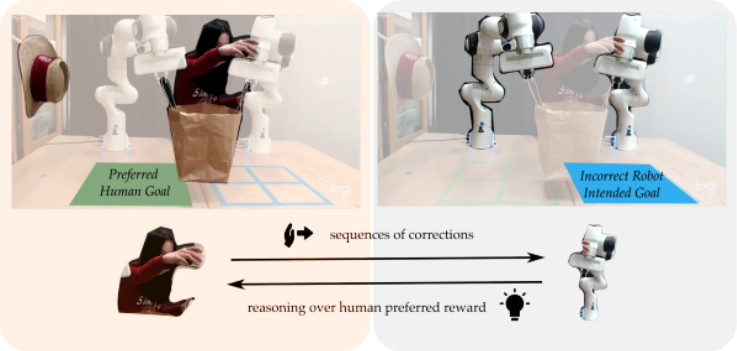
Alper Canberk - Research Statement

My long-term goal is to develop robots that can carry out **long-horizon manipulation tasks specified by human instructions** flexibly and reliably. To this end, I'm interested in combining language and video understanding to develop embodiment-agnostic, self-supervised methods for learning diverse manipulation skills.

Past Research and Future Research Interests



I was fortunate to work on my first research publication, *Learning Human Objectives from Sequences of Physical Corrections* [4] under Mengxi Li and Prof. Dorsa Sadigh at Stanford ILIAD. In this publication, we developed a model for understanding the trade-off between immediate (i.e. per-timestep) and episodic human user intentions through corrections to robot trajectories. By conditioning on sequences of corrections as opposed to assuming independent corrections, our method inferred intentions more accurately than prior works. Working on this project taught me a lot about research methodology and grew my interest in robotics. While human-robot interaction aligned with my goal of robots that understand task instructions, I realized that the real-world impact of understanding task instructions is currently bottlenecked by task execution performance.

That's why in my next project, I wanted to learn about the challenges of robot manipulation. Under Professor Shuran Song at CAIR Lab and Toyota Research Institute, I first authored *Cloth Funnels* [12], a bimanual multi-primitive learning framework for garment manipulation. By learning to 'funnel' garment states by unfolding and orienting them, our method could successfully execute downstream folding and ironing heuristics. Training such a model required tremendous effort because I had to **carefully engineer and tune action primitives** and **handcraft supervision signals,** which convinced me that reward and primitive engineering is not a scalable way to span all tasks. This led me to the question: *what if we could automate this process?*

Recently, large language models (LLMs) have found applications within robotic task planning. Through their exposure to large internet text corpora, LLMs exhibit common-sensical planning abilities that directly generalize to a wide variety of settings, ranging from tabletop object rearrangement [5] to high-level household tasks [3]. So far, these approaches have ignored the complexities of low-level manipulation by interfacing LLMs with manually

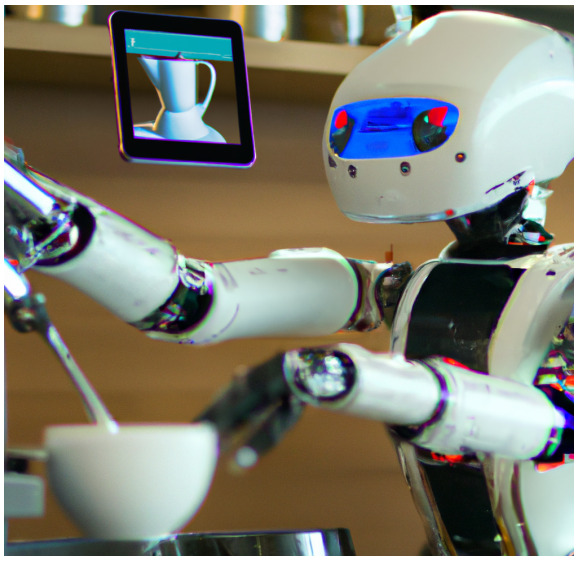
engineered or behavior-cloned *action primitives* (language-specifiable, coarse-grained manipulation skills such as “pick Y up”, “put X on Y”, etc.)[6]. Unfortunately, using behavior cloning means that for every new robot hardware, human operators have to tediously collect data for new models, which bottlenecks the learning of new action primitives.

To more efficiently generate robust action primitives, I want to investigate how language paired with vision data could give us *self-supervised* and *embodiment-agnostic* methods for learning action primitives. The key challenge is that LLMs, even when paired with VLMs, are insufficient sources of supervision for low-level control; large-scale interaction data is also required. Although no such robot interaction data yet exists online, we could use videos of *human interaction* for *any* embodiment to supervise *itself*. With this goal in mind, I propose to study the following:

1. **For learning and executing action primitives,** LLMs could propose useful primitives for a given task, and query relevant ego-centric human videos. From these videos, we can extract embodiment-agnostic reward functions for learning embodiment-specific manipulation abilities [6, 7], which we cannot possibly get from large pre-trained models. Finally, to generalize these abilities to an open vocabulary of objects without explicit training, VLMs could be used to highlight relevant spatial locations for an action primitive to act on (e.g. from the prompt “put my LEGO™ Technic Liebherr Excavator in the box”)
2. **For automatically deriving task-specific supervision**, there may not be enough available videos for the exact long-horizon task prompted. In this case, LLMs could decompose the task into key subtasks, have a model learn reward functions for each subtask separately, and chain them together.

In the long term, I would like to explore

1. Learning from non-ego-centric videos
2. Self-improvement in reset-free real-world environments
3. Seeking out online information for solving novel problems.



My vision for the future is a robot that can learn by watching YouTube and practicing on its own time. Given the prompt “make me coffee,” this robot will not just open cabinets and pick up mugs, but presented a new coffee machine, it will find and use instructional videos to operate the machine. By studying language and video understanding in the context of robotics, I hope to one day have a home robot that can zero-shot coffee-making out of the box.

References

[1] Ahn, Michael, et al. "Do as I can, not as I say: Grounding language in robotic affordances." arXiv preprint arXiv:2204.01691 (2022).

[2] Canberk, Alper, et al. "Cloth Funnels: Canonicalized-Alignment for Multi-Purpose Garment Manipulation." arXiv preprint arXiv:2210.09347 (2022).

[3] Huang, Wenlong, et al. "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents." arXiv preprint arXiv:2201.07207 (2022).

[4] Li, Mengxi, et al. “Learning Human Objectives from Sequences of Physical Corrections.” 2021 IEEE International Conference on Robotics and Automation (ICRA), May 2021. Crossref, https://doi.org/10.1109/icra48506.2021.9560829.

[5] Liang, Jacky, et al. "Code as policies: Language model programs for embodied control." arXiv preprint arXiv:2209.07753 (2022).

[6] Shao, Lin et al. "Concept2Robot: Learning Manipulation Concepts from Instructions and Human Demonstrations." Proceedings of Robotics: Science and Systems (RSS). 2020.

[7] Zakka, Kevin, et al. "Xirl: Cross-embodiment inverse reinforcement learning." Conference on Robot Learning. PMLR, 2022.