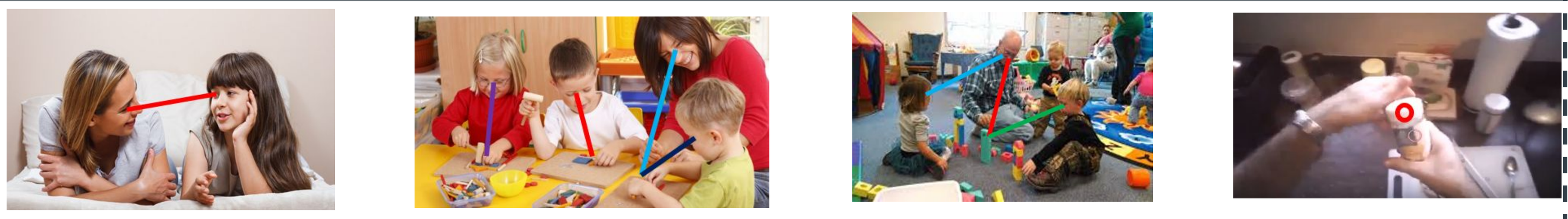
Understanding Teacher Gaze Patterns for Robot Learning

Akanksha Saran, Elaine Schaertl Short, Andrea Thomaz and Scott Niekum
asaran@cs.utexas.edu elaine.short@tufts.edu andrea.thomaz@ece.utexas.edu sniekum@cs.utexas.edu
University of Texas at Austin

Motivation

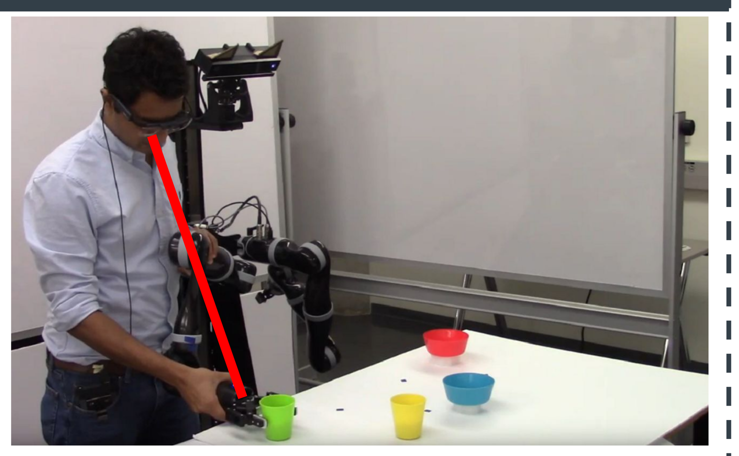


Prior work has shown gaze to convey goals, intent for future actions and mental load in human-human interaction, human manipulation and human-robot teleoperation [4, 5]. Is gaze from human demonstrations informative for learning?

Research Questions

Q. Do human demonstrations provided to robots contain interesting gaze patterns?

Q. Can these patterns be utilized to enhance robot learning?

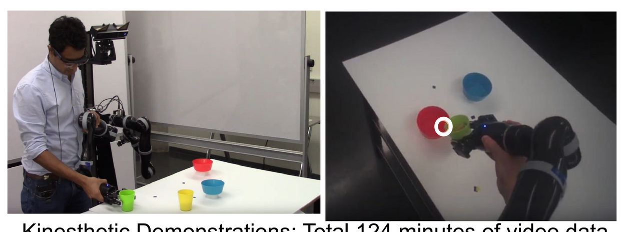


User Study Design

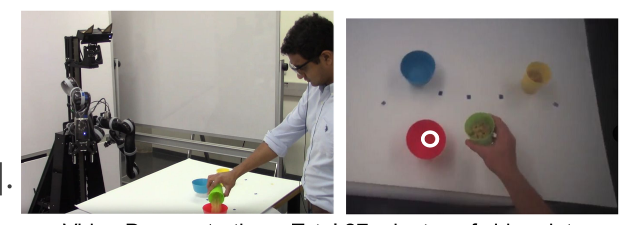
To understand the role of human gaze when humans teach robots, we study eye gaze behavior under two different Learning from Demonstration (LfD) paradigms:

(1) **Keyframe based Kinesthetic Demonstrations (KT)**: Users move the robot's end-effector and provide keyframes along a desired trajectory

(2) **Video/Observational Demonstrations**: Robot passively observes the human performing the task themself.



Kinesthetic Demonstrations: Total 124 minutes of video data



Video Demonstrations: Total 27 minutes of video data

We conducted a **2 x 2 mixed-design** human subjects study [**user type**: novice or expert x **gaze fixation area**: task relevant objects or task-irrelevant objects]. We recruited **20 subjects** (10 expert and 10 novice robot users) and collect first-person video (50Hz) and corresponding gaze coordinates via the Tobii Pro Glasses Eye Tracker 2.

Each user provides KT and Video Demonstrations for two tasks (placement and pouring) of varying complexity, under the same task layouts as shown. The order of demonstration types is counterbalanced across all users. Users provide 6 demos (3 KT, 3 video) for the pouring task and 4 demos (2 KT, 2 video) for the placement task. Each demo of the placement task is either spatially related to the bowl or the plate.

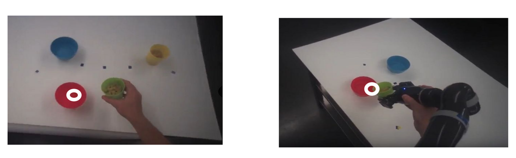
Placement Task (single step): Users are asked to place the green ladle either to the right of the yellow bowl or left of the red plate.

Pouring Task (multi-step): Users are asked to pour pasta from the green cup into the red bowl and from the yellow cup into the blue bowl.

Gaze Patterns in Human Demonstrations for Robots

Gaze Fixation Filtering

In our data, eye gaze movements can be characterized as:
Fixations: focus gaze on a single location (typically 100-500 ms)
Saccades: rapid, voluntary eye movements, abruptly change the point of fixation

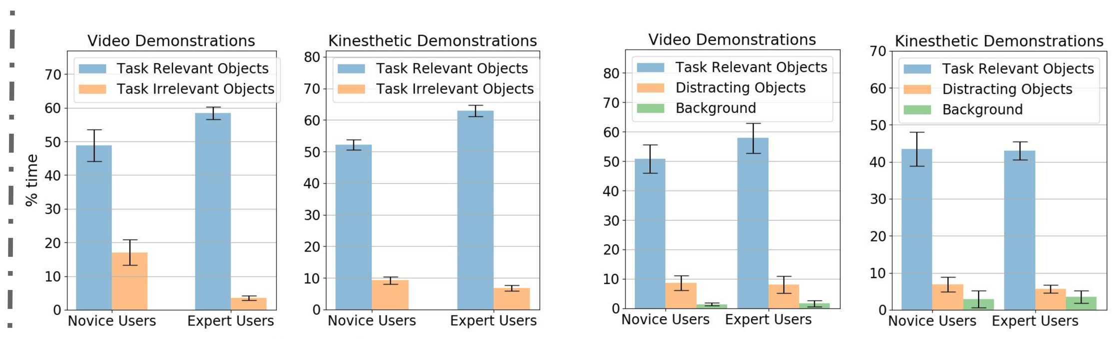


We use spatio-temporal features to filter out fixations:

- **Velocity**: discard gaze coordinates moving at high speeds
- **Area**: Color histogram in a 100-pixel radius circle around gaze location to determine object of fixation
- **Duration**: fixation > 100ms

Users rarely fixate on task-irrelevant objects

Users fixate more on the objects that are relevant to the task, i.e. gaze fixations on task-relevant objects and task-irrelevant objects come from different distributions ($p<0.01$ with mixed design ANOVA) for both demonstration types. This significant result holds for both tasks and for both user types individually as well.

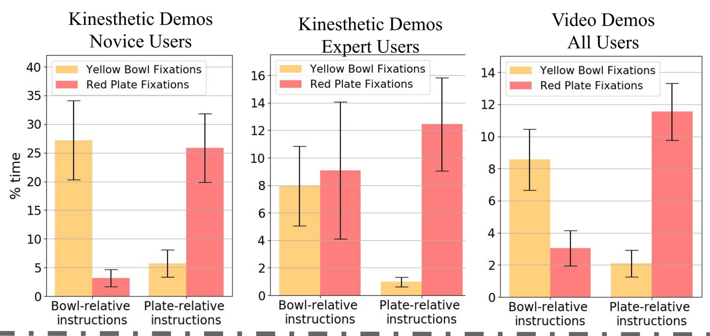


(a) Pouring Task (b) Placement Task

Gaze can identify intent for ambiguous actions

In the placement task, the ladle is placed either to the **left of the red plate** or **right of the yellow bowl**. This is ambiguous as both conditions refer to the same area between the bowl and the plate.

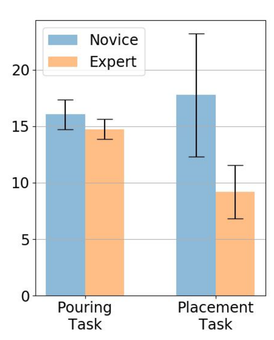
Users attend more to the object with respect to which they perform the placement task ($p<0.01$)



Novice robot users attend more to the robot's gripper during Kinesthetic Demonstrations

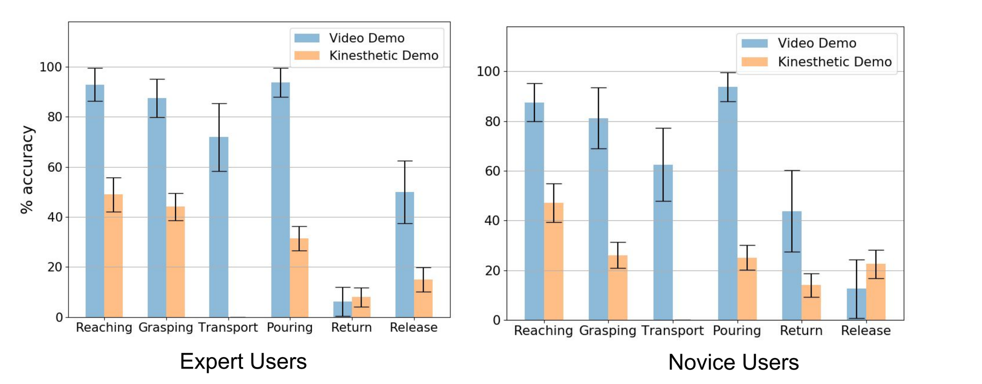
Novice users on average spend more time fixating on the gripper compared to expert users ($p=0.178$ for placement and $p=0.813$ for pouring). This effect is mostly as novice users struggle more with manipulating the robot's end effector.

Both user types spend gaze more at task-relevant objects than the gripper.

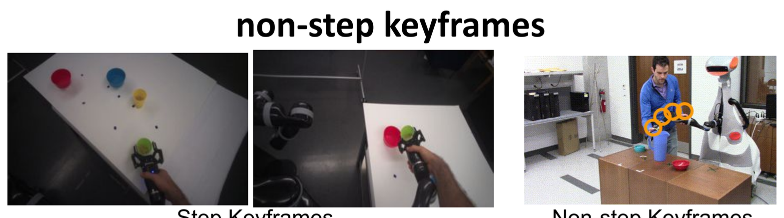


Gaze fixations can predict the target object of subtasks for video demonstrations

Gaze fixations within subtasks (reaching, grasping, pouring etc.) of the multi-step pouring task line up with their target reference frame (eg: reaching for the green cup) for video demonstrations (at least 75% of time for experts and 70% of the time with novice users). Video demos container a cleaner gaze fixation patterns than KT demos.



Gaze patterns differ between step and non-step keyframes



Step Keyframes Non-step Keyframes

Step Keyframes mark the boundaries of semantically different actions. **Non-step keyframes** are consecutive keyframes part of the same semantic action.

Hypothesis: It is more likely that users change their target of attention before and after step keyframes versus non-step keyframes

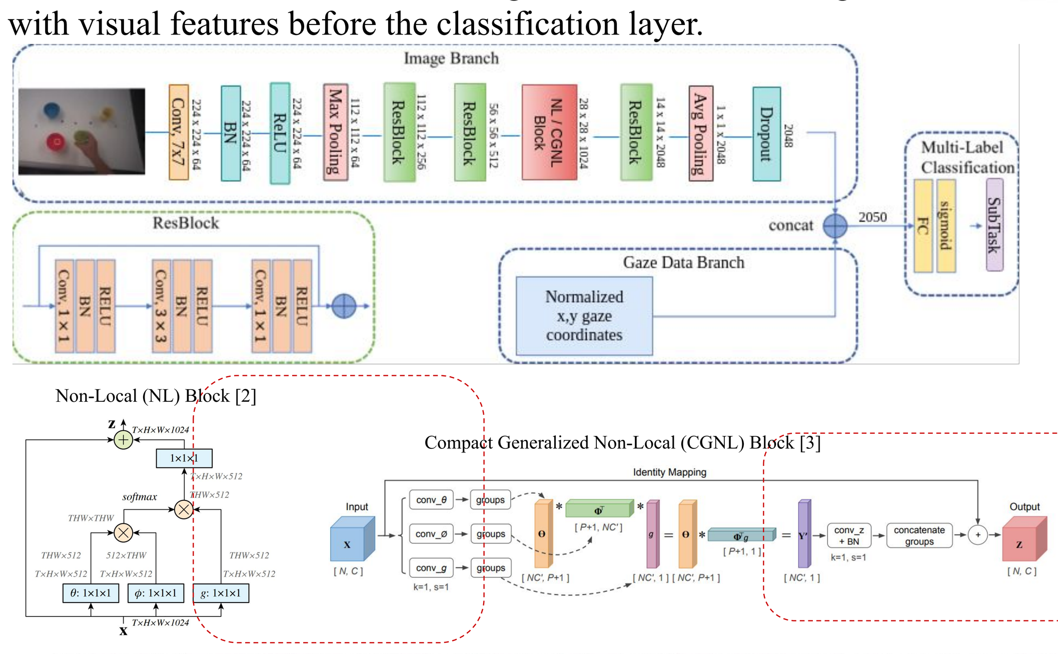
	Non-step KF	Step KF
Novices	19.44	**24.51**
Experts	15.79	**27.85**

Percentage of keyframes for which there was a change in the target object of attention 3 secs before and 3 secs after.

Gaze-enhanced Multi-step Task Segmentation

Automatic multi-step task segmentation (even with moderate noise) has been shown to be an effective intermediate step for multi-step policy learning [6]. We show that gaze helps to improve performance for task segmentation.

We use two different residual neural network architectures (ResNet-50) [2, 3] pretrained on ImageNet for subtask classification. These architectures have shown to perform well for activity recognition in videos. We concatenate normalized gaze coordinates in image frame with visual features before the classification layer.



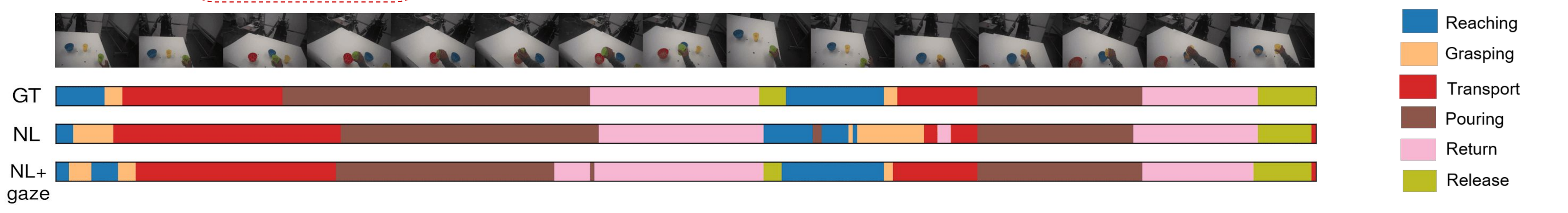
	Video Demos		KT Demos	
	NL	CGNL	NL	CGNL
Without Gaze	86.95	82.42	61.95	63.00
With Gaze	**87.63**	**88.88**	**62.71**	**64.11**

10-fold cross validation accuracy of subtask prediction.

Training Data:

- 3K images from video demos
- 12K images for KT demos

- Video Demos show more improvement with gaze than KT demos
- ResNet-50 with one CGNL block [3] shows more improvement over ResNet-50 with one NL block [2]



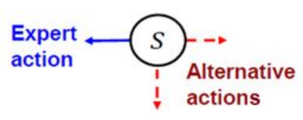
Gaze-augmented Bayesian Inverse Reinforcement Learning

Bayesian Inverse Reinforcement Learning (BIRL) [1]

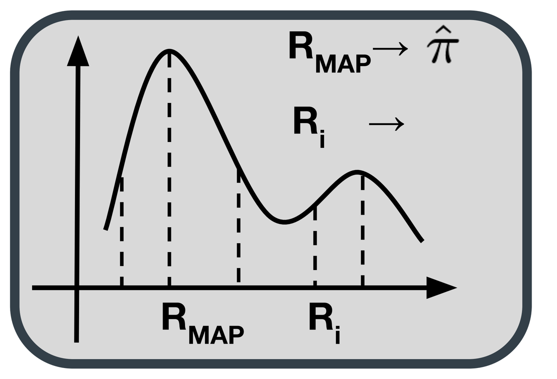
$$\underbrace{P(R|D)}_{\text{Posterior}} \propto \underbrace{P(D|R)}_{\text{Likelihood}}\underbrace{P(R)}_{\text{Prior}}$$

- Reasons over a distribution of reward functions
- Uses MCMC to sample from posterior
- Assumes softmax demonstrator likelihood
- MAP reward estimate: robot's best guess of the demonstrator's intent

$$P(D|R) = \prod_{(s,a)\in D} \frac{e^{Q^*_R(s,a)}}{\sum_{b\in A} e^{Q^*_R(s,b)}}$$



$$P((s,\leftarrow)|R) = \frac{e^{Q^*(s,\leftarrow,R)}}{e^{Q^*(s,\leftarrow,R)} + \sum e^{Q^*(s,\dashrightarrow,R)}}$$



Gaze-augmented Bayesian Inverse Reinforcement Learning (Gaze + BIRL)

$$P(R|D,G) \propto P(D|R)P(R|G)$$

- Gaze can help weed out unlikely reward functions better
- Differences in the amount of time spent looking at an object of interest can arise from the intent or internal reward of the demonstrator.
- Assumption on the Reward function: weighted sum of RBF kernels placed around each object. RBFs help capture the spatial relations around the objects.
- Penalize reward functions which violate the ranking of weights on object pairs in comparison to the amount of time fixated on them.

$$R(x) = \sum_{i=1,j=1}^{i=k,j=5} w_{ij} \cdot rbf(x, c_{ij}, \sigma_i^2)$$

$$\mathrm{rbf}(x,c,\sigma^2) = \exp(-\|x-c\|^2/\sigma^2)$$

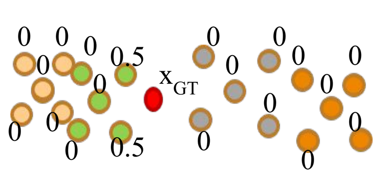
$$P(R|G) = -\sum_{i,j} I_{ij}\frac{l_i}{l_j}$$

l_i Fixation time on object i
w_i Weight on object i

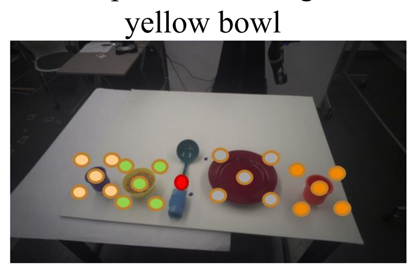
$$I_{ij} = \begin{cases} 1 & \begin{matrix} w_i < w_j \\ l_i > l_j \end{matrix} \\ 0 & \text{otherwise} \end{cases}$$

Gaze helps disambiguate among reward functions which are equally likely using state-action information alone.

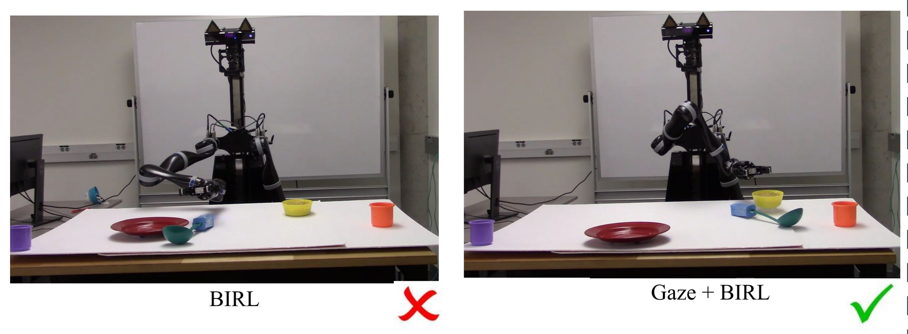
Weights of RBF kernels for the Ground Truth (GT) Reward Function



Ambiguous Demonstration
Ladle placed to the right of yellow bowl



Testing on the robot with unseen object configurations



BIRL Gaze + BIRL

Value of policy

$$V^\pi_R = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t R(s_t)\right]$$

Policy Loss

$$V^{\pi^*}_R - V^\pi_{R_{MAP}}$$

	5 Kinesthetic Demos		5 Video Demos		1 Kinesthetic Demo		1 Video Demo	
Instruction relative to	Bowl	Plate	Bowl	Plate	Bowl	Plate	Bowl	Plate
Without Gaze	0.619	0.081	0.678	0.036	0.073	0.666	0.486	0.184
With Gaze	**0.329**	**0.032**	**0.046**	**0.021**	**0.043**	**0.225**	**0.098**	**0.120**
Average improvement with Gaze	53.75%		**67.4%**		53.65%		**57.3%**	

TABLE I
AVERAGE POLICY LOSS WITH AND WITHOUT THE USE OF GAZE INFORMATION IN THE BIRL FRAMEWORK FOR THE PLACEMENT TASK.

$$\nabla_x R(x) = \sum_{i=1}^{k} w_i \cdot \mathrm{rbf}(x, c_i, \sigma_i^2)\left(\frac{-2x+2c_i}{\sigma_i^2}\right)$$

$$|x_{R_{GT}} - x_{R_{MAP}}|$$

	5 Kinesthetic Demos		5 Video Demos		1 Kinesthetic Demo		1 Video Demo	
Instruction relative to	Bowl	Plate	Bowl	Plate	Bowl	Plate	Bowl	Plate
Without Gaze	0.494	0.102	0.536	0.064	0.087	0.492	0.383	0.160
With Gaze	**0.291**	**0.063**	**0.068**	**0.045**	**0.066**	**0.191**	**0.102**	**0.122**
Average improvement with Gaze	39.66%		**58.5%**		42.66%		**48.56%**	

TABLE II
AVERAGE PLACEMENT LOSS WITH AND WITHOUT THE USE OF GAZE INFORMATION IN THE BIRL FRAMEWORK FOR THE PLACEMENT TASK.

Conclusions

- Distinct Gaze Patterns emerge between Expert / Novice Robot Users, Kinesthetic / Video Demonstrations, Step / Non-step keyframes
- Gaze helps identify ambiguous demonstrations
- Gaze improves performance on two learning tasks in support of statistical findings
- Gaze can be a useful source of information to exploit for other LfD approaches

Future Work

- Utilizing gaze for more complex tasks with increased ambiguities
- Incorporating noisy vision-based gaze detection systems for learning

The University of Texas at Austin
Department of Computer Science
College of Natural Sciences
THE UNIVERSITY OF TEXAS AT AUSTIN
UT ECE
ELECTRICAL & COMPUTER ENGINEERING

[1] Ramachandran, D., and Amir, E. 2007. Bayesian inverse reinforcement learning. In *Proceedings of International Joint Conference of Artificial Intelligence (IJCAI)*, 2007.
[2] Wang, X., Girshick, R., Gupta, A., & He, K. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018
[3] Yue, K., Sun, M., Yuan, Y., Zhou, F., Ding, E., & Xu, F. Compact generalized non-local network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
[4] Argyle, M. Non-verbal communication in human social interaction, 1972.
[5] Hayhoe, M., & Ballard, D. Eye movements in natural behavior. *Trends in cognitive sciences*, *9*(4), 2005.
[6] Goo, W., & Niekum, S. One-shot learning of multi-step tasks from observation via activity localization in auxiliary video. In IEEE *International Conference on Robotics and Automation (ICRA)*, 2019.

Video

Personal Autonomous Robotics Lab
PeARL
Socially Intelligent Machines Lab