Large Language Models as Proxies for Evaluating Collaborative Norms

Michelle Zhao[1], Hao Zhu[2], Reid Simmons[1], Yonatan Bisk[2], Henny Admoni[1]

ABSTRACT

As robots become more ubiquitous in our everyday lives, it is increasingly important to design robot agents that can interact fluently with everyday users. Prior work on collaborative fluency demonstrates that metrics outside of task success are important for human collaborators and influence their perceptions of the robot and willingness to continue to collaborate. As such, it is important for robots to evaluate collaborative interactions by factors, such as fluency, equitability, and safety. In order to enable robots to understand these concepts and synthesize them into a reward function, we turn to vision-language models (VLMs). In this work, we propose the use of VLMs to evaluate collaborative interactions and score them on along factors in collaborative norms. As a first step, we aim to determine whether the VLMs preference-based rankings of different interactions align with actual human ratings. This investigation will inform whether VLMs can be used a proxies for human evaluators of collaborative robot behavior, and will inform future work on allowing robots to reflect upon their interactions to update their policies.

CCS CONCEPTS

• **Human-centered computing → Interaction design process and methods**; **Activity centered design**.

KEYWORDS

human robot collaboration, reward design, vision language models

ACM Reference Format:
Michelle Zhao[1], Hao Zhu[2], Reid Simmons[1], Yonatan Bisk[2], Henny Admoni[1]. 2024. Large Language Models as Proxies for Evaluating Collaborative Norms. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3610978.3638155

1 BACKGROUND

As robots enter our homes, they will increasingly interact with one person, repeatedly. A home robot helping someone clean and fold laundry will aid in this endeavour several times a week over the course of months or years. Over each interaction, the robot should continuously learn to collaborate more fluently and efficiently with the human partner. A robot which only maximizes task success may not focus on the human partner's preferences, and may cause the team to reach a task outcome the human partner prefers. In value alignment [1], the task is exactly achieving the human's preferences. Cooperative inverse reinforcement learning, a paradigm for value alignment, defines the task as achieving the human's objective, captured by a reward function the robot must learn through observation and interaction with the human [3]. Even when the reward function is fully specified by human preferences, a collaborative robot who completely takes over the task or one who performs redundant tasks, is not a preferable teammate.

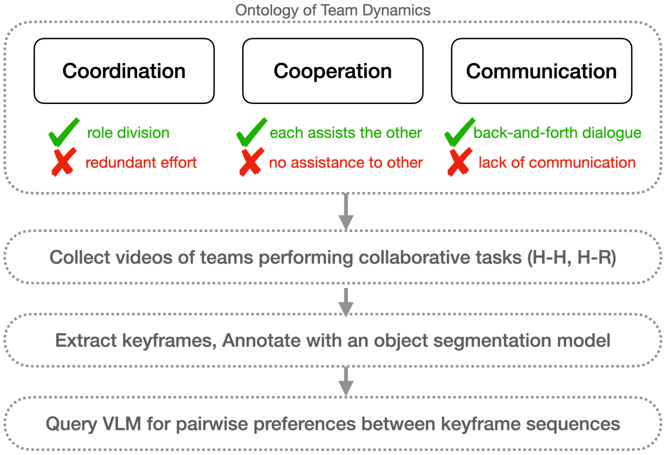
Ontology of Team Dynamics
Coordination
Cooperation
Communication
✓ role division
✗ redundant effort
✓ each assists the other
✗ no assistance to other
✓ back-and-forth dialogue
✗ lack of communication
Collect videos of teams performing collaborative tasks (H-H, H-R)
Extract keyframes, Annotate with an object segmentation model
Query VLM for pairwise preferences between keyframe sequences

Figure 1: We begin by constructing an ontology of team dynamics.

Collaborative fluency is a set of metrics that aims to capture how well team members perform a coordinated meshing of their joint actions [5]. Measures of fluency include the amount of idle time for each partner, the amount of concurrent activity, and the functional delay: the accumulated time between the completion of one agent's action and the beginning of the other agent's action in a sequential task. As we seek to develop a more collaborative robot, collaborative fluency metrics measure how effective an agent is at coordinating its behavior with its human partner. While fluency measures do not always directly correlate with task efficiency, they can change people's perception of collaboration.

Beyond fluency, effective robot collaborators should also maintain safe interactions [9], where the robot does not interfere in the human's workspace unless necessary for a coordinated task. The robot should also be performing predictable and legible actions [2], enabling the human to better plan ahead for their own contribution to the task. Adhering to such collaborative norms may synergistically lead the team towards more effective role division, sequential planning, and highly coordination fluency [10]. In situations in which both robot and human have similar action capabilities, an equitable collaboration occurs when each team member's contribution is similar in magnitude [8]. Though this may not be desirable

[1] Authors are with the Robotics Institute, Carnegie Mellon University and [2] the Language Technology Institute, Carnegie Mellon University. {mzhao2, rsimmons, hadmoni, hzhu2, ybisk}@andrew.cmu.edu.

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03
https://doi.org/10.1145/3610978.3638155

in situations in which the human team member has a disability or the robot has restricted capabilities, equitable contribution may also be an aspect of collaboration that human partners may prefer. Lastly, effective partners are cooperative, in that they will when necessary help team members on subtasks which are not their own objective.

These human intuitions on what norms and behaviors comprise effective collaborations (ie. what does it mean to be a good partner?) are inherently difficult for robots to capture, and are difficult for experimenters to define through a reward design [4]. As such, we propose to leverage the conceptual knowledge of foundation models [11] to enable robots to develop an understanding of collaborative norms. We propose to query a VLM for pairwise comparisons of annotated keyframes from videos of multiagent collaborations. In recent work, Li et al. [7] uses an LLM to generate and assign team member sub-goals for AI-AI coordination. Zhang et al. [12] proposes a collaborative agent for human-AI coordination leveraging the knowledge of the LLM to anticipate the human's decision-making.

2 APPROACH

2.1 Ontology of Behaviors in Teamwork

Our first step is to scope the team dynamics we wish to evaluate. Using the framework for classifying teamwork breakdowns in Wilson et al. [10], we investigate 3 areas in which teamwork can be effective or break down: Coordination, Cooperation, and Communication. We identify positive and negative examples of each component, and refine the scope of our task domain to be tabletop manipulation (see Figure 1). Communication involves the exchange of information between two people: one who sends the message and one who receives it. Therefore, communication failures occur when there's a delay or absence of the correct information being conveyed to the appropriate person at the necessary time. Coordination depends on accurate and timely efforts and inputs from every team member [6]. By using suitable coordination methods, team members can effectively organize, align, combine, and accomplish tasks while conserving important resources. Ultimately, when team members share similar attitudes and beliefs, they develop a consistent understanding of the task and environment. This results in improved collective knowledge, more efficient decision-making, and superior team performance. Failures in cooperation occur when team members lack the willingness and motivation to work together. Consequently, they do not engage with each other or anticipate one another's requirements, which are essential for fostering and sustaining a collective understanding.

2.1.1 *MDP Formulation.*

We formulate a collaborative task as a two-player Markov decision process (MDP) defined by tuple $\langle \mathcal{S}, \mathbb{A} = \{\mathbb{A}^1, \mathbb{A}^2\}, \mathcal{T}, R \rangle$. $\mathcal{S}$ is the set of states. The action space of a game with two agents is $\mathbb{A} = \mathbb{A}^1 \times \mathbb{A}^2$. The set of actions available to each team member i is $\mathbb{A}^i$. The transition function $\mathcal{T}$ determines how the state changes based on a joint action by both agents, $\mathcal{T} : \mathcal{S} \times (\mathbb{A}^1, \mathbb{A}^2) \rightarrow \mathcal{S}$. $R : \mathcal{S} \rightarrow \mathbb{R}$ is the team reward function.

2.1.2 *Data Collection.*

Given our team dynamics ontology, we record videos of two agents performing a collaborative task. The videos will be a combination of human-robot, and human-human teams. We construct a dataset $D_{raw} = \{\tau_i\}_{i=0}^{n}$, where each trajectory $\tau_i = \{(s_0, a_0^1, a_0^2), ..., (s_T, a_T^1, a_T^2)\}$ is a sequence of states and joint actions. In order to ensure diversity in positive and negative examples of teamwork, the data collection process may involve confederate team members. Given the videos, we will extract H keyframes from each, and annotate the H keyframes with object bounding boxes, extracted using an object sementation model, such as RAM++ [13]. $D_{annot} = \{\xi_i\}_{i=0}^{n}$, where each trajectory $\xi_i = \{s_{(0)}, ..., s_H\}$. The queries we provide to the VLM will be $Q = \{(\xi_i, \xi_j)\}_{i,j}^{n}$.

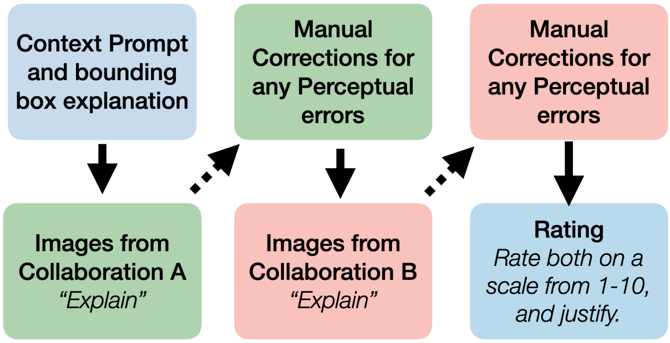
Context Prompt and bounding box explanation
Manual Corrections for any Perceptual errors
Manual Corrections for any Perceptual errors
Images from Collaboration A
"Explain"
Images from Collaboration B
"Explain"
Rating
Rate both on a scale from 1-10, and justify.

Figure 2: We prompt the VLM to generate a comparison between still frame image sequences from two different videos of multiparty collaborations.

2.1.3 *VLM Queries.*

Our final task is to query the VLM for a pairwise ranking of the trajectories ξ_i, ξ_j. We query the VLM using for a ranking between the two interactions and a scalar rating of the collaboration along the axes of collaborative fluency, safety, equitability, and efficiency. Our pipeline is delineated in Figure 2.

3 PROPOSED EVALUATION AND FUTURE WORK

We will evaluate our research question by comparing the VLM-generated pairwise rankings within our dataset of keyframes from collaborative task videos to human responses. We will run a user study to collect human responses on the same set of pairwise queries shown to the VLM. As a quantitative measure, we will evaluate the degree of overlap in human versus VLM responses. As a qualitative measure, we will additionally ask for users to explain their reasoning for each preferences, and compare whether the VLM-generated responses produced the same reasoning.

4 CONCLUSION

As robots increasingly integrate into our daily lives, it becomes essential to develop robotic agents capable of engaging smoothly with regular users. For robots interacting repeatedly with such users, it's crucial they learn from past interactions to enhance future communication and cooperation. Existing research on collaborative fluency has shown that success in tasks isn't the only metric that matters; how humans perceive robots and their willingness to continue working with them are also influenced by other aspects of the collaboration. Therefore, evaluating robotic interactions based on factors like fluency, fairness, and safety is vital. To equip robots with the ability to comprehend and apply these criteria, we aim to evaluate the ability of VLMs to assess collaborative efforts and rate them based on established collaborative standards. This study will help determine if VLMs can help robots assess interactions, and score interactions similarly to human evaluators.

REFERENCES

[1] Daniel S Brown, Jordan Schneider, Anca Dragan, and Scott Niekum. 2021. Value alignment verification. In *International Conference on Machine Learning*. PMLR, 1105–1115.

[2] Anca Dragan and Siddhartha Srinivasa. 2013. Generating legible motion. (2013).

[3] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 29 (2016).

[4] Jerry Zhi-Yang He and Anca D Dragan. 2021. Assisted robust reward design. *arXiv preprint arXiv:2111.09884* (2021).

[5] Guy Hoffman. 2019. Evaluating Fluency in Human–Robot Collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218. https://doi.org/10.1109/THMS.2019.2904558

[6] Steve WJ Kozlowski and Bradford S Bell. 2003. Work groups and teams in organizations. *Handbook of psychology: Industrial and organizational psychology* 12 (2003), 333–375.

[7] Wenhao Li, Dan Qiao, Baoxiang Wang, Xiangfeng Wang, Bo Jin, and Hongyuan Zha. 2023. Semantically Aligned Task Decomposition in Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2305.10865* (2023).

[8] Andrew D Moffat, Robin Fowler, Rebecca L Matz, and Madison Jeffrey. 2022. Is an Effective Team an Equitable Team? Protocol for a Scoping Review. In *2022 IEEE Frontiers in Education Conference (FIE)*. IEEE, 1–6.

[9] Marcello Valori, Adriano Scibilia, Irene Fassi, José Saenz, Roland Behrens, Sebastian Herbster, Catherine Bidard, Eric Lucet, Alice Magisson, Leendert Schaake, et al. 2021. Validating safety in human–robot collaboration: Standards and new perspectives. *Robotics* 10, 2 (2021), 65.

[10] Katherine A Wilson, Eduardo Salas, Heather A Priest, and Dee Andrews. 2007. Errors in the heat of battle: Taking a closer look at shared cognition breakdowns through teamwork. *Human factors* 49, 2 (2007), 243–256.

[11] Ilker Yildirim and LA Paul. 2023. From task structures to world models: What do LLMs know? *arXiv preprint arXiv:2310.04276* (2023).

[12] Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, et al. 2023. Proagent: Building proactive cooperative ai with large language models. *arXiv preprint arXiv:2308.11339* (2023).

[13] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. 2023. Recognize Anything: A Strong Image Tagging Model. *arXiv preprint arXiv:2306.03514* (2023).