Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Survey Track

A Comprehensive Survey on Physical Risk Control in the Era of Foundation Model-enabled Robotics

Takeshi Kojima[1†], **Yaonan Zhu**[1†], **Yusuke Iwasawa**[1†], **Toshinori Kitamura**[1], **Gang Yan**[1], **Shu Morikuni**[1], **Ryosuke Takanami**[1], **Alfredo Solano**[1], **Tatsuya Matsushima**[1], **Akiko Murakami**[2] and **Yutaka Matsuo**[1]
[1]The University of Tokyo
[2]Japan AI Safety Institute
{t.kojima, yaonan.zhu, iwasawa}@weblab.t.u-tokyo.ac.jp

Abstract

Recent Foundation Model-enabled robotics (FMRs) display greatly improved general-purpose skills, enabling more adaptable automation than conventional robotics. Their ability to handle diverse tasks thus creates new opportunities to replace human labor. However, unlike general foundation models, FMRs interact with the physical world, where their actions directly affect the safety of humans and surrounding objects, requiring careful deployment and control. Based on this proposition, our survey comprehensively summarizes robot control approaches to mitigate physical risks by covering all the lifespan of FMRs ranging from pre-deployment to post-accident stage. Specifically, we broadly divide the timeline into the following three phases: (1) pre-deployment phase, (2) pre-incident phase, and (3) post-incident phase. Throughout this survey, we find that there is much room to study (i) pre-incident risk mitigation strategies, (ii) research that assumes physical interaction with humans, and (iii) essential issues of foundation models themselves. We hope that this survey will be a milestone in providing a high-resolution analysis of the physical risks of FMRs and their control, contributing to the realization of a good human-robot relationship.

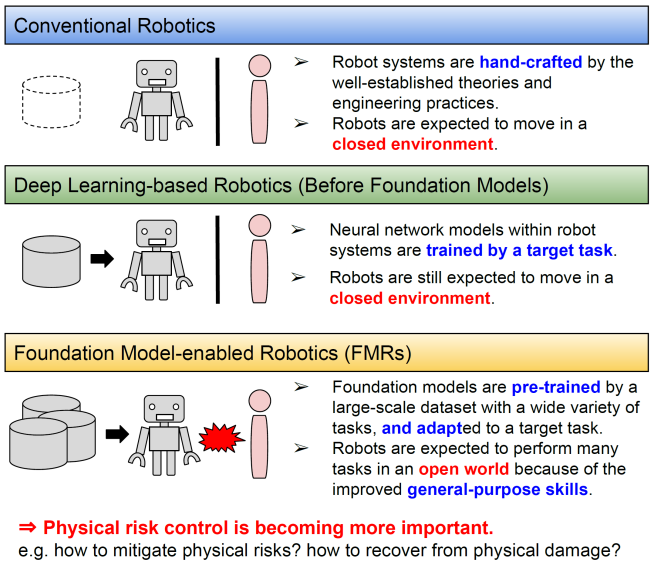

Figure 1: Motivation of our survey. Studies on conventional robotics and Deep Learning-based robotics (before foundation models) were mainly based on closed environments where physical risks are excluded or minimized by restricting the entry of humans and objects, e.g., inside a factory. Foundation Model-enabled robotics (FMRs) are expected to be utilized in an open world where physical risks inevitably exist because humans and robots are always in close proximity and physically interacting, e.g. a restaurant. Hence physical risk control is becoming more important in the era of FMRs.

1 Introduction

Since the emergence of foundation models, robotics has shown dramatic improvements in general-purpose and highly adaptable manipulation skills, indicating that it has entered a new era called Foundation Model-enabled robotics (FMRs). FMRs leverage large-scale pre-trained neural network models that integrate language, vision, and action modalities, enabling robots to generalize across diverse tasks [Firoozi *et al.*, 2023]. They comprise large language models (LLMs) and vision-language models (VLMs) for language comprehension and visual understanding, enhancing high-level task planning for long-horizon tasks [Liang *et al.*, 2023]. Additionally, they include robot transformers that integrate perception, decision-making, and action generation to process multimodal inputs and generate low-level motion commands for end-to-end control [Brohan *et al.*, 2022].

In contrast to conventional robotic engineering, which generally controls robots based on human-crafted rules, FMRs learn to control themselves from enormous amounts of data with a statistical approach. This paradigm shift has enhanced generalizability and long-horizon reasoning [Zawalski *et al.*, 2024], enabling FMRs to show promising advantages over classical methods in adapting to diverse tasks and unstructured environments. They have been successfully applied to various fields, including task planning [Ahn *et al.*, 2022], vi-

† Corresponding Authors.

sion language guided manipulation [O'Neill *et al.*, 2023], tactile perception [Zhao *et al.*, 2024a], locomotion [Bohlinger *et al.*, 2024], and navigation [Moroncelli *et al.*, 2024]. (➢ Section 2)

While we could enjoy economic benefits by replacing human labor with robots in various applications, we cannot completely avoid the risk of FMRs causing physical damage to surrounding humans or objects. As FMRs are engaged in more challenging tasks and environments, often requiring contact with people and objects, such as housework, surgery or nursing care, the risk of causing physical damage increases. Of course, even in such an environment, it is possible to mitigate some risks through the establishment of social rules and education, e.g., do not play near robots. However, we cannot completely eliminate accidents because of our misperception of the environment (e.g. robots in blind spots) or unexpected environmental changes (e.g., sudden hardware failures). (➢ Section 3)

Based on the premise that FMRs cannot completely eliminate the risk of causing physical damage, this survey comprehensively summarizes robot control approaches against physical risks by covering all the lifespan of FMRs, from pre-deployment to post-accident stage. Specifically, our study conducts this survey of physical risk controls by dividing the timeline into the following three phases: (1) pre-deployment phase, i.e., risk prevention phase when learning from data, (2) pre-incident phase, i.e., before an incident happens after deployment, and (3) post-incident phase, i.e., the recovery and improvement stage. (➢ Figure 3 and Section 4)

We emphasize that many classic robotic studies discussed safety control within a closed environment where humans could intervene and prevent hazardous incidents, such as inside a factory or laboratory. In such an environment, if a robot causes an incident, pressing the emergency stop button will solve the problem. In contrast, our survey focuses on the safety of FMRs assumed to act in an open world, where bigger physical risks often exist because humans and robots are always in close proximity and physically interacting, e.g. inside a home or a cafe. In this case, we need to consider how to recover or treat robots as well as surrounding humans and objects because "life still goes on" for both robots and humans after the incident. We also emphasize that prior surveys of FMRs focused on only some partial sections within the first two phases of safety control, i.e., pre-deployment and pre-incident phase in light of our classification scheme, and so have been insufficiently organized in detail [Bommasani *et al.*, 2021; Hu *et al.*, 2023; Firoozi *et al.*, 2023; Xiao *et al.*, 2023]. In other words, they only summarized partial sections of the mitigation strategies of physical risks before an incident, and generally have not covered post-incident recovery actions. (➢ Figure 2)

Throughout this survey, we have found that there is much room to study (i) pre-incident risk mitigation strategies, (ii) research that assumes physical interaction with humans, and (iii) essential issues of foundation models themselves. We emphasize that social measures such as legislation or insurance schemes are also important aspects to enhance mitigation of physical damage. (➢ Section 5)

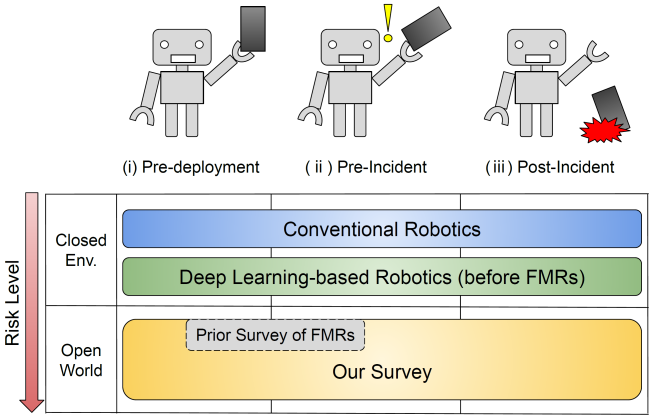

Figure 2: Scope of our survey. FMRs are expected to be utilized in an open world where humans and robots are interacting with each other. Based on this proposition, our study conducts a comprehensive suvey on physical risk control by categorizing the lifespan of FMRs into (1) pre-deployment phase, (2) pre-incident phase after deployment, and (3) post-incident phase. In contrast, prior surveys of FMRs mainly focused on partial sections of the first two phases.

2 Robotics and Foundation Models

2.1 Conventional Robotics Engineering

Conventional robotics has driven innovation for long, predating the rise of data-driven techniques. While modern robotics increasingly rely on deep learning and large-scale data, classical theories and engineering remain essential. In particular, it focuses on five pivotal directions: (1) innovative mechanical design [Sugiura *et al.*, 2024], (2) precise motion control [Sun *et al.*, 2018], (3) advanced perception systems [Mahler *et al.*, 2019], (4) planning and decision-making strategies [La Valle, 2011; La Valle, 2011; Hansel *et al.*, 2023], (5) adaptive learning and optimization methodologies [Zhang *et al.*, 2024; Saveriano *et al.*, 2023].

Robotics relies on robust mechanics and precise motion control for seamless interaction with the physical world. Engineers continually refine structures and control strategies to enhance performance, versatility, and user integration [Sugiura *et al.*, 2024; Yamamoto *et al.*, 2019; Zhu *et al.*, 2019; Yamamoto *et al.*, 2019; Hossain, 2023; Zhu *et al.*, 2019]. On the other hand, perception, planning, and adaptive learning drive robotic intelligence for open-world deployment. Advanced perception enables environmental awareness, while planning and decision-making allow navigation in complex scenarios [Chen *et al.*, 2019]. Adaptive learning techniques, including imitation learning, reinforcement learning, and deep learning-based approaches help robots acquire skills and adapt to changing task conditions [Zhang *et al.*, 2024; Saveriano *et al.*, 2023]. Together, these elements bridge mechanical capability with intelligent autonomy, enabling robots to operate effectively in dynamic environments.

2.2 Foundation Model-enabled Robotics (FMRs)

Despite advances in perception, planning, and adaptive learning, traditional methods often struggle with scalability, generalization, and handling of multimodal information in complex

environments. Even after the advent of Deep Learning-based Robotics, in which neural-network models within robot systems are trained on a specific target task, performance improvements were limited, so robots were still expected to stay in closed environments. Foundation models [Firoozi *et al.*, 2023] are large scale neural-network models which are pre-trained on broad data in self-supervised approaches and can be adapted to a wide range of downstream tasks by fine-tuning or prompting. They were firstly proposed in the natural language processing domain [Devlin *et al.*, 2019], eventually spreading to wide range of modalities including image, video, and audio because of their remarkable performance and generalizability [Dosovitskiy *et al.*, 2021; Arnab *et al.*, 2021; Baevski *et al.*, 2020].

FMRs have recently emerged as large-scale pre-trained models that integrate language, vision, and action modalities, enabling robots to generalize across diverse tasks [Firoozi *et al.*, 2023]. They comprise large language models (LLMs) and vision-language models (VLMs) for language comprehension and visual understanding, enhancing high-level task planning for long-horizon tasks [Liang *et al.*, 2023]. Additionally, they include robot transformers that integrate perception, decision-making, and action generation to process multimodal inputs and generate low-level motion commands for end-to-end control. Several studies such as RT-1 and RT-X [Brohan *et al.*, 2022; O'Neill *et al.*, 2023] have trained models with massive number of demonstration samples collected from the real world to realize generalization across different morphologies. π_0 [Black *et al.*, 2024] pre-trained a vision-language-action model on a diverse crossembodiment dataset with a variety of dexterous manipulation tasks, followed by fine-tuning with high quality data to enable complex multi-stage tasks, such as folding multiple articles of laundry or assembling a box.

3 Potential and Risks

Given their performance and adaptability, FMRs are poised to be used in an increasing number of real-world applications, thus speeding up the general adoption of robots in society. FMRs will no longer be confined to closed environments where physical risks are excluded or minimized by restricting the entry of humans and objects, such as factories or warehouses. Instead, FMRs will be able to perform activities in an open world where humans and robots are always in close proximity and physically interacting, e.g. inside a home, restaurant, or a public square. They are expected to reduce or replace many types of repetitive but complex manual labor, allowing people to pursue more engaging and rewarding activities. In countries with low birth rates and increasing aging populations, FMRs could help stabilize the workforce, creating a good impact on their economies.

While FMRs are expected to perform a wide variety of tasks in our open world, they will add many risks to our society, such as malicious usage (e.g. fraud), unintentional physical damage to humans or objects, environmental destruction due to their high power consumption and resource usage, or privacy information leakage caused by security vulnerabilities. This survey focuses on the risks of FMRs causing unintentional physical damage to humans or objects. As FMRs are engaged in more challenging tasks that often require contact with people and objects such as housework, surgery or nursing care, risks of physical damage increase. Of course, even in such an environment, it is possible to reduce certain risks through the establishment of social rules and education, e.g., do not play near robots. However, we cannot completely eliminate accidents because of misperception of the environment (e.g. robots in blind spots) or unexpected environmental changes (e.g., sudden hardware failures).

4 Control of Physical Risks

This section comprehensively summarizes robot control approaches against physical risks by covering all the lifespan of FMRs ranging from pre-deployment to post-accident stage. Specifically, we summarize physical risk control approaches by dividing the timeline of FMRs into the following three phases: (1) pre-deployment phase, i.e., risk mitigation phase, (2) pre-incident phase, i.e., a situation before an incident happens after deployment, and (3) post-incident phase, i.e., recovery and improvement stage. Although some of the surveyed papers include studies that do not use foundation models, we cite them as technologies that are expected to be utilized in future research of FMRs.

4.1 Pre-deployment Phase

Hardware and Software for Safety

Ensuring safety in robotic systems requires both robust hardware design and software-based safety limits. While physical mechanisms contribute to risk mitigation, software constraints play a crucial role in preventing hazardous behaviors and enforcing operational safety [Zacharaki *et al.*, 2020]. Safety-focused hardware includes force-limiting mechanisms such as series elastic actuators that absorb shocks and restrict excessive forces [Bodo *et al.*, 2023], collision detection sensors with safety prioritized control to stop the robot motion immediately upon contact [Haddadin *et al.*, 2008], and artificial skin to enable robots to autonomously sense the surroundings for enhanced safety while working near people [Bergner *et al.*, 2022]. Additionally, the use of compliant materials and soft robotics components helps reduce the risk of damage or injury during physical interactions [Truby *et al.*, 2019]. Safety standards like ISO/TS 15066 [Matthias and Reisinger, 2016] highlight the necessity of mechanical and electrical safety features, including emergency stop buttons and torque-limiting mechanisms [Lee *et al.*, 2009].

Software-based safety mechanisms complement hardware solutions by enforcing predefined constraints to prevent dangerous operations. Common approaches include velocity and torque limits that curb motor outputs to avoid excessive force application [Haddadin *et al.*, 2007; Ferraguti *et al.*, 2022; Haddadin *et al.*, 2007], virtual fences that restrict the robot's movement within designated safe areas, and fault monitoring systems that detect anomalies and trigger protective responses when necessary [Guiochet *et al.*, 2017]. Admittance control and other safety-aware algorithms dynamically adjust the robot's behavior based on external forces to enhance safe operation in unpredictable environments [Sun *et al.*, 2024b].

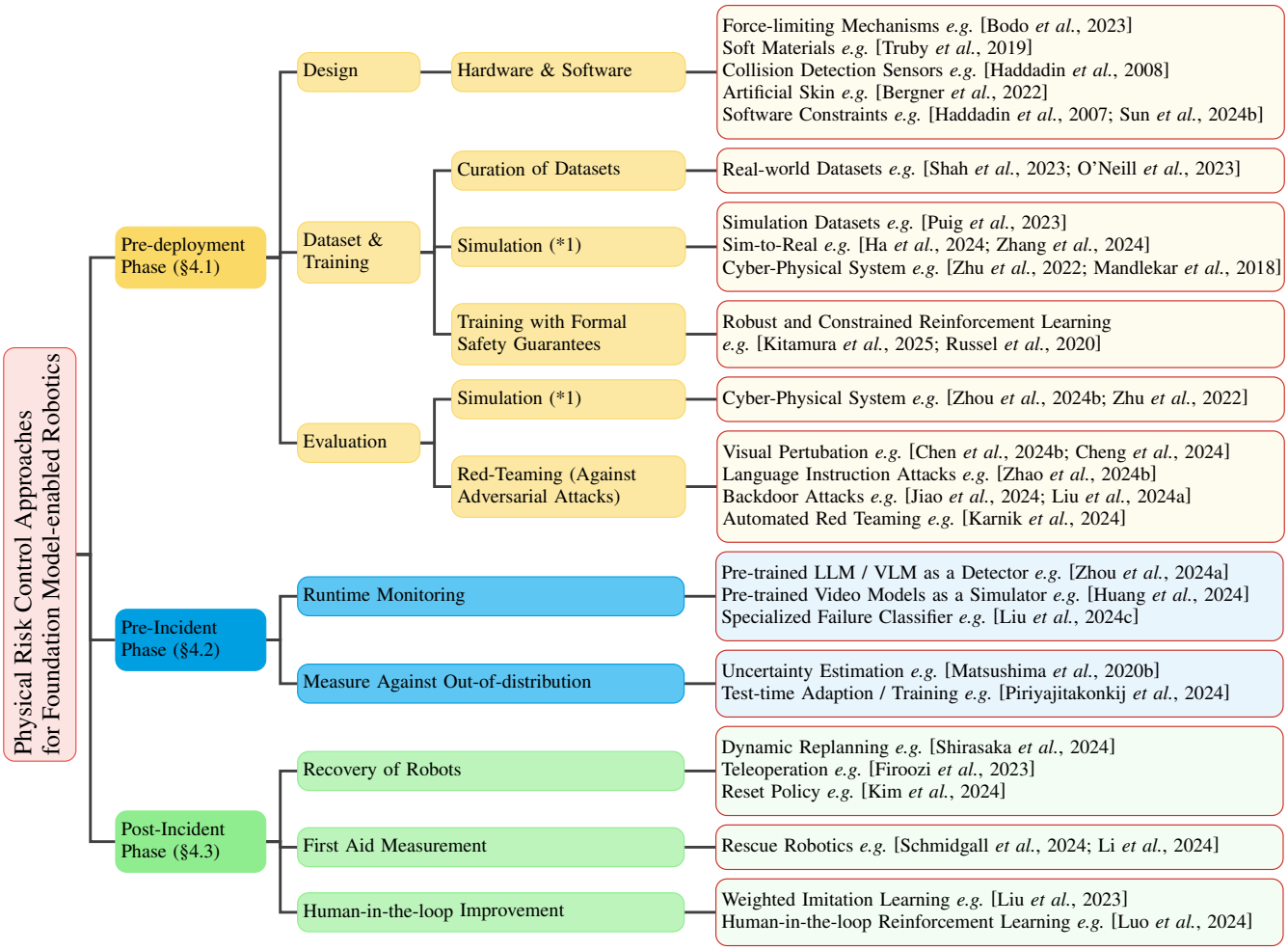

Figure 3: Categorization of physical risk control approaches for FMRs. (*1) Simulation plays a critical role as both a training environment and an evaluation framework. For simplicity, we integrated both of the contents into one subsection as "Simulation" in §4.1.

Together, these hardware and software measures form a robust safety framework, ensuring mechanical reliability and controlled operation for safe robotic deployment. Together, these hardware and software considerations create a comprehensive safety framework, ensuring both mechanical robustness and controlled operational behavior for reliable and safe robotic deployment.

Curation of Datasets

In the era of FMRs, curation of large-scale datasets with a wide variety of tasks is important for pre-training FMRs to increase generalization skills. Empirical evidence suggests that domain generalization abilities improve significantly when larger models are (pre-)trained on larger and more diverse datasets, indicating the reduction of risk of out-of-distribution (➢ Measure Against Out-of-distribution).

Real-world datasets are often the most intuitive and accurate source of information required for high performing FMRs. However, creating large scale and high quality real-world datasets for robotics is challenging due to its cost [Khazatsky *et al.*, 2024; O'Neill *et al.*, 2023]. For example, collecting a good set of demonstration data requires intense labour and skilled operators. The scaling cost increases proportional to the diversity of tasks, skills, objects, environments and embodiments. BREMEN [Matsushima *et al.*, 2020a] introduces a deployment-efficient model-based reinforcement learning approach that achieves policy learning with significantly fewer environment interactions by training a model of the environment and using offline optimization to update policies without excessive real-world data collection.

GNM [Shah *et al.*, 2023] is a recent notable effort which successfully integrated six different large-scale navigation datasets, formulated a unified navigation interface based on waypoints, and deployed it on different mobile platforms. Another significant effort on manipulation tasks is RT-X [O'Neill *et al.*, 2023], which is a joint collaboration among 34 laboratories with the goal to establish a standardized data format across 60 existing datasets with 22 robot embodiments. To further aid diversity in task and modality volumes, RH20T [Fang *et al.*, 2023] collected over 110k manipulation episodes, covering more than 140 contact-rich skills, including well calibrated RGB, depth, force-torque, tactile, proprioception, audio and language instruction.

Simulation

Simulation plays a pivotal role in robotics given the risks and costs associated with physical testing of robots, such as potential damage to the robot, harm to humans, or unintended environmental impacts. It offers a safe and controlled environment to design, test, and refine robotic systems before their deployment in real-world scenarios [Rohmer *et al.*, 2013; Huck *et al.*, 2023; Kargar *et al.*, 2024]. This is especially important in the era of FMRs, which is a probabilistic-based approach at its core aimed at a wide variety of tasks.

One of the critical applications of simulation is Sim-to-Real (Sim2Real) policy learning, which enables robots to develop and validate control policies in simulated environments [Bohlinger *et al.*, 2024]. Sim2Real helps mitigate safety concerns by enabling extensive testing under diverse and challenging conditions, allowing us to identify potential failures, refine safety constraints, and ensure robust real-world performance [Zhao *et al.*, 2020]. Advanced simulation frameworks like NVIDIA Isaac Sim, Isaac Lab [Mittal *et al.*, 2023], and Genesis [Genesis-Authors, 2024] help bridge the "reality gap" with accurate physics modeling, and high-fidelity graphics, which is essential for safety assurance in the pre-deployment phase.

Additionally, the robot in the simulator can receive control inputs from the physical world, creating a seamless cyber-physical system [Zhou *et al.*, 2024b]. This integration not only allows real-world devices, such as controllers or sensors, to interact with the virtual robot for testing and development in a simulation environment [Zhu *et al.*, 2022], but also facilitates the training of robot control policies through imitation learning. Demonstrations can be provided via teleoperation [Mandlekar *et al.*, 2018], enabling the robot to learn complex tasks in a safe and scalable manner within the simulation.

Simulation is also used to generate greater amounts of training data. Simulation can efficiently create diverse range of domain randomized data which is expected to promote generalization ability of models in FMRs. Generative simulations Genesis [Genesis-Authors, 2024], Gen2Sim [Katara *et al.*, 2024], and FACTORSIM [Sun *et al.*, 2024a] have emerged as a promising solution by automating the creation of diverse, scalable environments and facilitating broader coverage of training conditions. Habitat 3.0 [Puig *et al.*, 2023] and AI2THOR [Kolve *et al.*, 2017] are another line of effort for interactive environment frameworks focusing on scene realism for both navigation and manipulation tasks.

Red-teaming (Against Adversarial Attacks)

Red-teaming is the practice of simulating an enemy team attempting to perform some type of attack or other hostile action to the organization (i.e., blue team). It is common practice in the fields of defense, security and IT operations, where militaries and system administrators test their own systems in search of weak points that could be exploited. Recent LLM developers also organize red-teaming to assess the models' vulnerabilities by comprehensive stress-testing [Lin *et al.*, 2024]. In the case of FMRs, the practice is not yet as common, but it is expected to popularize as they continue to improve and the range of tasks they can perform reaches human or above level. As a pioneering example, [Karnik *et al.*, 2024] uses automated red teaming techniques with VLMs to generates diverse and challenging instructions. Experimental results show that state-of-the-art models frequently fail or behave unsafely on the tests, underscoring the shortcomings of current benchmarks.

One promising technique for stress-testing FMRs in red-teaming is adversarial attacks [Costa *et al.*, 2024]. It is well known that models utilizing deep neural networks (DNNs) are vulnerable to small input perturbations. Because FMRs are based on LLMs and VLMs, which themselves rely on DNNs, many similar problems have been reported. In the case of FMRs, it is especially important to understand which types of attacks they are vulnerable to, since these models not only make recognition errors but also act in the real world.

For example, as a direct attack on FMRs, [Chen *et al.*, 2024b] reports that both global perturbation attacks on Diffusion Policy and adversarial patches in a physical environment are effective in online and offline settings. [Cheng *et al.*, 2024] investigates the robustness of VLAs against various visual attacks such as Gaussian noise, changes in brightness, Adversarial Patches that modify part of an image, and Visual Prompts (e.g., adding the word "Stop" into images to control behavior). [Jiao *et al.*, 2024; Liu *et al.*, 2024a; Wang *et al.*, 2024] show vulnerabilities to backdoor attacks that use everyday objects (e.g., a yellow CD) as triggers to degrade behavior. [Zhao *et al.*, 2024b] have proposed to add adversarial suffixes to language inputs.

On the other hand, despite these demonstrated vulnerabilities, [Zhao *et al.*, 2024b] reports that many current FMRs have discrete action spaces, making standard attacks less effective. Still, in cases where attackers have access to internal features, using these features can increase the success rate of adversarial attacks, indicating the need for continued research on countermeasures. Because FMRs generally rely on LLMs and VLMs, it is necessary to verify the effectiveness of methods proven to work well in those models. Another challenge in research on adversarial attacks against FMRs is the lack of standard evaluation metrics. Although vulnerabilities have been studied in various components—ranging from simulations and real-world settings to planning vision modules—research on FMRs is still in its early stages, and a unified evaluation strategy remains insufficient. In particular, because the environments in which robots are expected to operate can be extremely diverse and because the dynamics of different robots vary, further validation is required to determine how generalizable the current findings are. As an earlier example of such attempts, [Lu *et al.*, 2024] proposes Harmful-RLbench, which evaluates the planning capabilities of LLMs in an environment featuring 25 distinct task scenarios. Moreover, [Zhao *et al.*, 2024b] develops VIMA-bench, an evaluation benchmark covering 13 types of robotic manipulation tasks.

Training with Formal Safety Guarantees

Safe controller design during the pre-deployment phase has been extensively studied in the field of *robust control theory*. Since the exact knowledge of the environment is unknown before deploying the robot, robust design accounts for environmental uncertainty and incorporates conservative risk

management into the robot controller.

Robust model predictive control and H_∞ optimal control [Bemporad and Morari, 2007; Zames, 1981; Doyle, 1982] are the representative robust control methods in linear dynamical systems, where the system is modeled as $x^{(t+1)} = Ax^{(t)} + Bu^{(t)}$. Here, $x^{(t)} \in \mathbb{R}^n$ and $u^{(t)} \in \mathbb{R}^m$ represent the system state and input signal, respectively. These robust controllers guarantee safety satisfaction even when the dynamics matrices (A, B) perturb from the nominal matrices.

However, linear models are unsuitable for modeling the recent nonlinear and high-dimensional input systems that FMRs aim to control (➢ Section 2.2). Robust reinforcement learning (RL) offers an alternative framework, capable of addressing robust nonlinear control problems when combined with function approximation techniques [Moos *et al.*, 2022]. However, robust RL alone is insufficient to achieve both high performance and safety, as ensuring safety typically involves solving constraint satisfaction problems (e.g., obstacle avoidance in self-driving systems [Altman, 1999]). While the RL community has recently begun exploring the combination of safety constraints and robustness [Mankowitz *et al.*, 2020; Russel *et al.*, 2020], theoretical results in this area remain limited. A recent result by [Kitamura *et al.*, 2025] presents an algorithm for computing a robust and constrained controller in a tabular Markov decision process setting. However, it does not account for the challenges posed by nonlinear dynamics, which are crucial for FMRs. In short, the theoretical question: "When and how can we realize robust constrained control in FMRs?" remains largely unanswered.

4.2 Pre-Incident Phase

Runtime Monitoring

Runtime monitoring is one of the fundamental tools for ensuring safety in robot policies. The monitoring systems are sometimes called "critics" of the policy analogical to actor-critic of reinforcement learning [Sutton and Barto, 2018].

Recently, LLMs, VLMs, and video prediction models have been utilized as critics of robot policies. Firstly, VLMs are utilized to detect success (or failure) in policy rollouts. For example, [Kanazawa *et al.*, 2023] proposes to leverage VLMs to detect state change of objects in cooking tasks, which is useful for executing task plans. Secondly, VLMs are used for constraint monitoring. Code-as-monitor [Zhou *et al.*, 2024a] leverages VLMs for generating programs for monitoring robot policy rollouts from robot image observations and descriptions of constraints generated by LLMs. Thirdly, [Huang *et al.*, 2024; Escontrela *et al.*, 2023] utilized log-likelihood of pre-trained video prediction models as a reward signal for the robot's actions to monitor in real time whether the state transitions in the environment are being properly learned.

Another approach to runtime monitoring involves training a failure classifier using human intervention data. Specifically, these methods leverage robot demonstration data to train a world model, enabling the learning of latent representations. By utilizing these latent representations along with human intervention flags, a failure detector can be trained [Liu *et al.*, 2024c]. Such methods are particularly effective as automated safety validators when robots are operating in parallel within an environment and sequentially learning policies [Liu *et al.*, 2024d].

Measure Against Out-of-distribution

After deploying a trained model in real-world scenarios, we may encounter out-of-distribution (OOD), in which robot inputs fall outside the data distribution used to train a model. Measures against this situation are crucial for safety when deploying robots in real-world scenarios. Specifically, the test data D_{test} is sampled from a distribution P_{test} , which invariably differs from the training distribution P_{train} . This discrepancy highlights the challenge of distributional shifts.

A key research to mitigate this risk is improving distributional robustness by optimizing the worst-case performance across various potential distributional shifts, thus ensuring dependable OOD performance [Ben-Tal *et al.*, 2012; Duchi and Namkoong, 2020]. However, since P_{test} is not directly accessible and the model $\boldsymbol{f}$ is learned from a finite set of training samples D_{train}, there is no guarantee that $\boldsymbol{f}$ will make accurate predictions during testing. Uncertainty estimation focuses on determining when and where the model individual predictions can be trusted, and, conversely, where confidence is lacking [Matsushima *et al.*, 2020b; Garnelo *et al.*, 2018b; Garnelo *et al.*, 2018a; Kingma, 2013]. Besides, causal inference is leveraged to address the root cause of poor generalization under distributional shifts. Learned models often rely on spurious correlations present in D_{train}, rather than capturing the invariant cause-and-effect relationships that drive the underlying process [Peters *et al.*, 2015; Arjovsky *et al.*, 2020; Pearl, 2009] In recent years, concepts such as test-time adaption/training [Wang *et al.*, 2021; Sun *et al.*, 2020; Piriyajitakonkij *et al.*, 2024; Park *et al.*, 2024] have been introduced into robotics research. Test-time adaptation allows a model to adjust its internal parameters or normalization statistics using the unlabeled data encountered during deployment, while test-time training leverages auxiliary self-supervised tasks to update the model during inference.

4.3 Post-incident Phase

Recovery of Robots

In the post-incident phase, FMRs play a vital role in autonomously detecting, assessing, and mitigating more hazardous risks [Chen *et al.*, 2024a]. These systems leverage real-time monitoring and fault detection to identify anomalies, such as hardware malfunctions or environmental changes [Shirasaka *et al.*, 2024]. Once a risk is detected, foundation models enable dynamic replanning to adjust trajectories or control policies, ensuring safe operation. However, when autonomous recovery is insufficient, errors can also be addressed through human intervention, such as teleoperation, ensuring flexibility and safety in complex scenarios [Liu *et al.*, 2024b].

Additionally, learning-based reset mechanisms play a crucial role in preventing robots from entering non-reversible states during reinforcement learning, improving safety, and reducing the need for manual intervention. For instance, a reset policy can reduce the number of entering non-reversible states, and manual resets to learn a task, while enhancing

safety and improving learning efficiency [Eysenbach *et al.*, 2018]. Similarly, reset-based deep ensemble methods enhance sample efficiency in safe RL by overcoming the limitations of the vanilla reset method [Kim *et al.*, 2024].

Foundation models also demonstrate the potential to resolve deadlocks in multi-agent robotic systems (MRS) by using high-level planners such as LLMs or VLMs [Garg *et al.*, 2024], ensuring smooth collaboration among agents. In safety-critical situations, the system may communicate risks to operators, ensuring effective recovery when autonomous methods fall short [Eder *et al.*, 2014]. By integrating proactive monitoring, adaptive planning, and contextual decision-making, foundation models enhance reliability and safety across dynamic environments [Firoozi *et al.*, 2023].

First Aid Measurement

When humans and objects are physically damaged by robots, immediate first aid measures are extremely important. In this situation, two types of first aid measures are possible. One is to call for help from rescue people/robots, the other is to call for help from the nearby robot that caused the damage itself to conduct basic first aid treatment.

Rescue robots [Delmerico *et al.*, 2019] are designed to help search and rescue people in the event of a disaster or emergency situation. They have been actively studied since before the advent of FMRs. Recently, several studies have developed robot models for human rescue by learning from data to improve quality and expand activity to more challenging situations, such as surgical robot systems or assistants [Yue *et al.*, 2023; Schmidgall *et al.*, 2024], and robot-assisted pedestrian evacuation in fire scenarios [Li *et al.*, 2024]. However, there is no guarantee that such specialized robots will always be near humans in emergency situations in our open world. Therefore, it will be necessary for robots that do not specialize in emergency rescue tasks to have the functionality to provide temporary aid, such as checking life signs, automatic call for an ambulance or rescue people/robots, provide useful information to nearby people, or stop bleeding with bandages.

Human-in-the-loop Improvement

Human-in-the-loop improvement aims to build a continuous improvement mechanism or pipeline that collects effective feedback from humans for model training by leveraging human intervention history or demonstrations. There are some pioneering studies on human-in-the-loop improvements that are expected to be applied to FMRs in the future. One such approach involves continuous human monitoring of policy deployment, where a human intervenes to stop the robot when a failure is imminent. The data immediately prior to the intervention is then used as negative examples for weighted imitation learning [Liu *et al.*, 2023]. The positive success data and negative failure data obtained through intervention in a sub-optimal policy are also highly compatible with RL. There are methods that leverage this by storing both successful policy rollouts and intervention data in an RL replay buffer, enabling the policy to learn from failures through reinforcement learning [Luo *et al.*, 2024].

5 Conclusion and Discussion

Our survey comprehensively summarized robot control approaches against physical risks by covering all the lifespan of FMRs ranging from pre-deployment to post-accident stage.

From this survey, we found that there is much room for future work of FMRs on the following three points. (i) Considering that there are a myriad of environments and task varieties in the real world, we need to pay more attention to risk mitigation before an actual incident happens (e.g., stress-testing as broadly as possible with red-teaming, promote generalizability of FMRs to prevent OOD, detection and intervention in failures at the earliest stage). (ii) Accelerating research that assumes physical interactions with humans in more realistic world settings (e.g., improving learning methods to strictly observe social rules, or research in the post-incident phase such as continuous improvement mechanisms and prevention of more hazards after an incident). (iii) While we can easily adapt pre-trained foundation models to a specific task with a small number of samples by fine-tuning or prompting, it becomes important to tackle essential issues of foundation models themselves when applying them to robotics (e.g., how to ensure the quality of large-scale pre-training datasets to prevent malfunction of trained models in robots, or how well do LLMs or VLMs understand the physical world in terms of predicting hazards, such as collision prediction between humans and robots through monitoring their motions and surrounding environment).

We also emphasize that in addition to the technical aspects, social measures such as legislation, insurance schemes, and ethical guidelines are important to enhance aftercare of physical damage in practice. We hope that this study will be a milestone in providing a high-resolution analysis of the physical risks of FMRs and their control, contributing to the realization of a good human-robot relationship.

Contribution Statement

Takeshi Kojima, Yaonan Zhu, and Yusuke Iwasawa have contributed equally to this study.

References

[Ahn *et al.*, 2022] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.

[Altman, 1999] Eitan Altman. *Constrained Markov Decision Processes*, volume 7. CRC Press, 1999.

[Arjovsky *et al.*, 2020] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv eprint 1907.02893*, 2020.

[Arnab *et al.*, 2021] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6836–6846, 2021.

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, volume 33, pages 12449–12460, 2020.

[Bemporad and Morari, 2007] Alberto Bemporad and Manfred Morari. Robust model predictive control: a survey. In *Robustness in Identification and Control*, pages 207–226. Springer, 2007.

[Ben-Tal *et al.*, 2012] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 2012.

[Bergner *et al.*, 2022] Florian Bergner, Emmanuel Dean-Leon, and Gordon Cheng. *Neuromorphic principles for large-scale robot skin*, pages 91–123. Institution of Engineering and Technology, January 2022.

[Black *et al.*, 2024] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[Bodo *et al.*, 2023] Giulia Bodo, Federico Tessari, Stefano Buccelli, Luca De Guglielmo, Gianluca Capitta, Matteo Laffranchi, and Lorenzo De Michieli. Customized series elastic actuator for a safe and compliant human-robot interaction: Design and characterization. In *2023 International Conference on Rehabilitation Robotics (ICORR)*, pages 1–6, 2023.

[Bohlinger *et al.*, 2024] Nico Bohlinger, Grzegorz Czechmanowski, Maciej Krupka, Piotr Kicki, Krzysztof Walas, Jan Peters, and Davide Tateo. One policy to run them all: an end-to-end learning approach to multi-embodiment locomotion. *arXiv preprint arXiv:2409.06366*, 2024.

[Bommasani *et al.*, 2021] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[Brohan *et al.*, 2022] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[Chen *et al.*, 2019] Hao Chen, Weiwei Wan, and Kensuke Harada. Combined task and motion planning for a dual-arm robot to use a suction cup tool. In *Humanoids*, pages 446–452. IEEE, 2019.

[Chen *et al.*, 2024a] Hongyi Chen, Yunchao Yao, Ruixuan Liu, Changliu Liu, and Jeffrey Ichnowski. Automating robot failure recovery using vision-language models with optimized prompts, 2024.

[Chen *et al.*, 2024b] Yipu Chen, Haotian Xue, and Yongxin Chen. Diffusion policy attacker: Crafting adversarial attacks for diffusion-based policies. *arXiv preprint arXiv:2405.19424*, 2024.

[Cheng *et al.*, 2024] Hao Cheng, Erjia Xiao, Chengyuan Yu, Zhao Yao, Jiahang Cao, Qiang Zhang, Jiaxu Wang, Mengshu Sun, Kaidi Xu, Jindong Gu, et al. Manipulation facing threats: Evaluating physical vulnerabilities in end-to-end vision language action models. *arXiv preprint arXiv:2409.13174*, 2024.

[Costa *et al.*, 2024] Joana C Costa, Tiago Roxo, Hugo Proença, and Pedro RM Inácio. How deep learning sees the world: A survey on adversarial attacks & defenses. *IEEE Access*, 2024.

[Delmerico *et al.*, 2019] Jeffrey Delmerico, Stefano Mintchev, Alessandro Giusti, Boris Gromov, Kamilo Melo, Tomislav Horvat, Cesar Cadena, Marco Hutter, Auke Ijspeert, Dario Floreano, et al. The current state and future outlook of rescue robotics. *Journal of Field Robotics*, 36(7):1171–1191, 2019.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[Doyle, 1982] John Doyle. Analysis of feedback systems with structured uncertainties. In *IEE Proceedings D-Control Theory and Applications*, 1982.

[Duchi and Namkoong, 2020] John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv eprint 1810.08750*, 2020.

[Eder *et al.*, 2014] Kerstin Eder, Chris Harper, and Ute Leonards. Towards the safety of human-in-the-loop robotics: Challenges and opportunities for safety assurance of robotic co-workers'. In *RO-MAN*, pages 660–665. IEEE, 2014.

[Escontrela *et al.*, 2023] Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and Pieter Abbeel. Video prediction models as rewards for reinforcement learning. *NeurIPS*, 2023.

[Eysenbach *et al.*, 2018] Benjamin Eysenbach, Shixiang Gu, Julian Ibarz, and Sergey Levine. Leave no trace: Learning to reset for safe and autonomous reinforcement learning. In *ICLR*, 2018.

[Fang *et al.*, 2023] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A robotic dataset for learning diverse skills in one-shot. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[Ferraguti *et al.*, 2022] Federica Ferraguti, Chiara Talignani Landi, Andrew Singletary, Hsien-Chung Lin, Aaron Ames, Cristian Secchi, and Marcello Bonfè. Safety and efficiency in robotics: The control barrier functions ap-

proach. *IEEE Robotics & Automation Magazine*, pages 139–151, 2022.

[Firoozi *et al.*, 2023] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *IJRR*, page 02783649241281508, 2023.

[Garg *et al.*, 2024] Kunal Garg, Jacob Arkin, Songyuan Zhang, Nicholas Roy, and Chuchu Fan. Large language models to the rescue: Deadlock resolution in multi-robot systems. *arXiv preprint arXiv:2404.06413*, 2024.

[Garnelo *et al.*, 2018a] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *ICML*, pages 1704–1713. PMLR, 2018.

[Garnelo *et al.*, 2018b] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

[Genesis-Authors, 2024] Genesis-Authors. Genesis: A universal and generative physics engine for robotics and beyond, December 2024.

[Guiochet *et al.*, 2017] Jérémie Guiochet, Mathilde Machin, and Hélène Waeselynck. Safety-critical advanced robots: A survey. *Robotics and Autonomous Systems*, 94:43–52, 2017.

[Ha *et al.*, 2024] Huy Ha, Yihuai Gao, Zipeng Fu, Jie Tan, and Shuran Song. Umi on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. *arXiv preprint arXiv:2407.10353*, 2024.

[Haddadin *et al.*, 2007] Sami Haddadin, Alin Albu-Schäffer, and Gerd Hirzinger. Safety evaluation of physical human-robot interaction via crash-testing. In *Robotics: Science and systems*, pages 217–224, 2007.

[Haddadin *et al.*, 2008] Sami Haddadin, Alin Albu-Schaffer, Alessandro De Luca, and Gerd Hirzinger. Collision detection and reaction: A contribution to safe physical human-robot interaction. In *IROS*, pages 3356–3363. IEEE, 2008.

[Hansel *et al.*, 2023] Kay Hansel, Julen Urain, Jan Peters, and Georgia Chalvatzaki. Hierarchical policy blending as inference for reactive robot control. In *ICRA*, pages 10181–10188, 2023.

[Hossain, 2023] Mokter Hossain. Autonomous delivery robots: A literature review. *IEEE Engineering Management Review*, 51(4):77–89, 2023.

[Hu *et al.*, 2023] Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Hao-Shu Fang, et al. Toward general-purpose robots via foundation models: A survey and meta-analysis. *arXiv preprint arXiv:2312.08782*, 2023.

[Huang *et al.*, 2024] Tao Huang, Guangqi Jiang, Yanjie Ze, and Huazhe Xu. Diffusion reward: Learning rewards via conditional video diffusion. In *ECCV*, pages 478–495. Springer, 2024.

[Huck *et al.*, 2023] Tom P. Huck, Martin Kaiser, Constantin Cronrath, Bengt Lennartson, Torsten Kröger, and Tamim Asfour. Reinforcement learning for safety testing: Lessons from a mobile robot case study, 2023.

[Jiao *et al.*, 2024] Ruochen Jiao, Shaoyuan Xie, Justin Yue, Takami Sato, Lixu Wang, Yixuan Wang, Qi Alfred Chen, and Qi Zhu. Exploring backdoor attacks against large language model-based decision making. *arXiv preprint arXiv:2405.20774*, 2024.

[Kanazawa *et al.*, 2023] Naoaki Kanazawa, Kento Kawaharazuka, Yoshiki Obinata, Kei Okada, and Masayuki Inaba. Recognition of heat-induced food state changes by time-series use of vision-language model for cooking robot. In *ICoIAS*, 2023.

[Kargar *et al.*, 2024] Seyed Mohamad Kargar, Borislav Yordanov, Carlo Harvey, and Ali Asadipour. Emerging trends in realistic robotic simulations: A comprehensive systematic literature review. *IEEE Access*, 12:191264–191287, 2024.

[Karnik *et al.*, 2024] Sathwik Karnik, Zhang-Wei Hong, Nishant Abhangi, Yen-Chen Lin, Tsun-Hsuan Wang, and Pulkit Agrawal. Embodied red teaming for auditing robotic foundation models. *arXiv preprint arXiv:2411.18676*, 2024.

[Katara *et al.*, 2024] Pushkal Katara, Zhou Xian, and Katerina Fragkiadaki. Gen2sim: Scaling up robot learning in simulation with generative models. In *ICRA*, pages 6672–6679. IEEE, 2024.

[Khazatsky *et al.*, 2024] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[Kim *et al.*, 2024] Woojun Kim, Yongjae Shin, Jongeui Park, and Youngchul Sung. Sample-efficient and safe deep reinforcement learning via reset deep ensemble agents. *NeurIPS*, 36, 2024.

[Kingma, 2013] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kitamura *et al.*, 2025] Toshinori Kitamura, Tadashi Kozuno, Wataru Kumagai, Kenta Hoshino, Yohei Hosoe, Kazumi Kasaura, Masashi Hamaya, Paavo Parmas, and Yutaka Matsuo. Near-Optimal Policy Identification in Robust Constrained Markov Decision Processes via Epigraph Form. In *ICLR*, 2025.

[Kolve *et al.*, 2017] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[La Valle, 2011] Steven M. La Valle. Motion planning. *IEEE Robotics & Automation Magazine*, 18(2):108–118, 2011.

[Lee *et al.*, 2009] Woosub Lee, Junho Choi, and Sungchul Kang. Spring-clutch: A safe torque limiter based on a spring and cam mechanism with the ability to reinitialize its position. In *IROS*, pages 5140–5145, 2009.

[Li *et al.*, 2024] Chuan-Yao Li, Fan Zhang, and Liang Chen. Robot-assisted pedestrian evacuation in fire scenarios based on deep reinforcement learning. *Chinese Journal of Physics*, 92:494–531, 2024.

[Liang *et al.*, 2023] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *ICRA*, pages 9493–9500. IEEE, 2023.

[Lin *et al.*, 2024] Lizhi Lin, Honglin Mu, Zenan Zhai, Minghan Wang, Yuxia Wang, Renxi Wang, Junjie Gao, Yixuan Zhang, Wanxiang Che, Timothy Baldwin, et al. Against the achilles' heel: A survey on red teaming for generative models. corr abs/2404.00629 (2024), 2024.

[Liu *et al.*, 2023] Huihan Liu, Soroush Nasiriany, Lance Zhang, Zhiyao Bao, and Yuke Zhu. Robot learning on the job: Human-in-the-loop autonomy and learning during deployment. In *Robotics: Science and Systems (RSS)*, 2023.

[Liu *et al.*, 2024a] Aishan Liu, Yuguang Zhou, Xianglong Liu, Tianyuan Zhang, Siyuan Liang, Jiakai Wang, Yanjun Pu, Tianlin Li, Junqi Zhang, Wenbo Zhou, et al. Compromising embodied agents with contextual backdoor attacks. *arXiv preprint arXiv:2408.02882*, 2024.

[Liu *et al.*, 2024b] Haokun Liu, Yaonan Zhu, Kenji Kato, Atsushi Tsukahara, Izumi Kondo, Tadayoshi Aoyama, and Yasuhisa Hasegawa. Enhancing the llm-based robot manipulation through human-robot collaboration. *RA-L*, 9(8):6904–6911, 2024.

[Liu *et al.*, 2024c] Huihan Liu, Shivin Dass, Roberto Martín-Martín, and Yuke Zhu. Model-based runtime monitoring with interactive imitation learning. In *ICRA*, 2024.

[Liu *et al.*, 2024d] 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, 2024.

[Lu *et al.*, 2024] Xuancun Lu, Zhengxian Huang, Xinfeng Li, Wenyuan Xu, et al. Poex: Policy executable embodied ai jailbreak attacks. *arXiv preprint arXiv:2412.16633*, 2024.

[Luo *et al.*, 2024] Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning, 2024.

[Mahler *et al.*, 2019] Jeffrey Mahler, Matthew Matl, Vishal Satish, Michael Danielczuk, Bill DeRose, Stephen McKinley, and Ken Goldberg. Learning ambidextrous robot grasping policies. *Science Robotics*, 4(26):eaau4984, 2019.

[Mandlekar *et al.*, 2018] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *CoRL*, pages 879–893. PMLR, 2018.

[Mankowitz *et al.*, 2020] Daniel J Mankowitz, Dan A Calian, Rae Jeong, Cosmin Paduraru, Nicolas Heess, Sumanth Dathathri, Martin Riedmiller, and Timothy Mann. Robust Constrained Reinforcement Learning for Continuous Control with Model Misspecification. *arXiv preprint arXiv:2010.10644*, 2020.

[Matsushima *et al.*, 2020a] Tatsuya Matsushima, Hiroki Furuta, Yutaka Matsuo, Ofir Nachum, and Shixiang Gu. Deployment-efficient reinforcement learning via model-based offline optimization. *arXiv preprint arXiv:2006.03647*, 2020.

[Matsushima *et al.*, 2020b] Tatsuya Matsushima, Naruya Kondo, Yusuke Iwasawa, Kaoru Nasuno, and Yutaka Matsuo. Modeling task uncertainty for safe meta-imitation learning. *Frontiers in Robotics and AI*, 7:606361, 2020.

[Matthias and Reisinger, 2016] Björn Matthias and Thomas Reisinger. Example application of iso/ts 15066 to a collaborative assembly scenario. In *Proceedings of ISR 2016: 47st international symposium on robotics*, pages 1–5. VDE, 2016.

[Mittal *et al.*, 2023] Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *RA-L*, 8(6):3740–3747, 2023.

[Moos *et al.*, 2022] Janosch Moos, Kay Hansel, Hany Abdulsamad, Svenja Stark, Debora Clever, and Jan Peters. Robust Reinforcement Learning: A Review of Foundations and Recent Advances. *Machine Learning and Knowledge Extraction*, 4(1):276–315, 2022.

[Moroncelli *et al.*, 2024] Angelo Moroncelli, Vishal Soni, Asad Ali Shahid, Marco Maccarini, Marco Forgione, Dario Piga, Blerina Spahiu, and Loris Roveda. Integrating reinforcement learning with foundation models for autonomous robotics: Methods and perspectives. *arXiv preprint arXiv:2410.16411*, 2024.

[O'Neill *et al.*, 2023] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[Park *et al.*, 2024] Daehee Park, Jaeseok Jeong, Sung-Hoon Yoon, Jaewoo Jeong, and Kuk-Jin Yoon. T4p: Test-time training of trajectory prediction via masked autoencoder and actor-specific token memory. *arXiv 2403.10052*, 2024.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[Peters *et al.*, 2015] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *arXiv eprint 1501.01332*, 2015.

[Piriyajitakonkij *et al.*, 2024] Maytus Piriyajitakonkij, Mingfei Sun, Mengmi Zhang, and Wei Pan. Tta-nav: Test-time adaptive reconstruction for point-goal navigation under visual corruptions. *arXiv 2403.01977*, 2024.

[Puig *et al.*, 2023] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023.

[Rohmer *et al.*, 2013] Eric Rohmer, Surya P. N. Singh, and Marc Freese. V-rep: A versatile and scalable robot simulation framework. In *IROS*, pages 1321–1326, 2013.

[Russel *et al.*, 2020] Reazul Hasan Russel, Mouhacine Benosman, and Jeroen Van Baar. Robust Constrained-MDPs: Soft-Constrained Robust Policy Optimization under Model Uncertainty. *arXiv preprint arXiv:2010.04870*, 2020.

[Saveriano *et al.*, 2023] Matteo Saveriano, Fares J Abu-Dakka, Aljaž Kramberger, and Luka Peternel. Dynamic movement primitives in robotics: A tutorial survey. *IJRR*, 42(13):1133–1184, 2023.

[Schmidgall *et al.*, 2024] Samuel Schmidgall, Ji Woong Kim, Alan Kuntz, Ahmed Ezzat Ghazi, and Axel Krieger. General-purpose foundation models for increased autonomy in robot-assisted surgery. *Nature Machine Intelligence*, pages 1–9, 2024.

[Shah *et al.*, 2023] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. Gnm: A general navigation model to drive any robot. In *ICRA*, pages 7226–7233, 2023.

[Shirasaka *et al.*, 2024] Mimo Shirasaka, Tatsuya Matsushima, Soshi Tsunashima, Yuya Ikeda, Aoi Horo, So Ikoma, Chikaha Tsuji, Hikaru Wada, Tsunekazu Omija, Dai Komukai, et al. Self-recovery prompting: Promptable general purpose service robot system with foundation models and self-recovery. In *ICRA*, pages 17395–17402. IEEE, 2024.

[Sugiura *et al.*, 2024] Sojiro Sugiura, Jayant Unde, Yaonan Zhu, and Yasuhisa Hasegawa. Variable grounding flexible limb tracking center of gravity for sit-to-stand transfer assistance. *RA-L*, 9(1):175–182, 2024.

[Sun *et al.*, 2018] Guanghui Sun, Ligang Wu, Zhian Kuang, Zhiqiang Ma, and Jianxing Liu. Practical tracking control of linear motor via fractional-order sliding mode. *Automatica*, 94:221–235, 2018.

[Sun *et al.*, 2020] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Morit Hardt. Test-time training with self-supervision for generalization under distribution shifts. *arXiv 1909.13231*, 2020.

[Sun *et al.*, 2024a] Fan-Yun Sun, SI Harini, Angela Yi, Yihan Zhou, Alex Zook, Jonathan Tremblay, Logan Cross, Jiajun Wu, and Nick Haber. Factorsim: Generative simulation via factorized representation. In *NeurIPS*, 2024.

[Sun *et al.*, 2024b] Jianwei Sun, Erik Harrison Kramer, and Jacob Rosen. A safety-focused admittance control approach for physical human–robot interaction with rigid multi-arm serial link exoskeletons. *IEEE/ASME Transactions on Mechatronics*, pages 1–12, 2024.

[Sutton and Barto, 2018] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[Truby *et al.*, 2019] Ryan L. Truby, Robert K. Katzschmann, Jennifer A. Lewis, and Daniela Rus. Soft robotic fingers with embedded ionogel sensors and discrete actuation modes for somatosensitive manipulation. In *RoboSoft*, pages 322–329, 2019.

[Wang *et al.*, 2021] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv 2006.10726*, 2021.

[Wang *et al.*, 2024] Xianlong Wang, Hewen Pan, Hangtao Zhang, Minghui Li, Shengshan Hu, Ziqi Zhou, Lulu Xue, Peijin Guo, Yichen Wang, Wei Wan, et al. Trojanrobot: Physical-world backdoor attacks against vlm-based robotic manipulation. *arXiv preprint arXiv:2411.11683*, 2024.

[Xiao *et al.*, 2023] Xuan Xiao, Jiahang Liu, Zhipeng Wang, Yanmin Zhou, Yong Qi, Qian Cheng, Bin He, and Shuo Jiang. Robot learning in the era of foundation models: A survey. *arXiv preprint arXiv:2311.14379*, 2023.

[Yamamoto *et al.*, 2019] Takashi Yamamoto, Koji Terada, Akiyoshi Ochiai, Fuminori Saito, Yoshiaki Asahara, and Kazuto Murase. Development of human support robot as the research platform of a domestic mobile manipulator. *ROBOMECH journal*, 6(1):1–15, 2019.

[Yue *et al.*, 2023] Wenxi Yue, Jing Zhang, Kun Hu, Qiuxia Wu, Zongyuan Ge, Yong Xia, Jiebo Luo, and Zhiyong Wang. Part to whole: Collaborative prompting for surgical instrument segmentation. *arXiv preprint arXiv:2312.14481*, 2023.

[Zacharaki *et al.*, 2020] Angeliki Zacharaki, Ioannis Kostavelis, Antonios Gasteratos, and Ioannis Dokas. Safety bounds in human robot interaction: A survey. *Safety Science*, 127:104667, 2020.

[Zames, 1981] George Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Transactions on Automatic Control*, 26(2):301–320, 1981.

[Zawalski *et al.*, 2024] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

[Zhang *et al.*, 2024] Chong Zhang, Nikita Rudin, David Hoeller, and Marco Hutter. Learning agile locomotion on risky terrains. In *IROS*, pages 11864–11871. IEEE, 2024.

[Zhao *et al.*, 2020] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *SSCI*, pages 737–744. IEEE, 2020.

[Zhao *et al.*, 2024a] Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward H Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. *arXiv preprint arXiv:2406.13640*, 2024.

[Zhao *et al.*, 2024b] Ke Zhao, Huayang Huang, Miao Li, and Yu Wu. Rethinking the intermediate features in adversarial attacks: Misleading robotic models via adversarial distillation. *arXiv preprint arXiv:2411.15222*, 2024.

[Zhou *et al.*, 2024a] 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. *arXiv preprint arXiv:2412.04455*, 2024.

[Zhou *et al.*, 2024b] Zhehua Zhou, Jiayang Song, Xuan Xie, Zhan Shu, Lei Ma, Dikai Liu, Jianxiong Yin, and Simon See. Towards building ai-cps with nvidia isaac sim: An industrial benchmark and case study for robotics manipulation. In *ICSE-SEIP*, page 263–274, 2024.

[Zhu *et al.*, 2019] Yaonan Zhu, Takayuki Ito, Tadayoshi Aoyama, and Yasuhisa Hasegawa. Development of sense of self-location based on somatosensory feedback from finger tips for extra robotic thumb control. *Robomech Journal*, 6:1–10, 2019.

[Zhu *et al.*, 2022] Yaonan Zhu, Jacinto Colan, Tadayoshi Aoyama, and Yasuhisa Hasegawa. Cutaneous feedback interface for teleoperated in-hand manipulation. In *IROS*, pages 605–611, 2022.