Problem Statement

- Collecting real world data is costly. Simulators can cheaply generate abundant data.
- To use sim data to train real world policies, we need to overcome the sim2real gap.
- Common approaches to do so (domain rand., manual sys. ID) are expensive & tedious.
- Can we instead improve sim2real transfer by leveraging natural language to learn domain-invariant visual representations?

Insight: Semantically Similar Images $\rightarrow$ Similar Actions

Both Images: “Gripper holding carrot above yellow square.”

Both Actions: Open gripper to place obj.

We want sim+real images with similar semantics to have similar representations for the policy to predict similar action distributions.

Our Approach

1. Image-Language Pretraining

2. Multitask, Multidomain Imitation Learning

How Do We Label Images with Language Descriptions at Scale?

We automatically label trajectory images with templated annotations either during scripted policy data collection, or with a VLM afterwards.

Tasks

Sim2Real Results

Our method outperforms all baselines across decreasing data regimes (columns $\rightarrow$) and increasing task difficulty and sim2real gap (rows 1). 

<table>
<thead>
<tr>
<th>Policy (sim)</th>
<th>Sim2Real</th>
<th>Pick+Place</th>
<th>Multi-step Pick+Place</th>
<th>Wrap Wire (Deformable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sim</td>
<td>Real</td>
<td>100</td>
<td>50</td>
<td>25</td>
</tr>
<tr>
<td>Multi-step</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrap Wire</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Main Takeaways

1. Language can bridge wide sim2real gaps with domain-invariant representations.
2. Our method enables leveraging low-fidelity sim data for sim2real transfer on deformable objects.