https://tinyurl.com/mutex-corl23

Learning Unified Policies from Multimodal Task Specifications

Rutav Shah, Roberto Martín Martín*, Yuke Zhu*

* Equal Advising

UT Robot Perception & Learning Lab

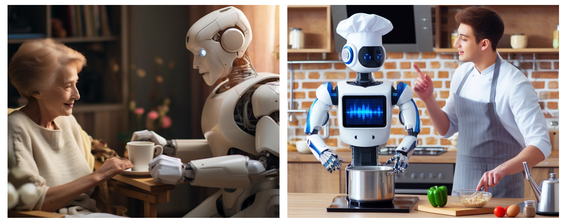
The University of Texas at Austin

RobIn ROBOT INTERACTIVE INTELLIGENCE LAB

TEXAS Robotics

One model that can interpret and execute tasks specified by humans in *any* modality.

Motivation



Effective Human-Robot Collaboration

Human Video Demonstration

Instructions

"Let's go to the kitchen. Get the pot from the cabinet. Carefully put it on the stove."

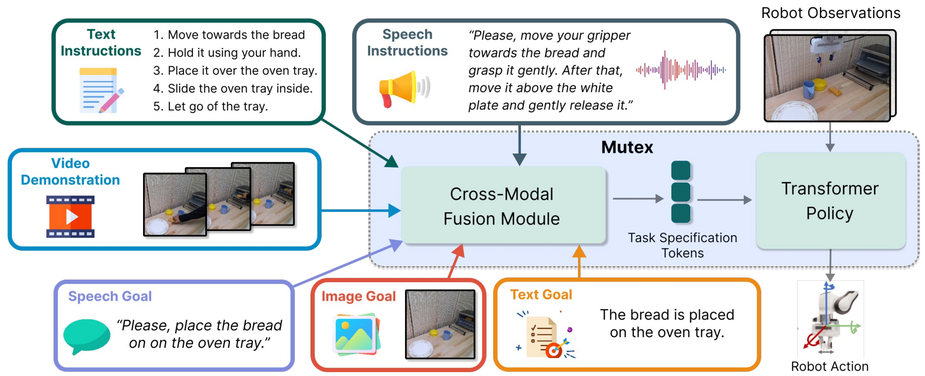
Image Goal

Goals

"Arrange the bread on the plate"

Human-Robot Interface must support multiple modalities.

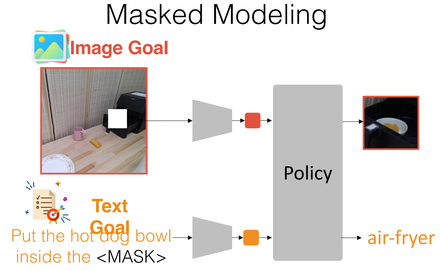
Key Idea



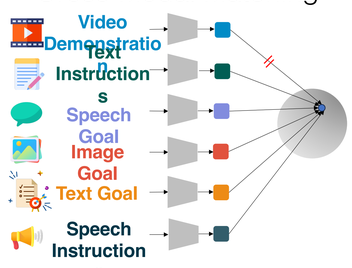
Training: Leverage all modalities. Evaluation: Use *any* modality.

Technical Overview

Masked Modeling

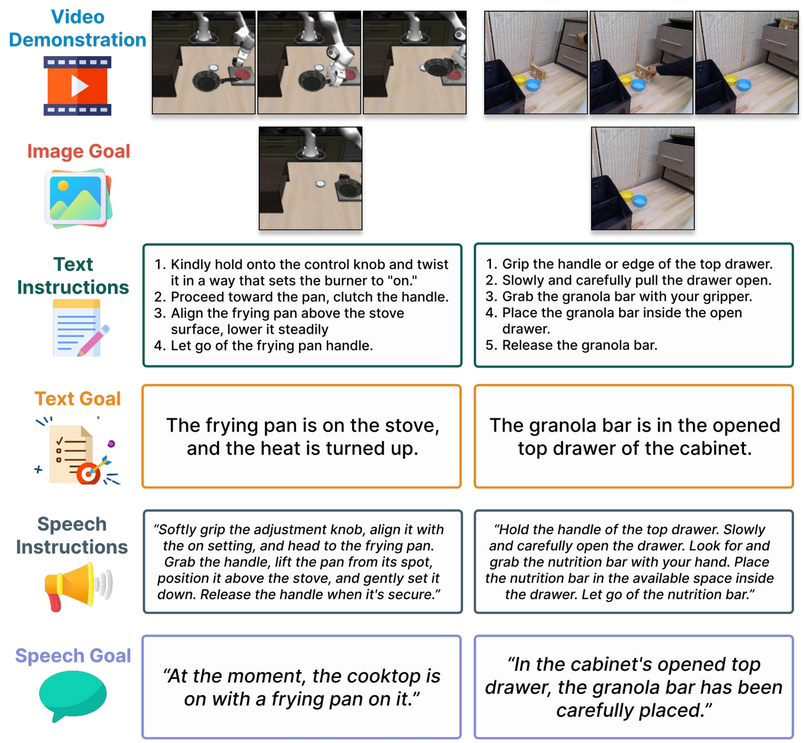


Cross-Modal Matching



Leverage cross-modal learning with behavior cloning to learn action-relevant representations better than training each modality individually.

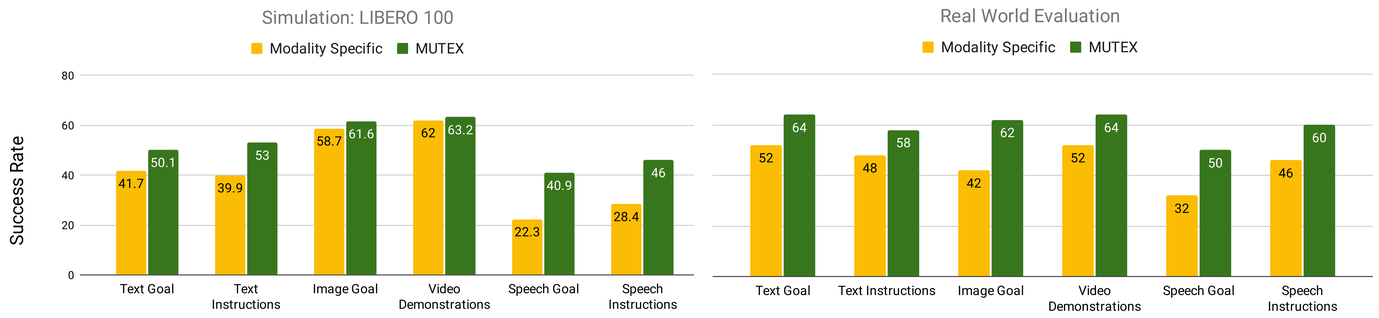
Dataset Visualization



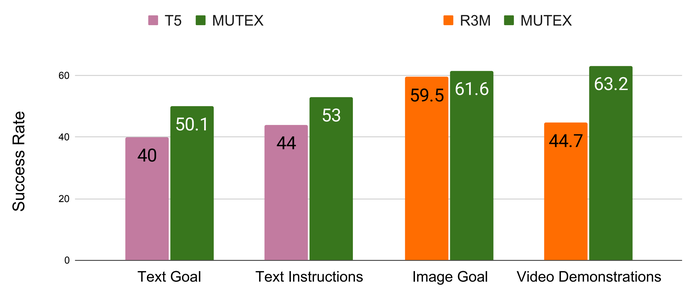
	# Tasks	# Expert Demonstrations	# Task Specifications per Task (train/test)
Real World	50	30	8/3
LIBERO-100	100	50	8/3

Experimental Results

MUTEX trained using all modalities consistently outperforms separate models trained for each modality.



MUTEX outperforms single-modality pretrained representations.



MUTEX can use multiple modalities during evaluation!