Natural Language Can Help Bridge the Sim2Real Gap

Albert Yu, Adeline Foote, Ray Mooney, Roberto Martín-Martín
University of Texas at Austin | albertyu@utexas.edu

Web

Problem Statement

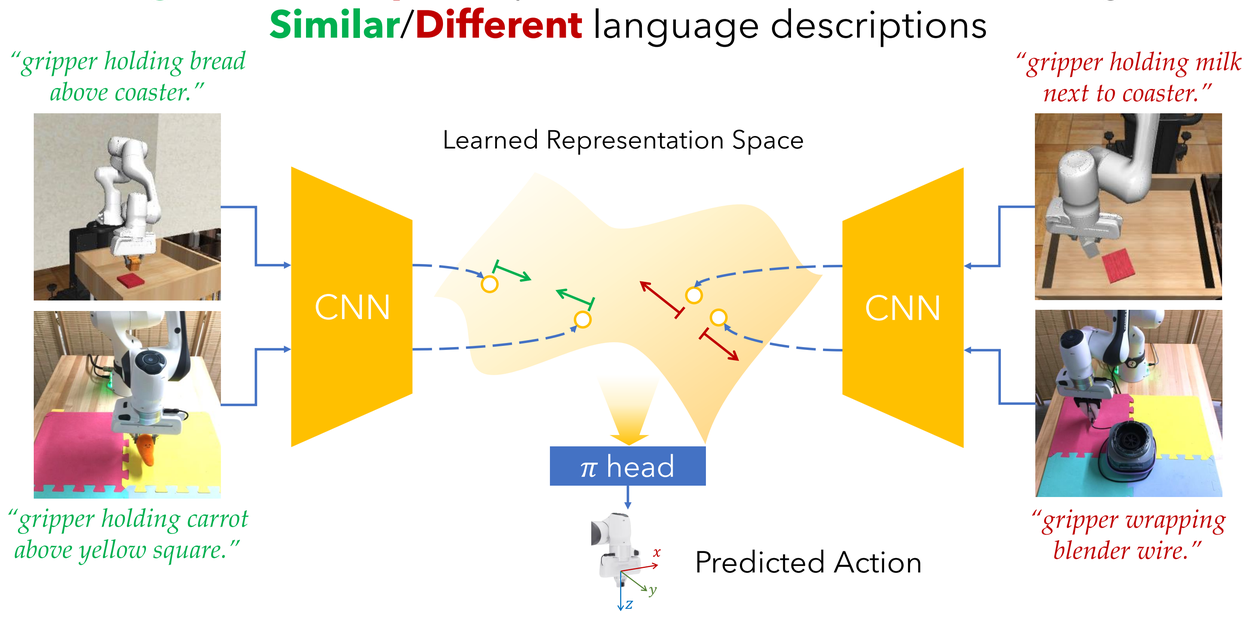
- Collecting real world data is costly. Simulators can cheaply generate abundant data.
- To use sim data to train real world policies, we need to overcome the sim2real gap.
- Common approaches to do so (domain rand., manual sys. ID) are expensive & tedious.
- Can we instead improve sim2real transfer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

Insight: Semantically Similar Images → Similar Actions

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

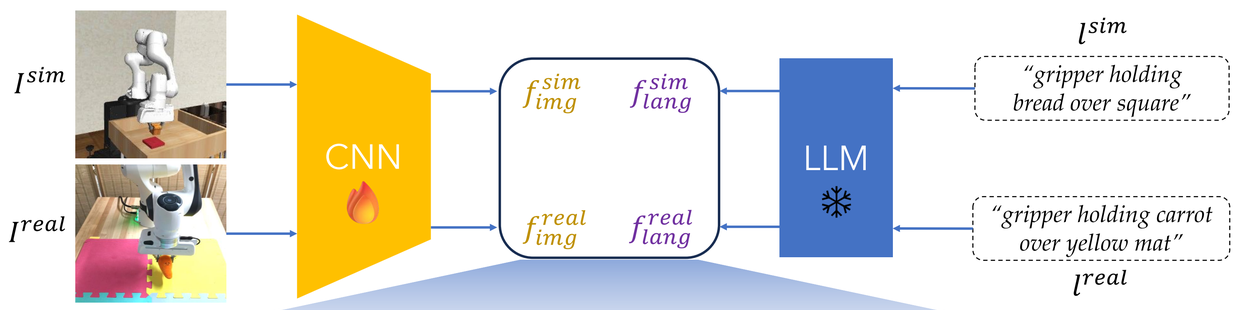
Language Links Sim+Real Visual Features via Semantic Similarity, Improving Sim2Real Transfer with Visuomotor Policies

Push together/**Pull apart** representations of sim and real images with **Similar**/**Different** language descriptions

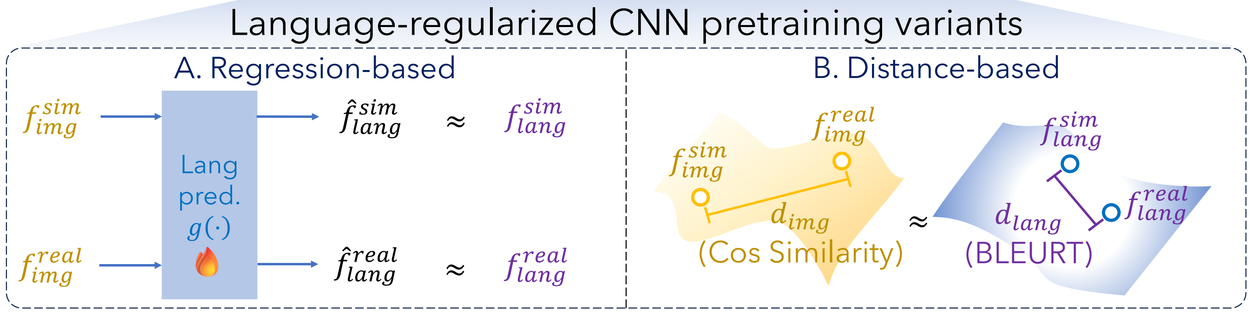


Our Approach

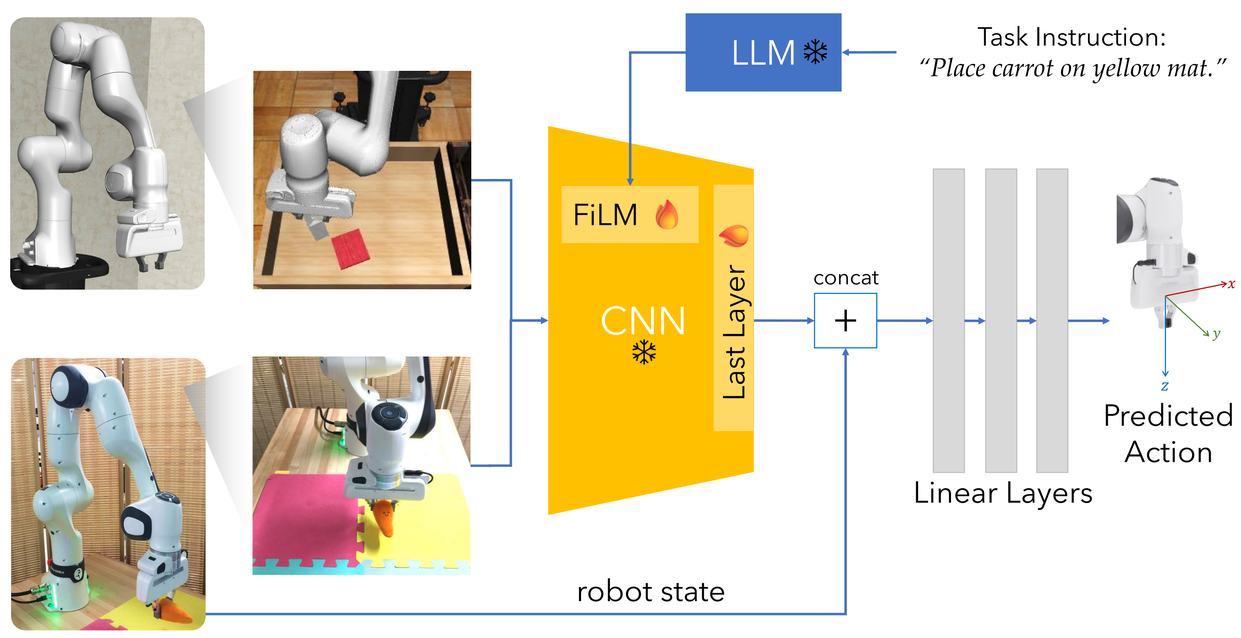
① Image-Language Pretraining



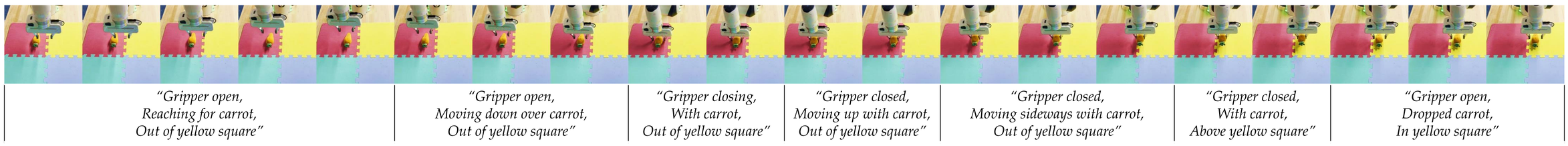
Language-regularized CNN pretraining variants



② Multitask, Multidomain Imitation Learning

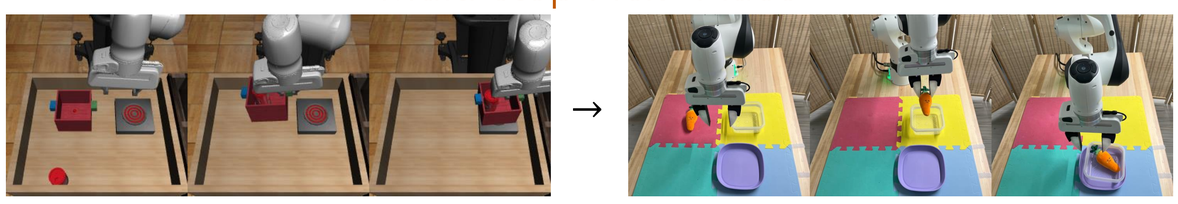


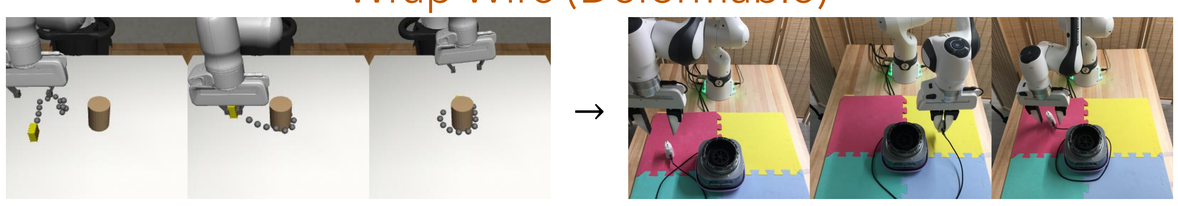
How Do We Label Images with Language Descriptions at Scale?



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 →) and increasing task difficulty and sim2real gap (rows ↓).

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.