Why Did the Robot Fail? Causal Concept Bottlenecks for Failure Detection and Explanation

Mihir Parekh
Department of Computer Science
Indiana University Bloomington
Bloomington, IN, USA
mihparek@iu.edu

Michelle Zhao
Department of Computer Science
Indiana University Bloomington
Bloomington, IN, USA
zhaomich@iu.edu

***Abstract*—Robots that learn from human feedback can detect failures and request additional demonstrations, corrections, or assistance, but existing failure detectors provide little insight into *why* a failure was detected or *what* should change to achieve success. In this work, we investigate causal concept bottleneck models (C2BMs) for explainable robot failure detection, grounding failure predictions in semantically meaningful concepts and their causal relationships. We propose a pipeline that uses vision-language models to identify failure-relevant concepts from robot rollout videos and construct a causal graph relating these concepts to task outcomes. Using this structure, we train C2BMs to predict robot failures through causally structured concept representations. Unlike black-box failure detectors, this representation exposes the concepts underlying a prediction and enables interventions and counterfactual reasoning about how changes to those concepts affect predicted task outcomes. In this extended abstract, we demonstrate these capabilities through a proof-of-concept robotics task, showing how causal concept-based failure detection enables robots to communicate not only that a failure has occurred, but why it may have occurred and what could change to achieve success. We further outline a proposed user study evaluating the utility of these failure explanations for interactive robot learning with diverse human teachers.**

***Index Terms*—human-robot interaction, failure prediction, explainable AI, causal modeling**

I. Introduction

Human feedback underlies the training data from which robots are aligned to end-user desires, explicitly getting their feedback, whether it be preferences [3, 15, 33], demonstrations [27], corrections [2], language [38], or implicitly communicative behaviors [41, 42, 43]. Even after alignment, learned policies may still encounter failures in out-of-distribution environments [34] or unexpected outcomes of actions [31]. As a result, robots are supervised by operators, such as safety drivers who monitor autonomous vehicles [7, 29, 35] and everyday people who constantly check if LLM outputs are reliable enough to use in their emails [5, 6, 17].

Robots can learn new tasks from human feedback, but learned policies inevitably encounter failures during deployment. Failure detection methods can help robots recognize when they need additional demonstrations, corrections, or assistance from a human teacher. However, these methods typically communicate only *that* the robot has detected a failure or requires help, leaving the human to determine why the robot will fail and how their feedback should address it. This places a significant burden on human teachers, particularly when they are non-experts or when multiple users contribute feedback.

In this work, we investigate how robot failure detectors can instead communicate the hypothesized causes of failure and support reasoning about what would need to change for a rollout to succeed. Our key insight is to formulate robot failure detection using causal concept bottleneck models (C2BMs) [10], such that failure predictions are mediated by semantically meaningful concepts organized according to causal structure.

We first consider a human-defined DAG as a sanity check for the proposed framework. Given a causal graph relating semantically meaningful task concepts to failure, we demonstrate how causal concept-based failure detection can support interactive robot learning. Rather than communicating only that it predicts a failure, the robot can communicate why by identifying the concepts and causal pathways contributing most to its prediction.

We then remove the assumption that the DAG is known a priori. Starting only from robot rollout videos and success/failure labels, we introduce a pipeline that uses VLMs to identify semantically meaningful, failure-relevant concepts and causal discovery to infer relationships among them. We use the resulting graph to train a C2BM that predicts rollout success or failure while respecting the discovered causal structure. We propose that these robot hypotheses as to why the policy failed can serve multiple roles in human-robot interaction, and serve to assist the user in reasoning about how the environment or policy must change to achieve success. We demonstrate through a proof of concept the types of explanations and interventions this representation enables, and outline a user study design plan to evaluate their impact on human understanding and feedback.

II. Related Work

A. Failure Prediction in Robot Learning

Failure prediction is increasingly used to identify when learned robot policies are likely to fail and, consequently, when additional human assistance or corrective data may be needed [24]. Recent approaches for generative robot policies formulate failure detection as a statistical prediction problem over policy observations or outputs. [31] identify impending failures from consecutive out-of-distribution observations

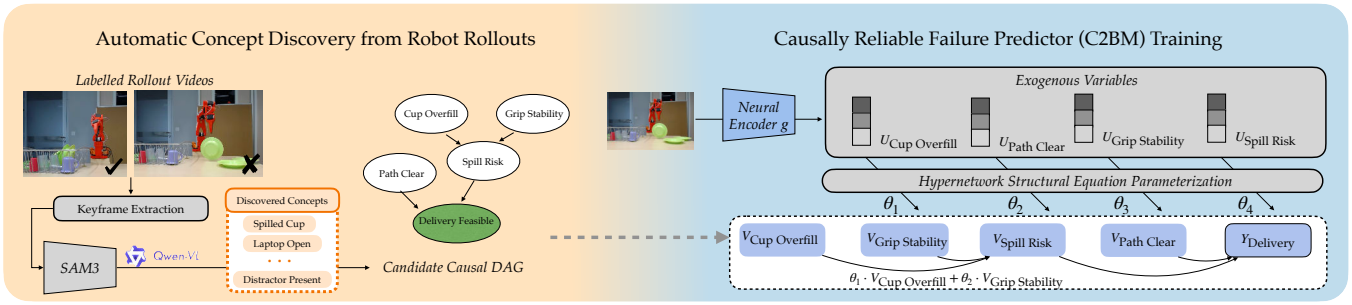


Fig. 1. We propose to extract from video rollouts the candidate causal DAG, and train our failure detector as a C2BM [10], imposing causal structure over human-interpretable concepts to enable interpretable failure predictions and causal reasoning.

and persistently high uncertainty in policy predictions, using conformal prediction [36] to calibrate failure thresholds. Similarly, Xu et al. [37] detect failures without requiring failure examples by scoring visual embeddings relative to successful demonstration data and learning time-varying thresholds using functional conformal prediction [12]. These approaches have the advantage of detecting failures without requiring an explicit specification of how a task can fail. However, their failure scores provide limited semantic information about the factors contributing to a prediction. When a detector signals that a rollout is likely to fail, the interpretation of why the failure occurred and what corrective feedback must be provided relies largely on the human teacher.

More recently, vision-language models (VLMs) have enabled failure detection at the semantic level. Prior work uses VLMs to produce binary success or failure judgments [13, 25, 39, 44] as well as textual descriptions or explanations of failures [8, 14]. Zhou et al. [45] formulate failure detection as spatio-temporal constraint satisfaction and use a VLM to determine from video whether task constraints have been violated. Such approaches provide semantically meaningful descriptions of task failures, but may rely on predefined task specifications, questions, or constraints that encode how successful execution should proceed. Our work bridges semantic and statistical failure detection by representing failure through *causally structured, semantic concepts* grounded in rollout videos extracted by the VLM's physical-world knowledge.

B. Explaining Robot Failures

Beyond detecting that a robot has failed, explanations can help users understand *why* a failure occurred and determine how to respond [20]. Prior work has therefore investigated generating semantic descriptions of robot behavior and failures. REFLECT [26], for example, uses a large language model to reason over hierarchical summaries of multisensory robot observations and generate failure explanations. Das et al. [9] instead learn an encoder–decoder model to autonomously generate natural-language explanations of robot behavior. While such approaches make failures more accessible to users, generated language provides an explanation of a particular execution rather than an explicit, inspectable model of the factors underlying the robot's failure prediction.

C. Causal Modeling

Causal models provide a structured representation of how variables relate to and influence one another, enabling forms of reasoning that extend beyond statistical prediction. Most closely related to our work, Diehl et al. [11] explicitly model causal relationships underlying robot failures. They learn a causal Bayesian network relating task executions to their consequences, and search the resulting causal model for nearby successful states to identify and avoid potential failures. This work demonstrates the utility of causal structure for reasoning beyond binary failure prediction. Our work builds on this motivation but considers causal failure models as an interface for *human-interactive robot learning*.

Concept bottleneck models (CBMs) [21, 22, 23] introduce human-interpretable concepts as an intermediate representation between model inputs and predictions. Given an input x, a CBM first predicts a set of semantic concepts $\mathbf{c}$ and then predicts an outcome y from this concept representation. For robot failure detection, we show that these concepts can correspond to visually identifiable properties of an execution, such as whether an object was securely grasped, whether it is near the edge of a surface, or whether an obstacle obstructs the robot's motion. An important property of CBMs is that their intermediate representations can be *intervened upon*: a user can replace a predicted concept value with a specified value and propagate the modified representation through the downstream predictor [21]. Such interventions provide a mechanism for probing a model's decision: given a failed manipulation rollout, a human could modify the concept *object near table edge* via physically changing the environment and observe how the model's predicted task outcome changes. However, interventions on a standard CBM do not necessarily correspond to *causal* interventions. Concepts may be statistically dependent, and independently modifying one concept while holding all others fixed can produce concept combinations that are inconsistent with the underlying data-generating process. Causally reliable concept bottleneck models (C2BMs) [10] incorporate an underlying causal directed acyclic graph (DAG) over concepts into the learning procedure. In this work, we leverage C2BMs to construct robot failure predictors whose intermediate representations are both semantically interpretable

and causally structured.

D. Robot Feedback to Human Teachers

Our aim is to use causal explanations in the context of interactive imitation learning (IIL) (e.g. Dataset Aggregation DAgger [32]). Robot-gated approaches can determine *when* such feedback is needed by requesting intervention based on uncertainty [18, 19, 28] or predicted failure [24]. However, these signals communicate only that the robot needs help, leaving the human to diagnose *why* and determine what corrective feedback to provide. More effective communication about the robot's learning, however, can increase transparency and human understanding of robot policies [16]. We seek failure explanations as a form of *robot-to-human feedback* for interactive learning. By representing failure through causally structured semantic concepts, a robot can communicate not only that it predicts failure, but also the factors it associates with that prediction and how changes to those factors may affect the predicted outcome. Importantly, these explanations describe the reasoning of a learned failure detector rather than ground-truth causes of policy behavior. Our opportunity is to treat them as *inspectable hypotheses*: a human can assess and probe the robot's explanation and use it to inform subsequent feedback.

III. Approach: Failure Prediction as Causally Reliable Concept Bottleneck Model

A C2BM is a concept-based architecture [10] that learns a "*causal bottleneck of concepts*" structured according to a provided causal graph. Let X denote noisy input features, $V = V_i i = 1^C$ a set of C semantically meaningful endogenous variables, and G a directed acyclic graph (DAG) specifying the causal relationships among variables in V. A C2BM learns an exogenous encoder g that models a distribution $P(U \mid X)$ over latent exogenous representations $U = U_i i = 1^C$, where each U_i is associated with an endogenous variable V_i, together with a parametric structural causal model (SCM) M_Θ.

Information flows through the architecture according to the causal structure specified by G. First, the encoder g infers the exogenous representations U from X. The SCM M_Θ then propagates information through G, predicting each endogenous concept V_i from its associated exogenous representation U_i and the values of its causal parents. This produces concept predictions that respect the specified causal structure and ultimately determine the model output.

We apply this architecture to robot failure prediction, where the output represents task success or failure and the intermediate concepts provide a causal hypothesis for why that outcome was predicted. We start with a hand-designed causal DAG, but also evaluate a DAG automatically extracted from the Botfails Dataset of robot rollout [4] and outputgenerated by a VLM [1].

Synthetic Causal Map.

We created a custom causal directed acyclic graph (Figure 2) to explicitly define concepts and edges that encode their functional causal relationships. Specifically, our DAG is a discrete Bayesian network, where each node, concept, corresponds to a random variable, and every node is paired with a conditional probability distribution (CPD). To evaluate [10] in a robotics setting, we define the classification problem as an assistive robot continuously verifying its capability to deliver coffee to a seated user in an at-home kitchen environment. Our leveraging [1] we aim to detect failure early, express what attributed to the failure,

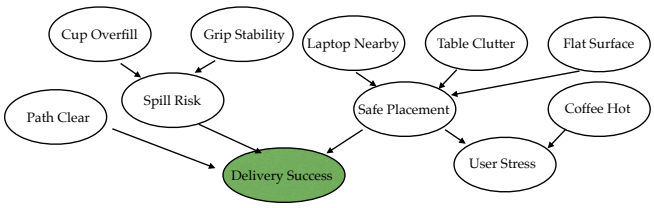


Fig. 2. We first validate our causal failure predictor training approach with a synthetic dataset generated from a ground truth causal DAG.

TABLE I
Prediction accuracy of task failure.

Model	Semantic Transp.	Causal Rel.	Toy DAG
OpaqNN	✗	✗	$71.57_{\pm 0.37}$
CBM_{+lin}	✓	✗	$70.92_{\pm 0.02}$
CBM_{+mlp}	✓	✗	$71.21_{\pm 0.30}$
CEM	✓	✗	$71.67_{\pm 0.29}$
SCBM	✓	✗	$70.43_{\pm 0.75}$
C^2BM	✓	✓	$70.13_{\pm 0.93}$

Following [10], we compare C2BM against an opaque neural network (OpaqNN) and several concept-based baselines, including CBMs, CEM, and SCBM, to evaluate the effect of introducing causal structure on task and concept prediction accuracy.

IV. Experimental Results

Takeaway 1: Prediction accuracy of our causally reliable failure predictor is on par with baselines.

Following [10], our first step consists of instantiating the C2BM with a causal DAG and Bayesian network. Table I shows that imposing this causal structure maintains failure prediction accuracy comparable to existing baselines.

We additionally find that the C2BM maintains high concept prediction accuracy, comparable to alternative concept-based architectures (Table II).

Takeaway 2: Structured causal predictions can be translated into human-interpretable failure explanations..

Beyond predicting whether a rollout will fail, the C2BM exposes the predicted states of semantically meaningful concepts and their causal relationships. We use a VLM to translate this structured representation into a concise verbal explanation

TABLE II
CONCEPT PREDICTION ACCURACY OF CONCEPT BOTTLENECK LAYERS.

MODEL	TOY DAG
OPAQNN$_M$	$77.95_{\pm 0.13}$
CBM$_{+lin}$	$77.94_{\pm 0.05}$
CBM$_{+mlp}$	$77.83_{\pm 0.80}$
CEM	$77.99_{\pm 0.06}$
SCBM	$78.11_{\pm 1.15}$
C^{2}BM	$77.98_{\pm 1.07}$

of the predicted failure. Importantly, the VLM does not independently infer the cause of failure from the rollout; rather, it communicates the causal hypothesis produced by the C2BM in natural language.

A. Generating Verbal Failure Explanations

We use Qwen3 [1] to translate the structured causal prediction produced by the C2BM into a verbal explanation for the human teacher. We provide the model with the predicted concept states and their causal relationships and prompt it to identify the primary predicted cause and relevant contributing factors. An example is shown below.

> **Prompt.** Translate the robot's causal failure prediction into a concise, non-technical explanation for a human teacher. Identify (1) the predicted outcome, (2) the primary causal factor, and (3) one or two relevant contributing factors. Do not report raw probabilities.
> **Predicted concepts:** `Path_Clear`: 0.92, `Table_Clutter`: 0.75, `Grip_Stability`: 0.08, `Spill_Risk`: 0.89, `Safe_Placement`: 0.14, ...
> **Generated explanation.** "I wasn't able to successfully move the cup. The main reason is that I couldn't get a secure grip on it, which made spilling very likely. The cluttered table also made it difficult to identify a safe place to set the cup down."

V. USER STUDY DESIGN PLAN

Given the C2BM's ability to express predicted failures through a bottleneck of human-interpretable concepts, we propose a user study examining how causal failure explanations can support human decision-making during interactive robot learning.

1) **RQ1: Workspace Intervention.** Can causal failure explanations help users modify the robot's workspace to improve task success? For example, if the robot attributes failure to workspace clutter, the user may remove obstacles or reorganize the scene rather than retrain the policy. We hypothesize that, when failure is attributable to modifiable environmental conditions, explanation-guided workspace interventions can provide a lower-cost path to task success than additional policy training.
2) **RQ2: Targeted Data Collection.** Can causal failure explanations help users collect more informative training data? Similar to guided data collection methods that identify underrepresented conditions [40], causal concepts can indicate specific conditions under which the policy frequently fails. For example, if cluttered scenes are repeatedly associated with failure, the robot can recommend collecting additional demonstrations in cluttered environments. We hypothesize that explanation-guided data collection will lead to more targeted and effective corrective demonstrations than unguided data collection.

Our proposed study begins with an initial round of policy training, followed by autonomous rollouts that produce both successful and failed trajectories. From these rollouts, we construct a causal concept graph, train the C2BM failure detector, and generate explanations for predicted failures. Participants then receive either an explanation-guided workspace intervention or an explanation-guided data-collection recommendation and act on that recommendation. We will evaluate whether these interventions improve subsequent task performance and compare against unguided interactive imitation learning, such as vanilla DAgger, and failure-triggered data collection approaches [24]. At the subjective level, we will evaluate via survey responses whether the explanations improve users' understanding of why the robot failed and their ability to select useful corrective actions.

VI. DISCUSSION AND LIMITATIONS

In this extended abstract, we present our approach and preliminary results. We first consider a hand-designed DAG encoding causal relationships among task-relevant concepts and train a C2BM to predict task success or failure through this structure. We then demonstrate how the resulting concept predictions and causal relationships can be translated by a VLM into semantically meaningful explanations for a human teacher.

Fit for Human-Robot Dialogue Workshop at IROS. Our work posits that human-robot dialogue in the form of causal hypothesis explanations will make interactive robot learning more collaborative. Ultimately, even failure predictions in an IIL setting are not necessarily the true cause of failure, and the robot can stop, report failure, and ask for help even when the human disagrees that failure has occurred. If the robot can explain the root causes of its failure, this could help increase transparency and trust with the user, and lead to targeted data collection.

Limitations. A primary limitation of our current approach is its reliance on a hand-designed DAG, which may be incomplete or encode incorrect causal relationships. Requiring this structure a priori also limits the generality of the approach. To address this limitation, we propose a pipeline using SAM2 [30] and VLMs [1] to identify task-relevant concepts directly from rollout videos and construct candidate causal graphs from observed successes and failures. Our future work will seek to make this process more iterative, allowing the robot to propose causal hypotheses and human teachers to validate or correct them, progressively refining the causal representation used for failure prediction and explanation.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] Andrea Bajcsy, Dylan P Losey, Marcia K O'Malley, and Anca D Dragan. Learning from physical human corrections, one feature at a time. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 141–149, 2018.

[3] Erdem Biyik and Dorsa Sadigh. Batch active preference-based learning of reward functions. *CoRR*, abs/1810.04303, 2018.

[4] Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, Steven Palma, Pepijn Kooijmans, Michel Aractingi, Mustafa Shukor, Dana Aubakirova, Martino Russi, Francesco Capuano, Caroline Pascal, Jade Choghari, Jess Moss, and Thomas Wolf. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. https://github.com/huggingface/lerobot, 2024.

[5] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[6] Joseph Chervenak, Harry Lieman, Miranda Blanco-Breindel, and Sangita Jindal. The promise and peril of using a large language model to obtain clinical information: Chatgpt performs strongly as a fertility counseling tool with limitations. *Fertility and sterility*, 120(3):575–583, 2023.

[7] Mary L. Cummings. What self-driving cars tell us about ai risks. *IEEE Spectrum*.

[8] Yinpei Dai, Jayjun Lee, Nima Fazeli, and Joyce Chai. Racer: Rich language-guided failure recovery policies for imitation learning. In *2025 IEEE international conference on robotics and automation (ICRA)*, pages 15657–15664. IEEE, 2025.

[9] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 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*, pages 351–360, 2021.

[10] Giovanni De Felice, Arianna Casanova Flores, Francesco De Santis, Silvia Santini, Johannes Schneider, Pietro Barbiero, and Alberto Termine. Causally reliable concept bottleneck models. *Advances in Neural Information Processing Systems*, 38:149099–149139, 2026.

[11] Maximilian Diehl and Karinne Ramirez-Amaro. A causal-based approach to explain, predict and prevent failures in robotic tasks. *Robotics and Autonomous Systems*, 162:104376, 2023.

[12] Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. The importance of being a band: Finite-sample exact distribution-free prediction sets for functional data. *arXiv preprint arXiv:2102.06746*, 2021.

[13] Yuqing Du, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando De Freitas, and Serkan Cabi. Vision-language models as success detectors. *arXiv preprint arXiv:2303.07280*, 2023.

[14] Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. Aha: A vision-language-model for detecting and reasoning over failures in robotic manipulation. *arXiv preprint arXiv:2410.00371*, 2024.

[15] Tesca Fitzgerald, Pallavi Koppol, Patrick Callaghan, Russell Quinlan Jun Hei Wong, Reid Simmons, Oliver Kroemer, and Henny Admoni. Inquire: Interactive querying for user-aware informative reasoning. In *6th Annual Conference on Robot Learning*, 2022.

[16] Soheil Habibian, Antonio Alvarez Valdivia, Laura H Blumenschein, and Dylan P Losey. A survey of communicating robot learning during human-robot interaction. *The International Journal of Robotics Research*, 44(4):665–698, 2025.

[17] Alaleh Hamidi and Kirk Roberts. Evaluation of ai chatbots for patient-specific ehr questions. *arXiv preprint arXiv:2306.02549*, 2023.

[18] Ryan Hoque, Ashwin Balakrishna, Ellen Novoseller, Albert Wilcox, Daniel S Brown, and Ken Goldberg. Thriftydagger: Budget-aware novelty and risk gating for interactive imitation learning. *arXiv preprint arXiv:2109.08273*, 2021.

[19] Ryan Hoque, Ashwin Balakrishna, Carl Putterman, Michael Luo, Daniel S Brown, Daniel Seita, Brijen Thananjeyan, Ellen Novoseller, and Ken Goldberg. Lazydagger: Reducing context switching in interactive imitation learning. In *2021 IEEE 17th international conference on automation science and engineering (case)*, pages 502–509. IEEE, 2021.

[20] Parag Khanna, Elmira Yadollahi, Mårten Björkman, Iolanda Leite, and Christian Smith. 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)*, pages 1829–1836. IEEE, 2023.

[21] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. Pmlr, 2020.

[22] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th international conference on computer vision*, pages 365–372. IEEE, 2009.

[23] Christoph H Lampert, Hannes Nickisch, Stefan Harmeling, et al. Learning to detect unseen object classes by between-class attribute transfer. In *Cvpr*, volume 1, page 3. Miami, FL, 2009.

[24] Huihan Liu, Yu Zhang, Vaarij Betala, Evan Zhang, James Liu, Crystal Ding, and Yuke Zhu. Multi-task interactive robot fleet learning with visual world models. *arXiv preprint arXiv:2410.22689*, 2024.

[25] Jiaming Liu, Chenxuan Li, Guanqun Wang, Xiaoqi Li, Sixiang Chen, Chuyan Xiong, Jiaxin Ge, Kaichen Zhou, and Shanghang Zhang. Self-corrected multimodal large language model for robot manipulation and reflection. 2023.

[26] Zeyi Liu, Arpit Bahety, and Shuran Song. Reflect: Summarizing robot experiences for failure explanation and correction. *arXiv preprint arXiv:2306.15724*, 2023.

[27] Shaunak A Mehta and Dylan P Losey. Unified learning from demonstrations, corrections, and preferences during physical human-robot interaction. *arXiv preprint arXiv:2207.03395*, 2022.

[28] Kunal Menda, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Ensembledagger: A bayesian approach to safe imitation learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5041–5048. IEEE, 2019.

[29] W Osten, C Bett, and G Situ. The challenge of making self-driving cars: may ai help to overcome the risks, or should we focus on reliable sensor technologies? In *Interferometry and Structured Light 2024*, volume 13135, pages 8–21. SPIE, 2024.

[30] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pages 28085–28128, 2025.

[31] Ralf Römer, Adrian Kobras, Luca Worbis, and Angela Schoellig. Failure prediction at runtime for generative robot policies. *Advances in Neural Information Processing Systems*, 38:7631–7670, 2026.

[32] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[33] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. *Active preference-based learning of reward functions*. 2017.

[34] Junwon Seo, Kensuke Nakamura, and Andrea Bajcsy. Uncertainty-aware latent safety filters for avoiding out-of-distribution failures. *arXiv preprint arXiv:2505.00779*, 2025.

[35] Amy L Stein. Assuming the risks of artificial intelligence. *BUL Rev.*, 102:979, 2022.

[36] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[37] Chen Xu, Tony Khuong Nguyen, Emma Dixon, Christopher Rodriguez, Patrick Miller, Robert Lee, Paarth Shah, Rares Ambrus, Haruki Nishimura, and Masha Itkina. Can we detect failures without failure data? uncertainty-aware runtime failure detection for imitation learning policies. *arXiv preprint arXiv:2503.08558*, 2025.

[38] Zhaojing Yang, Miru Jun, Jeremy Tien, Stuart Russell, Anca Dragan, and Erdem Biyik. Trajectory improvement and reward learning from comparative language feedback. In *8th Annual Conference on Robot Learning*.

[39] Zhouliang Yu, Jie Fu, Yao Mu, Chenguang Wang, Lin Shao, and Yaodong Yang. Multireact: Multimodal tools augmented reasoning-acting traces for embodied agent planning. 2024.

[40] Lihan Zha, Apurva Badithela, Michael Zhang, Justin Lidard, Jeremy Bao, Emily Zhou, David Snyder, Allen Z Ren, Dhruv Shah, and Anirudha Majumdar. Guiding data collection via factored scaling curves. *arXiv preprint arXiv:2505.07728*, 2025.

[41] Michelle Zhao, Reid Simmons, and Henny Admoni. Coordination with humans via strategy matching. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9116–9123. IEEE, 2022.

[42] Michelle Zhao, Reid Simmons, and Henny Admoni. The role of adaptation in collective human-ai teaming. *Topics in Cognitive Science*, 2022.

[43] Michelle Zhao, Reid Simmons, and Henny Admoni. Learning human contribution preferences in collaborative human-robot tasks. In *Conference on Robot Learning*, pages 3597–3618. PMLR, 2023.

[44] Zhi Zheng, Qian Feng, Hang Li, Alois Knoll, and Jianxiang Feng. Evaluating uncertainty-based failure detection for closed-loop llm planners. *arXiv preprint arXiv:2406.00430*, 2024.

[45] Enshen Zhou, Qi Su, Cheng Chi, Zhizheng Zhang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and He Wang. Code-as-monitor: Constraint-aware visual programming for reactive and proactive robotic failure detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6919–6929. IEEE, 2025.