Teaching Robots with Show and Tell: Using Foundation Models to Synthesize Robot Policies from Language and Visual Demonstrations

Michael Murray • Abhishek Gupta • Maya Cakmak

University of Washington

Overview

We introduce a modular, neuro-symbolic framework for teaching robots new skills through language and visual demonstration

Our approach, ShowTell, composes a mixture of foundation models to synthesize robot policies that are easy to interpret and generalize across a wide range of tasks and environments.

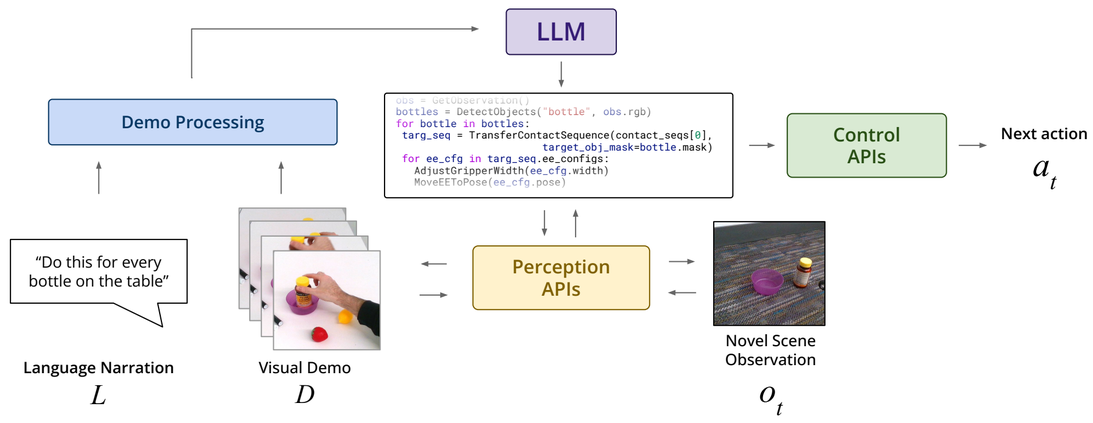

Figure 1. An overview of the ShowTell framework. First, the visual and spoken components of the demo are processed. An LLM synthesizes a modular program that can jointly reason about a provided demonstration and novel observations to transfer the demonstrated skill to new scenes.

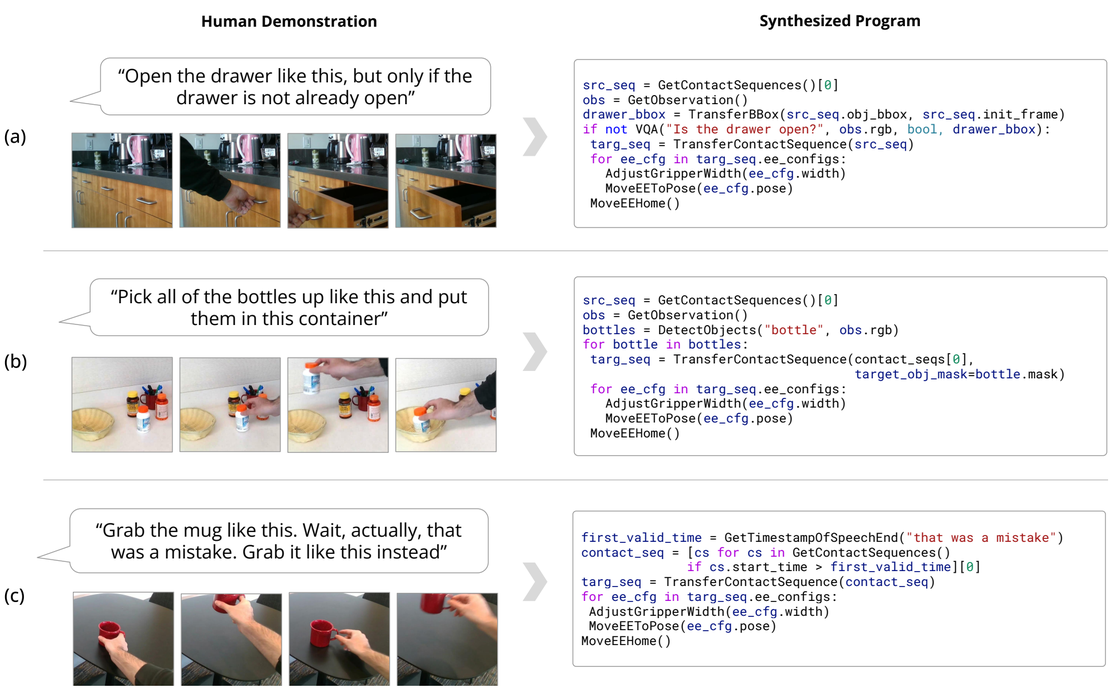

Figure 2. Examples of code synthesized by ShowTell for a set of representative demonstrations, showing the ability to follow high level logic including (a) conditionals (b) iteration and (c) segmentation.

Perception

`GetHandLandmarks, GetHandObjContactSequences, GetGestureRef, DetectObjects, TrackMask, VQA`

Novel Scene Transfer

`GetObservation, TransferBBox, TransferMask, TransferEEConfig, TransferContactSequence, AnyGrasp`

Speech Alignment

`GetTimestampOfSpeechStart, GetTimestampOfSpeechEnd`

Robot Control

`MoveEEToPose, MoveEERel, MoveEEGuarded, MoveEEHome, RotateEE, AdjustGripperWidth`

Figure 3. Taxonomy of modules available to the synthesized programs.

Diffusion-based Novel Scene Transfer

Figure 4. Diffusion features are used to transfer affordances across novel viewpoints, novel objects, and novel scenes.

Real World Robot Evaluation

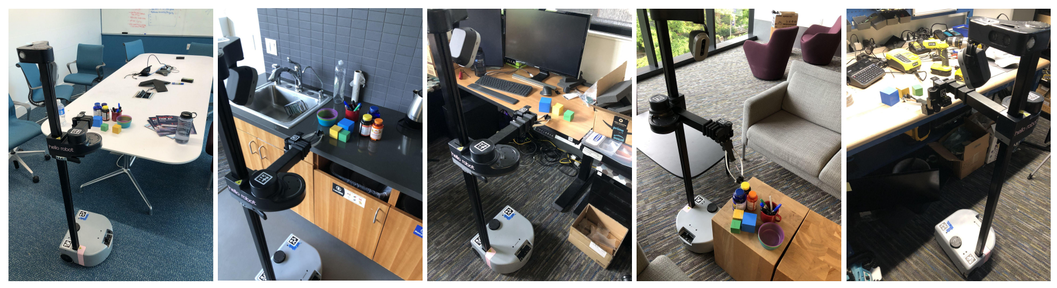

Figure 5. We evaluate using a Stretch RE2 robot to perform 16 real world manipulation tasks across 5 visually distinct environments

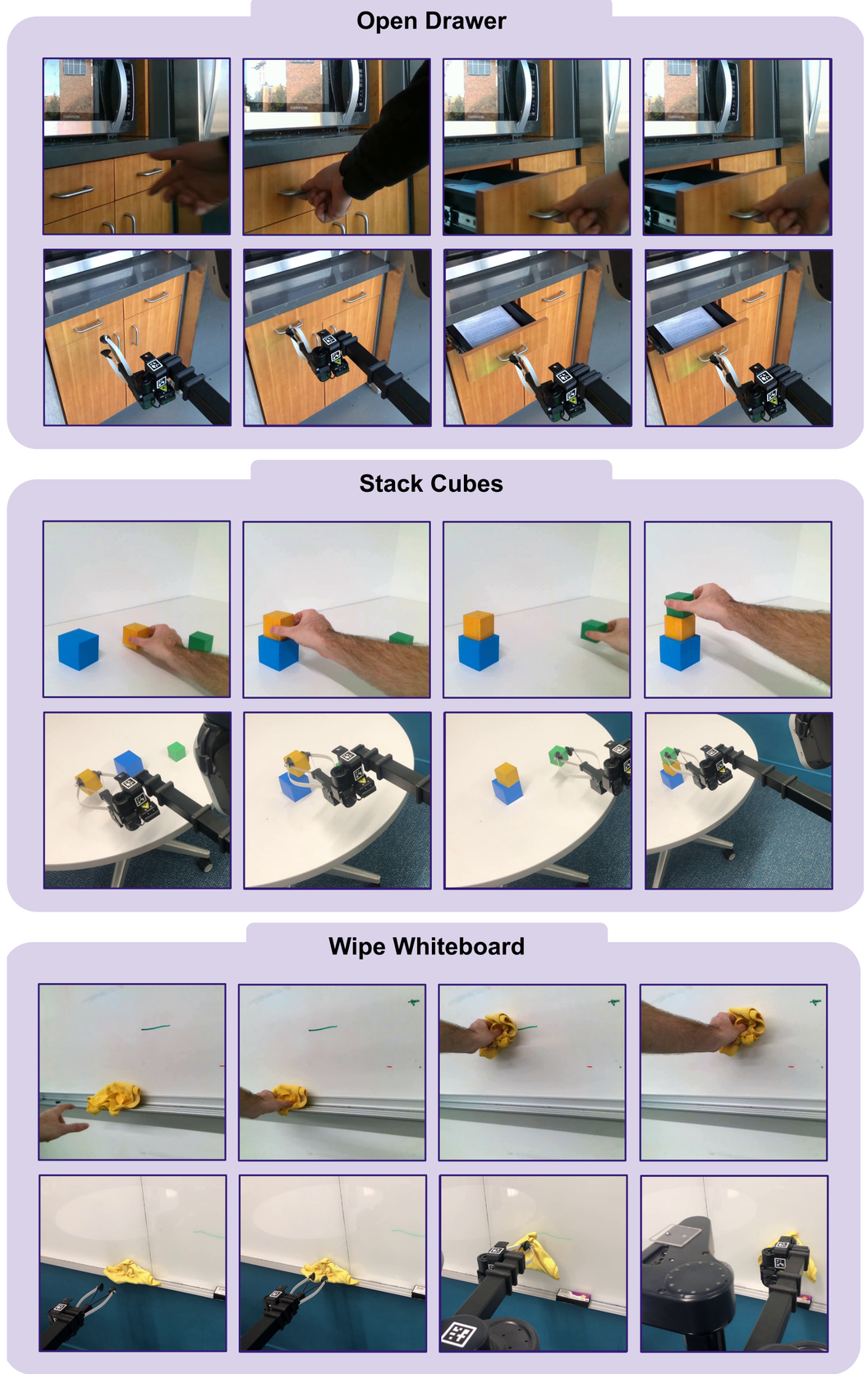

Figure 6. Representative human demonstrations and robot executions.

Quantitative Results

Demo Type	ShowTell-NoLang GCR	ShowTell-NoLang SR	ShowTell-NoVis GCR	ShowTell-NoVis SR	GPT4-V-Robot GCR	GPT4-V-Robot SR	ShowTell (ours) GCR	ShowTell (ours) SR
Simple	0.89	0.85	0.86	0.81	0.88	0.85	0.96	0.94
Iterative	0.31	0.12	0.61	0.59	0.43	0.22	0.94	0.85
Conditional	0.41	0.39	0.64	0.64	0.41	0.41	0.93	0.93
Segmented	0.12	0.06	0.48	0.42	0.20	0.18	0.93	0.91

Download

Get the code and supplementary video:

Learn more

Read the full pre-released paper: