NEURAL INFORMATION PROCESSING SYSTEMS

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

东方理工高等研究院 EASTERN INSTITUTE FOR ADVANCED STUDY

東南大學 SOUTHEAST UNIVERSITY

# Making Offline RL Online: Collaborative World Models for Offline Visual Reinforcement Learning

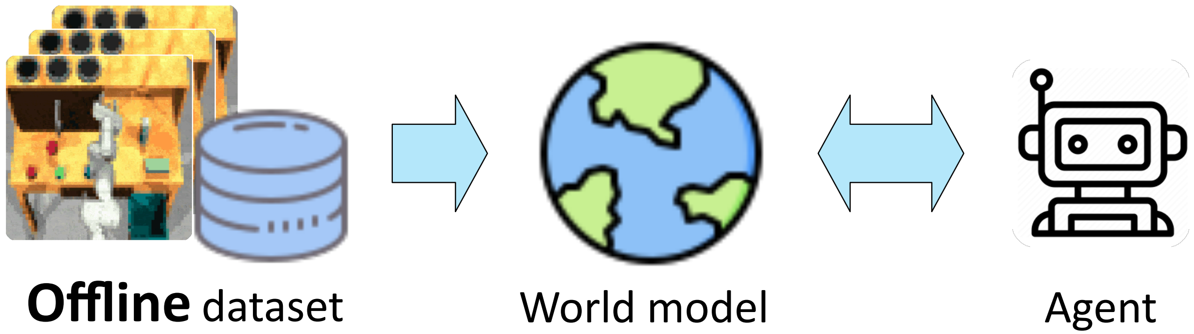

**Qi Wang*** **Junming Yang*** **Yunbo Wang**

**Xin Jin** **Wenjun Zeng** **Xiaokang Yang**

* Equal contribution

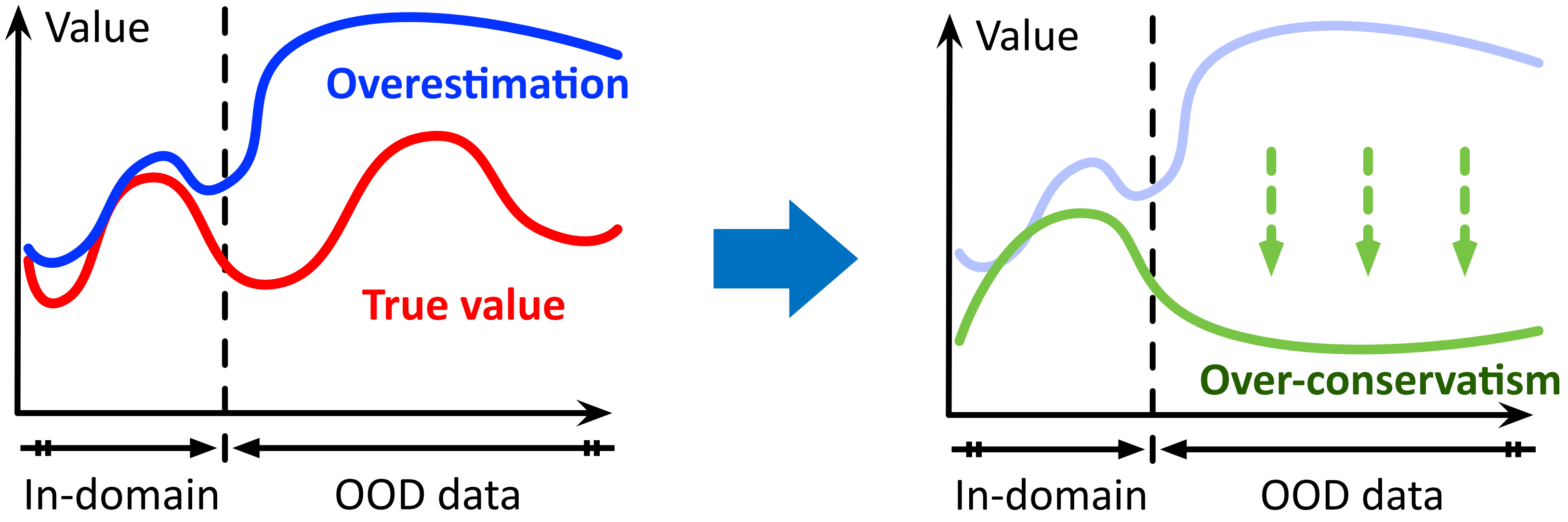
Correspondence to: Yunbo Wang

# Model-based RL for offline visual control



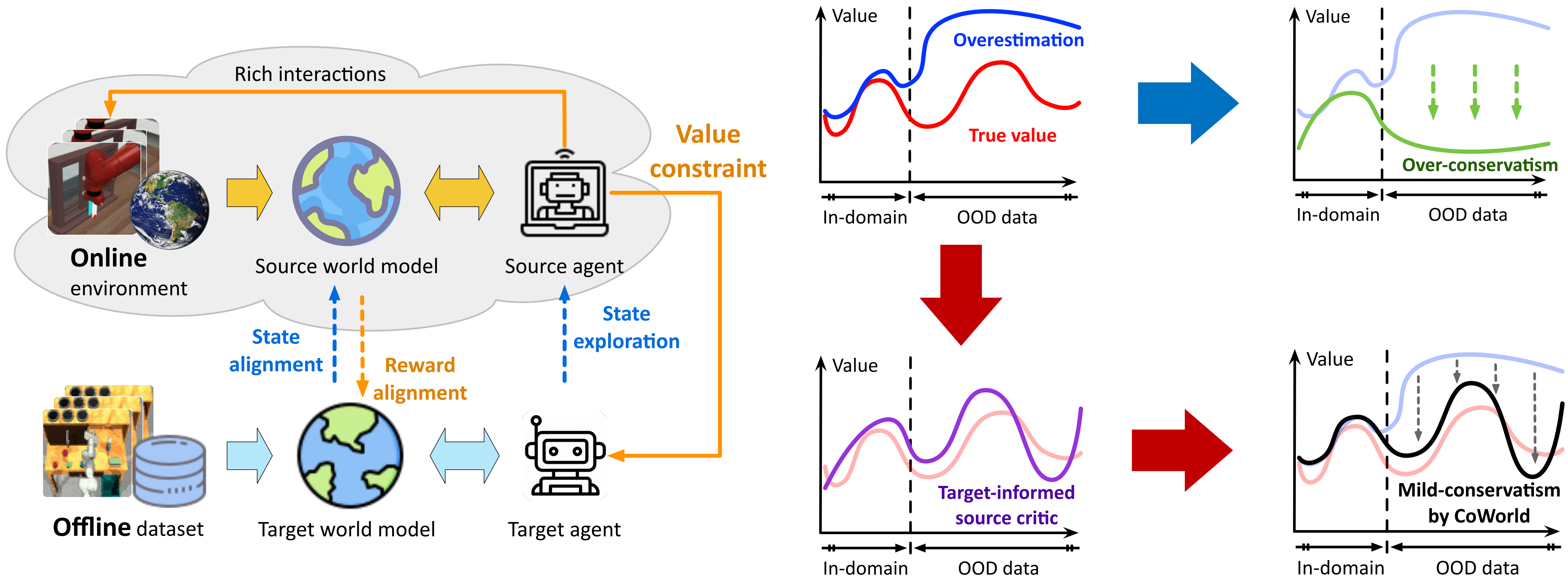

- Offline visual RL is a promising approach to learn an efficient control policy from visual observations, avoiding the need for high interaction costs with the physical world
- The benefits of using a world model for Offline RL are that the agent interacts with the model rather than directly with the dataset
- However, this approach cannot entirely solve the overestimation issue, as the world model may overfit the limited dataset, thereby introducing bias

# How to tackle value overestimation?




- Typical offline RL methods often penalize estimated values beyond the offline data distribution, leading to value over-conservatism
- This penalization can suppress the agent's exploration in the world model --- Exploration that may sometimes be valuable and at other times should indeed be suppressed
- How can we differentiate between the two? Address each case accordingly?

# A New Thought: Online Simulator as a Behavior "Test Bed"

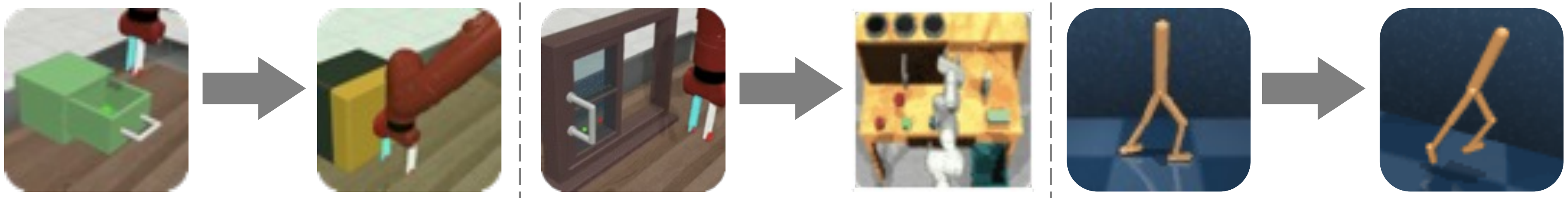



- CoWorld solves offline visual RL as an offline-online-offline transfer learning problem
- CoWorld leverages a target-informed source critic to provide mild constraints for target value estimation, without impeding state exploration with potential advantages

# Experimental Setups

**Meta-World** | **Meta-World → RoboDesk** | **DMC**

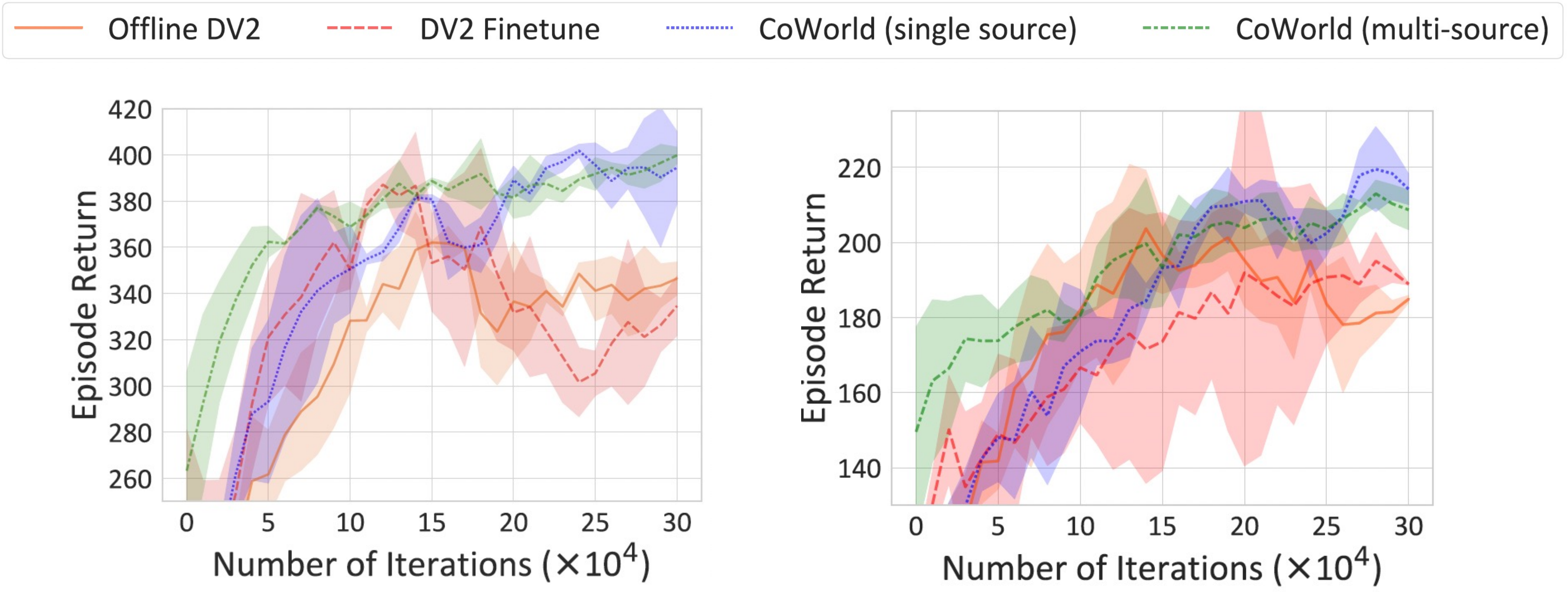
Drawer Close, Button Press | Window Close, Open Slide | Walker Walk, Walker Downhill



| | Source: *Meta-World* | Target: *RoboDesk* | Similarity / Difference |
|---|---|---|---|
| Task | Window Close | Open Slide | Related manipulation tasks |
| Dynamics | Simulated Sawyer robot arm | Simulated Franka Emika Panda robot arm | Different |
| Action space | Box(-1, 1, (4,), float64) | Box(-1, 1, (5,), float32) | Different |
| Reward scale | [0, 1] | [0, 10] | Different |
| Observation | Right-view images | Top-view images | Different view points |

- **Setup 1: Cross-Task** experiments on *Meta-World*
- **Setup 2: Cross-Environments** experiments from *Meta-World* to *RoboDesk*
- **Setup 3: Cross-Dynamics** experiments on *DeepMind Control Suite* (DMC)

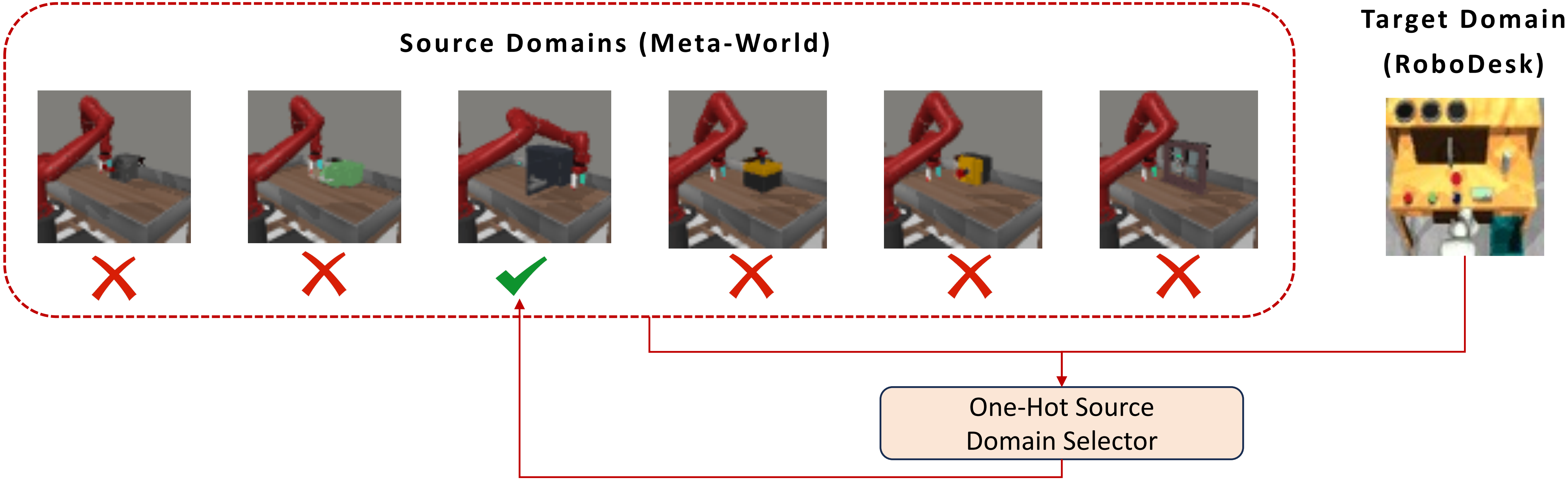
# Results: Meta-World→RoboDesk




*Button Press → Push Button* *Window Close → Open Slide*

- CoWorld outperforms Offline DV2 and DV2 Finetune by large margins
- Directly fine-tuning the source world model in this cross-environment setup, does not result in significant improvements in the final performance

# Results: Multi-Source CoWorld




- When there are notable distinctions between the source domain and target domain, multi-source CoWorld can adaptively selects a useful source task

# Results: Meta-World

| Model | BP→ DC* | DC → BP | BT→ WC | BP→ HP | WC→ DC | HP→ BT | Avg. |
|---|---|---|---|---|---|---|---|
| Offline DV2 | 2143±579 | 3142±533 | 3921±752 | 278±128 | 3899±679 | 3002±346 | 2730 |
| LOMPO | 2883±183 | 446±458 | 2983±569 | 2230±223 | 2756±331 | 1961±287 | 1712 |
| CoWorld | **3967±312** | **3623±543** | **4521±367** | **4570±677** | **4845±14** | **3889±159** | **4241** |

Offline DV2

- **Steps: 79**
- **Return： 3002**

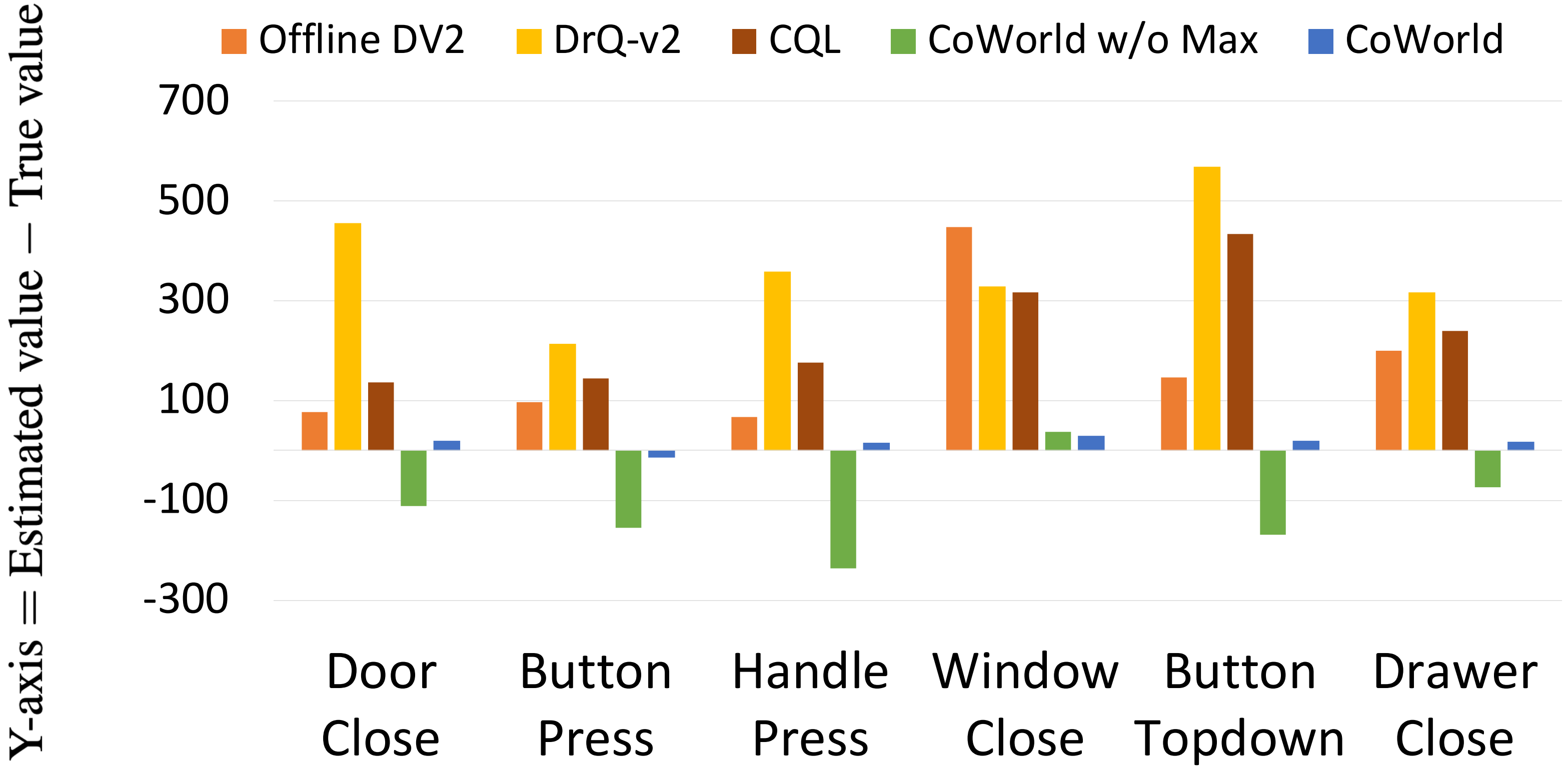
LOMPO

- **Unfinished**
- **Return： 1961**

CoWorld

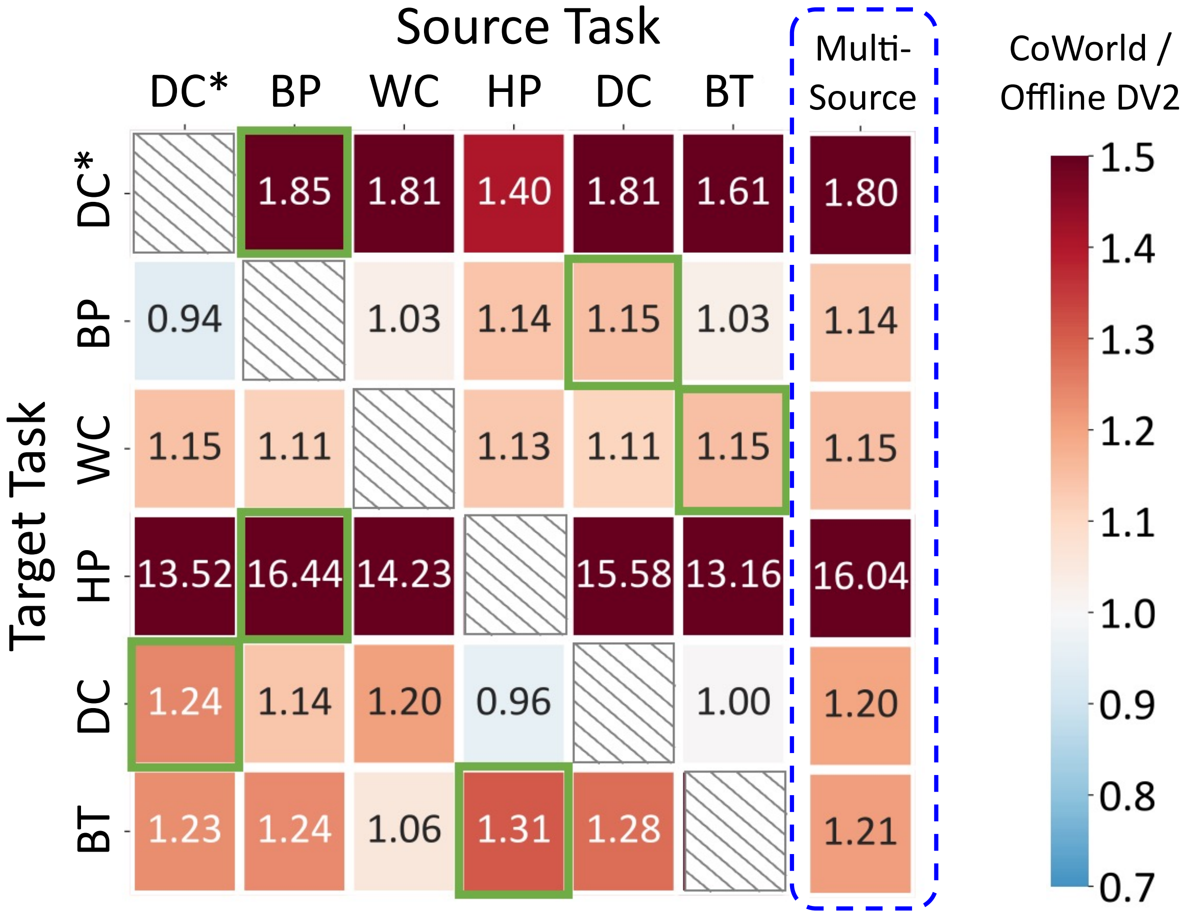
- **Steps： 59**
- **Return： 3889**

# Results: Meta-World




- Existing approaches often overestimate the value functions in the offline setup
- The values estimated by CoWorld are notably more accurate and more akin to the true values

# Results: Meta-World




- The value in each grid cell signifies the ratios of returns achieved by CoWorld compared to those achieved by the Offline DV2
- Cells highlighted with a green box represent the best-source tasks for transfer
- Notably, there are challenging cases with weakly related source and target tasks. In the majority of cases (26 out of 30), CoWorld outperforms Offline DV2

# Results: DeepMind Control Suite

| Model | WW → WD | WW → WU | WW → WN | CR → CD | CR → CU | CR → CN | Avg. |
|---|---|---|---|---|---|---|---|
| Offline DV2 | 435±22 | 139±4 | 214±4 | 243±7 | 3±1 | 51±4 | 181 |
| LOMPO | 462 ± 87 | 260±21 | **460±9** | 395±52 | 46±19 | 120±4 | 291 |
| CoWorld | **629±9** | **407±141** | 426±32 | **745±28** | **225±20** | **493±10** | **488** |

Offline DV2

LOMPO

CoWorld

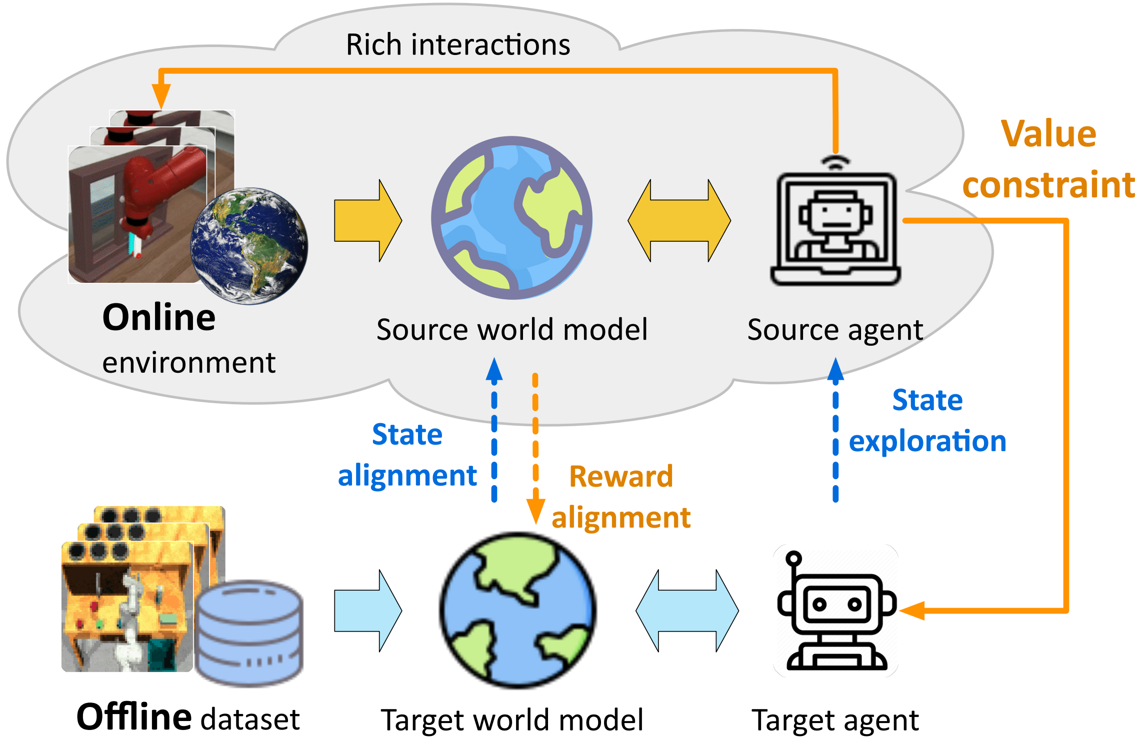
# Method

Step 1: Offline-to-Online State Alignment

Step 2: Online-to-Offline Reward Alignment

Step 3: Min-Max Value Constraint

*Please see our paper to find the technical details*



# Thanks!

https://qiwang067.github.io/coworld