Impact of Explanations on Transparency in HRI: A Study Using the HRIVST Metric

Nandu Chandran Nair[1][0000−0001−8284−0609], Alessandra Rossi[1][0000−0003−1362−8799], and Silvia Rossi[1][0000−0002−3379−1756]

Universita' degli Studi di Napoli Federico II, Naples, Italy

Abstract. This paper presents an exploration of the role of explanations provided by robots in enhancing transparency during human-robot interaction (HRI). We conducted a study with 85 participants to investigate the impact of different types and timings of explanations on transparency. In particular, we tested different conditions: (1) no explanations, (2) short explanations, (3) detailed explanations, (4) short explanations for unexpected robot actions, and (5) detailed explanations for unexpected robot actions. We used the Human-Robot Interaction Video Sequencing Task (HRIVST) metric to evaluate legibility and predictability. The preliminary results suggest that providing a short explanation is sufficient to improve transparency in HRI. The HRIVST score for short explanations is higher and very close to the score for detailed explanations of unexpected robot actions. This work contributes to the field by highlighting the importance of tailored explanations to enhance the mutual understanding between humans and robots.

Keywords: Social robotics · Explanations · Transparent HRI

1 Introduction

As robots become collaborative partners in various human-centered domains, researchers strive to facilitate clear understanding and control for people while enabling robots to interpret human cues, and accordingly adapt their behaviors [1, 2]. This is particularly important because transparent communication between humans and robots fosters comfortable and effective collaboration [3,4]. While previous studies have examined different cues to make human-robot interaction more transparent [5, 6], our focus is on verbal interaction and the role of explanations. Verbal communication is vital in enhancing the clarity and success of an HRI [7]. Our long-term goal is to explore strategies that empower effective communication between humans and robots both during activities performed in collaboration and not [8].

Focusing on evaluating the effectiveness of explanations in promoting transparency in HRI, we propose five different conditions of explanations that aim to evaluate the explanations based on the content and timing in which they are provided. We use the HRIVST (Human-Robot Interaction Video Sequencing Task) [9] measure to evaluate the legibility and predictability of very simple scenarios where a robot has "to fold" and "pick and place" some clothes. In this scenario, the robot may exhibit errors or have unexpected behaviors.

We recruited 85 participants and filtered the people in our quality check process. The outcomes indicate that the highest transparency score is attributed to providing short explanations in real-time. Notably, the second-highest score corresponds to situations where detailed explanations are provided upon a change in the robot's plan. In this study, we provide an evaluation of the multifaceted relationship between HRI explanations and transparency, and as a consequence, we identify which type of explanation contributes to improving people's trust in and understandability of robots through systematic experimentation and by using the HRIVST metric.

2 Related Work

Transparency allows humans to be aware of the state of a robot and assess the progress of tasks [10]. One way of providing transparent interaction is by using explanations. Explanations in HRI refer to a robot providing justifications or reasons for its decisions or actions. These explanations enhance user perceptions, justify the robot's reliability, and increase trust [11]. This paper addresses two main aspects of robot explanations: the content of the explanation and the timing of when the explanation is provided. It explores how the information conveyed in the explanation and the moment it is delivered impact the interaction between people and robots.

A relevant study [12] presented an experiment involving 366 participants to explore whether robots should provide explanations and examine the attributes of a desired explanation. These attributes encompass timing, the significance of engagement, resemblance to human explanations, and the act of summarization. The findings revealed a consensus among participants that robot behavior warrants explanation across the scenarios. It is to be noted that people's preferred mode of explanation aligns with how humans explain things in context. Participants appreciated concise summaries and preferred the robot to respond to only a limited number of follow-up questions.

While explanations alone may not significantly impact perceived competence intelligence, likeability, or safety ratings of the robot [11], they do contribute to the perception of the robot as more lively and human-like [11]. There are different types of explanations for HRI. One study evaluated the effectiveness of contrastive, causal, and example explanations in supporting human understanding of Artificial Intelligence (AI) in a hypothetical scenario [13]. Another study proposed a framework for generating explanations in autonomous robots focusing on presenting the minimum necessary information to understand an event [14]. Additionally, research on progressive explanations aims to improve understanding by limiting cognitive effort at each step [15]. Furthermore, human-like explanations based on the probability of success have been explored to make explanations more understandable for non-expert users [16].

Unlike prior research on robots explaining after being asked, this paper focuses on proactive explanations generated before actions are executed. The study investigates how these proactive explanations influence human-robot trust dynamics [17]. Prior work has shown that explanations, especially those of a complex nature, should be made in real-time during the execution of tasks. This helps spread the information to be explained and reduces the mental workload of humans in highly cognitively demanding tasks [18]. Moreover, the order in which the information is presented in an explanation or the progressiveness of the explanations can contribute to better learning and understanding [19].

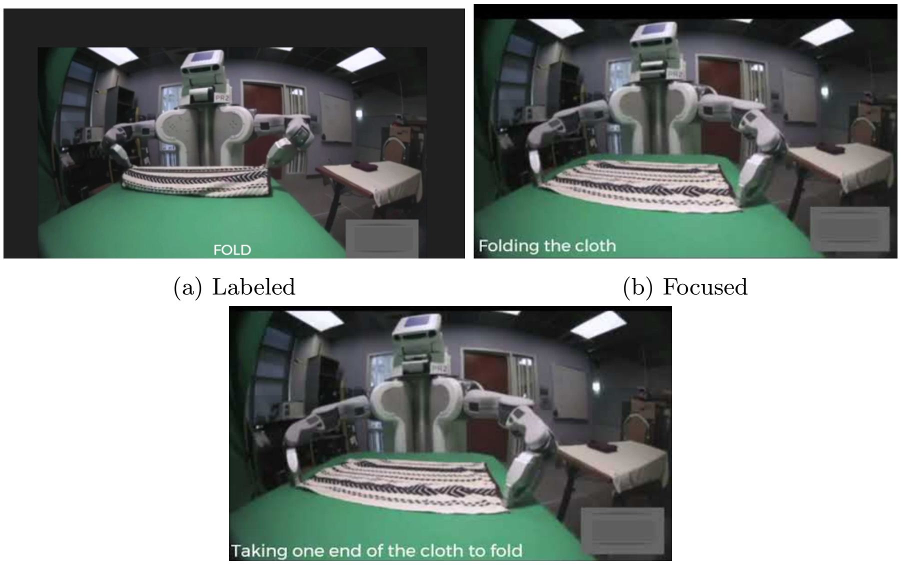
3 Explanation Types

In this work, we want to focus on different explanation types based on the content size. In particular, we considered the following types of explanations to identify the most effective strategies for enhancing mutual understanding and trust between humans and robots:

- **Labeled Explanation**: Labeled explanations are presented as succinct labels, where each label corresponds to a specific robot action. For example, if a robot is observed moving towards a door, the accompanying labeled explanation would be "MOVE". This concise explanation encapsulates the essence of the robot's action in a single keyword, making it an easily graspable reference (see Figure 1a).
- **Focused Explanation**: The focused explanation involves crafting sentences succinctly conveying the robot's actions while maintaining clarity and directness. For instance, if a robot is seen moving toward a door, the focused explanation would be, "Move towards the door". This approach provides a more detailed description than the labeled explanation while remaining concise and to the point (see Figure 1b).
- **Comprehensive Explanation**: Comprehensive explanations represent a more elaborate form of communication. In this type, sentences are constructed to encompass not only the robot's action but also additional contextual information that aids in understanding the intent and purpose behind the action. For example, if a robot is observed moving toward a door, the comprehensive explanation would provide a detailed description: "Move from the room's right side to the left to open the red door." This in-depth narrative offers a holistic view of the robot's actions and underlying motivations (see Figure 1c).

4 Methodology

For this study, we selected four videos where a robot performs a simple task. Each task consists of two or more actions. We use three videos in which the



(a) Labeled

(b) Focused

(c) Comprehensive

Fig. 1: Examples of explanation types for Video 4: (a) "Fold", (b)"Folding the cloth", and (c) "Taking one end of the cloth to fold".

robot picks and places a cloth and a fourth video in which the robot folds a cloth. Table 1 shows the description of the four videos.

Participants were tested with one of the following five different conditions:

- **No Explanation**: In this condition, the videos do not have any explanations. By observing participants' reactions and understanding when no explanation is provided, we establish a baseline for measuring the impact of explanations on transparency (see Figure 2e).
- **Focused Explanation**: The second condition involves providing participants with short explanations accompanying the videos. These explanations are designed to succinctly describe the robot's actions while the actions are being performed. This real-time provision of information aims to enhance transparency by offering immediate insights into the robot's intentions and tasks (see Figure 2a).
- **Comprehensive Explanation**: Participants will receive detailed explanations in this third condition. Similar to the above condition, these explanations will be delivered in real-time while the robot is engaged in its actions. The comprehensive nature of these explanations intends to provide a deeper understanding of the robot's actions, including contextual details that contribute to transparent communication (see Figure 2b).
- **Alerted Focused Explanation**: The fourth condition introduces a novel element of explanation timing. Here, the focused explanations will be provided to alert the human observer about the robot's actions. This condition

Table 1: Description of each video

No.	Videos	Description
1	Video 1	A robotic arm moves to the right side of the table. The robot picks up the folded cloth and moves to the left side of the table. The robot places the folded cloth on top of another cloth.
2	Video 2	A robotic arm moves to the right side of the table. The robot picks up the folded cloth. The cloth gets unfolded. The robot stops and moves toward the human. The robot hands over the cloth to fold it up.
3	Video 3	A robotic arm moves to the right side of the table. The robot picks up the folded cloth very slowly without unfolding it. The robot carefully places the folded cloth on top of another cloth.
4	Video 4	A robot folds the cloth vertically and flats it. Then, it folds the cloth horizontally from the left side and right side. Then, it picks up the cloth and moves it to the other side of the room to put it into another table.

is especially relevant in scenarios involving robot failures or changes in plan of action. By focusing on the importance of explanations during these critical moments, we aim to ascertain the impact of timely explanations on transparency and overall human-robot interaction (see Figure 2c).

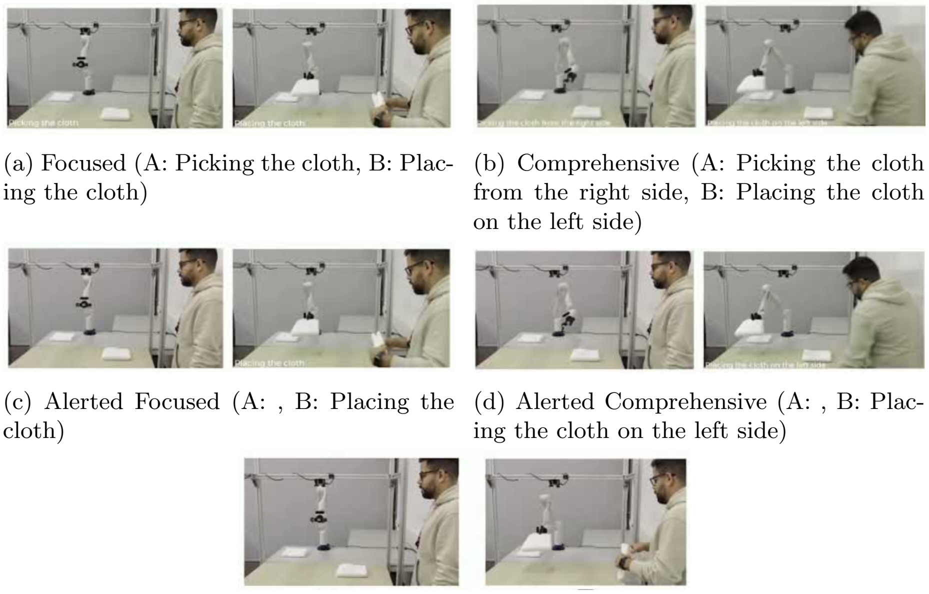
- **Alerted Comprehensive Explanation**: The fifth condition also focuses on the timing of the explanation. However, comprehensive explanations are provided instead of focused explanations to alert the observer. This condition aims to validate by giving details of the robot's actions to understand the change of plan (see Figure 2d).

4.1 Evaluation

We used the HRIVST to test if a robot's behavior is understandable to humans. The HRIVST metric is a subjective measure to evaluate the legibility of a robot's behavior by assessing individuals' capacity to discern goal-oriented actions [9]. The methodology involves segmenting the videos into several distinct clips, each corresponding to an action executed by the robot or the involved individuals during the interaction. Participants are prompted to view these video clips and arrange them in the order that reflects the chronological sequence of task actions. Participants could repeatedly watch the clips, enabling them to grasp the action sequence accurately and familiarize themselves with the task.

Participants were required to complete a brief questionnaire following each video clip to indicate the robot's intention, their expectation of the robot's actions, and their confidence level in attributing the robot's intention (i.e., whether it was difficult or easy).

The cumulative HRIVST score is derived from two components: the outcome of the logical sequence task, ranging from 0 to 6, and the responses provided in the questionnaire, which also have a potential range of 0 to 2. These two



(a) Focused (A: Picking the cloth, B: Placing the cloth)

(b) Comprehensive (A: Picking the cloth from the right side, B: Placing the cloth on the left side)

(c) Alerted Focused (A: , B: Placing the cloth)

(d) Alerted Comprehensive (A: , B: Placing the cloth on the left side)

(e) No Explanation (A: , B:)

Fig. 2: Comparison of explanation in the five conditions for Video 2.

components constitute the total HRIVST score for each video, yielding a possible score range of 0 to 8. The scoring mechanism for the logical sequence task is designed as follows:

- Both the first and last video clips are each assigned 2 points.
- For the others, if a participant correctly orders them in the sequence, they are awarded 2 points divided by the number of remaining clips.

For instance, in a video composed of 4 clips, the first and last clips would be worth 2 points each, while each centrally positioned clip would carry a potential value of 1 point if accurately sequenced.

5 Preliminary Results

The study was conducted online. The obtained participant distribution thus far is as follows: 30 participants in the "No Explanation" condition, 12 in the "Focused Explanation" condition, 21 in the "Comprehensive Explanation" condition, 10 in the "Alerted Focused Explanation" and 12 in the "Alerted Comprehensive Explanation" condition. Control questions were employed to ensure data quality. Consequently, participants providing incorrect responses were filtered out, resulting in the final participant counts of 21, 10, 14, 8, and 6 for the respective conditions. The final 58 participants included people of various nationalities.

Table 2: Descriptive Statistics of the HRIVST for different conditions. For each video, the highest value is reported in bold.

Conditions	**Video 1**	**Video 2**	**Video 3**	**Video 4**	**Average**
No Explanation Mean (SD)	6.23 (2.58)	7.00 (2.00)	6.71 (2.10)	6.57 (2.57)	6.63 (2.31)
Focused Mean (SD)	6.75 (2.10)	6.67 (2.13)	**7.50 (1.12)**	**7.86 (1.30)**	**7.19 (1.66)**
Comprehensive Mean (SD)	**6.76 (2.55)**	5.00 (2.35)	5.47 (2.03)	5.75 (2.61)	5.75 (2.38)
Alerted Focused Mean (SD)	5.5 (0.67)	5.2 (2.32)	6.67 (1.25)	6.45 (1.55)	5.95 (1.45)
Alerted Comprehensive Mean (SD)	6.58 (2.51)	**7.08 (2.03)**	6.38 (2.14)	6.86 (2.22)	6.73 (2.23)
Min - Max	0 - 8	0 - 8	1 - 8	1 - 8	

Table 3: T-Statistics for Videos

Video Pair	**Mean Difference**	**SD Difference**	**T-Statistics**	**p-value**
Video 1 vs Video 2	0.17	-0.43	0.34	0.74
Video 1 vs Video 3	-0.18	-0.18	-0.45	0.66
Video 1 vs Video 4	-0.33	-0.21	-0.80	0.44
Video 2 vs Video 3	-0.36	0.24	-0.63	0.54
Video 2 vs Video 4	-0.51	0.21	-0.89	0.39
Video 3 vs Video 4	-0.15	-0.03	-0.32	0.75

50.94 % of the participants have a Master's degree as an educational background. The participants are within the age group of 20 to 40 (avg. 26, st.dev. 4.98).

The HRIVST scores were computed for each video in all five conditions, as outlined in Table 2. Notably, the "Focused Explanation" condition yielded the highest HRIVST scores, indicating a higher level of legibility and understanding compared to the other conditions. However, the results are not statistically significant due to the limited number of participants. Hence, these findings provide only initial insights into the potential impact of different explanation strategies on transparency within HRI. Video 1 has a higher HRIVST score in "Comprehensive Explanation", Video 2 has a higher in "Alerted Comprehensive Explanation", Video 3, and Video 4 have higher HRIVST scores in the "Focused Explanation" condition.

By aggregating the evaluation of the different conditions for each video (see column Mean (SD) of Table 2), we can observe that Video 2 obtains the lower score using the HRIVST metrics and is evaluated as the less legible. The robot's actions in Video 2 are interrupted by an unexpected error, and it does not complete its task. This caused a certain uncertainty in understanding the final goal and made it less legible compared to other videos. The analysis suggests that while “Focused Explanations” are generally favorable, in the case of less legible behaviors, “Comprehensive Explanations” provided only at specific times could help transparency.

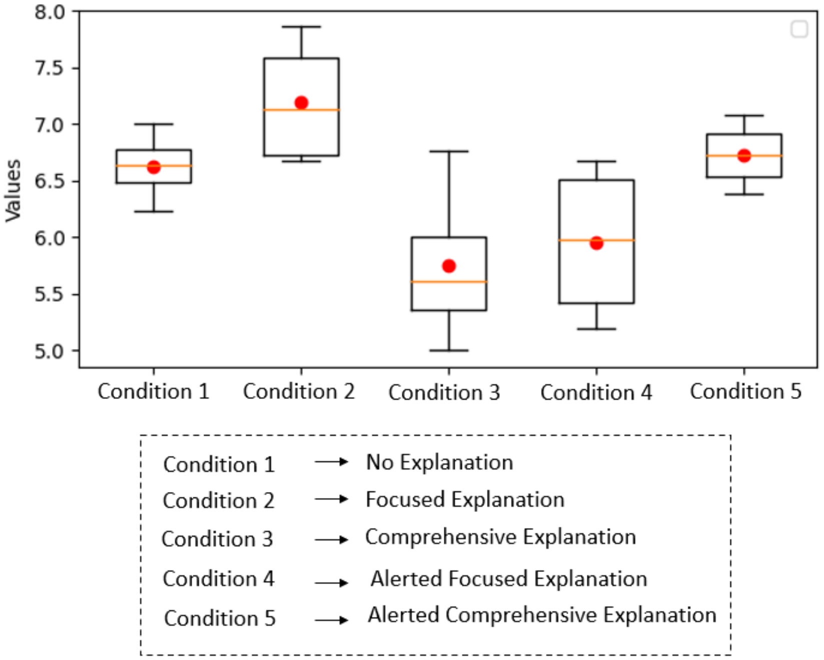


Fig. 3: Comparison of Different Explanation Types

Figure 3 shows the aggregated averages for each condition. While we have a higher average for the Focused Explanations condition, all the p-values in our analysis are above the significance level ($\alpha = 0.05$), which means we do not have statistically significant evidence to reject the null hypothesis for any of the comparisons we performed.

Based on our results, providing a “Focused Explanation” while performing actions and a “Comprehensive Explanation” as the alert explanation can potentially improve transparency in HRI.

6 Conclusions and Future Work

Establishing transparent communication channels is crucial to successful interactions in a world where human-robot collaboration is gaining momentum. Our exploration into the impact of explanations on transparency within HRI sheds light on the significance of effective verbal communication. By delving into the nuances of explanation types and timings, we have gained valuable insights that contribute to the overarching goal of seamless collaboration between humans and robots. The findings from our study, supported by the HRIVST metric, highlight the influence of "Focused Explanations" on transparency and are in line with [12]. The increase in transparency scores when providing short explanations underscores the power of clear, concise communication. Additionally, we observed that "Comprehensive Explanations" accompanying changes in the robot's plan could contribute significantly in less legible cases.

As a future work, expanding the participant pool would provide a more comprehensive perspective on the effectiveness of different explanation strategies. Moreover, a qualitative analysis of participant feedback could provide deeper insights into the subjective experiences and preferences surrounding explanation-driven transparency.

Investigating the interplay between explanations and other cues, such as non-verbal gestures and visual displays, could unveil synergies that amplify the transparency achieved in HRI. Additionally, the influence of contextual factors, such as task complexity and familiarity, warrants exploration, as these aspects could impact the relevance and reception of different explanation strategies.

By continuing to refine and expand our investigation, we aim to contribute to the ever-evolving understanding of how explanations can enrich human-robot interactions and pave the way for a future of harmonious collaboration.

Acknoledgment

This work has been supported by the Italian MUR and EU under the CHIST-ERA IV project COHERENT (PCI2020-120718-2) and Italian PON R&I 2014-2020 - REACT-EU (CUP E65F21002920003).

References

1. S. Rossi, G. Ercolano, L. Raggioli, E. Savino, and M. Ruocco, "The disappearing robot: An analysis of disengagement and distraction during non-interactive tasks," in *27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2018, pp. 522–527.
2. L. Raggioli, F. A. D'Asaro, and S. Rossi, "Deep reinforcement learning for robotic approaching behavior influenced by user activity and disengagement," *International Journal of Social Robotics*, 2023.
3. S. Sagheb, S. Gandhi, and D. P. Losey, "Should collaborative robots be transparent?" 2023. [Online]. Available: https://arxiv.org/abs/2304.11753
4. A. Hamacher, N. Bianchi-Berthouze, A. G. Pipe, and K. Eder, "Believing in bert: Using expressive communication to enhance trust and counteract operational error in physical human-robot interaction," 2016. [Online]. Available: https://arxiv.org/abs/1605.08817
5. G. Angelopoulos, A. Rossi, C. D. Napoli, and S. Rossi, "You are in my way: Non-verbal social cues for legible robot navigation behaviors," in *2022 IEEE/RSJ IROS*, 2022, pp. 657–662.

6. M. Matarese, A. Sciutti, F. Rea, and S. Rossi, "Toward robots' behavioral transparency of temporal difference reinforcement learning with a human teacher," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 6, pp. 578–589, 2021.
7. B. Hayes and J. A. Shah, "Improving robot controller transparency through autonomous policy explanation," in *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 2017, pp. 303–312.
8. S. Rossi, A. Rossi, and K. Dautenhahn, "The secret life of robots: Perspectives and challenges for robot's behaviours during non-interactive tasks," *International Journal of Social Robotics*, vol. 12, no. 6, pp. 1265–1278, 2020.
9. S. Rossi, A. Coppola, M. Gaita, and A. Rossi, "Human-robot interaction video sequencing task for robot's behaviour legibility," 7 2023. [Online]. Available: https://www.techrxiv.org/articles/preprint/Human-Robot_Interaction_Video_Sequencing_Task_for_Robot_s_Behaviour_Legibility/23696706
10. J. Patel, T. Ramaswamy, Z. Li, and C. Pinciroli, "Transparency in multi-human multi-robot interaction," 2021. [Online]. Available: https://arxiv.org/abs/2101.10495
11. J. Ambsdorf, A. Munir, Y. Wei, K. Degkwitz, H. M. Harms, S. Stannek, K. Ahrens, D. Becker, E. Strahl, T. Weber, and S. Wermter, "Explain yourself! effects of explanations in human-robot interaction," 2022. [Online]. Available: https://arxiv.org/abs/2204.04501
12. Z. Han, E. Phillips, and H. A. Yanco, "The need for verbal robot explanations and how people would like a robot to explain itself," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 4, pp. 1–42, 2021.
13. Z. Taschdjian, "Why did the robot cross the road? a user study of explanation in human-robot interaction," 2020. [Online]. Available: https://arxiv.org/abs/2012.00078
14. T. Sakai, K. Miyazawa, T. Horii, and T. Nagai, "A framework of explanation generation toward reliable autonomous robots," 2021. [Online]. Available: https://arxiv.org/abs/2105.02670
15. Y. Zhang and M. Zakershahrak, "Progressive explanation generation for human-robot teaming," 2019. [Online]. Available: https://arxiv.org/abs/1902.00604
16. F. Cruz, C. Young, R. Dazeley, and P. Vamplew, "Evaluating human-like explanations for robot actions in reinforcement learning scenarios," 2022. [Online]. Available: https://arxiv.org/abs/2207.03214
17. L. Zhu and T. Williams, "Effects of proactive explanations by robots on human-robot trust," in *Social Robotics: 12th International Conference, ICSR 2020*. Springer, 2020, pp. 85–95.
18. M. Zakershahrak, Z. Gong, N. Sadassivam, and Y. Zhang, "Online explanation generation for human-robot teaming," 2019. [Online]. Available: https://arxiv.org/abs/1903.06418
19. M. Zakershahrak, S. R. Marpally, A. Sharma, Z. Gong, and Y. Zhang, "Order matters: Generating progressive explanations for planning tasks in human-robot teaming," 2020. [Online]. Available: https://arxiv.org/abs/2004.07822