Enhancing Safety and Precision in Robotic Manipulation Tasks through Human Action Interventions

Jenny Wang*, Junyi Li*, Li-Chun Huang*, Xiao Huang*
Department of Computer Science, UCLA MEng Program, Los Angeles, CA, USA
{jianingwang99, lijunyi, lichhuan, xiaohuang}@g.ucla.edu
*These authors contributed equally to this work.

I. Introduction

In real-world robotic manipulation tasks, autonomous systems must prioritize safety and precision to avoid undesirable or dangerous outcomes. However, traditional reinforcement learning (RL) methods often rely on manually designed reward functions that fail to capture the complexity of real-world scenarios or fully align with human intentions. This misalignment frequently leads to suboptimal or unsafe behavior, such as prioritizing efficiency over safety constraints. Moreover, traditional RL feedback only contains either success or failure, which is too simple to capture human's complex intentions and provide useful information to agents, limiting the agent's overall performance.

Our motivation for this project is to overcome these limitations by incorporating active human involvement into the learning process. The reason is because the human-in-the-loop paradigm allows humans to provide guidance during the agent's training, which is particularly critical for ensuring safety and precision in situations where agents' wrong decisions may lead to significant risks. By intervening during training, humans can correct unsafe or suboptimal behaviors, enabling the agent to fulfill task requirements under safety guidance[1].

Building on previous works like Sirius, which emphasized real-time human feedback for robotic manipulation tasks, our approach integrates reinforcement learning with these insights and extends the Proxy Value Propagation (PVP) framework. Specifically, we assign a fixed weighting of human-to-agent ratio of 2:1 to data during training. This ensures that the agent prioritizes human guidance, which is of higher quality while still benefiting from an autonomous setting. By maintaining an optimal balance, we aim to enhance safety and precision in robotic manipulation tasks through targeted human interventions.

In this paper, we will explore two central questions:

(1) What is the most effective way for humans to provide feedback for RL agents?
(2) How can we model and interpret human action intervention to optimize performance?

By addressing these, we aim to demonstrate the advantages of real-time, action-based feedback, ultimately advancing the reliability and safety of RL in complex tasks.

II. Problem Formulation

In real-world robotic manipulation tasks, reinforcement learning (RL) agents face two major challenges:

High Precision Requirements and Safety: RL agents often struggle to meet human expectations for high precision and safety in tasks such as robotic arm manipulation. Current methods rely heavily on manually designed reward functions to guide agent behavior, but these are often inadequate for capturing the full complexity of human intentions. This limitation can lead to undesirable agent actions that deviate from expected outcomes, sometimes resulting in unsafe behaviors.

To address these challenges, we model safety compliance using a ground-truth compliance indicator function $C(s, a)$, defined as:

$$C(s,a) = \begin{cases} 1, & \text{if } a \text{ violates human intention,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

This indicator function is denoted as whether an agent's action a in state s violates human intentions or not. We further define the upper bound of the discounted occurrence of intent violations, S_{π_b}, as a measure of training time human intent compliance:

$$S_{\pi_b} = \mathbb{E}_{\tau \sim P_{\pi_b}} \sum_t \gamma^t C(s_t, a_t), \tag{2}$$

where S_{π_b} denotes the expected discounted sum of violations, γ represents the discount factor, and P_{π_b} is the probability distribution of trajectories deduced by the behavior policy π_b [1].

Lack of Real-Time Human Intervention: While human-in-the-loop approaches attempt to incorporate human interventions to enhance agent performance, current real-time intervention methods remain limited. Most approaches provide only passive human feedback when the agent is already performing

unsafe actions, which delays corrective feedback, thereby impacting both learning efficiency and behavioral accuracy.

To address this, we use a mixed behavior policy in the way that a human shares control with the learning agent (human-AI shared control):

$$\pi_b(a|s) = (1-I(s,\mu_\pi(s)))\delta(a-\mu_\pi(s))+I(s,\mu_\pi(s))\pi_h(a|s), \tag{3}$$

where $I(s,a)$ denotes human intervention, $\mu_\pi(s)$ represents a deterministic policy, $\pi_h(a|s)$ is the stochastic human policy, and δ means a deterministic action.
We make two important assumptions regarding the human subject[1]:

- **Assumption 1:** The probability of a human producing an undesired action is bounded by a small value $\epsilon < 1$ during human-AI shared control:

$$\mathbb{E}_{s\sim P_{\pi_b}, a\sim\pi_h(a|s)} C(s,a) \leq \epsilon. \tag{4}$$

- **Assumption 2:** The probability of the human failing to intervene when the agent's action is undesired is bounded by $\kappa < 1$ during human-AI shared control:

$$\mathbb{E}_{s\sim P_{\pi_b}} \left[(1 - I(s,\mu_\pi(s)))C(s,\mu_\pi(s))\right] \leq \kappa. \tag{5}$$

By using the upper bound of intent violation theorem, we define the intervention rate ψ as[1]:

$$\psi = \mathbb{E}_{s\sim P_{\pi_b}} I(s,a_n). \tag{6}$$

and the discounted occurrence of intent violation S_{π_b} as:

$$S_{\pi_b} \leq \frac{1}{1-\gamma}(\kappa + \epsilon\psi). \tag{7}$$

This is the formal mathematical framework for analyzing safety, learning, and human intervention in RL, which minimizes the undesired and unsafe actions while ensuring that agents learn policies that align with human preferences.

In short, the central question we aim to address is: How can we design an RL agent capable of safely, precisely, and efficiently learning and executing tasks that align with human expectations in high-demand, real-world environments?

III. Potential Questions and Responses

In this section, we aim to address some key questions that may be raised regarding our approach. By clarifying these questions, we hope to make clear the rationale behind our idea and the advantages of our approach.

A. Why use corrective actions instead of simpler positive/negative signals for human feedback?

While simpler positive/negative signals can provide general guidance, they lack precision, which is essential under safety-critical environments. In contrast, corrective actions allow human supervisors to intervene directly, providing specific adjustments to the agent's actions effectively and thus guaranteeing a safer and desirable outcome. This approach can not only reduce unsafe actions from the agents through human interventions but also ensure the outcomes align with human expectations.

B. Does the use of corrective actions limit the agent's exploration capabilities?

While the use of corrective actions has the potential to limit the agent's exploration capabilities, the costs of unsafe actions are too high in safety-critical environments. As the agent's performance improves through human interventions and the agent demonstrates reliability, we can gradually reduce the frequency of interventions, becoming less pessimistic and allowing the robot to explore more independently. In this case, the limitation of the agent's exploration capabilities can be alleviated to the greatest extent.

C. Why choose reinforcement learning over imitation learning in this context?

While imitation learning is effective, it performs well only for initial behavior modeling and lacks adaptability under unpredictable scenarios. In contrast, reinforcement learning with human interventions allows the agent to adjust and learn continuously from both environmental rewards and corrective feedback, making it more adaptable to changeable environments. In our context, this combined approach enables the agent to achieve better performance in complex and changeable settings.

D. How does this approach handle sparse reward environments?

We handle sparse reward environments by providing the agent with structured guidance even in the absence of frequent rewards. In our approach, the agent learns to interpret, adapt to, and respond to human feedback, which alleviates the limitation of sparse rewards and improves the overall learning efficiency.

IV. Related Work

In recent years, Reinforcement Learning (RL) has proven to be applied in robotic manipulation tasks. However, its reliance on manually designed reward functions often fails to align with complex human intentions, particularly in safety-critical and high-precision scenarios. This limitation has led to further research on exploring frameworks that integrate human feedback to improve learning efficiency and safety.

Traditional human-in-the-loop RL methods introduced how to use human feedback to guide agent behavior, such as trajectory ranking or preference-based advice [2] [3]. While these methods improved alignment with human intentions, they relied heavily on passive feedback. That is to say, intervention only appears after unsafe behaviors occur. This delayed intervention limited their ability to prevent unsafe actions proactively and often reduced learning efficiency.

The Proxy Value Propagation (PVP) framework introduced a significant step forward by leveraging human interventions to update the value function directly. It assigns a fixed value of +1 to human-corrected actions and -1 to autonomous actions requiring intervention [1]. This simple binary weighting effectively captures basic human intentions and improves policy learning efficiency.However, the fixed binary weighting (+1 and -1) oversimplifies the complexity of real-world scenarios. For instance, not all human corrections are equally critical. Sometimes, autonomous actions from the agent may provide more valuable exploratory behaviors. This lack of a balanced weighting to reflect the varying importance and quality of actions can lead to inefficient updates and suboptimal policies in practice. The Sirius framework emphasized real-time human interventions during robotic manipulation tasks, using behavior cloning to guide learning during task execution [4]. While Sirius introduced dynamic human feedback mechanisms, its reliance on supervised learning limited its adaptability, as it did not integrate reinforcement learning to propagate human feedback into policy updates.

Building on previous works, we try to aggregate the strengths of the existing approaches. Our approach combines the reinforcement learning foundation of PVP with the idea of reweighting from the paper "Robot Learning on the Job: Human-in-the-Loop Autonomy and Learning During Deployment". Specifically, we address the limitations of PVP's fixed binary weighting by introducing a reweighting mechanism that adjusts the importance of human corrections and autonomous actions based on their quality and contribution to learning. This approach will focus on solving: 1. Capturing Action Complexity: Adjusting weights to prioritize critical human corrections while preserving the exploratory value of certain autonomous actions. 2. Improving Learning Efficiency: Reducing unnecessary human interventions while making better use of feedback to optimize policy updates.

V. Proposed Method and Evaluation

The proposed method enhances the Proxy Value Propagation (PVP) framework by integrating a reweighting mechanism for agent-collected and human-collected data during training. The primary objective is to leverage the superior quality of human-collected data to align the agent's learning process with human intentions while retaining the exploratory benefits of agent-generated data. In our project, our proposed method is defined as PVP_reweight.

The implementation begins with data collection, where datasets are gathered from both human interventions and agent actions to ensure comparable size and diversity. while ensuring the human data is consistent, allowing for better comparison between our method and the baseline. During training, higher weights are assigned to human-collected data, reflecting its higher quality and critical insights. Specifically, a human-to-agent weighting ratio of 2:1 is applied, which allows the model to prioritize human data without completely suppressing the exploratory contributions of agent data. This balance ensures that the agent benefits from both high-quality guidance and self-directed learning.

The training process utilizes these reweighted datasets to fine-tune the agent's policy. By incorporating reweighted values directly into the PVP framework, the agent can achieve higher performance. This approach extends the applicability of PVP beyond traditional environments, enabling it to handle safety-critical tasks more efficiently and reliably.

The decision to prioritize human-collected data stems from its intrinsic quality and relevance in guiding the agent's learning. Human interventions encapsulate domain-specific expertise and safety considerations that are crucial for preventing undesirable actions in high-stakes environments. By reweighting the data, the proposed method ensures that these insights are leveraged effectively while maintaining the agent's ability to explore and learn independently.

The chosen ratio of 2:1 reflects a carefully considered trade-off between the cost of human interventions and the need for autonomous learning. It emphasizes human guidance where it is most impactful while preserving the agent's capacity for adaptation and exploration. This balance is particularly important in scenarios requiring both precision and safety. We also conducted a sensitivity analysis experiment on the reweight ratio. As shown in Figure 3, applying different weights to the human data has a significant impact on the experiment. This also sparked our thinking about whether, in the future, we can design an adaptive ratio that allows the agent to better leverage both human data and agent data.

The effectiveness of the proposed method is evaluated in the MetaDrive environment, which is an open-source driving simulator. To assess performance, we measure success rates both with and without the reweighting technique. The success rate reflects the agent's ability to complete tasks with or without human intervention. Based on the results, we will further refine the PVP approach to enhance better performance.

VI. Experiment

We start by evaluating the performance of PVP_reweight_TD3 (ours) compared to the baseline PVP_TD3 in the MetaDrive environment. The experiments are designed to test robots' performance on autonomous driving. The evaluation is conducted at regular training checkpoints to measure the agent's performance using two key metrics: (1) **Average Episode Reward**, which reflects the agent's ability to maximize cumulative rewards, and (2) **Average Success Rate**, which indicates the percentage of episodes in which the agent successfully achieves the task goal. To ensure robust results, each checkpoint evaluation includes multiple episodes.

The dataset used in our experiments is **self-collected** and consists of identical human-collected data, which is then used for training both PVP and PVP_reweight models for

comparison. To collect the data, we first run the PVP baseline to simulate tasks, and then we perform human interventions to manually intervene if there is any unsafe or suboptimal action. After that, we will save all human intervention data into a replay buffer, and we will then use the stored human intervention data as the shared dataset for training both the PVP and PVP_reweight models.

We also conducted a sensitivity analysis on the reweighting ratio. Using the shared dataset we manually collected, we adjusted the ratio of human data to agent data during the training process to 2:1, 3:1, and 4:1. The results of this experiment are shown in Figure 3, where the y-axis represents the Average Success Rate. We can observe that applying different weights to the human data has a significant impact on the agent's performance, with the 2:1 ratio yielding the best results in our experiments.

Below are the results we gathered during the experiment, illustrating the performance of PVP_reweight_TD3 and PVP_TD3 baseline across key metrics, with the human to agent's ratio comparison.

Fig. 1. Average episode reward across training checkpoints. The line represents the mean reward, and the shaded region indicates the standard deviation across episodes.

VII. Results and Discussion

In the MetaDrive environment, we compare the performance of our proposed PVP_reweight method to the PVP baseline. The results, as depicted in Figures 1–2, demonstrate that PVP_reweight consistently outperforms the baseline across key evaluation metrics, which are average episode reward and average success rate.

Although our proposed method can outperform the baseline, there still remains potential room for improvement. First, the assumption that agents always act poorly while humans always act well is too naive. In the early stage of training, the agent usually performs poorly since it is underfitting. However, in the later stages, when the agent's policy is well-trained, the assumption that the agent underperforms compared to humans becomes questionable. This can mislead the policy update direction and sometimes damage the trained policy. Therefore, reward scheduling might be a potential direction for improving our algorithm.

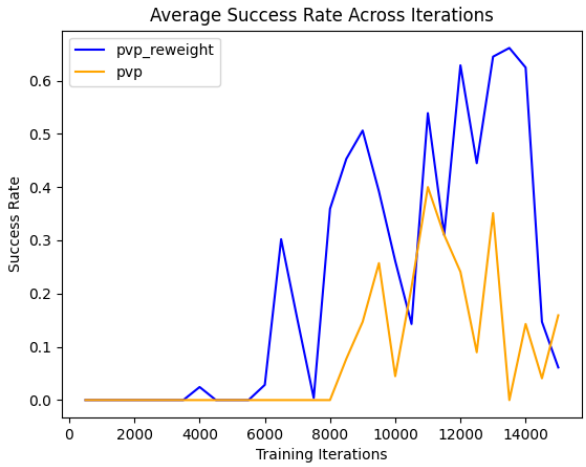

Fig. 2. Average Success Rate Across Iterations. The plot demonstrates the average success rate achieved by the agent over training iterations.

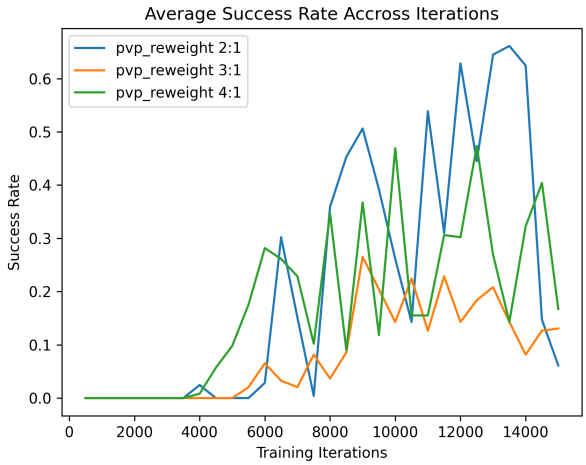

Fig. 3. Average Success Rate Across Iterations for different reweighting ratios. The plot conveys the information that Human:Agent = 2:1 showcases the best performance compared to ratios of 3:1 and 4:1.

Second, the quality of human actions is unstable and influenced by the intervention devices used, differences between individuals, and the concentration level throughout the entire training process. We discovered that sometimes we could not provide suitable human intervention on time, especially when the training process lasted for several hours. Whenever undesirable actions are provided, they cannot be detected or reverted in time. These undesired human actions will be loaded into the human action buffer and affect the policy.

Lastly, the algorithm we proposed involves the policy in the data collection process. In other words, the data quality is correlated with how we initialize the policy, update it, and sample data from the buffer. Moreover, since the data collection process also involves human interaction, the quality of the data is quite unstable. Whether it is suitable to compare various similar algorithms using self-collected data is a question. We believe it is necessary to include data quality measurement and a method to stabilize the data collection process.

VIII. Conclusion

In this research, we discovered that applying reweighted sampling to agent and human data significantly improves performance, especially when training a reward-free RL algorithm with human intervention. This framework requires considerably fewer human demonstrations compared to traditional imitation learning methods like **behavior cloning** or **inverse reinforcement learning**.

Since rewards in this approach are derived directly from human actions, the algorithm can faithfully capture desired human behavior without interference, ensuring alignment with human intention. While the method demonstrates impressive capabilities in imitating human actions, there is room for further improvement in algorithm performance.

Key areas for future explorations include:

- **Sample Weight Scheduling:** Dynamically adjusting the importance of human and agent data samples during training.
- **Reward Scheduling:** Optimizing the structure and timing of rewards on agent and human action to better guide learning.
- **Data Quality Measurement:** Establishing metrics to evaluate and ensure the quality of human demonstration data.
- **Enhanced Data Collection Procedures:** Improving the process to collect human intervention data for training.

This research represents an initial step into the promising field of human action learning with intervention in high-precision, safety-critical tasks. We are optimistic about the potential advancements and improvements that can be made to the proposed method. Looking ahead, this integrated human-in-the-loop method may broaden real-world applications for safer, more adaptive robotic systems.

Appendix: Team Contributions

Jenny Wang

- Drew the visualizations for the comparison of PVP and PVP_reweight using the average reward rate and average success rate metrics.
- Created the slides and presented the motivation, experiment results, and conclusion parts in the final presentation.
- Revised the paper format and wrote for the Introduction, Problem Formulation, Potential Questions and Responses, and the Experiment part in the final report.

Junyi Li

- Research about Sirius framework
- Present Prior Works, Key Insights and Contribution in the final Presentation
- Wrote Related Work section for the final report
- Checked and revised the Final Report

Li-Chun Huang

- Setup the environment and reproduce existing work (PVP).
- Implemented the proposed algorithm in PVP's environment.
- Performed the experiment for both PVP's baseline and proposed method.
- Presented the method part in the final presentation.
- Wrote Result and Conclusion part in the final report.

Xiao Huang

- Implemented the integration of the Sirius reweighting method into the PVP framework in code.
- Set up the PVP environment and conducted experiments for both PVP's baseline and the proposed method.
- Designed and executed experiments to analyze the proposed method under different reweighting ratios.
- Wrote Proposed Method and Evaluation, and Experiment part in the final report.

References

[1] Z. Peng, W. Mo, C. Duan, Q. Li, and B. Zhou, "Learning from active human involvement through proxy value propagation," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[2] C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz, "A survey of preference-based reinforcement learning methods," *Journal of Machine Learning Research*, vol. 18, no. 136, pp. 1–46, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-634.html

[3] T. Mandel, Y.-E. Liu, E. Brunskill, and Z. Popović, "Where to add actions in human-in-the-loop reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[4] H. Liu, S. Nasiriany, L. Zhang, Z. Bao, and Y. Zhu, "Robot learning on the job: Human-in-the-loop autonomy and learning during deployment," 2023. [Online]. Available: https://arxiv.org/abs/2211.08416