Foundation Models as Real-World Simulators

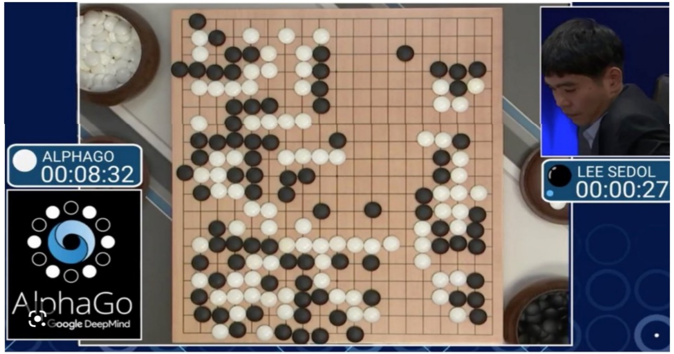
CVPR 2024 Workshop

Sherry Yang

Google DeepMind

Berkeley
UNIVERSITY OF CALIFORNIA

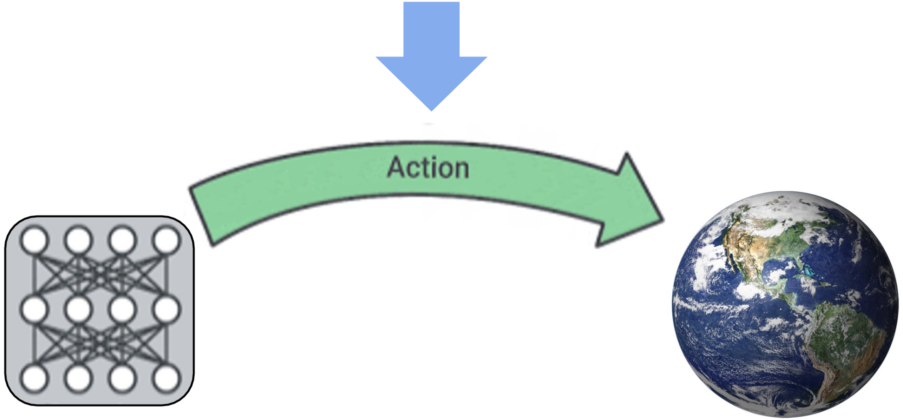
Advances in Machine Learning

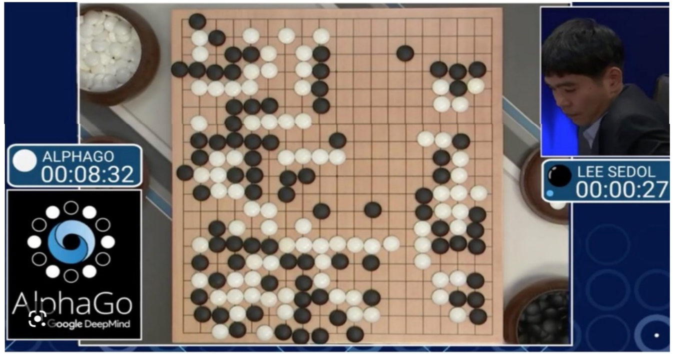


Outperforming humans in Go

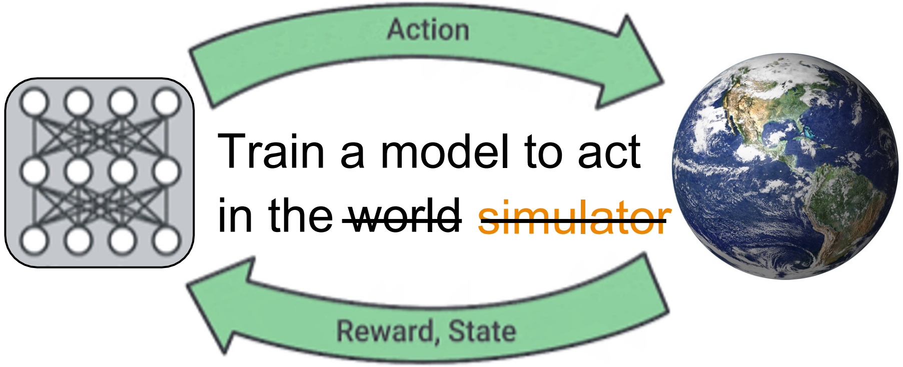
Generating language, image, and video

Decision Making

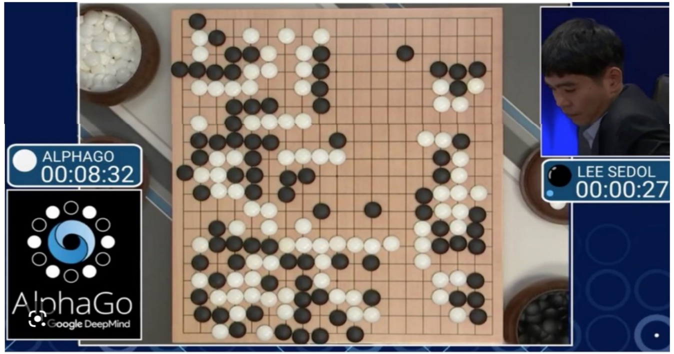




Decision Making and Internet-Scale Knowledge

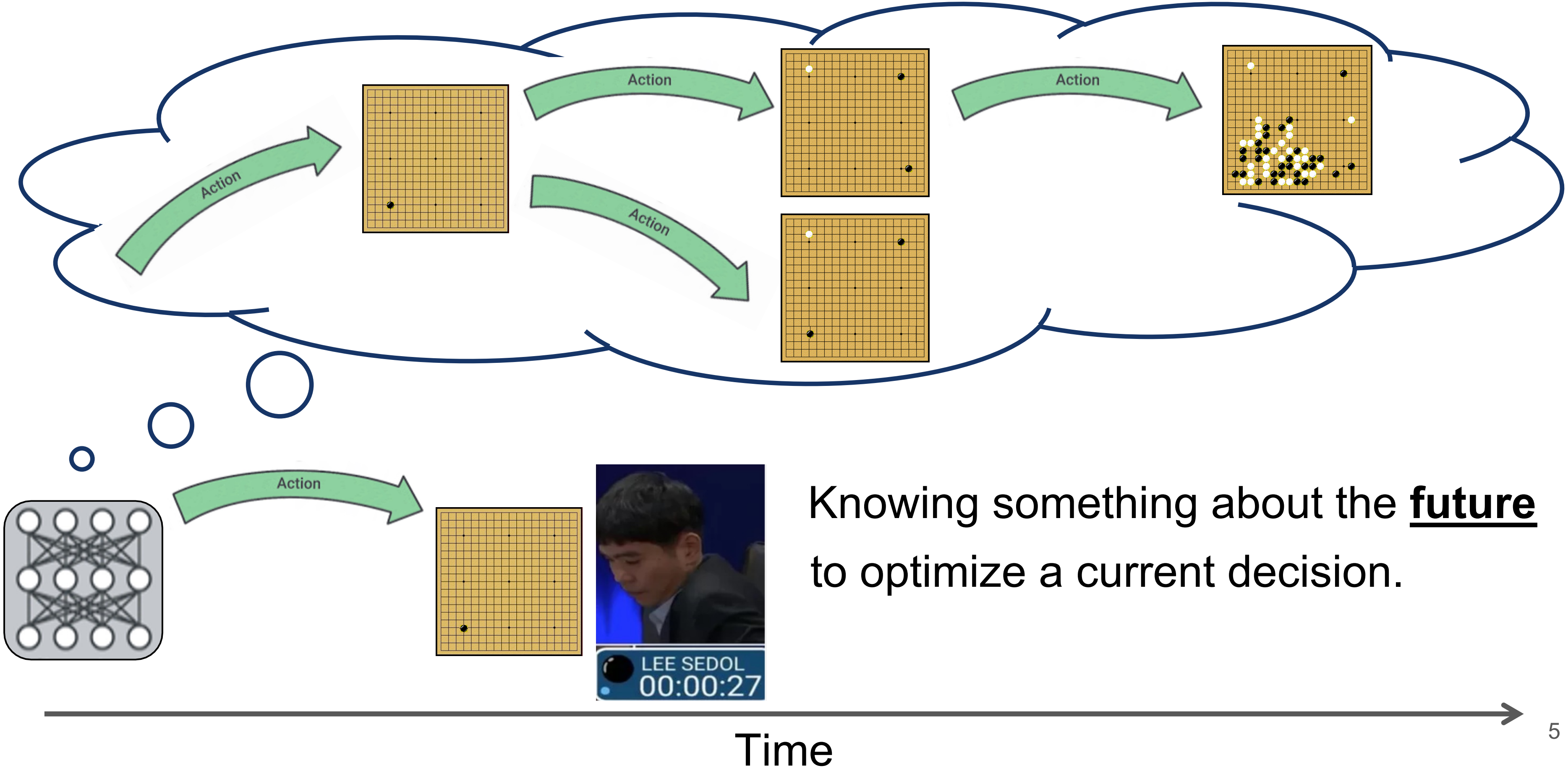


Train a model to act in the ~~world~~ ~~simulator~~



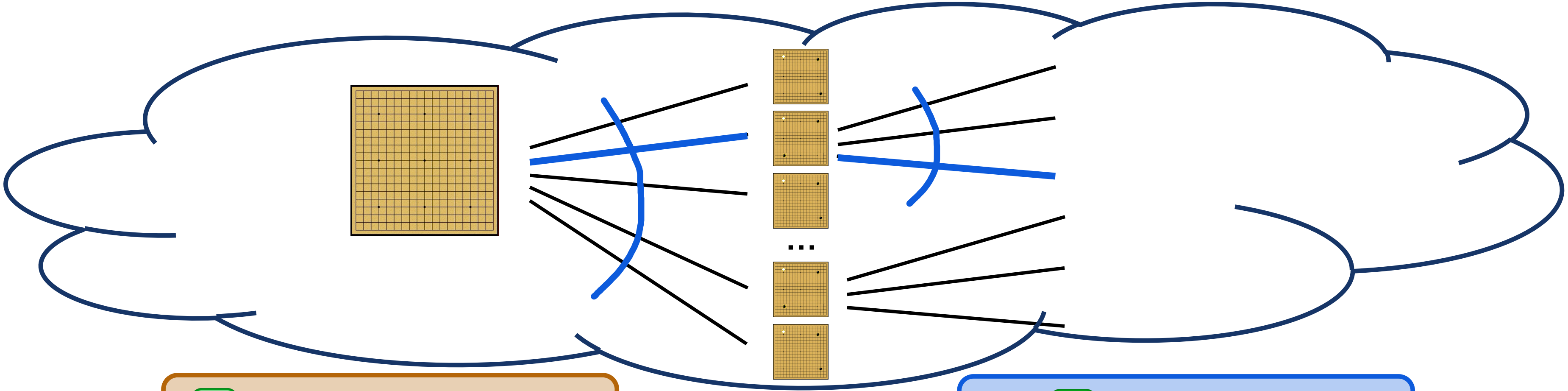
This talk: Use internet-scale data to simulate the real world

When Has Decision Making Worked?



Knowing something about the **future** to optimize a current decision.

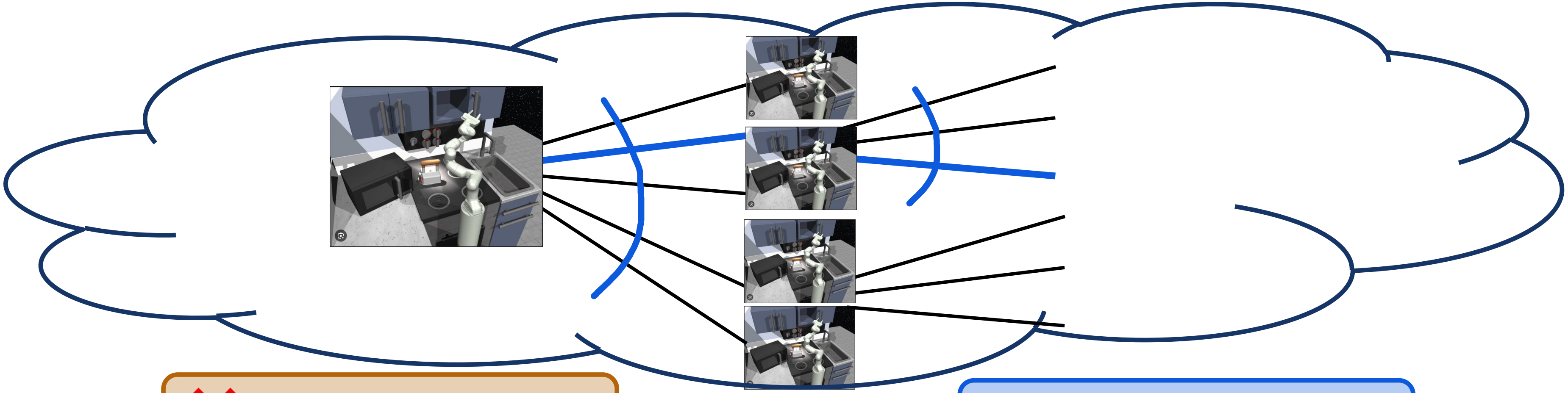
When Has Decision Making Worked?



✅ Perfect simulator

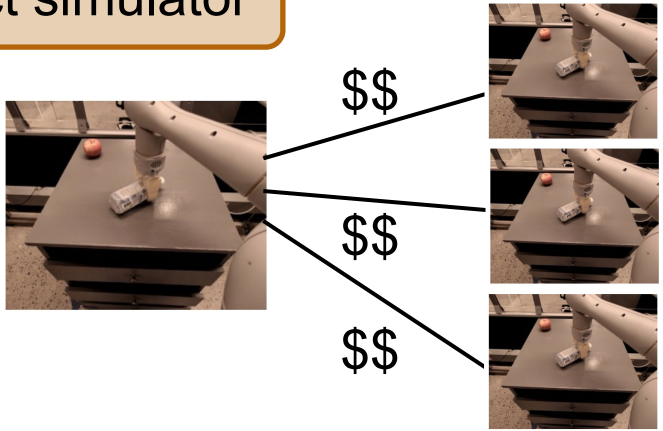
✅ Algorithms

When Has Decision Making Struggled?

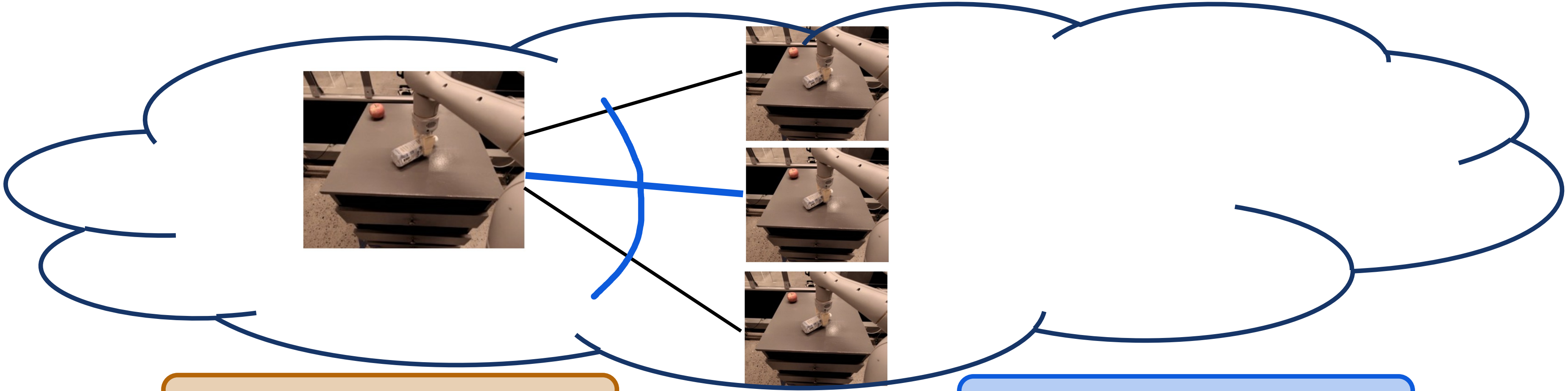


✗ Perfect simulator

? Algorithms

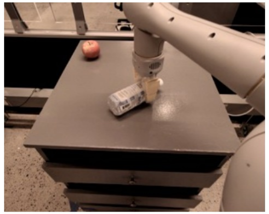


What if We Can Learn a Realistic Simulator?



✅ World model

✅ Algorithms



Definition: a learned simulator

Foundation Models as Real-World Simulators

☑ World model

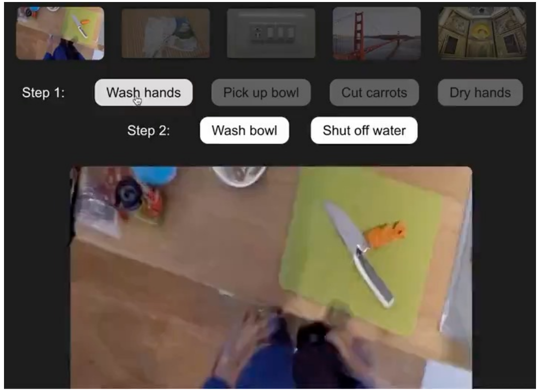
from internet data

☑ Algorithms

for decision making

☐ Challenges

and next steps



[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.
[2] **Yang** et al. Video as the New Language for Real-World Decision Making. ICML 2024.
[3] **Yang*,** Du*, et al. Learning Universal Policies via Text-Guided Video Generation. NeurIPS 2023.
[4] Du, **Yang**, et al. Video Language Planning. ICLR 2024.

Foundation Models as Real-World Simulators

✅ World model

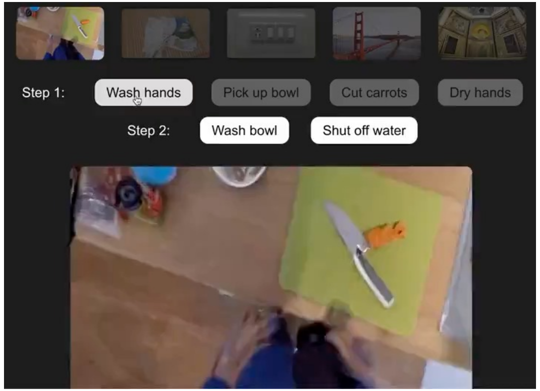
from internet data

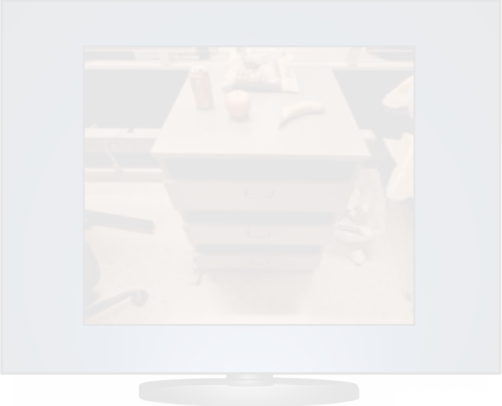
✅ Algorithms

for decision making

❑ Challenges

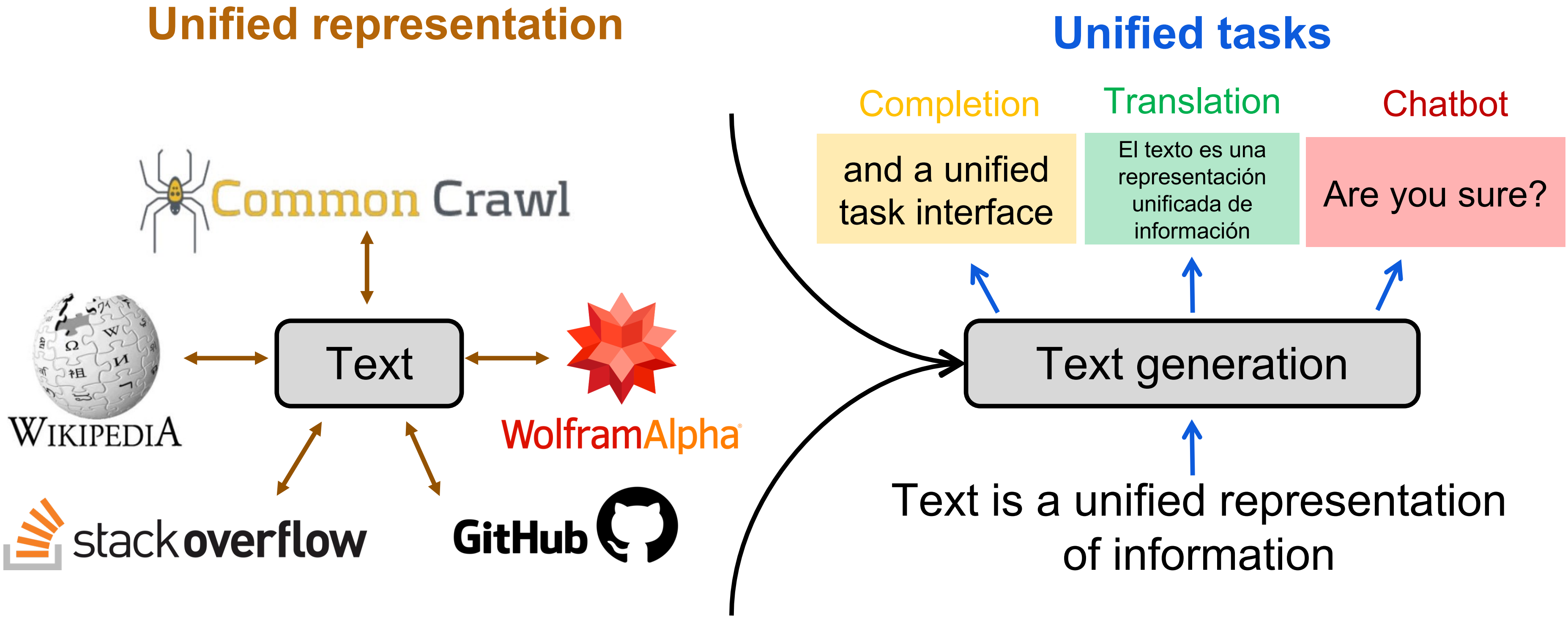
and next steps





[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.
[2] **Yang** et al. Video as the New Language for Real-World Decision Making. ICML 2024.
[3] **Yang*,** Du*, et al. Learning Universal Policies via Text-Guided Video Generation. NeurIPS 2023.
[4] Du, **Yang**, et al. Video Language Planning. ICLR 2024.

Text as Unified Representation and Task Interface



Unified representation

Text

Unified tasks

Completion

and a unified task interface

Translation

El texto es una representación unificada de información

Chatbot

Are you sure?

Text generation

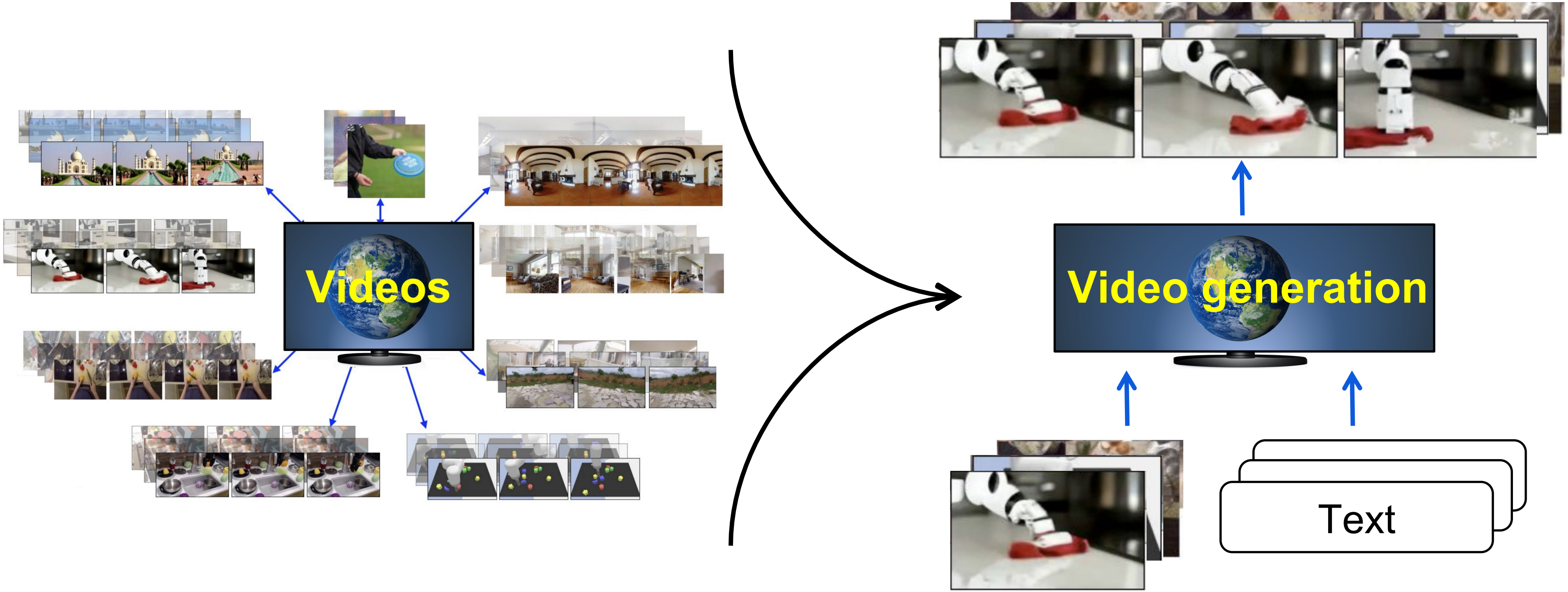
Text is a unified representation of information

[1] **Yang** et al. Video as the New Language for Real-World Decision Making. ICML 2024.

Video as Unified Representation and Task Interface

Unified representation

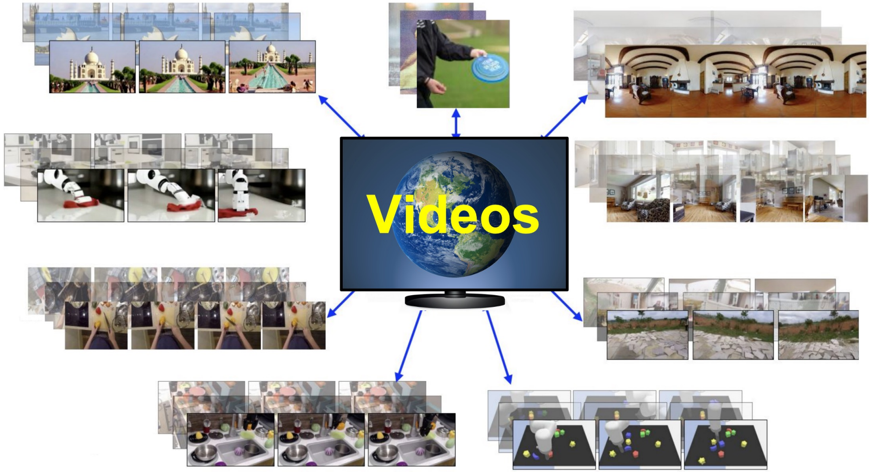
Unified tasks



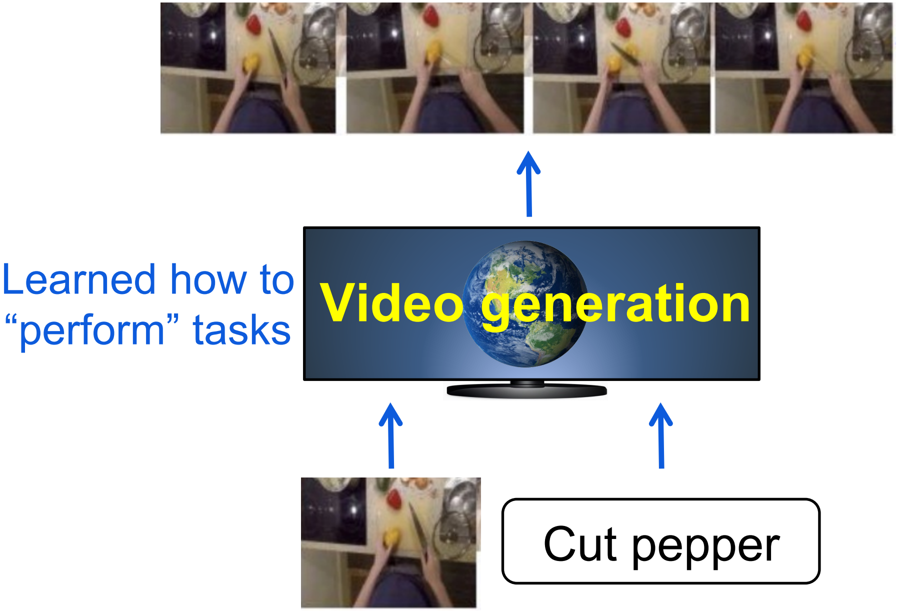
[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

Video as Unified Representation and Task Interface

Unified representation



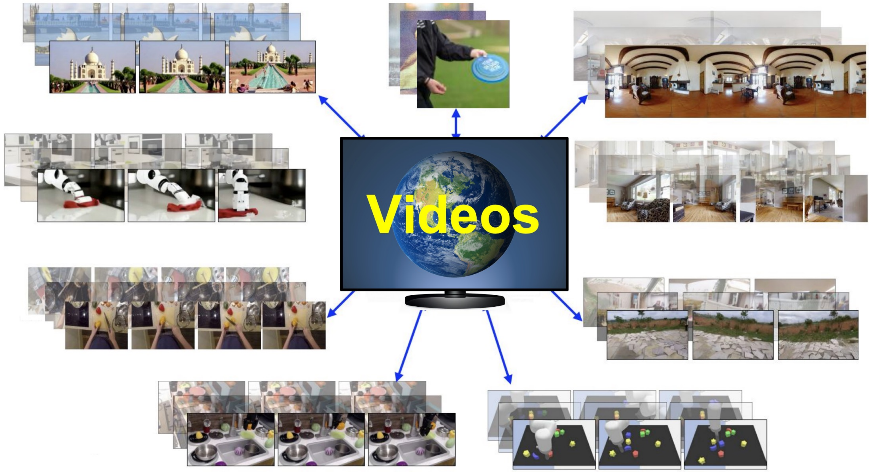
Unified tasks


Learned how to “perform” tasks

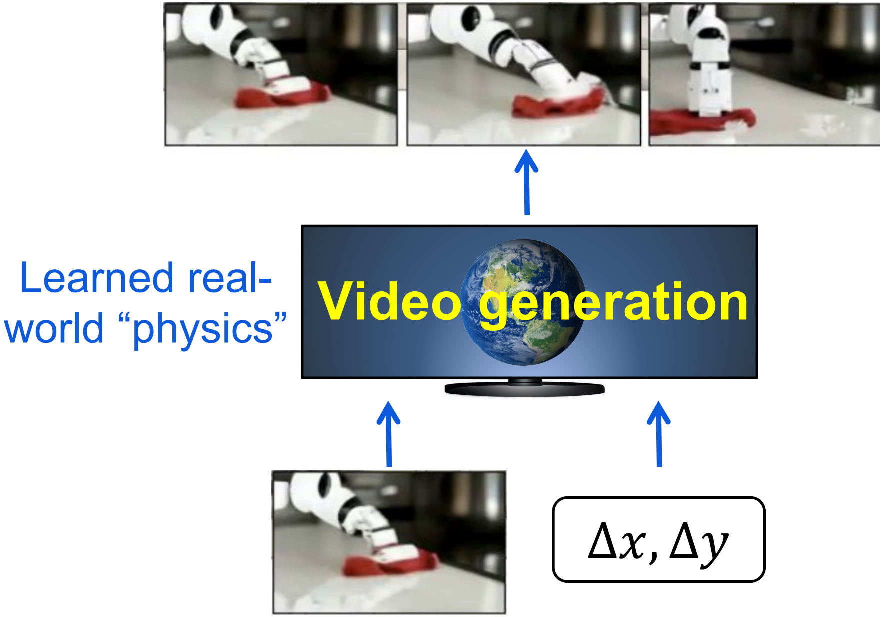
[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

Video as Unified Representation and Task Interface

Unified representation



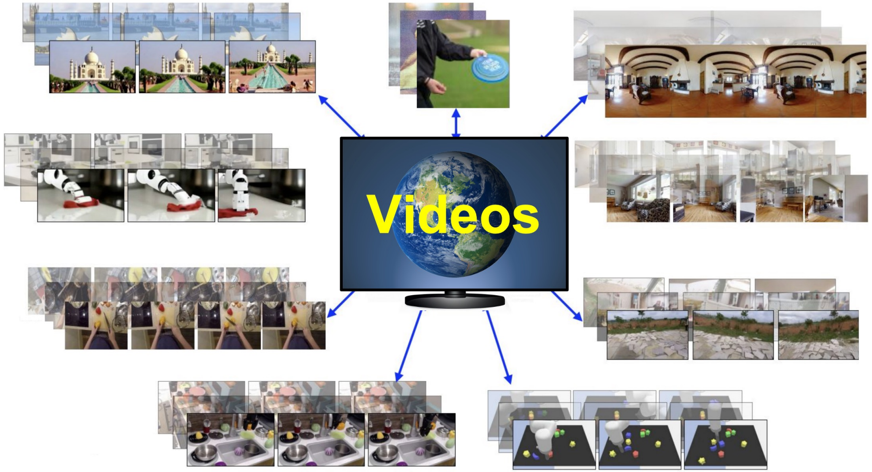
Unified tasks


Learned real-world “physics”

[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

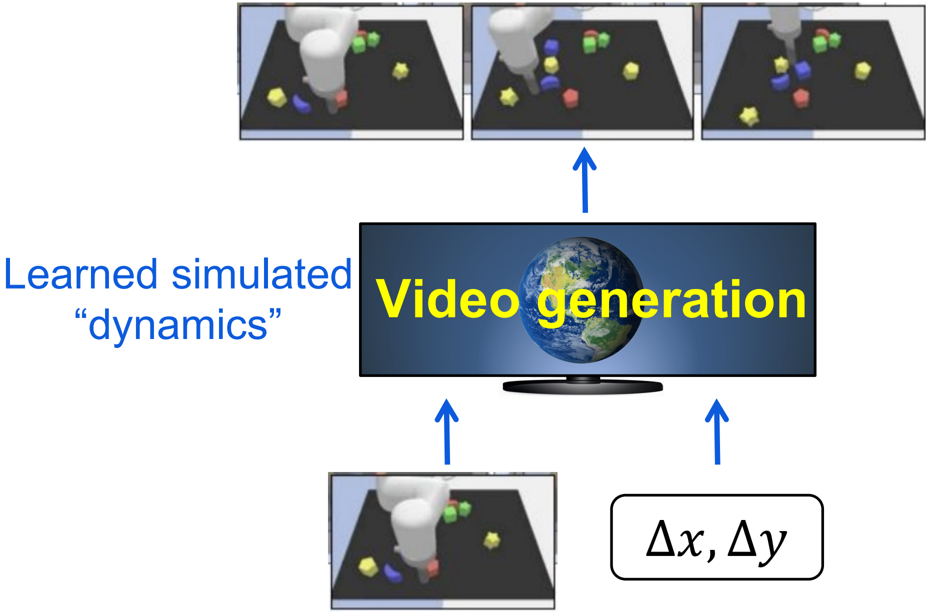
Video as Unified Representation and Task Interface

Unified representation



Videos

Unified tasks



Learned simulated “dynamics”

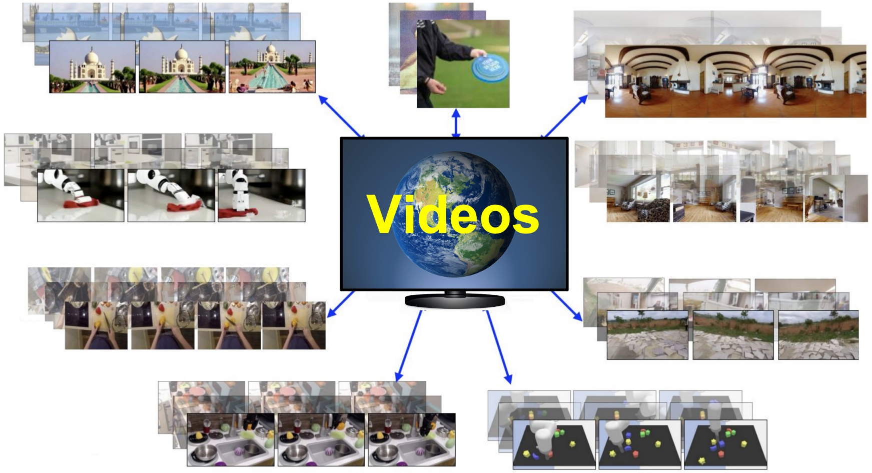
Video generation

$\Delta x, \Delta y$

[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

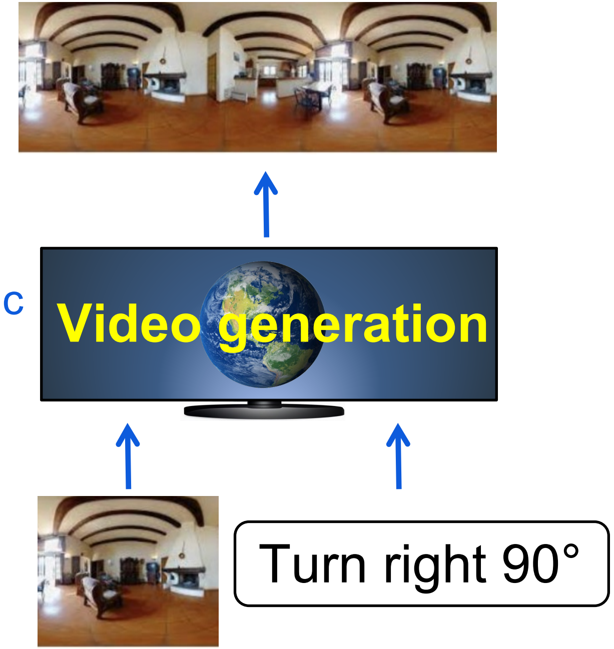
Video as Unified Representation and Task Interface

Unified representation



Learned egocentric movements

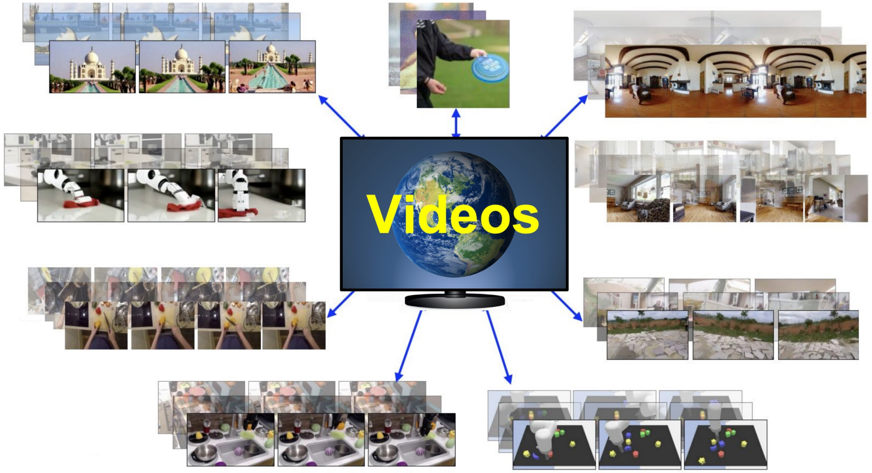
Unified tasks



[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

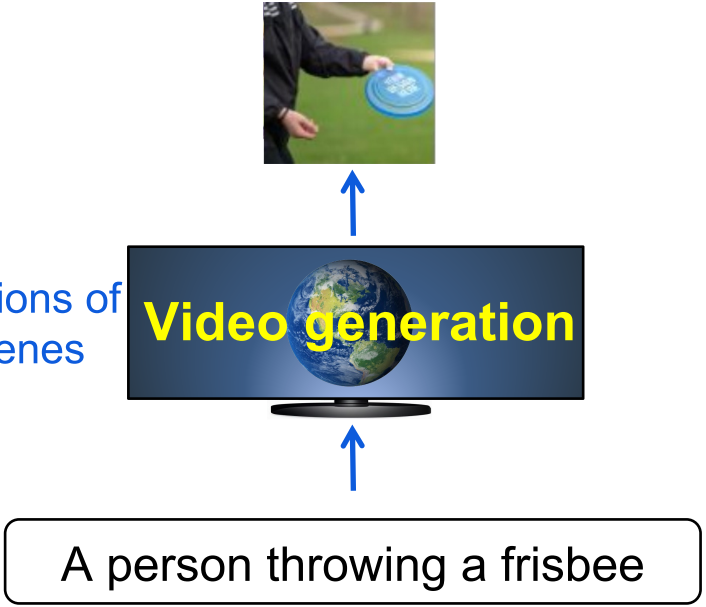
Video as Unified Representation and Task Interface

Unified representation



Unified tasks

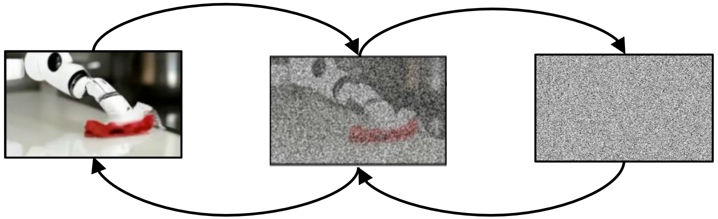
Learned notions of objects/scenes



[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

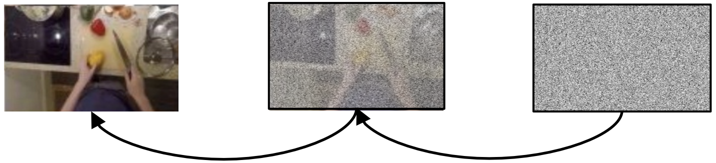
Background: Image Diffusion Models

Add Gaussian noise ϵ



Learn reverse schedule ϵ_θ(“robot”)

$$\min_\theta ||\epsilon - \epsilon_\theta||^2$$

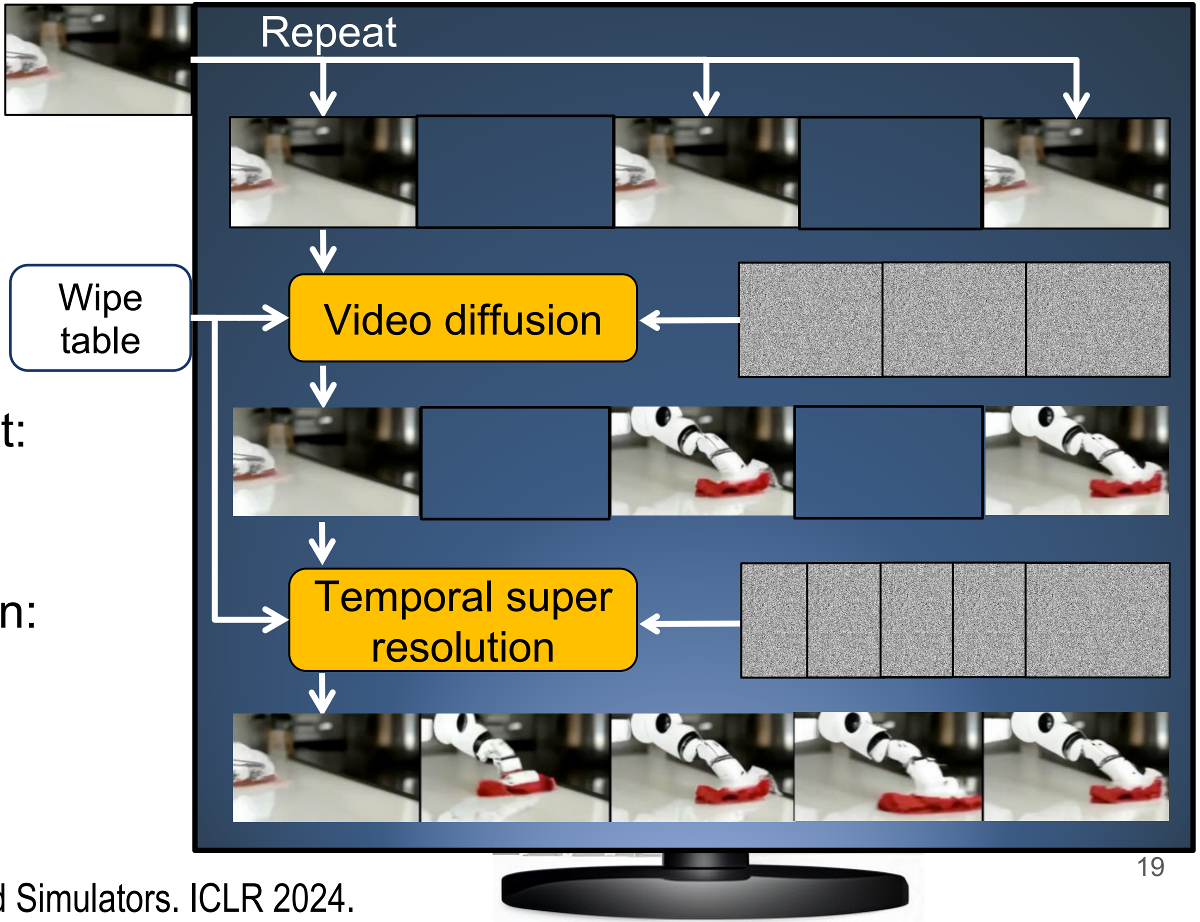


Denoise by subtracting ϵ_θ(“human”)

[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

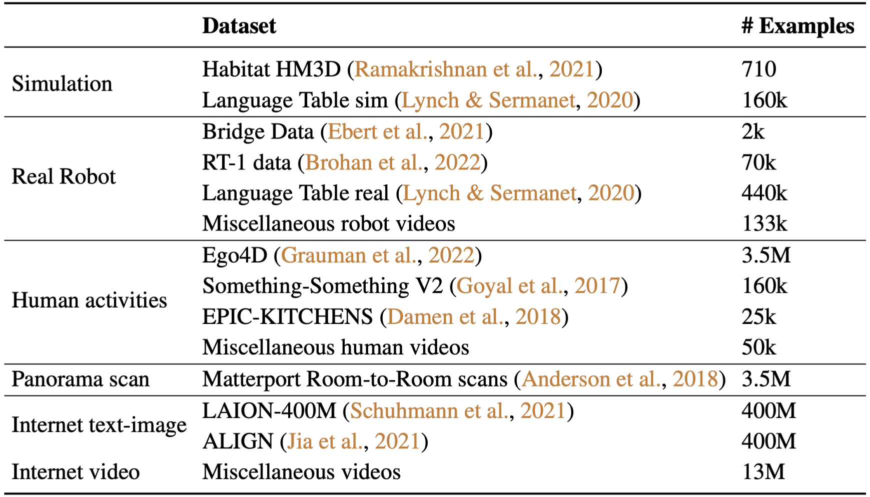
Adapting Diffusion for World Modeling

- ➢ Repeat the first frame: long-term consistency
- ➢ Condition on image & text: controllable generation
- ➢ Temporal super-resolution: flexible time horizon



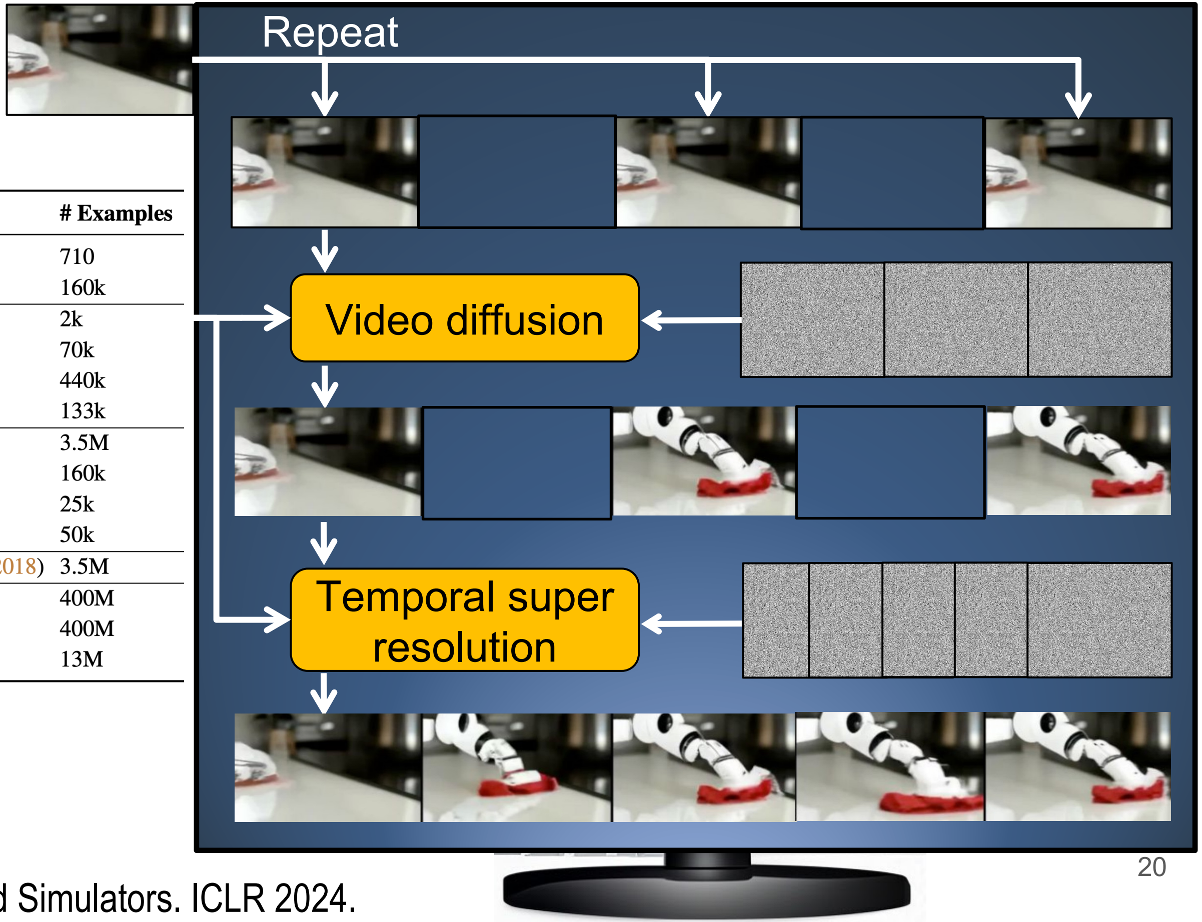
[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

Adapting Diffusion for World Modeling



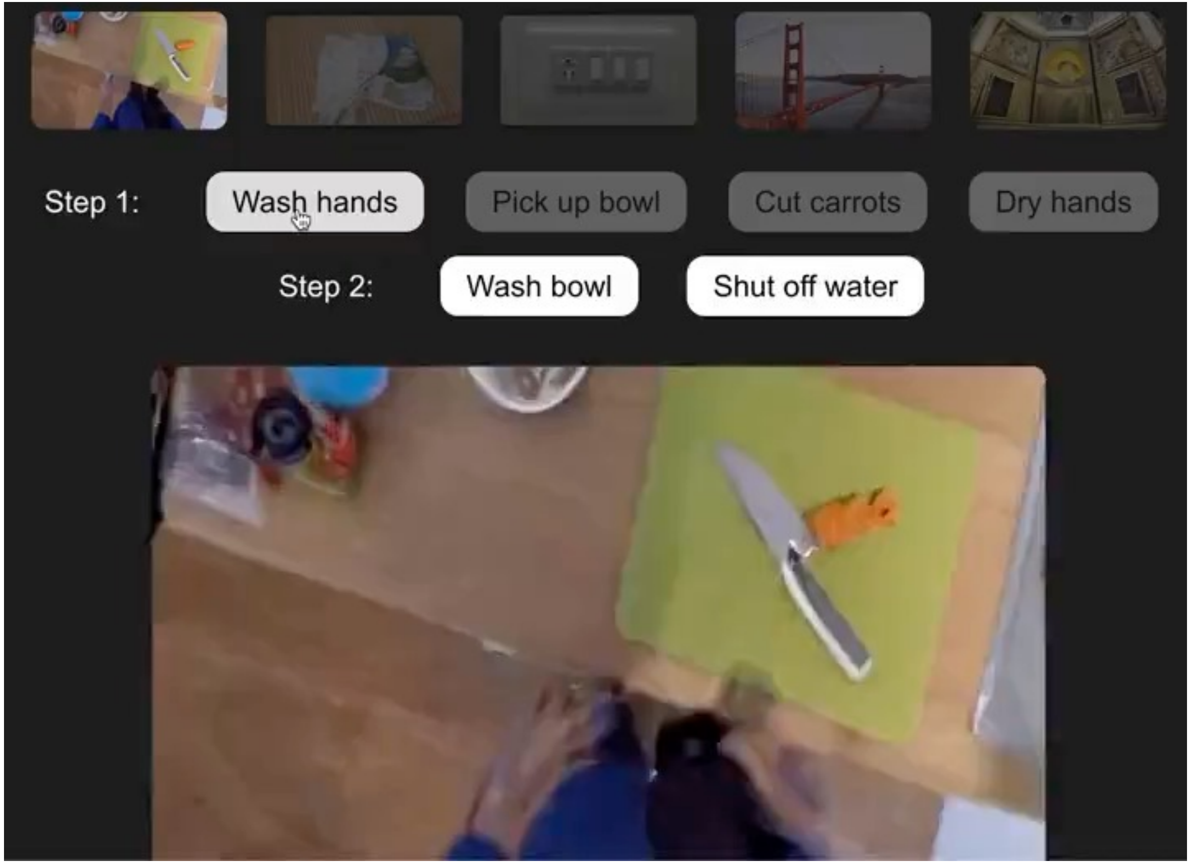
	Dataset	# Examples
Simulation	Habitat HM3D (Ramakrishnan et al., 2021)	710
	Language Table sim (Lynch & Sermanet, 2020)	160k
Real Robot	Bridge Data (Ebert et al., 2021)	2k
	RT-1 data (Brohan et al., 2022)	70k
	Language Table real (Lynch & Sermanet, 2020)	440k
	Miscellaneous robot videos	133k
Human activities	Ego4D (Grauman et al., 2022)	3.5M
	Something-Something V2 (Goyal et al., 2017)	160k
	EPIC-KITCHENS (Damen et al., 2018)	25k
	Miscellaneous human videos	50k
Panorama scan	Matterport Room-to-Room scans (Anderson et al., 2018)	3.5M
Internet text-image	LAION-400M (Schuhmann et al., 2021)	400M
	ALIGN (Jia et al., 2021)	400M
Internet video	Miscellaneous videos	13M

21M videos, 800M images



[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

UniSim: An Interactive Real-World Simulator



[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

Foundation Models as Real-World Simulators

✅ World model

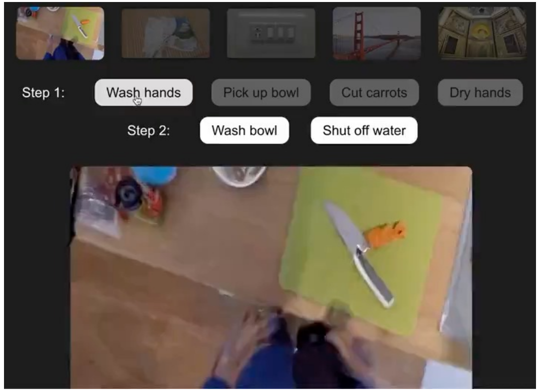
from internet data

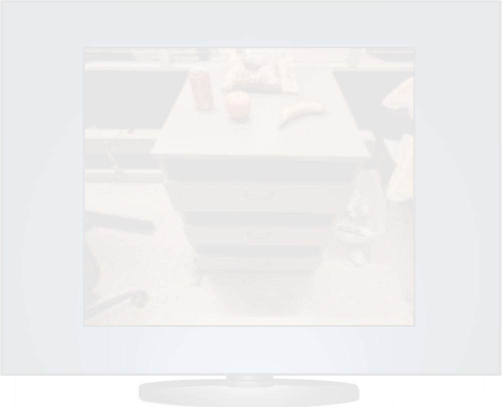
✅ Algorithms

for decision making

□ Challenges

and next steps





Takeaway: Unified repr & task interface

Foundation Models as Real-World Simulators

✅ World model

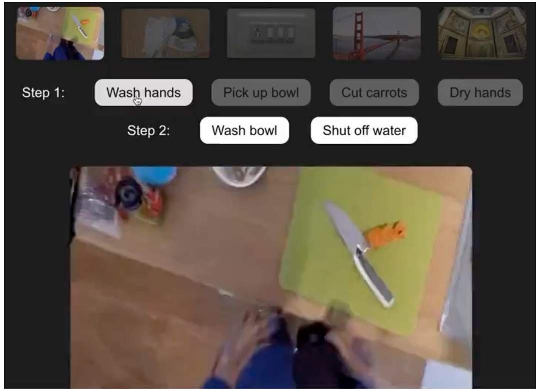
from internet data

✅ Algorithms

for decision making

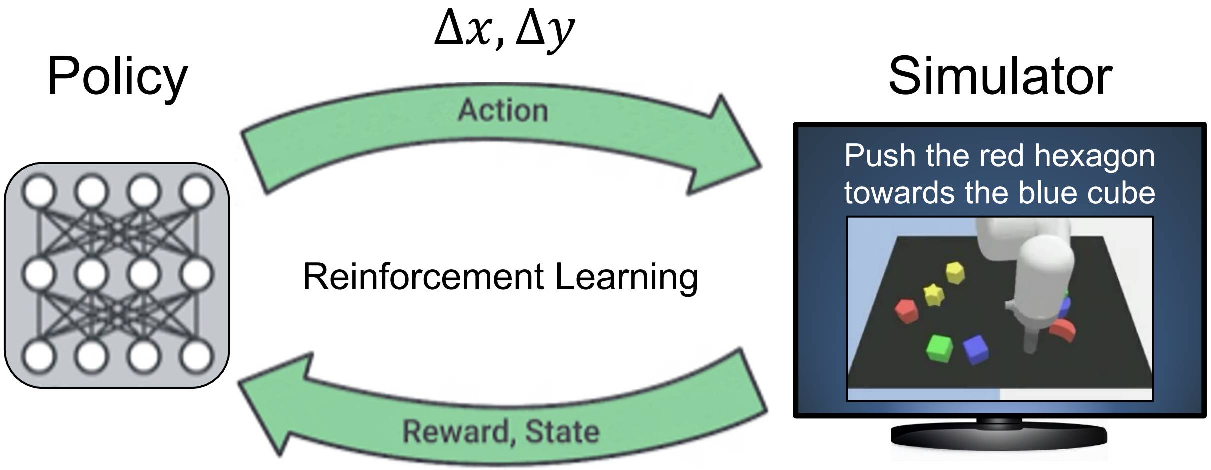
□ Challenges

and next steps

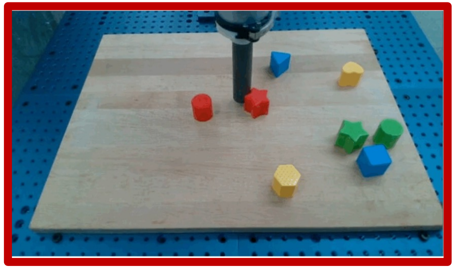


Takeaway: Unified repr & task interface

Reinforcement Learning with UniSim



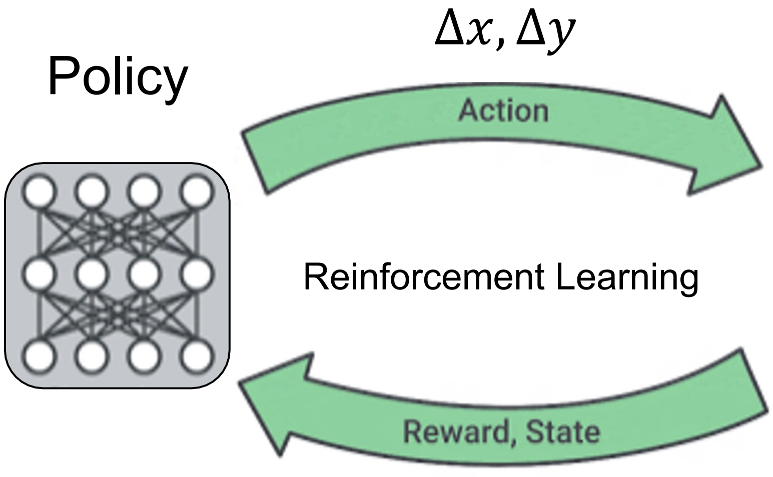
Real world



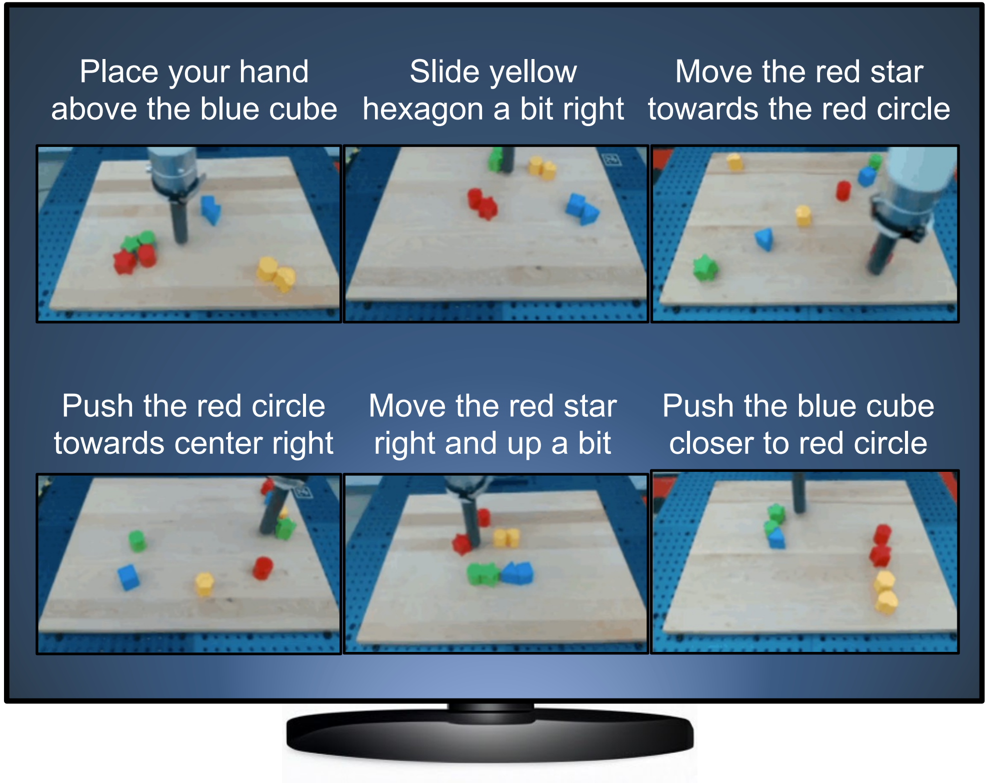
✗ Fail to transfer from sim to real

[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

Reinforcement Learning with UniSim



Simulator


Place your hand above the blue cube

Slide yellow hexagon a bit right

Move the red star towards the red circle

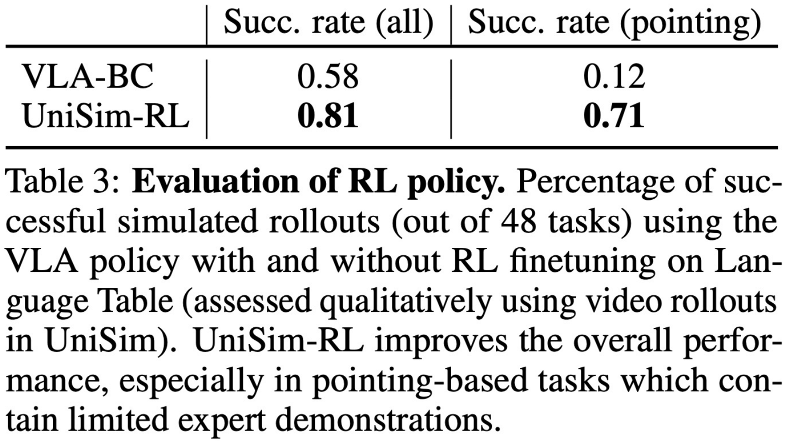
Push the red circle towards center right

Move the red star right and up a bit

Push the blue cube closer to red circle

[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

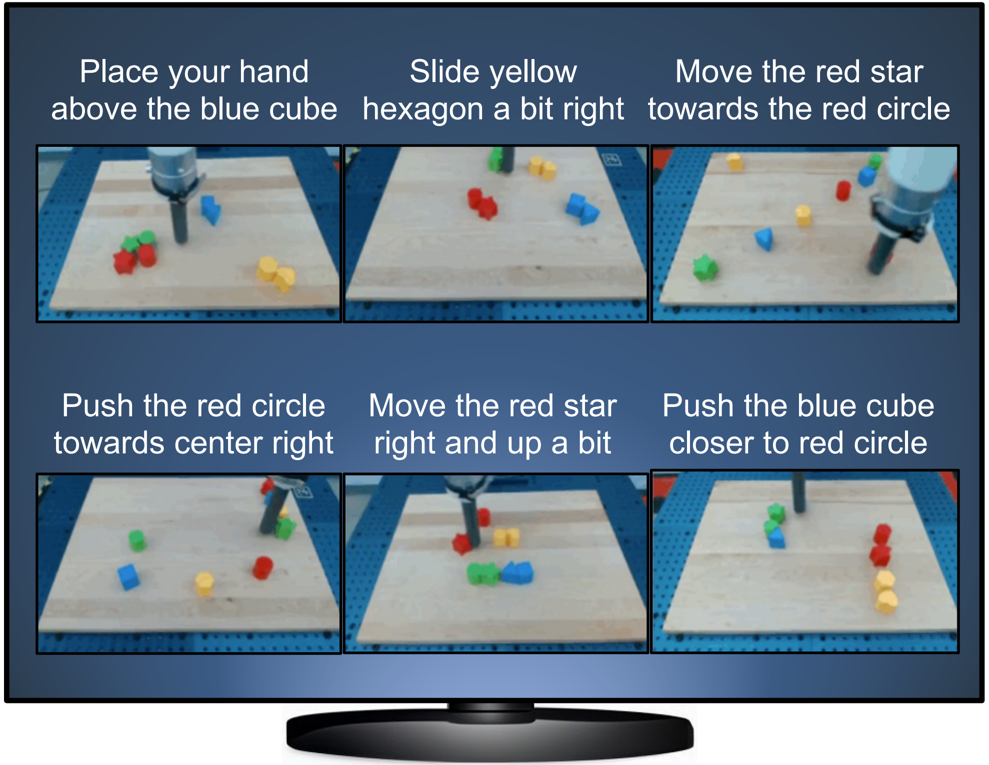
Reinforcement Learning with UniSim



	Succ. rate (all)	Succ. rate (pointing)
VLA-BC	0.58	0.12
UniSim-RL	**0.81**	**0.71**

Table 3: **Evaluation of RL policy.** Percentage of successful simulated rollouts (out of 48 tasks) using the VLA policy with and without RL finetuning on Language Table (assessed qualitatively using video rollouts in UniSim). UniSim-RL improves the overall performance, especially in pointing-based tasks which contain limited expert demonstrations.

Simulator



Place your hand above the blue cube

Slide yellow hexagon a bit right

Move the red star towards the red circle

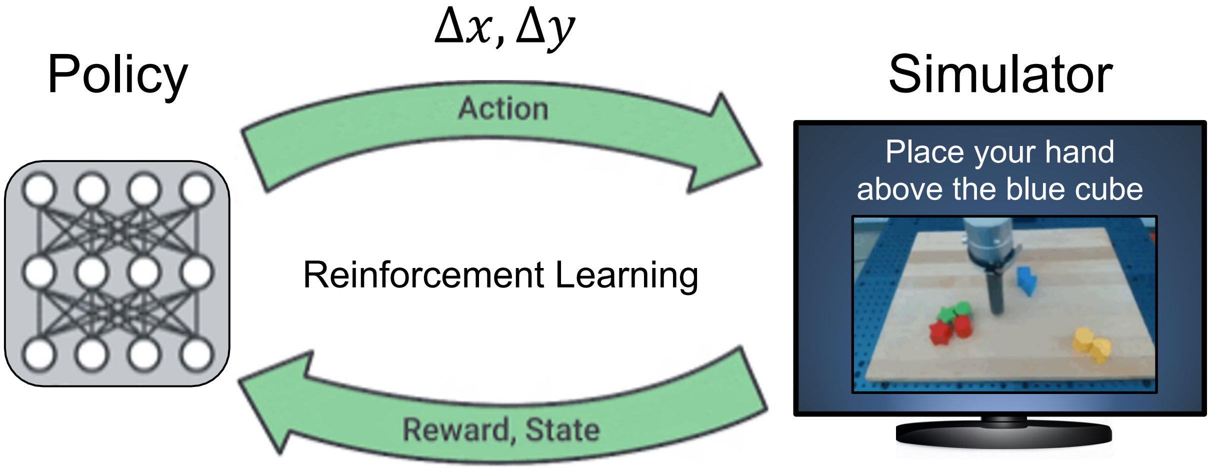
Push the red circle towards center right

Move the red star right and up a bit

Push the blue cube closer to red circle

[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

Reinforcement Learning with UniSim



Policy

$\Delta x, \Delta y$

Action

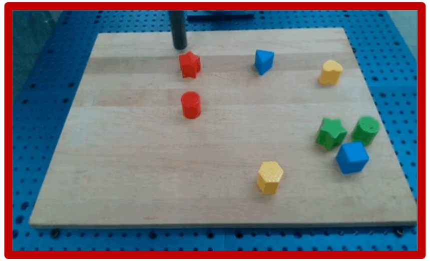
Reinforcement Learning

Reward, State

Simulator

Place your hand above the blue cube

Real world

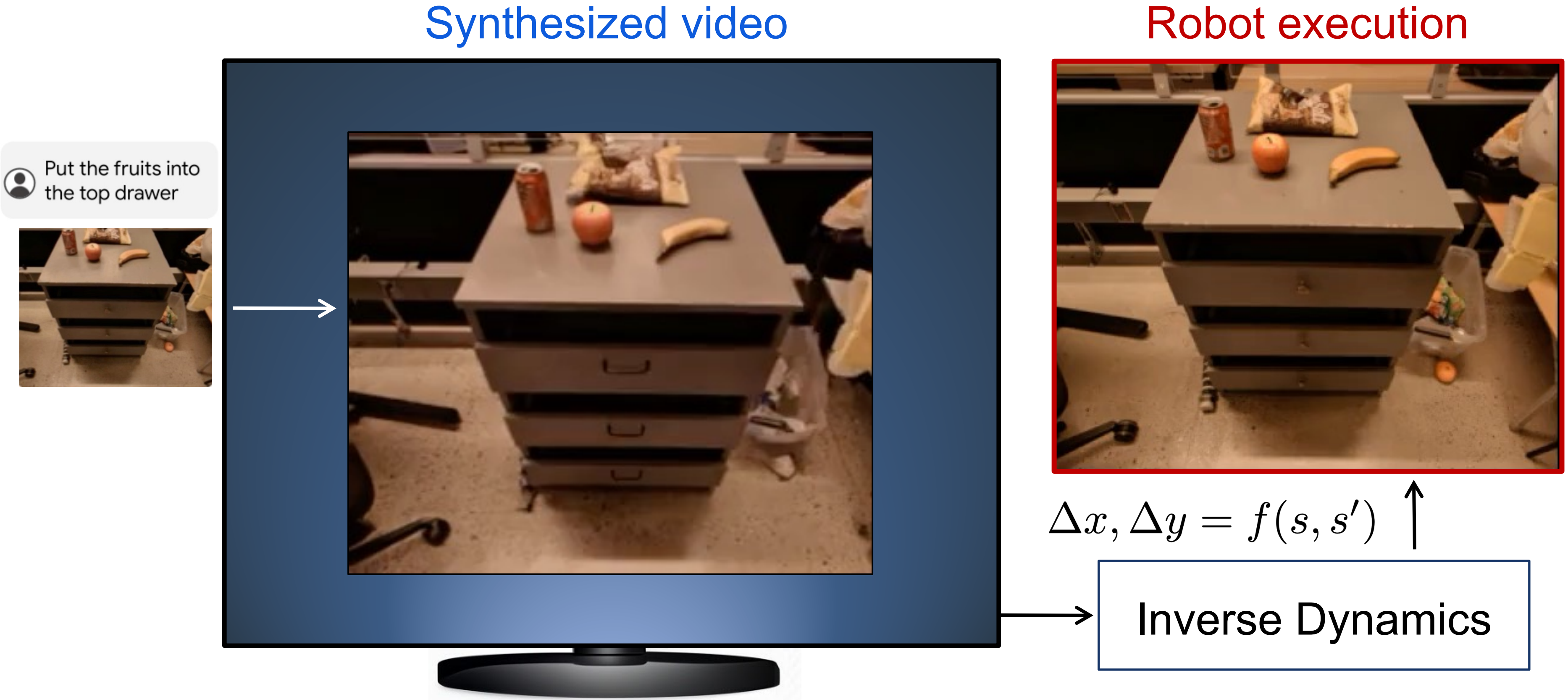


Task: Push the red star towards the blue cube

✅ Transfer from sim to real

[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

Planning with UniSim



Synthesized video

Robot execution

Put the fruits into the top drawer

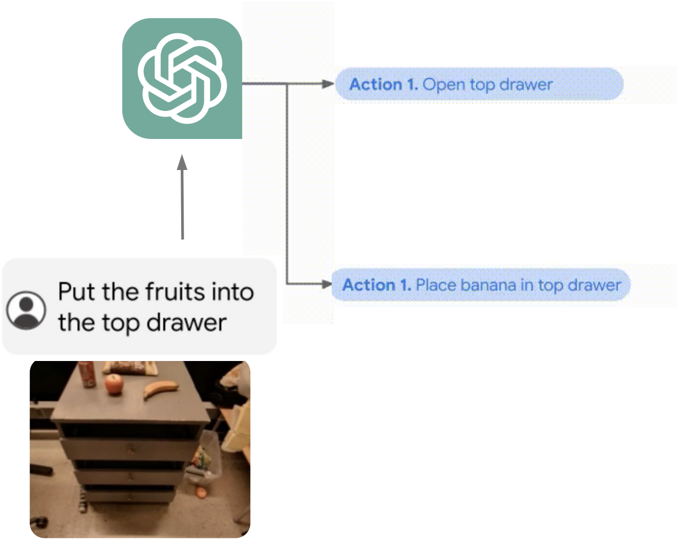
$\Delta x, \Delta y = f(s, s')$

Inverse Dynamics

[1] **Yang*,** Du*, et al. Learning Universal Policies via Text-Guided Video Generation. NeurIPS 2023.
[2] Du, **Yang**, et al. Video Language Planning. ICLR 2024.

Planning with UniSim

Vision language model

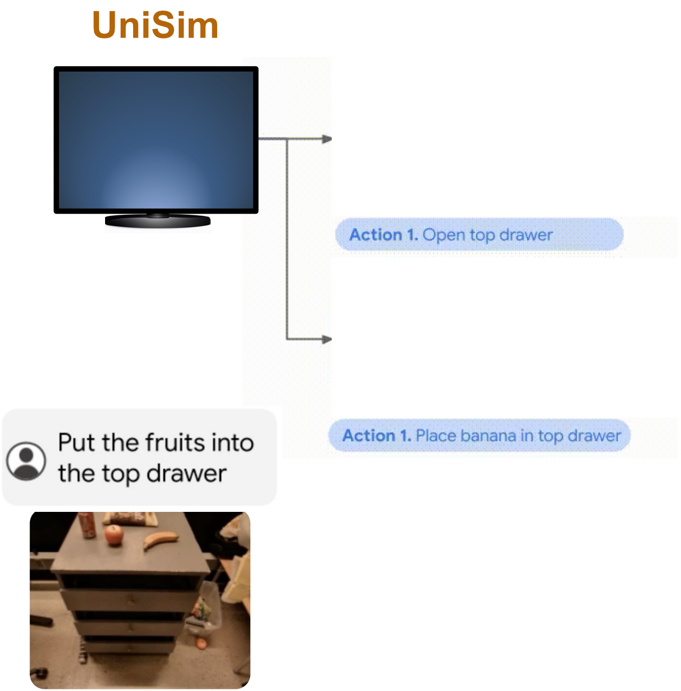


Action 1. Open top drawer

Action 1. Place banana in top drawer

Put the fruits into the top drawer

Planning with UniSim



UniSim

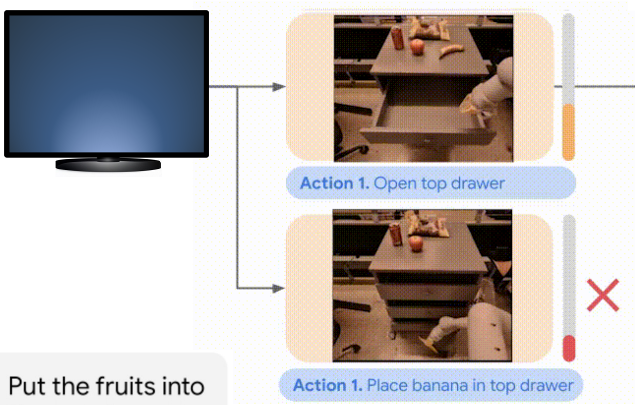
Action 1. Open top drawer

Action 1. Place banana in top drawer

Put the fruits into the top drawer

Planning with UniSim

Vision-language reward model

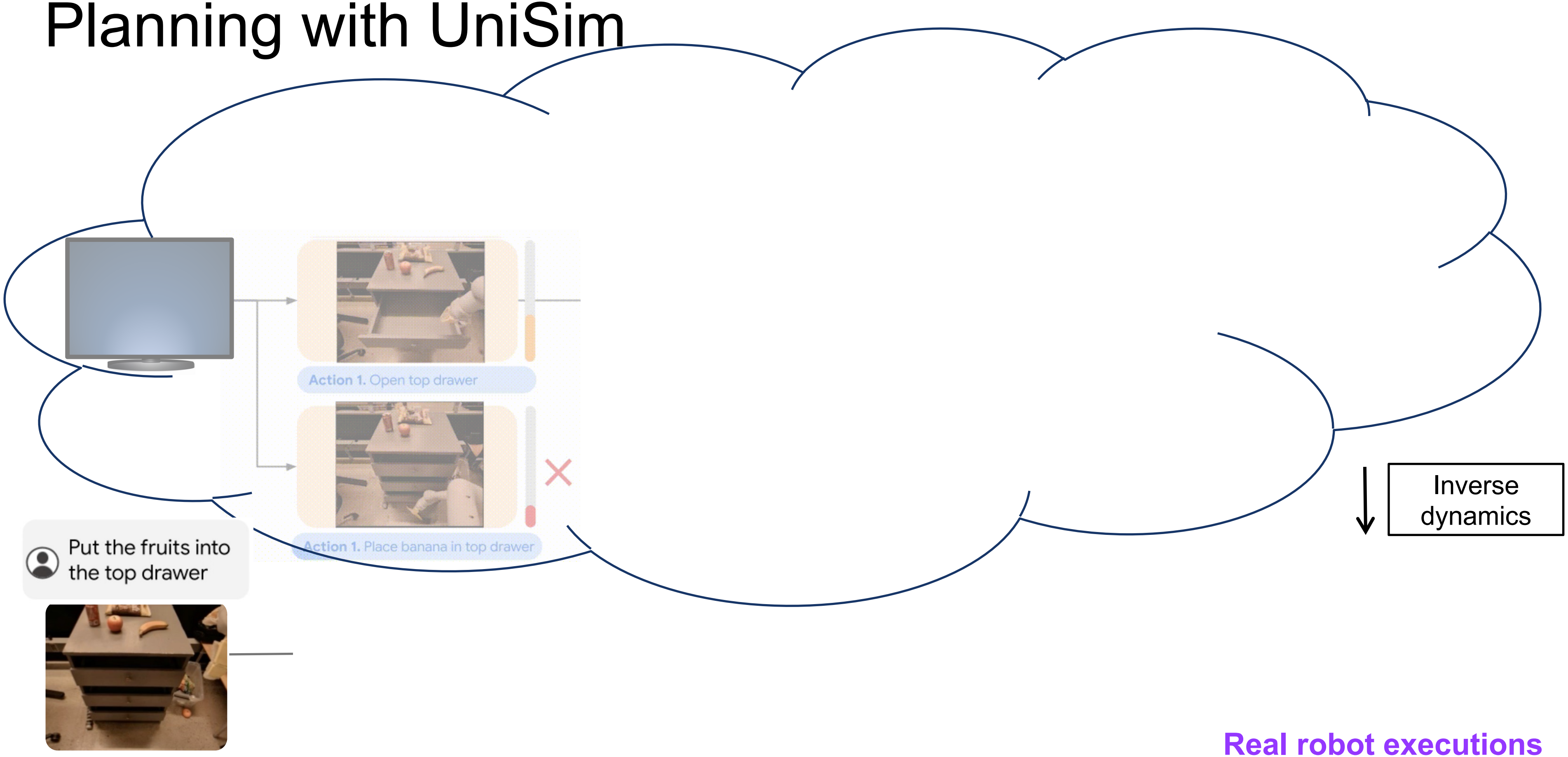


Action 1. Open top drawer

Action 1. Place banana in top drawer

Put the fruits into the top drawer

Planning with UniSim



Put the fruits into the top drawer

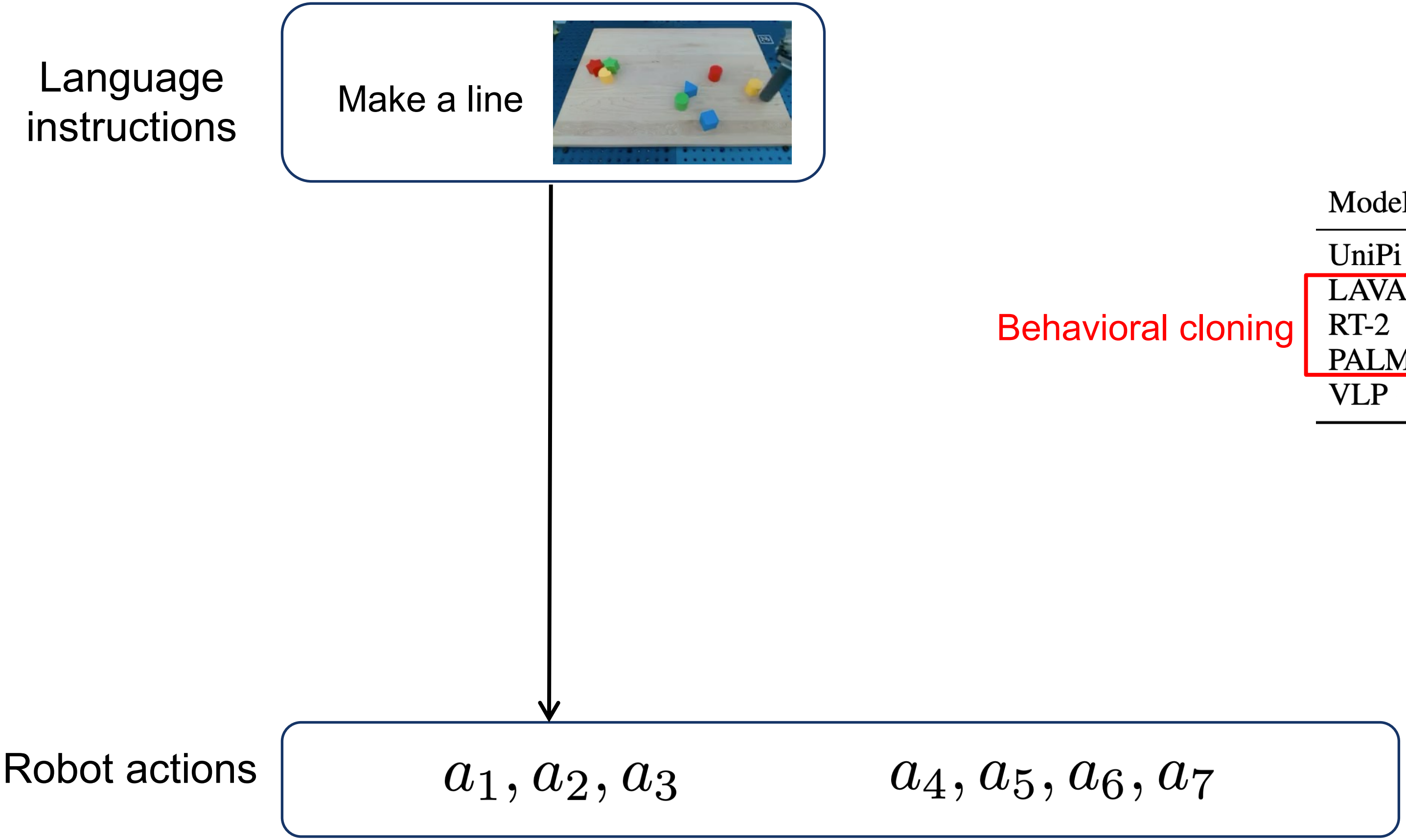
Action 1. Open top drawer

Action 1. Place banana in top drawer

Inverse dynamics

Real robot executions

Planning with UniSim – Why?



Language instructions

Make a line

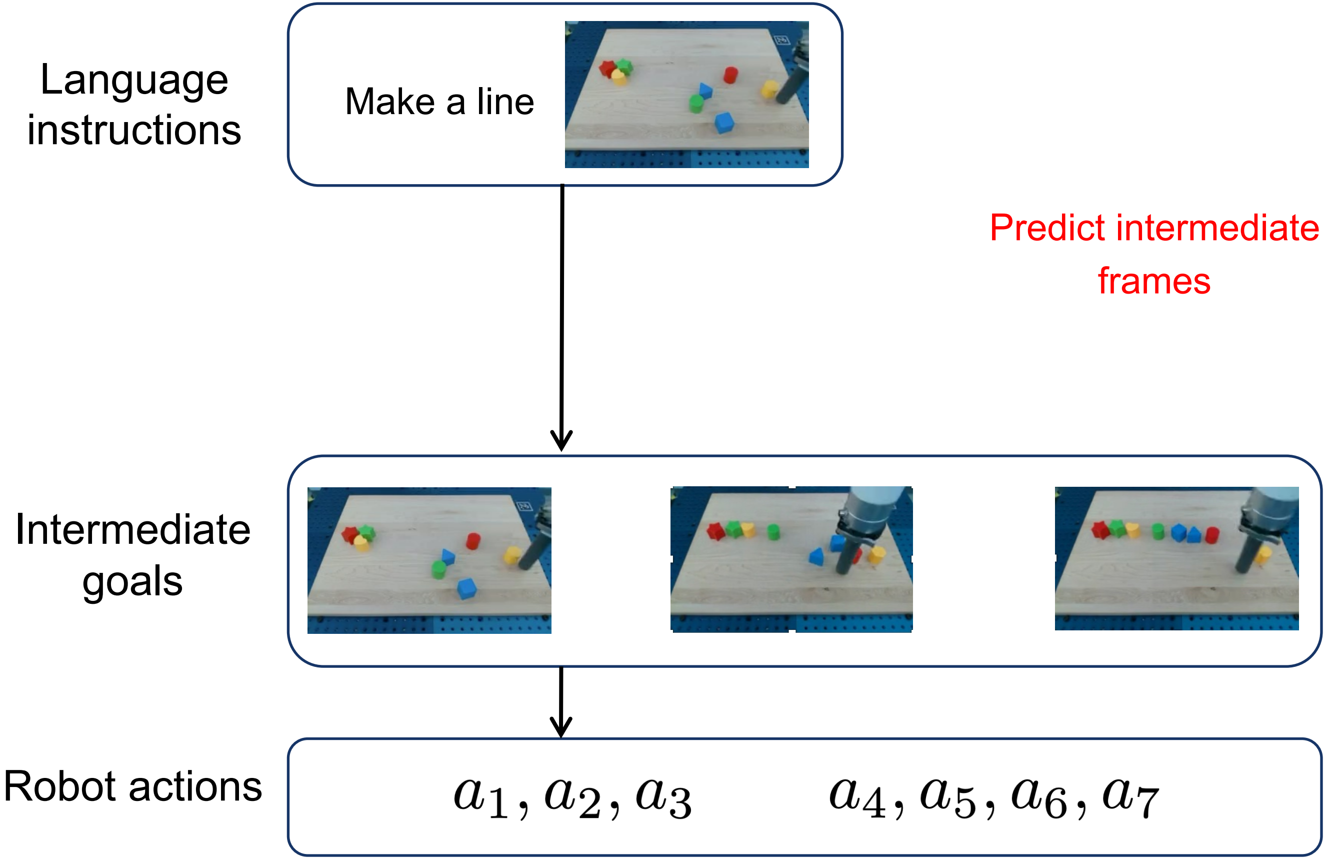
Robot actions

a_1, a_2, a_3 a_4, a_5, a_6, a_7

Behavioral cloning

	Make Line	
Model	Reward	Completion
UniPi	44.0	4%
LAVA	33.5	0%
RT-2	36.5	2%
PALM-E	26.2	0%
VLP	**65.0**	**16%**

Planning with UniSim – Why?



Language instructions

Make a line

Intermediate goals

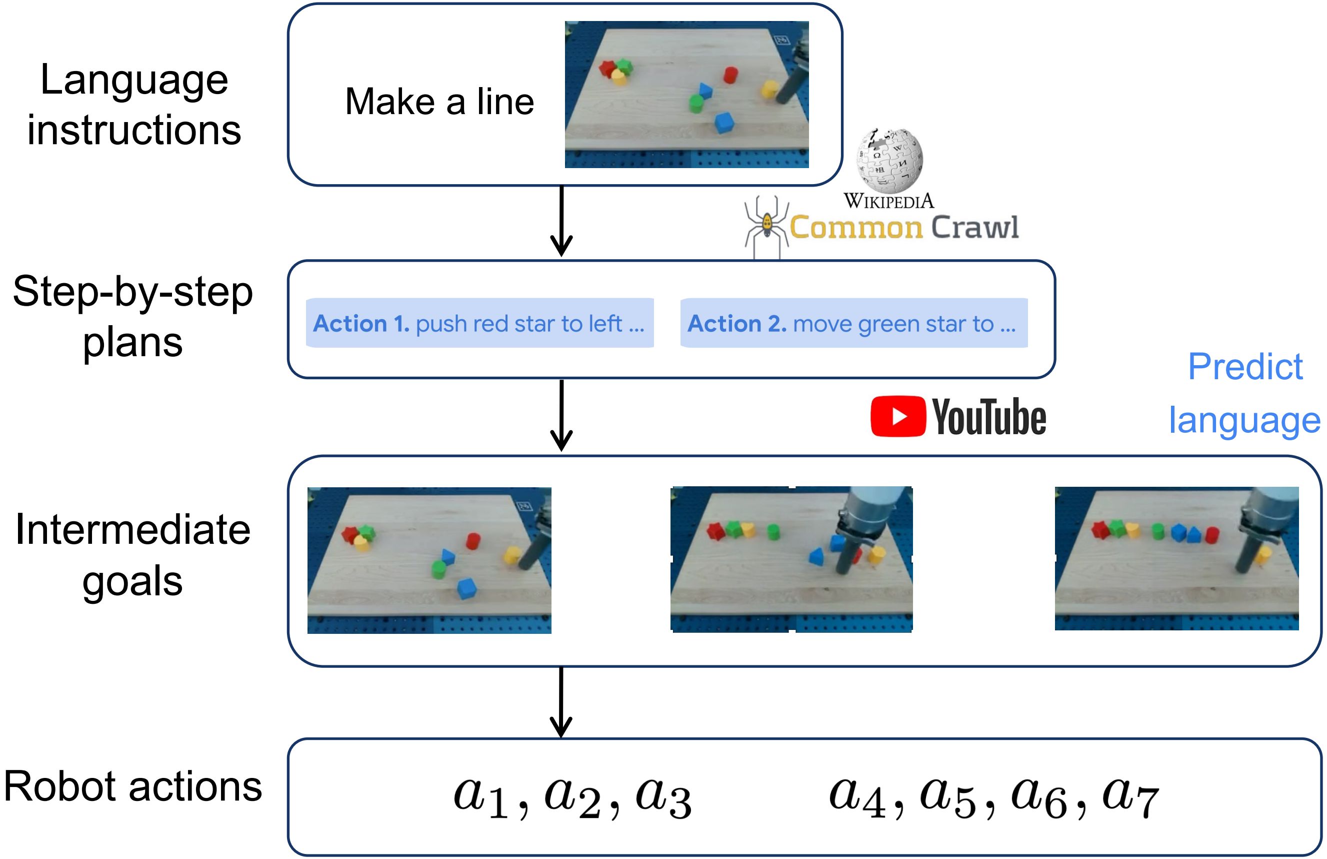
Robot actions

a_1, a_2, a_3 a_4, a_5, a_6, a_7

Predict intermediate frames

	Make Line	
Model	Reward	Completion
UniPi	44.0	4%
LAVA	33.5	0%
RT-2	36.5	2%
PALM-E	26.2	0%
VLP	**65.0**	**16%**

Planning with UniSim – Why?



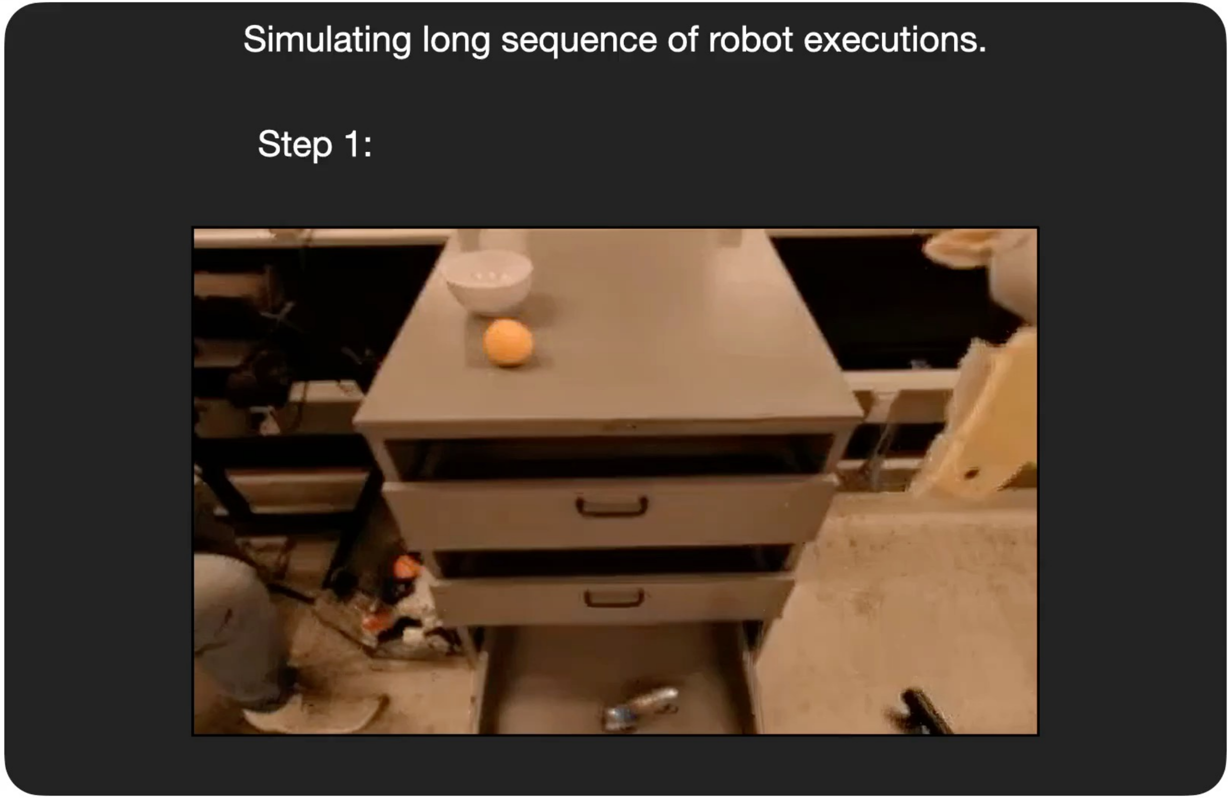
Predict language

	Make Line	
Model	Reward	Completion
UniPi	44.0	4%
LAVA	33.5	0%
RT-2	36.5	2%
PALM-E	26.2	0%
VLP	**65.0**	**16%**

Benefits:

(1) Internet-scale data
(2) Temporal flexibility
(3) Search, planning, verify at each level

Long-Horizon Planning with UniSim



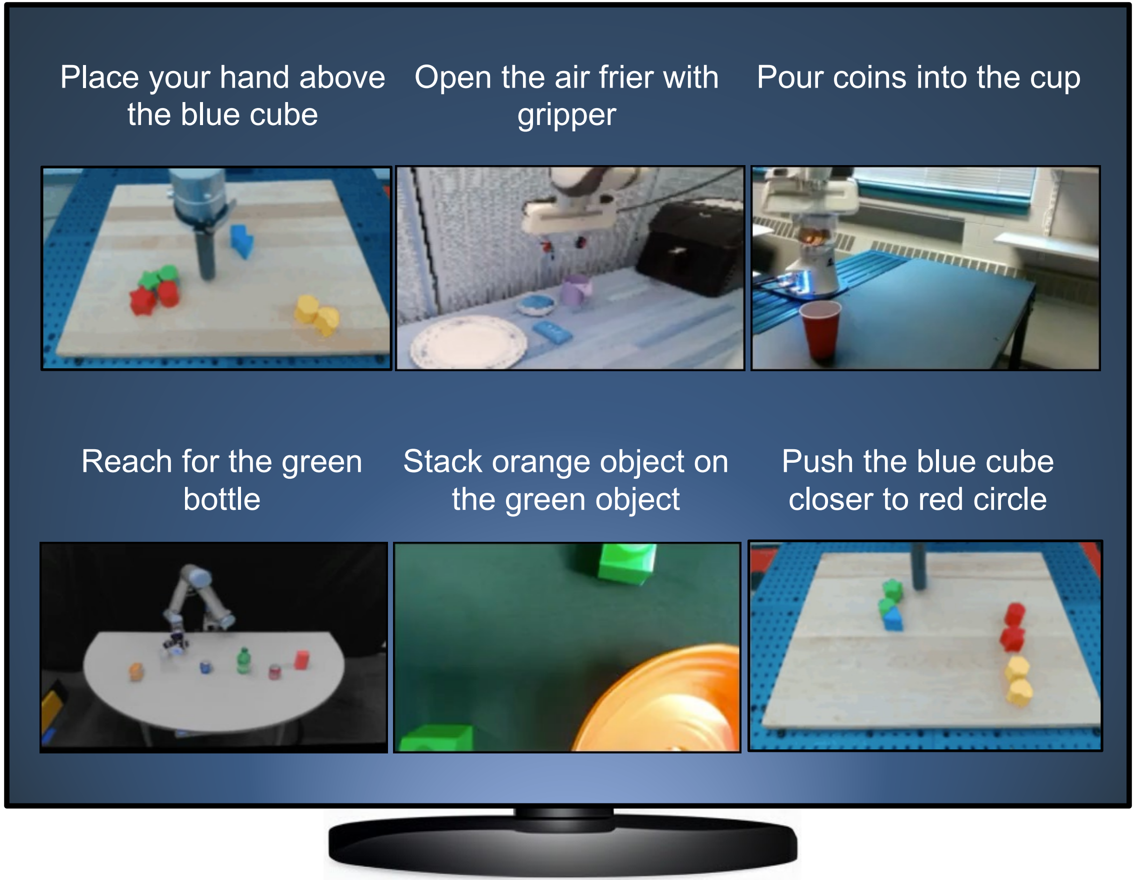
Simulating long sequence of robot executions.

Step 1:

[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

Multi-Task Planning with UniSim

Unified action & obs spaces



Place your hand above the blue cube

Open the air frier with gripper

Pour coins into the cup

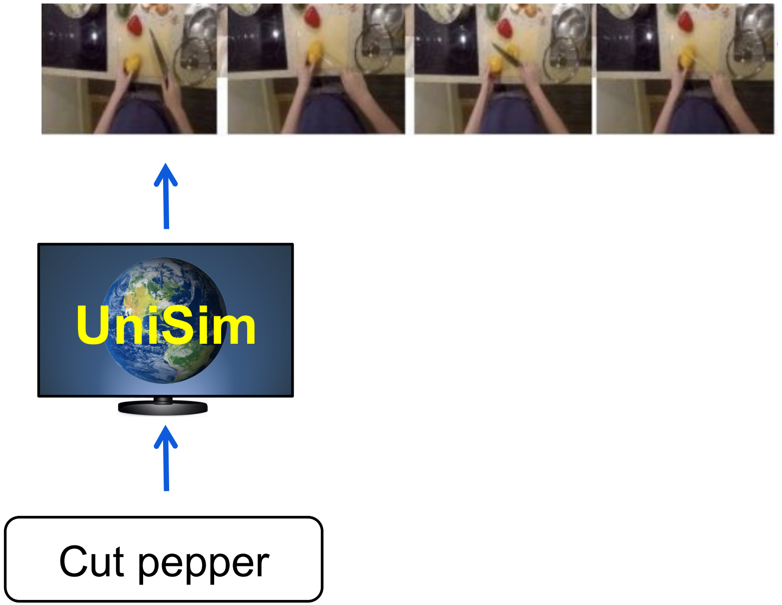
Reach for the green bottle

Stack orange object on the green object

Push the blue cube closer to red circle

[1] **Yang** et al. Video as the New Language for Real-World Decision Making. ICML 2024.

Generating Training Data for VLMs

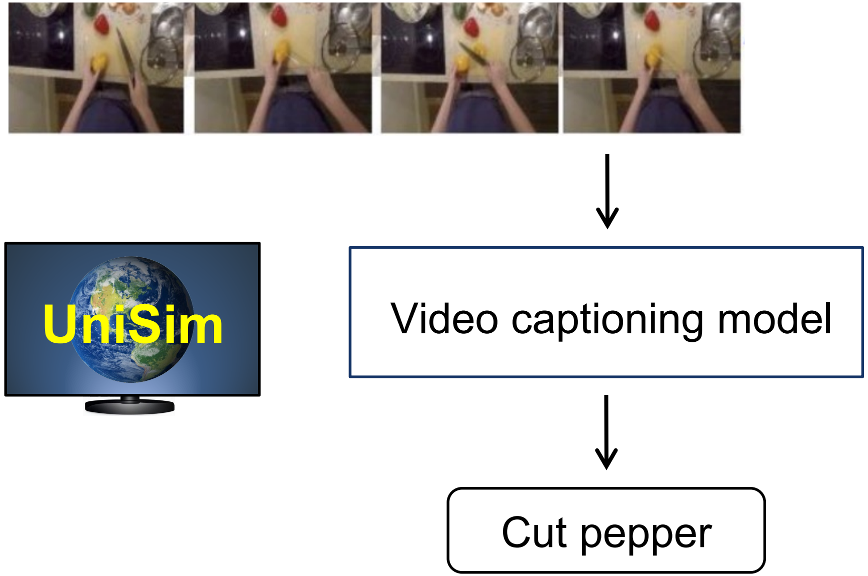


UniSim

Cut pepper

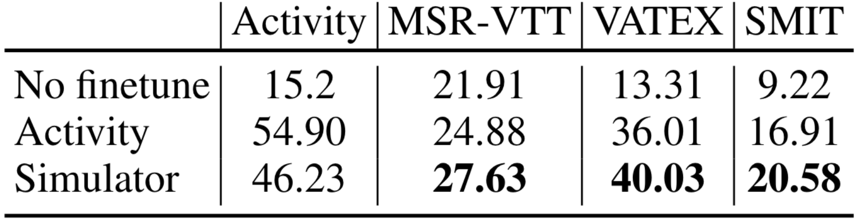
[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

Generating Training Data for VLMs



Video captioning model

Cut pepper



	Activity	MSR-VTT	VATEX	SMIT
No finetune	15.2	21.91	13.31	9.22
Activity	54.90	24.88	36.01	16.91
Simulator	46.23	**27.63**	**40.03**	**20.58**

Table 4: **VLM trained in the UniSim** to perform video captioning tasks. CIDEr scores for PaLI-X finetuned only on simulated data from the UniSim compared to no finetuning and finetuning on true video data from ActivityNet Captions. Finetuning only on simulated data has a large advantage over no finetuning and transfers better to other tasks than finetuning on true data.

[1] **Yang** et al. Learning Interactive Real-World Simulators. ICLR 2024.

Foundation Models as Real-World Simulators

✅ World model

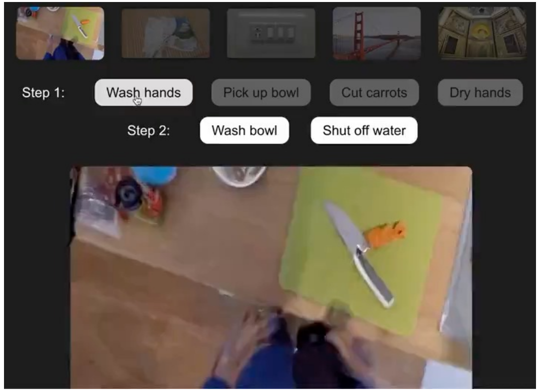
from internet data

✅ Algorithms

for decision making

☐ Challenges

and next steps



Takeaway: Unified repr & task interface

Takeaway: RL, planning in the world model

Foundation Models as Real-World Simulators

✅ World model

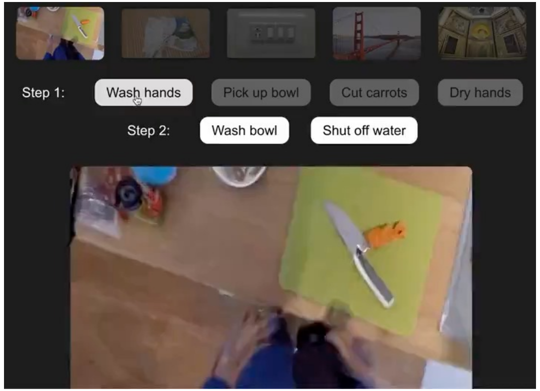
from internet data

✅ Algorithms

for decision making

□ Challenges

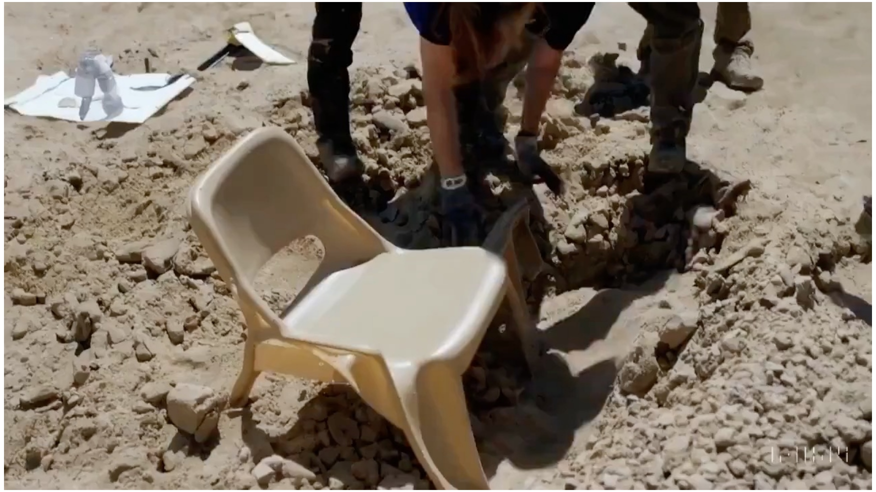
and next steps



Takeaway: Unified repr & task interface

Takeaway: RL, planning in the world model

Better World Models: Hallucination



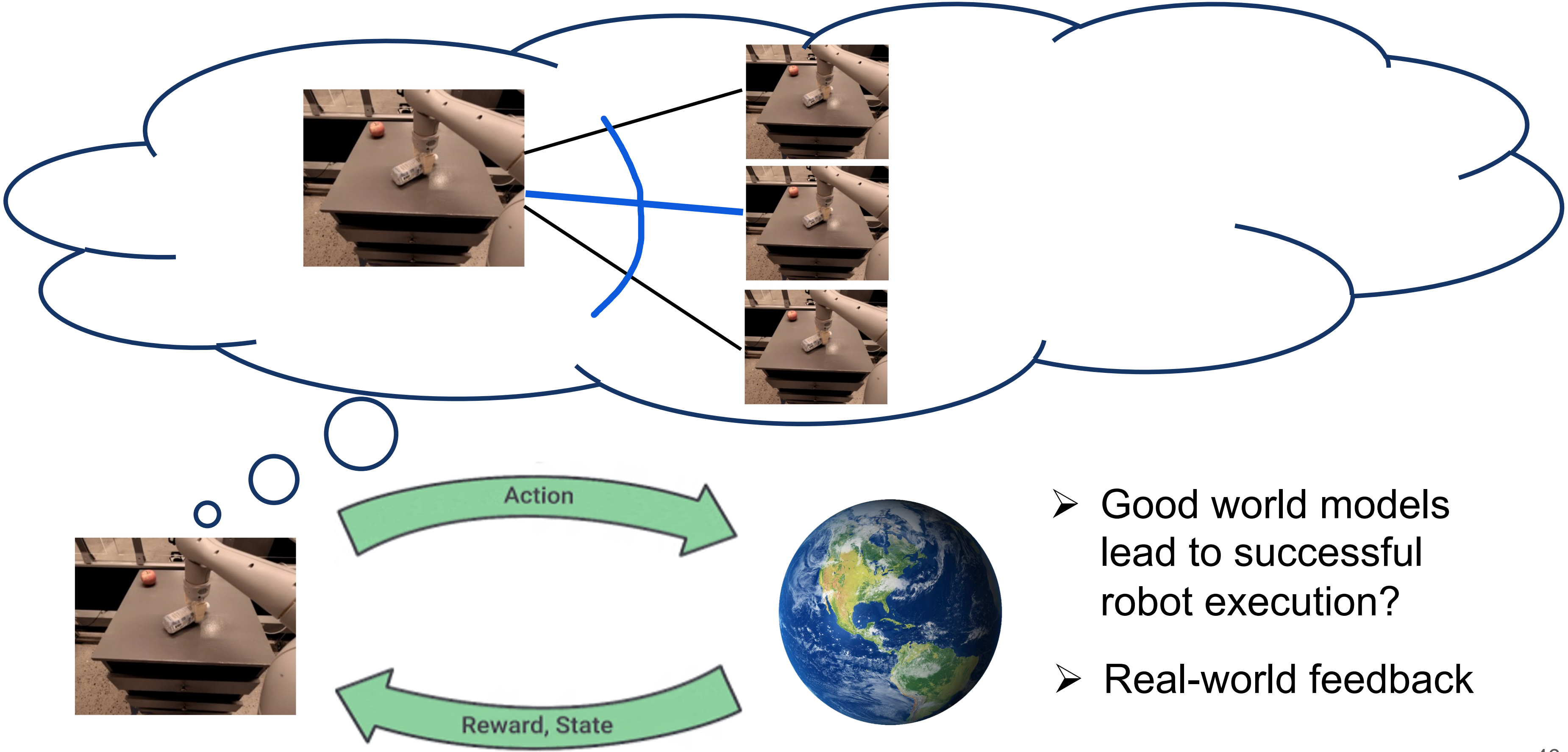
Better World Models: Hallucination

Better World Models: Hallucination

Text: Wash hands

Better World Models: Evaluation and Feedback

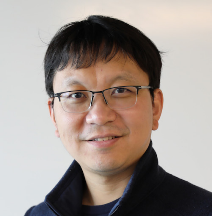
Better World Models: Evaluation and Feedback



- Good world models lead to successful robot execution?
- Real-world feedback

Collaborators

Yilun Du


Bo Dai

Hanjun Dai

Ofir Nachum

Kamyar
Ghasemipour

Jonathan Tompson

Leslie Kaelbling

Dale Schuurmans

Pieter Abbeel

& many others

Berkeley
UNIVERSITY OF CALIFORNIA

Google DeepMind

Thank You. Questions?