CSE 597_003, Vision and Language
The Pennsylvania State University
Fall 2024

Module #4 Paper critique

Jeongwon Bae (945397461)

Promptable Behaviors: Personalizing Multi-Objective Rewards from Human Preferences
(2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR))

The paper [1] introduces Promptable Behaviors, a framework designed to personalize robotic navigation by adapting to diverse human preferences with minimal interaction. Traditional reinforcement learning (RL) and multi-objective RL (MORL) [2] methods often face scalability issues because they require retraining for each new preference or combination of objectives. Promptable Behaviors overcomes these challenges by utilizing single-policy MORL, allowing the agent to adjust its behavior dynamically through a reward weight vector during inference. By conditioning the policy on weight vectors sampled from a simplex, the agent can generalize across multiple objectives without the need for retraining, thereby enhancing both scalability and practicality.

The framework is evaluated on two navigation tasks: Object-Goal Navigation (ObjectNav) and Flee Navigation (FleeNav) [3][4]. In ObjectNav, the agent must locate a specified object within one meter and within a maximum of 500 steps, balancing five objectives: time efficiency, path efficiency, house exploration, object exploration, and safety. FleeNav requires the agent to maximize its distance from the starting location, focusing on time efficiency, house exploration, and safety. The agent processes RGB image observations using a pre-trained CLIP [5] ResNet-50 encoder and incorporates a codebook module to map continuous reward weights to latent representations, facilitating generalization to unseen weight combinations.

Human preferences are integrated into the framework through three methods. First, human demonstrations involve users providing demonstration trajectories, from which the optimal weight vector is inferred by minimizing the negative log-likelihood of the demonstrated actions, aligning the agents behavior with user preferences. Second, preference feedback on trajectory comparisons allows users to compare trajectories generated with different reward weights. Using the Bradley-Terry model [6], the optimal weight vector is determined by maximizing the log-likelihood of these preferences. Group trajectory comparisons are employed to reduce the number of required interactions, achieving high confidence with fewer samples compared to traditional pairwise comparisons. Third, language instructions utilize large language models (LLMs) like ChatGPT [7] to translate natural language instructions into reward weight vectors. Techniques such as in-context learning [8] and chain-of-thought reasoning [9] enhance the models ability to handle complex instructions, enabling non-expert users to specify preferences naturally.

The evaluation metrics include success rate and Success weighted by Path Length (SPL) for ObjectNav, as well as distance ratios and Path Length weighted by Path Length (PLOPL) for FleeNav. Additional metrics assess sub-rewards such as time efficiency, exploration, path efficiency, and safety, all normalized to ensure fair comparisons. Preference prediction methods are evaluated using cosine similarity, the Generalized Gini Index (GGI) [10], and a human-based win rate [11].

Experimental results demonstrate that Promptable Behaviors effectively adapts to different preferences. In the ObjectNav task within the ProcTHOR environment, the framework achieves a 39.6% success rate and 29.8% SPL. When specific objectives are prioritized, performance metrics improve significantly, such as a 56.0% success rate with time efficiency and 65.0% with path efficiency. Prioritizing house exploration leads to a 68.0% success rate, while safety prioritization results in a high safety reward of 0.829, indicating effective collision avoidance. In the FleeNav task, prioritizing house exploration achieves a 73.7% success rate and a PLOPL of 0.861, while safety prioritization maintains a 71.1% success rate with high safety rewards.

Preference prediction methods show high effectiveness: human demonstrations achieve a 70.7% cosine similarity with a single demonstration, group trajectory comparisons reach 93.5% accuracy with only 25 comparisons, and language instructions via ChatGPT attain a 61.4% cosine similarity. Human evaluations reveal that group trajectory comparisons have the highest win rate of 65.0%, outperforming both pairwise comparisons and language instructions.

Qualitative analyses indicate that different prioritized objectives lead to distinct behaviors, such as more direct routes for path efficiency, thorough object inspections for object exploration, and active obstacle avoidance for safety. These behavioral adaptations underscore the frameworks capability to align navigation strategies with user preferences effectively.

Overall, Promptable Behaviors offers a flexible and scalable solution for personalized robotic navigation, efficiently balancing task completion with diverse user preferences through intuitive interaction methods.

The paper addresses a significant problem in robotics: the challenge of personalizing robotic behaviors to align with diverse human preferences without retraining or fine-tuning. In real-world applications, robots must adapt efficiently to varying user preferences, and the proposed framework effectively tackles this issue. By leveraging single-policy MORL, the authors provide a scalable solution that allows the agent to generalize across multiple objectives by conditioning on a reward weight vector. This approach extends beyond traditional RL methods and offers practical benefits in terms of computational efficiency and adaptability.

The literature review is comprehensive and situates the research within the context of existing work in MORL and preference learning. The authors clearly identify limitations in prior methods, such as scalability issues in multi-policy MORL and the inflexibility of single-objective RL approaches. By highlighting these gaps, they effectively motivate the need for their framework and demonstrate how it advances the state of the art.

Experimental validation is robust and thorough. The authors conduct extensive experiments on ObjectNav and FleeNav tasks in the AI2-THOR [12] simulation environment, designed to test the agent's ability to adapt to different user preferences across multiple objectives. Results show that the single-policy MORL agent consistently outperforms baseline models. For example, in ObjectNav, the agent achieves a success rate of 39.6% compared to 35.8% for the single-objective RL baseline, and an SPL of 29.8% versus 28.4%, demonstrating the framework's effectiveness in accommodating various preferences without retraining. The preference inference methods are convincingly demonstrated to be effective. Methods like group trajectory comparisons significantly reduce user interactions while maintaining high accuracy in aligning the agent's behavior with user preferences. This not only enhances user experience but also highlights the practical applicability of the framework in real-world scenarios.

The ablation studies are insightful and validate the proposed components of the framework. By examining the impact of the codebook module and different preference inference methods, the authors provide evidence of each component's significance. For instance, the codebook module improves training stability and generalization, reducing performance variance by 57.8% compared to raw weight encoding. The studies are designed to isolate individual contributions, strengthening the credibility of the findings.

Dataset usage is appropriate and reasonable. The AI2-THOR environment offers a diverse set of procedurally generated indoor scenes, providing a challenging and realistic testbed for evaluating the agent's performance. The use of both ProcTHOR and RoboTHOR environments ensures robustness across different scenarios, enhancing the generalizability of the results.

One of the main weaknesses is the assumption that human preferences can be fully captured through linear combinations of predefined objectives. In real-world scenarios, human preferences are often complex, non-linear, and context-dependent. This simplification may limit the framework's ability to handle nuanced or dynamic preference structures, potentially affecting its applicability in more sophisticated settings. For example, preferences that involve conditional dependencies or prioritization based on context might not be adequately represented by fixed weight vectors.

The reliance on LLMs like ChatGPT for interpreting language instructions introduces potential issues. LLMs may misinterpret ambiguous, colloquial, or context-specific language, leading to incorrect reward weight estimations. The paper does not extensively analyze how such misinterpretations impact the agent's performance or propose strategies to mitigate these risks. This dependence on LLMs could hinder the reliability of the preference inference process in practical applications, especially when users provide instructions that the model has not been trained to understand.

All experiments are conducted in simulated environments, which may not fully capture the complexities of real-world settings. Factors such as sensor noise, dynamic obstacles, unmodeled environmental conditions, and hardware limitations are absent in simulations but can significantly affect performance in real deployments. The lack of experiments on physical robots raises questions about the framework's robustness and practicality when transferred to real-world applications. Without testing in physical environments, it's unclear how the agent would handle unexpected variables that are commonplace outside simulations.

While the ablation studies validate certain components like the codebook module, they do not thoroughly examine all architectural choices and proposed components. For instance, the impact of different network architectures, hyperparameter settings, or alternative encoding methods for the reward weight vector is not fully explored. This limited scope may leave some questions unanswered regarding the optimal design choices for the framework. A more comprehensive exploration could identify potential improvements or reveal sensitivities in the system.

The datasets used, although diverse within indoor environments, do not cover outdoor scenarios or other types of environments where robots might operate. This restriction limits the assessment of the framework's generalizability to a broader range of real-world contexts. Additionally, the evaluation does not address how the agent would perform under varying environmental conditions such as different lighting, textures, or dynamic changes, which are common in real-world applications.

This empirical research focuses on developing and validating a novel framework through experimental evaluations in simulated environments. It presents a significant advancement in personalized robotics by addressing the challenge of adapting to diverse human preferences without retraining. Efficiently personalizing robotic behaviors is essential for integrating robots into everyday environments where user preferences vary and change over time.

The major findings demonstrate that a single-policy MORL agent can effectively generalize across multiple objectives by conditioning on a reward weight vector. The framework's integration of human preferences through demonstrations, preference feedback, and language instructions shows versatility and accessibility. Results indicate that the agent can adapt its behavior to align with user preferences, achieving higher success rates and better performance metrics compared to baseline models. Methods like group trajectory comparisons significantly reduce user burden while maintaining high accuracy in preference inference.

These findings offer a practical solution to scalability issues in RL and MORL, enabling robots to be more responsive to individual user needs without significant computational costs. The framework's adaptability and efficiency have the potential to impact various applications, from service robots in homes to autonomous navigation in dynamic environments.

To improve this research, future work could focus on modeling non-linear and context-dependent preferences to better capture the complexity of real-world human preferences. Incorporating more sophisticated preference learning algorithms or adaptive reward structures could enhance flexibility. Extending experiments to physical robots in real-world environments would provide valuable insights into robustness and practical applicability. Addressing limitations of using LLMs by developing more reliable natural language understanding mechanisms or incorporating interactive clarification processes could mitigate misinterpretation risks. Exploring passive or less burdensome preference inference methods could further reduce user effort and improve the user experience.

In conclusion, the paper presents a promising approach to personalizing robotic behaviors, offering a scalable and efficient framework that effectively integrates human preferences. By addressing identified weaknesses and expanding future research, this framework could significantly contribute to developing adaptive robotic systems capable of seamlessly integrating into diverse human environments and meeting individual user needs.

References

[1] M. Hwang, L. Weihs, C. Park, K. Lee, A. Kembhavi, and K. Ehsani, "Promptable Behaviors: Personalizing Multi-Objective Rewards from Human Preferences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[2] C. F. Hayes, R. Ruadulescu, E. Bargiacchi, J. Kallstrom, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, E. Howley, A. A. Irissappane, P. Mannion, A. Now'e, G. de Oliveira Ramos, M. Restelli, P. Vamplew, and D. M. Roijers, "A practical guide to multi-objective reinforcement learning and planning," *Autonomous Agents and Multi-Agent Systems*, vol. 36, 2021.

[3] M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford, L. Weihs, M. Yatskar, and A. Farhadi, "RoboTHOR: An Open Simulation-to-Real Embodied AI Platform," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3161–3171, 2020.

[4] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, J. Salvador, K. Ehsani, W. Han, E. Kolve, A. Farhadi, A. Kembhavi, and R. Mottaghi, "ProcTHOR: Large-Scale Embodied AI Using Procedural Generation," in *NeurIPS*, Outstanding Paper Award, 2022.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021.

[6] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, p. 324, 1952.

[7] O. J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H.-i. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I.-a. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, *et al.*, "Gpt-4 technical report," 2023.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20, Vancouver, BC, Canada: Curran Associates Inc., 2020, ISBN: 9781713829546.

[9] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *ArXiv*, vol. abs/2205.11916, 2022. [Online]. Available: `https://api.semanticscholar.org/CorpusID:249017743`.

[10] R. Busa-Fekete, B. Szörényi, P. Weng, and S. Mannor, “Multi-objective bandits: Optimizing the generalized Gini index,” in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, Jun. 2017, pp. 625–634.

[11] J. Jang, S. Kim, B. Y. Lin, Y. Wang, J. Hessel, L. Zettlemoyer, H. Hajishirzi, Y. Choi, and P. Ammanabrolu, “Personalized soups: Personalized large language model alignment via post-hoc parameter merging,” *arXiv preprint arXiv:2310.11564*, 2023.

[12] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, “AI2-THOR: An Interactive 3D Environment for Visual AI,” *arXiv*, 2017.