mathcal{D}_{\text{real}}$
end for

Janner, Michael, et al. When to trust your model: Model-based policy optimization. NeurIPS 2019.
Yarats, Denis, et al. Mastering visual continuous control: Improved data-augmented reinforcement learning. ICLR 2022.

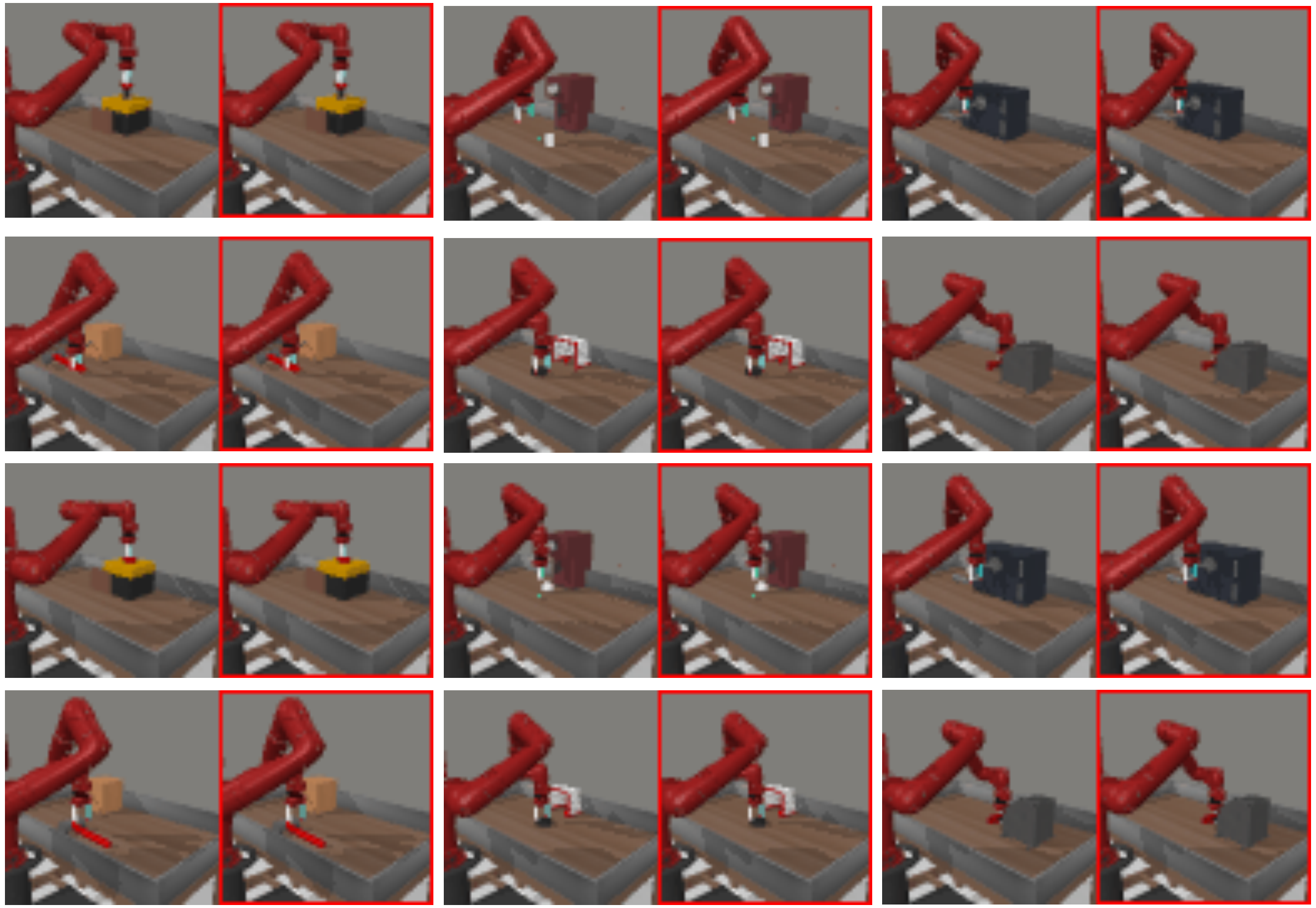
Visual Model-based RL: Meta-world

Six Meta-world manipulation tasks



- Empowered by iVideoGPT, simple MBPO not only remarkably improves the sample efficiency over its model-free counterpart but also matches or exceeds DreamerV3.
- To our knowledge, the first reported success of MBPO to visual continuous control.
- World models trained from scratch can degenerate the sample efficiency

Video Samples: Meta-world

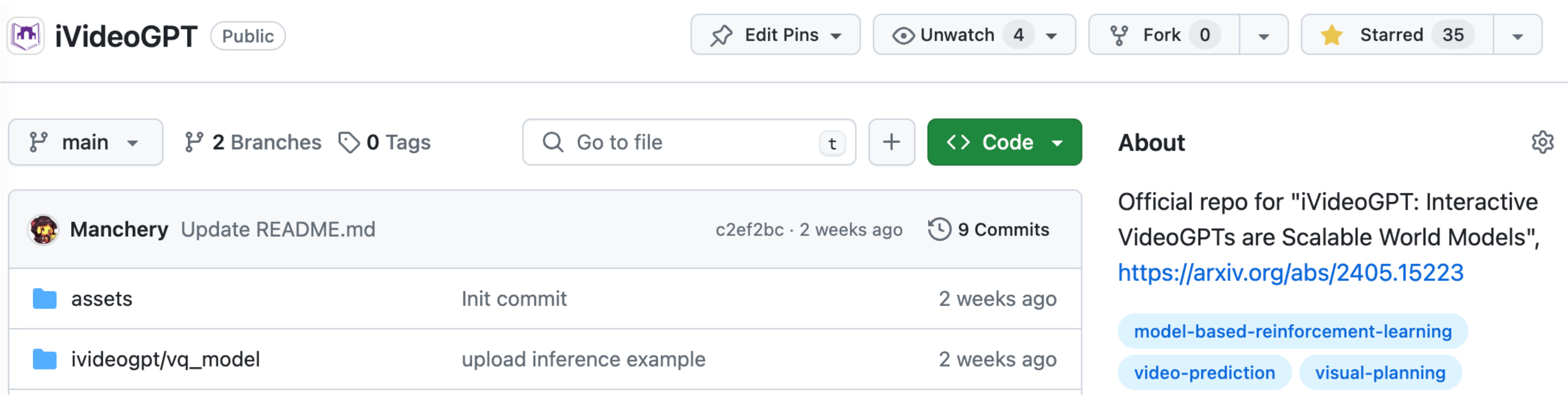


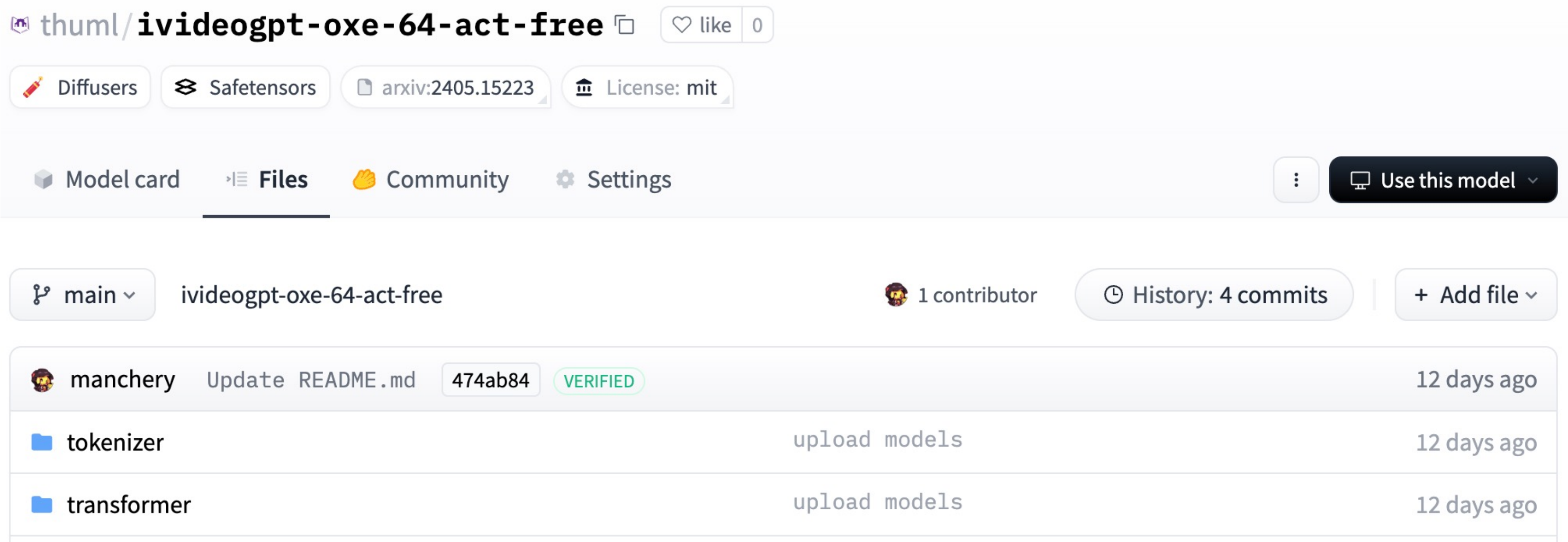
True and predicted rewards are labeled at the top left corner.

Summary

- **iVideoGPT**, a generic and efficient world model architecture based on compressive tokenization and autoregressive transformers
- Pre-trained on millions of human and robotic manipulation trajectories
- Adapted to a wide range of downstream tasks, particularly:
  - Accurate and generalizable video prediction
  - Simplified yet performant model-based RL

Open Source



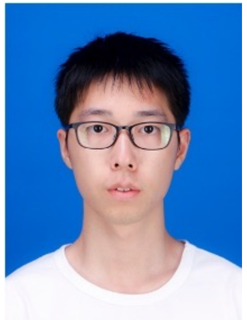


https://github.com/thuml/iVideoGPT

Pre-trained model and inference code released

王建民

龙明盛



吴佳龙

马浩宇

邓朝一

冯宁亚

尹绍沣

大数据系统软件国家工程研究中心

清华大学软件学院机器学习课题组