Yanjie Ze

yanjieze.com

Shanghai Jiao Tong Uni.

Visual Representations for Generalizable Robotic Manipulation

Oct 27th, 2023

[1] Hansen*, Yuan*, **Ze*** et al., *"On Pre-Training for Visuo-Motor Control: Revisiting a Learning-from-Scratch Baseline"*, **ICML** 2023.

[2] **Ze** et al., *"Visual Reinforcement Learning with Self-Supervised 3D Representations"*, **RA-L** 2023 & **IROS** 2023.

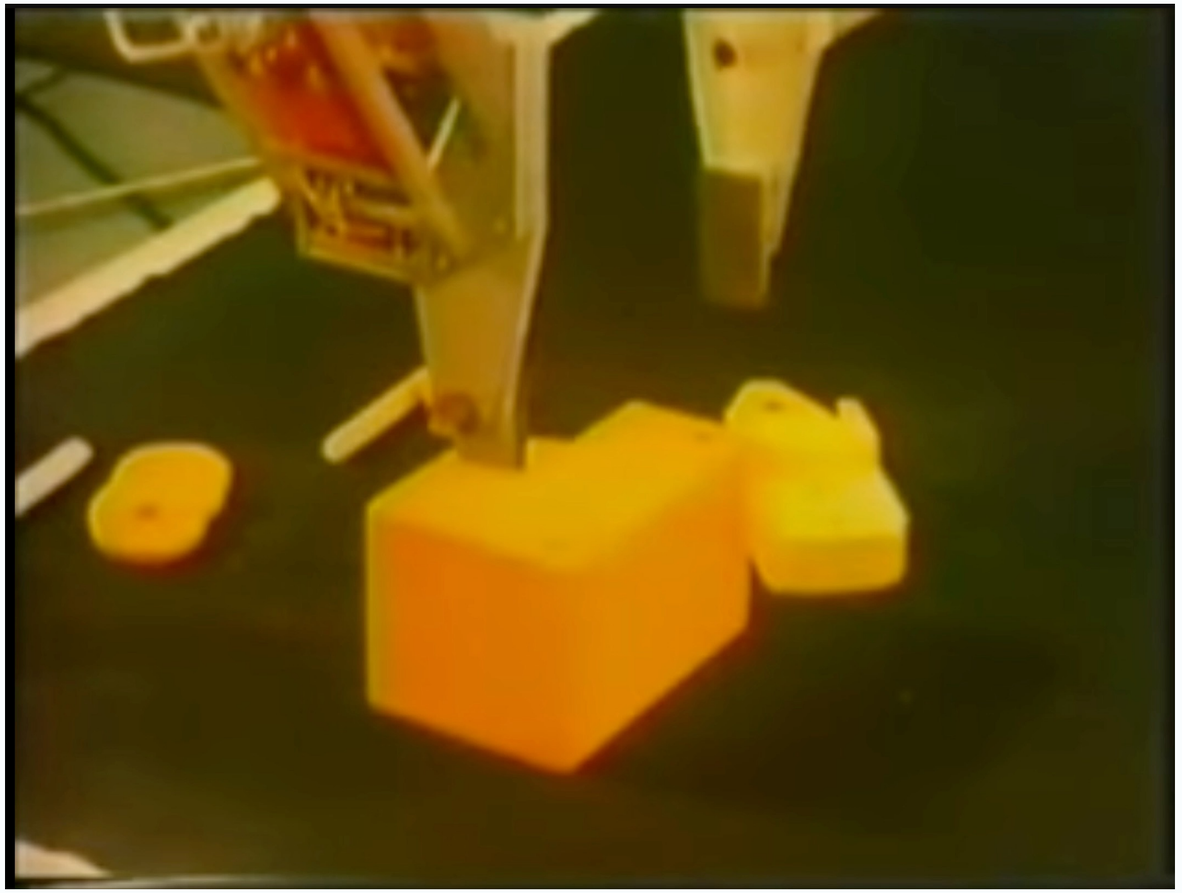
[3] **Ze** et al., *"GNFactor: Multi-Task Real Robot Learning with Generalizable Neural Feature Fields"*, **CoRL** 2023 **Oral.**

[4] **Ze** et al., *"H-InDex: Visual Reinforcement Learning with Hand-Informed Representations for Dexterous Manipulation"*, **NeurIPS** 2023.

[5] Yang*, **Ze*** et al., *"MoVie: Visual Model-Based Policy Adaptation for View Generalization",* **NeurIPS** 2023.

Source: https://youtu.be/03p2CADwGF8?si=AILVs_aCg9jWYkQ4

Generalization in Robotics



NOT JUST:

1 task

1 scene

1 object

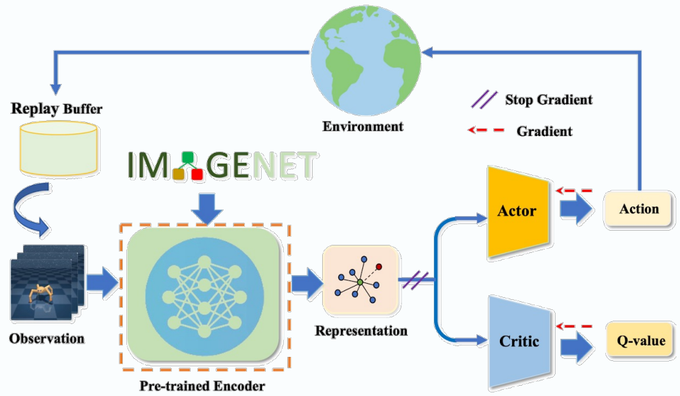
The 1973 Lighthill debate on Artificial Intelligence:
"The general purpose robot is a **mirage**"

Generalization in Computer Vision

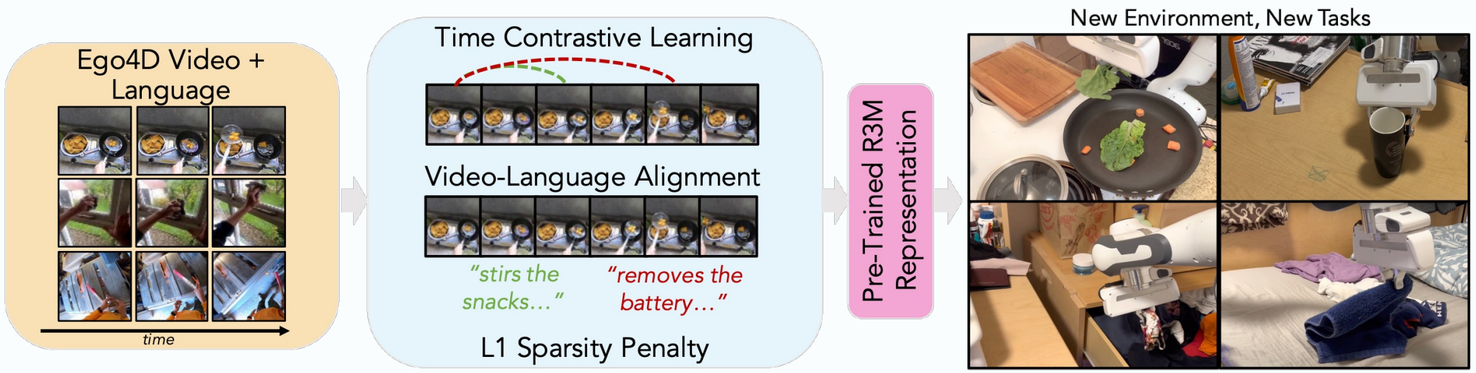
Segment Anything Model
Meta, 2023

DALL·E 3
OpenAI, 2023

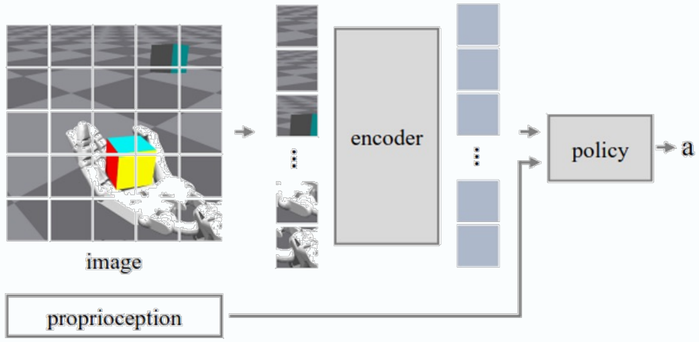
2D Visual Representations for Robotics



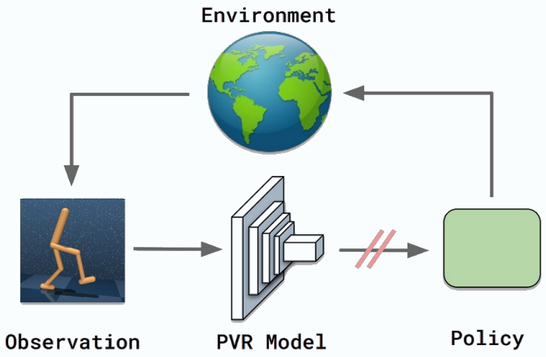
Yuan et al., 2022



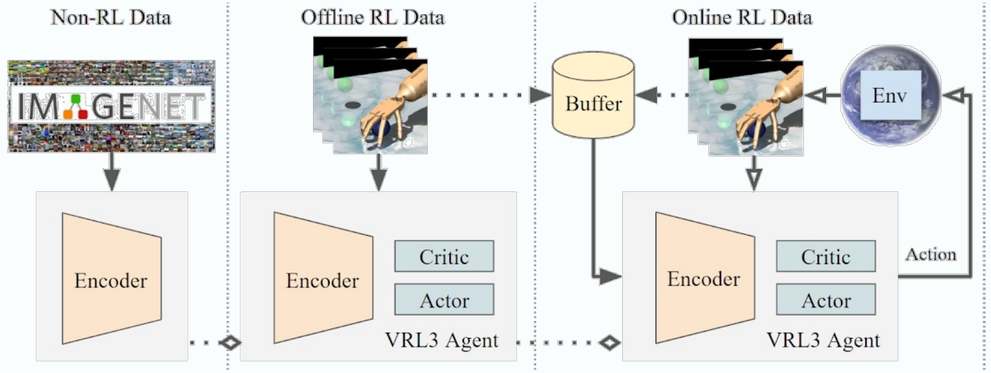
Nair et al., 2022



Xiao et al., 2022

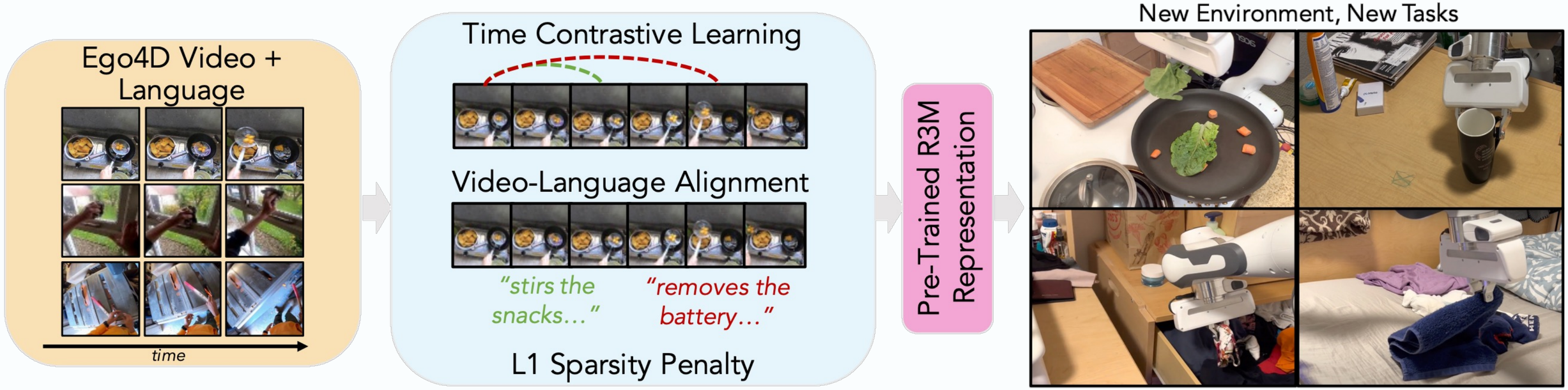


Parisi et al., 2022



Wang et al., 2022

2D Visual Representations for Robotics

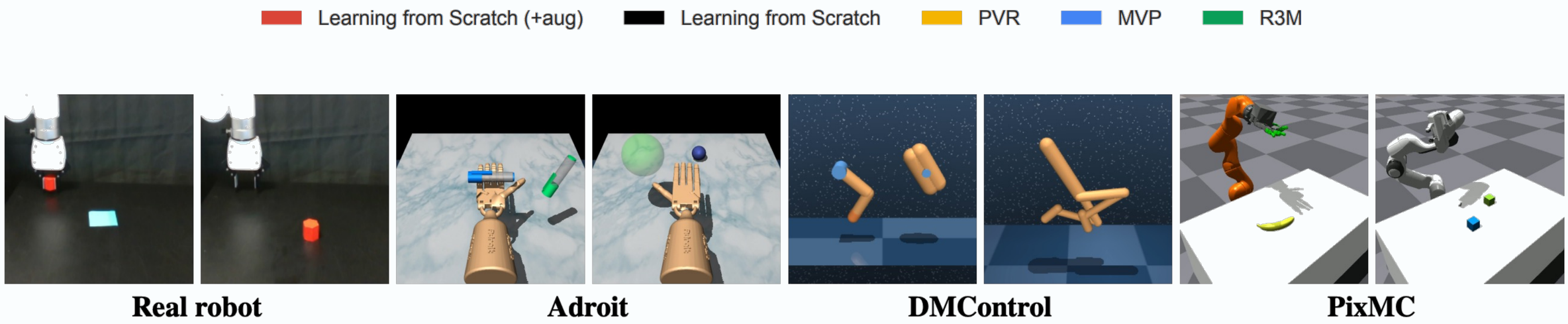


Nair et al., 2022

- How much are they helping manipulation?
- Actually, **not too much**.

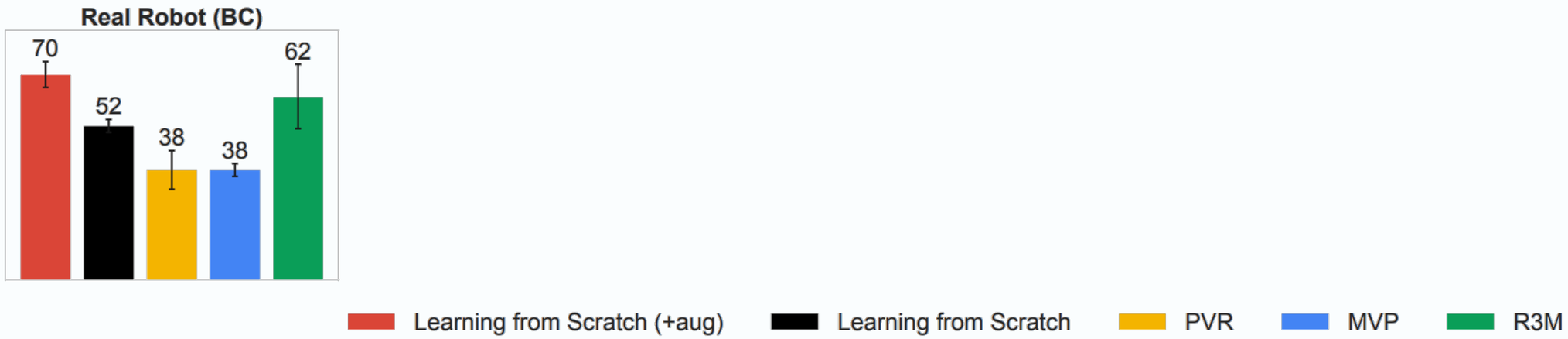
Revisiting Visual Representations for Visuomotor Control

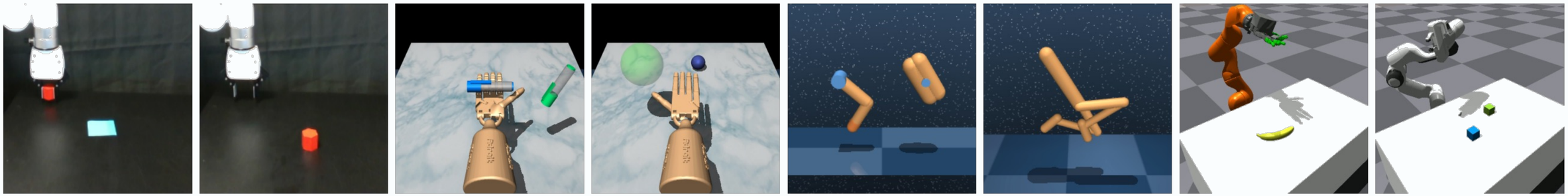
[1] Hansen*, Yuan*, **Ze*** et al., On Pre-Training for Visuo-Motor Control: Revisiting a Learning-from-Scratch Baseline, ICML 2023.



Revisiting Visual Representations for Visuomotor Control

[1] Hansen*, Yuan*, **Ze*** et al., On Pre-Training for Visuo-Motor Control: Revisiting a Learning-from-Scratch Baseline, ICML 2023.

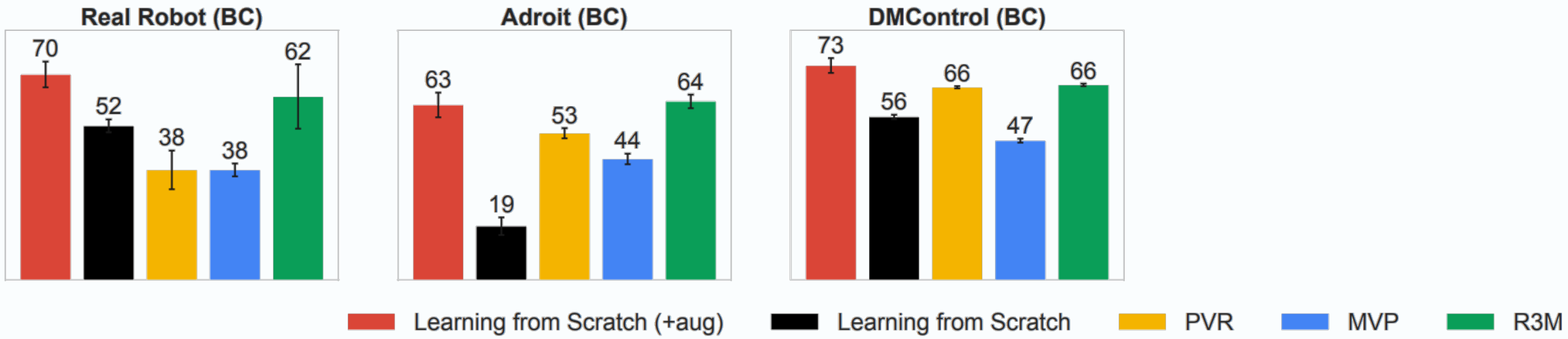


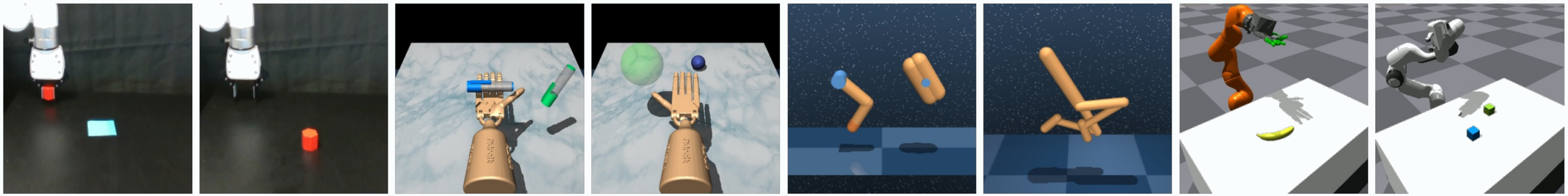


Real robot | Adroit | DMControl | PixMC

Revisiting Visual Representations for Visuomotor Control

[1] Hansen*, Yuan*, **Ze*** et al., On Pre-Training for Visuo-Motor Control: Revisiting a Learning-from-Scratch Baseline, ICML 2023.

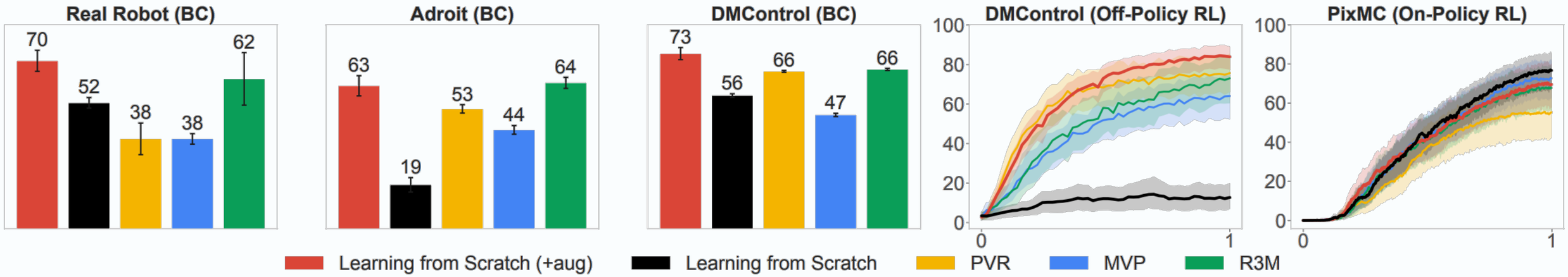


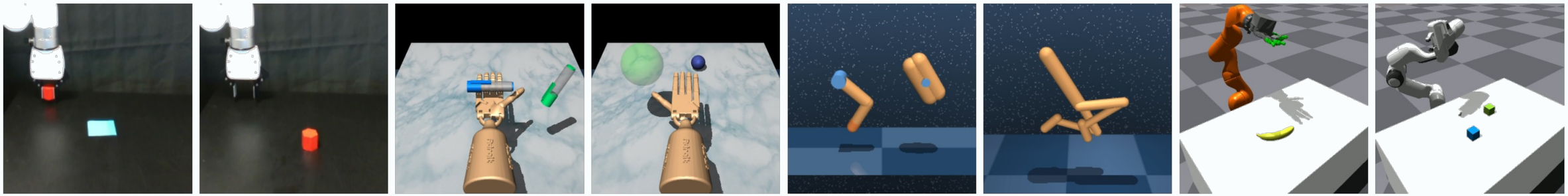


Real robot | Adroit | DMControl | PixMC

Revisiting Visual Representations for Visuomotor Control

[1] Hansen*, Yuan*, **Ze*** et al., On Pre-Training for Visuo-Motor Control: Revisiting a Learning-from-Scratch Baseline, ICML 2023.



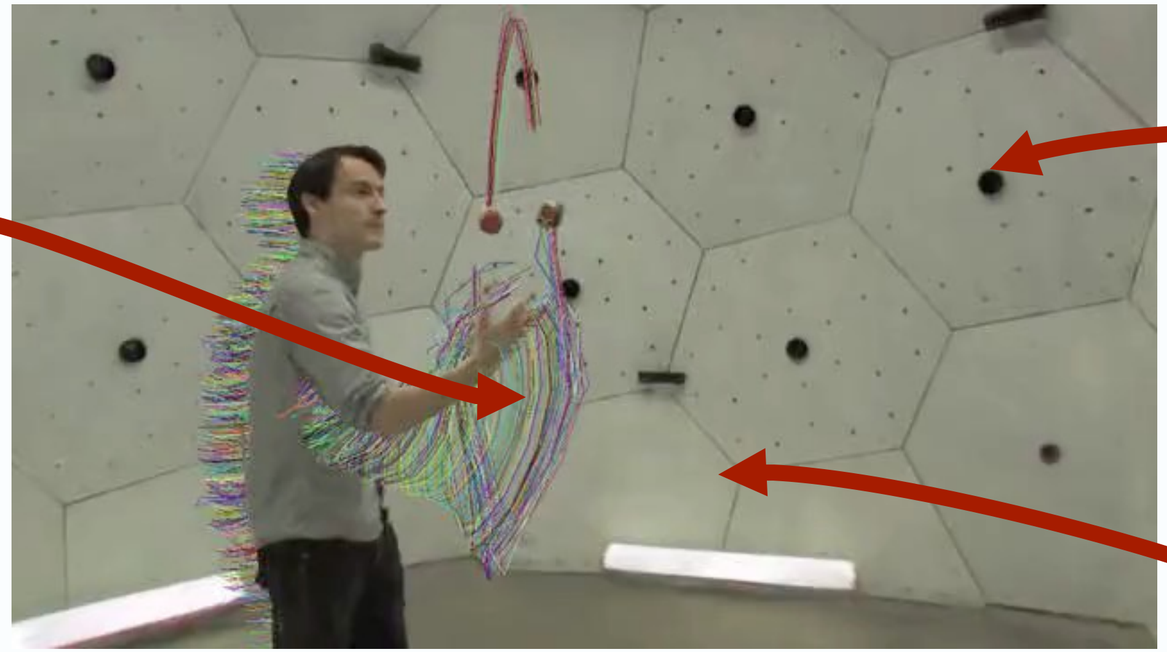


Real robot **Adroit** **DMControl** **PixMC**

What is the Missing Point?

The world itself contains rich **prior**.

Dynamics Prior
How environment transits



Geometric Prior
3D scene, object, …

Human Prior
body pose, hand pose, …

Luiten et al., 2023

Rich Visual Prior from the World

Geometric Prior

Human Prior

Dynamics Prior

Voxel

Sitzmann et al., 2019

Rich Visual Prior from the World

Geometric Prior

Human Prior

Dynamics Prior

NeRF

Barron et al., 2022

Rich Visual Prior from the World

Geometric Prior

NeRF

Barron et al., 2022

Human Prior

Human Pose

Cai et al., 2023

Dynamics Prior

Rich Visual Prior from the World

Geometric Prior

NeRF

Barron et al., 2022

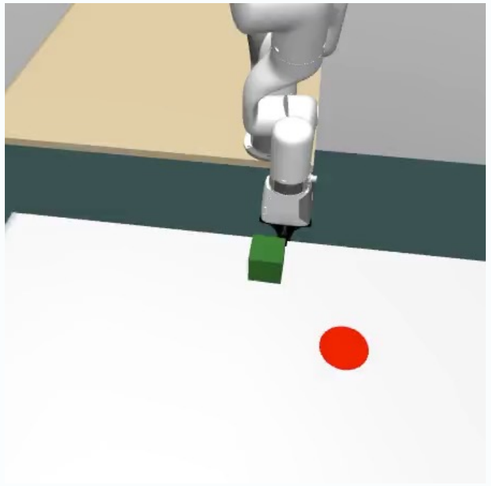
Human Prior

Human Pose

Cai et al., 2023

Dynamics Prior

Environment Transition



Yang et al., 2023

Visual Representations for Generalizable Robotic Manipulation

1

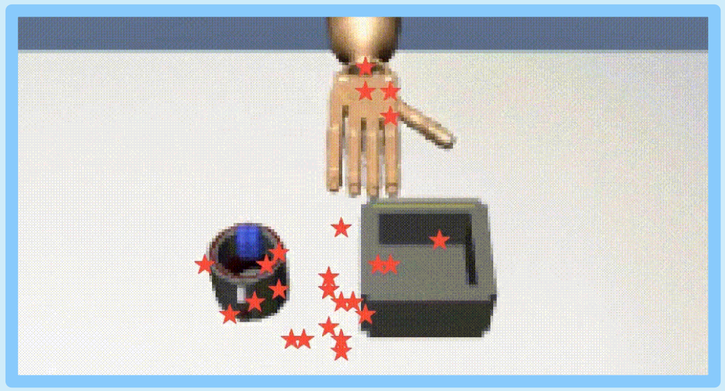
Geometric Prior

[2] **Ze** et al., “Visual Reinforcement Learning with Self-Supervised 3D Representations”, **RA-L** 2023 & **IROS** 2023.

[3] **Ze** et al., “GNFactor: Multi-Task Real Robot Learning with Generalizable Neural Feature Fields”, **CoRL** 2023 **Oral**.

2

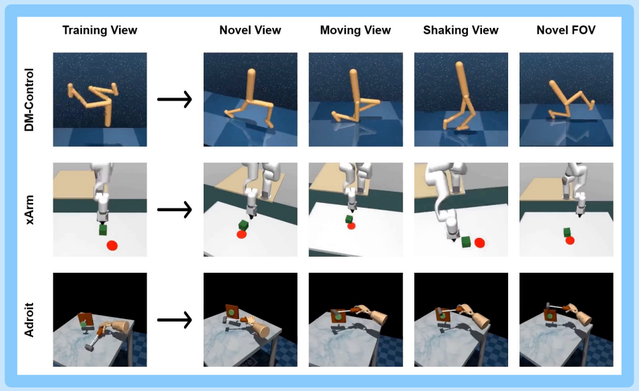
Human Prior



[4] **Ze** et al., *“H-InDex: Visual Reinforcement Learning with Hand-Informed Representations for Dexterous Manipulation”*, **NeurIPS** 2023.

3

Dynamics Prior



[5] Yang*, **Ze*** et al., *“MoVie: Visual Model-Based Policy Adaptation for View Generalization”,* **NeurIPS** 2023.

Visual Representations for Generalizable Robotic Manipulation

1

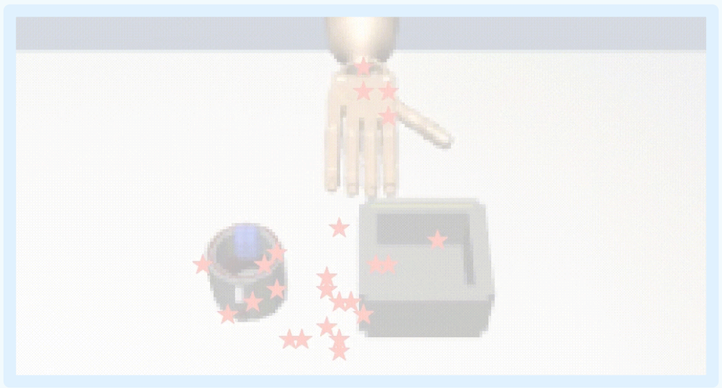
Geometric Prior

[2] **Ze** et al., “Visual Reinforcement Learning with Self-Supervised 3D Representations”, **RA-L** 2023 & **IROS** 2023.

[3] **Ze** et al., “GNFactor: Multi-Task Real Robot Learning with Generalizable Neural Feature Fields”, **CoRL** 2023 **Oral**.

2

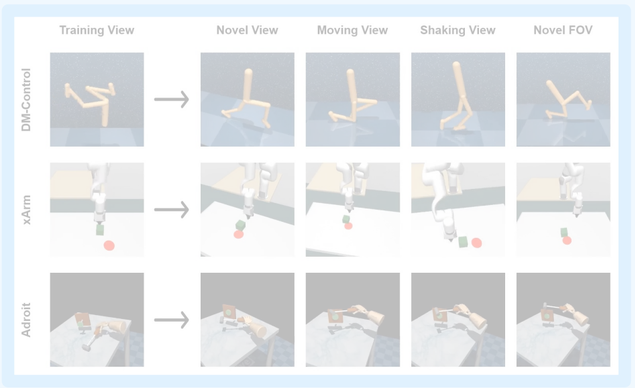
Human Prior



[4] **Ze** et al., *“H-InDex: Visual Reinforcement Learning with Hand-Informed Representations for Dexterous Manipulation”*, **NeurIPS** 2023.

3

Dynamics Prior



[5] Yang*, **Ze*** et al., *“MoVie: Visual Model-Based Policy Adaptation for View Generalization”*, **NeurIPS** 2023.

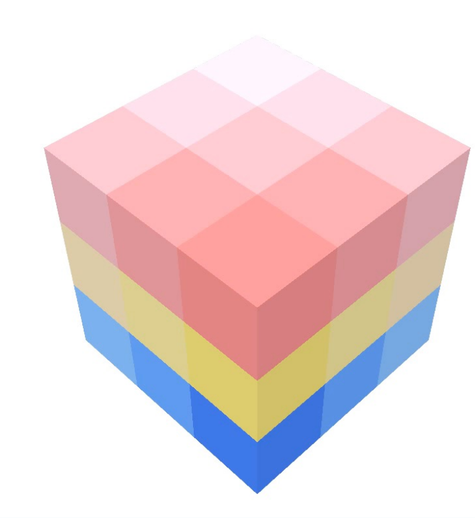
Visual Reinforcement Learning with Self-Supervised 3D Representations (RL3D)

Yanjie Ze[1*] Nicklas Hansen[2*] Yinbo Chen[2] Mohit Jain[2] Xiaolong Wang[2]

[1]Shanghai Jiao Tong University [2]UC San Diego

RA-L 2023 & **IROS** 2023

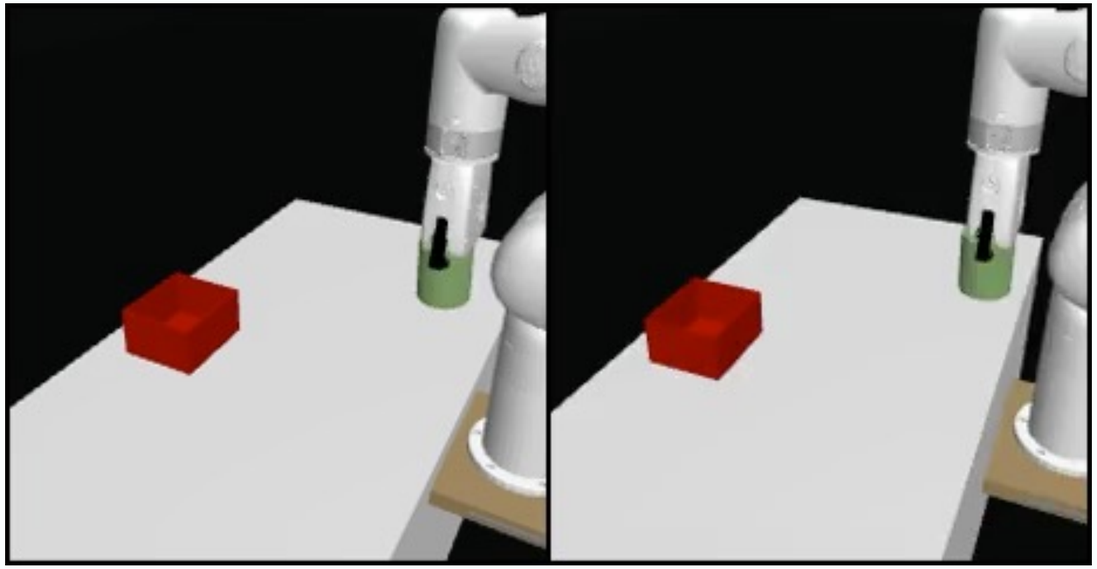
How about 3D pre-training for visuomotor control?



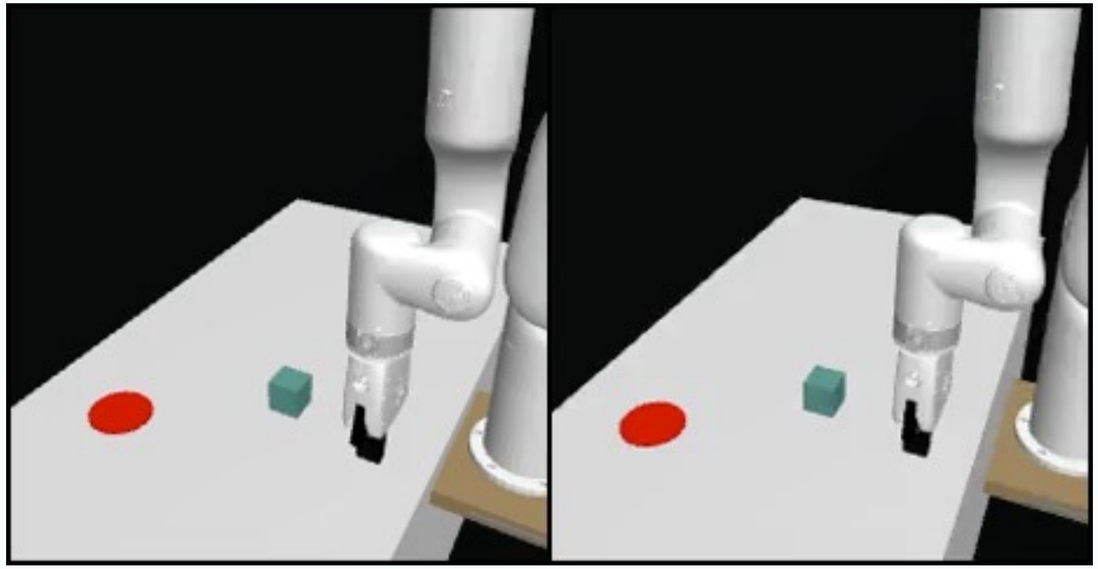
One Key Design

We design **two** cameras:
one **fixed** camera and one **dynamic** camera.

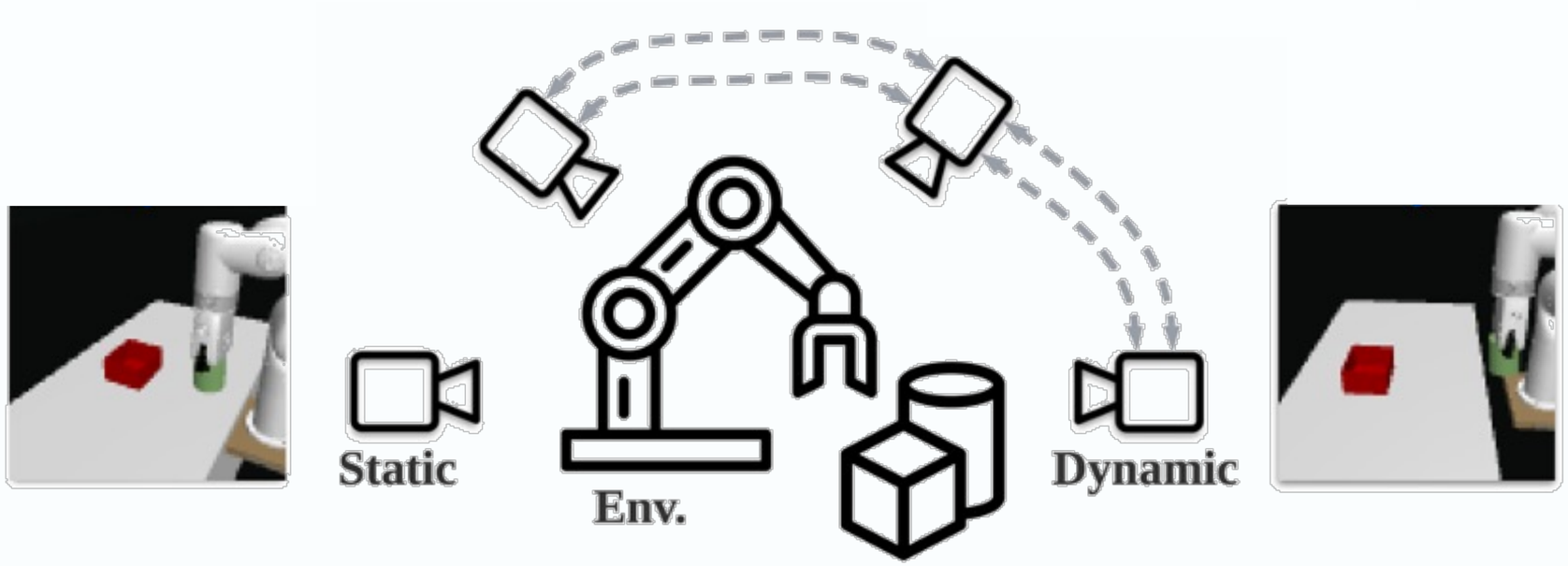
Static Dynamic



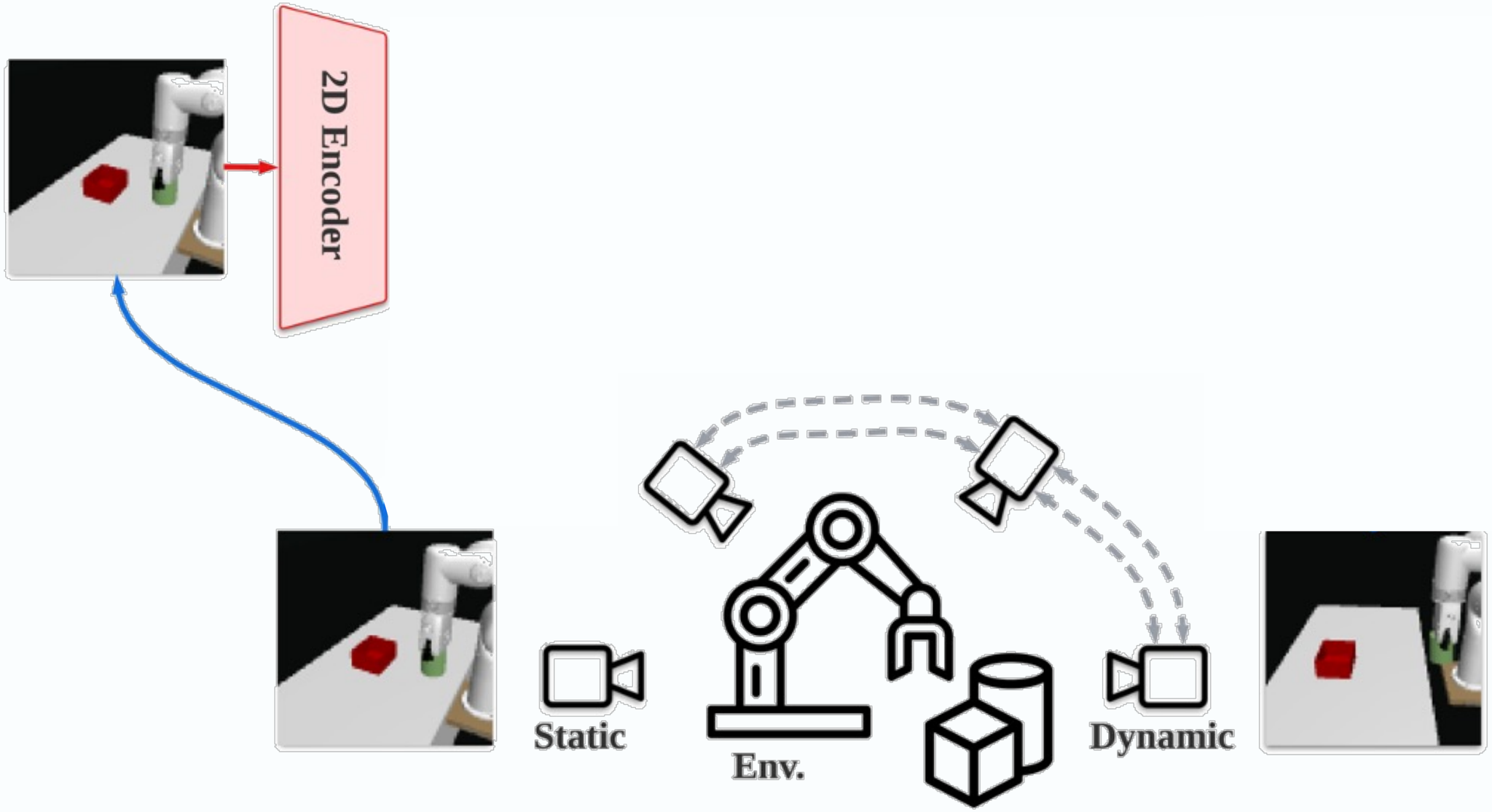
Static Dynamic



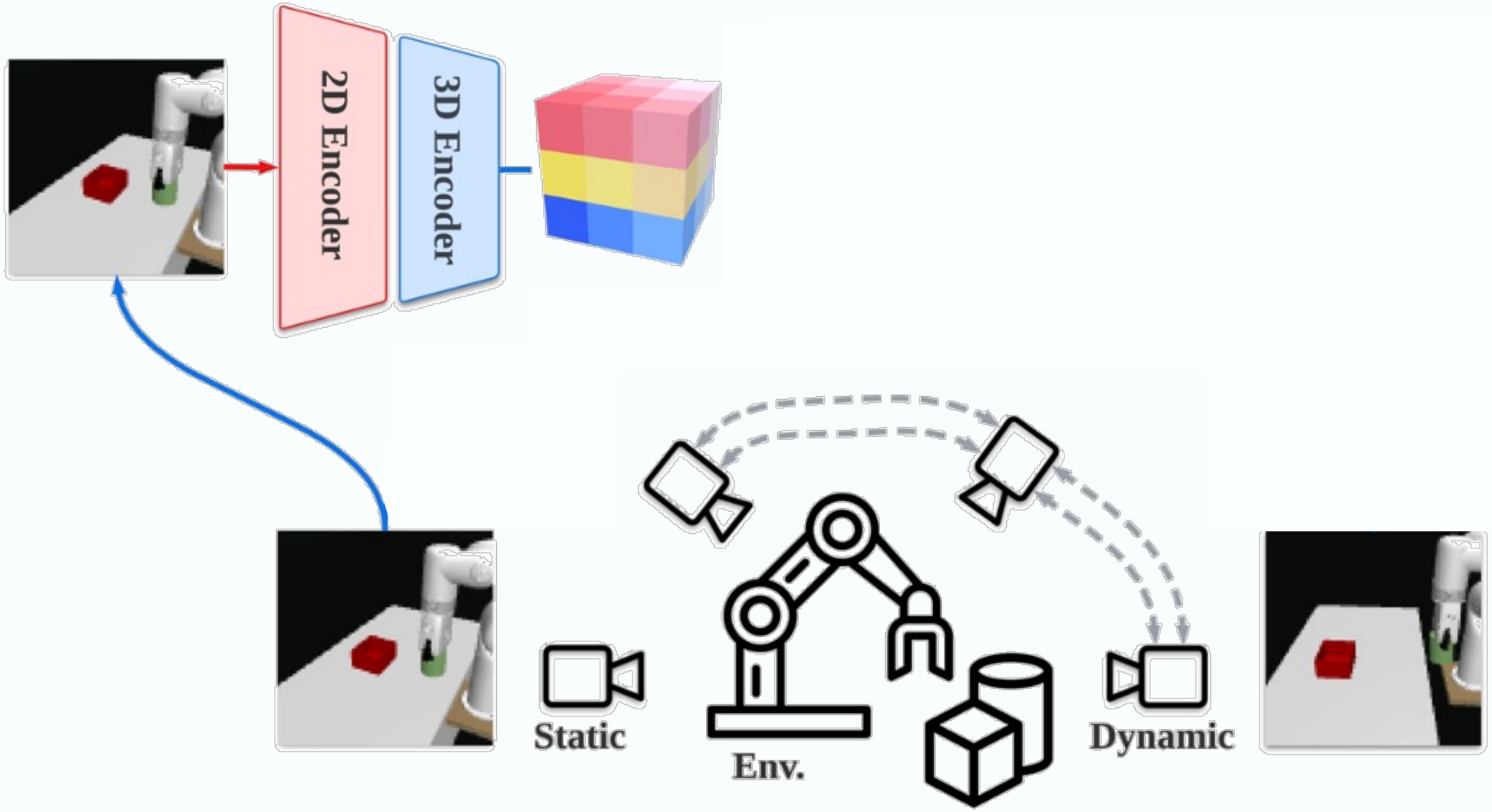
Algorithm Foundation: Video Auto-encoder



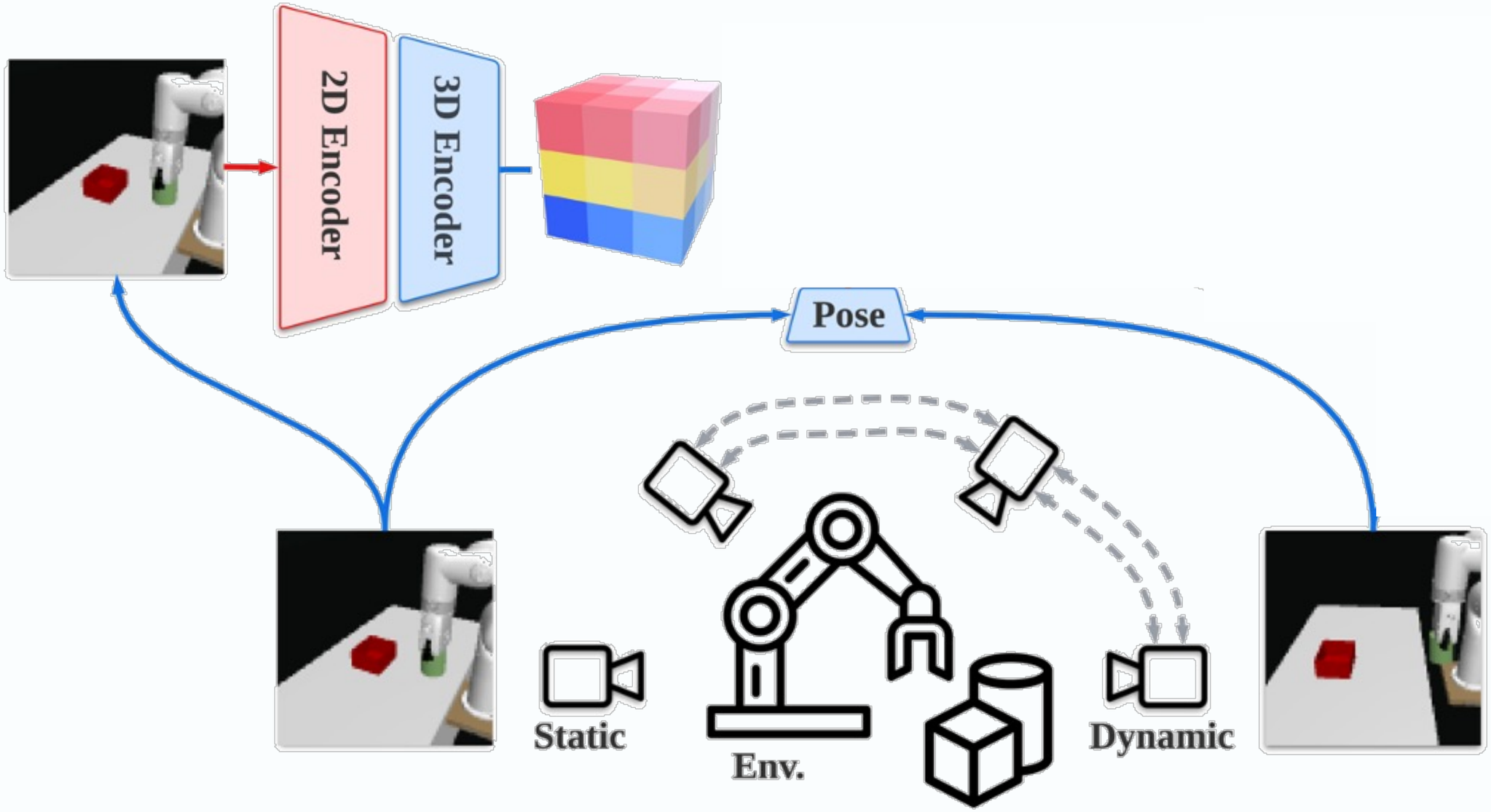
Algorithm Foundation: Video Auto-encoder



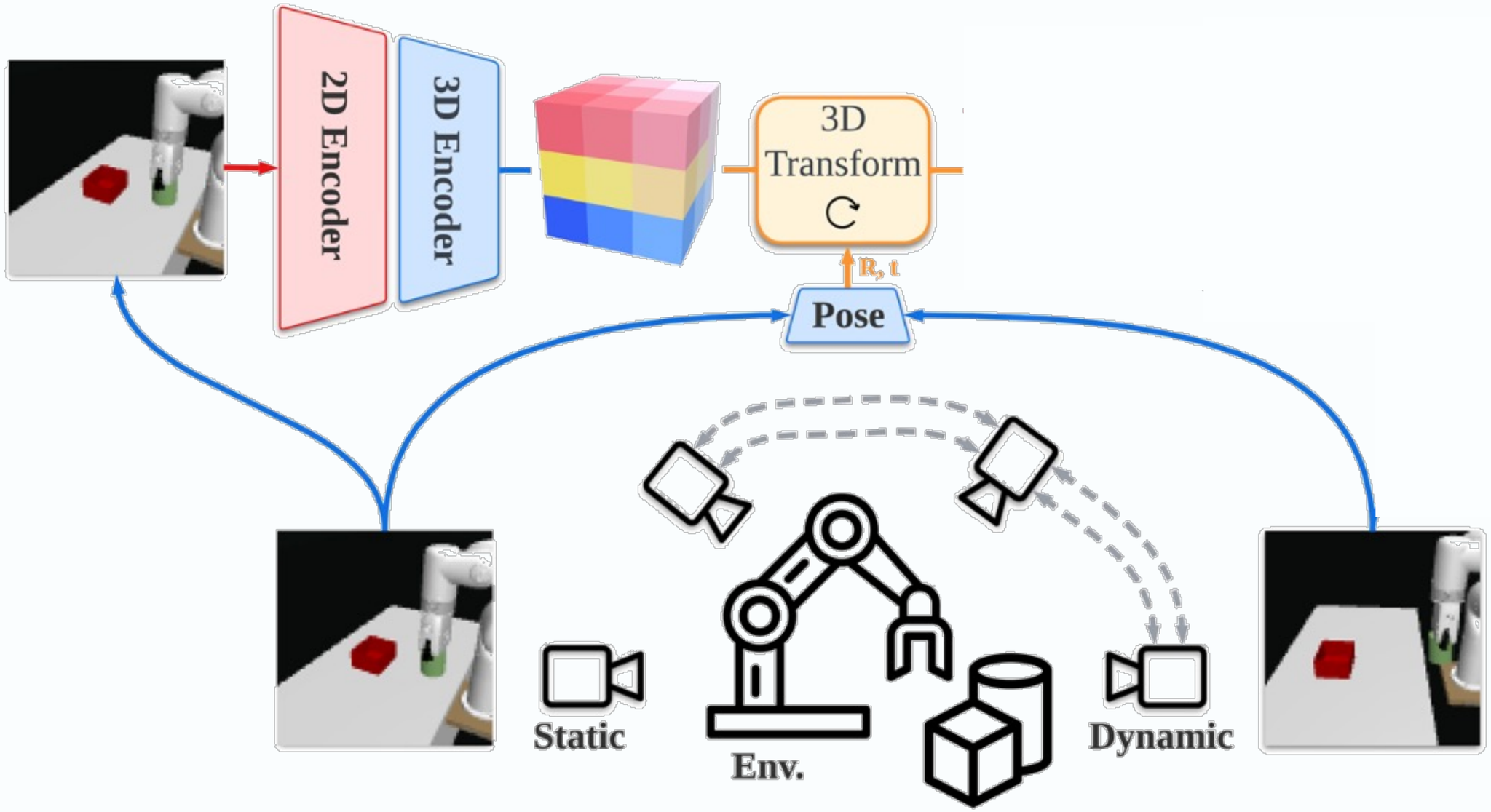
Algorithm Foundation: Video Auto-encoder



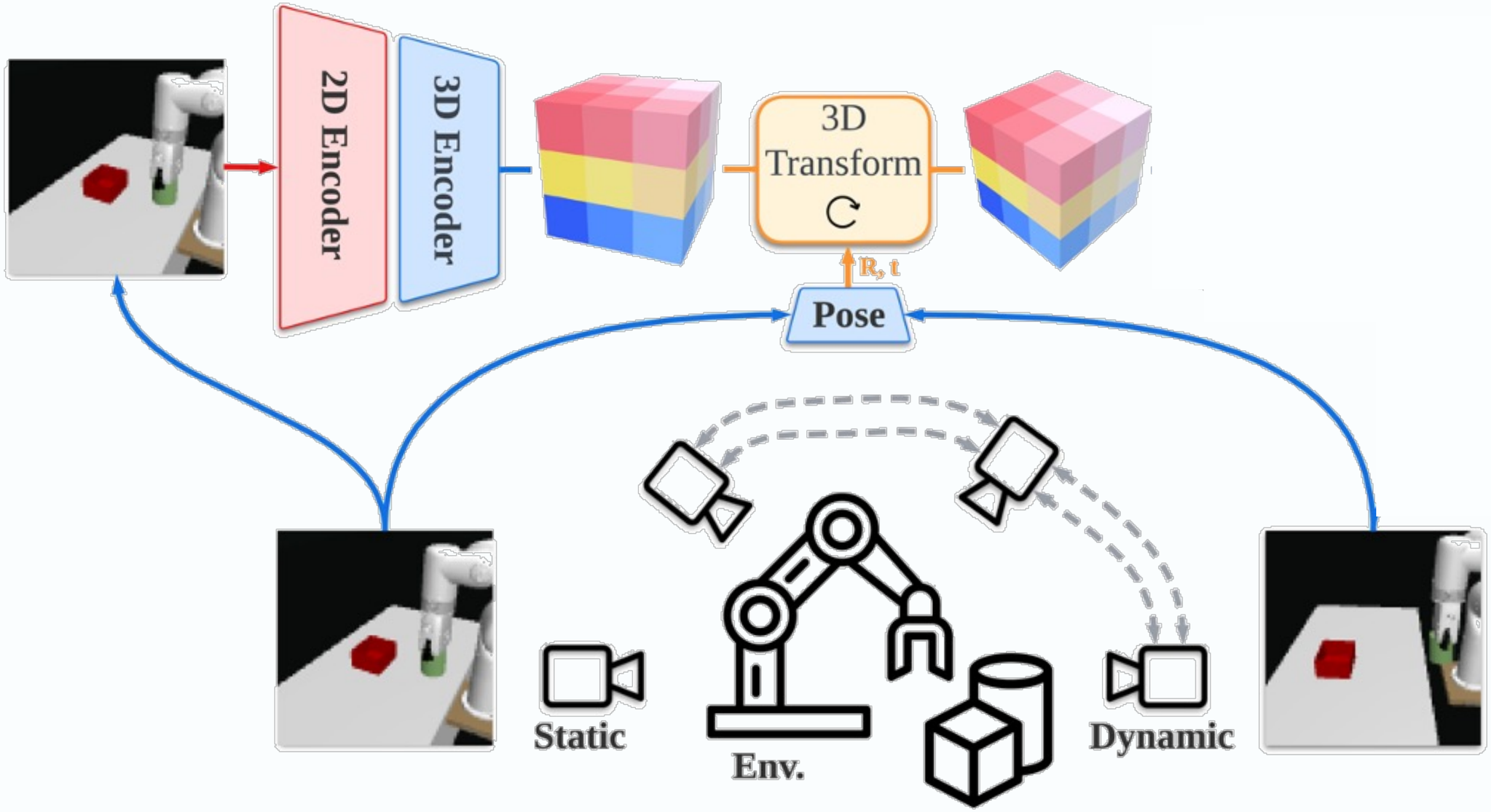
Algorithm Foundation: Video Auto-encoder



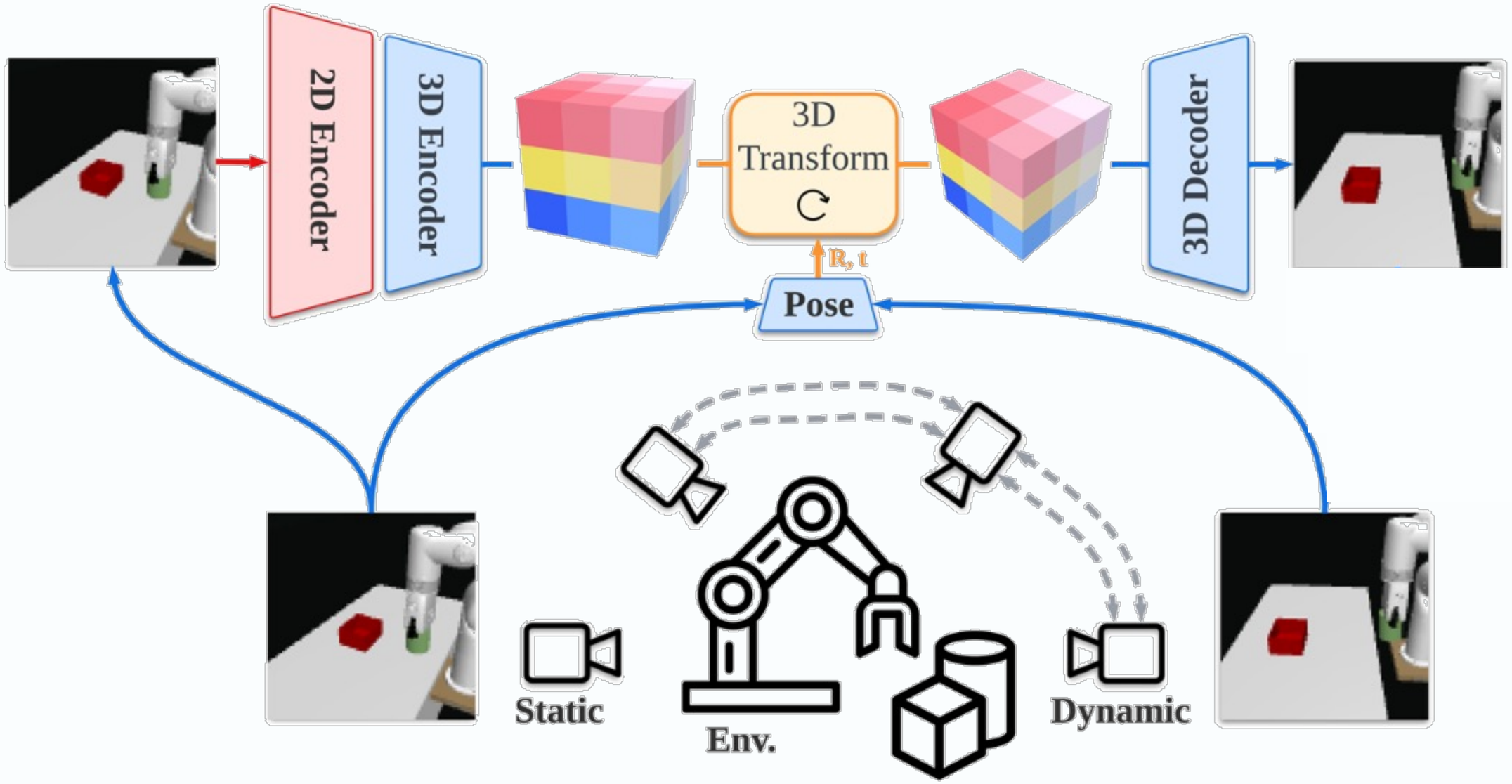
Algorithm Foundation: Video Auto-encoder



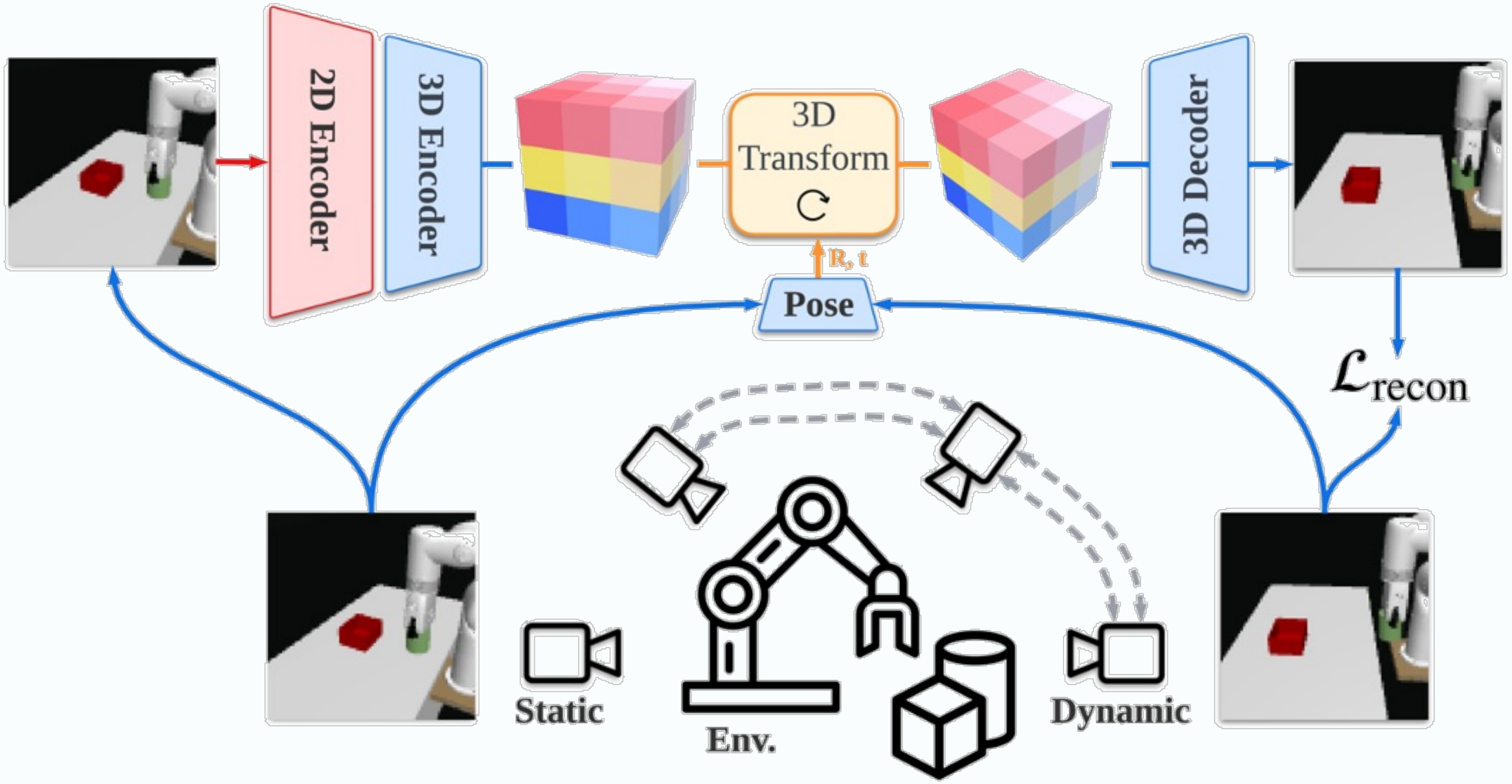
Algorithm Foundation: Video Auto-encoder



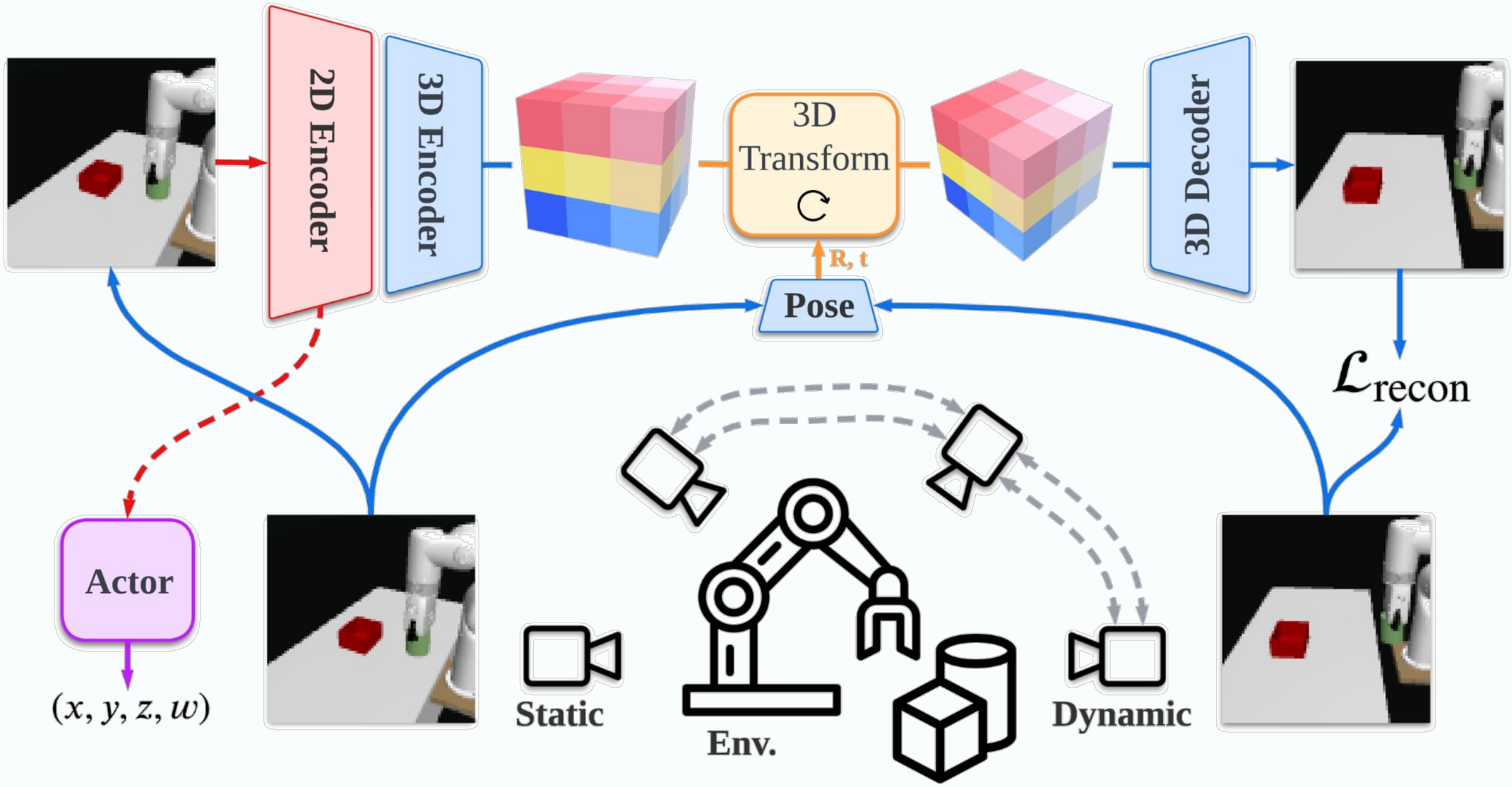
Algorithm Foundation: Video Auto-encoder



Algorithm Foundation: Video Auto-encoder



Algorithm Foundation: Video Auto-encoder



Framework of RL3D

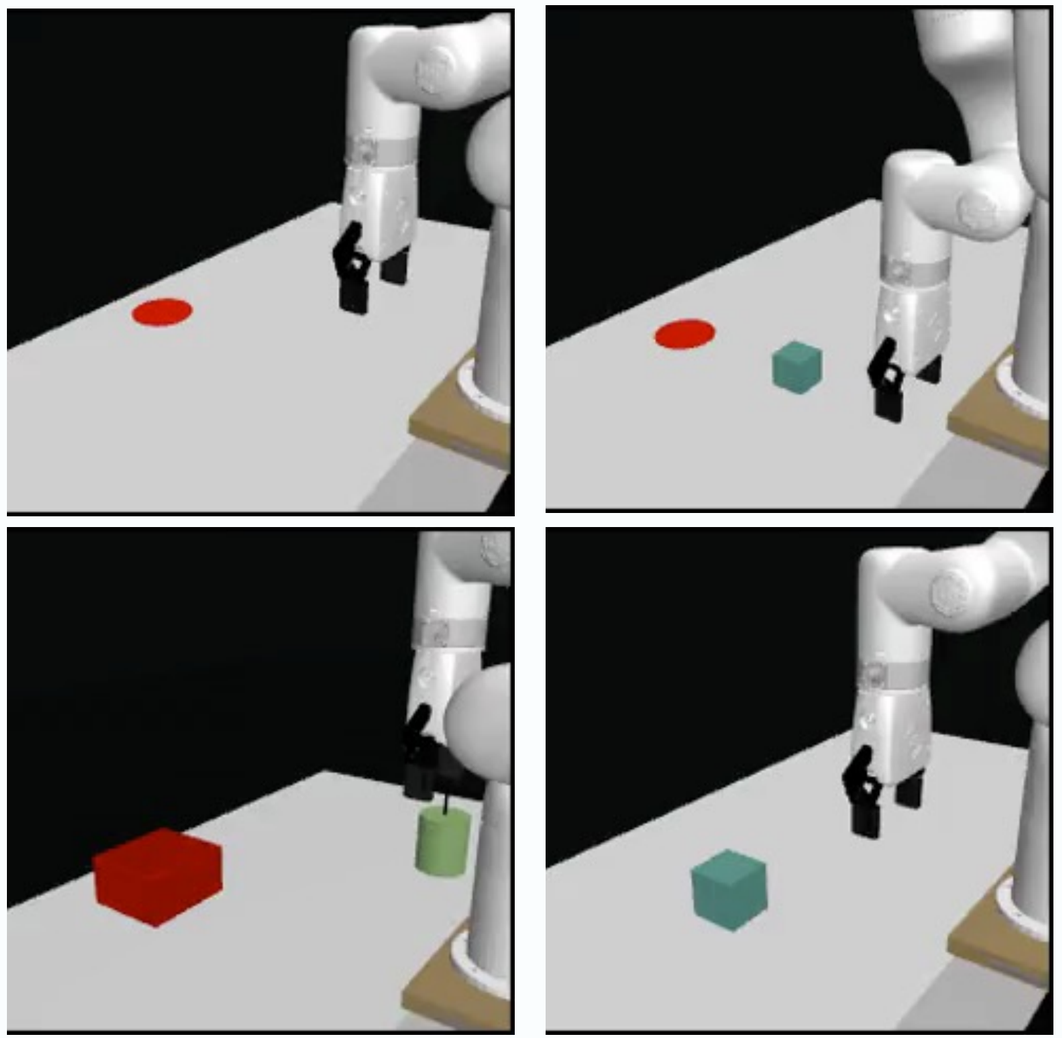
Pretrain on CO3D

Reizenstein et al, Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction, ICCV 2021.

Framework of RL3D

Pretrain on CO3D

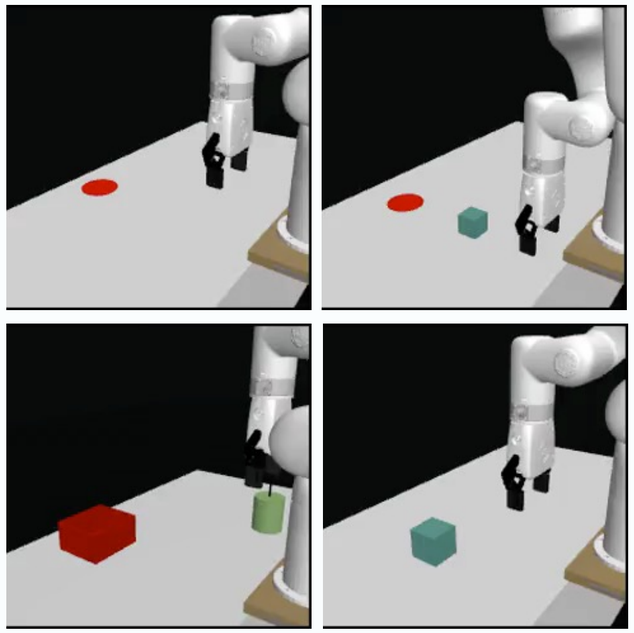
Finetune on Robot Env



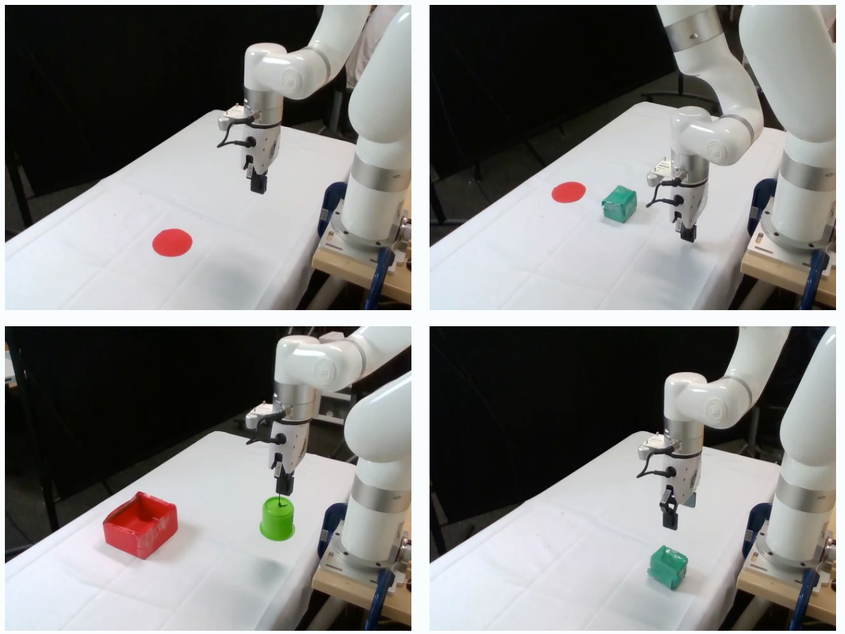
Framework of RL3D

Pretrain on CO3D

Finetune on Robot Env

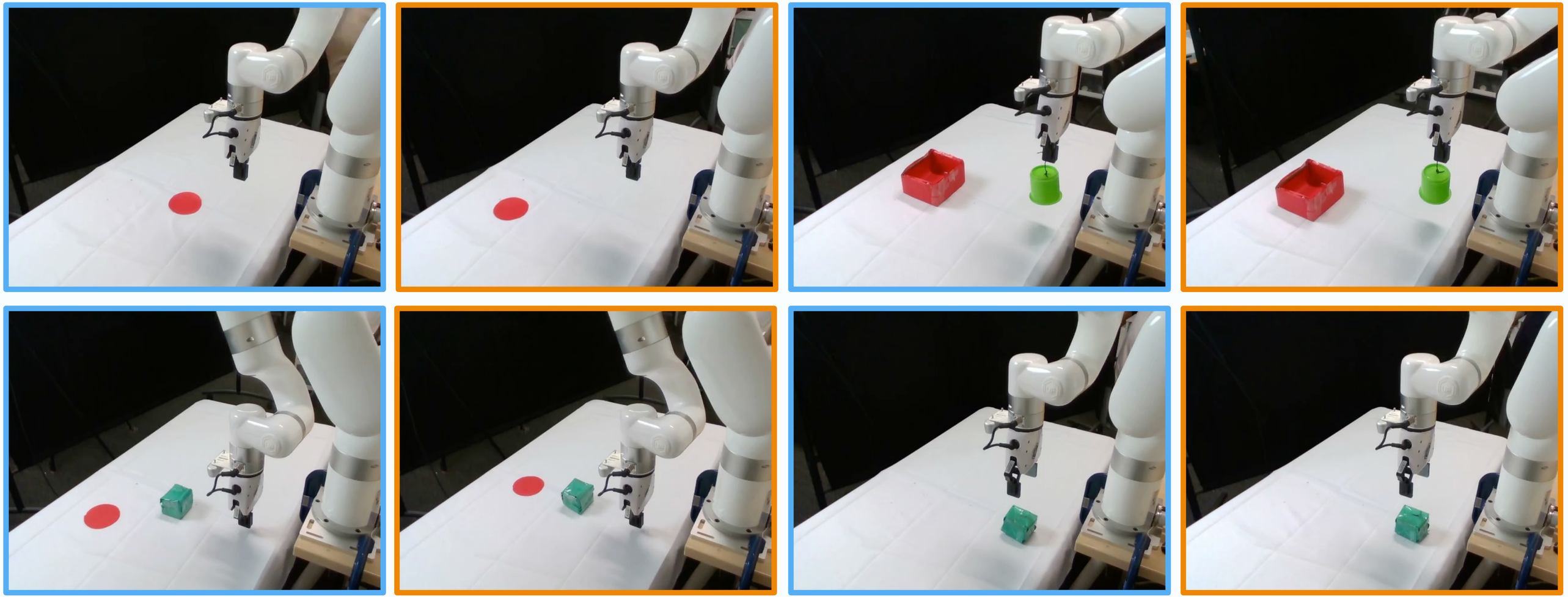


Sim-to-Real

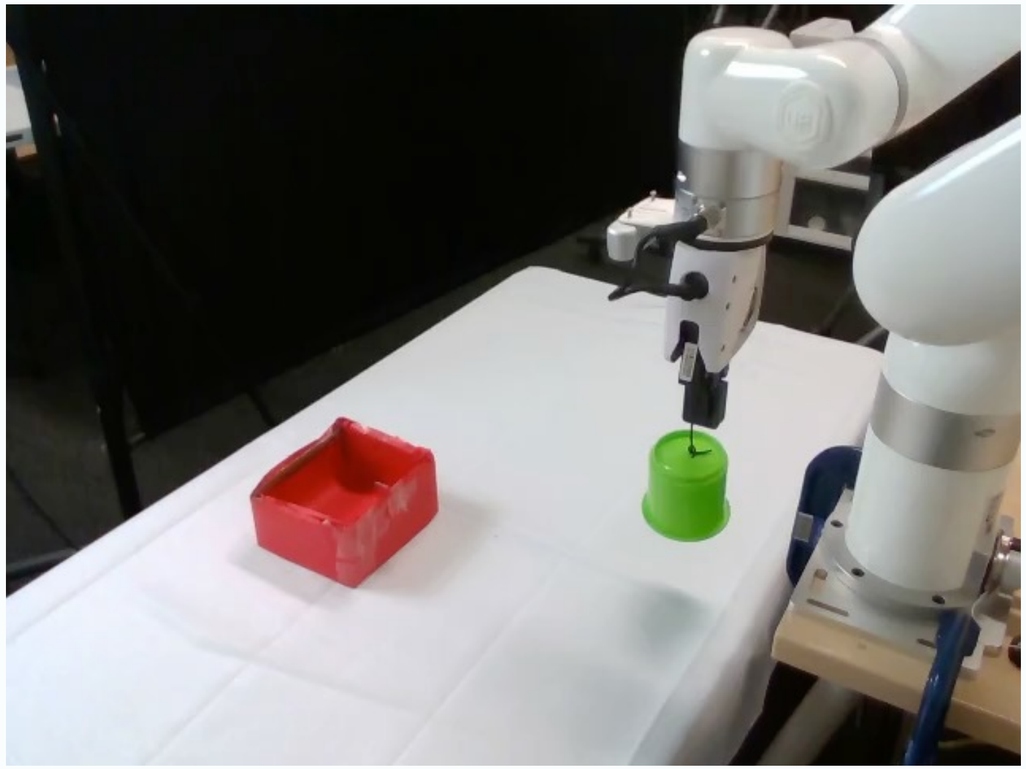


Sim-to-Real

- **Baseline** (MoCo)
- **RL3D**



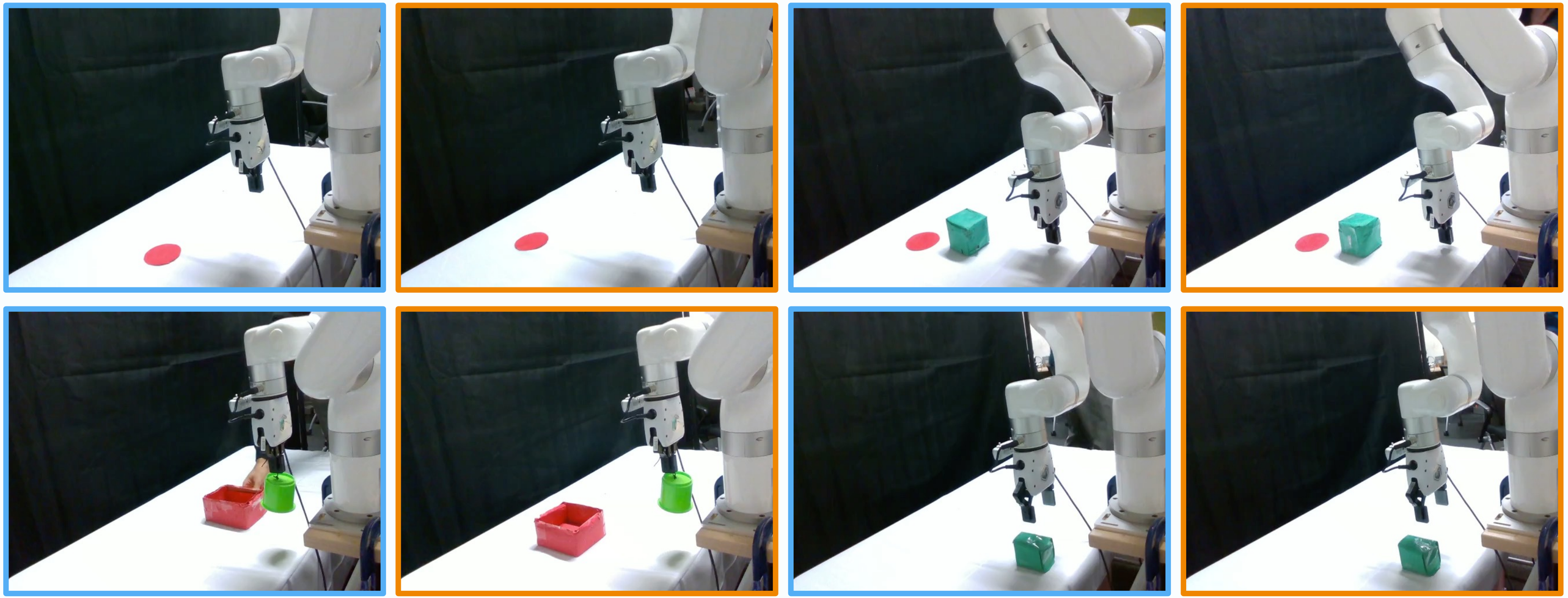
Generalize to New View



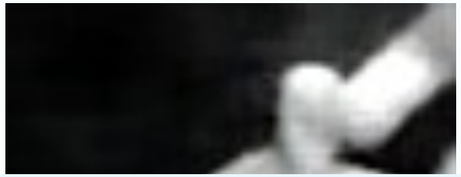
Generalize to New View

- **Baseline** (MoCo)
- **RL3D**

3D representations are more view-invariant!

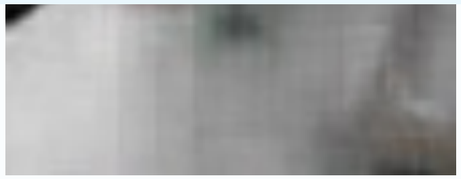


Real World View Synthesis

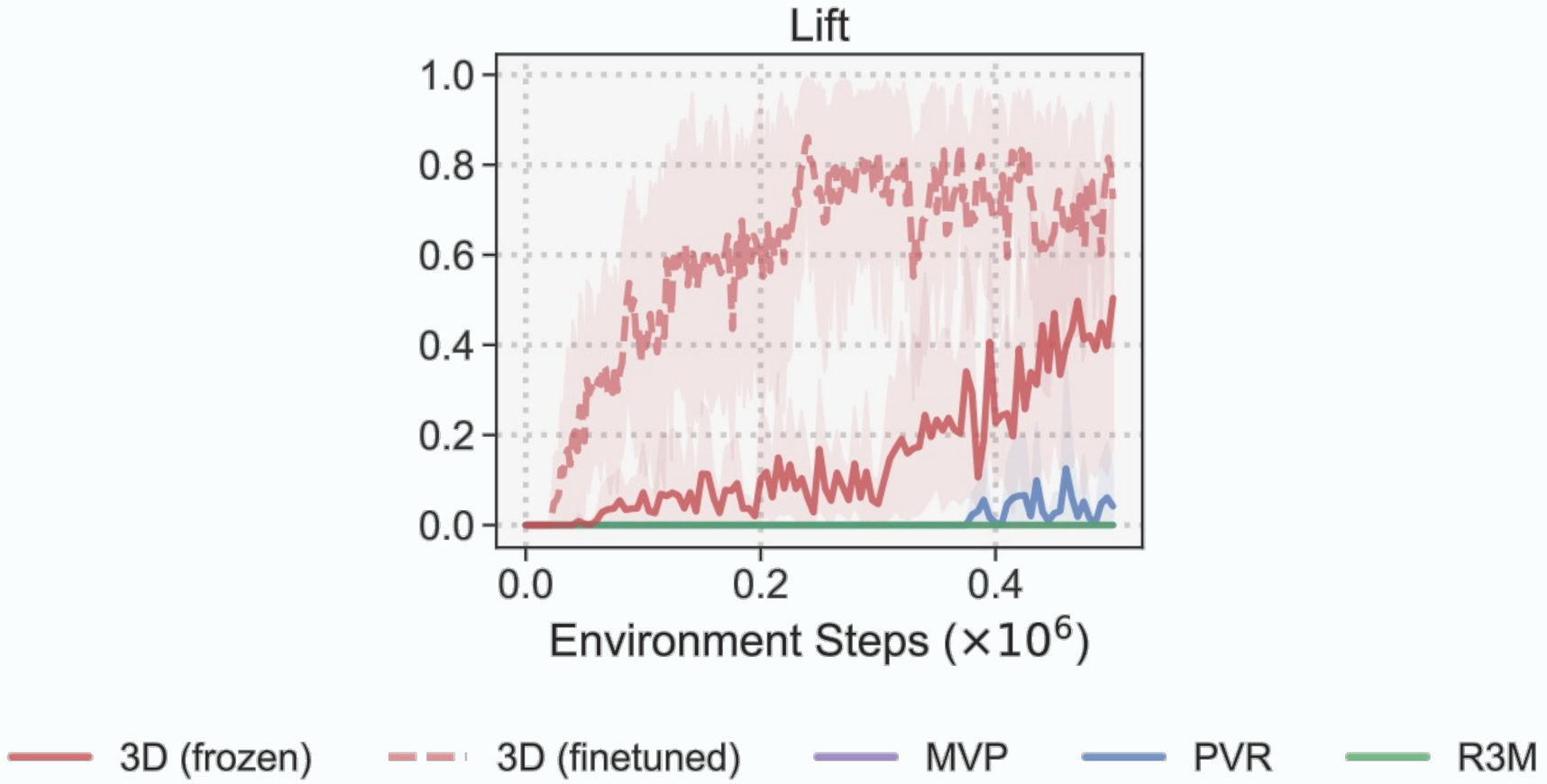



Reasons for low-quality synthesis:

- **Sim-to-real**
- **Joint training**
- **Voxel-based representations**

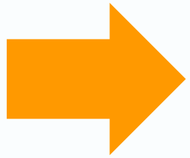



Finetuning Pre-trained Model is Critical !



Limitations of RL3D

- Only single-task agent
- Poor view synthesis
- Only easy tasks
- Only simplified scene structure
- Require millions of interactions with env



GNFactor!

GNFactor: Multi-Task Real Robot Learning with Generalizable Neural Feature Fields

Yanjie Ze[1]* Ge Yan[2]* Yueh-Hua Wu[2]* Annabella Macaluso[2]
Yuying Ge[3] Jianglong Ye[2] Nicklas Hansen[2] Li Erran Li[4] Xiaolong Wang[2]

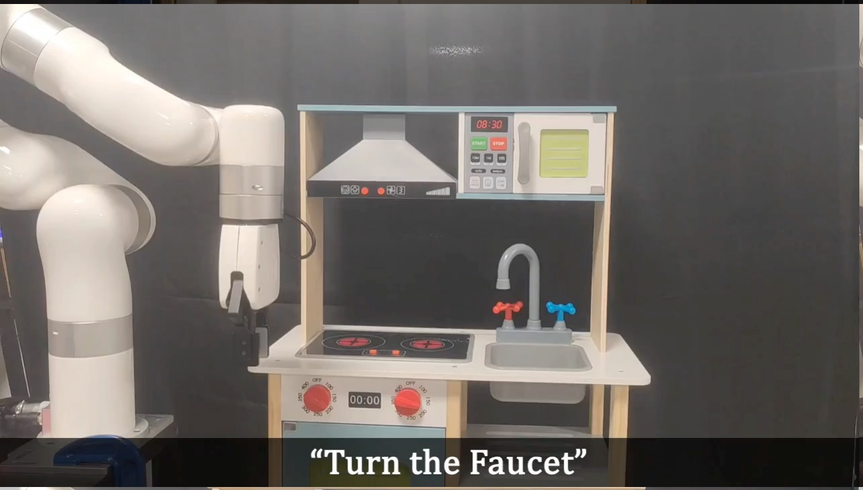
[1]Shanghai Jiao Tong University [2]UC San Diego [3]University of Hong Kong [4]AWS AI, Amazon

CoRL 2023 **Oral**


“Turn the Faucet”


“Place the Tea Pot on the Stove”


“Open the Top Microwave Door”

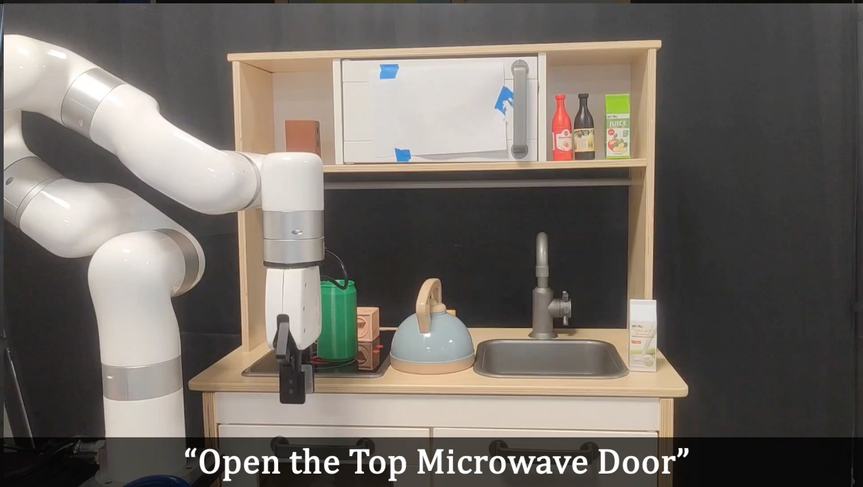

“Turn the Faucet”


“Place the Tea Pot on the Stove”


“Open the Top Microwave Door”

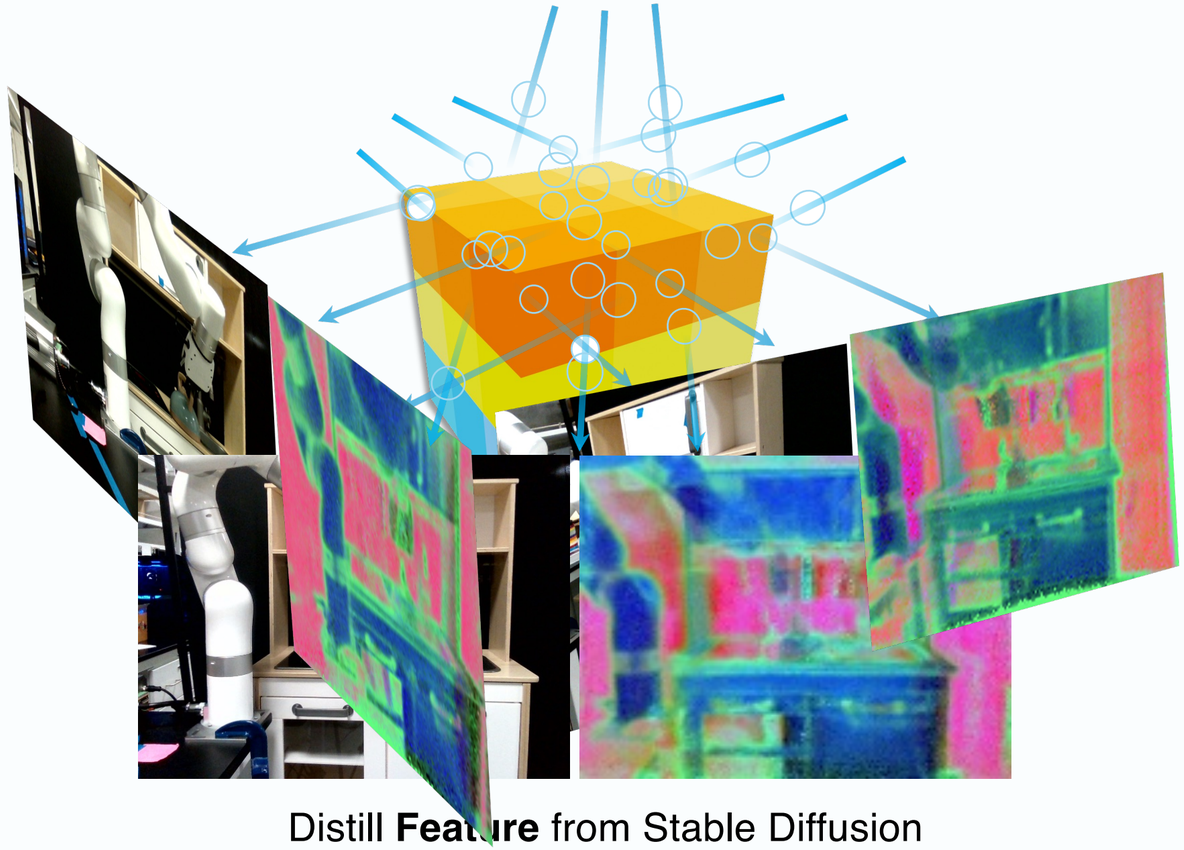

“Turn the Faucet”


“Place the Tea Pot on the Stove”


“Open the Top Microwave Door”

Overview of GNFactor

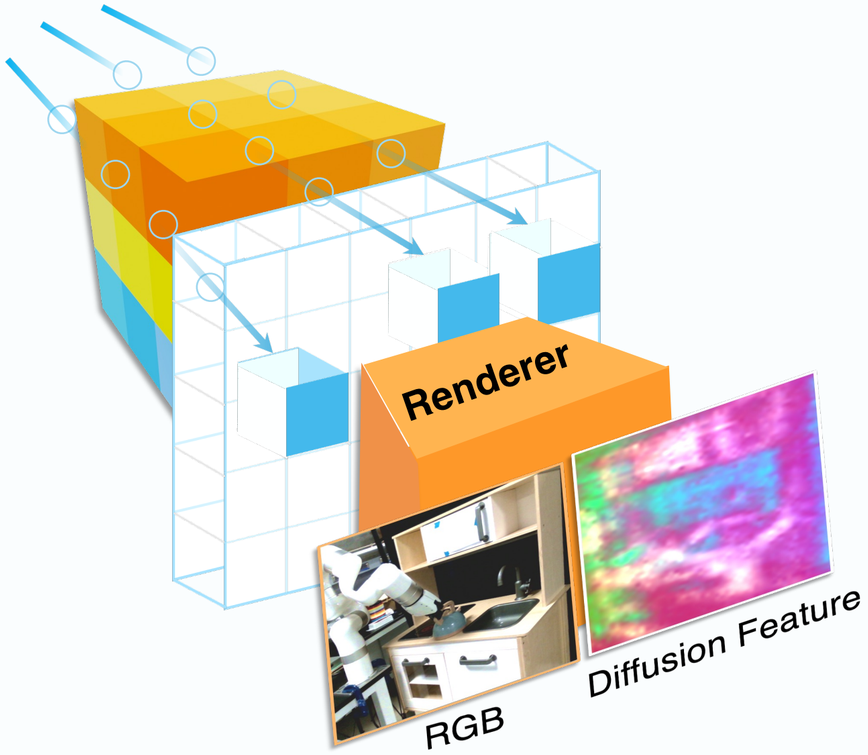
1. Train a Generalizable Neural Feature Field



Distill **Feature** from Stable Diffusion

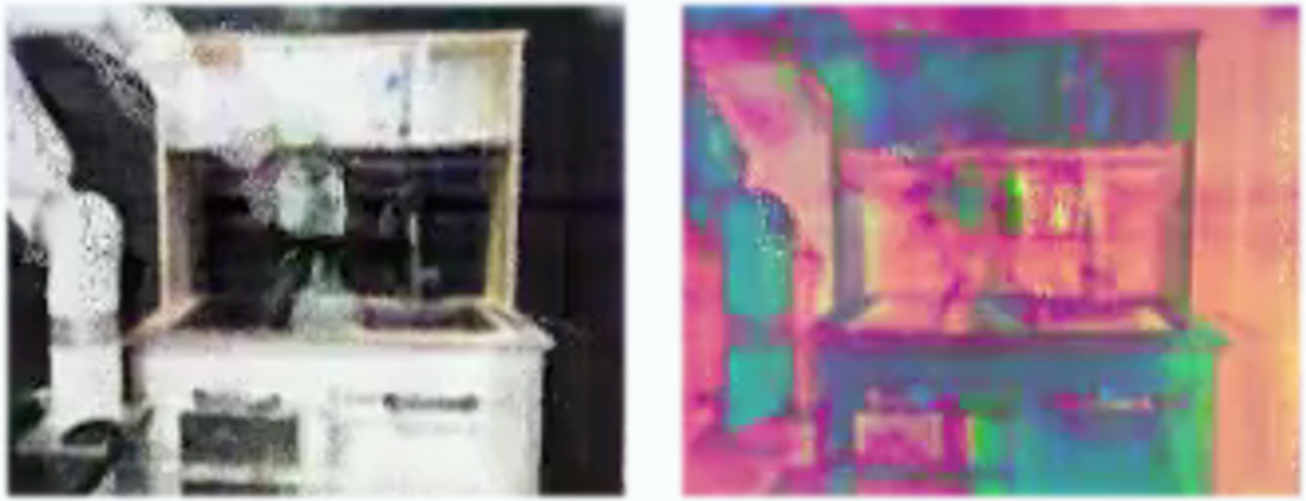
Overview of GNFactor

1. Train a Generalizable Neural Feature Field



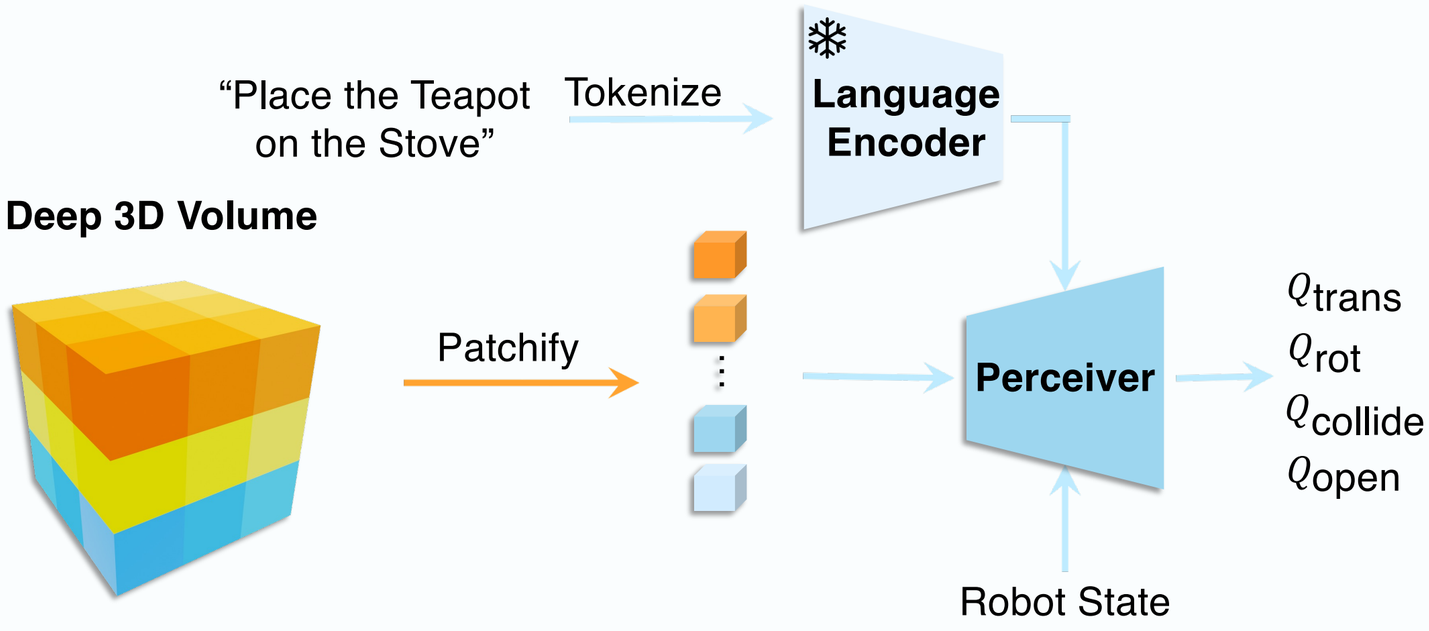
Novel View Synthesis

RGB and Feature



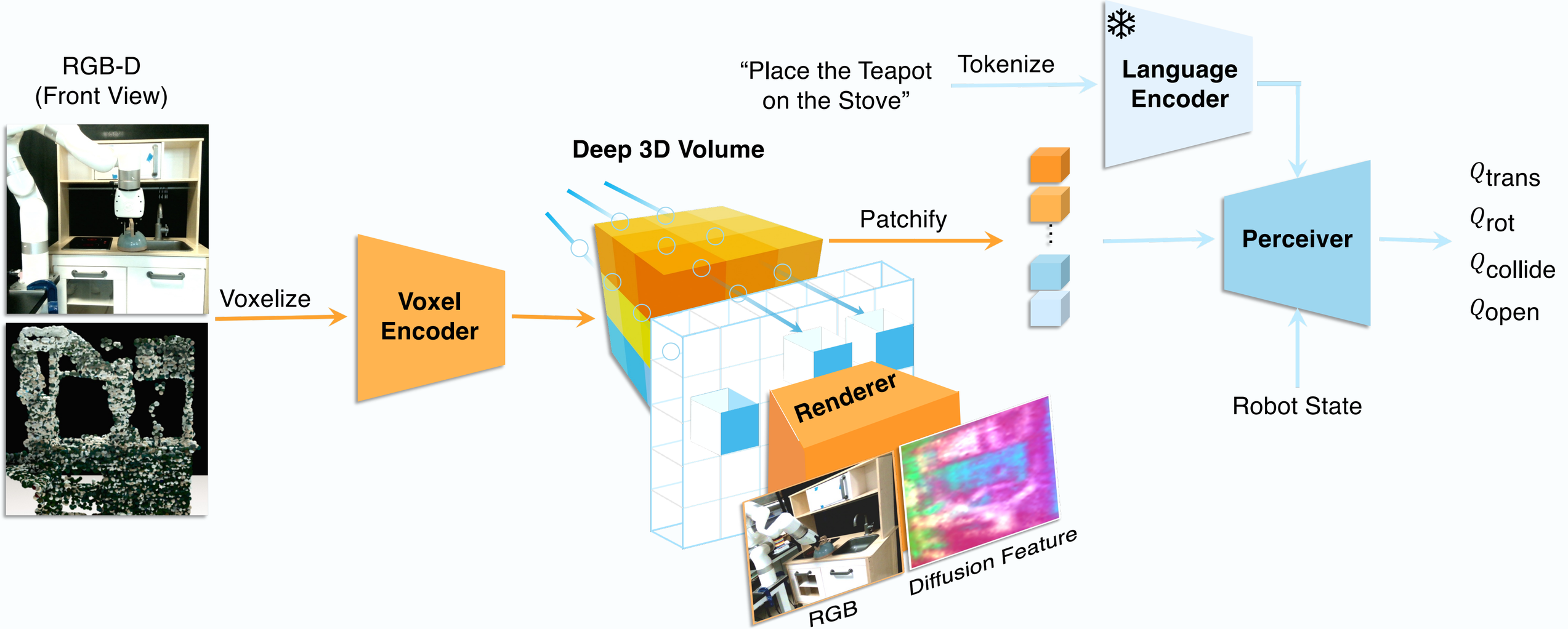
Overview of GNFactor

2. Train a Multi-Task Policy Jointly

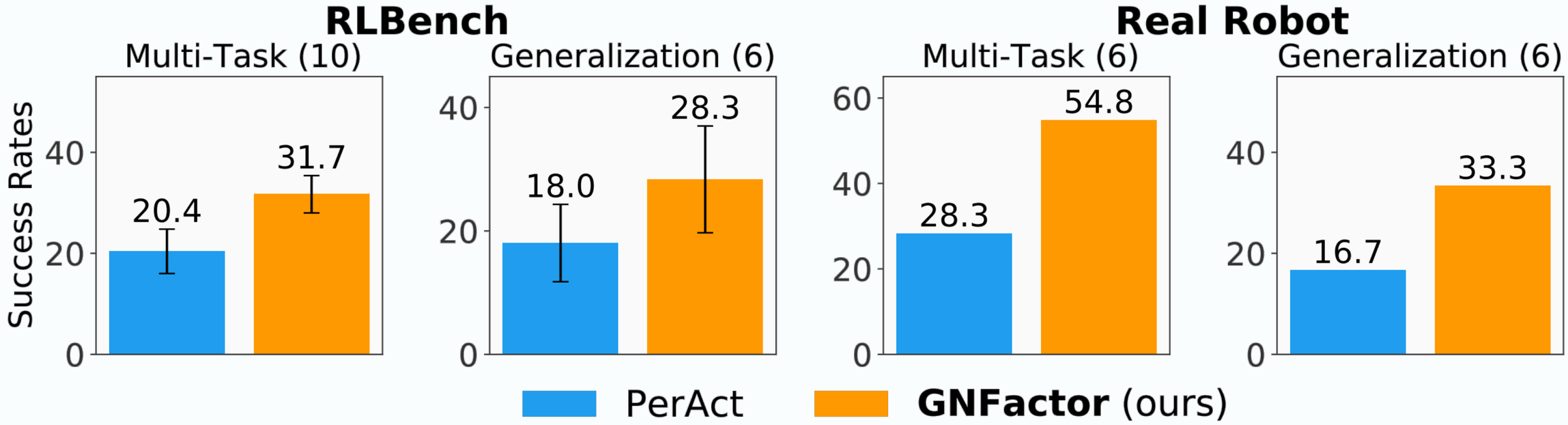


Grad-CAM in 3D

Overview of GNFactor



Strong Results in both Sim and Real



Shridhar et al, Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation, CoRL 2022.

Generalize to Cluttered Scenes

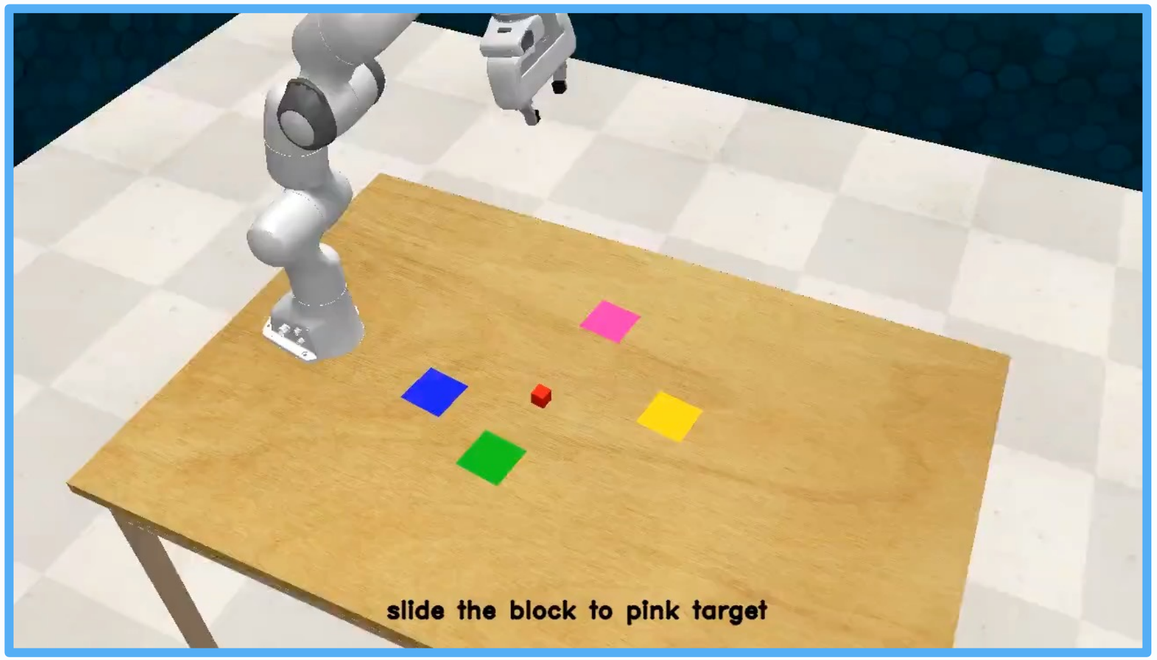
Training Scene

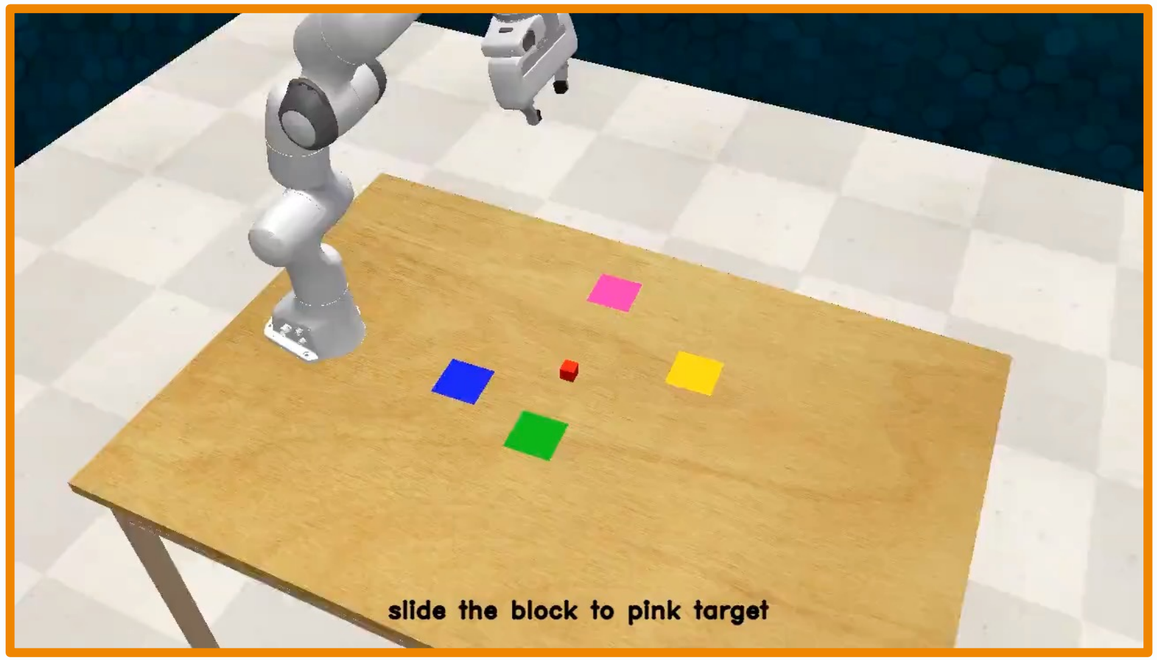
Cluttered Scene

Generalize to Unseen Scenarios

- PerAct
- GNFactor

Slide a smaller block to the target

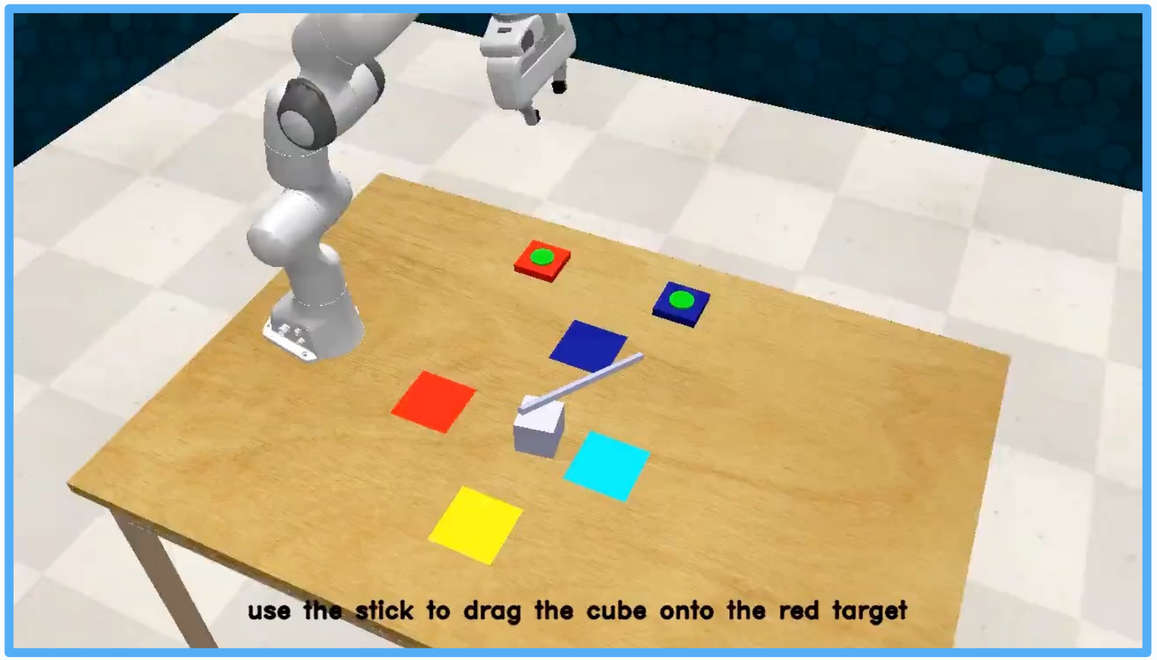


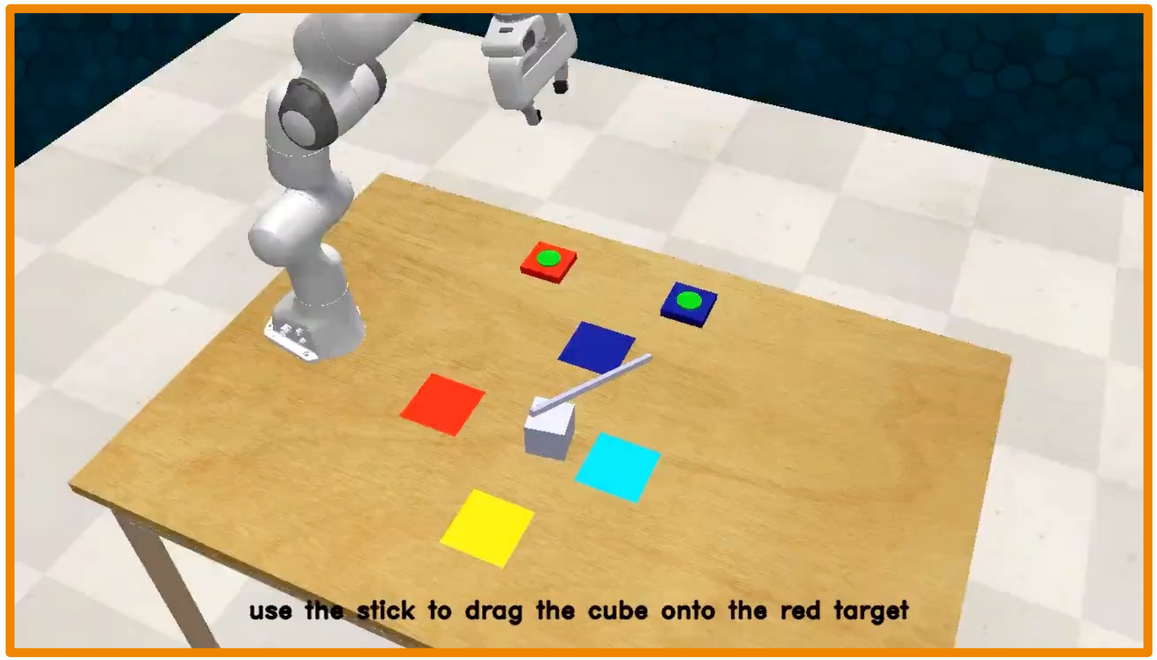


Generalize to Unseen Scenarios

- PerAct
- GNFactor

Drag the stick with distractors





Key Factors

Ablation	Success Rate (%)
GNFactor	**36.8**

Key Factors: Generalizable NeRF

Ablation	Success Rate (%)
GNFactor	**36.8**
w/o. GNF objective	24.2
w/o. RGB objective	27.2

Key Factors: Distilling Foundation Models

Ablation	Success Rate (%)
GNFactor	**36.8**
w/o. GNF objective	24.2
w/o. RGB objective	27.2
w/o. Diffusion	30.0
Diffusion $\rightarrow$ DINO	30.4
Diffusion $\rightarrow$ CLIP	32.0

Key Factors: Engineering

Ablation	Success Rate (%)
GNFactor	**36.8**
w/o. GNF objective	24.2
w/o. RGB objective	27.2
w/o. Diffusion	30.0
Diffusion $\rightarrow$ DINO	30.4
Diffusion $\rightarrow$ CLIP	32.0
w/o. DGS	29.2
w/o. skip connection	27.6

Key Factors: Parameter Tuning

Ablation	Success Rate (%)
GNFactor	**36.8**
w/o. GNF objective	24.2
w/o. RGB objective	27.2
w/o. Diffusion	30.0
Diffusion $\rightarrow$ DINO	30.4
Diffusion $\rightarrow$ CLIP	32.0
w/o. DGS	29.2
w/o. skip connection	27.6
$k = 19 \rightarrow 9$	33.2
$\lambda_{\text{feat}} = 0.01 \rightarrow 1.0$	35.2
$\lambda_{\text{recon}} = 0.01 \rightarrow 1.0$	35.2

Visual Representations for Generalizable Robotic Manipulation

1

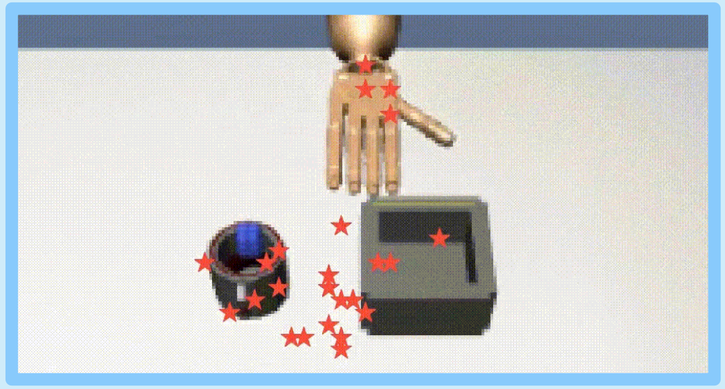
Geometric Prior

[2] **Ze** et al., “Visual Reinforcement Learning with Self-Supervised 3D Representations”, **RA-L** 2023 & **IROS** 2023.

[3] **Ze** et al., “GNFactor: Multi-Task Real Robot Learning with Generalizable Neural Feature Fields”, **CoRL** 2023 **Oral**.

2

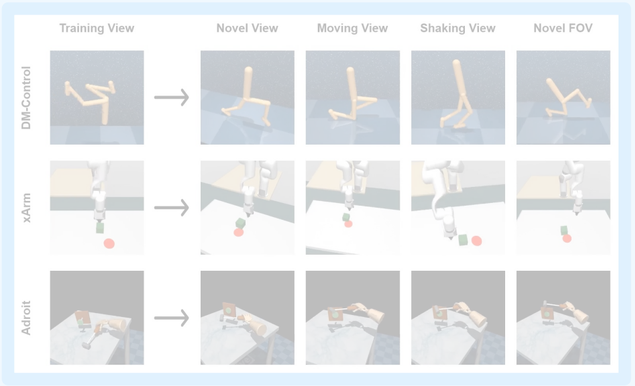
Human Prior



[4] **Ze** et al., *“H-InDex: Visual Reinforcement Learning with Hand-Informed Representations for Dexterous Manipulation”*, **NeurIPS** 2023.

3

Dynamics Prior



[5] Yang*, **Ze*** et al., *“MoVie: Visual Model-Based Policy Adaptation for View Generalization”*, **NeurIPS** 2023.

H-InDex: Visual Reinforcement Learning with Hand-Informed Representations for Dexterous Manipulation

Yanjie Ze[12] Yuyao Liu[3*] Ruizhe Shi[3*] Jiaxin Qin[4]
Zhecheng Yuan[31] Jiashun Wang[5] Huazhe Xu[316]

[1]Shanghai Qi Zhi Institute [2]Shanghai Jiao Tong University [3]Tsinghua University, IIIS
[4]Renmin University of China [5]Carnegie Mellon University [6]Shanghai AI Lab

NeurIPS 2023

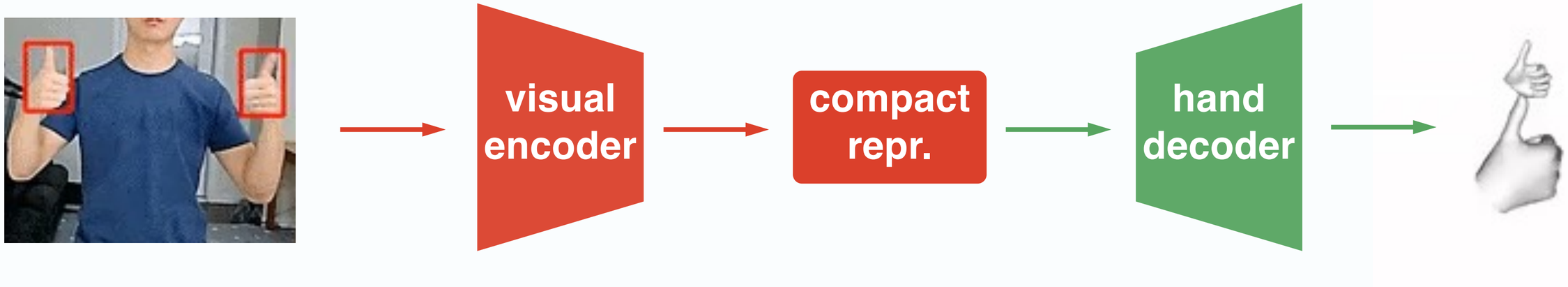




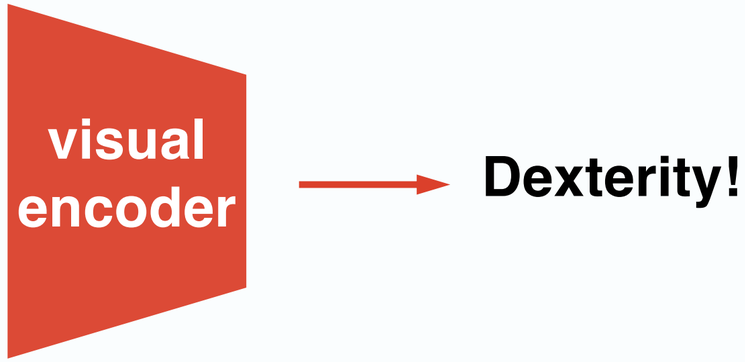


Can robotic hands leverage representations learned from human hands?

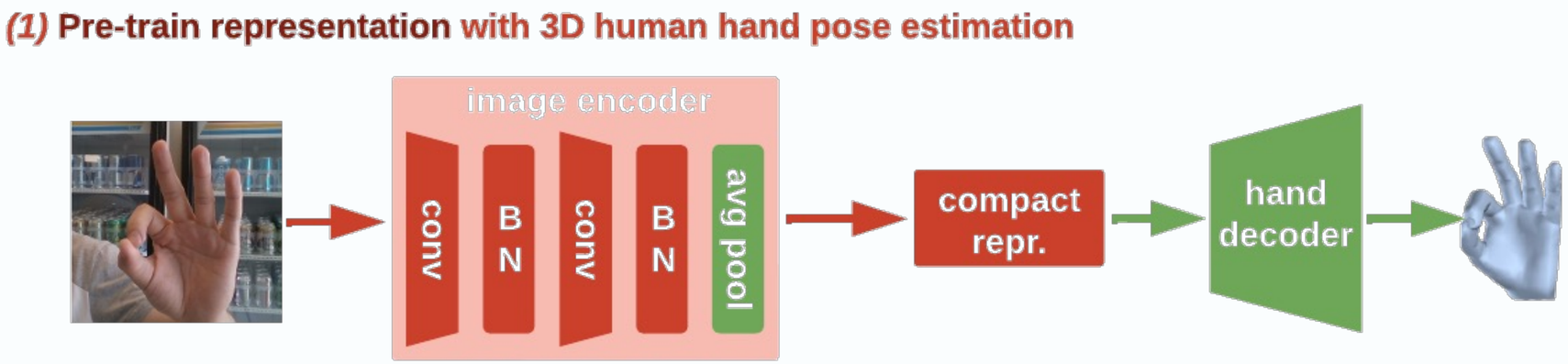
We directly adopt the pre-trained visual representation from 3D human hand pose estimation.



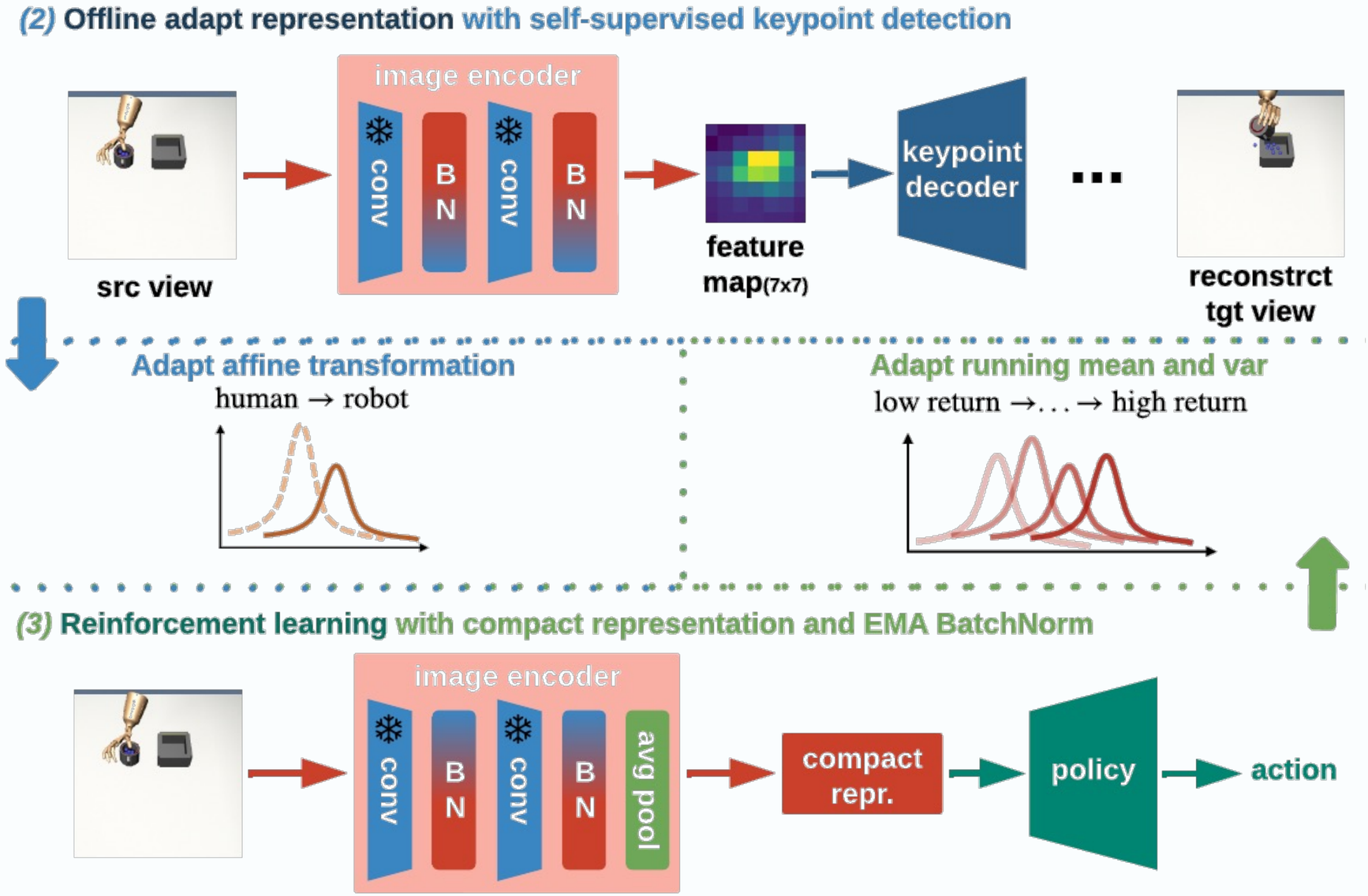
We directly adopt the pre-trained visual representation from 3D human hand pose estimation.



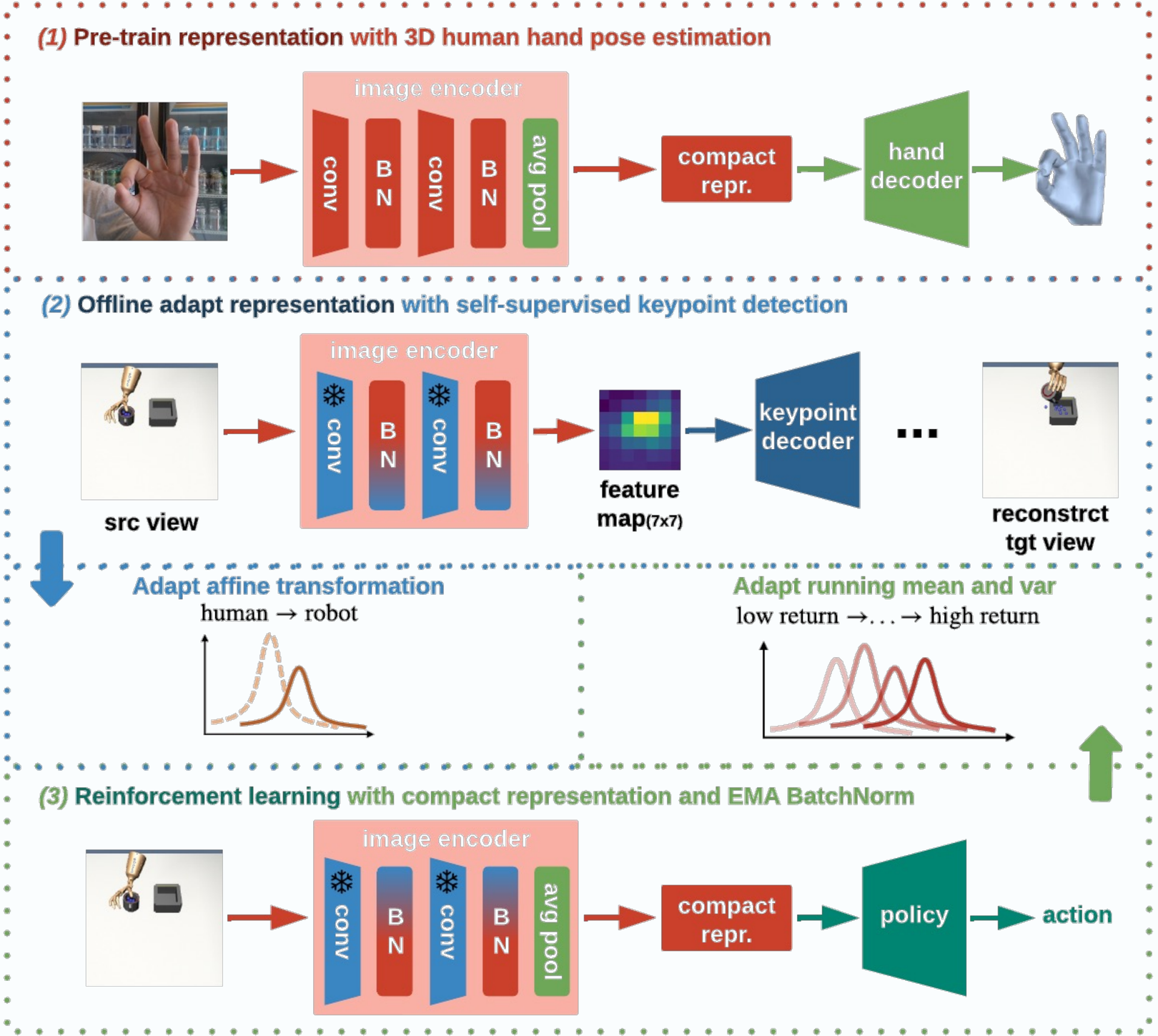
Overview of H-InDex



Overview of H-InDex

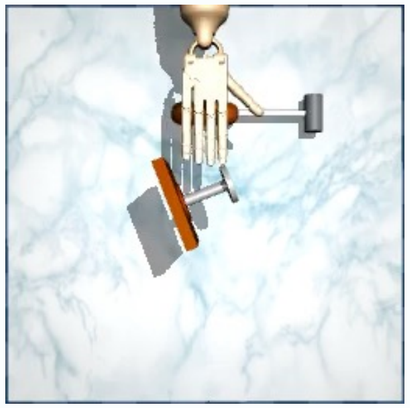


Overview of H-InDex

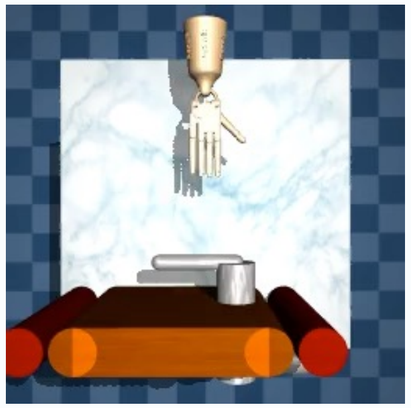


Strong Results in 12 Dexterous Manipulation Tasks

Hammer

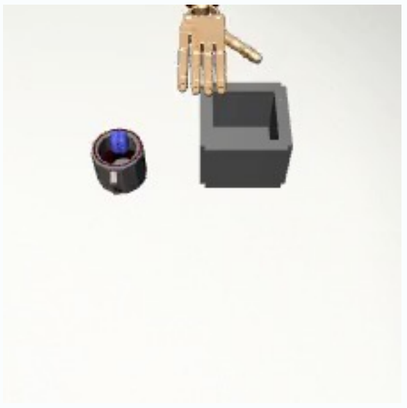


Door



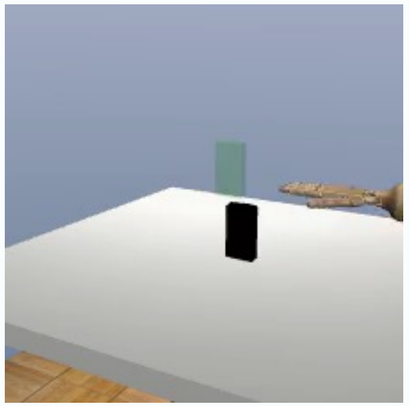
Pen

Pour

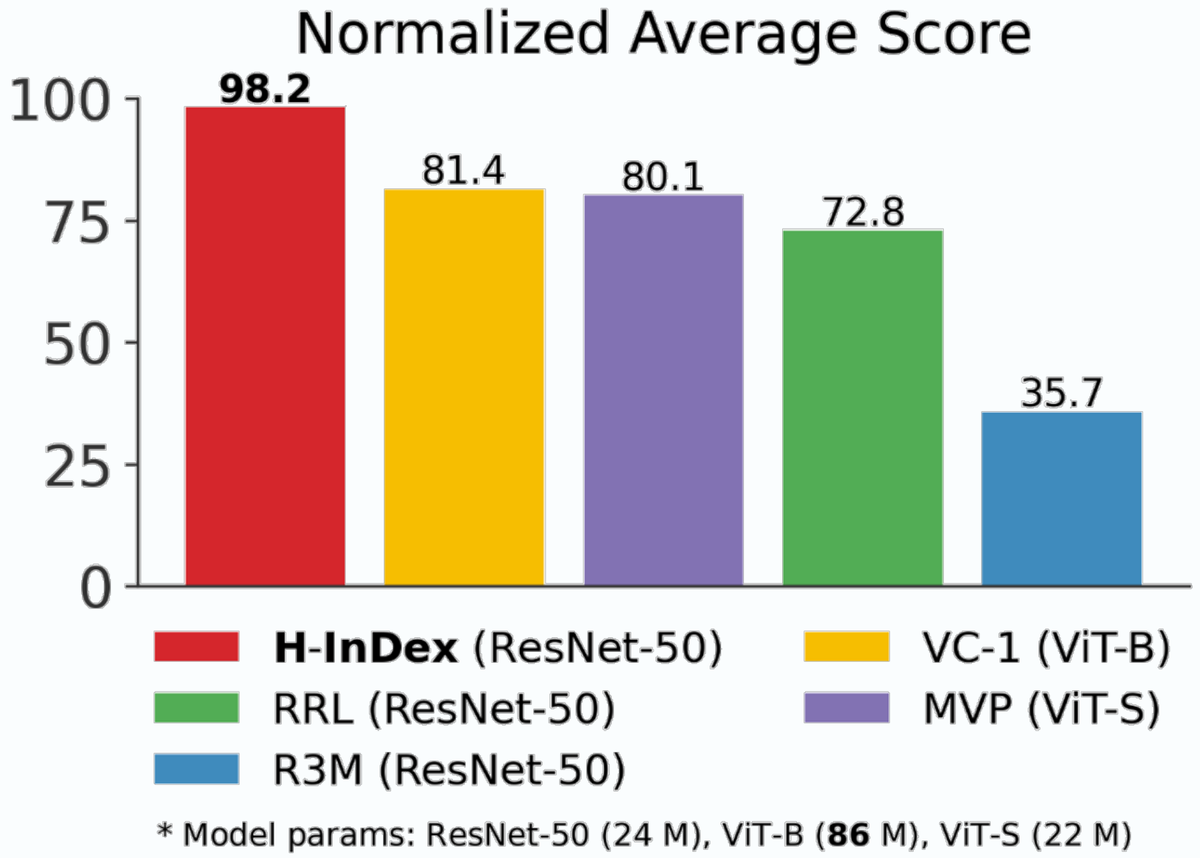


Place Inside

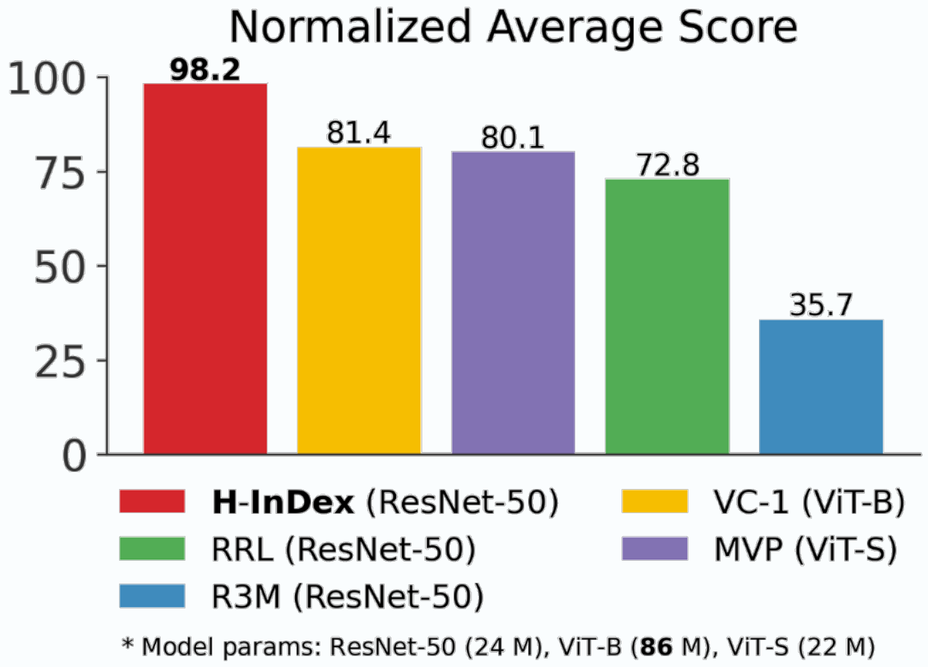
Relocate

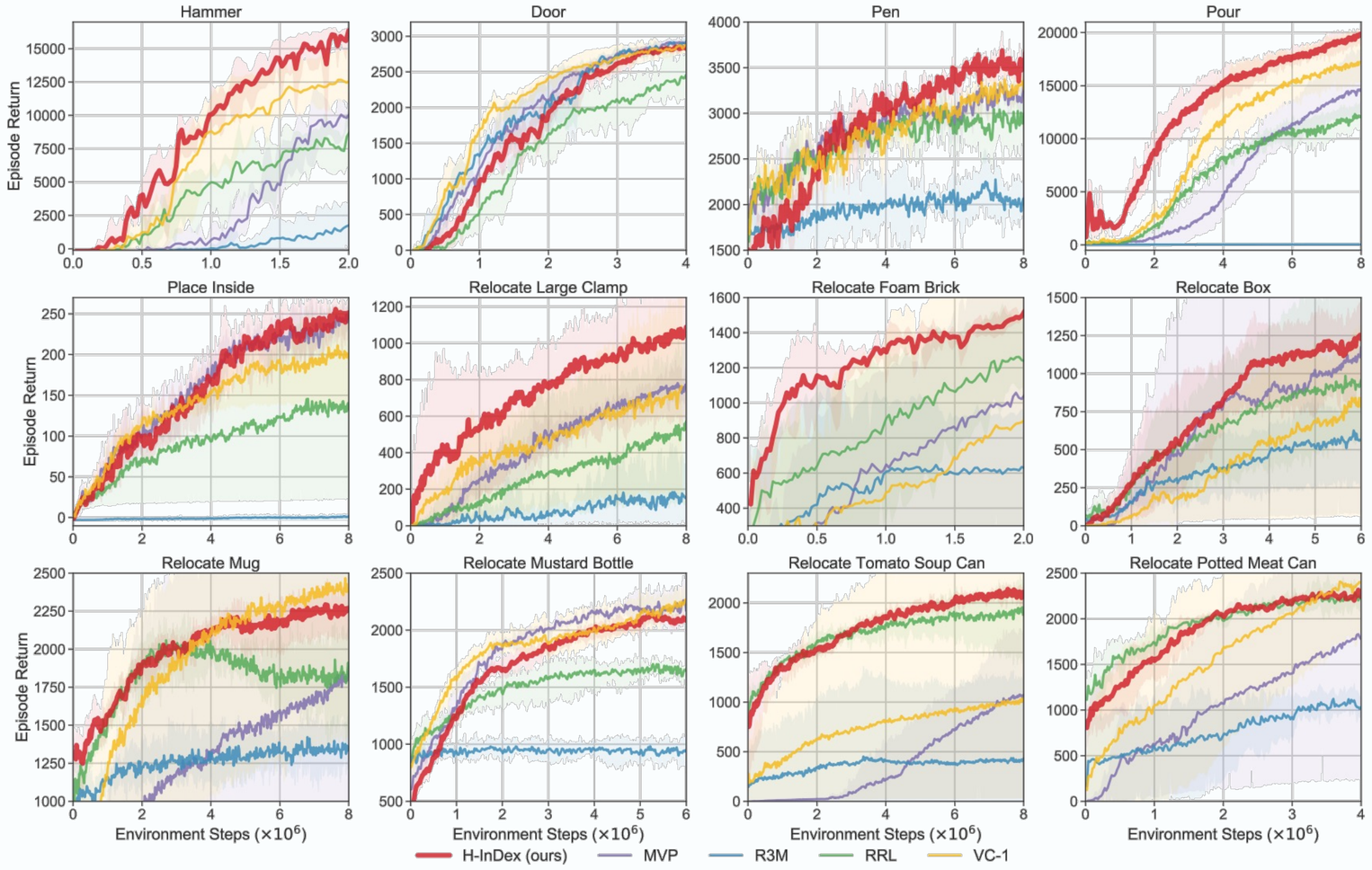


Strong Results in 12 Dexterous Manipulation Tasks

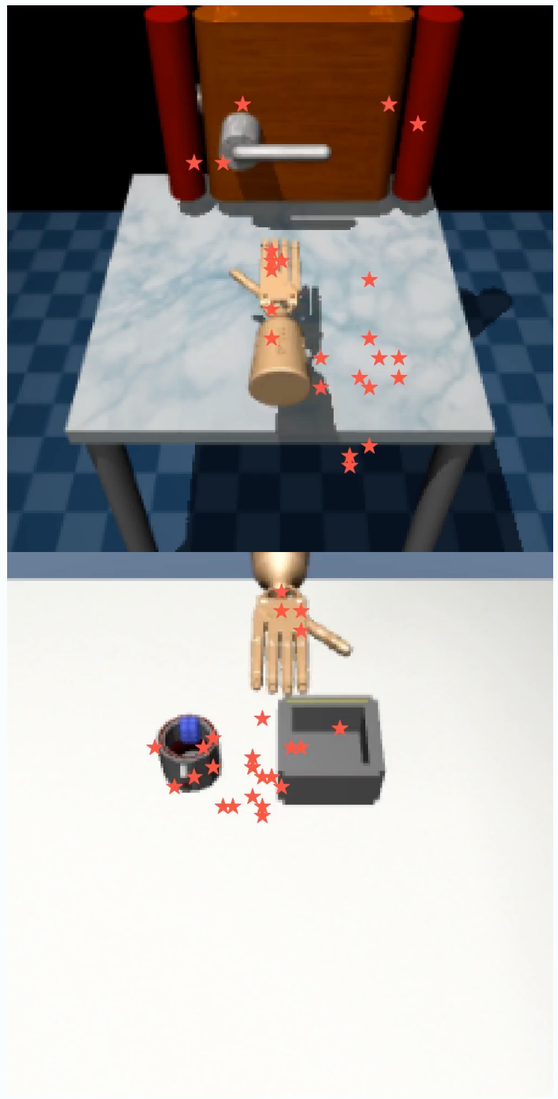


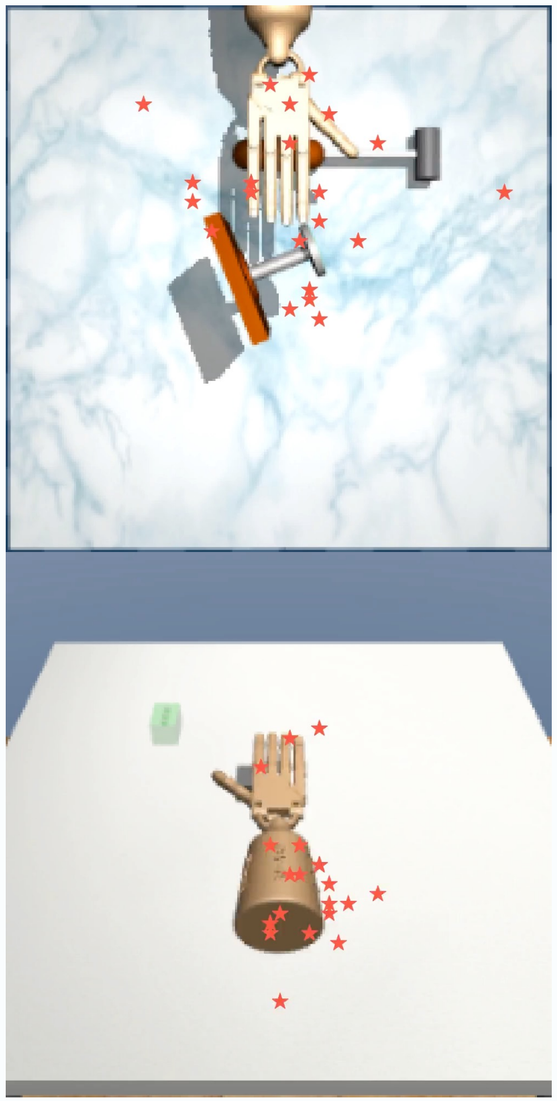
Strong Results in 12 Dexterous Manipulation Tasks

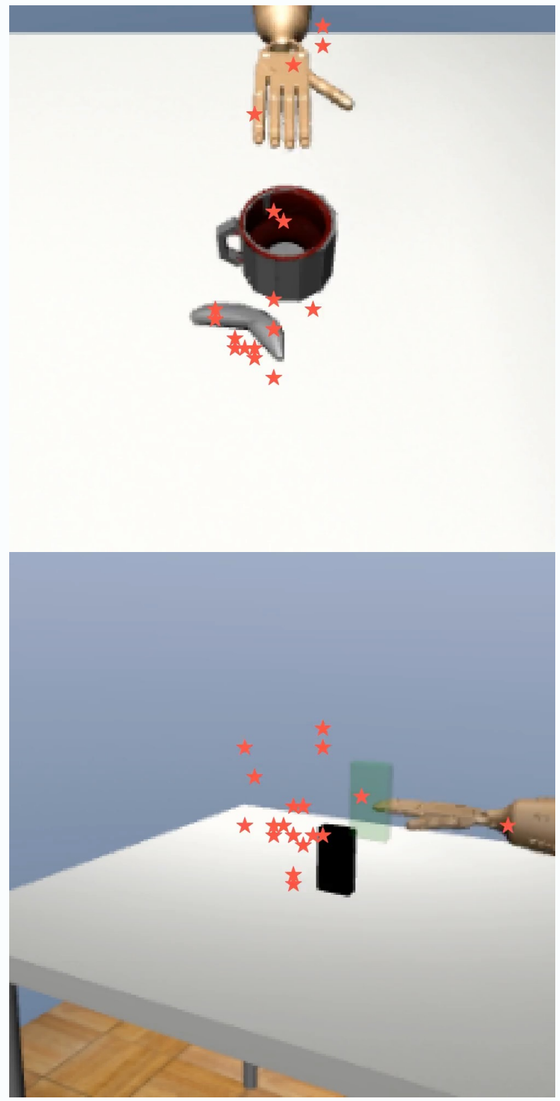




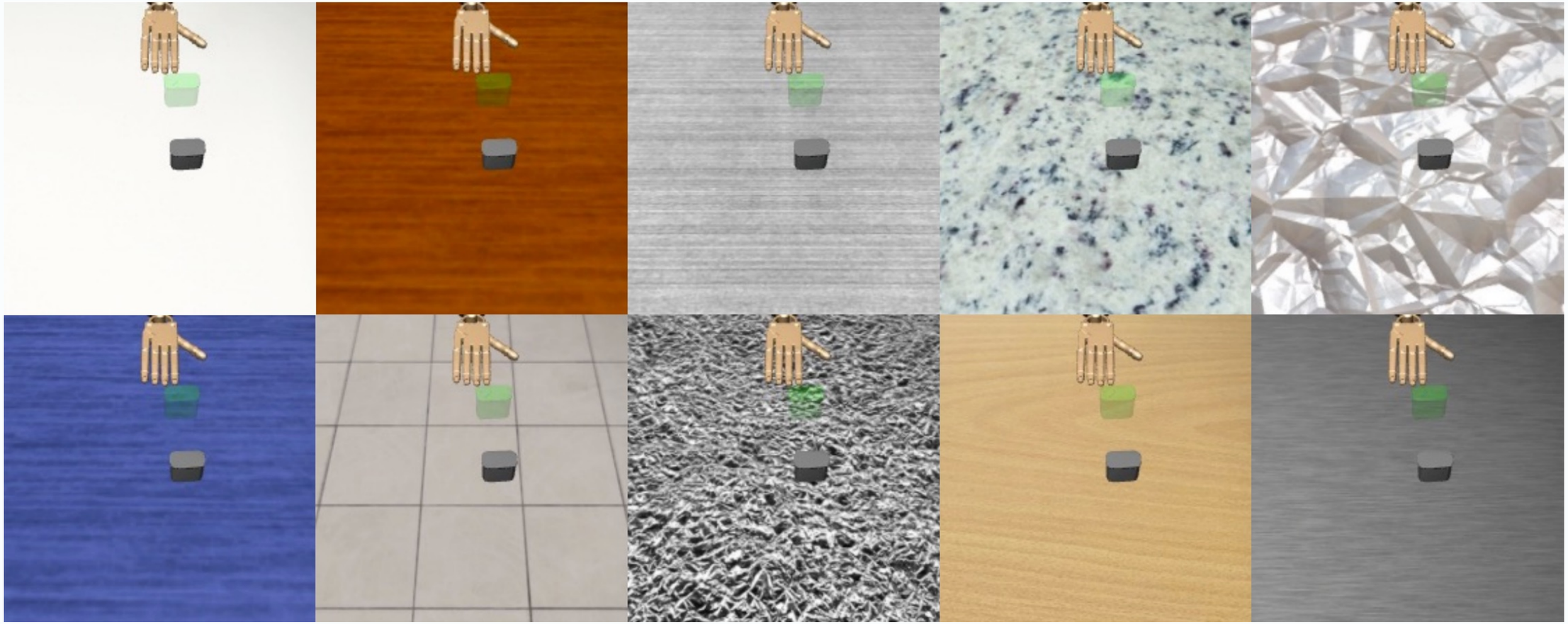
Self-Supervised Keypoint Detection



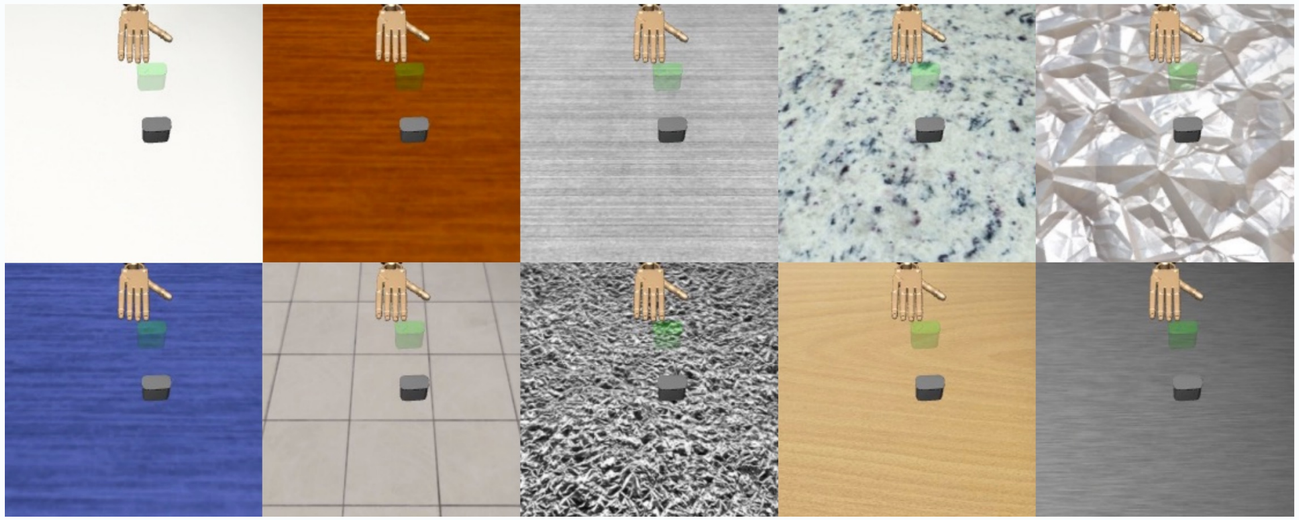




Generalize to Appearance



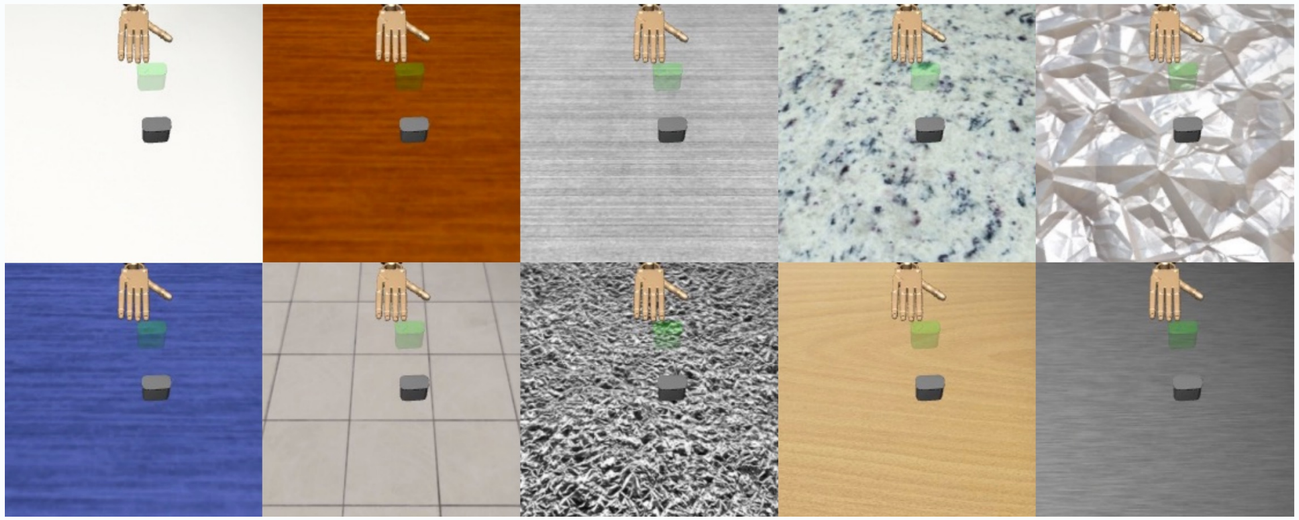
Generalize to Appearance



Scene ID / Method	VC-1 [17]	H-InDex
Origin	**2391.74**$_{\pm 602.83}$	2240.37$_{\pm 85.45}$

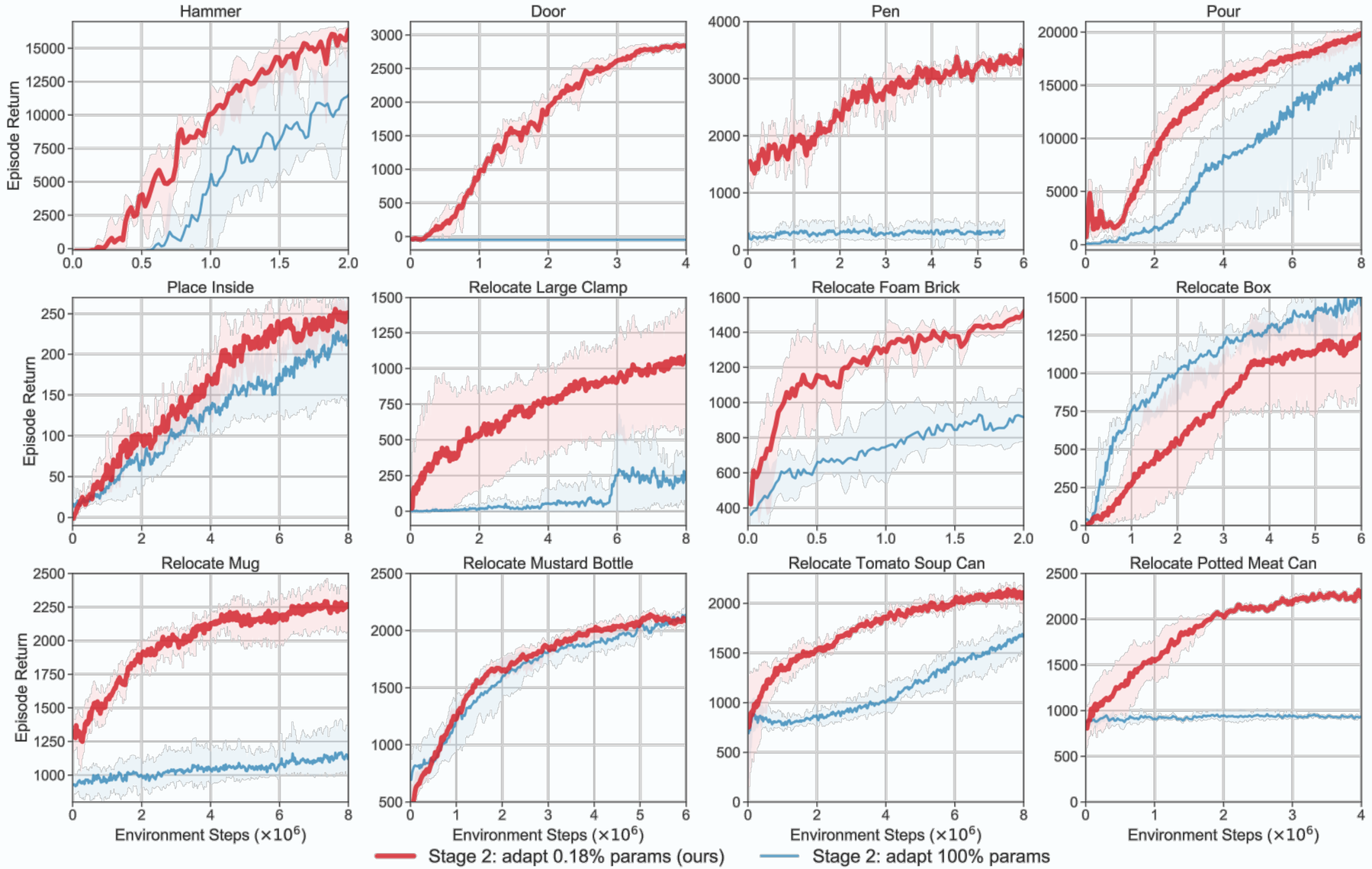
Generalize to Appearance

H-InDex helps visual generalization.



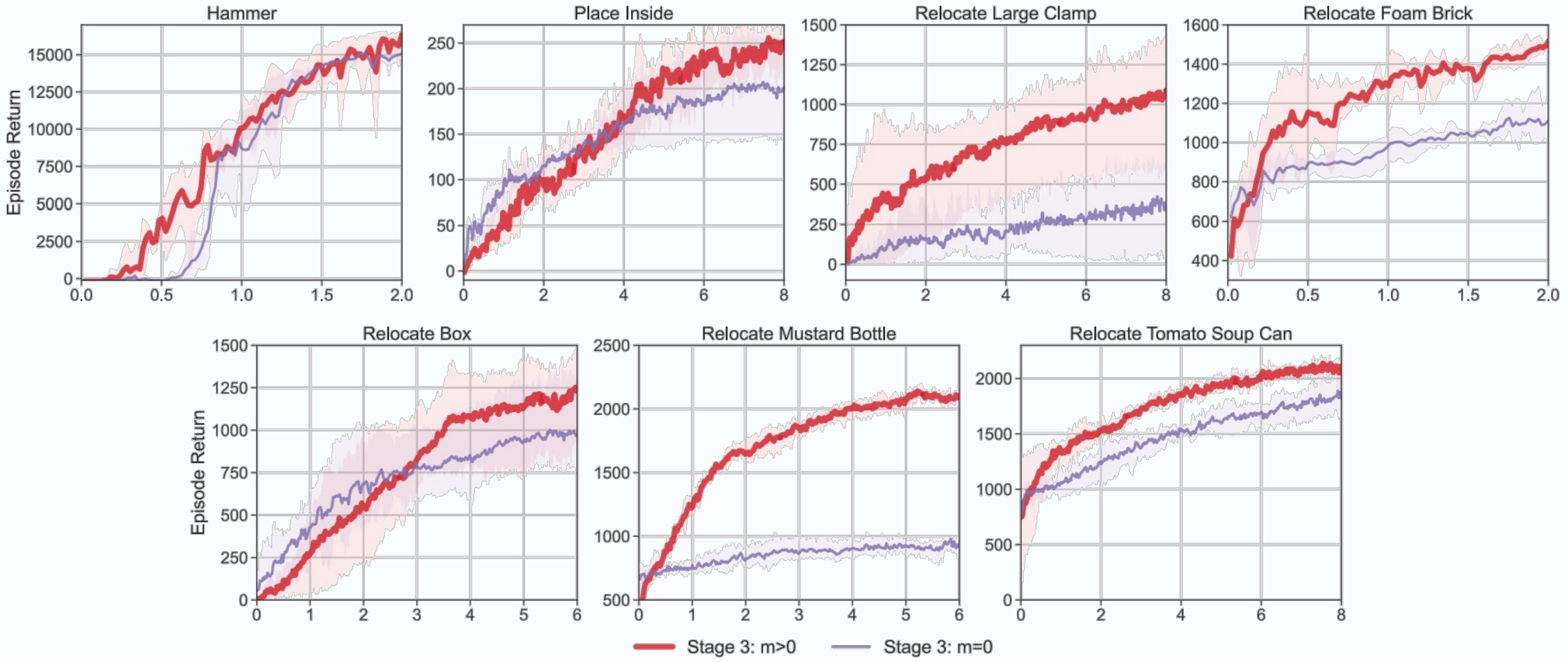
Scene ID / Method	VC-1 [17]	H-InDex
Origin	**2391.74**±602.83	2240.37±85.45
1	896.28±1006.55	**915.95**±922.65
2	603.26±920.48	**771.28**±793.51
3	451.36±839.45	**578.42**±764.09
4	360.21±772.64	**472.66**±715.03
5	300.02±718.05	**393.32**±676.41
6	256.80±673.16	**340.07**±639.68
7	224.21±635.56	**298.20**±608.54
8	226.59±610.58	**265.82**±581.03
9	214.30±581.76	**239.60**±556.80
Average	392.56	475.04

Full Finetuning in Stage 2 is Harmful



Updating BatchNorm Slightly during RL is Critical

$$\mu^{(i)} \leftarrow (1-m)\cdot\mu^{(i)} + m\cdot\mathbb{E}[x^{(i)}],$$
$$(\sigma^{(i)})^2 \leftarrow (1-m)\cdot(\sigma^{(i)})^2 + m\cdot\mathrm{Var}[x^{(i)}]$$



Visual Representations for Generalizable Robotic Manipulation

1

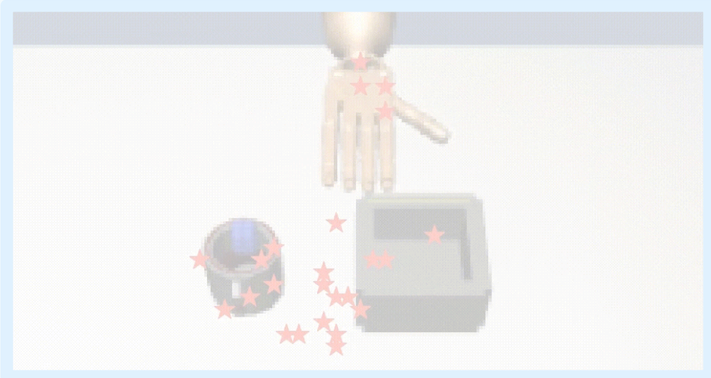
Geometric Prior

[2] **Ze** et al., “Visual Reinforcement Learning with Self-Supervised 3D Representations”, **RA-L** 2023 & **IROS** 2023.

[3] **Ze** et al., “GNFactor: Multi-Task Real Robot Learning with Generalizable Neural Feature Fields”, **CoRL** 2023 **Oral**.

2

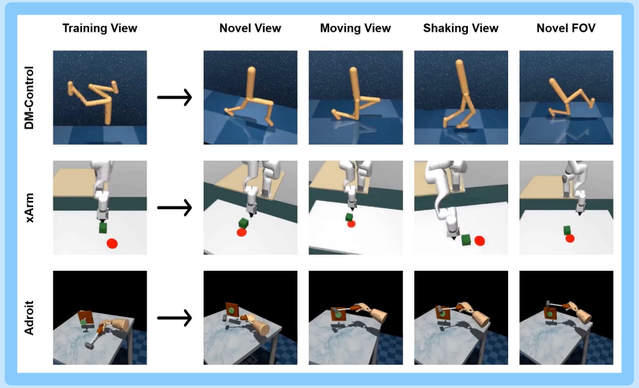
Human Prior



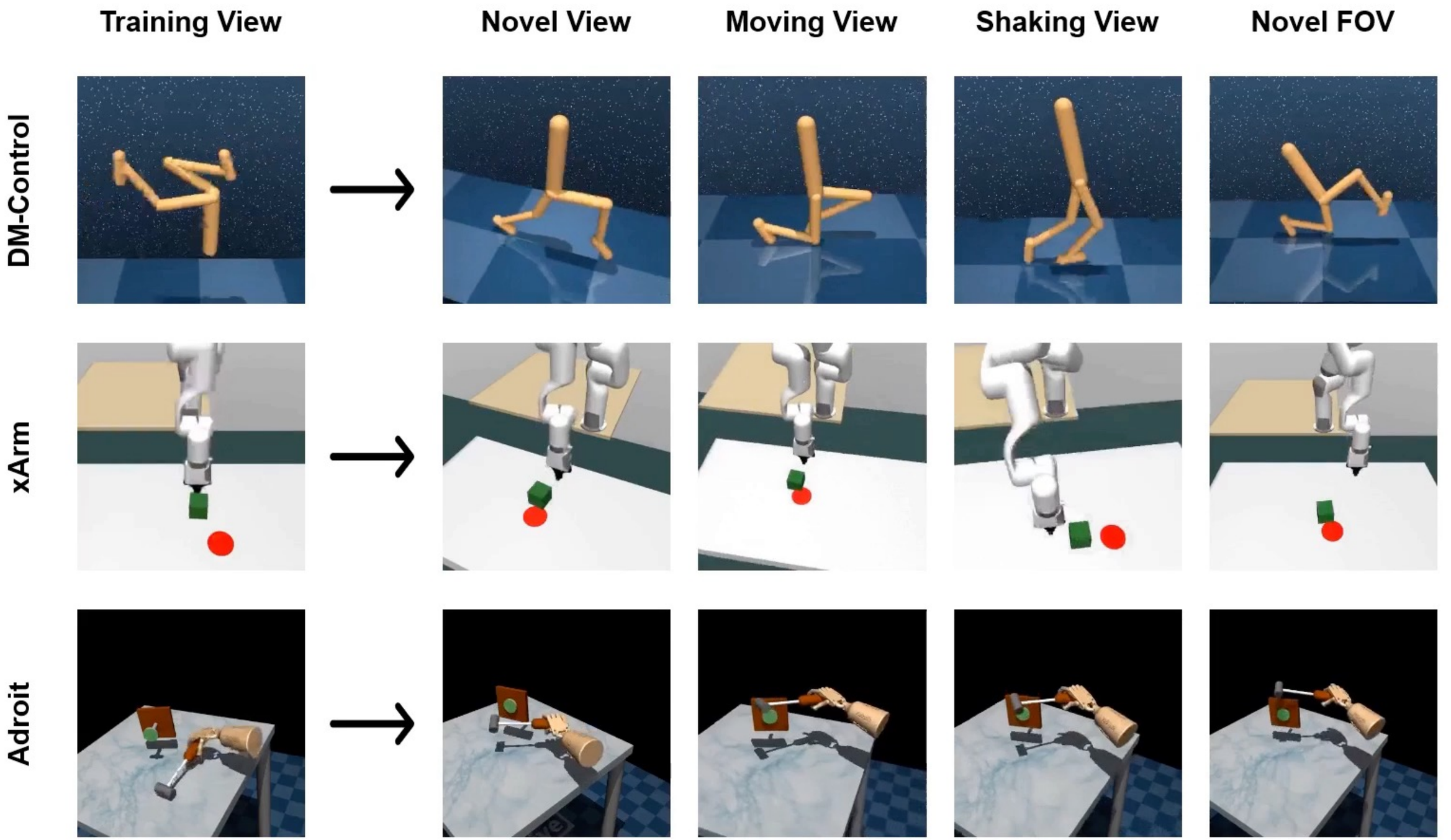
[4] **Ze** et al., *“H-InDex: Visual Reinforcement Learning with Hand-Informed Representations for Dexterous Manipulation”*, **NeurIPS** 2023.

3

Dynamics Prior



[5] Yang*, **Ze*** et al., *“MoVie: Visual Model-Based Policy Adaptation for View Generalization”*, **NeurIPS** 2023.


Training View
Novel View
Moving View
Shaking View
Novel FOV
DM-Control
xArm
Adroit

Only trained on one fixed camera view,
our agents generalize to diverse unseen views.

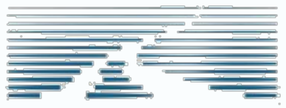
MoVie: Visual Model-Based Policy Adaptation for View Generalization

Sizhe Yang[12*] Yanjie Ze[13*] Huazhe Xu[415]

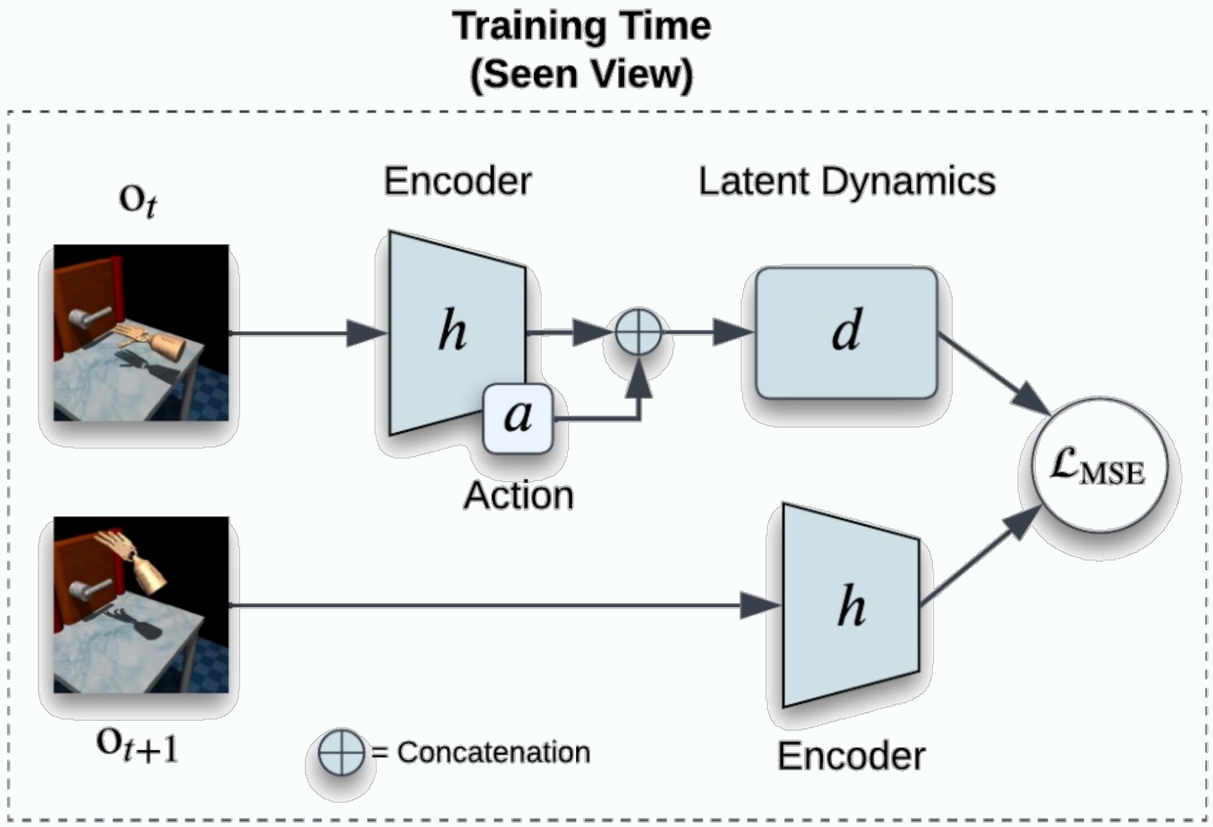
[1]Shanghai Qi Zhi Institute [2]UESTC [3]Shanghai Jiao Tong University
[4]Tsinghua University, IIIS [5]Shanghai AI Lab

NeurIPS 2023

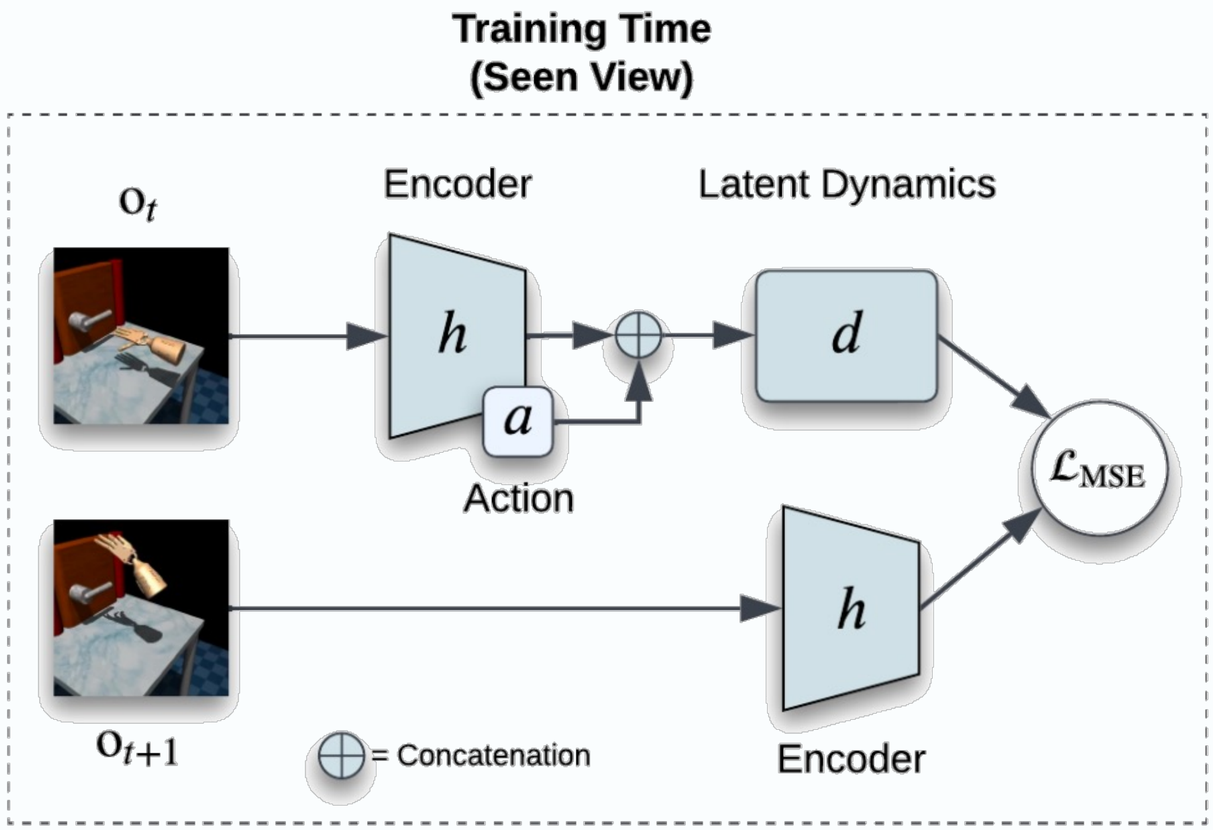


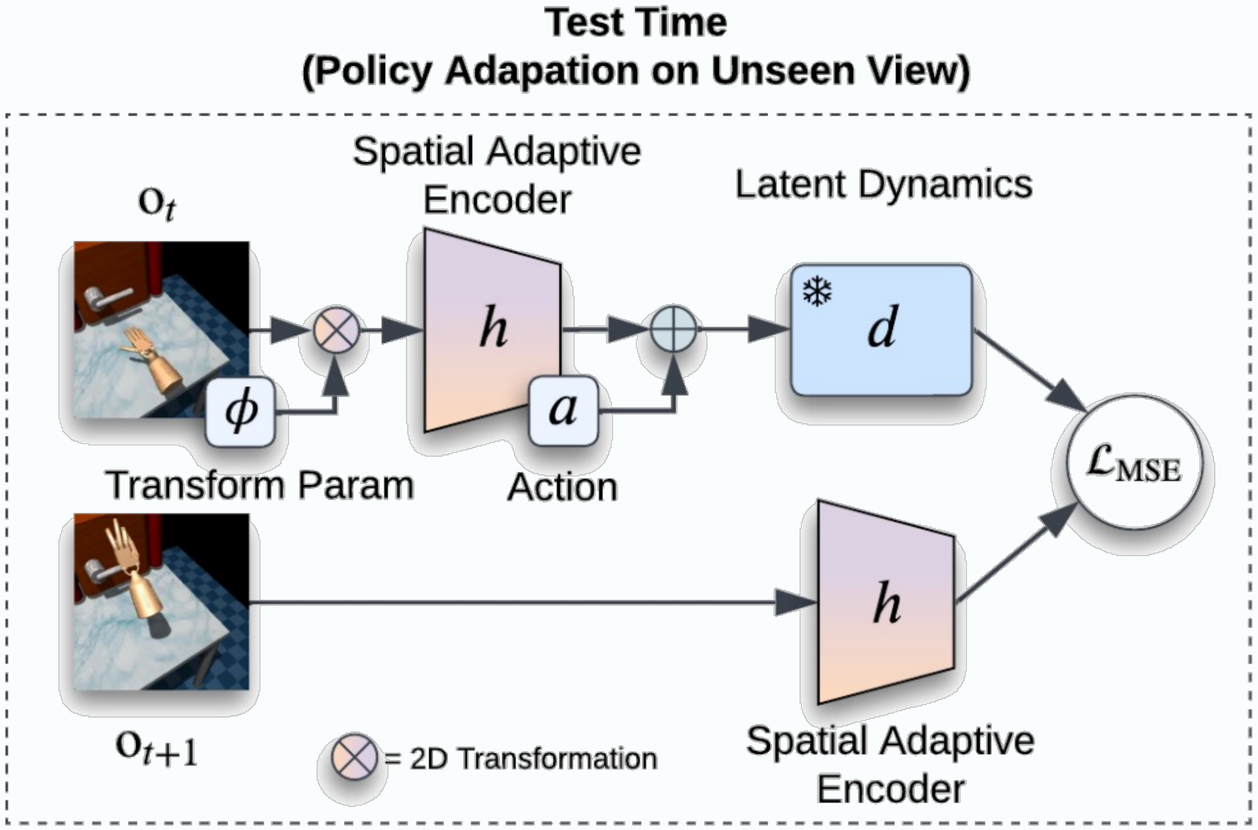


Test-Time Adaptation by Dynamics



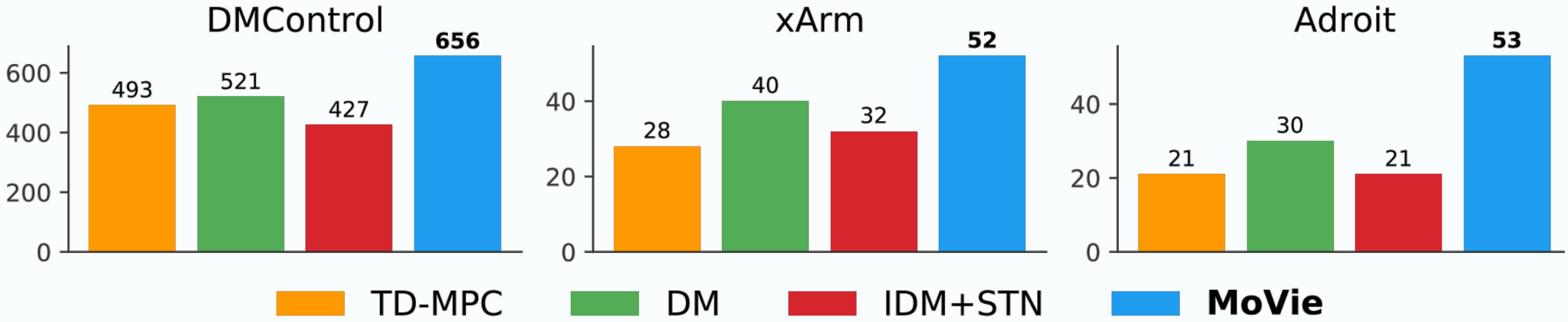
Test-Time Adaptation by Dynamics

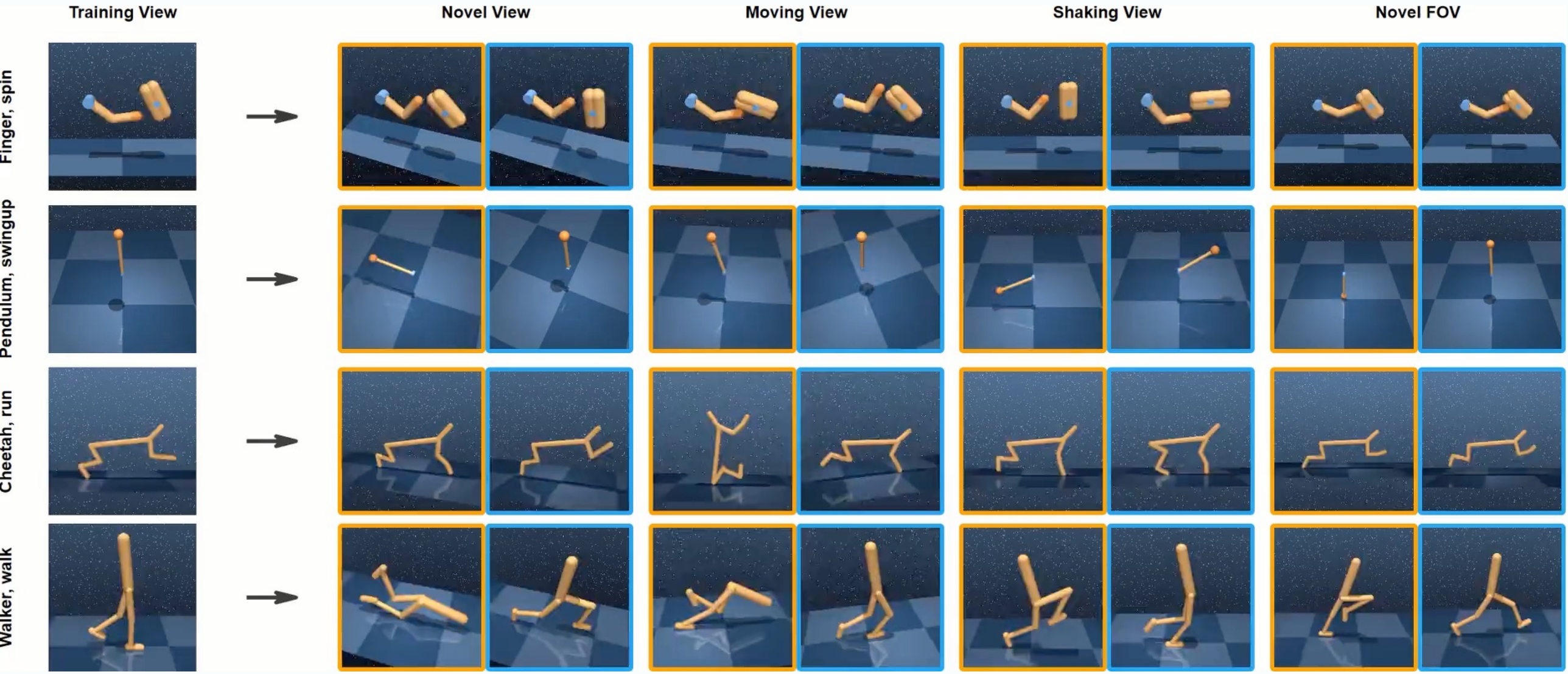


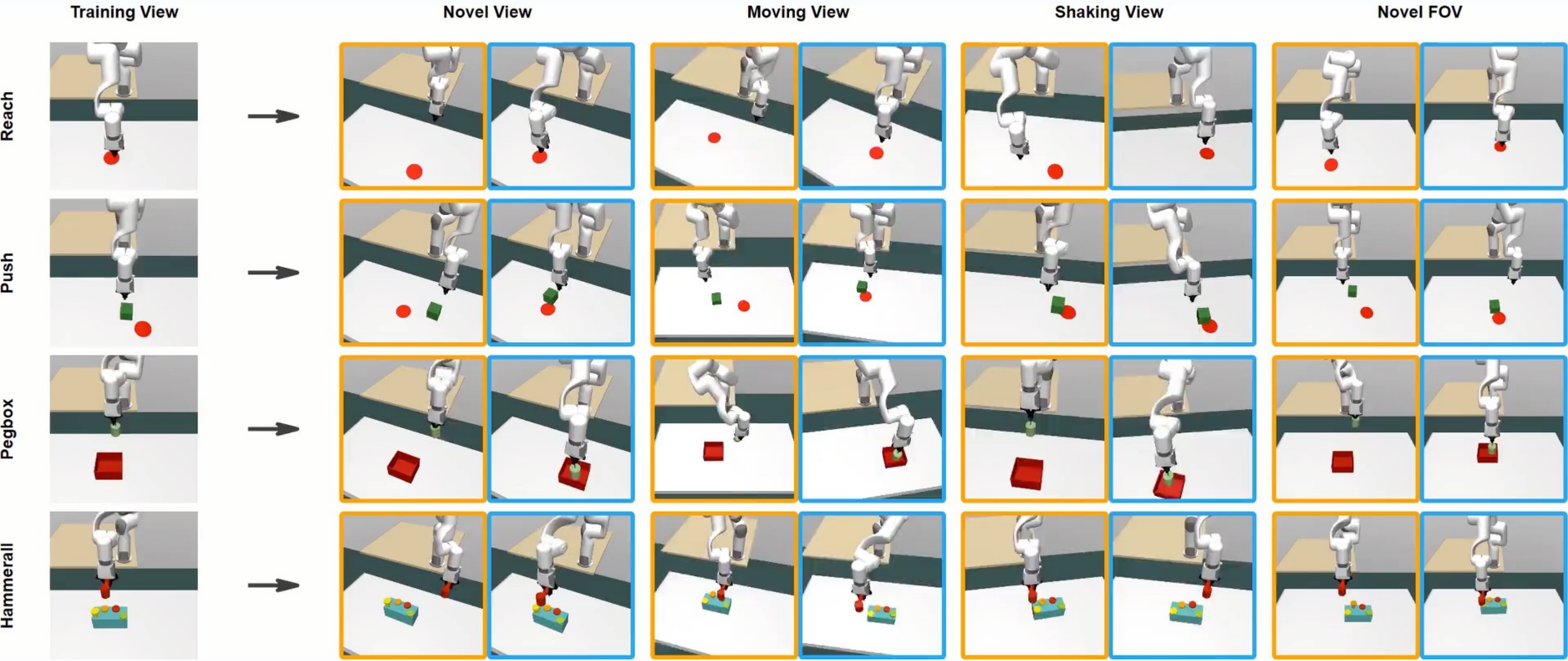


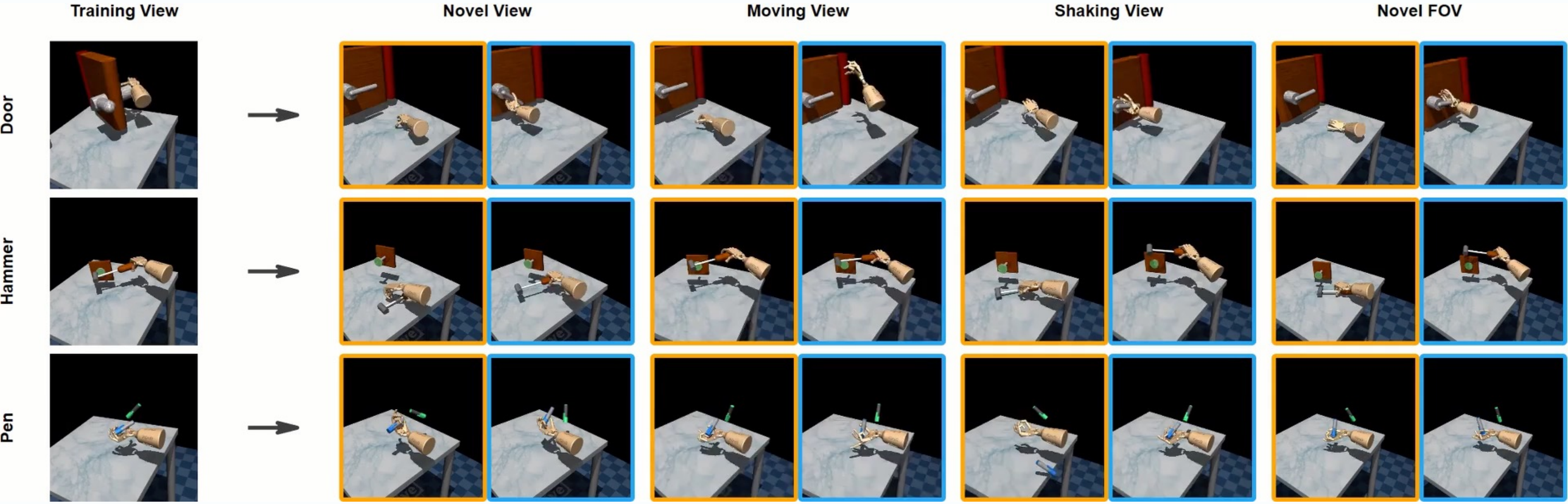
Experiments

18 tasks across 3 domains

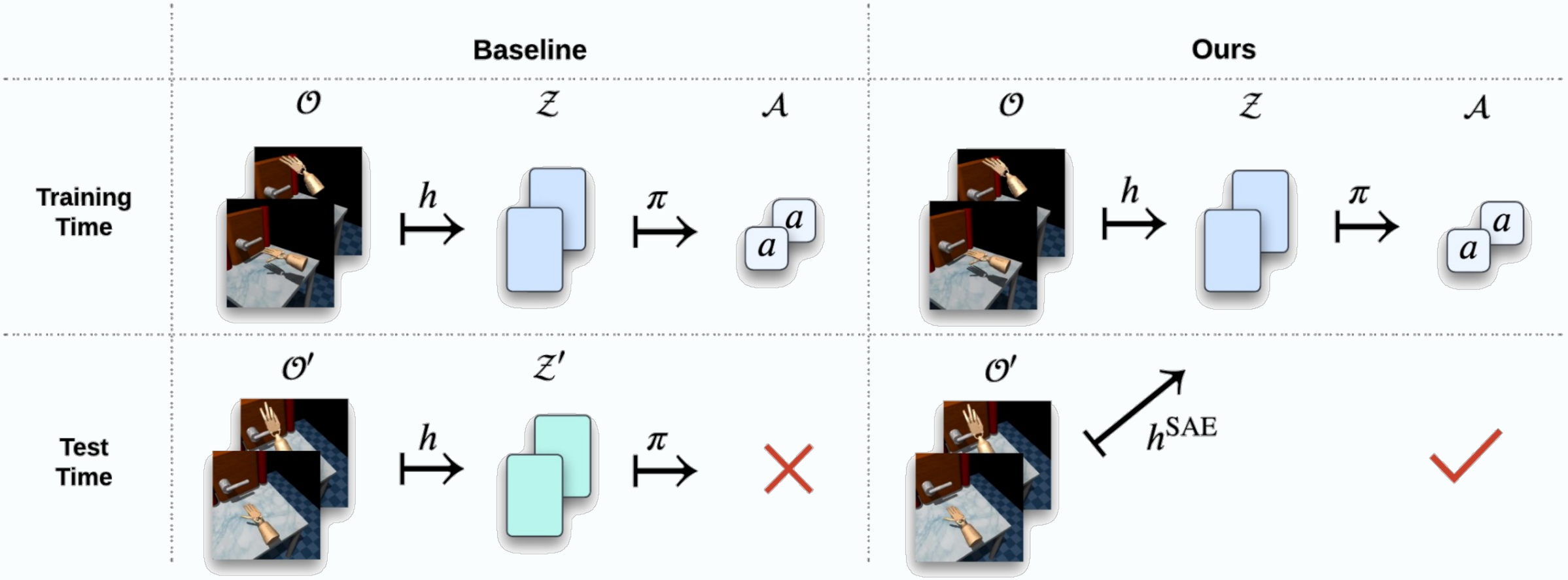



Training View
Novel View
Moving View
Shaking View
Novel FOV
Finger, spin
Pendulum, swingup
Cheetah, run
Walker, walk


Training View
Novel View
Moving View
Shaking View
Novel FOV
Reach
Push
Pegbox
Hammerall


Training View
Novel View
Moving View
Shaking View
Novel FOV
Door
Hammer
Pen

Explanation: Dynamics as Supervision



Explanation: Dynamics as Supervision

Fintuning dynamics → only use test-time information

Freezing dynamics → dynamics consistency between training and test

Cheetah-run	Novel View	Moving View	Shaking View	Novel FOV
MoVie	**342.39**$\pm$**54.95**	**365.22**$\pm$**42.66**	**493.54**$\pm$**56.80**	532.94$\pm$19.74
Finetune DM	273.01$\pm$11.29	331.55$\pm$22.22	476.25$\pm$30.25	**561.94**$\pm$**37.73**

Visual Representations for Generalizable Robotic Manipulation

1

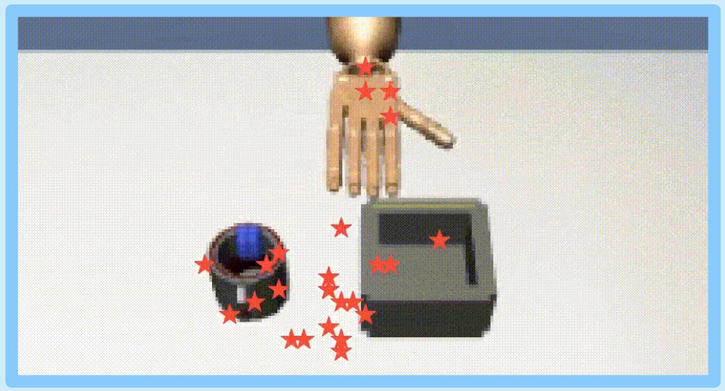
Geometric Prior

[2] **Ze** et al., “Visual Reinforcement Learning with Self-Supervised 3D Representations”, **RA-L** 2023 & **IROS** 2023.

[3] **Ze** et al., “GNFactor: Multi-Task Real Robot Learning with Generalizable Neural Feature Fields”, **CoRL** 2023 **Oral**.

2

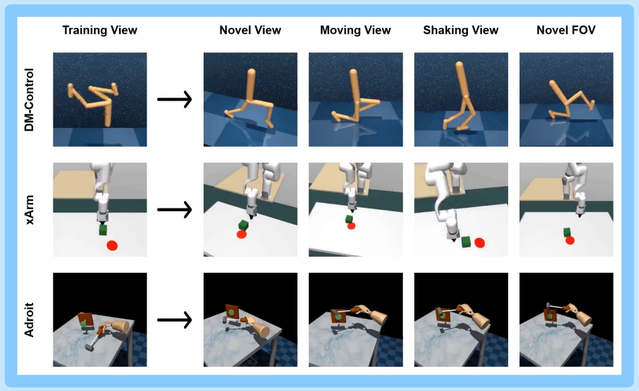
Human Prior



[4] **Ze** et al., *“H-InDex: Visual Reinforcement Learning with Hand-Informed Representations for Dexterous Manipulation”*, **NeurIPS** 2023.

3

Dynamics Prior



[5] Yang*, **Ze*** et al., *“MoVie: Visual Model-Based Policy Adaptation for View Generalization”,* **NeurIPS** 2023.

More info

https://yanjieze.com


To Be Continued