Effects of Incoherence in Multimodal Explanations of Robot Failures

Pradip Pramanick
pradip.pramanick@unina.it
University of Naples Federico II
Naples, Italy

Luca Raggioli
luca.raggioli@unina.it
University of Naples Federico II
Naples, Italy

Alessandra Rossi
alessandra.rossi@unina.it
University of Naples Federico II
Naples, Italy

Silvia Rossi
silvia.rossi@unina.it
University of Naples Federico II
Naples, Italy

Abstract

Providing explanations of a robot's behavior is a key enabler of trust in robots. Such explanations should be intuitive to people who are not experts in robotics. Prior research suggests that using multiple modalities to deliver explanations improves clarity. However, current methods for generating multimodal explanations neither assess nor ensure the coherence of the information across modalities. Here, we present an experiment to understand the effect of possible incoherence in multimodal explanations. We perform a user study asking participants to observe a series of robot failures and predict the reason for failure when provided with a controlled variation of multimodal explanations. Specifically, we present a methodology to compare incoherent and coherent explanations, aiming to understand their impact on perceiving robot failures.

CCS Concepts

• **Human-centered computing** → **User studies**; • **Computer systems organization** → External interfaces for robotics; • **Computing methodologies** → *Causal reasoning and diagnostics.*

Keywords

Multimodal Interaction, Multimodal Explanation, Explainable Artificial Intelligence, User Studies, Human-Robot Interaction

ACM Reference Format:
Pradip Pramanick, Luca Raggioli, Alessandra Rossi, and Silvia Rossi. 2024. Effects of Incoherence in Multimodal Explanations of Robot Failures. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI Companion '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3686215.3690155

1 Introduction

Explaining a robot's behavior is crucial for user acceptance and trust calibration [26]. These explanations must be intuitive for non-experts, especially when the robot behaves unexpectedly, i.e., during a failure. For non-experts, explanations can help understand what went wrong and form accurate expectations about a robot's abilities. However, this is challenging because it requires producing and presenting simple, yet sufficient reasoning for a sequence of decisions derived from increasingly complex systems. Thus, explaining using a single modality may lead to an information bottleneck [4].

Nonetheless, most known methods for automatic explanation generation of robot behavior use text as the sole modality. Further, text may not be suitable for conveying certain types of information. For example, Kwon et al. use motion as a modality to explain a robot's incapability of manipulation [11], which would be difficult to express using text. Angelopoulos et al. use gaze and gestures to explain a robot's directional intent [2]. Hence, using multiple modalities to explain different types of information appears beneficial. Recent studies on multimodal explanations also support this [1, 18, 29], although having very different scopes.

However, one important attribute of multimodal explanations has been largely overlooked. Current approaches to generating multimodal explanations neither evaluate nor ensure that the information provided across different modalities is coherent. Without such deliberations, incoherence can arise because most existing methods generate explanations independently across modalities [8, 13, 16]. For example, in [14] the black-box explainer of classifiers is decoupled from the dialogue policy. Similarly, in [8] and [16] the internal states of the robot are estimated and communicated independently. Furthermore, the automatic generation of explanation text is prone to hallucinations [4, 12], making simplistic combinations with other modalities potentially inconsistent. Further, in the case of approaches that jointly generate multimodal content, even though there is an attempt to maintain consistency of semantic content between the modalities, it is never guaranteed. Although less studied in robotics, this phenomenon is well-known in vision and language models [27, 28].

In this work, we study the effect of such incoherent explanations. More specifically, we propose an experiment to study two levels of incoherence - *Contradiction* and *Dissociation*. Contradiction occurs when an explanation in one modality contradicts others, while Dissociation refers to a lack of obvious semantic correlation between modalities. We provide examples in Figure 1b and a more formal description in Section 3.1.1. Our experiment aims to investigate how these incoherences affect people's ability to understand the robot's failures and their causes, as suggested in a recent work [17].

ICMI Companion '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0463-5/24/11
https://doi.org/10.1145/3686215.3690155

Even though our experiment is exploratory, we hypothesize that Dissociation will be perceived as incomplete and Contradiction as incorrect, compared to coherent explanations. We summarize our contributions in the following.

- We investigate how people are affected by incoherent multimodal explanations of robot failures. To our knowledge, this is the first experiment on this topic.
- Our initial findings indicate that incoherence may impact the correctness and sufficiency of multimodal explanations.

2 Related Works

Significant prior research on robot behavior explanations [3, 22, 24] falls into two categories based on the need for explanation. The first assumes explanations are needed when a robot's plan differs from a human's typical world model, making the plan seem sub-optimal. Existing approaches to this focus on counterfactual explanations, and are generally based on formal reasoning over task plans [19–21]. We focus on a second category which assumes explanations are needed when a robot fails to achieve its task goal. This necessitates reasoning on both the plan and the robot's observations. Failure explanations are crucial for non-experts because people perceive the need for explanations to be significantly higher in failure scenarios than when the robot's plan seems sub-optimal [23]. Failure to fulfill a task goal is usually caused by *flawed planning* from an incorrect world model [12], *limitations of sensing capabilities* [9, 12], and *physical constraints* [6, 9]. Prior works typically do not address all three causes together.

In contrast, we study all of these causes of failure in our experiment. In this regard, a previous experiment closest to ours is by Das et al. [4], which does consider the three causes. However, this study is based on unimodal explanations using text. Several other experiments study various aspects of failure explanations provided as natural language descriptions, either spoken or shown as text. Khanna et al. find a positive correlation of detailed explanations with increasing task complexity [10]. Melsion et al. compare the effect of explanation on trust in high and low-stake scenarios [15]. Hald et al. study post-explanation trust repair after mistakes made by a virtual robot. [5, 12] propose multimodal reasoning-based explanation generation approaches, but the produced explanations are communicated as text.

User studies on multimodal explanations are largely unexplored. Some relevant works include [16] which uses multiple modalities to provide transparency about a robot's scene and language understanding. Robb et al. find improvements in several metrics in a study with an interactive multimodal interface, compared to a non-interactive one [18]. In a non-robotic setting, Alipour et al. find that the helpfulness of multimodal explanations is correlated with the system's accuracy in a visual question-answering task [1]. However, no prior studies on multimodal explanations have been conducted in a robot-failure context, nor do they measure the effect of incoherence in such explanations.

3 Method

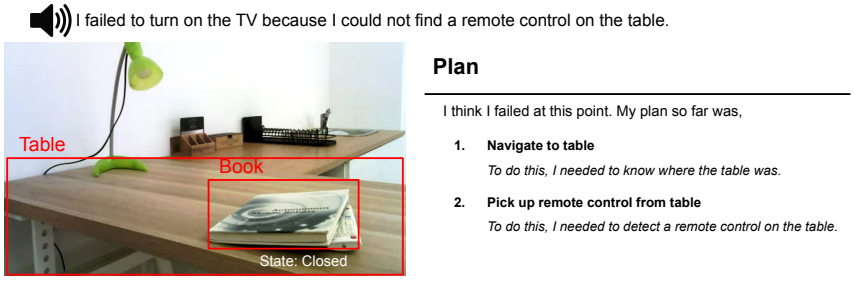
3.1 Design

Our choice of modalities for the explanation is partly motivated by prior approaches to automated robot failure explanations and user studies. We envision a scenario where a robot explains its past failures to a nearby human user. We consider natural language as a modality due to its extensive use for explaining failure [4, 12] and plan optimality [7, 23]. Further, we choose to communicate concise language-based explanations using speech to complement other visual elements. The second modality is a visualization of the robot's perception that includes names and state of the objects mentioned in the plan. The motivation is again, twofold. Firstly, as we consider sensing limitations as one of the reasons for failure, we posit that such limitations can be well-explained by overlaying the robot's egocentric observations on the image. This is also motivated by previous attempts at providing transparency of the robot's perception, such as [16] and [25]. Similarly, the third modality, which is shown beside the perception-visualization, is based on our consideration of planning failures and incorrect world models which may lead to unmet preconditions of the actions. This is somewhat similar to the explanation strategy chosen for plan suboptimality [20].

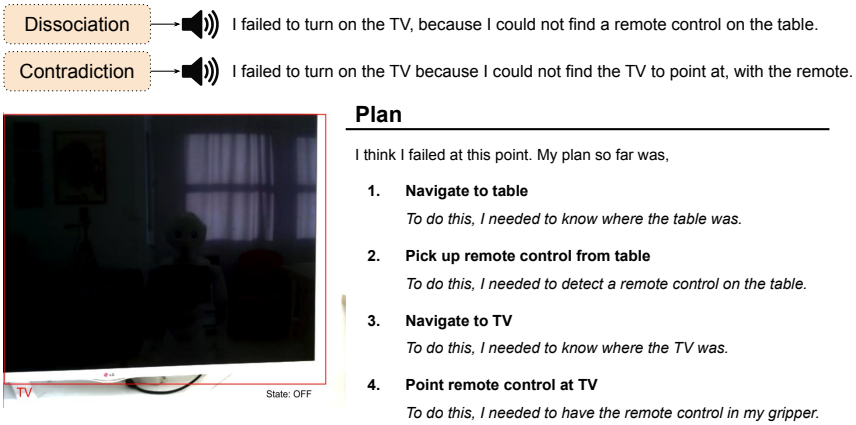
3.1.1 Coherence Conditions. In this study, we consider three coherence conditions. As introduced earlier, we use the two conditions - Contradiction (**C1**) and Dissociation (**C2**) as two levels of incoherent explanations. We also include Coherence (**C3**) as the third condition, which refers to a scenario with an obvious semantic correlation between all pairs of modalities and no Contradiction. Figure 1b shows examples of Contradiction and Dissociation for the task of turning on a TV. The corresponding Coherence condition is the same as the example in Figure 1a. For a more formal definition, let us consider that an explanation $\mathcal{E}$ is a conjunction of n propositions, i.e., $\mathcal{E} = \mathcal{P}_1 \wedge \mathcal{P}_2, \wedge \cdots \wedge \mathcal{P}_n$. In general, given any pair of explanations in two modalities from a set of total k modalities, denoted as $\{\mathcal{E}^{m_1}, \mathcal{E}^{m_2}\} \in \mathcal{E}^{m_1^k}$, we provide a formal definition of the conditions as the following.

$$C1 \equiv \exists \mathcal{P}_i \in \mathcal{E}^{m_1}, \exists \mathcal{P}_j \in \mathcal{E}^{m_2} : \mathcal{P}_i \perp \mathcal{P}_j.$$
$$C3 \equiv \exists \mathcal{P}_i \in \mathcal{E}^{m_1}, \exists \mathcal{P}_j \in \mathcal{E}^{m_2} : \mathcal{P}_i \models \mathcal{P}_j \wedge \neg(C1).$$
$$C2 \equiv \neg(C1 \wedge C3).$$

The constraints defined above can be recursively applied to all pairs of modalities to decide the level of coherence. $\mathcal{P}_i \perp \mathcal{P}_j$ denotes that the proposition $\mathcal{P}_i$ logically contradicts the proposition $\mathcal{P}_j$. As an example, consider the Contradiction example in Figure 1b. The proposition made by the robot about not finding the TV in the spoken modality contradicts the visualization of its perception - which shows a bounding box detected around a TV. $\mathcal{P}_i \models \mathcal{P}_j$ denotes that the proposition $\mathcal{P}_i$ logically entails the proposition $\mathcal{P}_j$. This constraint is necessary, but not sufficient for C3, which also requires an absence of contradiction between other pairs of propositions. Considering the example in Figure 1a, the absence of a detected remote control in the graphical modality entails the spoken proposition about not finding a remote on the table. This proposition also entails the precondition for picking up the remote control, as shown in the explanation of the plan. Further, there are no contradictory pairs of propositions. Similarly, $\neg(C1 \wedge C3)$ denotes an absence of both entailing and contradictory propositions. Thus, we posit that this denotes the absence of an obvious semantic correlation, i.e., C2. This happens when the primary cause of the failure and its detection by the robot do not coincide. In the example of Dissociation in Figure 1b, the spoken proposition of not finding



(a) On the top - a textual representation of the spoken explanation, left - a visualization of the robot's perception, right - a structured text that describes the robot's plan.



(b) Dissociation (C1) and Contradiction (C2): the content of the graphical and structured text modality (Plan) remains the same, while the spoken contents are different.

Figure 1: Examples of the multimodal explanations used in the experiment.

the remote and the corresponding precondition in the plan cannot be verified in the visualization of perception. However, there are no contradictory propositions either. Thus, we manually develop the explanation variants to prevent confounding from automatic generation during the experiment.

3.1.2 Tasks & Causes of Failure. We inject three failure types (see Section 2) in each of the robot's three tasks, set in a mock-up living area with slight object re-arrangement per task. Figure 2 shows cropped snapshots of the robot's working area for the three tasks detailed below.

(1) Turning off a light - The robot aims to turn off a ceiling light using a switch which is blocked by a chair. This represents *physical constraints*.
(2) Turning on a TV - The robot aims to turn on a TV using a remote on a table, partially occluded by a book. It cannot detect the remote, representing *limited sensing capabilities*.
(3) Setting dinner table - The robot aims to pick up a plate from a rack (based on an incorrect world model), but it is in the sink instead, leading to *flawed planning*.

3.2 Procedure

The experiment was approved by the ethical committee of the University of Napoli Federico II. We assign a random triplet of

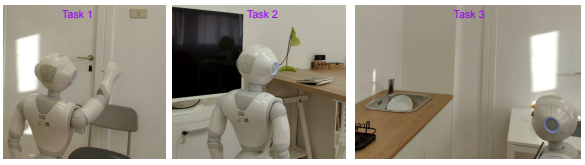


Figure 2: A Pepper robot attempting the three tasks.

conditions × *task types* such that each participant receives the three distinct conditions for the three different tasks. The participants start with a brief introduction to the procedure, followed by a demographic questionnaire. Then the participants are asked to perform the following three times for the three task types.

(1) The participants watch a video that displays two synchronized views: one from the robot's camera and another from a static camera behind the robot, capturing the entire workspace where the robot attempts tasks.
(2) The participants answer an initial set of questions related to the measures described in Section 3.3.
(3) The participants watch another video based on the assigned condition, which we mention to be produced by the robot. Essentially, the video plays a recorded audio of the explanation, while showing the perception-visualization and the

structured description of the plan, as the example shown in Figures 1a and 1b.

(4) The participants answer another set of questions on the completeness and correctness of the explanations, as detailed in Section 3.3. They are then asked to rank the three modalities in order of their helpfulness.

Finally, the participants complete an attention-check questionnaire.

3.3 Measure

In the following, we describe the primary questionnaire. For all the questions (Q1-Q5), we asked the participants to answer from (Yes/No/Maybe) and provide justifications for their answers in free-form text fields. Firstly, after watching the robot attempt to perform the task, and before receiving any explanation, we asked the participant for an evaluation of the robot's behavior as follows:

- (Q1) Do you think the robot completed the task?
- (Q2) Do you think that something went wrong while the robot performed the task?

After watching the explanations, the participants answer the following questions:

- (Q3) Do you think something went wrong while the robot performed the task?
- (Q4) Do you believe that the information provided by the robot was sufficient to understand what happened?
- (Q5) Do you believe that the information was a correct description of what happened?
- Which type of information was most helpful in understanding what caused the robot to fail the given task (most helpful to least helpful):
  - The text description of the robot's actions in its plan
  - The spoken information
  - The image shown

3.4 Participants

The attention checks excluded 28 participants, resulting in 74 valid participants (29 Female, 44 Male, 1 Other), aged 19-49 years (M = 27.97, SD = 4.8). Most were Italian (61%), followed by Indian (8%), German (4%), and others including French, Turkish, Romanian (8% combined), and 13 other nationalities, one from each. One participant preferred not to say. When asked about their prior experience with robots, 44.6% of participants stated no prior exposure. From the rest, 40% interacted with a robot before, 36% seen robots on social media, 27% were involved in a study with robots, and 24% were roboticists. As the participants were randomly assigned to conditions × task scenario, we obtained 74 responses for "turning off the light" scenario; 65 responses for the "switching the TV on" scenario; and 74 responses for the "setting dinner table" scenario.

4 Results

Firstly, we aim to understand if the participants recognized the failures without explanations. We observe that most participants (Q1 - 87% and Q2 - 52%, respectively) agreed that the robot failed to complete its task and did something wrong. Next, to understand the effect of the incoherent explanations in the different experimental conditions, we conduct a series of chi-square tests of independence between the responses to questions Q3-Q5 (after manipulation) with the conditions (C1-C3). Table 1 summarizes the results. We found no statistically significant association between people recognizing that something went wrong during the tasks with the different explanation variants (Q3, p>.05). However, we observed statistically significant associations between the conditions with the participant's belief in the explanation being sufficient (Q4, Cramér's V = 0.28) and the perceived correctness of the explanations (Q5, Cramér's V=0.27). As in this report, our primary objective is to highlight the perceived correctness of incoherent explanations, we further analyze the responses from Q5 to understand the contribution of each condition to the χ^2 statistic, as shown in Table 2. The results indicate that C1 (contradiction) shows a relatively stronger contribution to the correlation, compared to C2 and C3.

Question	χ^2 Value & Significance ($p < .05$)
Q3	$\chi^2(4) = 4.62, p = .329$
Q4	$\chi^2(4) = 34.14, p < 0.001*$
Q5	$\chi^2(4) = 30.75, p < 0.001*$

Table 1: χ^2 results for Q1, Q2, and Q3 responses.

Table 2: Adjusted standardized residuals of the Crosstabulation between Q5 responses and the conditions.

Condition	No	Maybe	Yes
C1	1.88	0.00	-0.51
C2	-1.20	-0.58	0.44
C3	-0.80	0.51	0.12

5 Conclusion

In this report, we present a novel experiment to understand how incoherence in multimodal explanations of robot failures affects the quality of the explanations along several attributes, and a preliminary analysis of the results. We introduce two types of incoherence in multimodal explanations - dissociation and contradiction. Our initial findings indicate that the coherence levels of multimodal explanations are associated with the participants' belief about the explanation's correctness and sufficiency. In the future, we aim to report a detailed analysis of the experiment. Particularly, we will compare the participants' responses before and after receiving explanations, categorize and compare justification texts with the answers, conduct post hoc analysis, and compare the responses between the different failure types. Finally, we aim to find which modality helped the most in comprehending the failures.

Acknowledgments

This work is supported by the European Union's Horizon Europe program under the TRAIL project, Marie Skłodowska-Curie grant agreement No 101072488, and the Italian Ministry for Universities and Research (MUR) with the PNRR Project FAIR PE0000013.

References

[1] Kamran Alipour, Jurgen P Schulze, Yi Yao, Avi Ziskind, and Giedrius Burachas. 2020. A study on multimodal and interactive explanations for visual question

answering. *arXiv preprint arXiv:2003.00431* (2020).
[2] Georgios Angelopoulos, Alessandra Rossi, Claudia Di Napoli, and Silvia Rossi. 2022. You are in my way: non-verbal social cues for legible robot navigation behaviors. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 657–662.
[3] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 1078–1088.
[4] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*. 351–360.
[5] Devleena Das and Sonia Chernova. 2021. Semantic-based explainable ai: Leveraging semantic scene graphs and pairwise ranking to explain robot failures. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3034–3041.
[6] Maximilian Diehl and Karinne Ramirez-Amaro. 2022. Why did i fail? a causal-based method to find explanations for robot failures. *IEEE Robotics and Automation Letters* 7, 4 (2022), 8925–8932.
[7] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th international conference on intelligent user interfaces*. 263–274.
[8] Helen Hastie, Francisco Javier Chiyah Garcia, David A Robb, Pedro Patron, and Atanas Laskov. 2017. MIRIAM: a multimodal chat-based interface for autonomous systems. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 495–496.
[9] Arda Inceoglu, Eren Erdal Aksoy, and Sanem Sariel. 2024. Multimodal Detection and Classification of Robot Manipulation Failures. *IEEE Robotics and Automation Letters* 9, 2 (2024), 1396–1403.
[10] Parag Khanna, Elmira Yadollahi, Mårten Björkman, Iolanda Leite, and Christian Smith. 2023. Effects of Explanation Strategies to Resolve Failures in Human-Robot Collaboration. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1829–1836.
[11] Minae Kwon, Sandy H Huang, and Anca D Dragan. 2018. Expressing robot incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 87–95.
[12] Zeyi Liu, Arpit Bahety, and Shuran Song. 2023. REFLECT: Summarizing Robot Experiences for Failure Explanation and Correction. In *Conference on Robot Learning*. PMLR, 3468–3484.
[13] Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2023. ConvXAI: a system for multimodal interaction with any black-box explainer. *Cognitive Computation* 15, 2 (2023), 613–644.
[14] Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2023. ConvXAI: a System for Multimodal Interaction with Any Black-box Explainer. *Cognitive Computation* 15, 2 (March 2023), 613–644. https://doi.org/10.1007/s12559-022-10067-7
[15] Gaspar Isaac Melsion, Rebecca Stower, Katie Winkle, and Iolanda Leite. 2023. What's at Stake? Robot explanations matter for high but not low-stake scenarios. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2421–2426.
[16] Leah Perlmutter, Eric Kernfeld, and Maya Cakmak. 2016. Situated Language Understanding with Human-like and Visualization-Based Transparency.. In *Robotics: Science and Systems*, Vol. 12. 40–50.
[17] Pradip Pramanick and Silvia Rossi. 2024. Multimodal Coherent Explanation Generation of Robot Failures. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE.
[18] David A Robb, Francisco J Chiyah Garcia, Atanas Laskov, Xingkun Liu, Pedro Patron, and Helen Hastie. 2018. Keep me in the loop: Increasing operator situation awareness through a conversational multimodal interface. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 384–392.
[19] Maayan Shvo, Toryn Q Klassen, and Sheila A McIlraith. 2022. Resolving misconceptions about the plans of agents via theory of mind. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 32. 719–729.
[20] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. 2021. Foundations of explanations as model reconciliation. *Artificial Intelligence* 301 (2021), 103558.
[21] Stylianos Loukas Vasileiou, Ashwin Kumar, and William Yeoh. 2023. DR-HAI: Argumentation-based Dialectical Reconciliation in Human-AI Interactions. In *ICAPS 2023 Workshop on Human-Aware Explainable Planning*.
[22] Lennart Wachowiak, Oya Celiktutan, Andrew Coles, and Gerard Canal. 2023. A Survey of Evaluation Methods and Metrics for Explanations in Human–Robot Interaction (HRI). In *ICRA2023 Workshop on Explainable Robotics*.
[23] Lennart Wachowiak, Andrew Fenn, Haris Kamran, Andrew Coles, Oya Celiktutan, and Gerard Canal. 2024. When Do People Want an Explanation from a Robot?. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 752–761.
[24] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2021. Explainable Embodied Agents Through Social Cues: A Review. *ACM Transactions on Human-Robot Interaction* 10, 3 (Sept. 2021), 1–24.
[25] Chao Wang and Anna Belardinelli. 2022. Investigating explainable human-robot interaction with augmented reality. In *5th International Workshop on Virtual, Augmented, and Mixed Reality for HRI*.
[26] Ning Wang, David V Pynadath, and Susan G Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 109–116.
[27] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, et al. 2023. Llava-grounding: Grounded visual chat with large multimodal models. *arXiv preprint arXiv:2312.02949* (2023).
[28] Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. 2021. Consensus graph representation learning for better grounded image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 3394–3402.
[29] Hongbo Zhu, Chuang Yu, and Angelo Cangelosi. 2022. Affective Human-Robot Interaction with Multimodal Explanations. In *International Conference on Social Robotics*. Springer, 241–252.