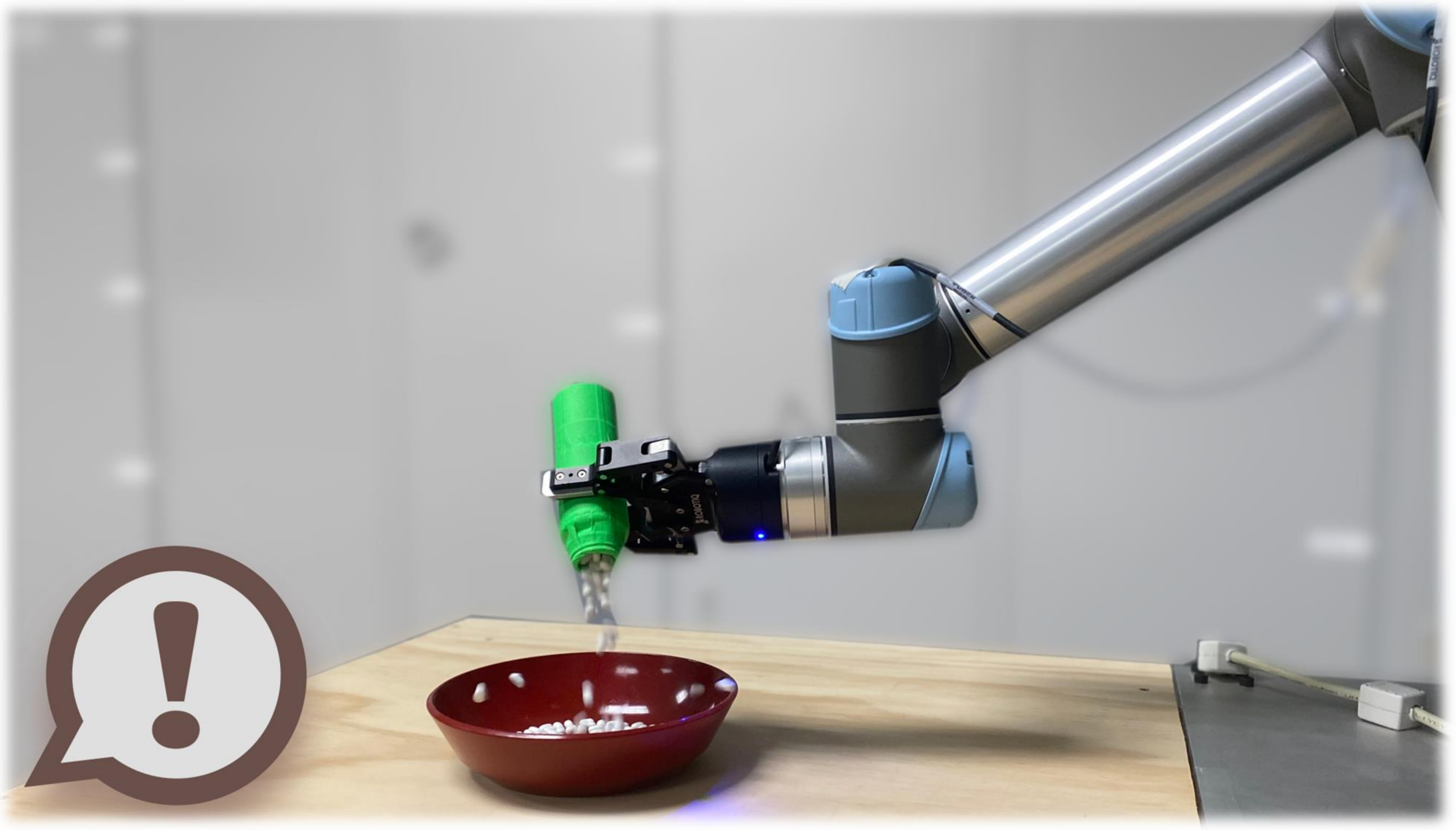
Language-Conditioned Imitation Learning for Robot Manipulation Tasks

Simon Stepputtis, *sstepput@asu.edu*, Arizona State University
Joseph Campbell, *jacampb1@asu.edu*, Arizona State University
Mariano Phielipp, *mariano.j.phielipp@intel.com*, Intel AI Labs
Stefan Lee, *leestef@oregonstate.edu*, Oregon State University
Chitta Baral, *chitta@asu.edu*, Arizona State University
Heni Ben Amor, *hbenamor@asu.edu*, Arizona State University

NEURAL INFORMATION PROCESSING SYSTEMS



Introduction

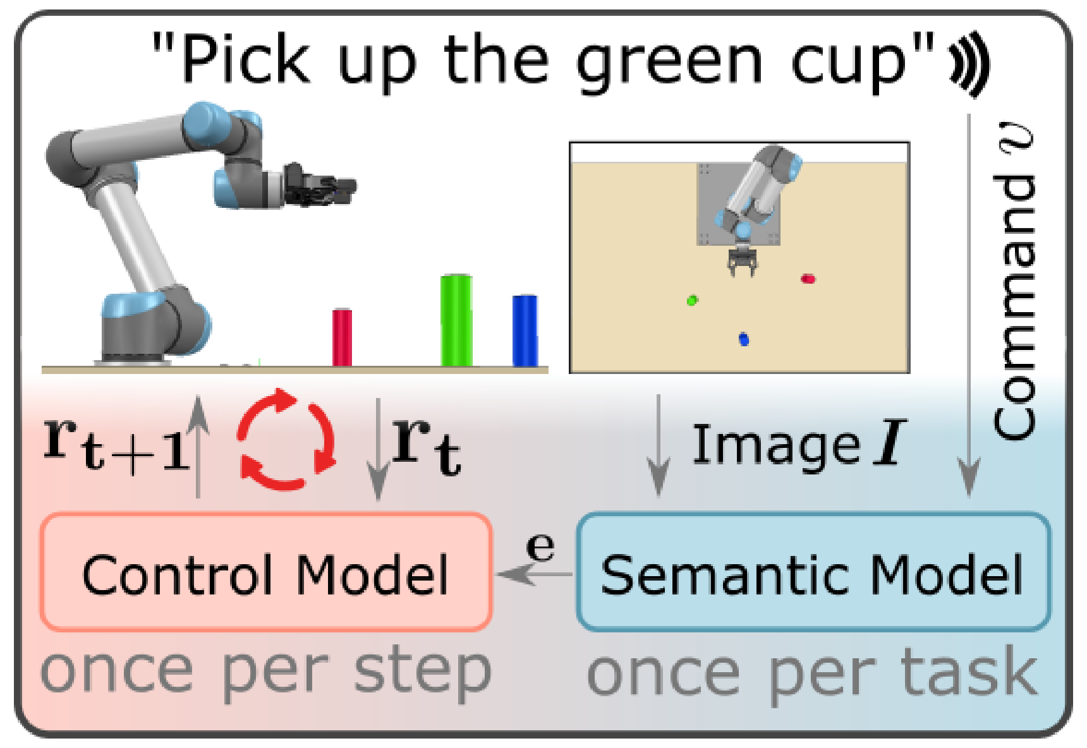
We consider language as a flexible goal specification for imitation learning in manipulation tasks, providing a communication channel between the human expert and the robot, allowing users to convey the intent of a task. While training, our model learns to interrelate language, vision and motion control to capture the correlations between them, generating a language conditioned control policy that is specific to the defined task.

To summarize our contributions, we:

- introduced a language-conditioned manipulation task
- provide robot task specifications in an intuitive fashion through language
- developed an end-to-end, language-conditioned control policy for
- Integrate language, vision, and control within a single framework

Model

We treat policy generation as a translation process from language and vision. While our approach is an end-to-end approach, we can conceptually divide it into two parts, namely semantic and control model:



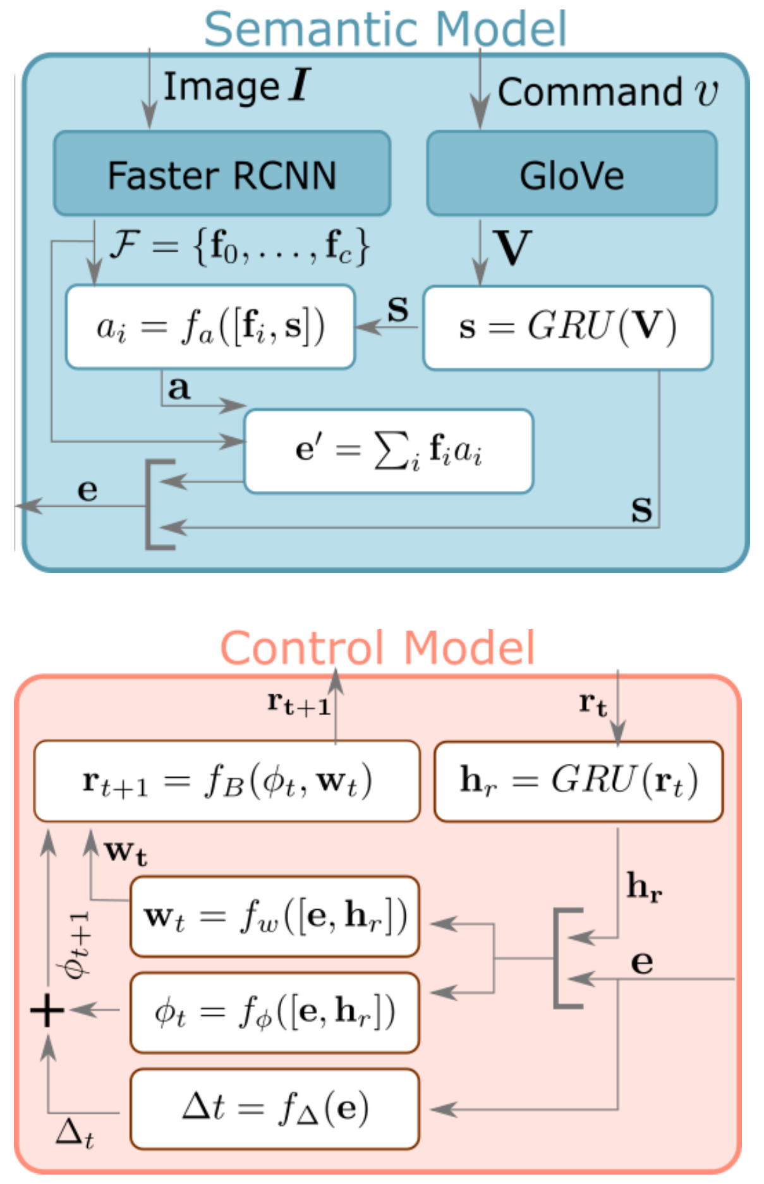
Semantic Model: Creates a unique task representation from language and vision.

Control Model: Translates the task representation into a task-specific control policy while taking the current robot state into account.

Model Integration

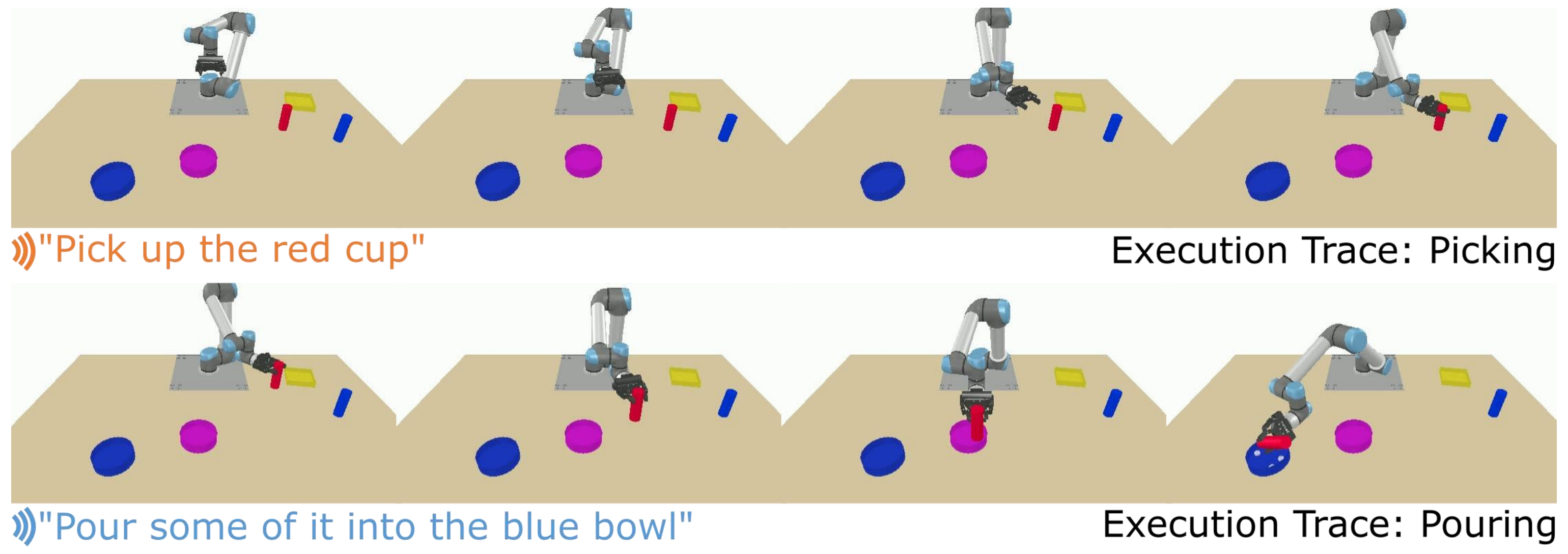
The semantic model combines the image (using FRCNN for object detection) and the language (using GloVe) in an attention model to identify the target object. Finally, the target object is combined with the sentence representation ***s*** into the final task embedding ***e***. This module is running once per task.

The control model translates the task embedding ***e*** together with the current robot state ***r*** into the hyper-parameters for a motion primitive that is specific to the desired task. This module is running once per step during the execution of the task over *t* time-steps.


Semantic Model
Image I
Command v
Faster RCNN
GloVe
$\mathcal{F} = \{\mathbf{f}_0, \dots, \mathbf{f}_c\}$
$\mathbf{V}$
$a_i = f_a([\mathbf{f}_i, \mathbf{s}])$
$\mathbf{s}$
$\mathbf{s} = GRU(\mathbf{V})$
$\mathbf{a}$
$\mathbf{e}' = \sum_i \mathbf{f}_i a_i$
$\mathbf{e}$
$\mathbf{s}$
Control Model
$\mathbf{r_{t+1}}$
$\mathbf{r_t}$
$\mathbf{r}_{t+1} = f_B(\phi_t, \mathbf{w}_t)$
$\mathbf{h}_r = GRU(\mathbf{r}_t)$
$\mathbf{w_t}$
$\mathbf{h_r}$
$\mathbf{w}_t = f_w([\mathbf{e}, \mathbf{h}_r])$
ϕ_{t+1}
$\mathbf{e}$
+
$\phi_t = f_\phi([\mathbf{e}, \mathbf{h}_r])$
$\Delta t = f_\Delta(\mathbf{e})$
Δ_t

We train our model on 40,000 synthetically generated scenarios. Task descriptions are generated by a templating system based on 200 human expert descriptions (5 experts), utilizing a synonym replacement approach.

Evaluation



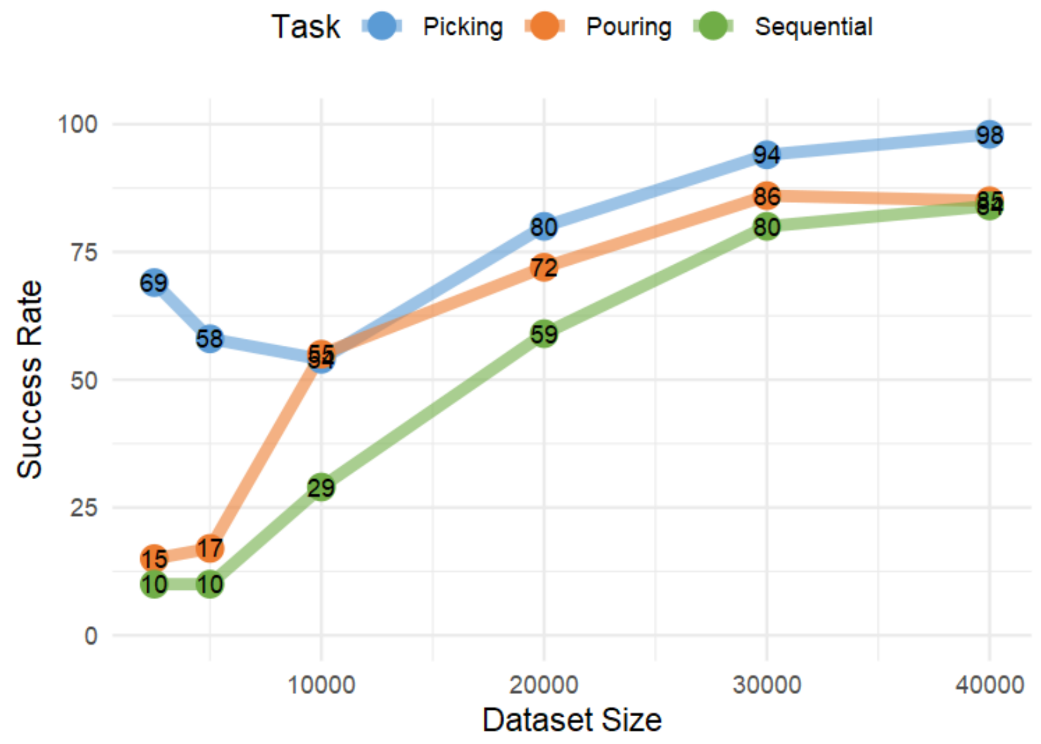
We evaluated our approach in a simulated robot task with a table-top setup, in which the robot is taught by an expert how to perform a combination of picking and pouring behaviors specified by language (see execution traces):

- **Picking**: Grasping (and lifting) one of three possible cups
- **Pouring**: Pouring a specified quantity into one of 20 possible bowls

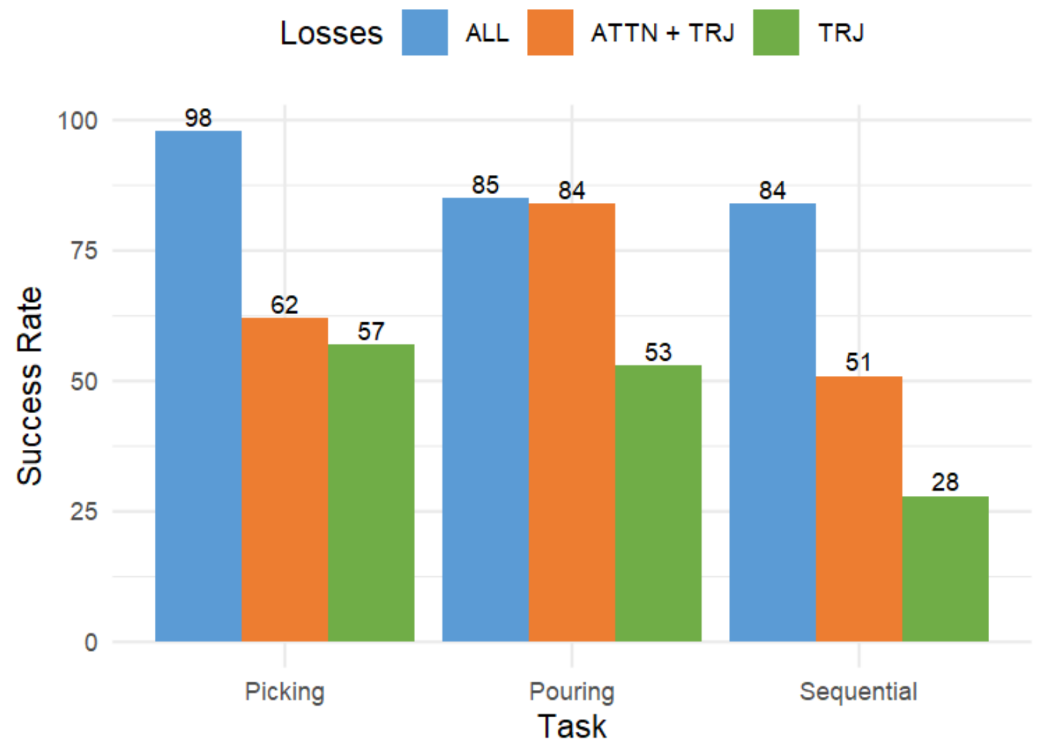
Model	Picking	Pouring	Sequential	Detection	In Target	MAE
Ours	98%	85%	84%	94%	94%	0.05°
PayAttention	23%	8%	0%	66%	41%	0.53°
Simple RNN	58%	0%	0%	52%	6%	0.30°

Model Performance: Our model compared with two baselines (PayAttention! [1] and a simple RNN architecture). The table shows the Picking, Pouring and Sequential task success. Furthermore, we show the Detection rate of the correct object, the percentage of how much of a cup's contend is in the correct target bowl and the MAE of the robot's joint configuration.

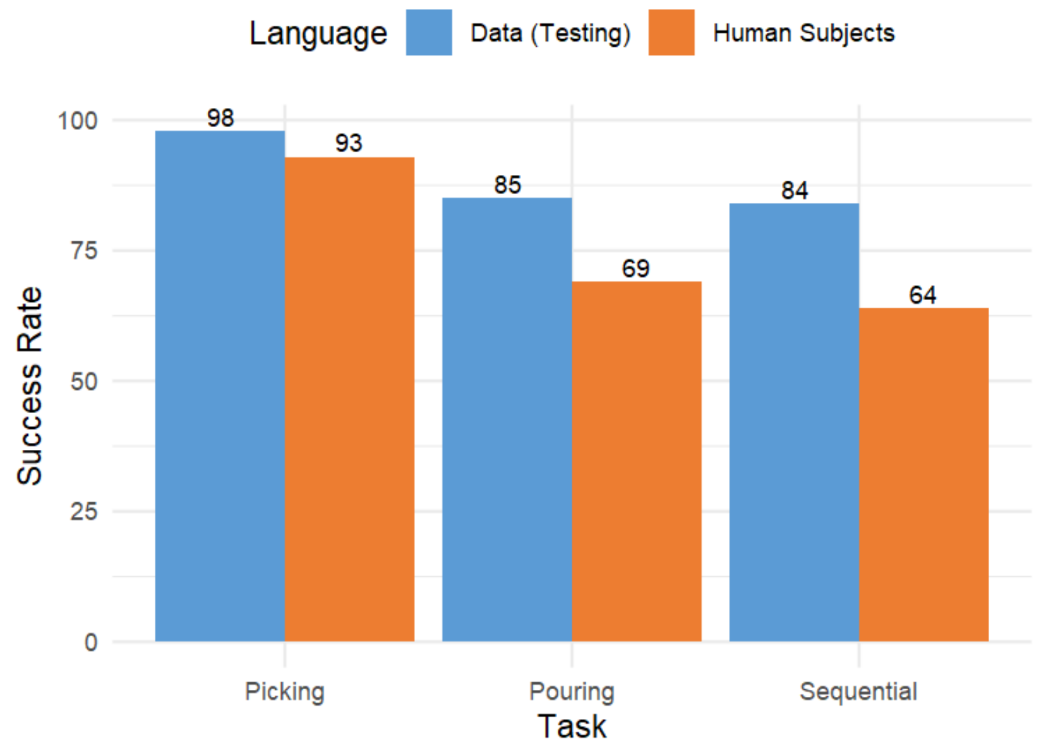
Results



Dataset Size: Influence of the dataset size on the success rate. Experiments show that significant performance increases can be seen by increasing the dataset size from 2,500 to 30,000. In our experiments, we use a dataset size of 40,000.



Losses: We utilize auxiliary losses to complement the generated robot control signal. Guiding the object detection (ATTN) helps in the pouring task, guiding the policy generation yields significant performance increases in the pouring task.



Human Interaction: We also evaluate our model with five new human participants issuing commands and compare it to our synthetic language. Overall, our model responds well to new natural language from new human operators.

Conclusion

We present an approach for end-to-end imitation learning of robot manipulation policies that combines language, vision, and control. Empirically, we showed that our approach significantly outperformed alternative methods, while also generalizing across a variety of experimental setups.

References

1. Pooya Abolghasemi et al.(2019). *"Pay attention!-robustifying a deep visuomotor policy through task-focused visual attention."* In: CVPR: IEEE Conference on Computer Vision and Pattern Recognition; p. 4254 - 4262

Full paper available at: https://arxiv.org/abs/2010.12083
Code and Dataset available at: https://github.com/ir-lab/LanguagePolicies