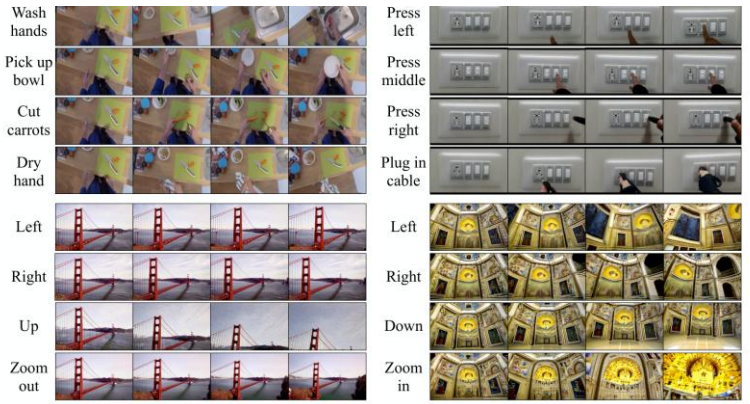
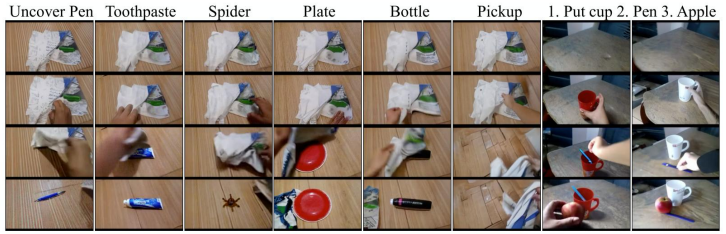
UniSim: Learning Interactive Real-World Simulators



Sherry Yang[12], Yilun Du[3], Kamyar Ghasemipour[2], Jonathan Tompson[2], Leslie Kaelbling[3], Dale Schuurmans[2], Pieter Abbeel[1]

[1] UC Berkeley [2] Google DeepMind [3] MIT

Outstanding Paper Award at ICLR 2024



Presented by Hyokyung Choi

November 17, 2026

Why learn a real-world simulator?

Realistic video ≠ interactive world

A simulator must respond to actions and preserve their consequences.

Generative video model

Creates plausible clips
from prompts or other conditions

Interactive simulator

Predicts what happens next
after each action

Why it matters: robot training in the real world is costly and risky.

What was missing in prior research?

PREVIOUS APPROACHES

VIDEO GENERATION

limited control
over repeated interactions

WORLD MODELS

Often one system,
compact states

TRAINING DATA

Images: diverse scenes, little motion
Robot data: controls, limited scale

UNISIM'S RESPONSE

Action-conditioned prediction
across successive video segments

Combine different environments
in one framework

Combine complementary datasets

What is UniSim? A visual world simulator

A learned model of what the camera might see after an action

RECENT FRAMES → **ACTION** → **UNISIM** → **NEXT VIDEO**

STATE | underlying world

Full physical situation

OBSERVATION | what the agent sees

Camera images or video frames

HIGH-LEVEL ACTION

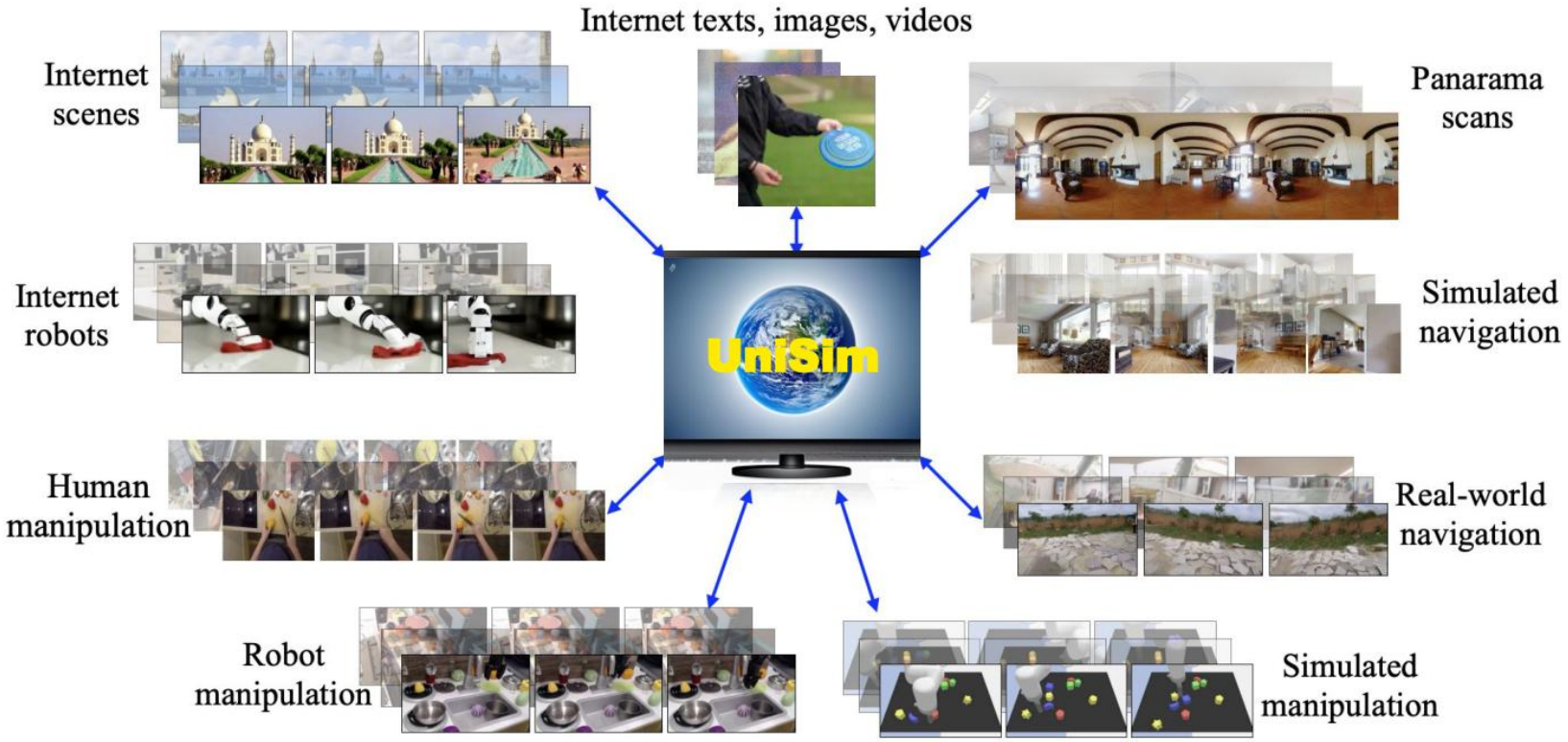
Language instruction

LOW-LEVEL ACTION

Specific motor control

A policy chooses the action. UniSim predicts its possible visual result.

Key idea: complementary datasets



Broad scene knowledge + action-rich data + precise controls

UniSim has 5.6 billion parameters and was trained on diverse complementary datasets for 20 days using 512 TPU-v3 chips.

One interface for different action formats

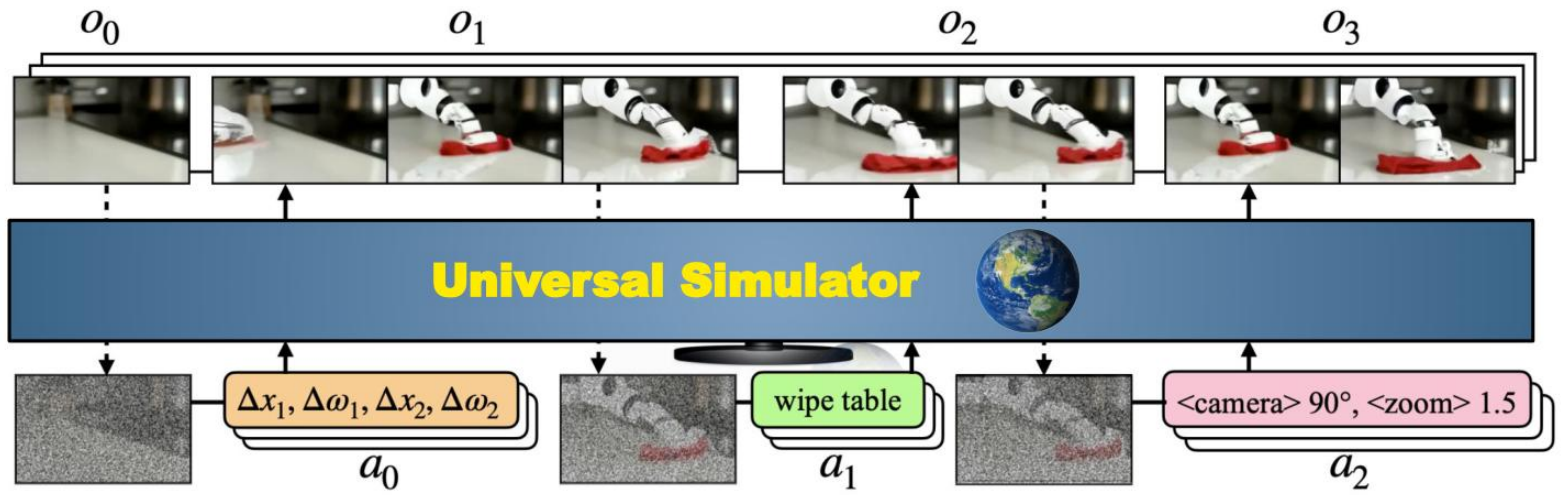
Language / activity labels	T5 text embeddings
Robot motor controls	Normalized + discretized values
Panorama camera poses	Navigation actions (e.g., “turn left”)
Static image + caption	One-frame video + text condition

Low-level controls are discretized into 4,096 bins (Appendix B).

How UniSim connects observations and actions

$$p(o_t \mid h_{t-1}, a_{t-1})$$

Predict the next video segment from recent frames and the chosen action



o_0, o_1, o_2, o_3 = observation segments | a_0, a_1, a_2 = actions between segments

Diffusion generates each video segment

TRAINING

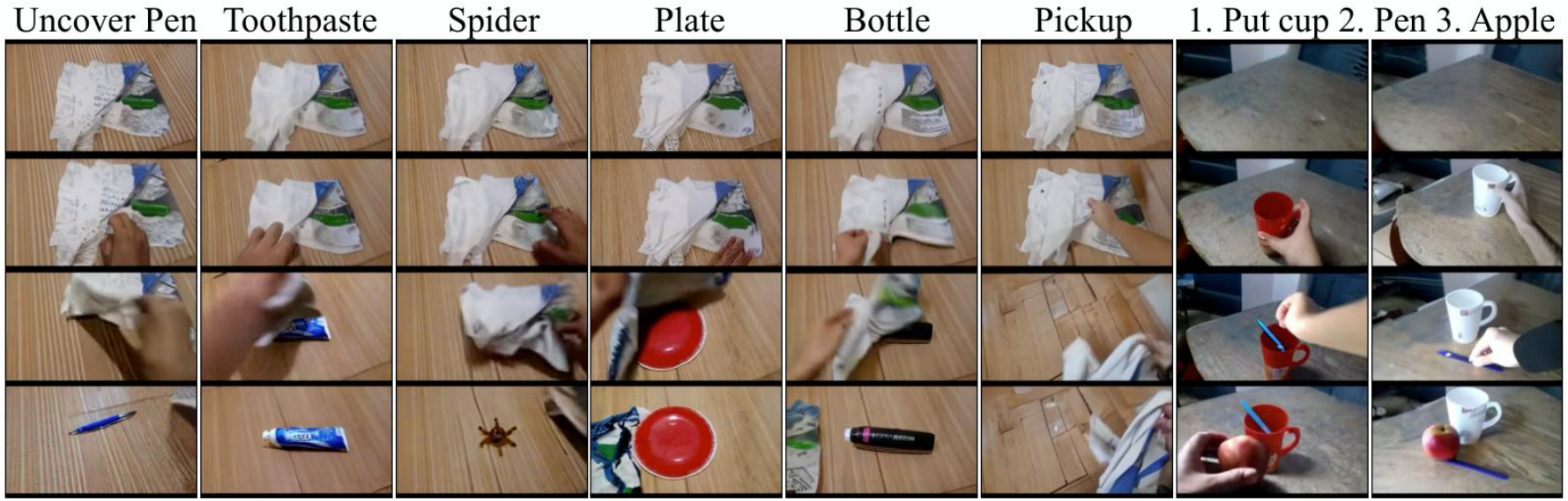
Real video + noise
→ predict the added noise

GENERATION

Start from noise
→ denoise into future video

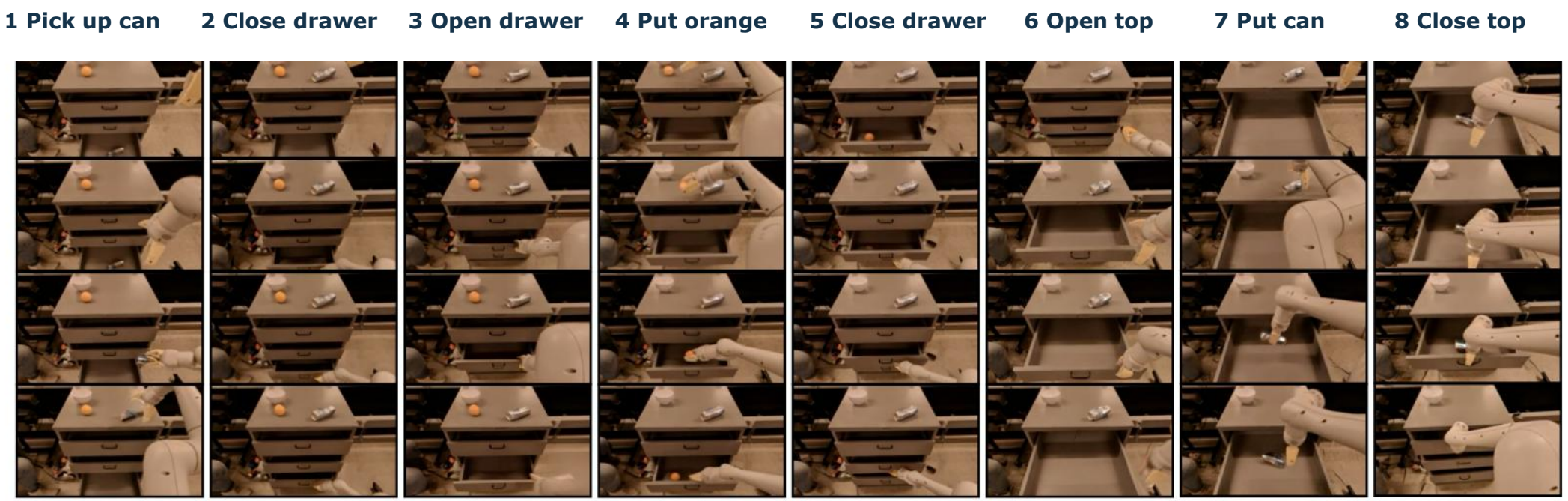
Denoising Network : U-Net • strengthen the action condition by using classifier-free guidance(CFG)

Left: specified objects revealed • Right: diverse results from actions



Eight actions: does the world stay consistent?

Eight connected interactions, read from left to right



The generated video preserves objects across multiple actions.

History ablation (FVD ↓): 1 frame 315.69 → 4 recent frames 211.30

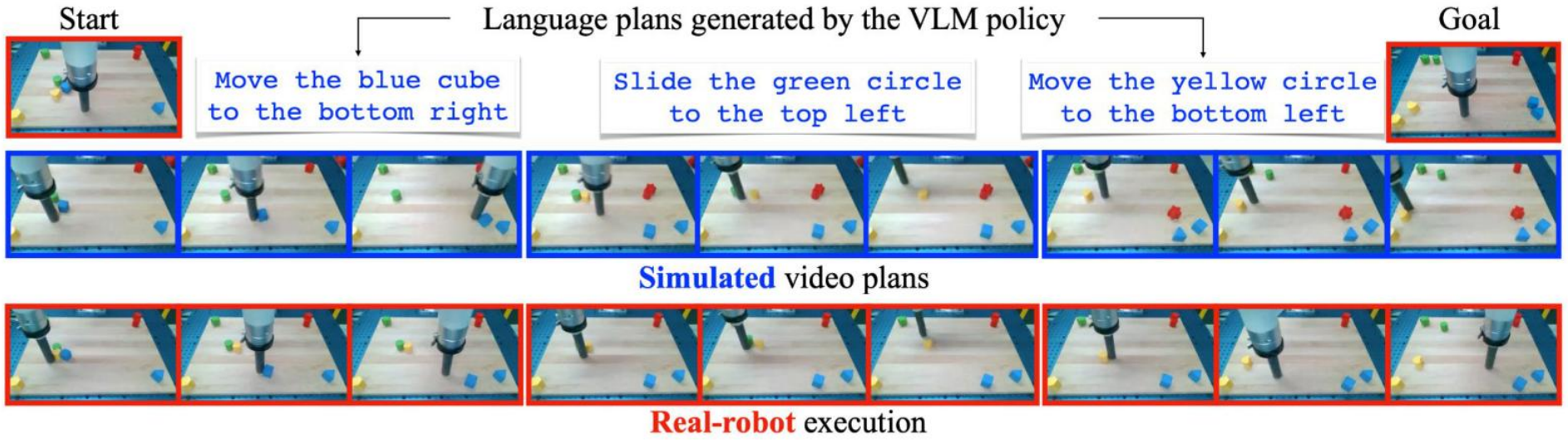
Simulating long sequence of robot executions.

Step 1:



Application 1: learning high-level decisions

10K simulated trajectories → hindsight goals → VLM policy



TOP: VLM language plan | MIDDLE: UniSim video plan | BOTTOM: real robot

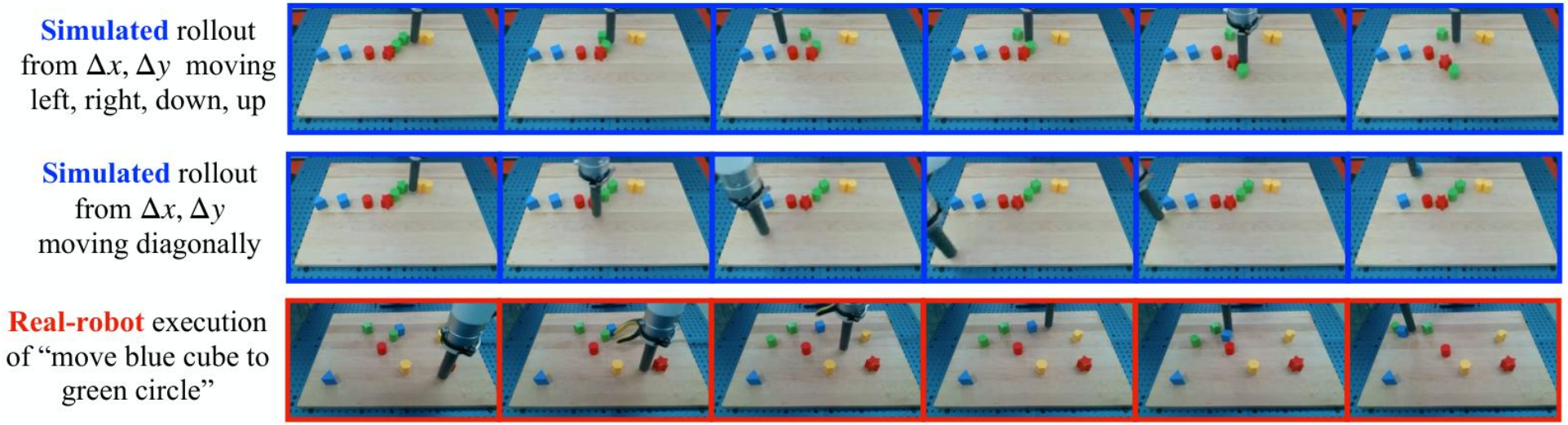
RDG(all blocks), simulated evaluation

0.07 → 0.34

The figure also shows a qualitative real-robot execution; RDG numbers are from simulation.

Application 2: low-level control with RL

Behavioral Cloning policy → Δx, Δy actions → UniSim → reward → REINFORCE



TOP 2 ROWS: simulated motor controls | BOTTOM: real robot moves blue cube

All tasks

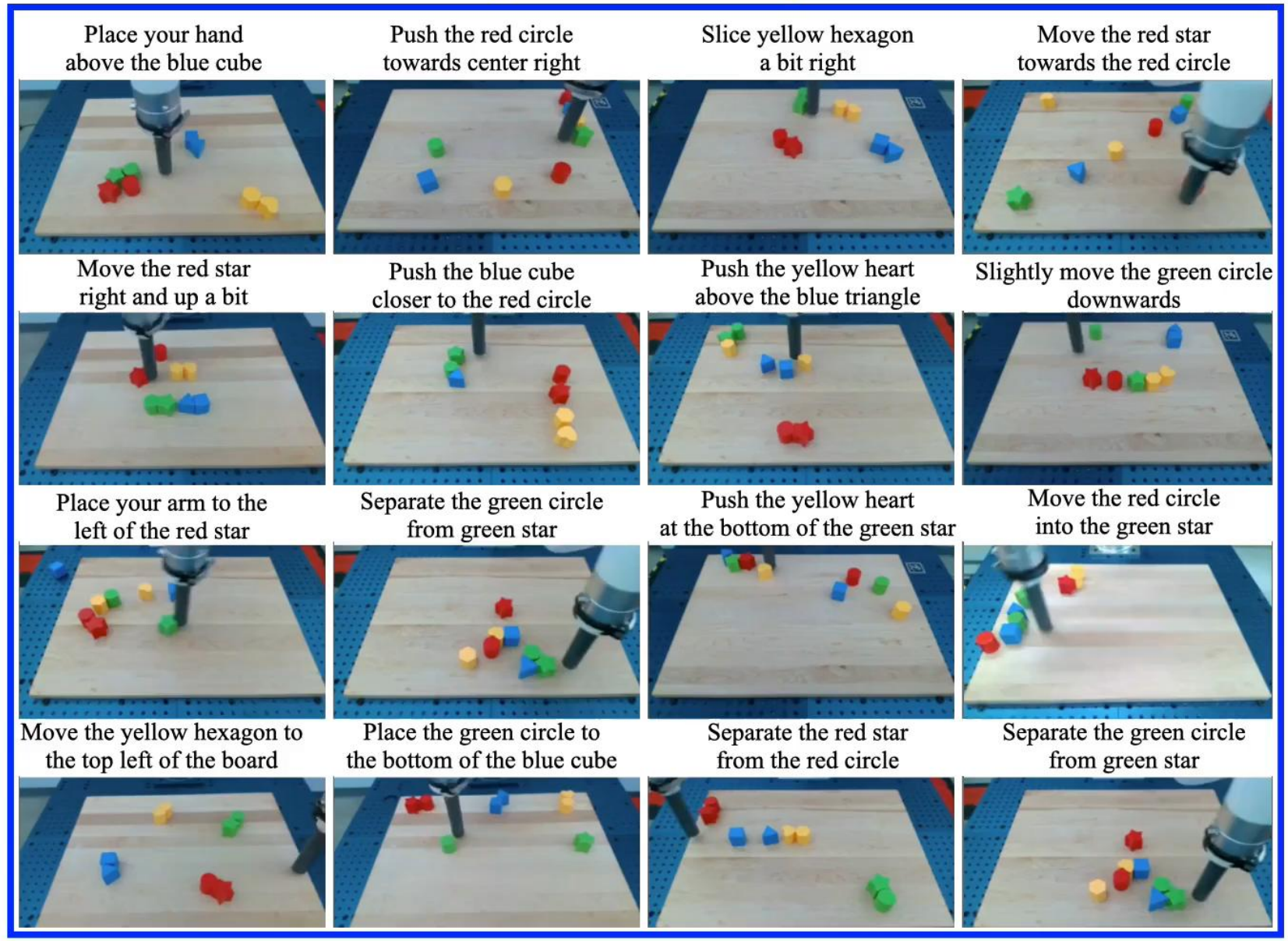
58% → 81%

Pointing tasks

12% → 71%

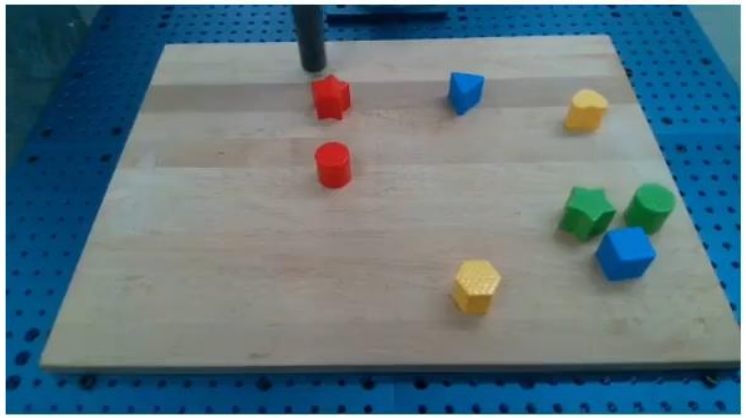
Success rates are from 48 simulated tasks; real-robot results are qualitative.

Simulated rollouts

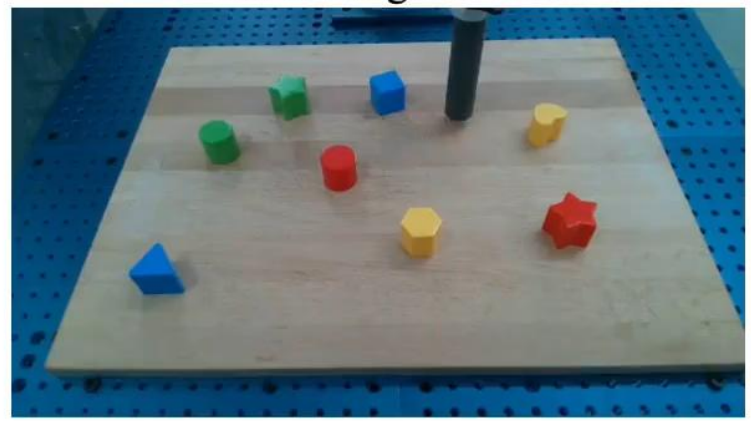


Real executions

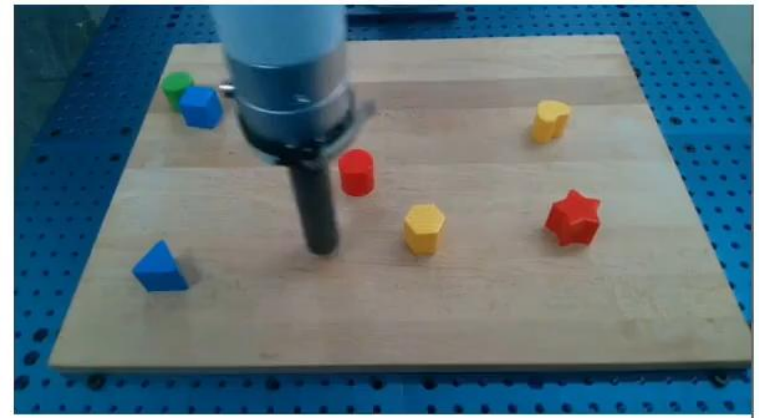
Put the red star
towards the blue cube



Move the blue cube
next to the green circle



Put the yellow pentagon
towards the blue cube



Beyond robotics: synthetic data for captioning

ActivityNet captions → **UniSim videos** → **PaLI-X fine-tuning**

(Video Captioning Dataset) (Generate synthetic video) (VLM; Train with captions and videos)

Fine-tuning data	ActivityNet	MSR-VTT	VATEX	SMIT
None	15.20	21.91	13.31	9.22
Real ActivityNet video	54.90	24.88	36.01	16.91
UniSim-generated video	**46.23**	**27.63**	**40.03**	**20.58**

Synthetic data transfers better to three other captioning datasets.

CIDEr ↑ (higher is better). ActivityNet itself favors real-video fine-tuning.

Limitations and main takeaway

LIMITATIONS

• **Hallucination**	May generate impossible actions
• **Limited memory**	Only recent history is conditioned on
• **Out-of-domain generalization**	Unseen robots and scenes remain difficult
• **Visual only**	No touch, force, or sound simulation

Takeaway: generated experience can help train agents, not just produce realistic-looking videos.

More quantitative real-world validation is still needed.

Questions for Discussion

1. How would you go about adapting a model of this scale to more commercially available hardware systems, and is there any way to optimize this kind of model to maintain its performance in, say, a quantized state?

2. A visually realistic simulator may still make incorrect predictions about the effects of actions. How can we determine whether a generative world model is reliable enough for training agents? Are there promising ways to prevent agents from learning behaviors that exploit errors in the simulator?

3. To what extent have the method developed in this paper been adopted or built upon within the wider field?

4. How can UniSim remember objects from earlier actions when it only uses a few recent frames?