Towards Embodied Agent Intent Explanation in Human-Robot Collaboration: ACT Error Analysis and Solution Conceptualization

Amanuel Ergogo[1] and Zhao Han[2]

Abstract— **Collaborative robots must not only perform competently but also communicate their intents transparently to ensure safety and efficiency in shared task environments. However, state-of-the-art robot policies such as Action Chunking Transformer (ACT) models are opaque, which may make it difficult for human partners to interpret or predict their actions and intent to facilitate task coordination.**

To confirm this, we conducted a two-condition comparative study in a collaborative medication-dispensing scenario, showing inaccurate estimation of an ACT robot's intent led to miscoordination, duplicate medicine retrievals, and safety risks such as simultaneous access to shared shelf space. Specifically, we trained an ACT agent on human-human demonstration data and tested it in a human-agent condition. Compared to the human-human baseline, the opaque agent had a 36% drop in task success from 97% to 62%, a 17-fold increase in safety incidents from 2% to 34%, i.e., simultaneous access to shared shelf space and incorrect medication delivery, as well as a 44% increase in task completion (18s to 26s). This evidenced critical coordination breakdowns due to the lack of transparent intent.

In this work in progress, we thus conceptualize model-agnostic CRIE (Contextual Robot Intent Explanation) that predicts robot intention and explains in natural language without modifying the underlying policy itself. By analyzing multimodal contextual features—such as task phase, spatial configuration, and action trajectories—CRIE aims for real-time transparency about the robot's future actions. Our results will demonstrate how contextual, policy-agnostic intent explanations help close the gap between high-performing but opaque policies and transparent, human-compatible robot teamwork.

I. Introduction

Recent advances in robot learning, e.g., large-scale foundational models [1], [2], have empowered robots to acquire complex skills by imitating human demonstrations [3], [4]. Reducing reliance on manual programming, these learned policies show strong potential for deployment in dynamic, human-centered settings, including collaborative healthcare [5], assisted living [6], and warehouse operations [7], where adaptability and task generalization are important. In these human environments, however, autonomous robots must also make their *intent* transparent to human collaborators who can understand what the robot is currently doing, why it is acting in a particular way, and what it intends to do next. This ability to generate and convey interpretable intent explanations contributes to safety and coordination in human-robot collaboration [8].

Yet, generating such explanations remains a central challenge. Most learning systems model behavior as a Markov Decision Process (MDP), where each action is conditioned not only on the current state but also on future, uncertain observations [9]. This formulation introduces strong state-action coupling, complicating long-term intent prediction due to sequential dependencies between subsequent states and actions. Furthermore, deep neural policies encode decision logic in high-dimensional latent representations, making the robot reasoning process inaccessible and causing them to remain black-boxed about their intentions [10].

One line of research in explainable robotics has attempted to address this gap by specifying robot tasks with inherently interpretable structures. For instance, the recent development of behavior trees (BTs) [11] offers inherent interpretability through their modular, hierarchical design, allowing for subgoal tracing and structured explanation [12]–[14]. However, such symbolic methods require hand-crafted behavior logic, which is infeasible for black-box policies.

On the other hand, some learned policy architectures have begun to integrate a built-in structure that offers limited interpretability. For example, Action Chunking Transformers (ACT) [15] introduce a form of built-in structure by segmenting continuous behaviors into discrete action chunks, which may support short-horizon predictions of the robot's next actions. However, this structure is restricted to explaining ACT outputs and cannot generalize to other policy architectures. Furthermore, commonly seen in black-boxed policies, they also lack symbolic subtask labels and natural language interpretations, both of which are essential for communicating robot intent. Moreover, relying solely on action sequences without incorporating contextual environmental information can lead to inaccurate or misleading intent estimates since the same action can have different meanings depending on the environmental context.

In response to the opaqueness and model-specificity, researchers have begun exploring explainable AI (XAI) techniques within robotics, aiming to improve transparency and interpretability in data-driven systems. Examples include visual saliency maps [16], symbolic policy distillation [17], and causal reasoning frameworks [18]. While promising, these approaches are often post hoc, static, or assume access to model internals—making them unsuitable for real-time, model-agnostic intent explanation in collaborative scenarios.

In this preliminary work, we seek to answer the following research question: *How can we enable robots to generate interpretable, real-time intent explanations—without modifying their underlying opaque policies?* First, we conducted an empirical evaluation showing that an opaque ACT policy we trained on human-human demonstrations results in reduced task success, elevated safety incidents, and decreased fluency

All authors are with RARE Lab, Bellini College of Artificial Intelligence, Cybersecurity and Computing, University of South Florida, Tampa, FL, USA. aergogo@usf.edu[1] zhaohan@usf.edu[2]

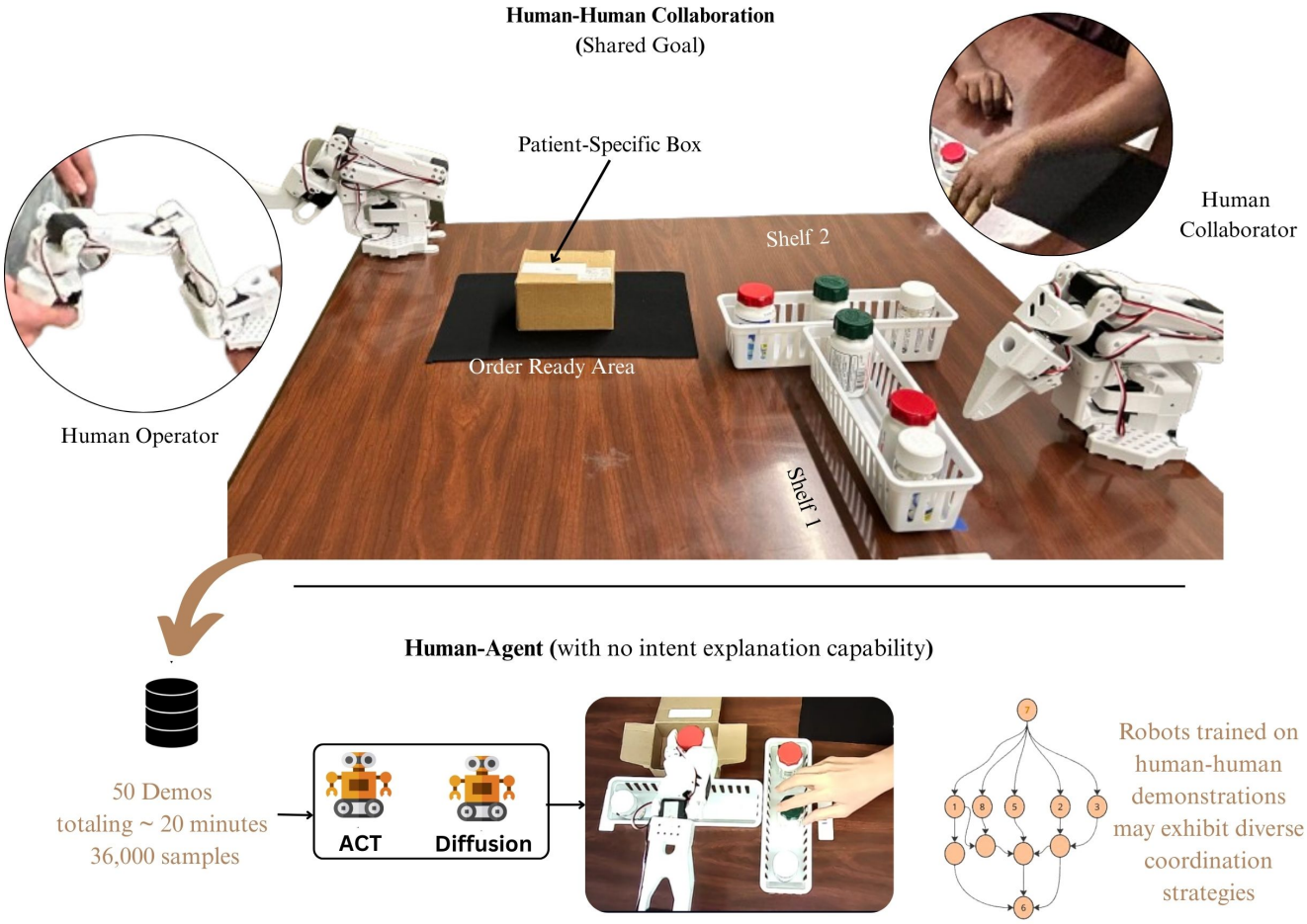

Fig. 1: Overview of Human-Human (top) and Human-Agent (bottom) collaboration in a medication-dispensing task. Data from 50 Human-Human demonstrations, featuring diverse coordination strategies, were used to train ACT and Diffusion policies. In the Human-Agent scenario, the robot executes tasks without intent explanation while collaborating with a human partner.

in human-agent collaboration. Secondly, we conceptualize CRIE (Contextual Robot Intent Explanation), a novel framework that infers and verbalizes symbolic robot intent as subtasks from multimodal context—including environmental dynamics, policy outputs, and high-level task goals. CRIE will be *policy-agnostic* and operate in *real time alongside any learned robot policies*, including those trained on natural human- human demonstrations that capture diverse coordination strategies. Its architecture combines a Transformer-based encoder-decoder with a Conditional Variational Autoencoder (CVAE), enabling the generation of subtask-level intent predictions that can be conveyed to human teammates via natural language or display interfaces. CRIE fills a critical gap by offering real-time symbolic explanations without requiring access to internal policy structure, thereby supporting transparent collaboration in high-stakes environments.

II. Related Work

A. Symbolic Robot Behavior Explanation

Some efforts in explainable robotics primarily relied on symbolic controllers such as behavior trees (BTs), which offer *inherent interpretability* through explicit control logic [12], [19]. This data structure supports explanation through modular, hierarchical representations of subgoals and action sequences and has been demonstrated in both a kitting manipulation task and a taxi simulation [12]. Specifically, it identifies the overall goal from the root node, clarifies the current subgoal through parent-node backtracing, and uses depth-first search to project the steps required to complete a task. However, such methods require access to the robot's internal behavior definitions and are tightly coupled to the specific control structure, making them unsuitable for black-box policies learned from data. As a result, recent approaches have begun incorporating built-in intent generation within policy models.

B. Robot Learning and Explanable AI

Recent robotic systems increasingly employ Learning from Demonstration (LfD) and deep reinforcement learning, using Markov Decision Processes (MDPs) to model decisions [9]. While these approaches offer adaptability, the state-action coupling inherent in MDPs complicates prediction and explanation of future behavior [20]. Morden models like Action Chunking Transformers (ACT) attempt to address this by segmenting behavior into action chunks, enabling short-horizon action forecasting [21]. However, it lacks high-level symbolic labeling and natural language explanations, which are essential for intent communication. Furthermore, its interpretability is model-specific and does not generalize across different policy architectures.

To support broader generalization, researchers have explored model-agnostic explainability approaches, such as saliency maps [22], surrogate models like LIME and SHAP [23], and attention-based visualizations [24]. Causal reasoning and symbolic policy distillation have also been used to extract human-interpretable abstractions from learned policies [25]. While these methods offer promising directions, most provide post-hoc, static explanations and require access to internal model components—making them impractical for real-time, dynamic applications in robotics [26].

In our conceptualization, CRIE is designed to bridge the gap between model performance and human interpretability. Unlike methods that rely on internal access or handcrafted logic, CRIE will be policy-agnostic and operates in real time. It will generate high-level, symbolic intent explanations by observing low-level actions, task goals, and contextual dynamics such as spatial layout, object movements, and task progression. By leveraging multimodal input streams, including task specifications (for example, "Fulfill prescription order A and B"), visual observations of the shared workspace, and robot actions (that is, policy actions at time t such as joint positions or end-effector poses), along with structured temporal representations, CRIE will support transparent and proactive collaboration in environments where continuous communication of robot intent is critical.

III. Evaluating ACT in Collaborative Task

We first conducted a baseline experiment to perform an empirical analysis of the opaque ACT model in a collaborative medication-dispensing task under two conditions: Human-Human and Human-Agent (ACT), as shown in Figure 1. Each trial involved a shared goal: fulfilling an order for two medicines by coordinating labeling, retrieval, verification, and delivery to the Ready Order Area.

A. Task Setup

In the task, a team of two workers fulfill an order for two medicines (e.g., Medicine A and B). The shared workspace includes a shelf with bottles, a labeling station, and a designated "Ready Order Area." This safety-critical task requires dynamic coordination, which introduces variability in robot behavior and makes future actions hard to anticipate without explicit intent explanations.

B. Collaboration Strategies

In the Human-Human condition, two people jointly completed the medication-dispensing task, providing baseline demonstrations for training. A total of 50 demonstrations were recorded for training. We randomly selected 15 trials to compare with the human-agent condition. Observing these sessions revealed three distinct collaboration strategies: (1) *Role-based*: The collaborator labeled the patient-specific box while the operator retrieved medicine A, followed by the collaborator retrieving medicine B, verifying the order, and completing the delivery; (2) *Concurrent*: Both participants retrieved one medicine each in parallel, enabling concurrent task execution; (3) *Delegated:* The operator retrieved both medicines, while the collaborator handled labeling and final delivery, forming a fully sequential workflow. These diverse interaction styles were used to train the ACT model to learn a range of natural human coordination strategies.

In the Human-Agent condition, the ACT-trained robot autonomously executed subtask sequences to achieve the shared goal. No explicit intent explanation was provided to the human partner, allowing us to evaluate the effectiveness and limitations of the agent in real-time collaboration.

C. Measures

To assess collaborative performance, we measured three primary metrics. The *Task Success Rate* represented the percentage of trials in which both medicines were successfully retrieved, labeled, verified, and transferred to the Ready Order Area. The *Safety Incident Rate* quantified the proportion of subtask transitions or workspace actions that resulted in unsafe overlaps or conflicts—such as simultaneous access to the same shelf space or patient-specific box. Finally, the *Task Completion Time* measured the total duration from labeling the patient-specific box to retrieving both medicines, packing them, and completing delivery to the Ready Order Area.

D. Quantitative Performance Results

The results from 15 matched trials are shown in Figure 3. The Human-Human condition achieved a task success rate of 97%, averaged 2 safety incidents per trial, and completed tasks in 18 seconds on average. In contrast, the Human-Agent (ACT) condition yielded a success rate of 62%, averaged 34 safety incidents per trial, and had a longer average task completion time of 26 seconds. These findings demonstrated the ACT-based agent's limitations in coordination and transparency compared to human collaboration.

E. Failure Analysis

In addition to model-level errors (e.g., missed pickups, misgrasped items), we observed several recurring failure modes in the Human-Agent condition:

- **Redundant Retrievals**: Due to the robot's opaque intent, human partners often duplicated retrieval actions, unaware that the robot was pursuing the same subtask.
- **Timing Delays**: Humans frequently paused, unsure of whether the robot intended to retrieve medicine A or

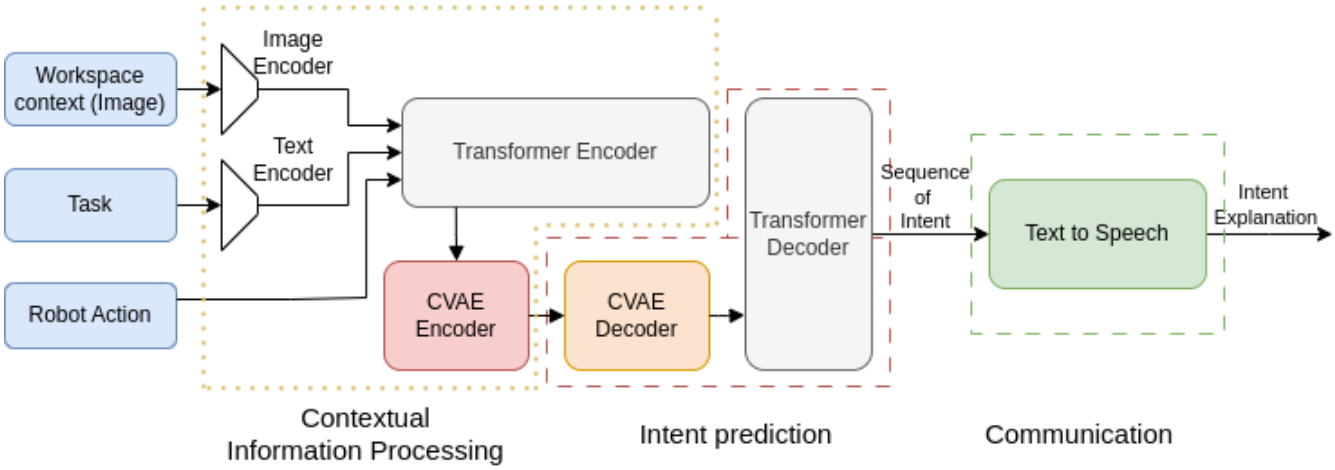

Fig. 2: Overview of the CRIE Transformer with a Conditional Variational Autoencoder (CVAE) for robot sequence of intent prediction. Workspace dynamics captured by the camera, robot action, and task goal are fused in the encoder to produce a latent distribution over subtask labels. The decoder then generates symbolic intent predictions.

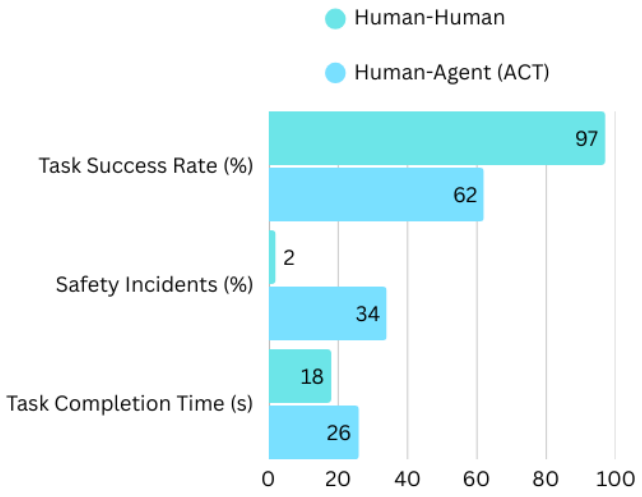

Fig. 3: Performance comparison between Human-Human and Human-Agent (ACT) teams in a collaborative medication-dispensing task. Metrics include task success rate (%), safety incidents rate(%), and task completion time (in seconds).

B, leading to missed coordination opportunities and timeouts.

- **Safety Conflicts**: Multiple trials showed both agents accessing the same shelf or patient-specific box simultaneously, increasing the risk of physical interference.

F. Insight

These results highlighted the limitations of opaque policies in collaborative human-robot tasks. Lack of transparency impaired coordination, increased error rates, and reduced efficiency. These findings motivated the conceptual development of CRIE, a symbolic intent explanations framework to improve task fluency, safety, and shared decision-making.

IV. Contextual Robot Intent Explanation (CRIE) Conceptualization

CRIE aims to address the challenge of enhancing transparency in robot actions, particularly in making a robot's next actions understandable to human collaborators. CRIE will enable real-time intent explanation by converting opaque, low-level control outputs into symbolic, high-level subtasks. It is policy-agnostic and integrates seamlessly with existing LfD models. As shown in Figure 2, the system will consist of three core components: the *contextual information processing module*, the *intent prediction network*, and the *human communication and integration interface*.

A. Contextual Information Processing

In addition to control outputs from opaque robot models (joint positions and end-effector poses) and high-level task goals (prescription objectives to fulfill medication orders for items A and B), CRIE will ingest contextual information about task progression and environmental changes due to robot action. While environmental context could be represented using knowledge-driven approaches such as ontologies or graphs, CRIE will use a learned latent representation of continuous RGB frames from the shared workspace, making it more flexible and less task-dependent. CRIE will capture environmental changes caused by the robot's actions, as well as the previous sequence of robot actions, to infer task progress. Inputs such as images, robot actions, and task goals will be synchronized and segmented into temporal windows to form a unified input representation. This representation will enable CRIE to more accurately interpret robot behavior by integrating temporal, spatial, and contextual dimensions of intent, enabling a more nuanced and accurate understanding of robotic actions.

B. Intent Prediction Network

To predict human-understandable subtask labels from numeric policy outputs, CRIE will employ a Transformer-based

encoder-decoder architecture augmented with a Conditional Variational Autoencoder (CVAE). The encoder will fuse the environmental state, control actions, and goal context into a latent embedding. During training, ground-truth subtask labels will guide the CVAE to learn a compact latent representation of task segmentation. This latent distribution helps to capture uncertainty in subtask boundaries or variations across demonstrations. The decoder will generate symbolic intent labels aligned with the robot's projected behavior. The training loss will combine cross-entropy for classification accuracy and KL divergence for latent space regularization:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \beta \cdot D_{KL}(q(z|x, y)||p(z)) \quad (1)$$

C. Communication and Integration Interface

CRIE is designed to operate passively alongside any LfD policy, reading policy outputs and environment observations to produce intent labels in real time. It will support various policies, such as Action Chunking Transformers (ACT) [15] and diffusion-based controllers [27]. After generating subtask predictions, they will be communicated to human collaborators using various modalities—such as textual displays, audio announcements, or other interfaces—enhancing human understanding and coordination. [14] showed that the effectiveness of communication modalities depends on task type. Their findings recommend verbal explanations for manipulation and object placement tasks. Accordingly, CRIE will use speech-based explanations in its primary application—medication dispensing—to enhance clarity, trust, and coordination. By matching modality to task semantics, CRIE will improve transparency and enable proactive human collaboration.

V. Conclusion and Ongoing Work

We conducted an empirical evaluation of a collaborative medication-dispensing task comparing Human-Human and Human-Agent teams, with the agent operating under an opaque ACT policy. Our results showed significant limitations in transparency, coordination, and safety when deploying black-box policies in real-time collaborative settings: task success rates dropped, safety violations increased, and task completion times were extended due to user hesitation and action overlap.

To address these issues, we conceptualized the CRIE framework—designed to enhance transparency by generating real-time intent explanations from observed robot actions and environment context. CRIE will be model-agnostic and will be easily integrated into existing learned policies without modifying the policy architecture itself.

In future work, we will integrate CRIE with the ACT policy to assess its impact on safety, efficiency, and coordination. We also plan to extend the evaluation to new task scenarios and robot policy models to better understand its generalizability. Additionally, we aim to incorporate diverse communication modalities to support a wider range of human-robot collaboration settings.

These evaluations will quantify CRIE's impact and further explore how transparency mechanisms can improve the reliability of learned robotic behavior in shared environments.

References

[1] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[2] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[3] A. Mandlekar, D. Xu, J. Wong *et al.*, "What matters in learning from offline human demonstrations for robot manipulation," *arXiv preprint arXiv:2108.03298*, 2021.

[4] D. Park, Y. Hoshi, and S. S. Srinivasa, "Toward active robot-assisted feeding with a general-purpose mobile manipulator: Design, evaluation, and lessons learned," *arXiv preprint arXiv:1909.09652*, 2019.

[5] W. Liu, D. Zhao, Y. Wang, and S. Song, "Advancing healthcare through intelligent human-robot collaboration: Overview, challenges, and opportunities," *arXiv preprint arXiv:2403.10835*, 2024.

[6] S. Liu, A. Lee, P. Mo *et al.*, "Robots in assisted living facilities: Scoping review," *Journal of Medical Internet Research*, vol. 25, p. e42753, 2023.

[7] U. Arora, R. Patel *et al.*, "Ai-driven warehouse automation: A comprehensive review of systems," *arXiv preprint arXiv:2403.15971*, 2024.

[8] M. Scheutz, B. F. Malle, G. Briggs, and E. Kraemer, "Transparency for robots and autonomous systems: Fundamentals, technologies and applications," *International Journal of Human-Computer Interaction*, vol. 38, no. 14, pp. 1329–1344, 2022.

[9] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[10] N. Jakobi, R. Feng, and H. Zhao, "Interpreting deep reinforcement learning policies in robotics: Challenges and future directions," *Engineering Applications of Artificial Intelligence*, 2024.

[11] M. Colledanchise and P. Ögren, *Behavior trees in robotics and AI: An introduction*. CRC Press, 2018.

[12] Z. Han, D. Giger, J. Allspaw, M. S. Lee, H. Admoni, and H. A. Yanco, "Building the foundation of robot explanation generation using behavior trees," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 3, pp. 1–31, 2021.

[13] G. LeMasurier, A. Gautam, Z. Han, J. W. Crandall, and H. A. Yanco, "Reactive or proactive? how robots should explain failures," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 413–422.

[14] Z. Han and H. Yanco, "Communicating missing causal information to explain a robot's past behavior," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 1, pp. 1–45, 2023.

[15] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," 2023.

[16] A. Singh and R. Kumar, "Explaining visual reasoning in robotic perception via saliency mapping," *arXiv preprint arXiv:2404.00682*, 2024.

[17] J. Lee, X. Gao, and Z. Han, "Distilling robot policies into symbolic explanations for human understanding," *arXiv preprint arXiv:2403.14328*, 2024.

[18] L. Rossi and Y. Yang, "Causal models for explainable robot behavior in human-robot collaboration," *Robotics and Autonomous Systems*, vol. 166, p. 104397, 2023.

[19] M. Colledanchise and P. Ögren, *Behavior trees in robotics and AI: An introduction*. CRC Press, 2018.

[20] Z. Erickson, V. Gangaram, A. Kapusta, and et al., "Assistive gym: A physics simulation framework for assistive robotics," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 169–10 176.

[21] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[22] R. R. Selvaraju, M. Cogswell, A. Das *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[23] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[24] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[25] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "Explainable reinforcement learning through a causal lens," in *AAAI*, 2020.

[26] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.

[27] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.