ICML
International Conference
On Machine Learning

Multi-View Masked World Models for Visual Robotic Manipulation

Younggyo Seo*, Junsu Kim*, Stephen James, Kimin Lee,
Jinwoo Shin, Pieter Abbeel

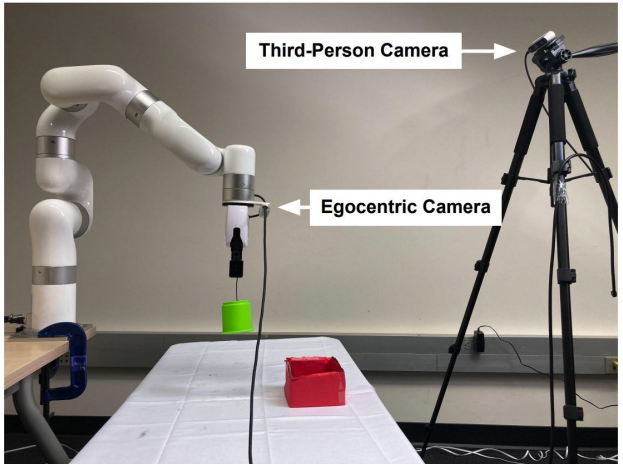
KAIST dyson Google BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

*Equal Contribution

Extension to Multi-View Inputs

- Multiple cameras have often been used for visual robotic manipulation

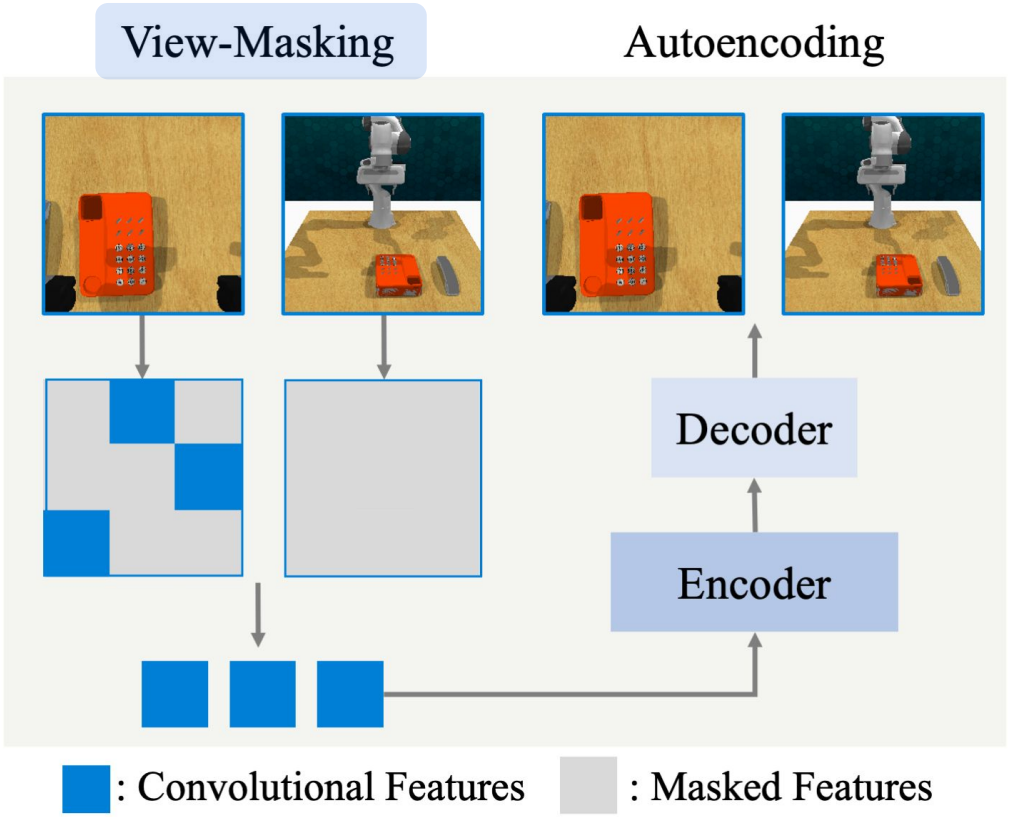
[Akkaya et al., 2019]



[Jangir et al., 2022]

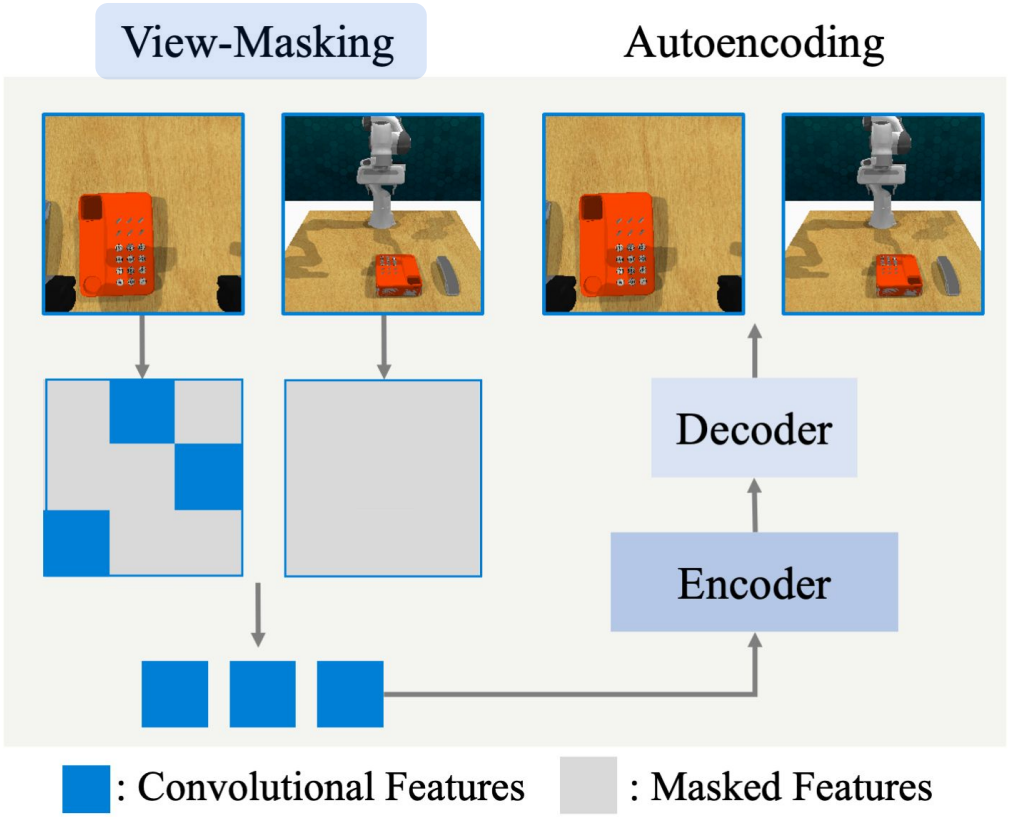
Akkaya, Ilge, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino et al. "Solving rubik's cube with a robot hand." *arXiv preprint arXiv:1910.07113* (2019).
Jangir, Rishabh, Nicklas Hansen, Sambaran Ghosal, Mohit Jain, and Xiaolong Wang. "Look Closer: Bridging Egocentric and Third-Person Views With Transformers for Robotic Manipulation." *IEEE Robotics and Automation Letters* 7, no. 2 (2022): 3046-3053.

Multi-View Masked Autoencoder (MV-MAE)



Main Idea:
Reconstruct masked viewpoints to learn cross-view information

Multi-View Masked Autoencoder (MV-MAE)

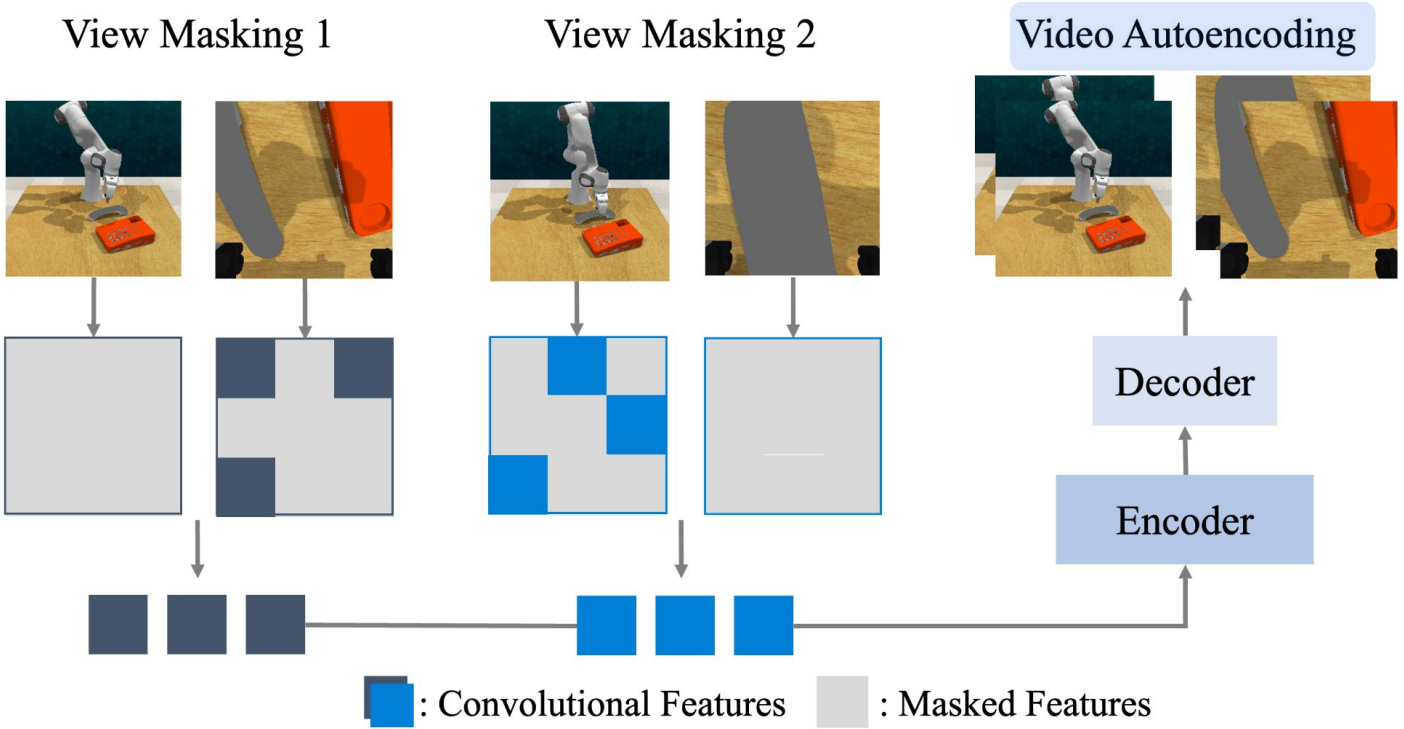


Main Idea:
Reconstruct masked viewpoints to learn cross-view information

Challenge:
Objective might be too difficult for the model

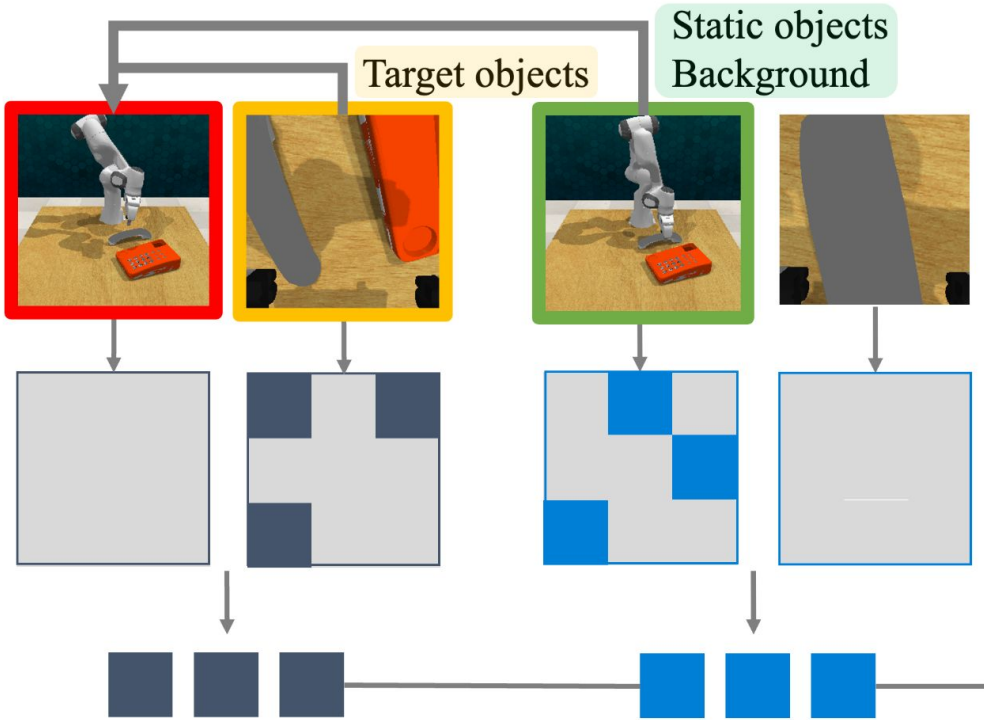
Multi-View Masked Autoencoders (MV-MAE)

- **Masked view reconstruction** with View-Masking and **Video Autoencoding**



Multi-View Masked Autoencoders (MV-MAE)

- **Masked view reconstruction** with View-Masking and **Video Autoencoding**

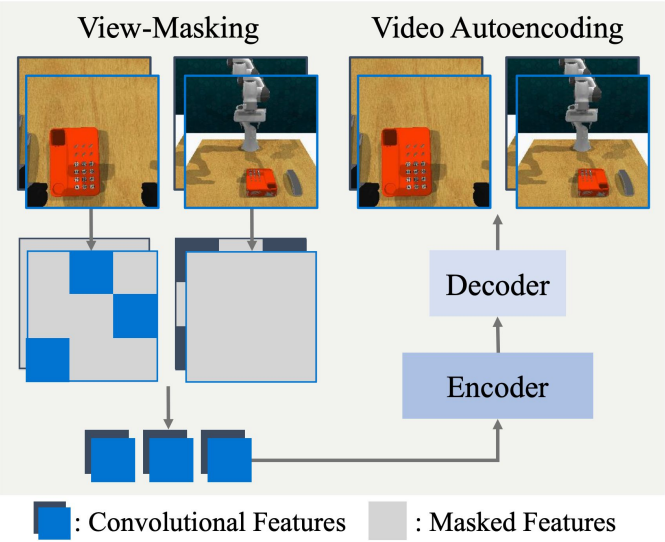


Why? 🤔

- Makes it easy to reconstruct masked viewpoints

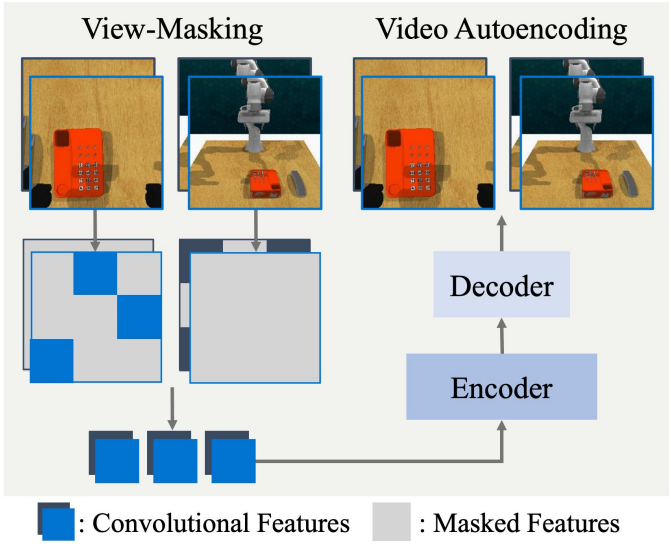
Multi-View Masked World Models (MV-MWM)

MV-MAE can extract both ***multi-view*** and ***single-view*** representations

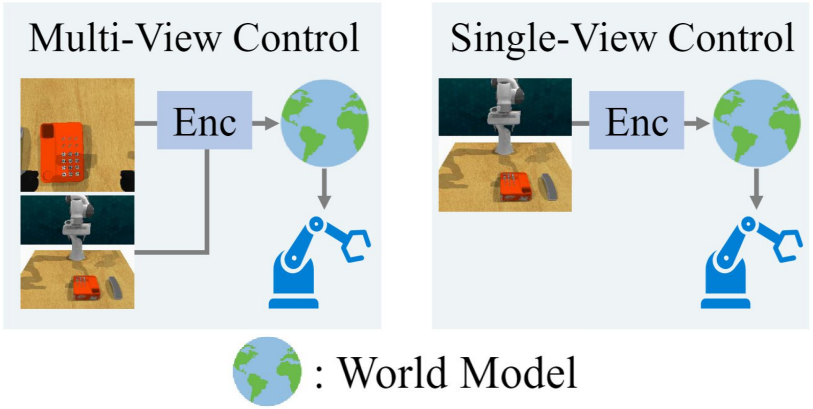


Multi-View Masked World Models (MV-MWM)

MV-MAE can extract both ***multi-view*** and ***single-view*** representations

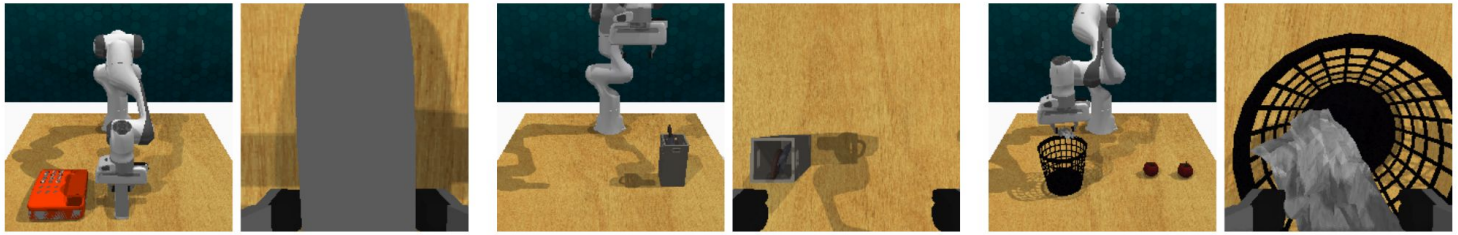


Visual robotic manipulation with **multi-view** or **single-view** data



Experiments 1: Setup

- RLBench [James et al., 2020] with **front** and **wrist** cameras
  - Widely-used camera configuration



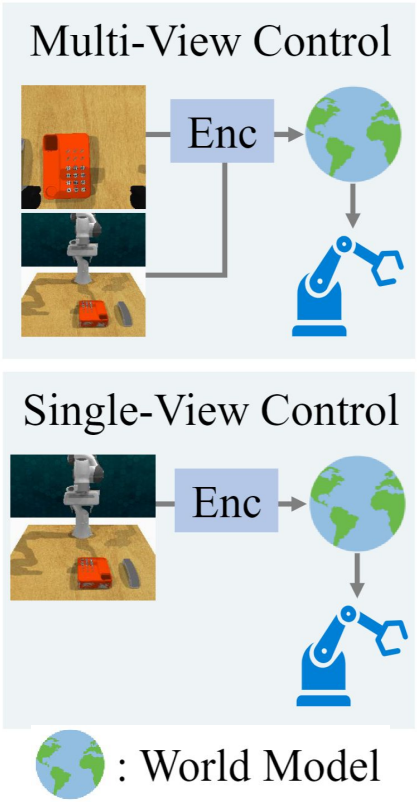
(a) Phone On Base (b) Take Umbrella Out of Stand (c) Put Rubbish in Bin

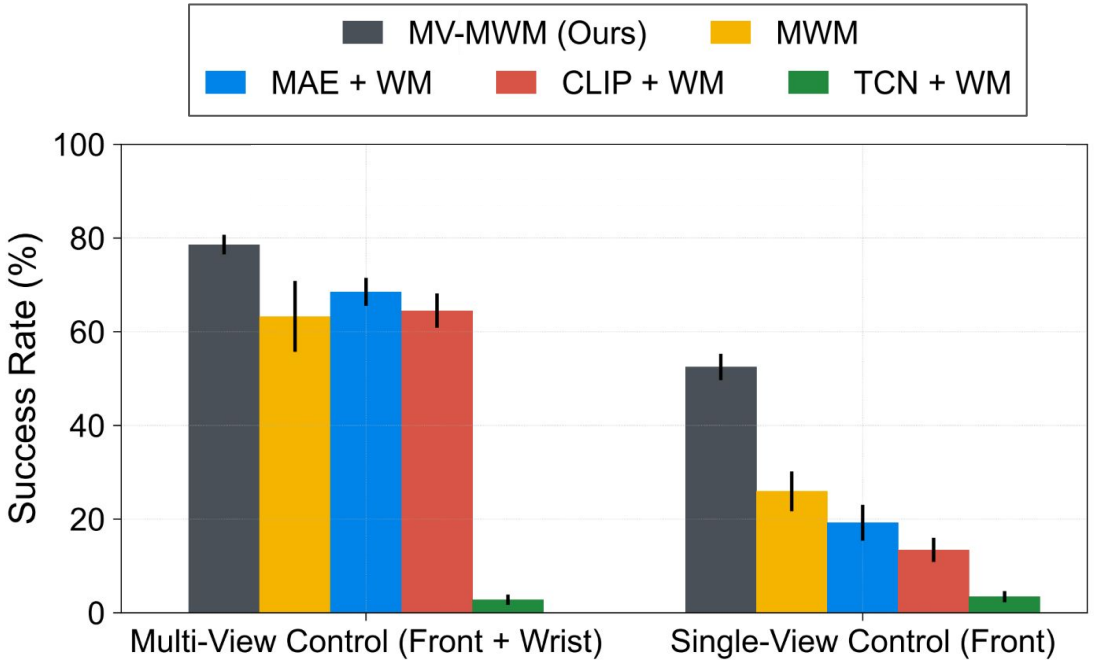
(d) Pick Up Cup (e) Stack Wine

James, Stephen, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. "RLbench: The robot learning benchmark & learning environment." *IEEE Robotics and Automation Letters* 5, no. 2 (2020): 3019-3026.

Experiments 1: Multi-View and Single-View Control

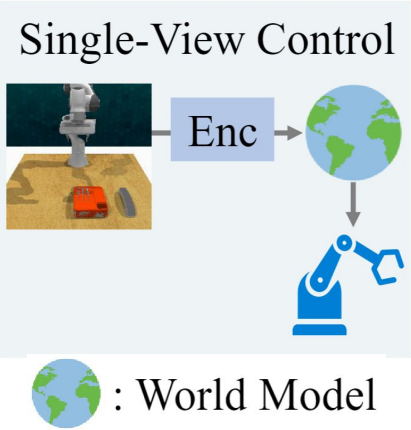
- MV-MWM outperforms both single-view and multi-view baselines





Experiments 2: Imitation Learning

- MV-MWM is also outperforming baselines in imitation learning setup



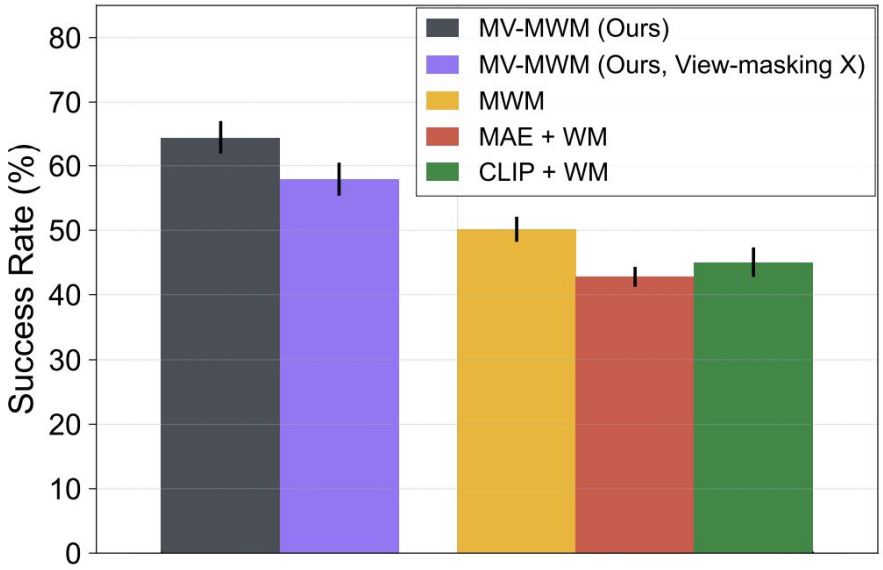
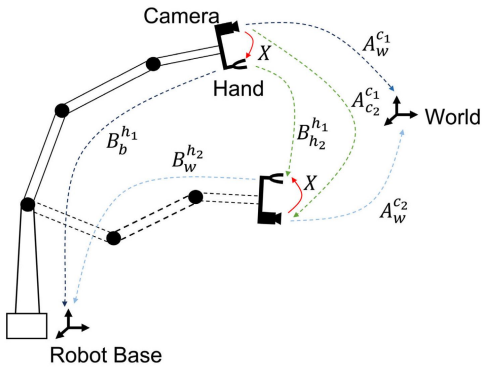


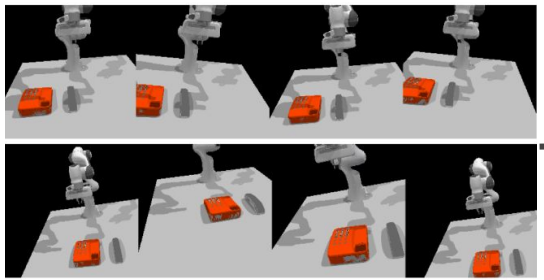
Figure 8. Aggregate success rate of imitation learning agents on five single-view control tasks. The result shows the mean and stratified bootstrap confidence interval across 20 runs.

Experiments 3: Setup



Motivation:
Camera calibration is a tedious procedure

- **Solution:** Training a **viewpoint-robust** policy with viewpoint randomization



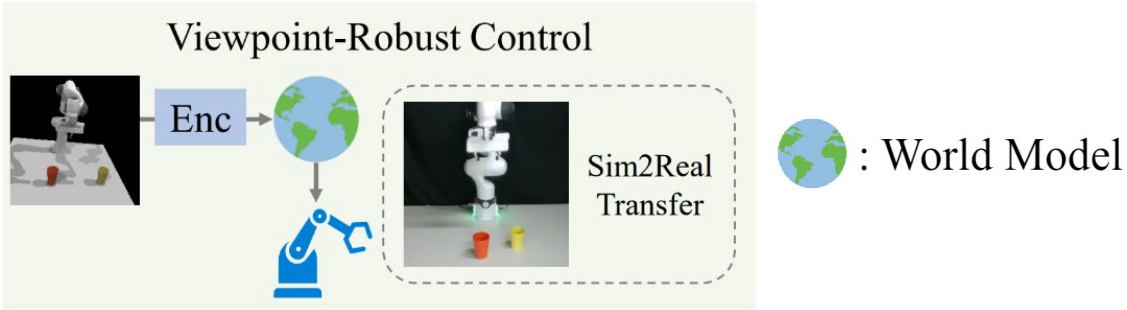
Viewpoint randomization

: World Model

Experiments 3: Viewpoint-Robust Control

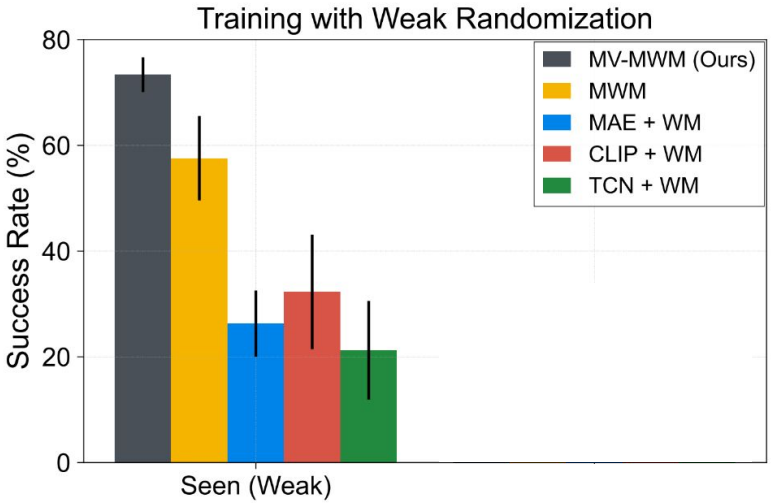
- **Step 1:** Multi-view representation learning with viewpoint randomization

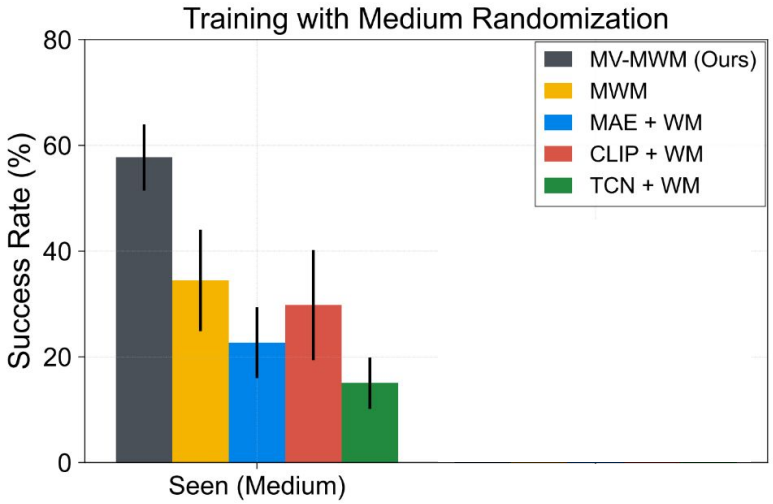
- **Step 2:** Learn a world model for viewpoint-robust control

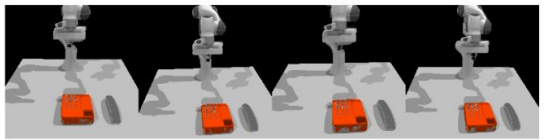


Experiments 3: Viewpoint-Robust Control

- **MV-MWM** learns a policy with aggressive viewpoint randomization





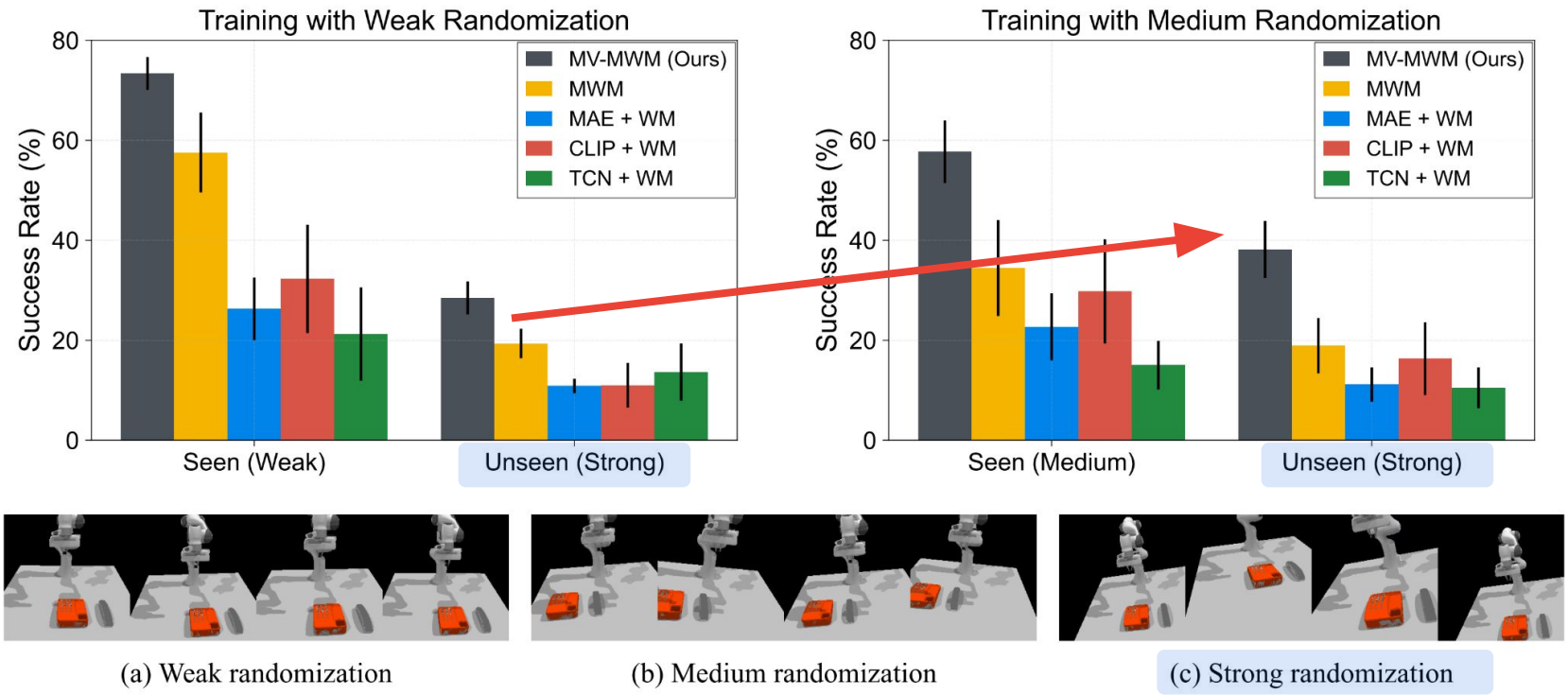


(a) Weak randomization

(b) Medium randomization

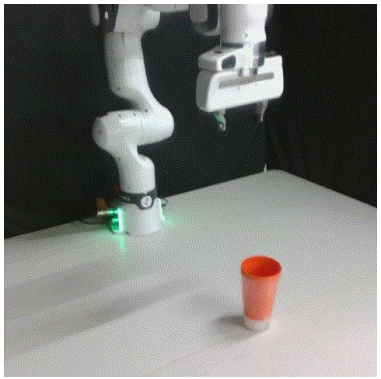
Experiments 3: Viewpoint-Robust Control

- **MV-MWM** learns to solve tasks under unseen viewpoints

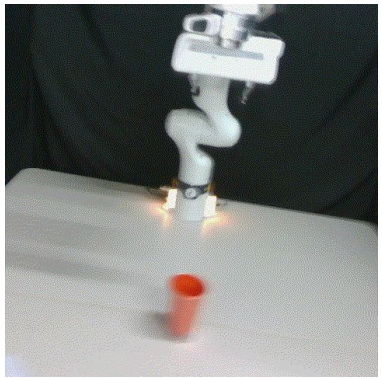


Experiments 3: Viewpoint-Robust Control

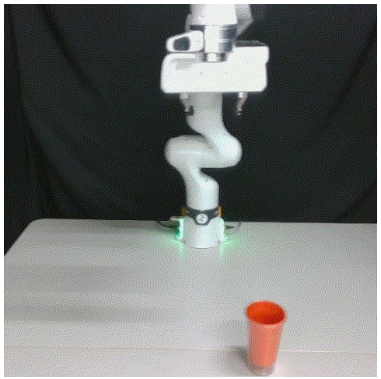
- **Zero-Shot Sim2Real Transfer with Hand-held Cameras**
  - Without proprioceptive states, depth, and adaptation



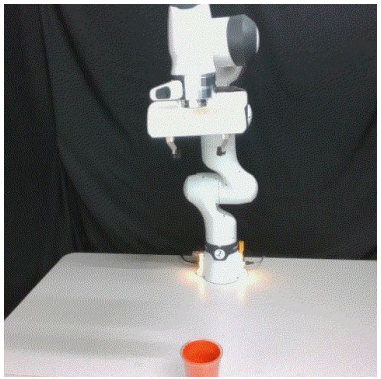
Rotation



Shake



Translation



Zoom

Contributions

- We present Multi-View Masked World Models, which can utilize multi-view data for both representation and dynamics learning
- Representation learning with multi-view data is helpful for both multi-view and single-view control setups
- Sim2Real viewpoint-robust control demonstrations