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

Jialong Wu*, Haoyu Ma*, Chaoyi Deng, Mingsheng Long

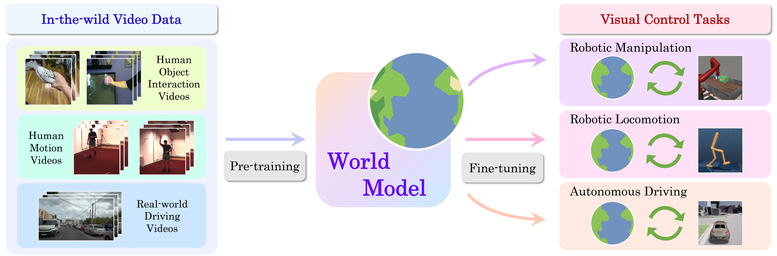
NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

- **World models**: Internal models of how the world works
- Previous work on **Model-based RL** (MBRL):
  - Learn domain-specific world models from scratch
  - Pre-train world models with simulated video data
- **In-the-wild videos**: Promising sources of general world knowledge

Research Problem

Can world models pre-trained on diverse in-the-wild videos benefit sample-efficient learning of downstream visual control tasks?



Challenges towards a General World Model

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

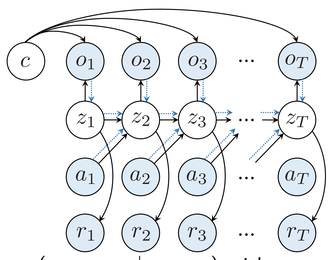
- **Contributions**:
  - A paradigm of pre-training world models with in-the-wild videos
  - Contextualized World Models, which explicitly separates contexts
  - Significant improvement of sample efficiency on various domains

Contextualized Latent Dynamics Models

Our Insight

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

- **Contextualized generative models**: utilizing rich contextual information c beyond the expressiveness of latent dynamics variables z_t
- **Latent dynamics inference**: concentrates on essential temporal variations



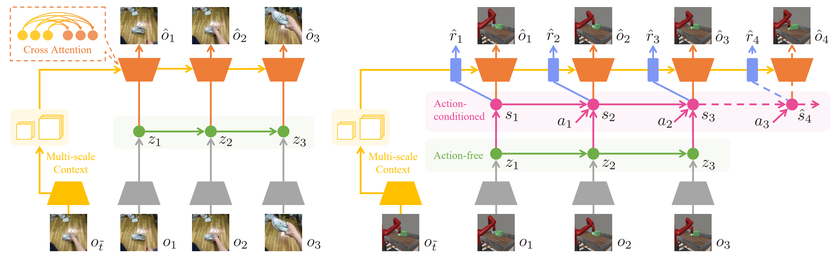
- **Optimization with ELBO** of conditional $\ln p_\theta\left(o_{1:T}, r_{1:T} \mid a_{1:T}, c\right)$ without the need to model the context distribution $p(c)$:

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

Contextualized World Model Architectures

Overview of Architecture

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



- **Context formulation**: a random single frame from the trajectory segment is chosen as the context c. By random selection, the context encoder learns to be robust to temporal variations.

$$c \doteq o_{\tilde{t}},\ \tilde{t} \sim \text{Uniform}\{1, 2, \ldots, T\}$$

- **Muti-scale cross-attention conditioning**: ContextWM adopts U-Net-style multi-scale feature shortcuts. Instead of naive concatenation forcing a spatial alignment, an adaptive cross-attention mechanism is utilized.

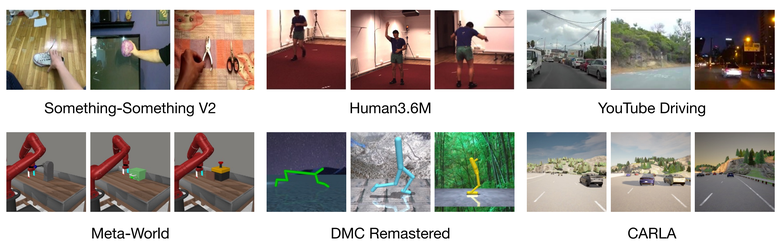
$$\begin{aligned}\text{Flatten: } & Q = \text{Reshape}(X),\ K = V = \text{Reshape}(Z) \in \mathbb{R}^{hw \times c} \\ \text{Cross-Attention: } & R = \text{Attention}\left(QW^Q, KW^K, VW^V\right) \in \mathbb{R}^{hw \times c} \\ \text{Res-Connection: } & X = \text{ReLU}\left(X + \text{BatchNorm}\left(\text{Reshape}(R)\right)\right) \in \mathbb{R}^{c \times h \times w}.\end{aligned}$$

- **Dual reward predictors**:
  - Behavioral reward predictor predicts exploratory reward $r_t + \lambda r_t^{int}$ for behavior learning.
  - Representative reward predictor predicts pure task reward r_t for enhancing task-relevant representation learning.
- **Overall objective**:

$$\mathcal{L}^{\text{CWM}}(\phi, \varphi, \theta) \doteq \underbrace{\mathbb{E}_{q_\phi(s_{1:T}|a_{1:T},z_{1:T}),q_\theta(z_{1:T} \mid o_{1:T})}}_{\text{context-unaware latent inference}} \Bigg[\sum_{t=1}^{T} \Bigg(\underbrace{-\ln p_\theta(o_t|s_t, c)}_{\text{contextualized image loss}} \underbrace{-\ln p_\phi(r_t + \lambda r_t^{\text{int}}|s_t)}_{\text{behavioral reward loss}} \underbrace{-\beta_r \ln p_\varphi(r_t|s_t)}_{\text{representative reward loss}} \underbrace{+\beta_z \operatorname{KL}\left[q_\theta(z_t|z_{t-1}, o_t) \,\|\, p_\theta(\hat{z}_t|z_{t-1})\right]}_{\text{action-free KL loss}} \underbrace{+\beta_s \operatorname{KL}\left[q_\phi(s_t|s_{t-1}, a_{t-1}, z_t) \,\|\, p_\phi(\hat{s}_t|s_{t-1}, a_{t-1})\right]}_{\text{action-conditional KL loss}} \Bigg)\Bigg].$$

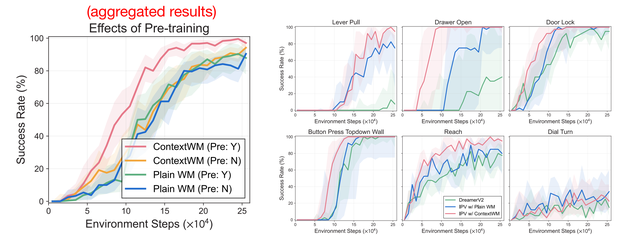
Experimental Setup

- Diverse pre-training datasets & visual control tasks

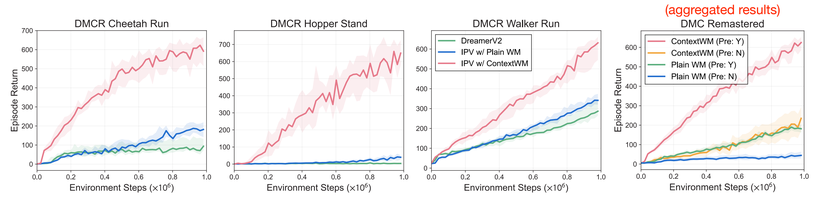


Main Results

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

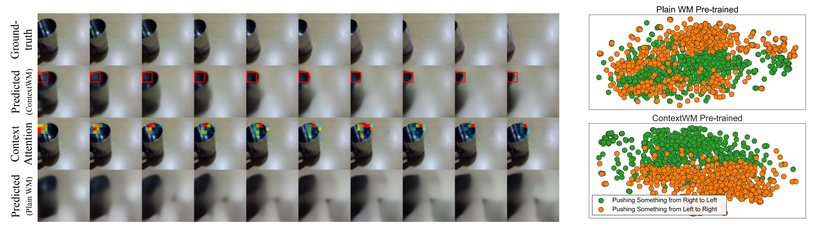


- **DMC Remastered**: ContextWM further unleashes performance boost on visual generalization benchmark with pre-training from in-the-wild videos.

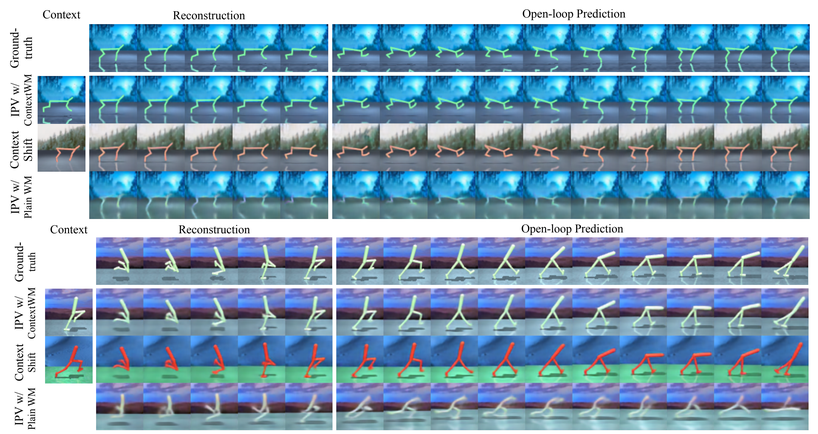


Qualitative Analysis

- **Video prediction and representations**: ContextWM effectively captures the shape and motion of the object, while the plain WM fails.



- **Compositional decoding**: ContextWM successfully learned disentangled representations of contexts and dynamics, while the plain WM suffers from learning entangled representations and thus makes poor predictions.



School of Software - BNRist - Tsinghua University - China
Code: https://github.com/thuml/ContextWM
Mail: wujialong0229@gmail.com