Foundation Models as Real-World Simulators

CVPR 2024 Workshop

Sherry Yang
Advances in Machine Learning

Outperforming humans in Go

Generating language, image, and video
Decision Making

Train a model to act in the world simulator
Decision Making and Internet-Scale Knowledge

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

$\quad$

$\quad$

$\quad$
What if We Can Learn a Realistic Simulator?

Definition: a learned simulator
Foundation Models as Real-World Simulators

✅ World model from internet data

✅ Algorithms for decision making

☑ Challenges and next steps

---

Foundation Models as Real-World Simulators

- World model from internet data
- Algorithms for decision making
- Challenges and next steps

Text as Unified Representation and Task Interface

Unified representation

- Common Crawl
- Wikipedia
- Stack Overflow
- GitHub
- Wolfram Alpha

Unified tasks

- Completion
- Translation
- Chatbot

Text is a unified representation of information

Text is a unified representation and task interface

Are you sure?

El texto es una representación unificada de información

Video as Unified Representation and Task Interface

Unified representation

Unified tasks

Videos

Video generation

Text

Video as Unified Representation and Task Interface

Unified representation

Unified tasks

Learned how to “perform” tasks

Video generation

Cut pepper

Video as Unified Representation and Task Interface

Unified representation

Unified tasks

Learned real-world “physics”

Video generation

Video as Unified Representation and Task Interface

Unified representation

Unified tasks

Videograms

Learned simulated “dynamics”

Video generation

Δx, Δy

Video as Unified Representation and Task Interface

Unified representation

Unified tasks

Video as Unified Representation and Task Interface

Unified representation

Learned notions of objects/scenes

Unified tasks

Video generation

A person throwing a frisbee

Background: Image Diffusion Models

Add Gaussian noise $\epsilon$

Learn reverse schedule $\epsilon_\theta$ ("robot")

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

Denoise by subtracting $\epsilon_\theta$ ("human")

Adapting Diffusion for World Modeling

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

Adapting Diffusion for World Modeling

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

21M videos, 800M images

UniSim: An Interactive Real-World Simulator

Step 1: Wash hands Pick up bowl Cut carrots Dry hands

Step 2: Wash bowl Shut off water

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

Takeaway: Unified repr & task interface

Questions:

- Challenges
- and next steps
Reinforcement Learning with UniSim

\[ \Delta x, \Delta y \]

Policy

Action

Reinforcement Learning

Simulator

Reward, State

Real world

Push the red hexagon towards the blue cube

\[ \times \box{Fail to transfer from sim to real} \]

Reinforcement Learning with UniSim

$$\Delta x, \Delta y$$

Policy

Action

Reinforcement Learning

Reward, State

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

Reinforcement Learning with UniSim

<table>
<thead>
<tr>
<th></th>
<th>Succ. rate (all)</th>
<th>Succ. rate (pointing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLA-BC</td>
<td>0.58</td>
<td>0.12</td>
</tr>
<tr>
<td>UniSim-RL</td>
<td><strong>0.81</strong></td>
<td><strong>0.71</strong></td>
</tr>
</tbody>
</table>

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.

Reinforcement Learning with UniSim

Policy $\Delta x, \Delta y$

Reinforcement Learning

Action

Simulator

Place your hand above the blue cube

Real world

Task: Push the red star towards the blue cube

✅ Transfer from sim to real

Planning with UniSim

Synthesized video

Robot execution

Put the fruits into the top drawer

Inverse Dynamics

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


Planning with UniSim

Vision language model

Put the fruits into the top drawer

Action 1. Open top drawer

Action 1. Place banana in top drawer
Planning with UniSim

UniSim

Action 1. Open top drawer

Put the fruits into the top drawer

Action 1. Place banana in top drawer
Planning with UniSim

Vision-language reward model

Put the fruits into the top drawer
Planning with UniSim

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Reward</th>
<th>Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniPi</td>
<td>44.0</td>
<td>4%</td>
</tr>
<tr>
<td>LAVA</td>
<td>33.5</td>
<td>0%</td>
</tr>
<tr>
<td>RT-2</td>
<td>36.5</td>
<td>2%</td>
</tr>
<tr>
<td>PALM-E</td>
<td>26.2</td>
<td>0%</td>
</tr>
<tr>
<td>VLP</td>
<td>65.0</td>
<td>16%</td>
</tr>
</tbody>
</table>
Planning with UniSim – Why?

Language instructions

Make a line

Intermediate goals

Predict intermediate frames

Robot actions

\( a_1, a_2, a_3 \quad a_4, a_5, a_6, a_7 \)

<table>
<thead>
<tr>
<th>Model</th>
<th>Reward</th>
<th>Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniPi</td>
<td>44.0</td>
<td>4%</td>
</tr>
<tr>
<td>LAVA</td>
<td>33.5</td>
<td>0%</td>
</tr>
<tr>
<td>RT-2</td>
<td>36.5</td>
<td>2%</td>
</tr>
<tr>
<td>PALM-E</td>
<td>26.2</td>
<td>0%</td>
</tr>
<tr>
<td>VLP</td>
<td>65.0</td>
<td>16%</td>
</tr>
</tbody>
</table>
Planning with UniSim – Why?

**Language instructions**

Make a line

**Step-by-step plans**

Action 1. push red star to left ...

Action 2. move green star to ...

**Intermediate goals**

**Robot actions**

\[ a_1, a_2, a_3 \]

\[ a_4, a_5, a_6, a_7 \]

**Benefits:**

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

**Make Line**

<table>
<thead>
<tr>
<th>Model</th>
<th>Reward</th>
<th>Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniPi</td>
<td>44.0</td>
<td>4%</td>
</tr>
<tr>
<td>LAVA</td>
<td>33.5</td>
<td>0%</td>
</tr>
<tr>
<td>RT-2</td>
<td>36.5</td>
<td>2%</td>
</tr>
<tr>
<td>PALM-E</td>
<td>26.2</td>
<td>0%</td>
</tr>
<tr>
<td>VLP</td>
<td>65.0</td>
<td>16%</td>
</tr>
</tbody>
</table>
Long-Horizon Planning with UniSim

Simulating long sequence of robot executions.

Step 1:

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

Generating Training Data for VLMs

Generating Training Data for VLMs

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.

Foundation Models as Real-World Simulators

✅ World model from internet data

✅ Algorithms for decision making

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
Thank You. Questions?