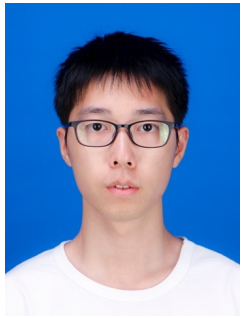
NeurIPS | 2023

Thirty-seventh Conference on 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



Jialong Wu

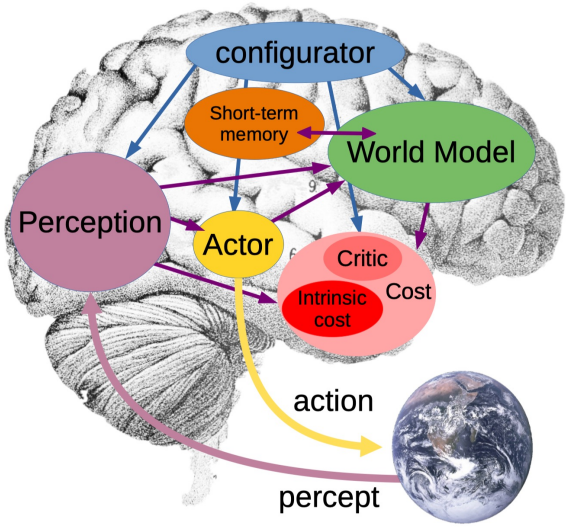
Haoyu Ma

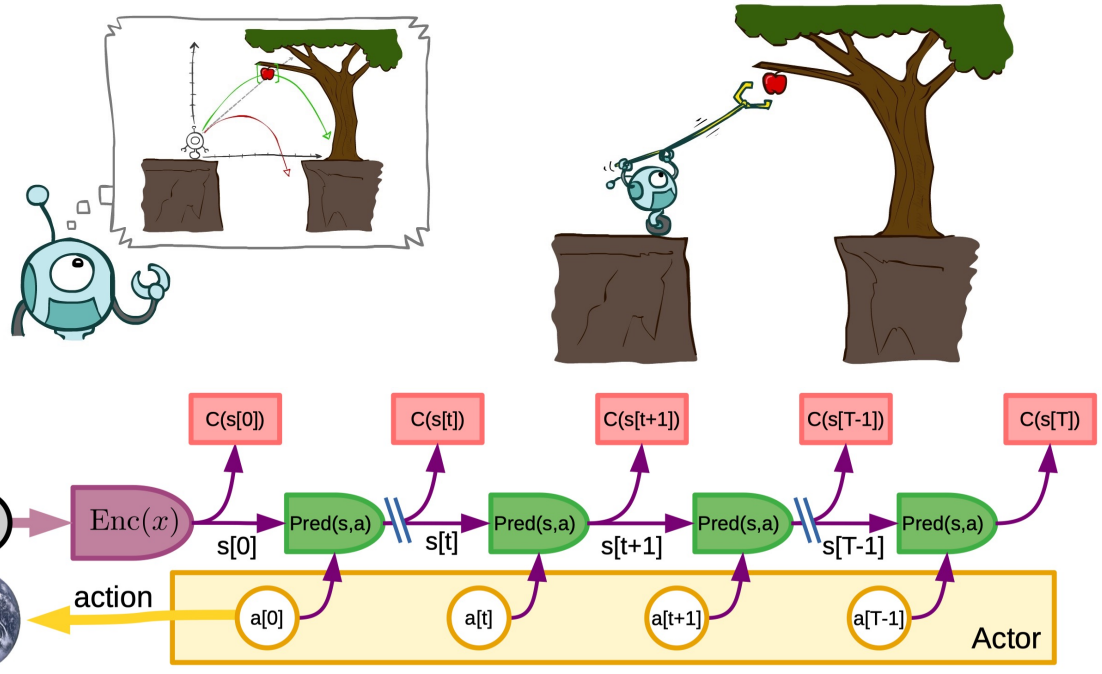
Chaoyi Deng

Mingsheng Long

Tsinghua University

World Models





World Models:

internal models of how the world works

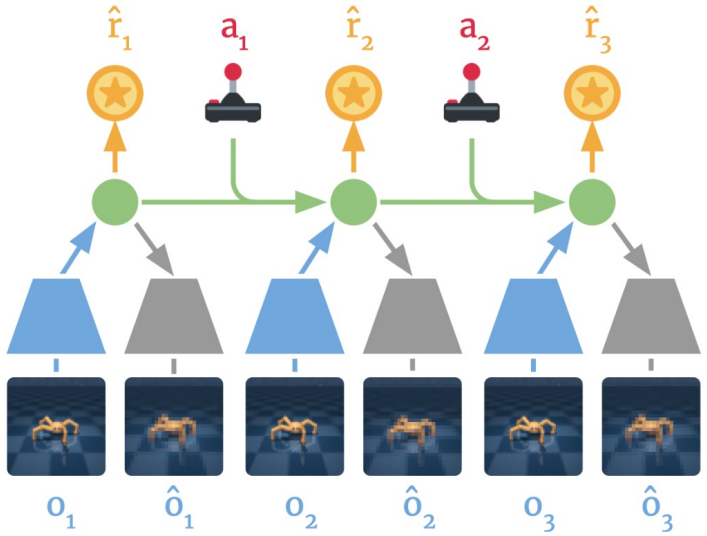
Model-based Agents:

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.

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

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

Behavior Learning: Purely on imaginary latent trajectories

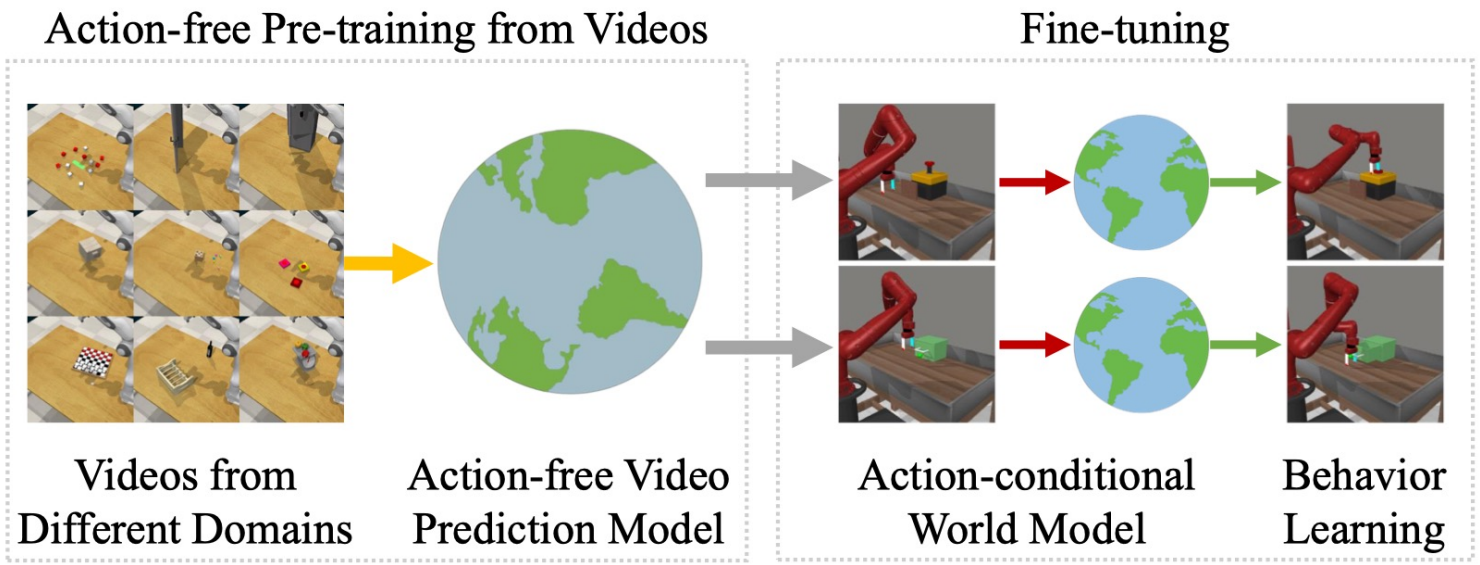
Hafner, Danijar, et al. Dream to control: Learning behaviors by latent imagination. ICLR 2020.

Hafner, Danijar, et al. Mastering atari with discrete world models. ICLR 2021.

APV: Action-free Pre-training from Videos

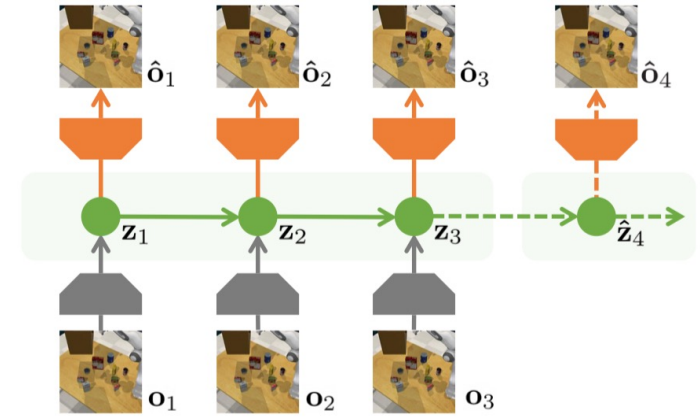
How to represent and acquire prior knowledge for RL?

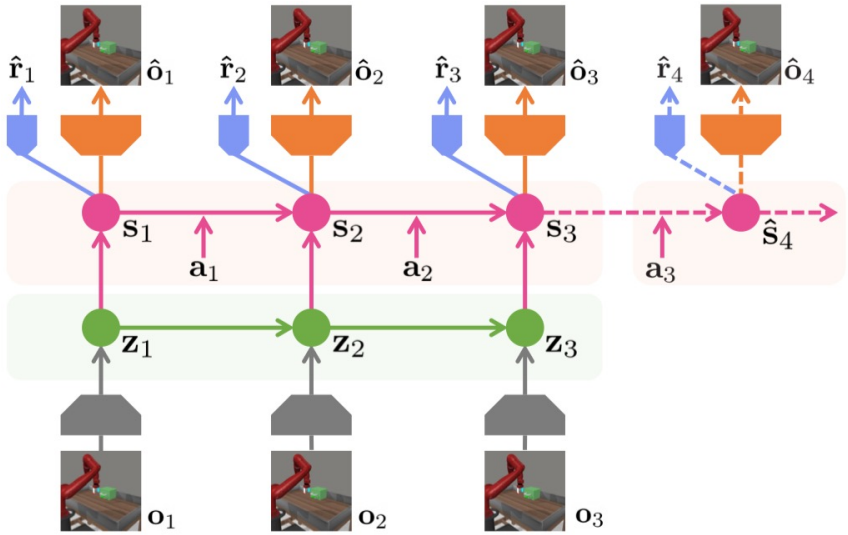
Learning representations useful for understanding the dynamics via generative pretraining on videos



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

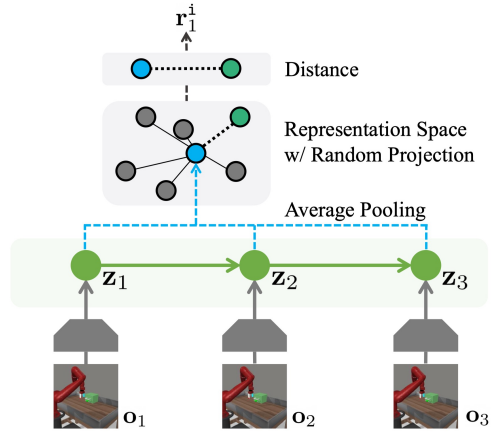
Action-conditional

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

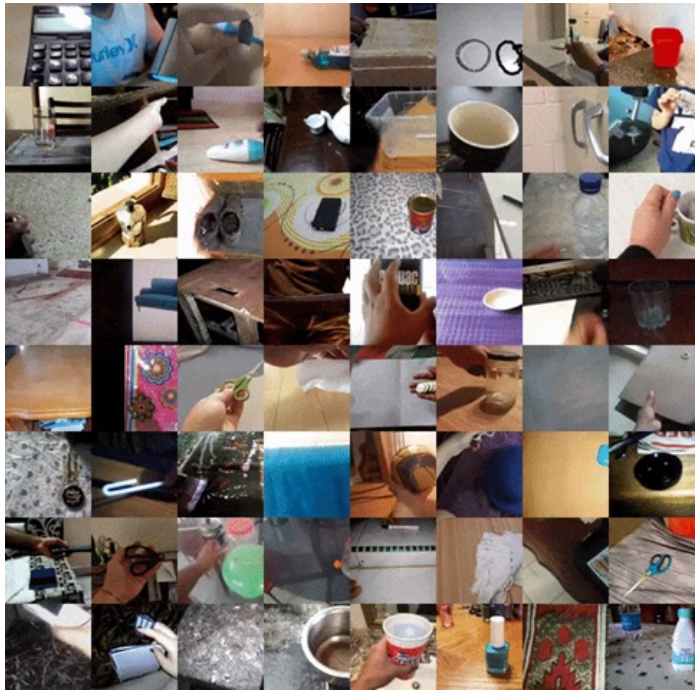
1. Pre-train an action-free latent video prediction model
2. Stack an action-conditional model when fine-tuned for MBRL
3. Video-based intrinsic bonus for better exploration



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

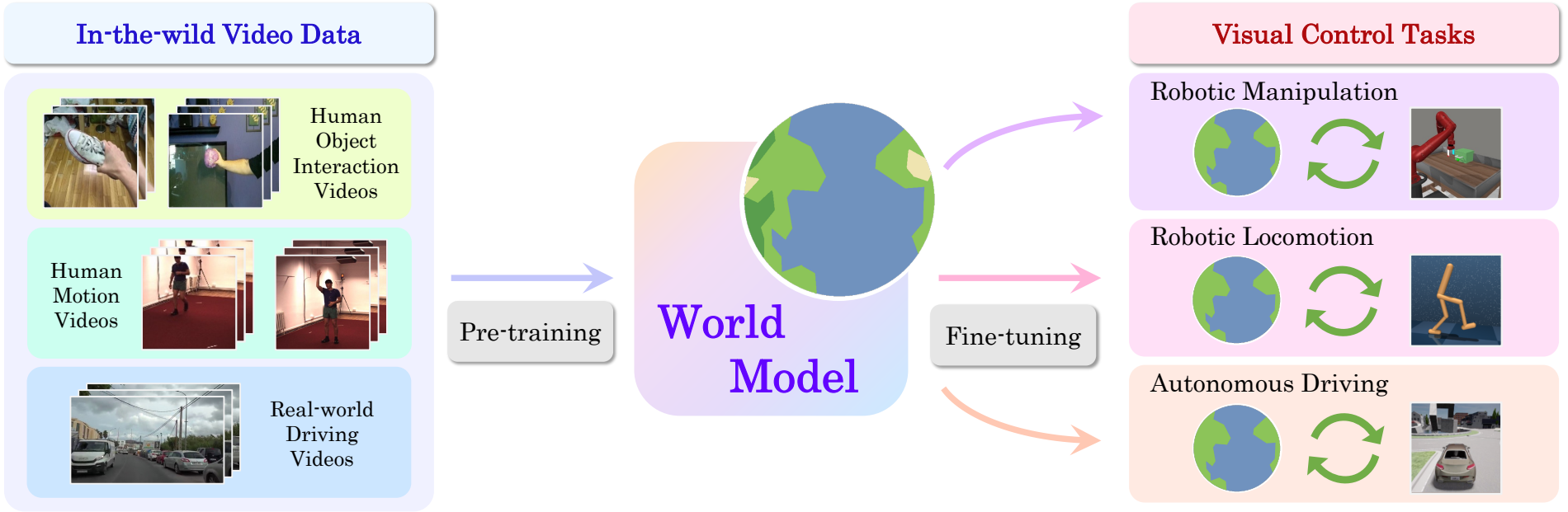
Ego4D

Grauman et al., Facebook AI. CVPR 2022

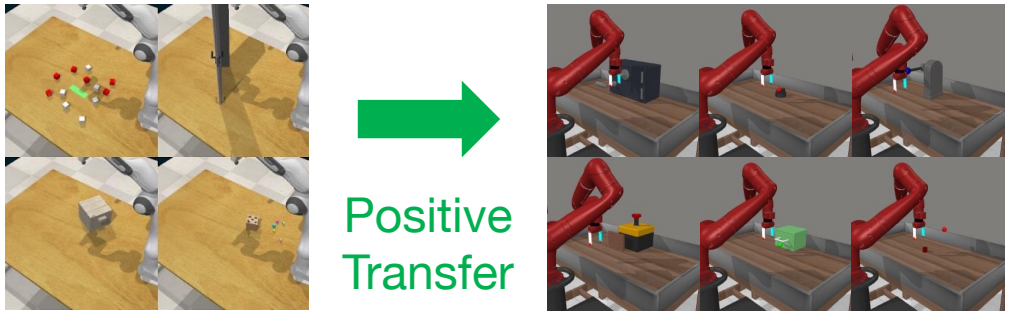
IPV: In-the-wild Pre-training from Videos

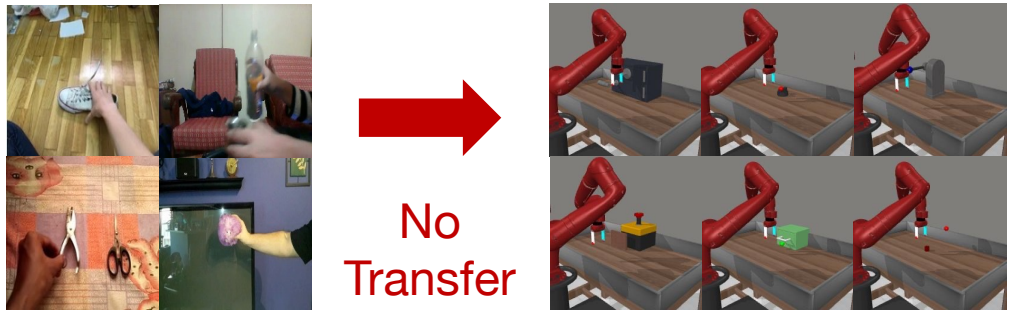
Towards a general world model:

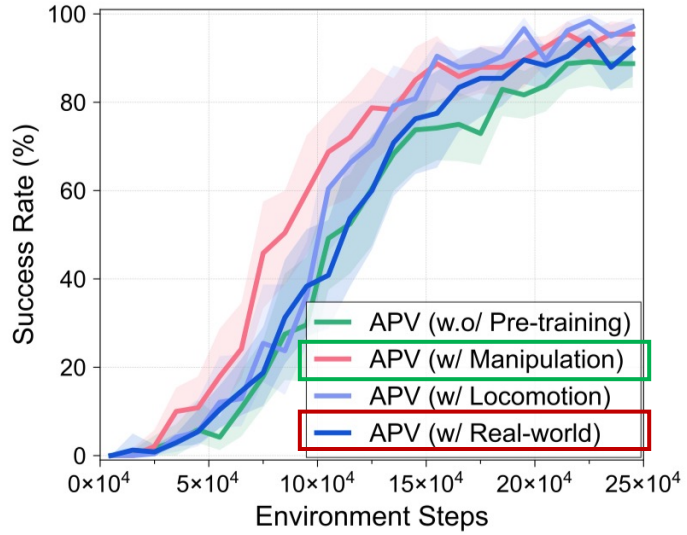
- How to overcome the visual complexity and diversity?
- What is the shared knowledge transferable from in-the-wild video domain to visual control tasks?



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







Why pre-training fails?

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

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

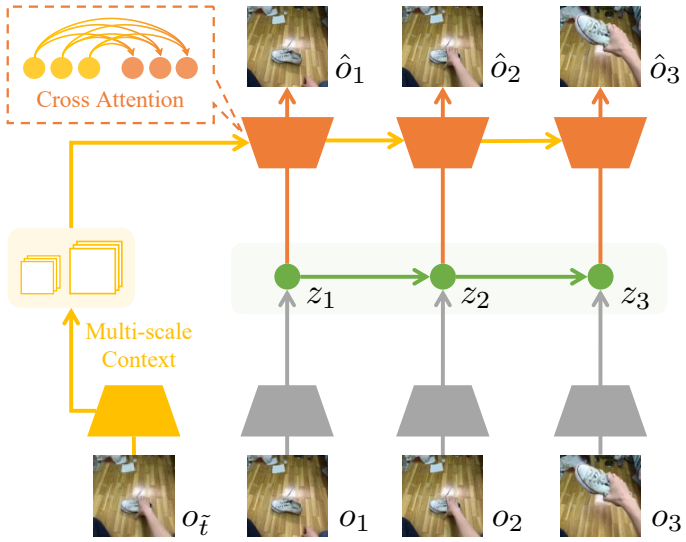
Seo, Younggyo, et al. Reinforcement learning with action-free pre-training from videos. ICML 2022.

Contextualized World Models (ContextWM)

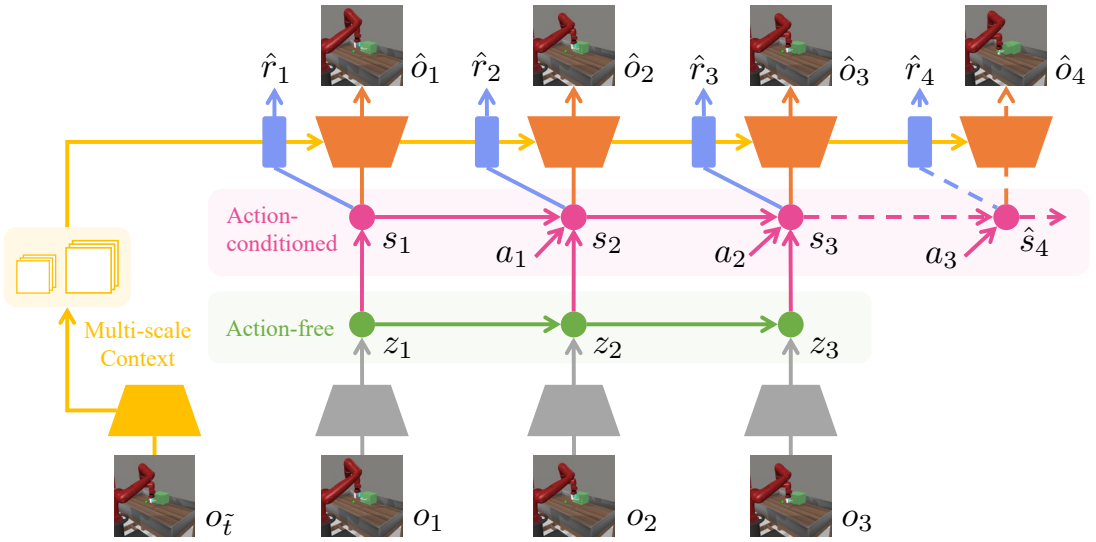
Overview:

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



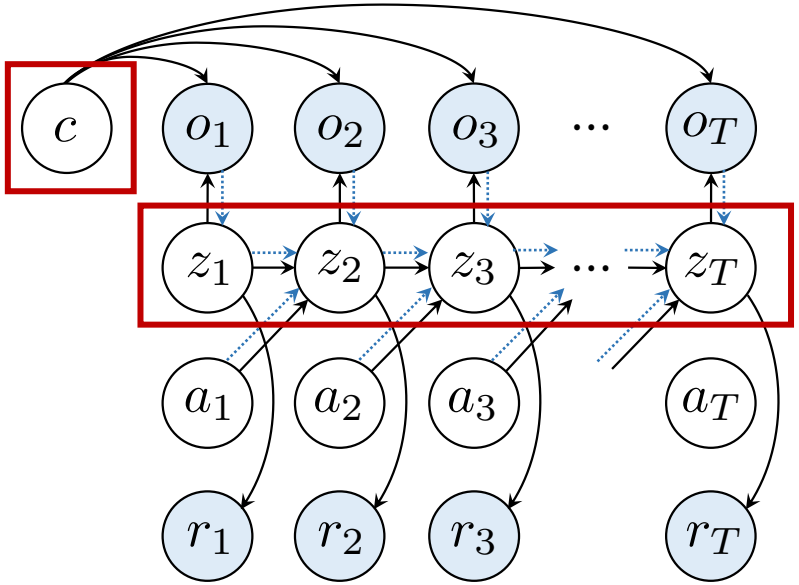
Step 2. Fine-tuning on downstream visual control tasks with MBRL

Contextualized Latent Dynamics Models

Our insight:

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

Distinct **contexts**



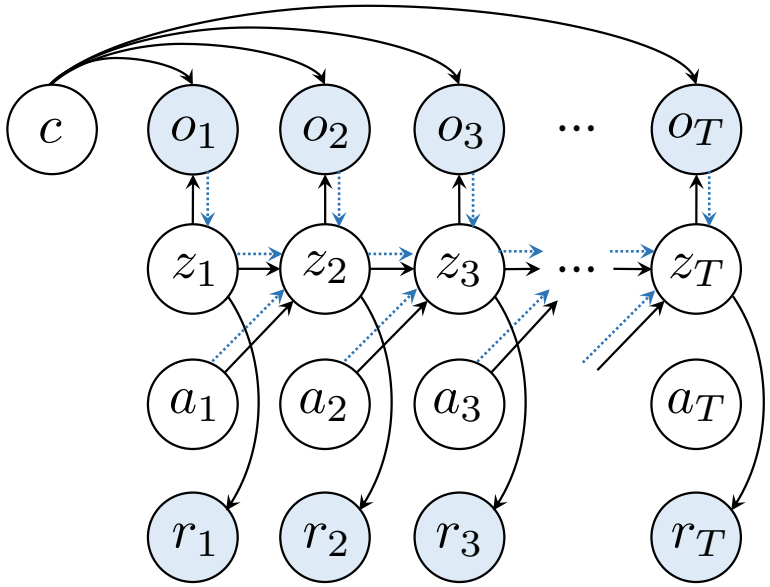
Similar **dynamics**

Contextualized Latent Dynamics Models

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

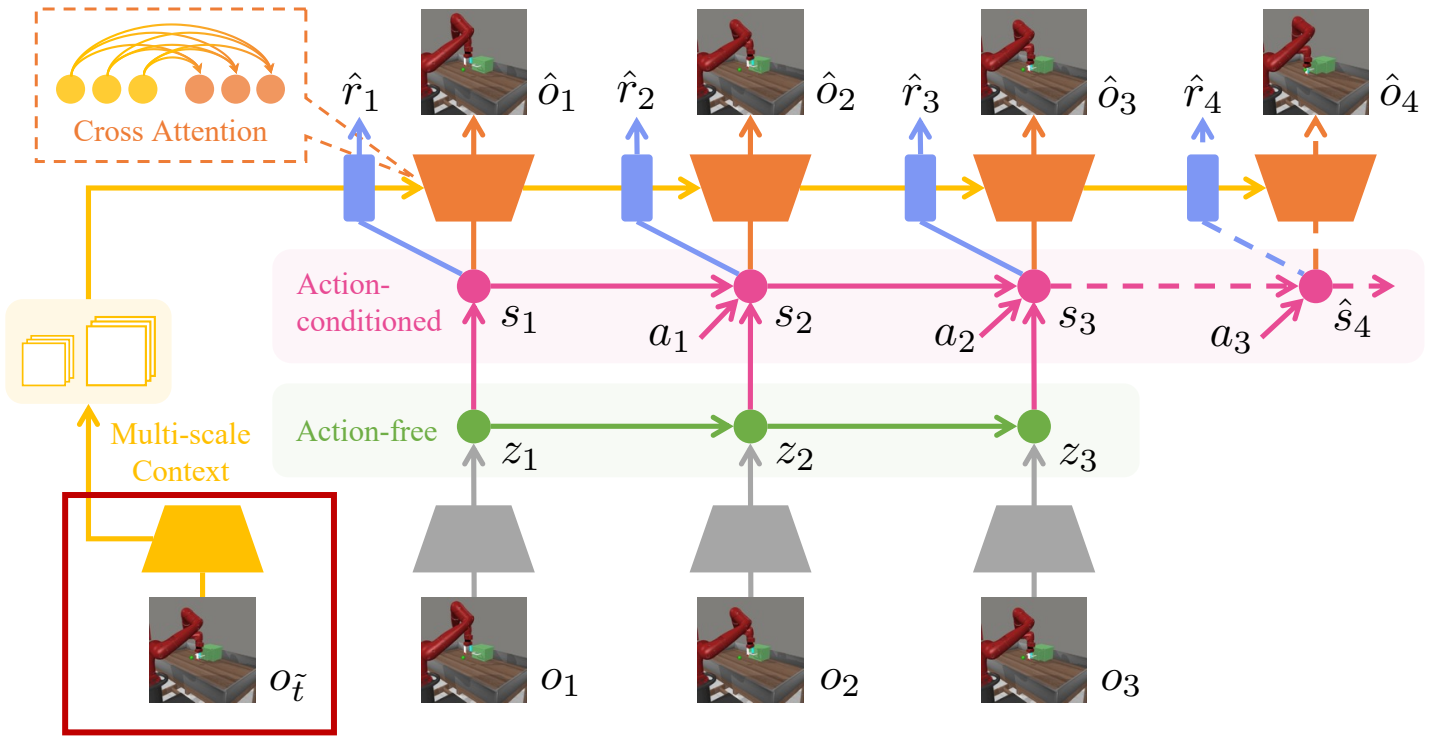
context-unaware latent inference

contextualized image loss



- Learn with ELBO of conditional $\ln p_\theta\left(o_{1:T}, r_{1:T} \mid a_{1:T}, c\right)$ 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



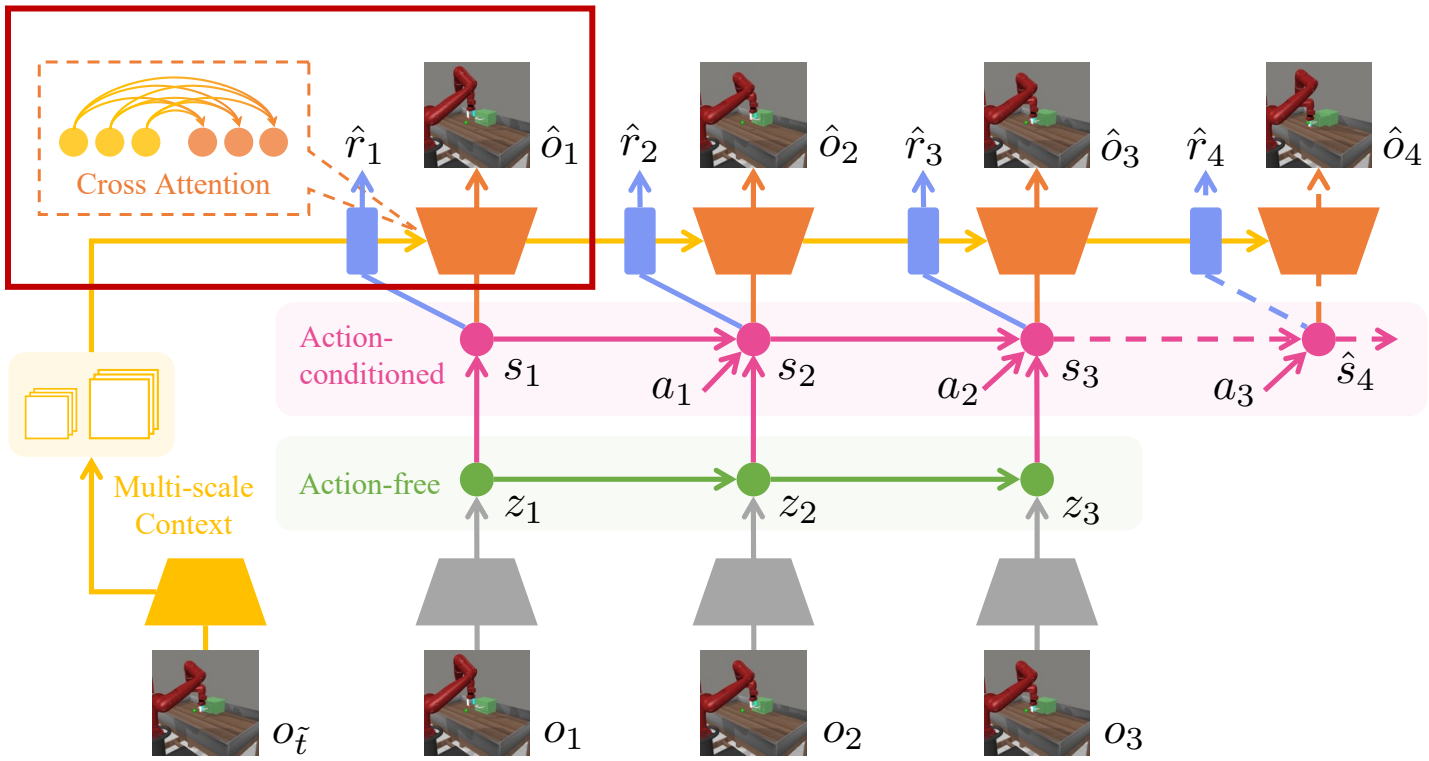
Context formulation:

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:

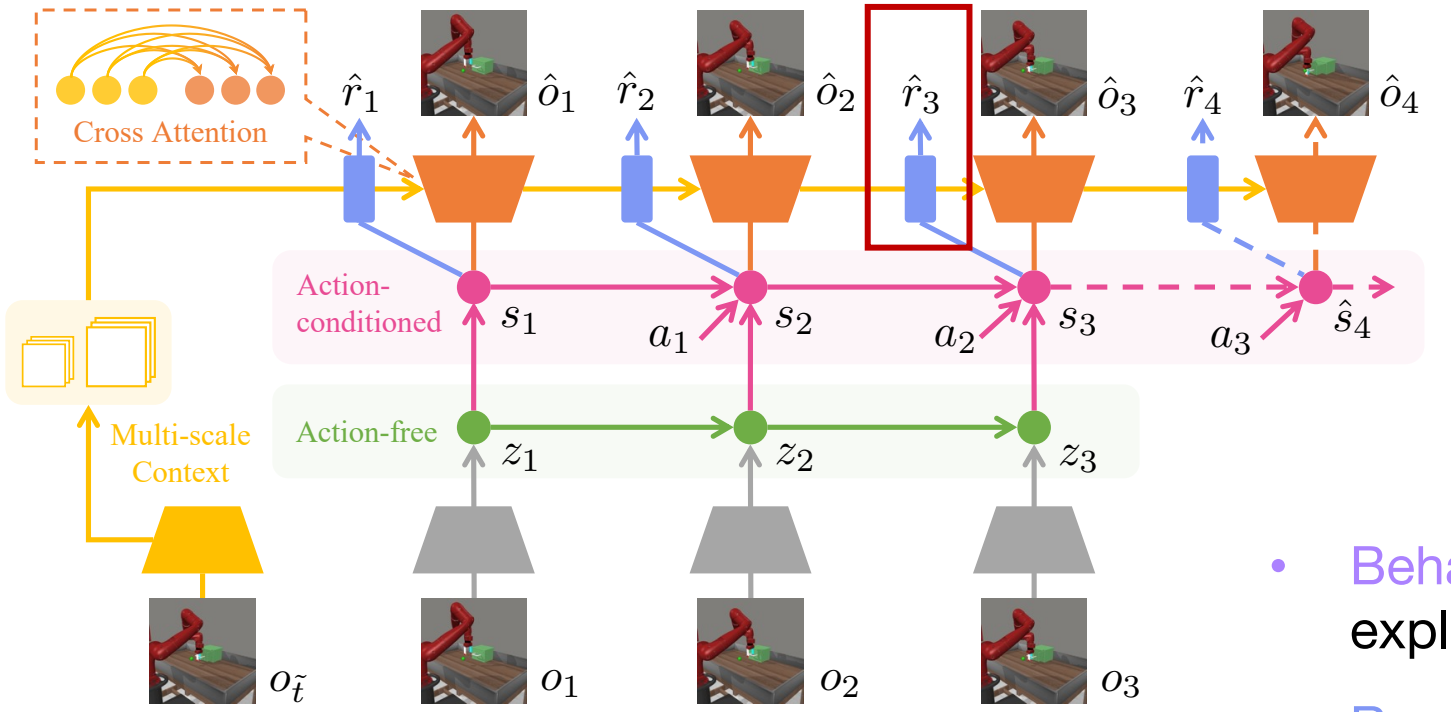
1. U-Net-style multi-scale feature shortcuts
2. Instead of naive concatenation forcing a spatial alignment, adaptive cross-attention mechanism is utilized

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

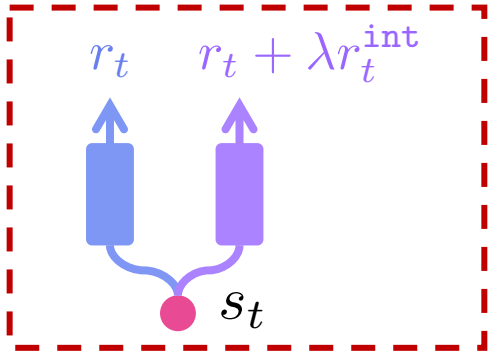
$$\text{Cross-Attention: } R = \text{Attention}\left(QW^Q, KW^K, VW^V\right) \in \mathbb{R}^{hw \times c}$$

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

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

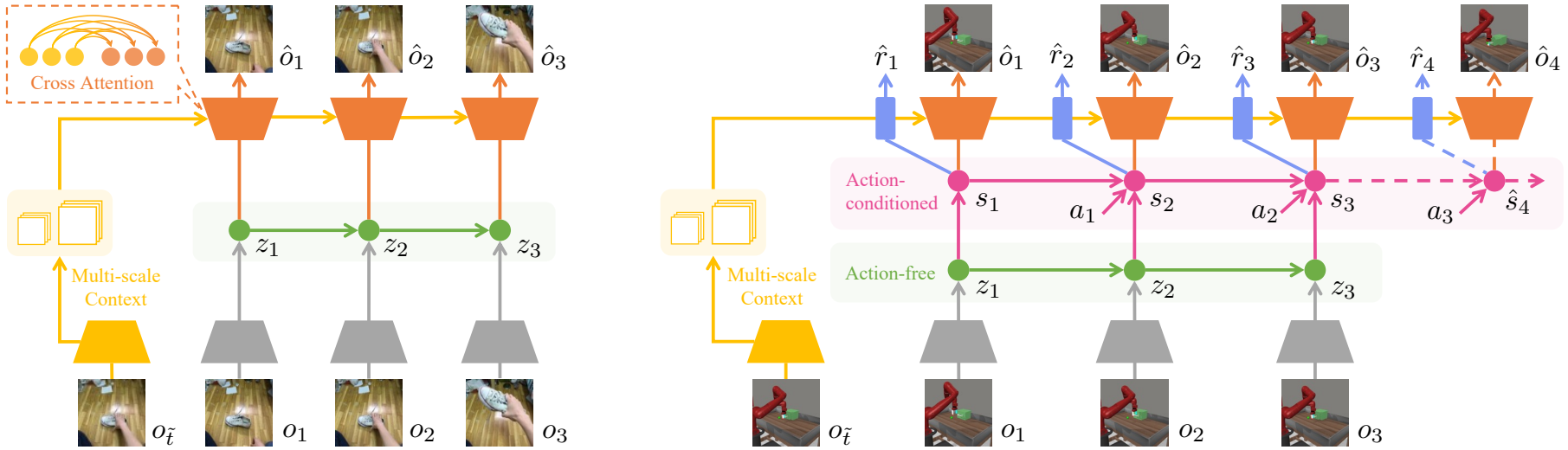
Contextualized World Models: An Implementation

Overall objective:

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

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

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



Experiments: Diverse Datasets & Tasks

Something-Something V2

Goyal et al. ICCV 2017

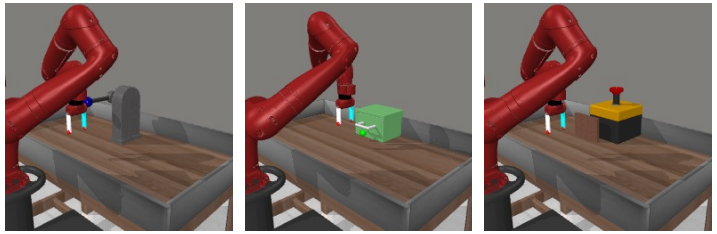
Human3.6M

Ionescu et al. TPAMI 2014

YouTube Driving

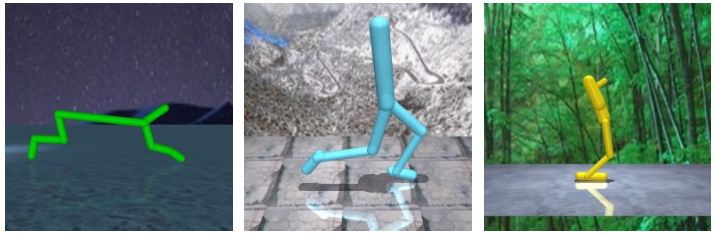
Zhang et al. ECCV 2022

Transfer



Meta-World

Yu et al. CoRL 2020



DMC Remastered

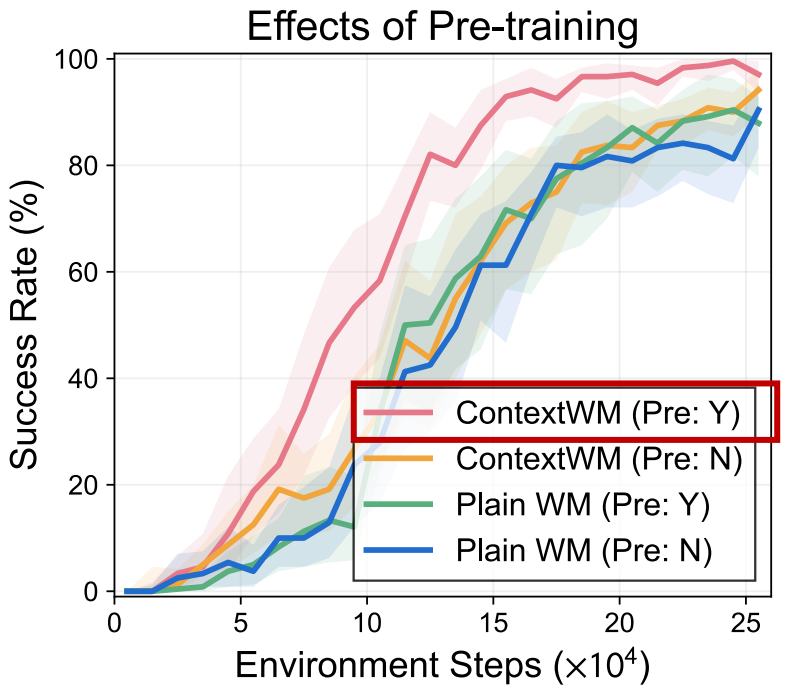
Grigsby et al. 2020

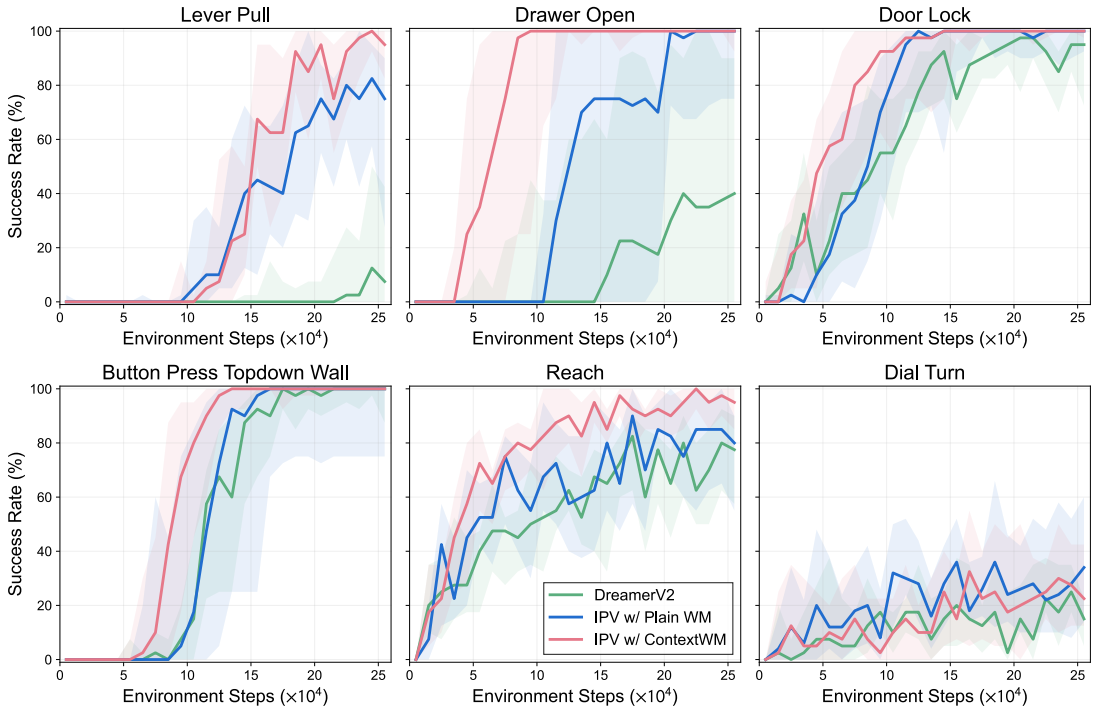
CARLA

Dosovitskiy et al. CoRL 2017

Main Results: Meta-world

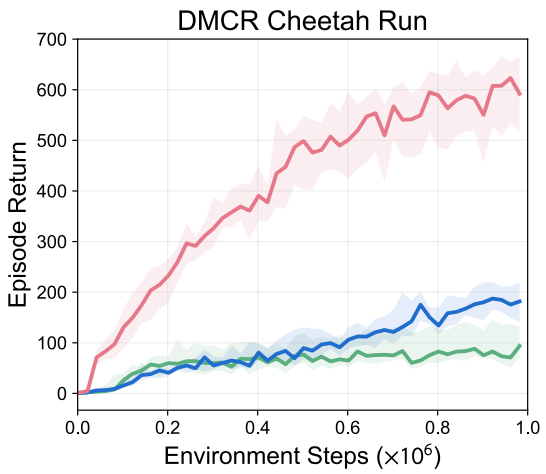
(aggregated results)

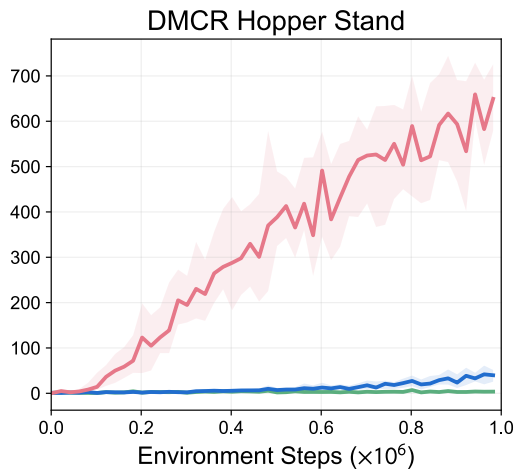


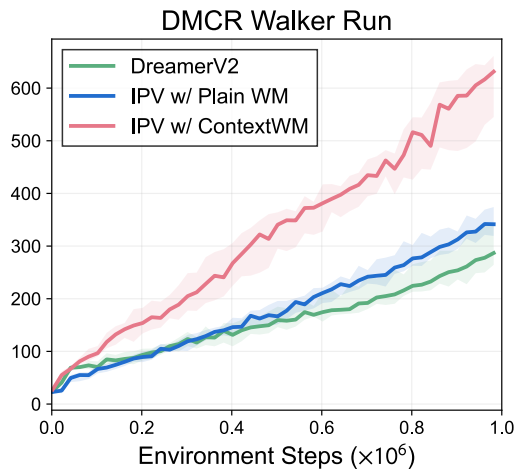


On six Meta-world tasks, ContextWM achieves significant positive transfer (from SSv2) in terms of sample efficiency, while a plain WM fails.

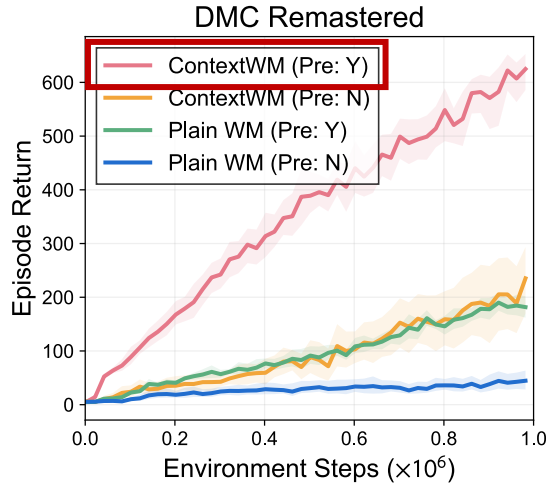
Main Results: DMC Remastered



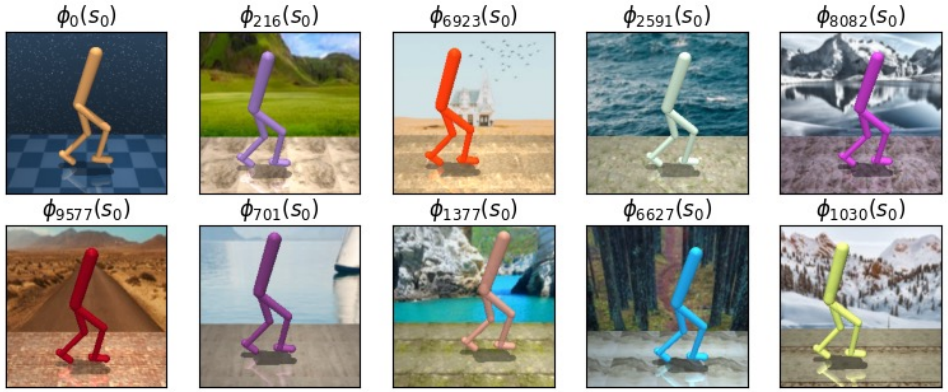




(aggregated results)

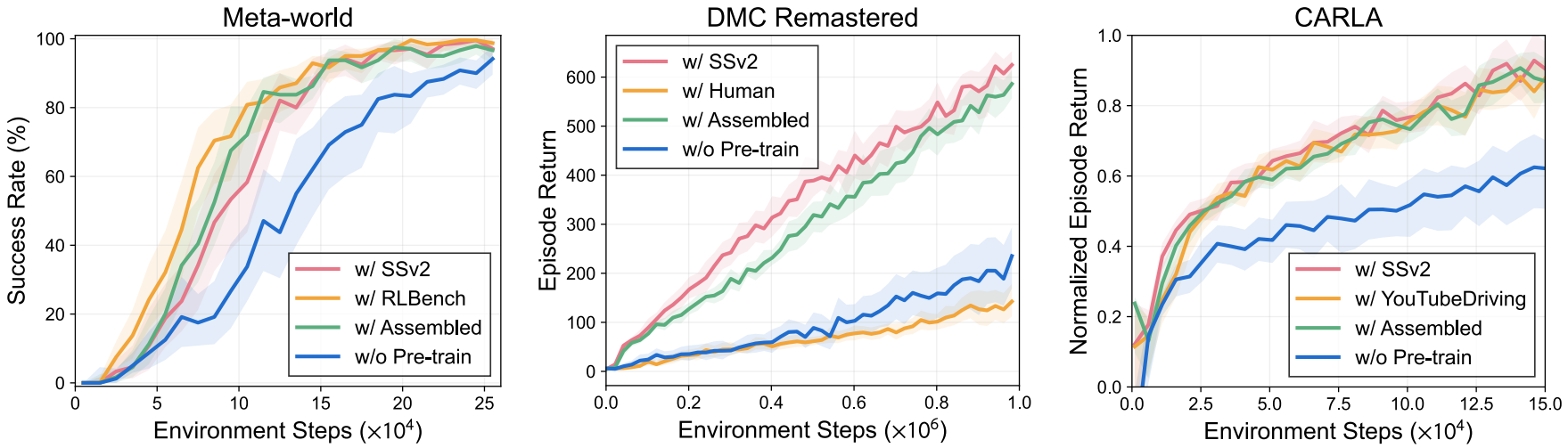


On visual generalization benchmark, pre-training from in-the-wild videos (SSv2) incurs significant performance boost, which is further unleashed by ContextWM.



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

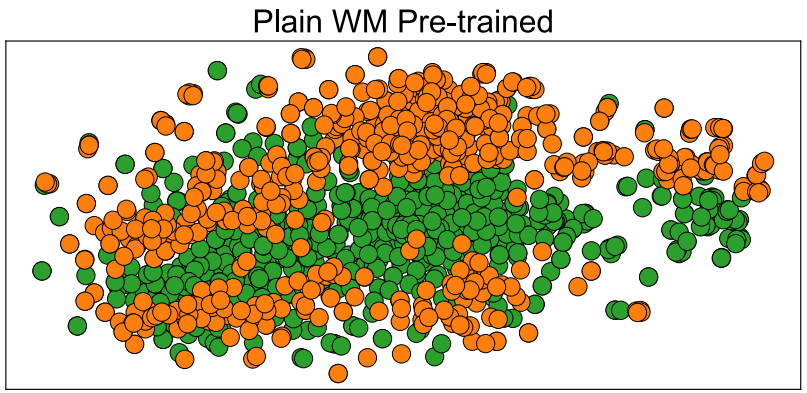
Effects of Pre-training Dataset Domain

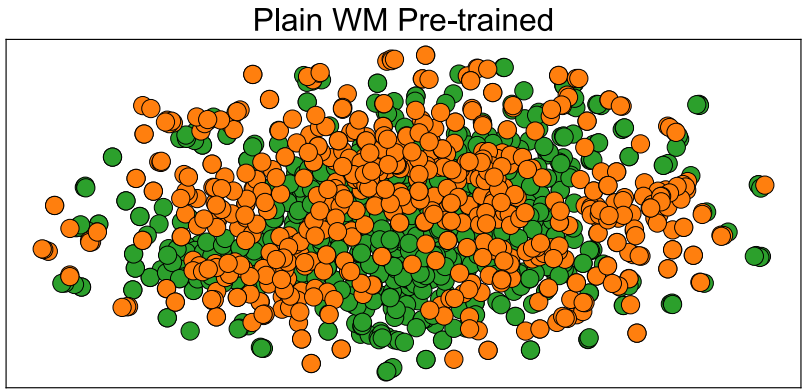


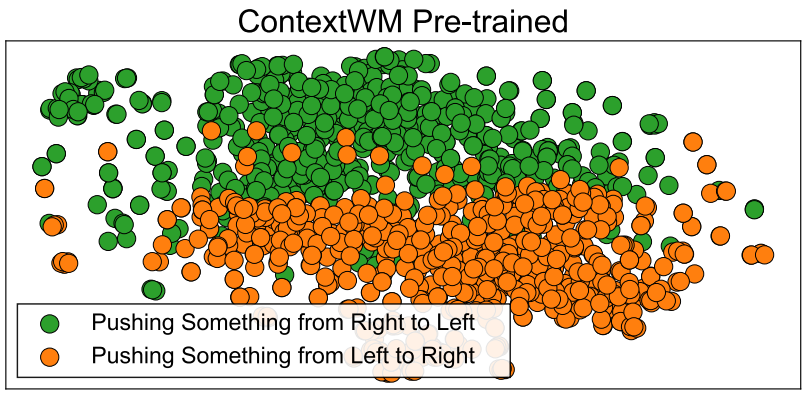
Takeaways:

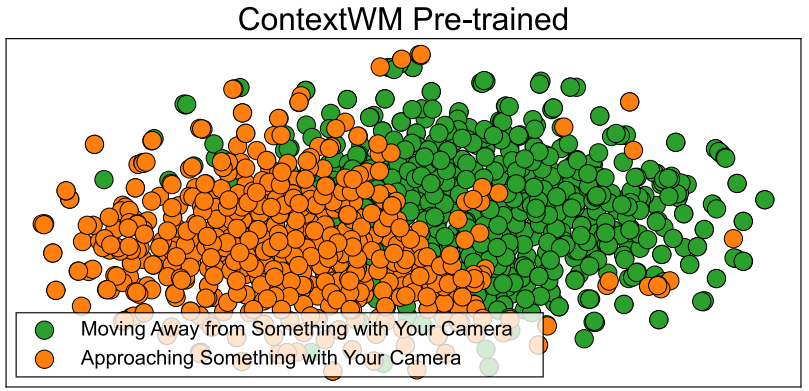
1. Human-object interaction data (SSv2) are generally beneficial.
2. A more similar domain (e.g. RLBench) is more useful, but more diverse datasets can serve as promising scalable alternatives.
3. Pre-training data lack of diversity (Human3.6M) can even be harmful.

Qualitative Evaluation: Video Representations



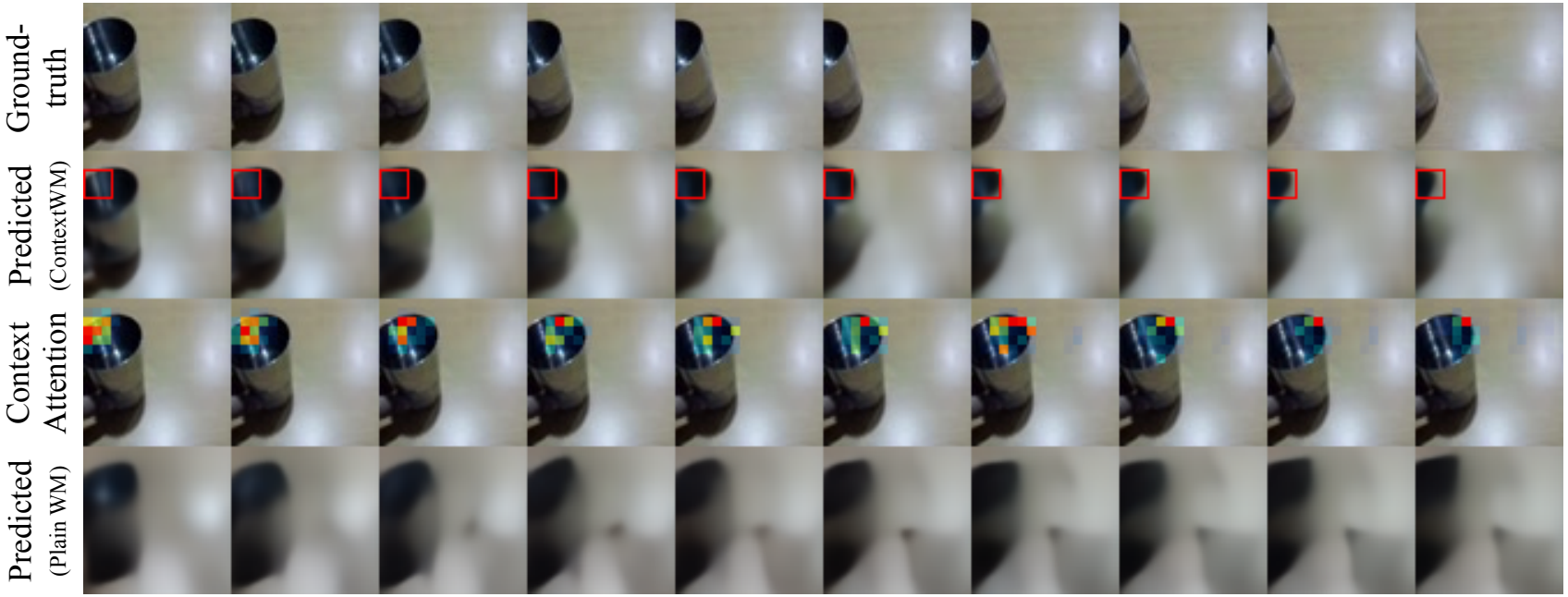






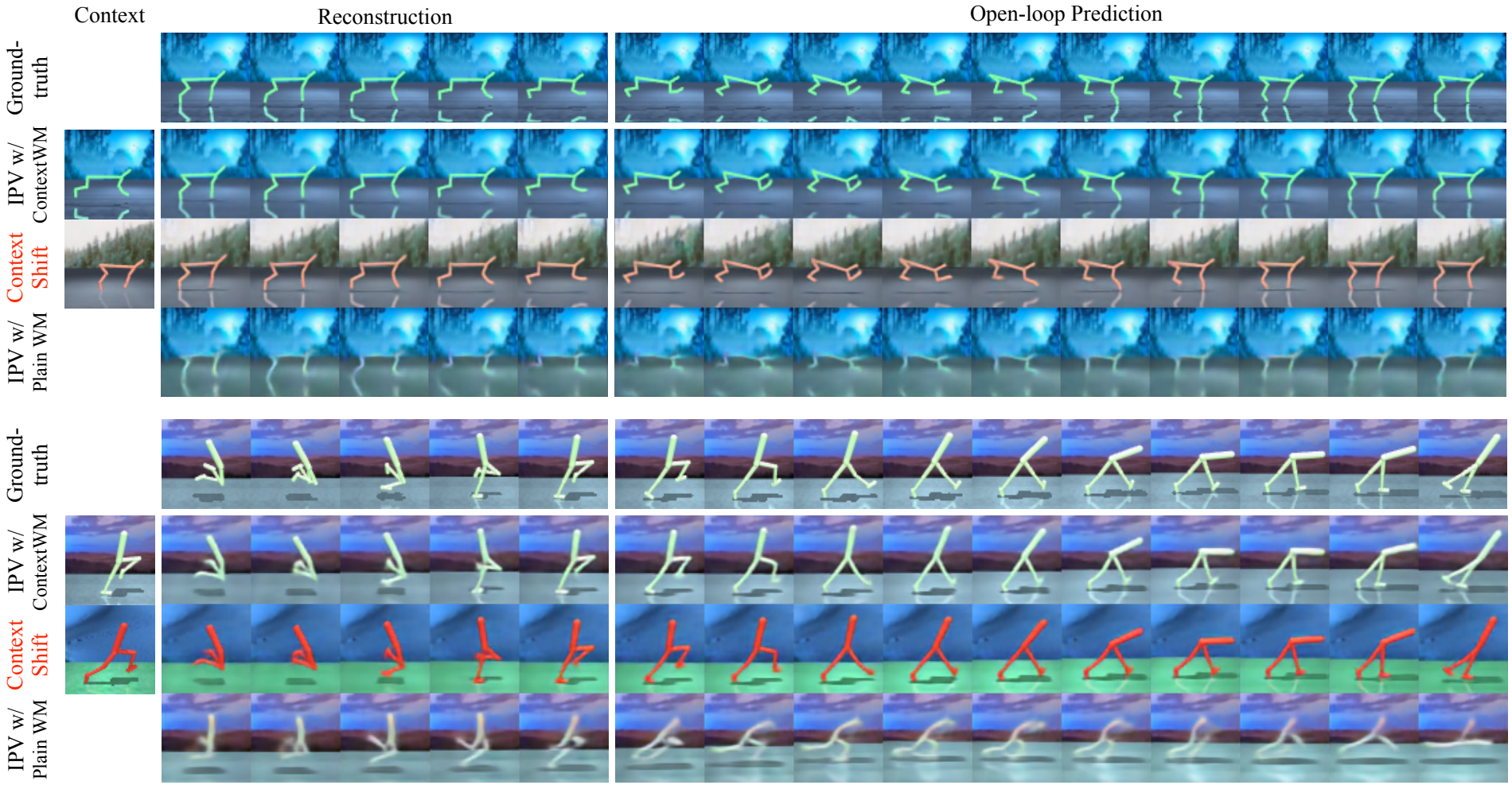
ContextWM learns representations well distributed according to different directions of motion, while not utilizing any labels of the videos in pre-training

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.

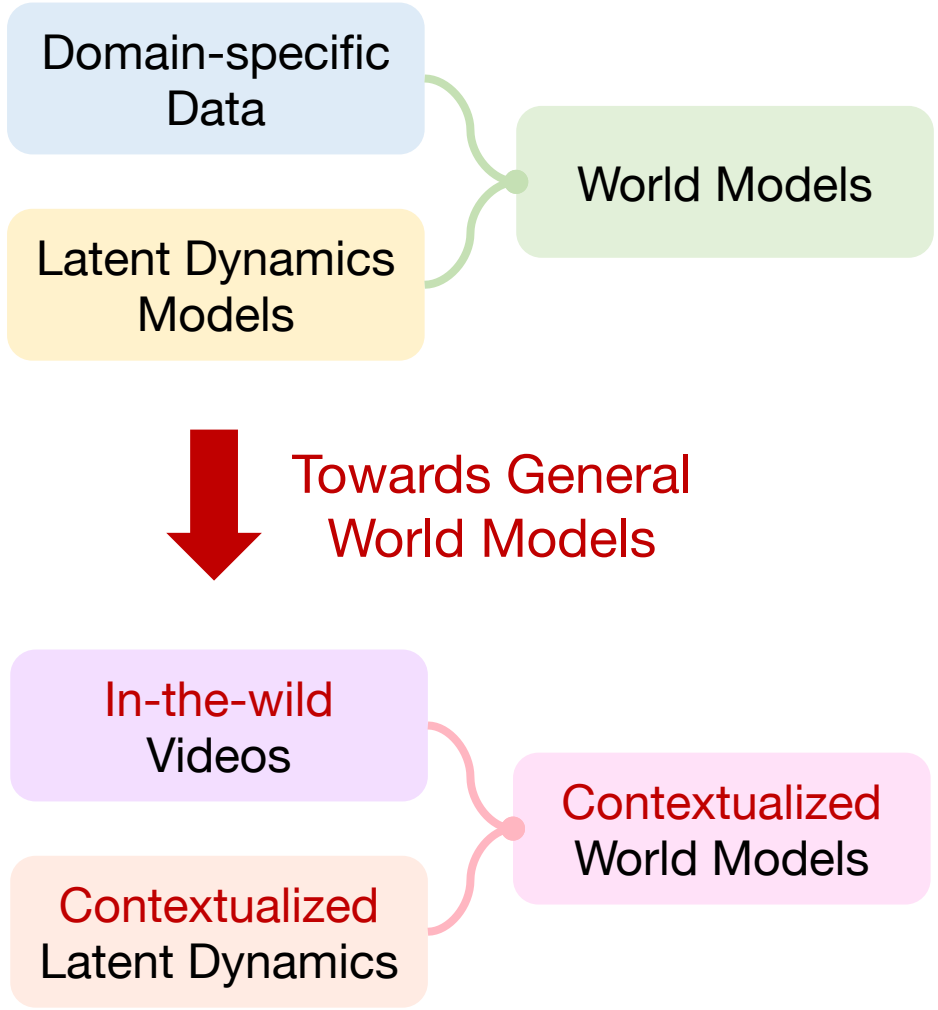
Qualitative Evaluation: Compositional Decoding



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

Summary

- A world model is a internal models of how the world works, which empowers the agent to reason and plan.
- Our work:
  - 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



Thank You!

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

Contact: wujialong0229@gmail.com

Machine Learning Group, School of Software, Tsinghua University

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

清華大學
Tsinghua University