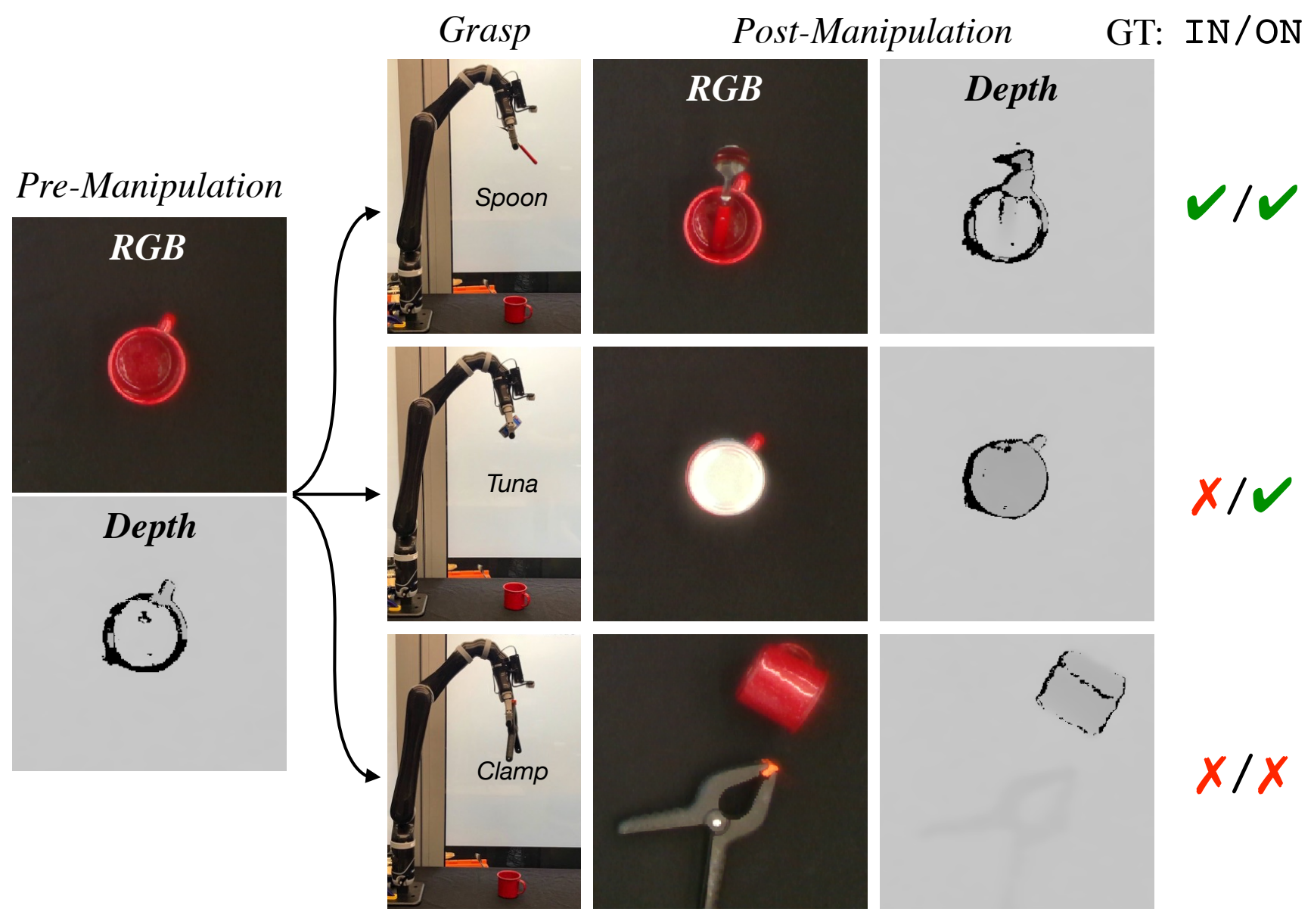
Improving Robot Success Detection using Static Object Data

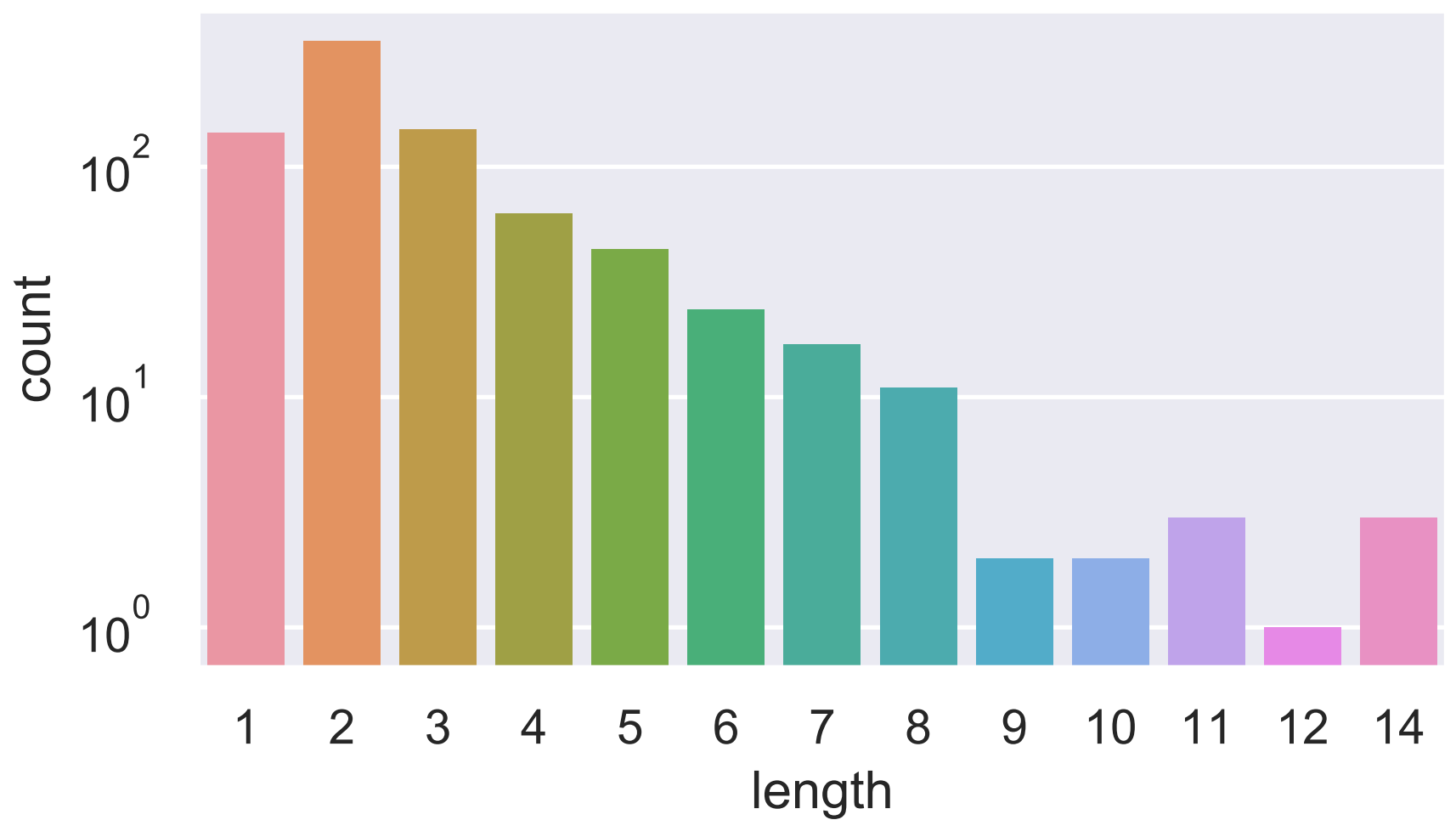
Rosario Scalise, **Jesse Thomason**, Yonatan Bisk, and Siddhartha Srinivasa
University of Washington

Success Detection

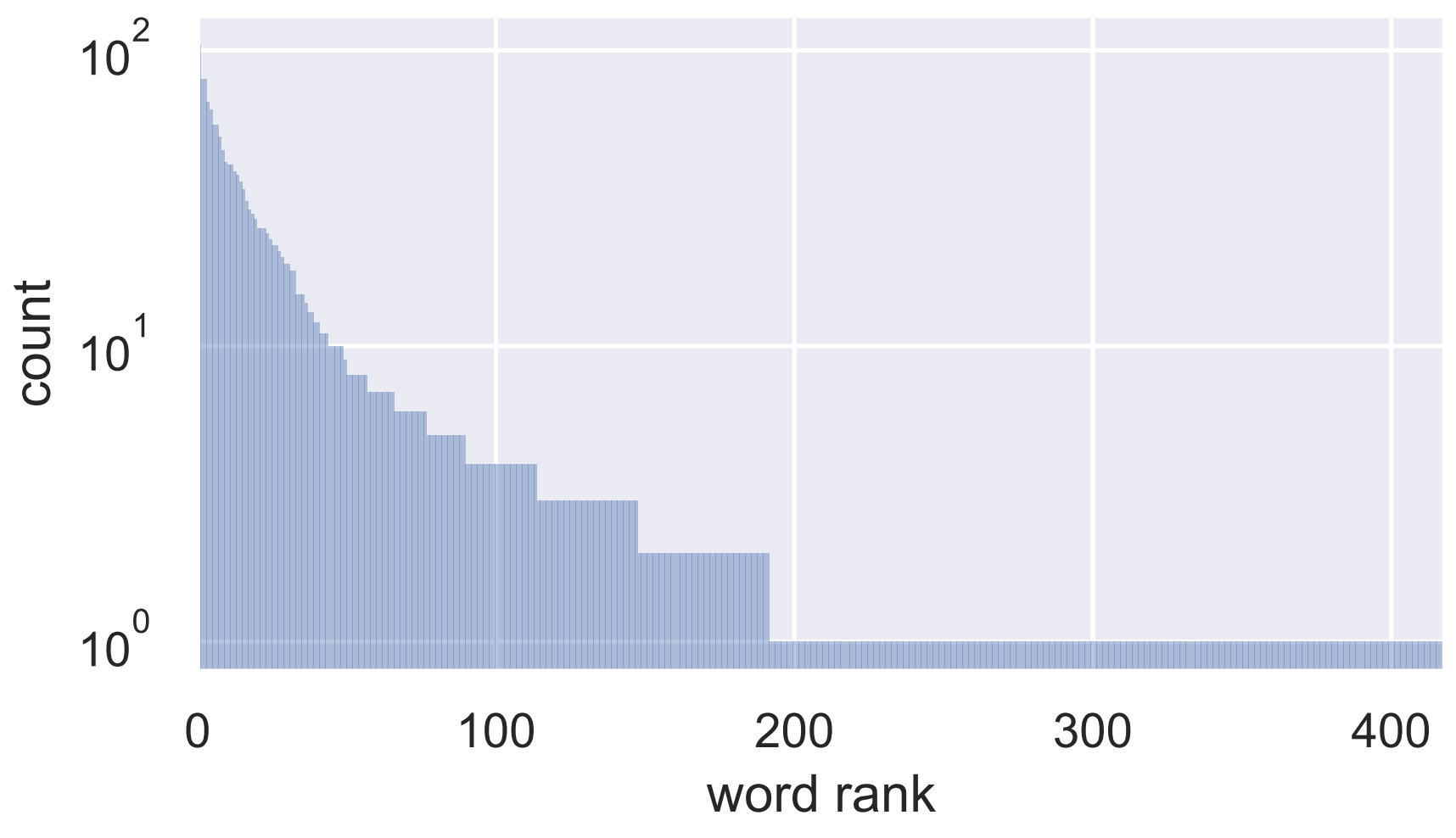
Stacking and nesting actions are necessary for robots clearing a dining table or packing a bin. Using an RGB-D camera to detect success is insufficient: same-colored objects can be difficult to differentiate, and reflective silverware cause noisy depth camera perception. We collect over 13 hours of egocentric manipulation data and show that adding static data about the objects themselves improves the performance of an end-to-end pipeline for classifying action outcomes.



The task is to detect whether dropping one object onto another resulted in the first being *in* or *on* the second using RGB-D scans of the workspace pre- and post- action.



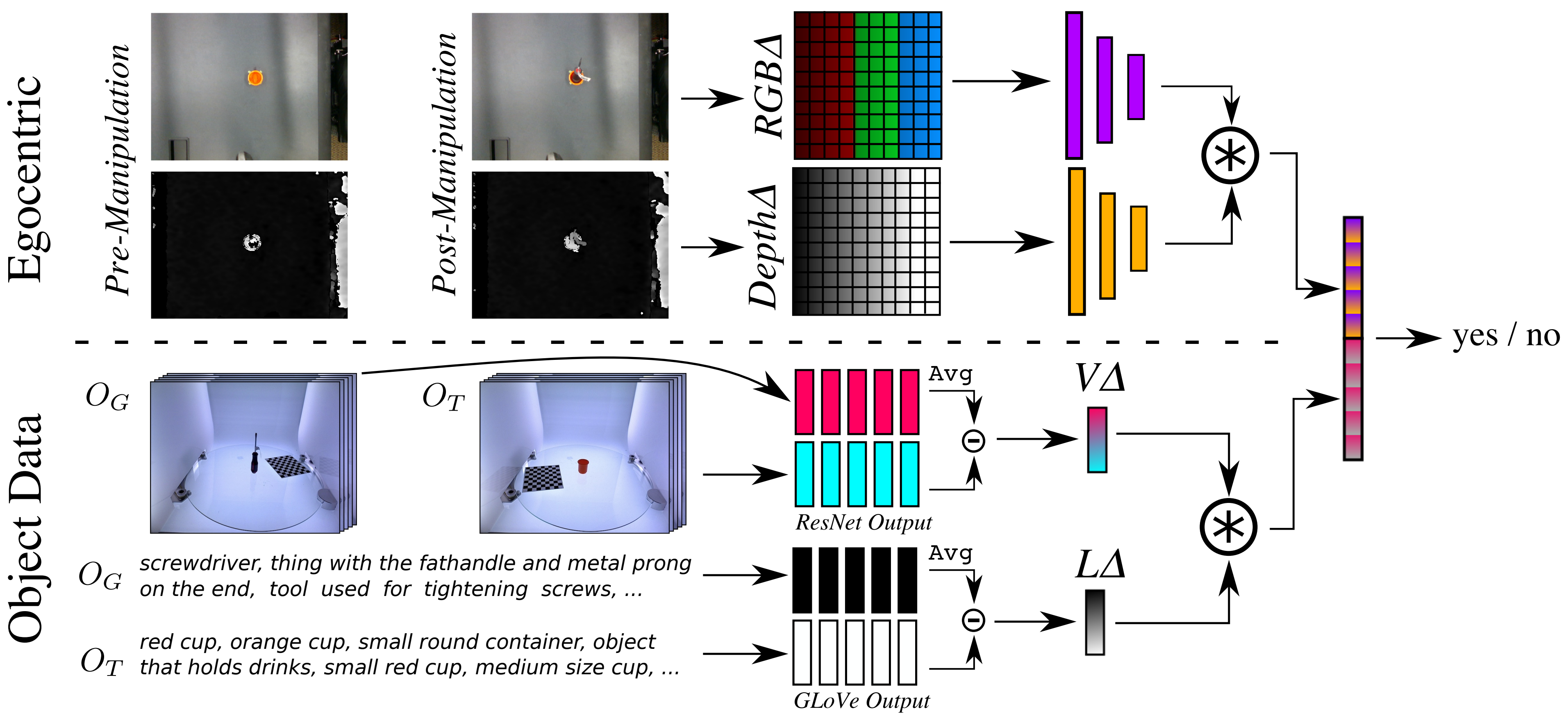
Referring expressions vary in length but are mostly short.



Words are Zipfian distributed across the referring expressions.

Punchline

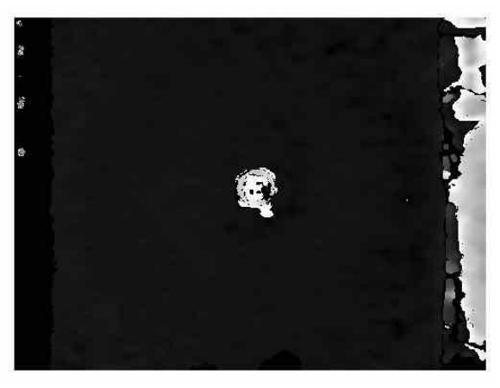
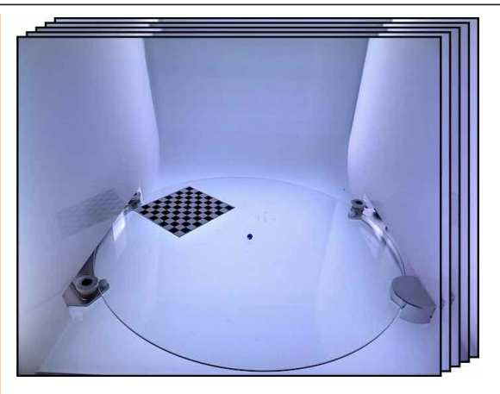
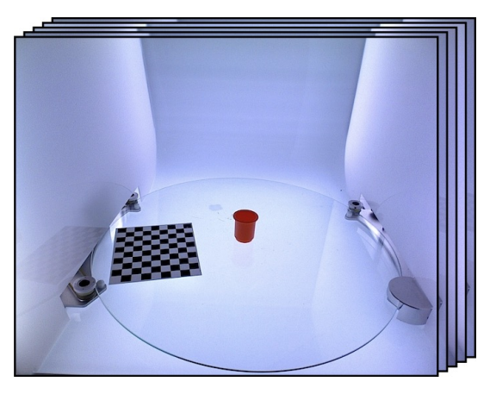
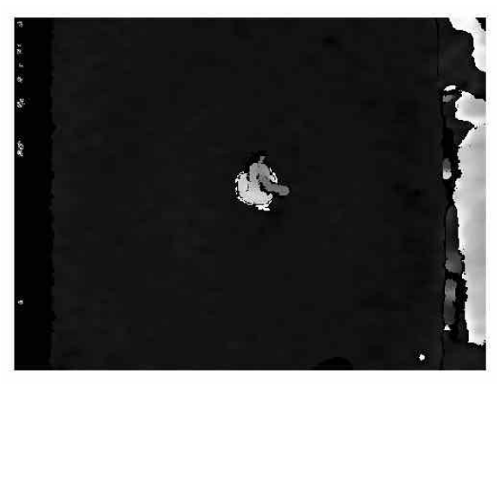
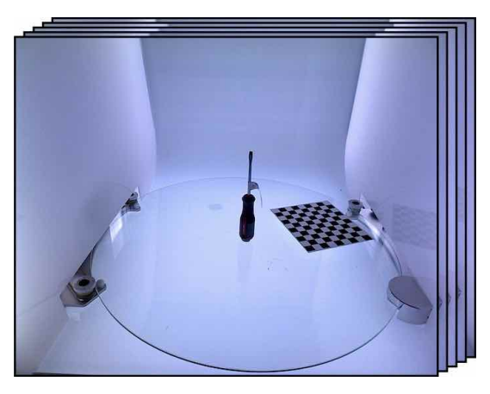
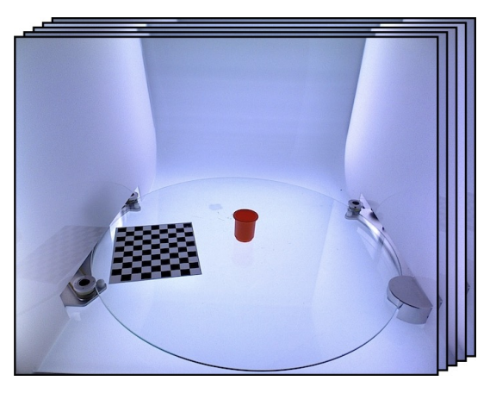
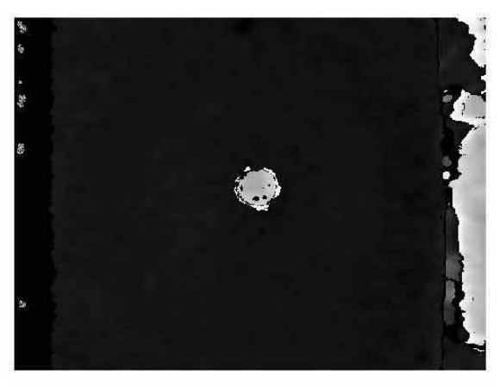
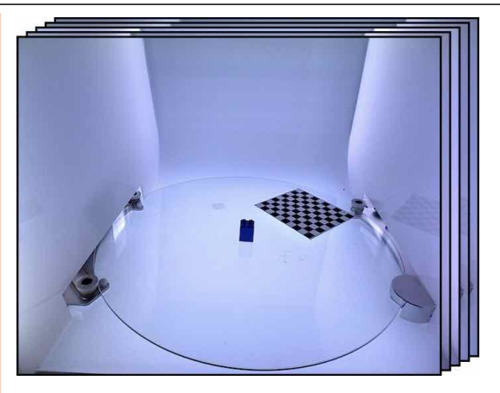
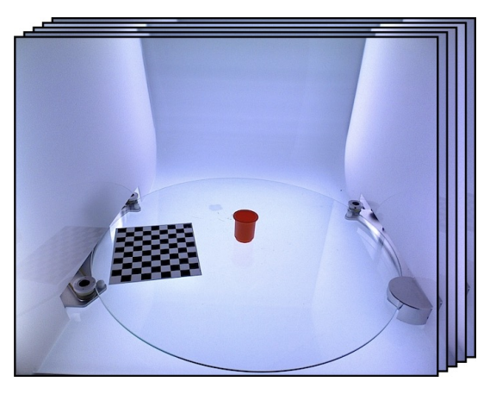
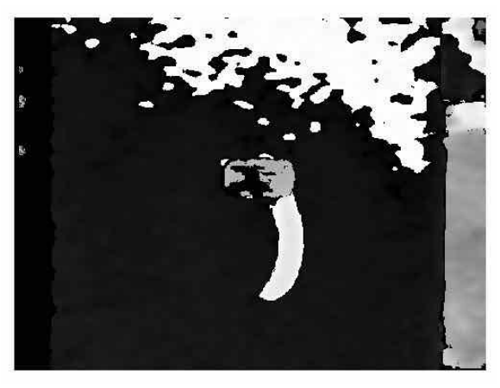
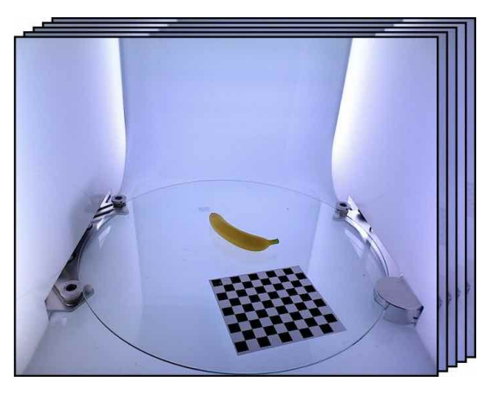
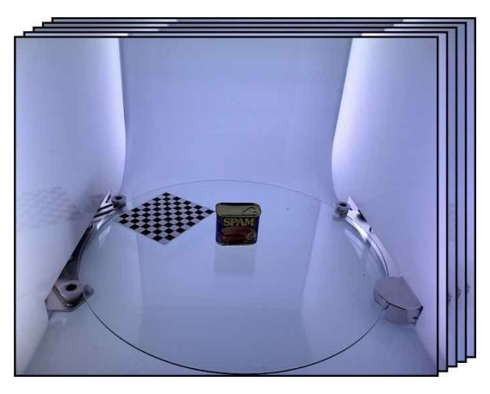
Referring expressions and pictures of individual objects improve robot success detection for object stacking and nesting!



O_G screwdriver, thing with the fathandle and metal prong on the end, tool used for tightening screws, ...

O_T red cup, orange cup, small round container, object that holds drinks, small red cup, medium size cup, ...

The full architecture takes egocentric visual input, vision embeddings from multiple static image viewpoints of each object, and language embeddings from referring expressions for each object.

Egocentric ⇒	Pred	Egocentric + Pretrained Object		⇒ Pred	Truth
 	✗In	 	*small black sphere, round black item, small marble, the blue object, round object, tiny object, tiny dot, blue round object, little ball* *red cup, orange cup, small round container, object that holds drinks, small red cup, red cup, medium size cup without handles, red plastic thing, red cylinder*	✓In	✓In
 	✗In	 	*screwdriver, thing with the fat handle and metal prong on the end, tool used for tightening screws, screw driver with long tip, screwdriver, plastic handle screw driver, non phillips screw driver, tool, black screwdriver* *red cup, orange cup, small round container, object that holds drinks, small red cup, red cup, medium size cup without handles, red plastic thing, red cylinder*	✓In	✓In
 	✗On	 	*blue thing, blue plastic rectangle, blue plastic block, blue cube, lego piece, blue plastic thing, blue block, small square block, little blue block* *red cup, orange cup, small round container, object that holds drinks, small red cup, red cup, medium size cup without handles, red plastic thing, red cylinder*	✓On	✓On
 	✗On	 	*yellow thing, long yellow item, soft yellow thing, yellow curved cylinder, yellow fruit, the object that is mostly yellow with slight green at one of the tips, yellow long fruit, yellow banana, banana* *spam, canned meat, metal can, can of spam, aluminum cube, blue and gold cube, rectangular can, square, glass circle*	✓On	✗On

Evaluation and Ablations

The dataset consists of pairs of YCB objects *Y* and containers *C*, split into folds. For a subset of pairs, we have egocentric, **Robot** manipulation data.

	Objects		Robot		All	
Fold	*Y*	*C*	*in*	*on*	*in*	*on*
Train	51	17	191	191	800	2500
Dev	20	5	47	58	100	400
Test	19	6	60	60	114	361

✓indicates signal was included, while "pre" indicates models with object features pretrained from **All Pairs** of available objects.

		Model (*M*)		Detection Correct ↑	
		Object Data			
	Ego	Lang	Vis	*in*	*on*
Dev Fold	$\vec{0}$	✓	$\vec{0}$	.70 ± .03	.56 ± .10
	$\vec{0}$	pre	$\vec{0}$	.72 ± .04	.57 ± .09
	$\vec{0}$	$\vec{0}$	✓	.71 ± .08	.50 ± .06
	$\vec{0}$	$\vec{0}$	pre	.72 ± .07	.53 ± .05
	$\vec{0}$	✓	✓	.76 ± .08	.58 ± .05
	$\vec{0}$	pre	pre	**.78** ± .08	.60 ± .04
	✓	✓	$\vec{0}$	.67 ± .08	.60 ± .08
	✓	pre	$\vec{0}$	.68 ± .08	**.62** ± .08
	✓	$\vec{0}$	✓	.70 ± .10	.58 ± .11
	✓	$\vec{0}$	pre	.72 ± .08	.59 ± .13
	✓	✓	✓	.70 ± .09	.59 ± .07
	✓	pre	pre	.73 ± .09	.62 ± .07
	Baseline (MC)			.32 ± .00	.36 ± .00
	Baseline (Rand)			.49 ± .06	.50 ± .06
Test Fold	$\vec{0}$	✓	$\vec{0}$	.79 ± .02	.45 ± .05
	$\vec{0}$	pre	$\vec{0}$	.79 ± .02	.48 ± .07
	$\vec{0}$	$\vec{0}$	✓	**.80** ± .04	.46 ± .09
	$\vec{0}$	$\vec{0}$	pre	.81 ± .04	.48 ± .06
	$\vec{0}$	✓	✓	**.80** ± .03	.55 ± .04
	$\vec{0}$	pre	pre	.79 ± .04	.55 ± .04
	✓	✓	$\vec{0}$	.75 ± .06	.54 ± .10
	✓	pre	$\vec{0}$	**.80** ± .02	.57 ± .07
	✓	$\vec{0}$	✓	.75 ± .11	.57 ± .10
	✓	$\vec{0}$	pre	**.80** ± .05	.56 ± .10
	✓	✓	✓	.74 ± .07	**.59** ± .08
	✓	pre	pre	.77 ± .05	**.59** ± .06
	Baseline (MC)			.20 ± .00	.32 ± .00
	Baseline (Rand)			.52 ± .05	.51 ± .07

Performance on **Robot Pairs**.

	Model (*M*)		Prediction Correct ↑	
	Object Data			
	Lang	Vis	*in*	*on*
Dev Fold	✓	$\vec{0}$	.86 ± .02	.76 ± .01
	$\vec{0}$	✓	.94 ± .01	.79 ± .01
	✓	✓	.86 ± .04	.78 ± .01
	Baseline (MC)		.87 ± .00	.73 ± .00
	Baseline (Rand)		.49 ± .07	.50 ± .03
Test Fold	✓	$\vec{0}$	.86 ± .01	.83 ± .01
	$\vec{0}$	✓	.88 ± .02	.82 ± .01
	✓	✓	.87 ± .02	.83 ± .01
	Baseline (MC)		.84 ± .00	.83 ± .00
	Baseline (Rand)		.51 ± .06	.51 ± .03

Performance on **All Pairs**.

jdtho@cs.washington.edu https://arxiv.org/abs/1904.01650