Leveraging Vision and Language Models for Zero-Shot, Personalization of Household Multi-Object Rearrangement Tasks

Benjamin A. Newman
Carnegie Mellon University
Meta
Pittsburgh, PA, USA
newmanba@cmu.edu

Pranay Gupta
Carnegie Mellon University
Pittsburgh, PA, USA
pranaygu@andrew.cmu.edu

Yonatan Bisk
Carnegie Mellon University
Meta
Pittsburgh, PA, USA
ybisk@andrew.cmu.edu

Kris Kitani
Carnegie Mellon University
Meta
Pittsburgh, PA, USA
kmkitani@andrew.cmu.edu

Henny Admoni*
Carnegie Mellon University
Pittsburgh, PA, USA
henny@cmu.edu

Chris Paxton*
Meta
Pittsburgh, PA, USA
cpaxton@meta.com

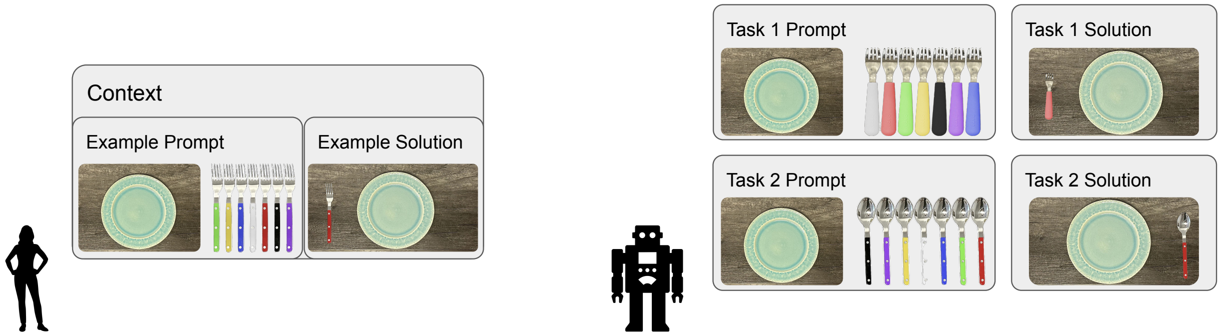

Figure 1: A single-step table setting task. Left: a person gives a single context example, which consists of an initial state, a set of objects, and a final state. From this final state, you can see the person likes red utensils placed according to traditional Western table setting rules. Right: an assistive robot uses this example to generalize this preference to different sets of objects, such as babyforks and spoons, and place them appropriately on the table.

ABSTRACT

Robots should adhere to personal preferences when performing household tasks. Many household tasks can be posed as multi-object rearrangement tasks, but solutions to these problems often target a single, hand defined solution or are trained to match a solution drawn from a distribution of human demonstrated data. In this work, we consider using an internet-scale pre-trained vision-and-language foundation model as the backbone of a robot policy for producing personalized task plans to solve household multi-object rearrangement tasks. We present initial results on a one-step table setting task that shows a proof-of-concept for this method.

Human-Large Language Model Interaction Workshop HRI'24, March 11, 2024, Boulder, CO

CCS CONCEPTS

• **Computer systems organization** → *Robotics*; • **Human-centered computing** → Interaction design theory, concepts and paradigms.

KEYWORDS

Assistive Human-Robot Collaboration, Zero-Shot Collaboration, Multi-Object Rearrangement

ACM Reference Format:
Benjamin A. Newman, Pranay Gupta, Yonatan Bisk, Kris Kitani, Henny Admoni*, and Chris Paxton*. 2024. Leveraging Vision and Language Models for Zero-Shot, Personalization of Household Multi-Object Rearrangement Tasks. In *Proceedings of Human-Large Language Model Interaction Workshop HRI'24.* ACM, New York, NY, USA, 4 pages.

*

1 INTRODUCTION

Assistive robots operating in people's homes should complete tasks in ways that align with their personal preferences [22]. These preferences are highly subjective, and can be abstract or eccentric. We

[*] denotes equal advising

	fork	babyfork	spoon
fork	0.986 ± 0.03	0.929 ± 0.06	0.914 ± 0.07
babyfork	0.814 ± 0.09	0.886 ± 0.08	0.986 ± 0.03
spoon	0.943 ± 0.06	0.986 ± 0.03	0.986 ± 0.03

Table 1: Color Selection Accuracy. Context objects are shown in the rows, while Prompting objects are shown in the columns. Error is reported as 95% confidence bound.

	fork	babyfork	spoon
fork	0.986 ± 0.03	0.986 ± 0.03	0.929 ± 0.06
babyfork	0.986 ± 0.03	0.987 ± 0.03	0.7 ± 0.11
spoon	0.943 ± 0.06	0 ± 0.00	0.986 ± 0.03

Table 2: Location Selection Accuracy. Context objects are shown in the rows, while Prompting objects are shown in the columns. Error is reported as 95% confidence bound.

aim to develop a generalizable planning approach for preference-aligned multi-object rearrangement.

Prior research on personalized household object rearrangement collects task-specific datasets of simulated or human demonstrations and tries to matching preferences present within this dataset [1, 15, 16, 23]. However, curating large datasets of human demonstrations with diverse preferences is challenging. The space of possible preferences is effectively unbounded. Preferences are highly subjective and depends on the physical and mental qualities of the individual. Thus collecting a dataset representative of all user preferences is challenging.

Furthermore, these preferences can be complex and abstract. For example, someone's preferred table-top setting might be grounded in accessibility, visual aesthetics or cultural and traditional rules. Hence learning or modelling these preferences in a generalizable fashion is non-trivial. Finally, preferences are often underspecified. A command such as "Help me set the table for dinner" is commonly issued, but does not indicate that a person prefers to use ceramic dishes for everyone except their child whose place should be set with silicone. Exhaustively and explicitly communicating such preferences in operationalizable ways can be tedious and require precise language that is difficult for people to produce.

We wish to develop a method for generalizable personalized household rearrangement that 1) has low sample complexity 2) is able to model abstract and complex preferences about object rearrangement, and 3) develop these task plans even with under specified instructions.

Recent advances into vision-and-language foundation models (VLMs) provide solutions to all three of these issues. Large-language models (LLMs) and VLMs pretrained on internet scale data have been shown to effectively solve myriad tasks for which they weren't explicitly trained. Specifically, combining LLMs with in-context learning [3] has made tremendous strides in developing task plans that solve general multi-object rearrangement tasks and in solving these tasks according to easily specified human preferences [32] in a few shots.

We present an initial method that takes advantage of these recent advancements in internet-scale pretrained VLMs in order to solve multi-object rearrangement tasks according to personal preferences, even when those preferences are not fully specified. We present the initial results of this method on a single-step table setting task and find proof-of-concept for our method.

2 RELATED WORKS

Foundation Models for Robotics: VLMs pre-trained on large scale datasets have shown commonsense reasoning abilities. Researchers have leveraged these abilities to perform planning and control for robotics [8, 10]. Many prior works [2, 12, 13, 18–20, 24, 28, 30, 31, 35] have used pre-trained LLMs to generate actionable natural language plans for robots. VLMs have also been used to generate subgoals for navigation [4, 6, 9, 11, 25, 26] and manipulation [5, 27] tasks. Additionally, prior works have also leveraged LLMs to directly generate low-level executable policy code for robots [17, 29]. Another line of works, has also used LLMs to generate rewards, which can be for RL [14, 21, 34]. In our work, we use a VLM to generate the policy code to accomplish a continuous preference aligned novel goal state.

3 METHOD

We seek a robot policy π to solve a multi-object rearrangement tasks. We query the policy with an initial language instruction l_0 to ground the task, a context variable c that gives N examples of a completed task that implicitly defines a person's preference, and a prompt, p, which is an example of an incomplete task that the policy must solve.

The context and prompt are both comprised of both image and language inputs. For each example in the context, we provide an image of the initial state, s_0, an image of a set of context objects O_c overlaid with spatial reference marks (which have been shown to improve a foundation model's object detection capabilities [33]) and an image of the state after the desired action is performed s_1. We also provide a code file l_1 that uses information from these images and preprogrammed robot actions to solve each example in the context $c = [s_0, O_c, s_1, l_1]$. A prompt p is similar to an example from the code, but contains a different set of objects O_c, does not include s_1, and contains a partial code file $\hat{l}_2$ that must be completed by the policy, e.g. $p = \left[s_0, O_p, \hat{l}_2\right]$.

We parameterize our policy as a vision-and-language model, specifically GPT4-V [7], and test it in initial experiments. We develop a small dataset of four household objects commonly used in table setting: plate, fork, spoon, baby fork. The plate is present in all initial states. Each of the fork, spoon and baby fork and take on one of seven different colors: white, red, black, yellow, green, blue, purple.

4 INITIAL RESULTS

To test our method, we give develop a single-step object rearrangement task. Each experiment consists of two context examples that contain prior placements for the context objects. The context implicitly encode two preferences: the preferred location of the context object, and the preferred context object color. Using this context, the policy should be able to select the appropriate placement and color of a new object in a new table setting. We test the full combination of objects and colors being provided in either the context or prompt, for a total of 63 experiments. This is akin to asking the question: "If I like red forks placed to the left of the place, how do I like to place red spoons?", for each combination of color, context, and prompting objects. We run each experiment ten times. We report color selection and location selection accuracy, broken down by context object and prompt object in Tables 1 and 2, respectively.

These results show a few interesting trends. First, our method exceeds chance in both color (0.14) and location (0.50) selection in all tasks except for predicting the placement of the babyfork from the placement of the spoon. This is likely due to two factors: the tines of the babyfork are wider than those of a normal fork, giving it a more spoon-like appearance, and that the placement of the babyfork is not as strictly bound by traditional Western table-setting decorum as the other objects in this dataset.

5 CONCLUSION AND FUTURE WORK

In this work we present a novel use case for VLMs: using them for assistive human robot collaboration in multi-object rearrangement tasks. We show initial results on a single-step table setting task, which we believe show a proof-of-concept for the current direction. We plan to extend this to multi-step table setting scenarios, and expand the scope of the types of preferences present in our dataset. Finally, we plan to involve human data and perform rigorous testing to determine how well all user preferences can be captured, as opposed to only those that match the distribution of training data well.

REFERENCES

[1] Nichola Abdo, Cyrill Stachniss, Luciano Spinello, and Wolfram Burgard. Robot, organize my shelves! tidying up objects by predicting user preferences. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1557–1564. IEEE, 2015.

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11509–11522. IEEE, 2023.

[5] Yuchen Cui, Scott Niekum, Abhinav Gupta, Vikash Kumar, and Aravind Rajeswaran. Can foundation models perform zero-shot task specification for robot manipulation? In *Learning for Dynamics and Control Conference*, pages 893–905. PMLR, 2022.

[6] Vishnu Sashank Dorbala, Gunnar Sigurdsson, Robinson Piramuthu, Jesse Thomason, and Gaurav S Sukhatme. Clip-nav: Using clip for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2211.16649*, 2022.

[7] OpenAI et. al. Gpt-4 technical report, 2023.

[8] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *arXiv preprint arXiv:2312.07843*, 2023.

[9] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23171–23181, 2023.

[10] Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Zhibo Zhao, et al. Toward general-purpose robots via foundation models: A survey and meta-analysis. *arXiv preprint arXiv:2312.08782*, 2023.

[11] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023.

[12] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022.

[13] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[14] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[15] Mincheul Kang, Youngsun Kwon, and Sung-Eui Yoon. Automated task planning using object arrangement optimization. In *2018 15th International Conference on Ubiquitous Robots (UR)*, pages 334–341. IEEE, 2018.

[16] Ivan Kapelyukh and Edward Johns. My house, my rules: Learning tidying preferences with graph neural networks. In *Conference on Robot Learning*, pages 740–749. PMLR, 2022.

[17] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

[18] Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: From natural language instructions to feasible plans. *arXiv preprint arXiv:2303.12153*, 2023.

[19] Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023.

[20] Zeyi Liu, Arpit Bahety, and Shuran Song. Reflect: Summarizing robot experiences for failure explanation and correction. *arXiv preprint arXiv:2306.15724*, 2023.

[21] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931*, 2023.

[22] Benjamin A. Newman, Reuben M. Aronson, Kris Kitani, and Henny Admoni. Helping people through space and time: Assistance as a perspective on human-robot interaction. *Frontiers in Robotics and AI*, 8, 2022. ISSN 2296-9144. doi: 10.3389/frobt.2021.720319. URL https://www.frontiersin.org/articles/10.3389/frobt.2021.720319.

[23] Benjamin A. Newman, Christopher Jason Paxton, Kris Kitani, and Henny Admoni. Towards online adaptation for autonomous household assistants. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, page 506–510, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399708. doi: 10.1145/3568294.3580136. URL https://doi.org/10.1145/3568294.3580136.

[24] Shreyas Sundara Raman, Vanya Cohen, Eric Rosen, Ifrah Idrees, David Paulius, and Stefanie Tellex. Planning with large language models via corrective re-prompting. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

[25] Dhruv Shah, Michael Robert Equi, Błażej Osiński, Fei Xia, Brian Ichter, and Sergey Levine. Navigation with large language models: Semantic guesswork as a heuristic for planning. In *Conference on Robot Learning*, pages 2683–2699. PMLR, 2023.

[26] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, pages 492–504. PMLR, 2023.

[27] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[28] Tom Silver, Varun Hariprasad, Reece S Shuttleworth, Nishanth Kumar, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Pddl planning with pretrained large language models. In *NeurIPS 2022 foundation models for decision making workshop*, 2022.

[29] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE, 2023.

[30] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023.

[31] Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*, 2023.

[32] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *arXiv preprint arXiv:2305.05658*, 2023.

[33] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

[34] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, et al. Language to rewards for robotic skill synthesis. *arXiv preprint arXiv:2306.08647*, 2023.

[35] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.