Unified World Models: Coupling Video and Action Diffusion for Pretraining on Large Robotic Datasets

Robotics: Science and Systems (*RSS*) 2025

Imitation Learning Is Bottlenecked by Action Labels

Behavior Cloning (BC): learns a policy from **human operator (observation → action) pairs**.

🤦**Two problems at scale:**

- **Expensive and slow:** every demo is teleoperated on a real robot.
- **Brittle OOD:** the policy learns no world dynamics.

On the contrary, **Video** of people and robots is cheap and abundant.

Problem: *the cheapest data is the data BC cannot use.*

Action-labeled robot data

Contains the actions BC imitates.
Tied to one robot embodiment.
This paper: **2,000 DROID trajectories**.

Action-free video

Diverse scenes and behaviors.
Shows how objects move and interact.
No action labels to imitate.

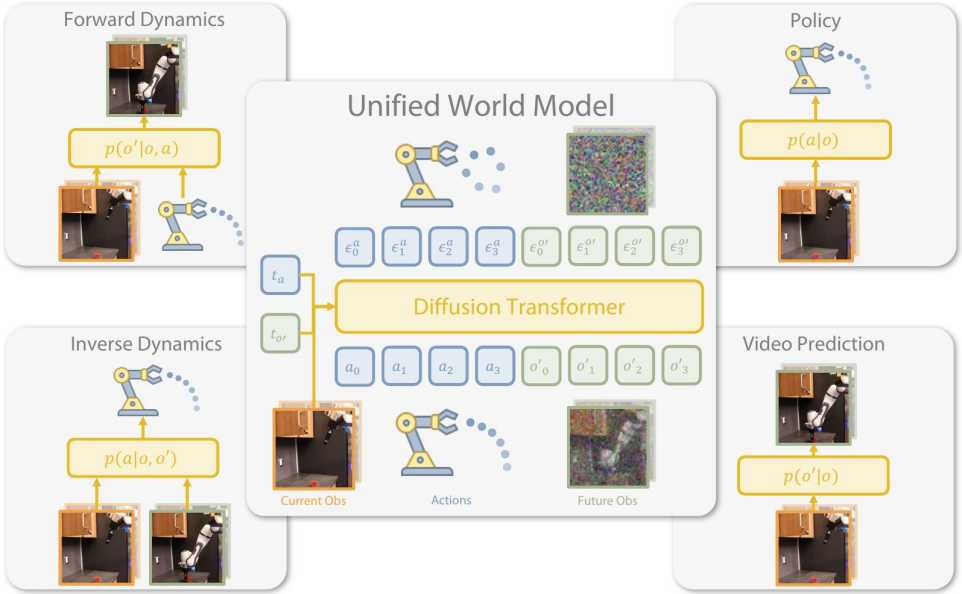
RQ: Can one model learn from both?

Existing Paradigms

Paradigm	Training data	Outputs	Limitation
Imitation learning *(BC, Diffusion Policy)*	Action-labeled demos	Policy $p(a \mid o)$	No dynamics; brittle OOD
World models *(video diffusion)*	Robot data + raw video	Dynamics $p(o' \mid o, a)$	No direct controller
Joint video–action *(GR-1, PAD)*	Both, via a mask token	Policy + video	One shared timestep: joint generation only
UWM *(this paper)*	Both, via timesteps	Policy, dynamics, video	**Independent t_a, $t_{o'}$**: any conditional

Key Insight: Controlling Information with Diffusion Timesteps

- **Standard diffusion:** one timestep t shared by every variable.
- **UWM:** two independent timesteps, t_a for the action and $t_{o'}$ for the next frame.
- **t = T** (pure noise): the variable is masked out, i.e. marginalized.
- **t = 0** (clean data): the variable is given as input, i.e. conditioned on.



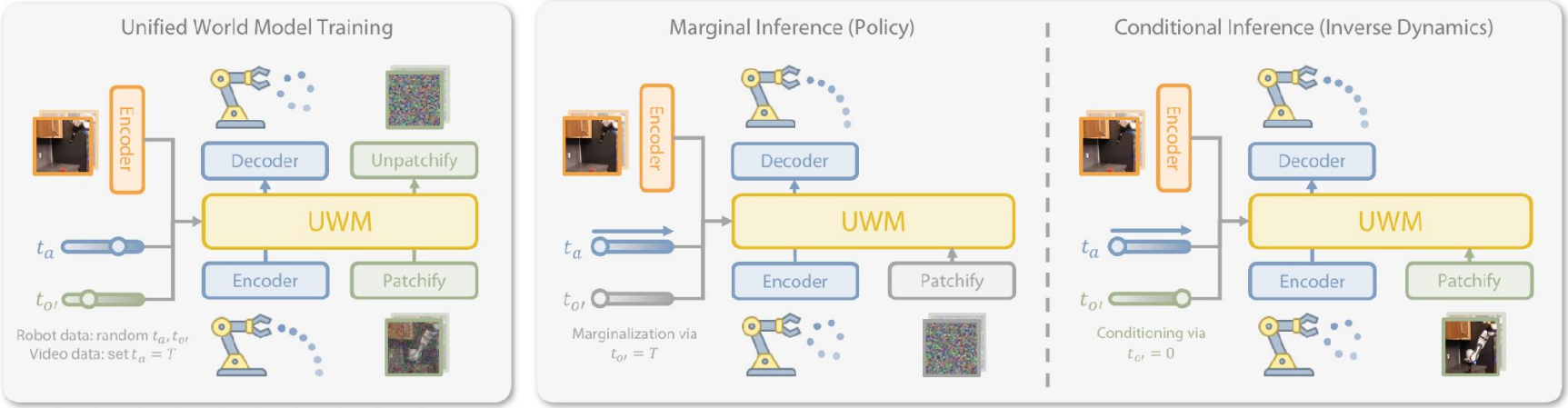
Choosing (t_a, $t_{o'}$) selects one of **four inference modes** from one network, with no extra modules or loss terms.

One Network, Four Inference Modes

Inference mode	Distribution	Action timestep t_a	Observation timestep $t_{o'}$	Denoises
Policy	$p(a \mid o)$	$T \to 0$	fixed at **T** (marginalized)	actions
Video prediction	$p(o' \mid o)$	fixed at **T** (marginalized)	$T \to 0$	future frame
Forward dynamics	$p(o' \mid o, a)$	fixed at **0** (conditioned)	$T \to 0$	future frame
Inverse dynamics	$p(a \mid o, o')$	$T \to 0$	fixed at **0** (conditioned)	actions

Pin a timestep to **T** to marginalize that variable, or to **0** to condition on it.

Training: Independent Timestep Sampling

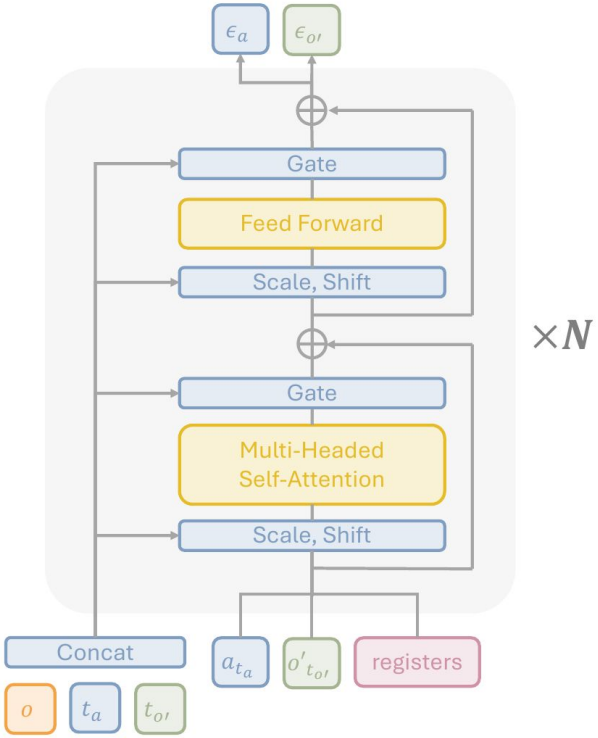


$$\ell(\theta) = E[w_a \|\varepsilon^{\theta}_a - \varepsilon_a\|^2 + w_{o'} \|\varepsilon^{\theta}_{o'} - \varepsilon_{o'}\|^2]$$

- Sample $t_a, t_{o'} \sim U(0, T)$ **independently** for every training example.
- The model sees every combination of known or unknown action and known or unknown future frame.

vs. PAD: one shared timestep only trains joint generation. Two timesteps train every conditional.

Architecture: A Diffusion Transformer with Registers



Backbone

- Diffusion Transformer (DiT) with AdaLN conditioning.

Conditioning (via AdaLN)

- **Current frames o:** ResNet-18 per camera, h_o = 2 frames.
- **Timesteps t_a, $t_{o'}$:** sinusoidal embeddings.

Tokens

- **Actions:** h_a = 16 steps, shallow MLP embedding.
- **Future frames:** frozen SDXL VAE, 224×224×3 → 28×28×4 latents, patch size (4, 4, 2).
- **Registers:** 8 learnable tokens, discarded at the output; shared scratch space for both modalities (Book-Caddy: 0.81 → 0.88).

Co-Training on Action-Free Video

For a video clip with no actions

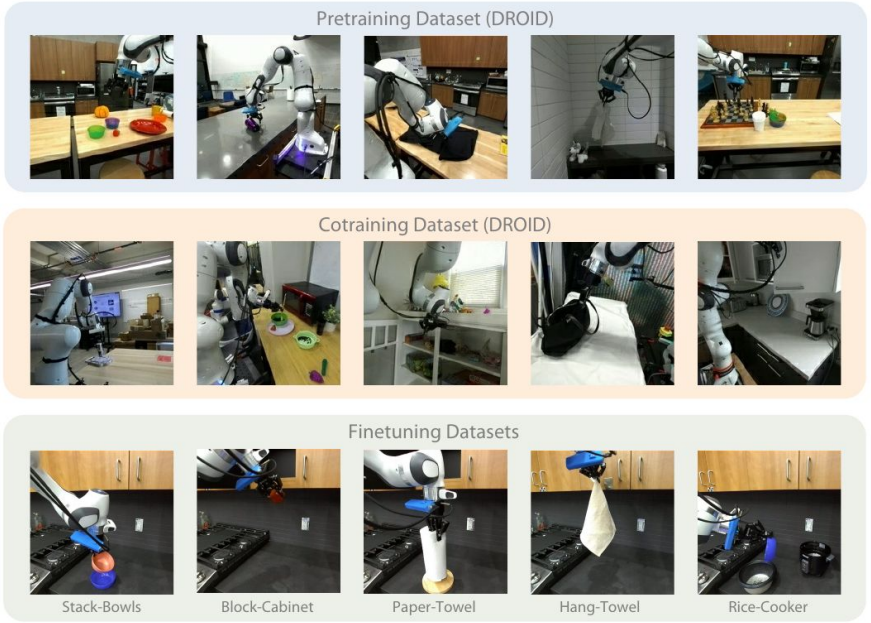
1. Fill the action slots with noise and pin $t_a = T$.
2. Train with the same denoising loss.

Compared with the baselines

- No pseudo-labels from an inverse-dynamics model.
- No learned mask token (GR-1 and PAD use one).

Data mix

- 2,000 action-labeled DROID trajectories plus a disjoint 2,000 with actions removed, mixed in every batch.



Experimental Setup: Real Franka + LIBERO

Pretraining

- 2,000 DROID trajectories, 100K steps.
- Co-training adds 2,000 action-free DROID videos.

Real robot (Franka)

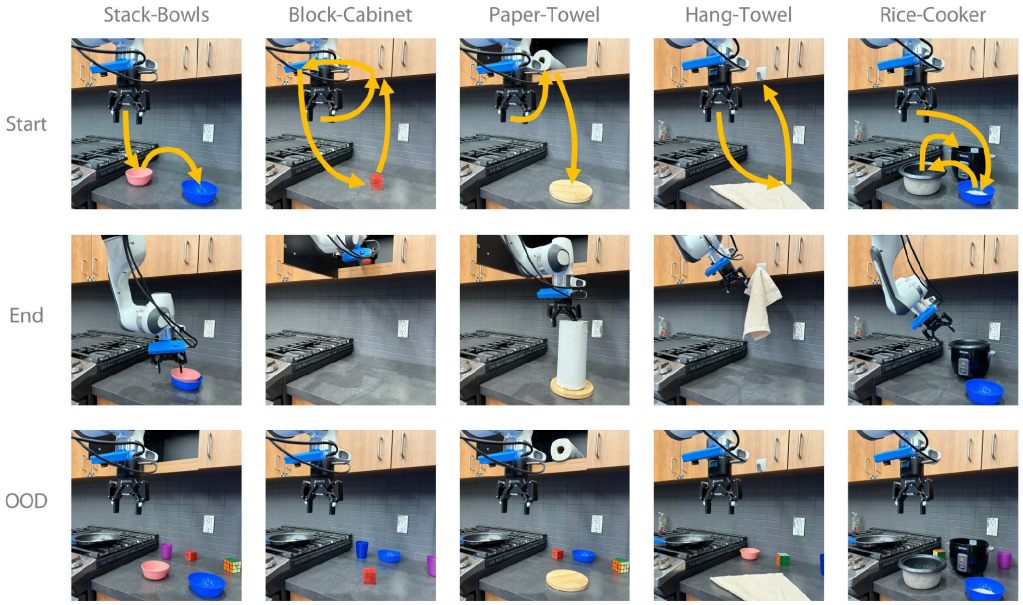
- 5 tasks, 50–150 demos each.
- Tested in-distribution (ID) and with unseen distractors (OOD).

Simulation

- LIBERO-90 pretraining → 5 LIBERO-10 tasks.

Baselines (re-implemented by the authors)

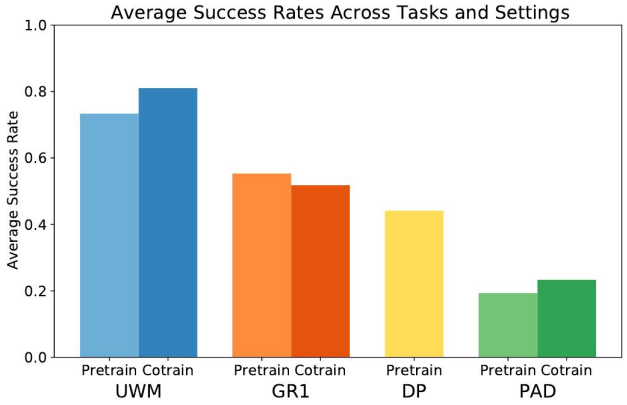
- **DP:** action-only diffusion.
- **PAD:** joint diffusion, one shared timestep.
- **GR-1:** joint prediction by regression.



Real-Robot Results: UWM Leads on All Five Tasks

TABLE I
EVALUATION RESULTS ACROSS REAL ROBOT TASKS (PRETRAIN / COTRAIN)

(Pretrain / Cotrain)	Stack-Bowls ID	Stack-Bowls OOD	Block-Cabinet ID	Block-Cabinet OOD	Paper-Towel ID	Paper-Towel OOD	Hang-Towel ID	Hang-Towel OOD	Rice-Cooker ID
UWM (Ours)	**0.86 / 0.92**	**0.76 / 0.84**	**0.76 / 0.84**	**0.60 / 0.72**	**0.78 / 0.86**	**0.78 / 0.84**	**0.82 / 0.86**	**0.64 / 0.76**	**0.60 / 0.65**
DP	0.48 / –	0.36 / –	0.60 / –	0.26 / –	0.52 / –	0.48 / –	0.64 / –	0.28 / –	0.35 / –
PAD	0.08 / 0.20	0.08 / 0.12	0.00 / 0.00	0.00 / 0.00	0.42 / 0.42	0.34 / 0.44	0.52 / 0.54	0.30 / 0.38	0.00 / 0.00
GR1	0.66 / 0.62	0.48 / 0.38	0.66 / 0.74	0.44 / 0.64	0.60 / 0.46	0.60 / 0.46	0.66 / 0.66	0.48 / 0.44	0.40 / 0.25



- **Best on every task, ID and OOD:** up to +0.20 in-distribution, +0.28 OOD (Stack-Bowls).
- **Video co-training helps UWM everywhere:** all 9 entries improve; mean 0.73 → 0.81.
- **GR-1:** co-training hurts on 3 of 5 tasks (Stack-Bowls, Paper-Towel, Rice-Cooker).
- **PAD:** 0.00 on Block-Cabinet and Rice-Cooker; the authors attribute this to raw-pixel conditioning.

Video Co-Training Improves OOD Robustness

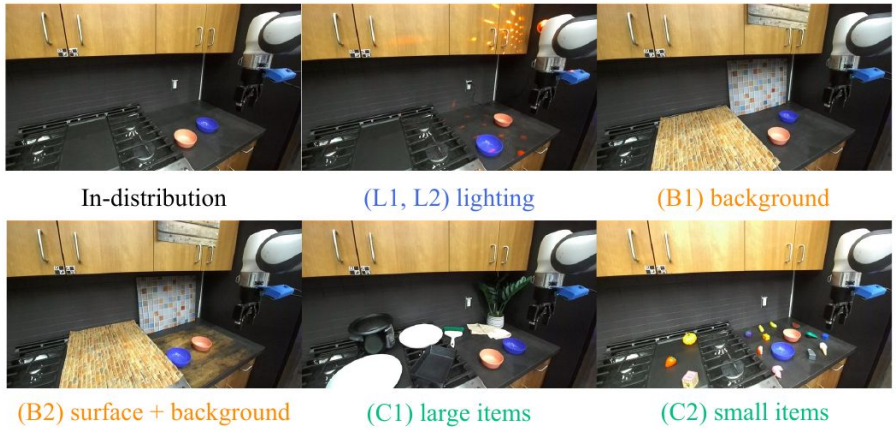


TABLE IV
OOD PERFORMANCE ON STACK-BOWLS AND BLOCK-CABINET

	Stack-Bowls			Block-Cabinet		
	UWM (Co)	UWM (Pre)	DP	UWM (Co)	UWM (Pre)	DP
L1	4/5	4/5	2/5	5/5	5/5	3/5
L2	3/5	2/5	2/5	4/5	0/5	0/5
B1	4/5	3/5	3/5	4/5	3/5	2/5
B2	3/5	1/5	2/5	1/5	0/5	0/5
C1	3/5	2/5	2/5	0/5	0/5	0/5
C2	4/5	3/5	1/5	1/5	0/5	0/5
All	**21/30**	**15/30**	**12/30**	**15/30**	**8/30**	**6/30**

6 OOD shifts: lighting (L1, L2), background (B1, B2), clutter (C1, C2); 5 trials each.

Total successes out of 30

- **Stack-Bowls:** UWM co-trained 21, pretrained 15, DP 12.
- **Block-Cabinet:** UWM co-trained 15, pretrained 8, DP 6.

Action-free video nearly doubles OOD success on Block-Cabinet (8 → 15) with **no extra action labels**.

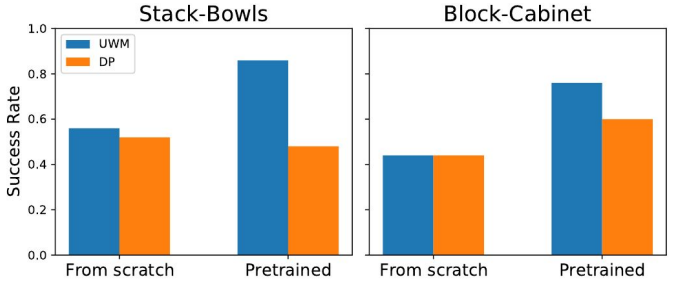
LIBERO Results and Scaling with Pretraining

LIBERO (simulation)

TABLE II
EVALUATION ON LIBERO BENCHMARK.

	Book-Caddy	Soup-Cheese	Bowl-Drawer	Moka-Moka	Mug-Mug	Average
UWM (Ours)	**0.91** ± 0.07	**0.93** ± 0.01	**0.80** ± 0.02	**0.68** ± 0.02	**0.65** ± 0.01	**0.79** ± 0.11
DP	0.73 ± 0.10	0.88 ± 0.02	0.77 ± 0.02	0.65 ± 0.03	0.53 ± 0.05	0.71 ± 0.12
PAD	0.78 ± 0.04	0.47 ± 0.04	0.74 ± 0.05	0.59 ± 0.08	0.25 ± 0.04	0.57 ± 0.19
GR1	0.77 ± 0.03	0.65 ± 0.05	0.62 ± 0.03	0.46 ± 0.04	0.38 ± 0.05	0.58 ± 0.14

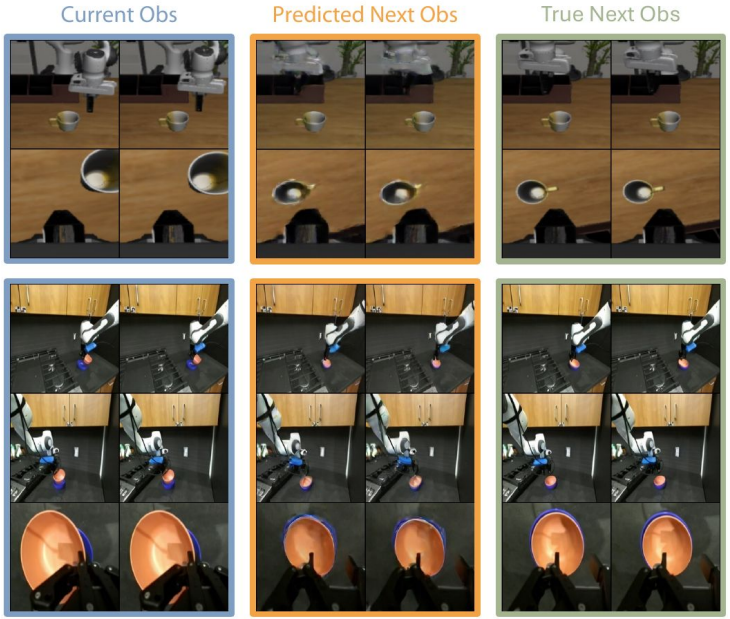
Real robot: from scratch vs. pretrained



- **LIBERO average:** UWM 0.79, DP 0.71, GR-1 0.58, PAD 0.57.
- DP is second in simulation; on the real robot, GR-1 was.
- Smaller OOD margin than on the real robot; the authors suggest simpler simulated dynamics.
- **From scratch, UWM ≈ DP.** After pretraining, UWM pulls ahead (Stack-Bowls 0.86 vs. 0.48).

UWM's advantage is a **pretraining** advantage: trained from scratch, it matches DP.

Evaluating World-Model Capabilities



Forward dynamics $p(o' \mid o, a)$

- Pin $t_a = 0$ with the true action; denoise o'.
- Predicted frames match the true robot and object poses.

Inverse dynamics $p(a \mid o, o')$

- Pin $t_{o'} = 0$ on a reference frame; denoise a.
- With the same time budget as the reference, it tracks better than the policy (0.65 vs. 0.47, Book-Caddy).

TABLE III
TRAJECTORY TRACKING EXPERIMENTS

	Book-Caddy	Soup-Cheese
Policy (1000 steps)	1.00 $\pm$ 0.00	0.97 $\pm$ 0.01
Policy (trajectory length)	0.47 $\pm$ 0.02	0.26 $\pm$ 0.02
Inverse dynamics (trajectory length)	0.65 $\pm$ 0.01	0.55 $\pm$ 0.02

Ablations: Key Design Choices

TABLE VII
ABLATION OF DESIGN CHOICES

	Book-Caddy	Soup-Cheese
UWM w/ 8 registers	**0.88** ± 0.04	**0.90** ± 0.02
UWM w/ 4 registers	0.83 ± 0.05	0.86 ± 0.03
UWM w/o registers	0.81 ± 0.07	0.85 ± 0.03
Cross attention UWM	0.78 ± 0.05	0.86 ± 0.04

Registers and AdaLN help

- 8 / 4 / 0 registers: 0.88 / 0.83 / 0.81
- AdaLN 0.88 vs. cross-attention 0.78

LIBERO Book-Caddy

TABLE VIII
ABLATION OF LEARNING OBJECTIVES

	Stack-Bowls	Block-Cabinet
UWM Reconstruct Future Obs	0.86	0.76
UWM Reconstruct Current Obs	0.70	0.66
DP (No Reconstruction)	0.48	0.60

Predicting the future matters most

- Next frame: 0.86 / 0.76
- Current frame: 0.70 / 0.66
- No image prediction (DP): 0.48 / 0.60

Stack-Bowls / Block-Cabinet

TABLE IX
COTRAINING ON INTERNET VIDEOS

	Stack-Bowls	Block-Cabinet
UWM Robot Data + Robot Videos	0.92	0.84
UWM Robot Data + Internet Videos	0.88	0.80
UWM Robot Data	0.86	0.76

In-domain video helps most

- Robot video: 0.92 / 0.84
- Internet video: 0.88 / 0.80
- No video: 0.86 / 0.76

Stack-Bowls / Block-Cabinet

Dynamics is the key ingredient: predicting the next frame beats reconstructing the current one.

Thank You!