



Università di Genova

DRIM

Ph.D. Program of National Interest in Robotics and Intelligent Machines

Administrative Headquarters: Università di Genova

Towards Safe Embodied Intelligence: Learning Contact-Rich Manipulation from Impedance Control to Multimodal Foundation Models

by

Heng Zhang

Thesis submitted for the degree of *Doctor of Philosophy* (38° cycle)

April 2026

Dr. Arash Ajoudani

Dr. Yu She

Dr. Antonio Sgorbissa

Supervisor

Co-advisor during visiting

Head of the PhD program

Thesis Jury:

Prof. Heni Ben Amor, *Arizona State University*

Prof. Ding Zhao, *Carnegie Mellon University*

Dr. Lorenzo Natale, *Italian Institute of Technology*

External examiner

External examiner

Internal examiner



Ph.D. program of national interest in Robotics and Intelligent Machines (DRIM)



Human-Robot Interfaces and Interaction Lab (HRI² Lab)

To my family, for their unconditional love and endless support throughout this journey.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Heng Zhang
June 2026

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Arash Ajoudani, for his invaluable guidance, unwavering support, and insightful feedback throughout my PhD. It was Dr. Ajoudani who truly lit my passion and enthusiasm for robotics and AI — his vision, creativity, and genuine love for science are contagious in the best possible way. He has always been encouraging and supportive of my pursuit of bold and unconventional ideas, never letting me settle for the easy path when a more ambitious one was within reach. His advice has shaped not only the direction of my research but also my identity as a scientist. I am deeply grateful for the intellectual freedom he granted me to explore new directions, while always being there with the right guidance at the right moment.

I would like to express my sincere gratitude to Prof. Yu She at Purdue University for hosting me as a visiting scholar and for his exceptional mentorship during my time there. The visiting period at Purdue was an intellectually enriching and transformative experience that significantly broadened my research perspective. I am deeply grateful to Prof. She for the stimulating discussions, the collaborative spirit, and the warm welcome extended by his entire research group. The work on tactile sensing and the TacVLA project was greatly shaped by the insights and expertise I gained during this visit. I would also like to thank all the members of Prof. She's lab at Purdue for their friendship, technical discussions, and for making my stay both productive and enjoyable.

I would like to thank my colleagues at the HRI² Lab at IIT. In particular, I am grateful to Dr. Gokhan Solak, whose collaboration and technical discussions were instrumental in shaping the SRL-VIC, Bresa, and Passivity-Centric Safe RL works, etc. His deep knowledge of variable impedance control and safe robot learning has been invaluable throughout this journey. I would also like to thank Dr. Gustavo JG Lahr, Sebastian Hjorth, Qiyi Tong, and Youyuan Long for their contributions and all other current and former members of the HRI² lab for their camaraderie, valuable discussions, and support throughout my doctoral studies.

I am grateful to the collaborators who contributed to the research presented in this thesis, some of them from Purdue, Humanoids Lab at iit, and Germany: they are Pokuang Zhou,

Haddy Alchaer, Kaidi Zhang, Rui Dai, Puze Liu and the co-authors. Their expertise and contributions significantly enriched the research and broadened its scope.

I would like to acknowledge the financial support from the European Union's Horizon, IIT, and program of National Interest in Robotics and Intelligent Machines (DRIM), which made this research possible. I am also grateful to Purdue University for supporting my visiting research period.

I am deeply grateful to my family for their unconditional love and support throughout my doctoral studies. Their encouragement, patience, and belief in me have been a constant source of strength during the challenging moments of this journey. Finally, I would like to thank my friends and colleagues, both in Genova and at Purdue, for the memorable experiences, the stimulating discussions, and for making this journey not only rewarding but also truly enjoyable.

Last but not least, I would like to express my sincere gratitude to the researchers in robotics and AI whose work has continually inspired and motivated me. I am equally thankful for the open-source community and the spirit of sharing that defines it. The collective efforts of this scientific community have been a constant source of inspiration, pushing me to explore new ideas and expand the boundaries of knowledge in this exciting field.

Abstract

Safe physical interaction remains one of the key barriers to deploying robots in unstructured real-world environments. Contact-rich tasks such as insertion, assembly, tool use, and fine manipulation require robots not only to complete a task, but to do so while regulating force, adapting to uncertainty, and avoiding damage to themselves, their surroundings, and the objects they handle. This PhD thesis addresses that challenge by developing approaches for safe learning in contact-rich robotic tasks, spanning classical compliant control, safe reinforcement learning, and emerging foundation-model-based manipulation.

The thesis first establishes the problem landscape through a comprehensive survey of safe learning for contact-rich robotics, organizing prior work into safe exploration and safe execution. Building on this foundation, it introduces SRL-VIC, a variable stiffness-based safe reinforcement learning framework that combines a safety critic, a recovery policy, and variable impedance control to enable safer exploration in unknown environments. It then presents Bresa, a bio-inspired reflexive safe reinforcement learning approach that decouples task learning from safety learning and enables rapid intervention through a high-frequency safety critic. To address the instability of conventional reinforcement learning policies in physical interaction, the thesis further proposes a passivity-centric safe reinforcement learning method that enforces control stability via passivity theory while improving energy efficiency through passivity-aware training.

Beyond reinforcement learning, this PhD thesis investigates how high-level semantic reasoning can be coupled with low-level compliant control to unlock safer and more adaptive manipulation. To this end, it proposes OmniVIC, a vision-language-model-enhanced variable impedance controller that uses retrieval-augmented generation and in-context learning to generate adaptive impedance parameters. It further introduces a CompliantVLA-adaptor that augments state-of-the-art vision-language-action (VLA) models with VLM-guided compliant control, and finally presents TacVLA, a tactile-enhanced VLA framework that integrates compact tactile sensing into a transformer-based policy through a contact-aware gating mechanism for robust fine-grained manipulation.

Extensive simulation and real-world experiments show that the proposed methods consistently improve task success, reduce unsafe force violations, and enhance the robustness of physical interaction. Collectively, the contributions of this PhD thesis advance the state of the art in safe contact-rich robotic learning and provide a pathway toward reliable, safety-aligned, and foundation-model-enabled robots that can operate with improved operational capability, adaptability, and trustworthiness in complex real-world settings.

Table of contents

List of figures	xiii
List of tables	xxi
1 Introduction	1
1.1 Motivation	1
1.2 Research Challenges	2
1.2.1 Safe Exploration in Contact-Rich Tasks	2
1.2.2 Safe Execution with Variable Impedance Control	2
1.2.3 Passivity and Energy Management	3
1.2.4 Foundation Models for Contact-Rich Manipulation	3
1.3 Research Contributions	3
1.4 List of Publications	5
1.5 Thesis Outline	7
2 Background and Related Work	9
2.1 Introduction to Contact-Rich Robotic Manipulation	9
2.2 Definition of Safety in Contact-Rich Tasks	11
2.3 Problem Formulation and Learning Paradigm	12
2.3.1 Problem Setting	12
2.3.2 Learning Paradigm	13
2.4 Categorization of Methods in Related Works	14
2.4.1 Safe Exploration Methods	15
2.4.2 Safe Execution Methods	16
2.4.3 Foundation Models in Contact-Rich Tasks	17
2.4.4 Safety Enforcement Spaces	18
2.4.5 Sensing Modalities for Safe Contact-Rich Learning	21

2.4.6	Safety Evaluation Metrics	23
2.4.7	Application Domains	24
2.5	Key Challenges	26
2.5.1	Fragile Safety and Stability Guarantees	26
2.5.2	Data Scarcity and Biased Supervision	26
2.5.3	Limited Robustness and Generalisation	27
2.5.4	Safe Exploration Under Physical Risk	27
2.5.5	Safe Execution and Performance Trade-offs	27
2.5.6	Benchmark and Dataset Limitations	27
2.5.7	Safety Risks Specific to VLMs and VLAs	28
2.5.8	Human-Robot Collaboration Safety	28
2.6	Positioning of This Thesis	28
2.7	Summary	29
3	Safe Reinforcement Learning with Variable Impedance Control for Contact-Rich Robotic Tasks	37
3.1	Introduction	37
3.2	Approach	39
3.2.1	Impedance Control in Task Space	39
3.2.2	Safe RL for Contact-Rich Tasks	41
3.3	Evaluation	43
3.3.1	Experiment Setup	43
3.3.2	Simulation Experiments	44
3.3.3	Real-World Experiments	47
3.4	Summary	52
4	Reflexive Safe Reinforcement Learning for Contact-Rich Robotic Tasks	53
4.1	Introduction	53
4.2	Method: Bio-inspired Hierarchical Reflexive Safe RL	55
4.2.1	Safe RL Problem Formulation	55
4.2.2	Hierarchical Safe RL Baseline	56
4.2.3	Reflex Mechanism	56
4.2.4	Trajectory Control and VIC	57
4.3	Evaluation	58
4.3.1	Navigation Task	58
4.3.2	Maze Exploration Task	59

4.3.3	Model Test	61
4.3.4	Parameter Studies	62
4.3.5	Real-World Experiments	63
4.4	Summary	66
5	Passivity-Centric Safe Reinforcement Learning for Contact-Rich Robotic Tasks	67
5.1	Introduction	67
5.2	Method: Passive Safe Reinforcement Learning	69
5.2.1	Safe RL for Contact-Rich Tasks	69
5.2.2	Cartesian Impedance Control	71
5.2.3	Passivity-Based Approach for Variable Stiffness Control	72
5.2.4	Passivity-Aware Safe RL	73
5.2.5	Passivity-Ensured Control	74
5.3	Evaluation	75
5.3.1	Passivity-Aware Policy Training	75
5.3.2	Model Tests	77
5.3.3	Passivity-Ensured Deployment	78
5.3.4	Real-World Experiments	79
5.4	Discussion	80
5.5	Summary	82
6	A Self-Improving Variable Impedance Controller with Vision-Language In-Context Learning	83
6.1	Introduction	83
6.2	Methodology	85
6.2.1	VLM-Informed Variable Impedance Control	85
6.2.2	RAG Data Acquisition and Storage	86
6.2.3	RAG Retrieval	87
6.2.4	ICL-Enhanced Impedance Parameter Prediction	87
6.3	Experiments	88
6.3.1	Simulation Experiments	88
6.3.2	Simulation Results	89
6.3.3	Real-World Experiments	91
6.4	Summary	93

7	VLM-Guided Variable Impedance Action for Safe Contact-Rich Manipulation	94
7.1	Introduction	94
7.2	Methodology	96
7.2.1	Problem Formulation	96
7.2.2	CompliantVLA-Adaptor	98
7.2.3	Hybrid VLA-VIC Control Architecture	99
7.3	Experiments	100
7.3.1	Simulation Tasks	100
7.3.2	Simulation Results	101
7.3.3	Real-World Experiments	102
7.3.4	Real-World Results	103
7.4	Discussion and Limitations	104
7.5	Summary	105
8	Contact-Aware Tactile Fusion for Robust Vision-Language-Action Manipulation	107
8.1	Introduction	107
8.2	Methodology	109
8.2.1	Overall Architecture	109
8.2.2	Contact-Aware Gating Tactile Fusion	111
8.2.3	Training Procedure	111
8.3	Experiments	112
8.3.1	Experimental Setup	112
8.3.2	Data Collection	113
8.4	Results and Analysis	114
8.4.1	Main Results	114
8.4.2	Robustness Evaluation	116
8.4.3	Ablation Study	117
8.5	Discussion and Limitations	119
8.6	Summary	119
9	Conclusions and Future Work	120
9.1	Summary of Contributions	120
9.1.1	Safe Reinforcement Learning with Variable Impedance Control (SRL-VIC)	120
9.1.2	Bio-Inspired Reflexive Safe RL (Bresa)	121
9.1.3	Passivity-Centric Safe RL	121

9.1.4	OmniVIC: VLM-Guided Variable Impedance Control	121
9.1.5	CompliantVLA: VLA with Compliant Control	122
9.1.6	TacVLA: Tactile-Enhanced VLA	122
9.2	Connections Between Contributions	123
9.3	Limitations	123
9.4	Future Research Directions	124
9.4.1	Scalable Safe Learning for Contact-Rich Tasks	124
9.4.2	Foundation Models with Physical Intelligence	124
9.4.3	Multimodal Contact Sensing	124
9.4.4	Safe Deployment in Human-Robot Collaboration	124
9.4.5	Benchmarks and Evaluation	125
9.5	Concluding Remarks	125
References		126
Appendix A Simulation Environment and Implementation Details		145
A.1	MuJoCo Simulation Setup	145
A.1.1	Robot Model	145
A.1.2	Maze Environment	146
A.2	Simulation Parameters	146
A.3	Real-World Robot Setup	146
A.4	Neural Network Architectures	147
A.5	Hyperparameter Sensitivity Analysis	147
A.5.1	SRL-VIC Hyperparameters	147
A.5.2	Bresa Hyperparameters	148
A.5.3	Passivity-Centric Safe RL Hyperparameters	148
A.6	Training Curves and Convergence Analysis	148
A.6.1	SRL-VIC Training Curves	148
A.6.2	Bresa Training Curves	149
A.7	Real-World Experiment Details	149
A.7.1	Cable Routing Task Setup	149
A.7.2	Peg-in-Hole Task Setup	149
A.7.3	Disassembly Task Setup	150
Appendix B Foundation Model Details and Prompts		151
B.1	OmniVIC Prompt Design	151

B.1.1	Task Description Prompt	151
B.1.2	RAG Knowledge Base	152
B.1.3	In-Context Learning Examples	152
B.2	Extended Prompt Examples for OmniVIC	153
B.2.1	Peg-in-Hole Insertion Prompt	153
B.2.2	Cable Routing Prompt	153
B.3	CompliantVLA Adaptor Architecture	154
B.3.1	VLM-based Impedance Generator	154
B.3.2	Force Feedback Module	154
B.4	TacVLA Implementation Details	155
B.4.1	Tactile Sensor Specification	155
B.4.2	Tactile Tokenization	155
B.4.3	Contact-Aware Gating	155
B.5	Training Details for Foundation Model Experiments	156
B.5.1	OmniVIC Training	156
B.5.2	CompliantVLA Training	156
B.5.3	TacVLA Training	156

List of figures

2.1	Common examples of contact-rich robotic tasks: (a) peg insertion or assembly; (b) human assistance (e.g., dressing); (c) tool-use (e.g., vegetable cutting); (d) wiping, polishing or similar; (e) opening doors or drawers. Adapted from [1].	10
2.2	Publication trend showing the annual number of publications related to imitation learning for contact-rich robotic tasks, with a clear growth trend in recent years.	10
2.3	Six-dimensional taxonomy of safe learning methods for contact-rich robotic tasks. The six axes capture the learning paradigm, safety mechanism, enforcement space, sensing modality, task type, and evaluation approach. . . .	11
2.4	Detailed breakdown of the six-dimensional taxonomy showing the distribution of approaches across each dimension in the safe learning literature. . .	12
2.5	Structural overview of safe learning approaches for contact-rich robotic tasks, organised by safety mechanism, enforcement space, and learning paradigm	31
2.6	Illustration of how large language models can be used to generate task descriptions and safety constraints for contact-rich manipulation tasks. . . .	32
2.7	The metrics in the contact-rich safe learning literature can be categorised according to their objectives: safety, task, and efficiency. There are trade-offs between these objectives. The researchers aim to develop safe systems with a good trade-off for task success and efficiency. Some metrics, such as force, velocity and energy serve to measure multiple objectives.	32
2.8	Three-dimensional bubble chart showing the distribution of safe learning works across safety, task, and efficiency metrics.	34
2.9	Three levels of result analyses in safe learning for contact-rich tasks: global (high-level), variational (mid-level), and exemplar (low-level).	34

2.10	Application domains of contact-rich robotic tasks, spanning industrial assembly, household manipulation, and healthcare assistance.	35
2.11	Temporal evolution of application domains in the contact-rich robotic tasks literature.	35
2.12	Comparison of different imitation learning approaches for contact-rich tasks across key performance dimensions, including data efficiency, generalisation, and contact handling.	36
3.1	Maze exploration is an unstructured contact-rich task that has practical applications such as cable-laying in walls. The agent does not have access to vision, thus it navigates using only contact information.	38
3.2	The proposed framework: we combine the recovery-based safe RL approach [2] with VIC to solve the contact-rich maze exploration task. We first use an automated procedure to collect the offline data and pre-train our safety critic and recovery policy. Then, we train all learning components using online data. The risk value ϵ_{risk} is used to activate either the task or recovery policy. The action a_t is chosen by the activated component, and it is fed to the VIC. Our action includes a relative position change and desired stiffness vector $\{K_x, K_y\}$	40
3.3	Trajectories of the end-effector during different stages of <i>training</i> , in top-down view. The orange circle denotes goal point of the task, the cyan circles and yellow arrows along the trajectory represent the stiffness ellipsoids and the motion directions respectively.	45
3.4	Learning curves for maze exploration. Our framework (SRL-VIC) outperforms others, particularly in terms of ratio of successes/violation and cumulative violations, i.e., our framework is safer than others. Furthermore, it learns faster, achieving success at an earlier stage.	46
3.5	Different behaviours when encountering obstacles at training episode 1000. Top (SRL-K300): The robot cannot move forward and fails due to low stiffness. Bottom (SRL-VIC): VIC approach switches to high stiffness when there are some obstacles on the way. The method learns to safely push the obstacles without exceeding the force threshold.	47

- 3.6 Relation between the risk value and state/action values in the offline test. Top: The relationship of the risk ε_{risk} to stiffness $\|K\|$ and desired move $\|\Delta P\|$ values. Bottom: Plot of ε_{risk} and contact force $\|F\|$. The force magnitude is scaled down by 1/100. 48
- 3.7 Real-world experiments. (a) *Maze shape-1*, without obstacles. The yellow overlay shows the original maze from the simulation. (b) The robot meets the pile of bolts halfway. (c) The obstacles create congestion at the curve. (d) The SRL-VIC policy overcomes the obstacles by increasing the stiffness. (e, f) The *complex task scenario* with small wall segments at new positions. (g) *Maze shape-2*. (h) *Maze shape-3*, failed at red X marker. 48
- 3.8 Trajectories of the end-effector during executing task in the real world. The orange circle denotes goal point of the task, the cyan circles and yellow arrows along the trajectory represent the stiffness ellipsoids and the motion directions respectively. (a-c) *Maze shape-1* (Fig.3.7.a): (a) *no obstacle*, (b) *small obstacles*, (c) *mixed obstacles*. (d, e) *Complex task scenario* with discrete walls, (f) *Maze shape-2* (Fig.3.7.g). 50
- 3.9 Force (F_y), move size (ΔP_x), and stiffness (K_y) values during the real-robot experiments plotted against the y- position. The pink highlight shows the approximate placement of obstacles. 51
- 4.1 a) Bresa framework. The RL agent operates at the *decision loop*, planning the high-level action **a** that is executed by the trajectory controller. The controller operates at a high-frequency *control loop*, executing the low-level action $\hat{\mathbf{a}}$ based on the state feedback $\hat{\mathbf{s}}$ at each control step. The reflex mechanism gives the system a quick reaction capability by interrupting the control loop in the case of high risk. b) A simplified illustration of the human central nervous system. While high-level decisions are made in the brain, safety-related reflexes are managed by the spinal cord, allowing for faster responses that override slower, more complex decision-making processes. 54
- 4.2 Flowchart of the Bresa algorithm. We color-coded the *decision loop*, *control loop* and *reflex* for comparison to Fig. 4.1. The safety critic operates at the control loop frequency, enabling real-time reflexive responses. The maze exploration environment in the MuJoCo simulator is shown on the right: the robot physically interacts with the maze walls and the obstacles through an end-effector flange equipped with F/T sensor. 57

4.3	a) Reflex mechanism on an obstacle avoidance scenario. Even when the high-level state-action pair $(\mathbf{s}, \mathbf{a})$ is evaluated to be safe, an intermediate state-action pair $(\hat{\mathbf{s}}, \hat{\mathbf{a}})$ may entail high risk ($\epsilon_{\text{risk}} > \epsilon_{\text{safe}}$) and trigger the reflex mechanism. The stochasticity of the environment leads to a drift in the outcomes of minor actions $\hat{\mathbf{a}}$. c) Maze exploration environment in the Mujoco simulator. The robot physically interacts with the maze walls and the obstacles through an end-effector flange equipped with F/T sensor. . . .	58
4.4	Reflexive mechanism in Navigation task during training. The risk value along with exploration is plotted in a colormap showing fine-grained risk prediction in our method while it is coarse-grained in the baseline. Blue annotations show single high-level actions.	59
4.5	Offline data collection locations in both tasks. Left: navigation task. The green and yellow circles indicate start and goal points, and red dots indicate the sampled start positions. Upper right: histogram of exponentially sampled action sizes in the maze exploration task. Lower right: sampled action locations on the maze.	60
4.6	Overall performance of Bresa in contact-rich maze exploration task in terms of cumulative successes and violations. Bresa significantly outperforms the baseline in both metrics.	61
4.7	Illustration of model performance from a specific episode. Upper: reflexive response of <i>risk critic</i> according to the current state and action. Lower: trajectory of end-effector where the ellipsoids indicate stiffness and red color shows high-risk value.	62
4.8	Success-violation ratio results over navigation task training with different <i>action size</i> a_{max} , <i>noise</i> σ , γ_{safe} and γ_{task} parameters shown in subfigure (a), (b), (c) and (d) respectively.	64
4.9	Real-world setup. Top: Overview of the entire setup. Bottom: Closeups of the robot in action. Left: The robot makes contact with a wall. Center: A stick applies random external forces during motion. Right: The robot interacts with large, stationary obstacles. Bottom: the trajectory of the external-force run shows robot behavior, where the ellipses indicate the stiffness. The red ellipse indicates the recovery action.	65

5.1	<i>Passivity-ensured</i> Safe RL: The proposed method adds both <i>energy-based constraints</i> C_e to enable learning the passivity-awareness and an active <i>passivity-filtering layer</i> between the RL agent and the low-level controller for guaranteed passive control. The signals are the state s , reward r , action a , force constraint c_f , energy-based constraints C_e , and control command u	70
5.2	<i>Passivity-agnostic</i> Safe RL: the baseline RL framework that we extend. The baseline includes the force-based safety constraints c_f in addition to the usual state s , action a and reward r observations of an RL framework.	71
5.3	<i>Passivity-aware</i> Safe RL: We introduce the energy constraints ($c_{eb}, c_{ef} \in C_e$) to enable the RL agent to learn passive behavior.	74
5.4	Experiments setup in MuJoCo. Right top: the robot explores the turn area with contact. Right bottom: three sphere obstacles.	76
5.5	Passivity-aware policy training experiment results. We present the total numbers of successes and violations in the first two plots. The third plot shows the ratio of mean success to mean violation.	76
5.6	These plots visualize the effect that the changes of the stiffnesses $k_{t_x}, k_{t_y} \in \mathbf{K}_T$ during parts of the training have on the energy flow $\dot{E}_T$ and consequently on the energy stored in the energy tank E_T	77
5.7	Plots illustrate the effect that the introduction of the <i>energy tank constraint</i> has on the agent's energy expenditure at different stages during training.	78
5.8	Policy performance in deployment experiments. The bars show how the models (agnostic, Eb4, Eb8, Eb8-Ef7) perform in terms of successes and violations with and without the passivity layer in simulation for 100 episodes.	79
5.9	(a) Energy budget progress during passivity-ensured deployment experiments. The curves indicate the mean of 300 runs (100 per seed) and the shades indicate the standard deviation. (b) Histogram of the run times, i.e., the number of runs that finished in time duration intervals.	80
5.10	Snapshots and result analysis in real-world experiments. (a) the overview of the maze exploration task. (b) closeup when the robot moves along with contact force. (c) failure case agnostic but non-passive behavior at the turnaround area. (d) shows the energy budget consumption among different setups, demonstrating our method outperforms in terms of passivity and energy efficiency.	81

6.1	Overview of the proposed OmniVIC. The system integrates a VLM with a VIC to enable safe and adaptive manipulation in contact-rich tasks. The VLM processes multimodal inputs, including visual observations, natural language instructions, and real-time force/torque feedback, to generate context-aware impedance parameters (stiffness and damping) for the VIC.	85
6.2	Tasks used in simulation experiments.	89
6.3	Simulation results: Task success rates of baseline (π_0 with position control) vs. OmniVIC (with VLM-informed variable impedance control).	90
6.4	Real-world experiments.	91
6.5	Task-1: The robot gently closes the top drawer using OmniVIC variable impedance control. The system adapts stiffness and damping parameters based on visual and language context to ensure safe interaction and avoid excessive force.	92
6.6	Task 2 (top) and Task 3 (bottom): The robot moves along the negative Y-axis while maintaining the same height in the Z-axis. The shaded regions indicate the contact phase as identified by OmniVIC.	92
7.1	Existing VLA systems lack force-awareness, leading to unsafe interactions in physical tasks involving contact or uncertainty. CompliantVLA-adaptor addresses this challenge by augmenting VLA models with VLM-guided variable impedance control.	95
7.2	Overview of the CompliantVLA-adaptor. A VLM processes visual observations, language instructions, and real-time force feedback $\mathcal{F}$ to generate context-aware impedance parameters $\mathcal{K}, \mathcal{D}$. These parameters modulate a variable impedance controller (VIC) that executes actions generated by a VLA model, ensuring safe and adaptive contact-rich manipulation.	97
7.3	Contact-rich tasks in simulation. Task 1 and Task 2 are from ManiSkill for RDT, while the remaining tasks come from LIBERO for Pi0 and OpenVLA. Together, they form a contact-centric task suite.	100
7.4	Evaluation results of task success rates under contact force constraint across 8 related tasks in simulation. The title of each bar diagram indicates a different setup, where “T” with a value denotes task number, and the suffix “-R”, “-P” and “-O” means using RDT, Pi0 and OpenVLA-of-t model, respectively. Different colors indicate the run with and without our CompliantVLA-adaptor.	102

7.5	Real-world experiment setups. Left: box-pushing setup, where the robot pushes a 3 kg box toward a marked target region. Right: whiteboard-wiping setup, where the robot removes marker traces on the board.	103
7.6	Real-world wiping task execution.	103
7.7	Wiping task: adaptive stiffness regulation and measured contact force. Background colors indicate task phases.	104
7.8	Real-world pushing task execution.	104
7.9	Pushing task: adaptive stiffness regulation and measured contact force. Background colors indicate task phases.	105
8.1	Overview of TacVLA. (a) Input modalities including visual observations, language instructions, and tactile measurements. (b) TacVLA architecture, consisting of modality-specific encoders and tokenizer, a pretrained VLM backbone, an action expert, and the contact-aware gating module. (c) The proposed contact-aware gating module that selectively activates tactile tokens based on the contact state, enabling adaptive multimodal fusion during contact-rich manipulation. (d) Experimental evaluation on contact-rich constraint-locked disassembly and in-box picking tasks, together with robustness tests under camera occlusion and human disturbance.	109
8.2	Hardware setup: a 7 DoF Franka robotic platform equipped with tactile sensors and two cameras for visual input to evaluate TacVLA on contact-rich manipulation tasks.	110
8.3	Four contact-rich constraint-locked disassembly tasks with diverse geometric constraints: (a) Task 1: tight shaft; (b) Task 2: press clip; (c) Task 3: shaft rotation; (d) Task 4: slide pull.	113
8.4	The real-world experimental setup and procedures for four constraint-locked disassembly and in-box picking tasks. The experiments demonstrate the capability of TacVLA in contact-rich fine-grained manipulation, as well as its robustness to visual occlusion in the in-box picking scenario.	114
8.5	Robustness Evaluation: the performance of TacVLA under conditions of visual occlusion and runtime disturbance, demonstrating its ability to adapt and maintain performance in challenging scenarios.	116

8.6	Success rates under block camera disturbance across four constraint-locked disassembly tasks. We compare finetuned Pi0.5, finetuned Pi0.5 with tactile input (w/o gating), and TacVLA (ours). Our method consistently improves performance under visual occlusion.	117
8.7	Failure cases of <i>Finetuned Pi0.5 + Tactile (w/o Gating)</i> during constraint-locked disassembly. Without contact-aware gating, tactile tokens remain active even in non-contact phases, leading to unstable behaviors.	118

List of tables

2.1	Sensing and policy modalities for safe learning in contact-rich tasks	33
4.1	Model performance in maze exploration task	61
4.2	Performance in real world	63
5.1	Comparison of passivity modes in RL	69
5.2	Results of model tests. Performance evaluation over 100 episodes for each of three randomly selected seeds.	77
6.1	Tasks in simulation experiments	88
7.1	List of tasks in simulation	101
8.1	Performance comparison across Constraint-locked Disassembly and In-Box Picking tasks	115
8.2	Ablation study on the effect of the contact-aware gating module	118
A.1	Simulation parameters for maze exploration experiments	146
A.2	Neural network architectures for RL experiments	147
A.3	SRL-VIC hyperparameter sensitivity analysis	147
A.4	Bresa hyperparameter sensitivity analysis	148
A.5	Passivity-Centric Safe RL hyperparameter sensitivity	148

Chapter 1

Introduction

1.1 Motivation

The next generation of robots will not be judged only by how precisely they move in free space, but by how safely and effectively they can make contact with the physical world. Many societally relevant applications, such as assembly, insertion, cable routing, surface treatment, healthcare assistance, and tool use, fundamentally require sustained interaction with objects, surfaces, or humans [3, 4]. These are not tasks that can be solved by avoiding contact; they demand that contact be created, regulated, and exploited intelligently.

This requirement immediately exposes a central tension in modern robot learning. On the one hand, learning-based methods offer the adaptability needed for unstructured and dynamic environments, where precise analytical models are often unavailable or too brittle to rely on [5]. On the other hand, the same trial-and-error processes that make learning powerful also make it dangerous in contact-rich settings [6]. A robot that learns by exploration may collide too hard, jam against constraints, damage a workpiece, or behave unpredictably under distribution shift. As a result, contact-rich manipulation turns safety from an auxiliary consideration into a first-order design requirement.

A contact-rich task, as considered in this thesis, is one whose successful completion requires dynamic and sustained physical interaction with the environment, such that motion and force are tightly coupled through contact constraints. This distinguishes contact-rich manipulation from simple pick-and-place tasks, where contact is largely resolved after grasping, and from impact-dominated behaviours such as throwing or catching, where contact is brief. Typical examples include peg-in-hole insertion, polishing, massage, multi-finger manipulation, vegetable cutting, drawer opening, and constrained pushing.

What makes these tasks difficult is the convergence of several challenges. Contact dynamics are discontinuous and hard to model accurately. Interaction forces can quickly become damaging if not regulated. Successful execution requires continuous coordination between motion generation and compliance adaptation. At the same time, the robot must remain safe not only after training, but also during learning itself. In other words, the robot must be able to learn from physical interaction without treating the environment, the hardware, or nearby humans as disposable.

This thesis is motivated by the need to close that gap between capability and safety. Its central premise is that contact-rich intelligence must be both physically grounded and safety-aware: physically grounded because the robot must reason through force, compliance, and contact; safety-aware because these interactions unfold in regimes where failure is costly. To this end, the thesis develops a progression of methods spanning from safe reinforcement learning for contact-rich tasks to foundation-model-based approaches with multimodal sensing. The work is organized around two connected themes: (1) safe learning for contact-rich manipulation, where safety is addressed during both exploration and execution; and (2) foundation-model-based manipulation with physical intelligence, where VLA systems are extended with compliance control and tactile sensing for robust contact-rich behavior.

1.2 Research Challenges

The research challenges addressed in this thesis can be summarized as follows:

1.2.1 Safe Exploration in Contact-Rich Tasks

Standard reinforcement learning (RL) methods explore the environment by trial and error, which can lead to unsafe interactions during contact-rich tasks. The challenge is to develop exploration strategies that are both efficient and safe, minimizing the risk of unsafe actions while still allowing the robot to learn effective policies. Key questions include: How can we define and enforce safety constraints during exploration? How can we balance exploration efficiency with safety? How can we leverage prior knowledge about the task to guide safe exploration?

1.2.2 Safe Execution with Variable Impedance Control

Variable impedance control (VIC) is a powerful framework for contact-rich manipulation, as it allows the robot to adapt its stiffness and damping online to meet task requirements. However,

ensuring safety during execution with VIC is challenging, as the impedance parameters must be carefully tuned to avoid instability and excessive forces. The challenge is to develop methods that can learn safe impedance control policies that are robust to uncertainty and disturbances.

1.2.3 Passivity and Energy Management

Passivity is a fundamental property of physical systems that ensures stability and safety during interaction. In contact-rich manipulation, maintaining passivity is crucial for safe interaction with the environment. The challenge is to develop methods that can enforce passivity constraints during learning and execution, while still allowing the robot to perform complex manipulation tasks.

1.2.4 Foundation Models for Contact-Rich Manipulation

Recent advances in Vision-Language-Action (VLA) models have demonstrated remarkable potential for generalized robotic manipulation. However, these models are primarily designed for free-space manipulation and do not account for the physical constraints and safety requirements of contact-rich tasks. The challenge is to extend VLA models with physical intelligence, including compliance control and tactile sensing, to enable safe and effective contact-rich manipulation.

1.3 Research Contributions

This thesis makes the following contributions:

1. **Safe Reinforcement Learning with Variable Impedance Control (SRL-VIC):** The first main contribution of this thesis is a safe RL framework that integrates variable impedance control with predictive safety assessment for contact-rich manipulation. The key idea is to combine a safety critic and recovery policy with impedance adaptation, so that the robot can explore while explicitly regulating interaction risk rather than treating contact as merely a disturbance. This contribution establishes the foundational safe-learning formulation of the thesis and demonstrates, in both simulation and real-world experiments, that safety and task performance can be improved jointly.
2. **Bio-Inspired Reflexive Safe RL (Bresa):** This introduces a reflexive safety mechanism inspired by biological motor reflexes. Instead of relying only on decision-level safety

checks introduced in the SRL-VIC, Bresa uses a two-level architecture in which a high-level RL policy plans task behaviour while a faster low-level reflex mechanism intervenes during execution when danger emerges. The main contribution here is the shift from predictive safety alone to reflexive safety during action execution, enabling greater robustness under unexpected disturbances.

3. **Passivity-Centric Safe RL:** The third contribution extends safe RL from constraint satisfaction to physical stability by introducing a passivity-centric formulation. This approach uses energy tanks and energy-flow constraints to monitor and regulate how the controller exchanges energy with the environment during interaction. The main contribution is to connect safe reinforcement learning with passivity-based control, showing that contact-rich safety must also include guarantees on energetic behaviour, not only collision avoidance or force thresholds.
4. **OmniVIC: VLM-Guided Variable Impedance Control:** This work marks the transition from safe RL to foundation-model-based control for contact-rich manipulation. OmniVIC integrates Vision-Language Models (VLMs) with variable impedance control, allowing impedance parameters to be generated from language instructions, visual observations, and retrieved prior experience. The main contribution is to show that semantic reasoning from foundation models can be used not only for high-level planning, but also to inform low-level compliant interaction in a flexible and generalizable way.
5. **CompliantVLA: VLA with Compliant Control:** The fifth contribution brings physical compliance directly into the Vision-Language-Action (VLA) paradigm, which is usually ignored in the state-of-the-art VLA models. CompliantVLA augments generalist VLA policies with a lightweight compliance adaptor that transforms rigid position-oriented action generation into safer contact-aware execution. The main contribution is a practical bridge between the generalization capabilities of VLA models and the physical requirements of contact-rich manipulation, enabling safer deployment without redesigning the entire policy architecture.
6. **TacVLA: Tactile-Enhanced VLA:** The further contribution extends VLA-based manipulation beyond vision and language by incorporating tactile sensing into the policy. TacVLA introduces compact tactile tokenization together with a contact-aware gating mechanism, so that tactile feedback is integrated when it is physically informative and suppressed when it is not. The main contribution is to show that robust contact-rich manipulation, especially under visual occlusion and fine-grained

physical interaction, requires not only compliant control but also explicit contact-aware multimodal perception.

7. **Comprehensive Surveys:** The final contribution is the development of two comprehensive surveys on safe learning for robotic contact-rich tasks and on contact-rich imitation learning. These surveys synthesize the state of the art, identify the major open challenges, and provide a structured perspective on where the field is heading. Their main contribution is to position the technical developments of this thesis within a broader research landscape and to clarify how the individual chapters address gaps identified in prior work.

1.4 List of Publications

The research presented in this thesis has been published in or submitted to the following venues:

1. **H. Zhang**, G. Solak, G. J. Lahr, A. Ajoudani, “SRL-VIC: A variable stiffness-based safe reinforcement learning for contact-rich robotic tasks,” *IEEE Robotics and Automation Letters (RA-L)*, 2024. (Chapter 3)
2. **H. Zhang**, G. Solak, S. Hjorth, A. Ajoudani, “Towards Passive Safe Reinforcement Learning: A Comparative Study on Contact-rich Robotic Manipulation,” *IEEE Robotics and Automation Letters (RA-L)*, under review. (Chapter 5)
3. **H. Zhang***, G. Solak*, A. Ajoudani, “Bresa: Bio-inspired Reflexive Safe Reinforcement Learning for Contact-Rich Robotic Tasks,” *IEEE Robotics and Automation Letters (RA-L)*, under review. (Chapter 4)
4. **H. Zhang**, R. Dai, G. Solak, P. Zhou, Y. She, A. Ajoudani, “Safe Learning for Contact-Rich Robot Tasks: A Survey from classical Learning-Based Methods to Safe Foundation Models,” *International Journal of Robotics Research (IJRR)*, under review, 2026. (Chapter 2)
5. **H. Zhang**, W.-H. Huang, G. Solak, A. Ajoudani, “OmniVIC: A Self-Improving Variable Impedance Controller with Vision-Language In-Context Learning for Safe Robotic Manipulation,” *IEEE International Conference on Robotics and Automation (ICRA)*, under review, 2026. (Chapter 6)

6. **H. Zhang**, W.-H. Huang, Q. Tong, G. Solak, P. Liu, K. Zhang, S. Liu, J. Peters, Y. She, A. Ajoudani, “CompliantVLA-adaptor: VLM-Guided Variable Impedance Action for Safe Contact-Rich Manipulation,” *IEEE International Conference on Robotics and Automation (ICRA)*, under review, 2026. (Chapter 7)
7. K. Zhang*, **H. Zhang***, Z. Xu, Z. Zhang, M. R. I. Prince, L. Xiang, X. Han, Y. Zhou, A. Ajoudani, Y. She, “TacVLA: Contact-Aware Tactile Fusion for Robust Vision-Language-Action Manipulation,” *Under Review*, 2026. (Chapter 8)
8. **H. Zhang**, H. Alchaer, A. Ajoudani, Y. She, “Reward-Zero: Language Embedding Driven Implicit Reward Mechanisms for Reinforcement Learning,” *Under Review*, 2026.
9. P. Zhou, Y. Zhou, Q. Luu, S. Han, **H. Zhang**, B. Huang, Y. Li, A. Ajoudani, Z. Xu, Y. She, “Learning Tactile-Aware Quadrupedal Loco-Manipulation Policies,” *Under Review*, 2026.
10. P. Guo, Z. Mai, Z. Xu, K. Zhang, **H. Zhang**, Z. Miao, A. Ajoudani, Z. Kingston, Q. Qiu, Y. She, “AgenticLab: A Real-World Robot Agent Platform that Can See, Think, and Act,” *Robotics: Science and Systems (RSS)*, under review, 2026.
11. Y. Long, G. Solak, S. Zeynalpour, **H. Zhang**, A. Ajoudani, “Self-supervised Physics-Informed Manipulation of Deformable Linear Objects with Non-negligible Dynamics,” *IEEE Transactions on Robotics (T-RO)*, under review, 2026.
12. J. Wang, W. Wang, B. Deng, **H. Zhang**, R. Dai, N. Tsagarakis, “INTENTION: Inferring Tendencies of Humanoid Motion Through Physical Intuition and Grounded VLM,” *IEEE-RAS International Conference on Humanoid Robots*, 2025.
13. S. Liu, Z. Li, W. Wang, H. Sun, **H. Zhang**, H. Chen, Y. Qin, A. Ajoudani, Y. Wang, “ActivePose: Active 6D Object Pose Estimation and Tracking for Robotic Manipulation,” *IEEE International Conference on Robotics and Automation (ICRA)*, under review, 2026.
14. P. Zhang, **H. Zhang**, et al., “aiXiv: A Next-Generation Open Access Ecosystem for Scientific Discovery Generated by AI Scientists,” *The 40th Annual AAAI Conference on Artificial Intelligence*, under review, 2026.

15. H. Wu, **H. Zhang**, P. Zhang, J. Wang, C. Wang, “HiBerNAC: Hierarchical Brain-emulated Robotic Neural Agent Collective for Disentangling Complex Manipulation,” *Under Review*.
16. P. Zhang*, **H. Zhang***, H. Xu, R. Xu, Z. Wang, C. Wang, A. Garg, Z. Li, A. Ajoudani, X. Liu, “Scaling Laws in Scientific Discovery with AI and Robot Scientists,” *Nature Machine Intelligence*, in submission.
17. K. Chen, J. Xiao, J. Liu, Q. Tong, **H. Zhang**, R. Liu, J. Zhang, A. Ajoudani, S. Chen, “Semantic visual simultaneous localization and mapping: A survey,” *IEEE Transactions on Intelligent Transportation Systems*, 2025.

1.5 Thesis Outline

The remainder of this thesis is organized as follows:

Chapter 2 provides a comprehensive literature review covering the background on contact-rich robotic manipulation, safe reinforcement learning, variable impedance control, and foundation models for robotics. We review existing methods, identify their limitations, and motivate the research contributions of this thesis.

Chapter 3 presents the SRL-VIC framework for safe reinforcement learning with variable impedance control. We describe the problem formulation, the proposed method, and experimental results on contact-rich manipulation tasks.

Chapter 4 presents the Bresa framework for bio-inspired reflexive safe reinforcement learning. We describe the bio-inspired reflex mechanism, the two-level control architecture, and experimental results demonstrating robustness to unexpected disturbances.

Chapter 5 presents the passivity-centric safe RL framework. We describe the energy tank formulation, the passivity enforcement mechanism, and experimental results on contact-rich manipulation tasks.

Chapter 6 presents OmniVIC, the VLM-guided variable impedance control framework. We describe the VLM integration, the impedance parameter generation, and experimental results on a variety of contact-rich manipulation tasks.

Chapter 7 presents CompliantVLA, the VLA with compliant control framework. We describe the compliance adaptor, the integration with VLA models, and experimental results on contact-rich manipulation tasks.

Chapter 8 presents TacVLA, the tactile-enhanced VLA framework. We describe the tactile tokenization scheme, the contact-aware gating mechanism, and experimental results on contact-rich manipulation tasks with visual occlusion.

Chapter 9 concludes the thesis by summarizing the contributions, discussing limitations, and outlining future research directions.

Taken together, this chapter has defined the central problem of the thesis and outlined the two main research thrusts pursued in the following chapters: safe reinforcement learning for contact-rich tasks, and foundation-model-based manipulation with physical intelligence. The next chapter reviews the relevant literature and positions these contributions within the broader research landscape.

Chapter 2

Background and Related Work

2.1 Introduction to Contact-Rich Robotic Manipulation

Robots are increasingly expected to operate in unstructured and dynamic environments, where interaction with the physical world is inevitable [3, 4]. Among such interactions, contact-rich tasks such as assembly, insertion, cutting, surface polishing, or tool use are particularly challenging due to their complex, discontinuous dynamics and the need for precise force and motion control. Traditional model-based control methods often struggle in these scenarios due to inaccurate or hard-to-obtain models, while learning-based methods promise flexibility and adaptability by leveraging data-driven representations [5].

In robotics, contact-rich manipulation refers to tasks that involve continuous and complex interactions between the robot and its environment, often requiring sophisticated control of forces at one or several contact points [1]. This is why direct and indirect force control techniques, such as impedance and admittance control, have been widely exploited to address this problem. Pioneering contributions in this field include the formulation of operational space, hybrid and impedance control [7–9], as well as advancements in contact modelling and multi-contact control [10, 11].

Figure 2.1 illustrates common contact-rich robotic tasks spanning industrial assembly, household manipulation, and healthcare assistance.

The publication trend in this area has grown significantly in recent years, as shown in Figure 2.2, reflecting the increasing importance and research interest in contact-rich manipulation.

A major challenge in contact-rich manipulation is the scalability of classical control solutions to varying task conditions and contact dynamics [12]. For example, even a seemingly simple assembly task, such as peg-in-hole, involves varying force or impedance requirements

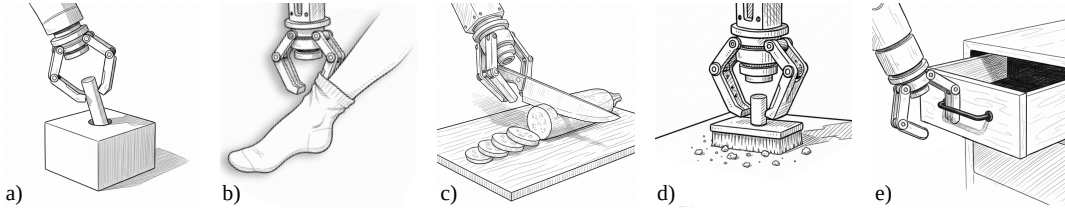


Figure 2.1 Common examples of contact-rich robotic tasks: (a) peg insertion or assembly; (b) human assistance (e.g., dressing); (c) tool-use (e.g., vegetable cutting); (d) wiping, polishing or similar; (e) opening doors or drawers. Adapted from [1].

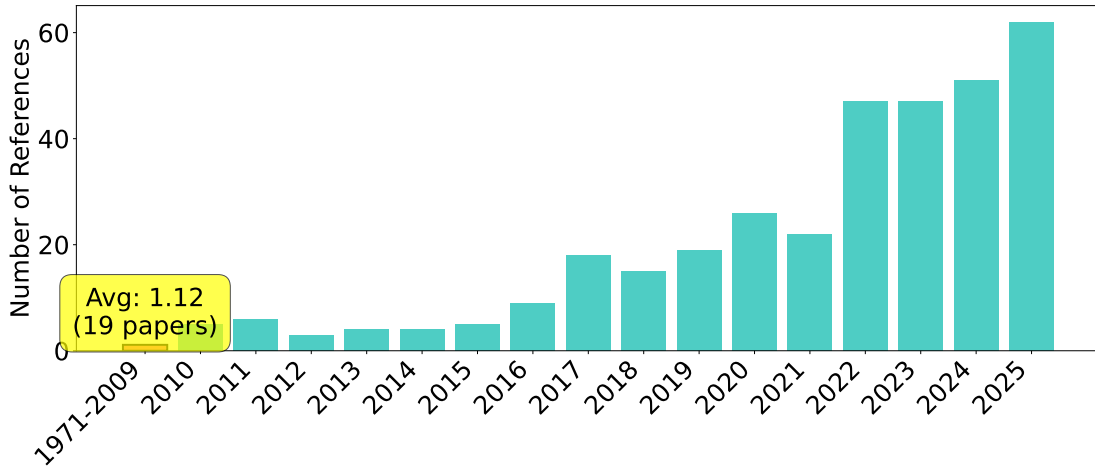


Figure 2.2 Publication trend showing the annual number of publications related to imitation learning for contact-rich robotic tasks, with a clear growth trend in recent years.

depending on factors like the position and orientation of the parts, their material properties, and the clearance between them. Similarly, in tasks that need continuous contact like wiping a surface or polishing, changes in surface geometry, friction, or compliance call for constant adaptation of the contact-related robot dynamics to maintain stable and effective interaction.

Recent advancements in machine learning have accelerated the resolution of this problem, by enabling robots to learn intricate interaction dynamics directly from data [13]. This shift from traditional control methods to data-driven learning has yielded significant improvements in robustness and adaptability. Techniques such as reinforcement learning (RL) [5], imitation learning (IL) [1], and adaptive model predictive control are paving the way for robots to achieve human-like dexterity in manipulating contacts.

However, applying learning-based methods to contact-rich tasks raises critical safety concerns. Exploration during training may lead to unsafe actions that damage the robot, the environment, or nearby humans. Similarly, the execution of learned policies in deployment must guarantee constraint satisfaction and robustness, especially under distributional shifts.

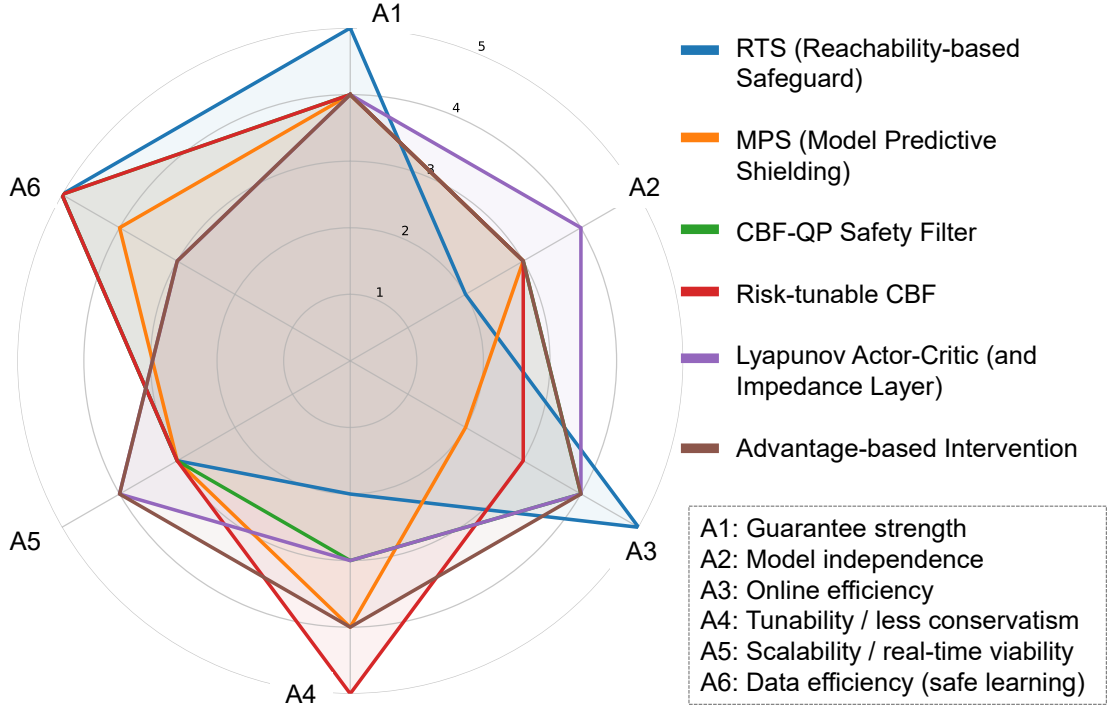


Figure 2.3 Six-dimensional taxonomy of safe learning methods for contact-rich robotic tasks. The six axes capture the learning paradigm, safety mechanism, enforcement space, sensing modality, task type, and evaluation approach.

Therefore, safety must be addressed both during exploration and during execution to ensure the reliable operation of autonomous systems in the real world.

2.2 Definition of Safety in Contact-Rich Tasks

This thesis draws on two comprehensive surveys co-authored during this doctoral research: a survey on safe learning for robotic contact-rich tasks [13], and a survey on imitation learning for contact-rich tasks [1]. Together, these surveys provide a systematic and up-to-date review of the field, which is directly relevant to understanding the motivation and context for the technical contributions presented in subsequent chapters. Building on the thesis motivation introduced in Chapter 1, this chapter consolidates the conceptual foundations needed for the rest of the thesis. In particular, it clarifies how safety, compliance, sensing, and learning interact in contact-rich manipulation, thereby motivating the specific methodological choices made in the subsequent technical chapters. Figure 2.3 illustrates the six-dimensional taxonomy of safe learning methods for contact-rich tasks, covering the learning paradigm, safety mechanism, enforcement space, sensing modality, task type, and evaluation approach.

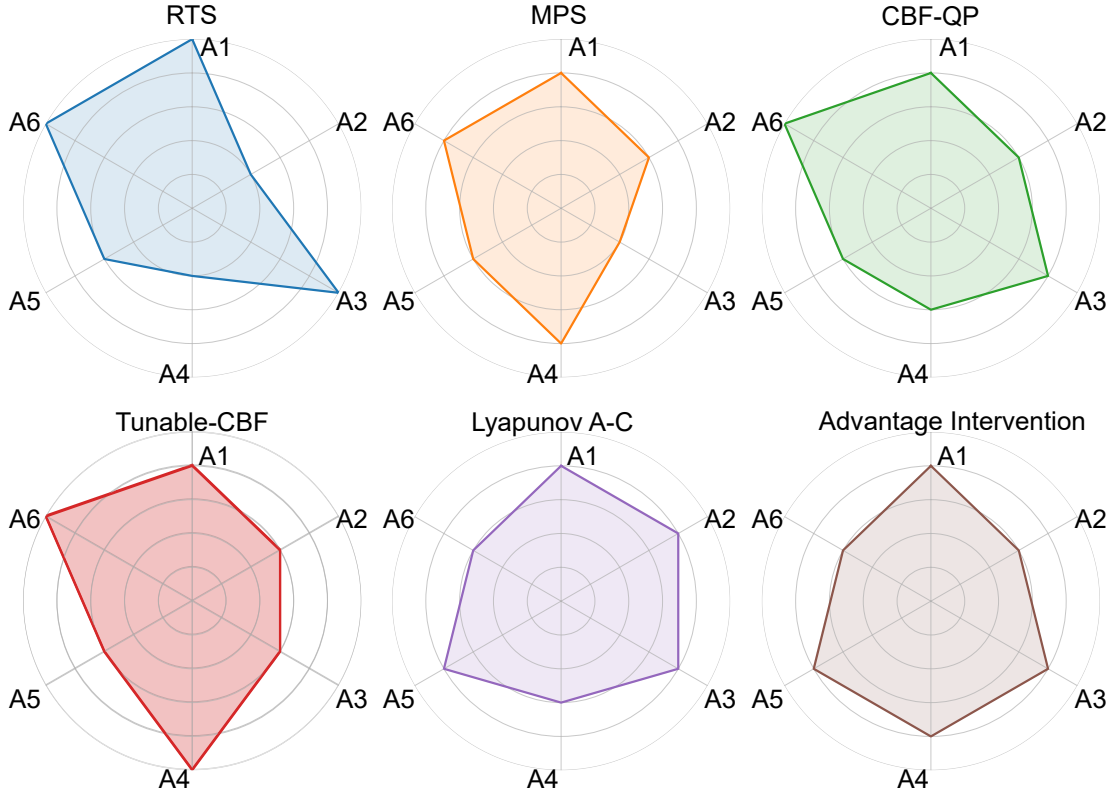


Figure 2.4 Detailed breakdown of the six-dimensional taxonomy showing the distribution of approaches across each dimension in the safe learning literature.

Figure 2.5 presents the overall structural overview of safe learning approaches for contact-rich robotic tasks.

2.3 Problem Formulation and Learning Paradigm

2.3.1 Problem Setting

Contact-rich manipulation arises across assembly, insertion, forceful surface interaction, and human–robot collaboration. We consider a general formulation that does not commit to a single mathematical model. Tasks are described by (i) a state representation x_t that includes robot kinematics/dynamics, latent contact modes, and measured or estimated wrenches; (ii) a control signal u_t that can be low-level torques/velocities or higher-level references such as force targets and impedance/admittance gains; and (iii) safety and feasibility specifications involving force/torque envelopes, friction and approach cones, motion limits, and human-centric constraints [4, 14, 15]. Depending on assumptions and requirements, the underlying

decision model may be an unconstrained Markov decision process (MDP) or a partially observable Markov decision process (POMDP), a constrained variant (e.g., constrained Markov decision process (CMDP)), or a hybrid-systems/control-theoretic problem with explicit contact transitions. In all cases, safety is operationalized through limits on contact forces/pressures and motion, potentially augmented by risk- or semantics-derived rules [15–18].

Observability and abstraction are key design choices. When direct force/torque sensing is available, controllers can track force references and tighten safety envelopes; otherwise, observers or learned estimators provide wrench/contact-state cues for policy inputs and online monitoring [14]. Action abstraction often improves safety and sim-to-real transfer: outputting pose/force targets or impedance parameters instead of raw torques yields compliant interaction and clearer hooks for runtime safety mechanisms [14, 4, 15]. Finally, semantics (VLM/VLA) supply high-level task structure and safety rules that must be grounded into formal constraints and checked during execution [16–20].

2.3.2 Learning Paradigm

We summarize five main learning paradigms and how they interface with safety in contact-rich manipulation. Fig. 2.5 provides an overview of this field.

Reinforcement learning: RL learns policies for variable-impedance forceful skills and insertions, often using action spaces that output force targets or stiffness, damping to preserve compliant interaction. Safety is enhanced by conservative objectives, recovery strategies, and shielded execution (e.g., barrier, reachability, MPC layers) [21–26]

Imitation learning: In IL, initial policies of contact sequencing and force behavior (approach, touch, slide) are learned from demonstrations. It is common to refine the policy by RL or supervised updates; hybrid trajectory-force learning is prevalent in insertion and assembly settings. IL shares the same safety-friendly action abstractions and leverages compliant controllers during execution [27–29]. In this survey, we focus on methods with a notable online learning component, and direct the readers to recent surveys for pure IL methods [1].

Learning for model-based and control-theoretic methods: Impedance and admittance regulation, robust and adaptive force controllers, MPC with contact constraints, and barrier, reachability filters operationalize limits on contact force, torque, friction cones, and etc. Learning augments these by identifying dynamics, adapting gains, or shaping feasible sets, yielding transparent enforcement at runtime [30–33, 30, 34]

Foundation model-based methods: Foundation models for robotics leverage large-scale vision-language (and action) pretraining to parse open-ended goals, infer affordances, and impose safety-aware constraints before execution. In contact-rich settings, these methods are typically embedded in a plan-parameterize-enforce stack: (i) a VLM/VLA module interprets natural-language goals and scene context to propose task sketches and safety rules (e.g., force threshold, no-touch regions, speed limitation); (ii) the semantic plan is compiled into formal constraints and parameterized skill references (pose, force, impedance) rather than raw torques to preserve compliant interaction [35]; and (iii) low-level safety layers (e.g., MPC/CBF filters, rate limiters) execute with verifiable bounds on contact forces and approach geometry [16–18]. Data and evaluation rely on open toolchains and safety-oriented benchmarks to quantify violation rates, refusal/recovery behavior, and generalization under distribution shift [36, 37, 19, 20]. Practically, safety guardrails for foundation-model pipelines include structured prompting and tool schemas, text-to-constraint grounding with pre-execution checks, conservative default limits (force/velocity caps), and “plan-then-filter” action projection through certified controllers; these mitigations address hallucination and mis-grounding that could otherwise lead to overloads at contact [16, 18, 38].

Hybrid integration Planning and skill graphs (potentially guided by VLM/VLA) provide structure; learned policies output pose/force/impedance parameters; safety layers project actions to satisfy certified constraints. Semantic modules further align high-level intent and safety rules with the certified execution stack [19, 20, 18, 38]. Evaluation commonly reports violations, near misses, and generalization using standardized suites and simulators [36, 39, 37].

2.4 Categorization of Methods in Related Works

This section organizes existing safe learning approaches for contact-rich robot tasks along several complementary dimensions to clarify their underlying assumptions and design trade-offs. By examining methods through perspectives such as task characteristics, sensing and policy modalities, data acquisition strategies, simulation environments and benchmarks, safety evaluation metrics, safety abstraction levels, and safety enforcement spaces, the survey highlights common patterns and gaps that are not apparent from any single taxonomy. This multi-view categorization also helps connect classical control-theoretic techniques with modern learning-based and foundation-model-based approaches, providing a unified lens for analyzing current methods and guiding future developments.

2.4.1 Safe Exploration Methods

Overview

Safe human-robot cooperation (HRC) is a major application branch within the broader class of contact-rich tasks, and it has traditionally been approached through path planning algorithms such as rapidly-exploring random tree star (RRT*) [40] for dynamic human-robot systems, or artificial potential field (APF) [41] and virtual fixtures [42] for real-time adjustment of the robot’s trajectory. However, these methods often yield slow trajectory computation and limited generalisation, since they require task-specific tuning, assume simplified environments, and adapt poorly to dynamic or unstructured settings [42].

Control Barrier Functions and Safety Filters

Recent approaches enforce safety constraints via control barrier functions (CBFs), where quadratic programs compute minimal trajectory deviations to maintain safety [43–45]. A CBF $h : \mathcal{X} \rightarrow \mathbb{R}$ defines a safe set $\mathcal{C} = \{x \in \mathcal{X} : h(x) \geq 0\}$ and ensures forward invariance of this set by requiring:

$$\dot{h}(x, u) \geq -\alpha(h(x)) \quad (2.1)$$

for some class- $\mathcal{K}$ function α . Extensions to uncertain CBFs (U-CBFs) handle environmental uncertainty [46, 47], though worst-case assumptions often yield conservative designs. Stochastic frameworks using chance constraints offer less conservative alternatives [48, 49], but remain computationally intractable for most HRC applications due to non-convex feasible regions [50].

Reachability-Based Methods

Across reachability-based methods, a common theme is to shift most of the computational burden into the offline construction of trajectory-parameterised reachable sets based on a reasonably accurate model. Reachability-based Trajectory Safeguard (RTS) [51] is an offline-online safety layer for continuous-control RL that precomputes trajectory-parameterised forward reachable sets and, at runtime, projects the RL agent’s chosen plan parameters to a nearest safe plan while preserving learning efficiency.

Model predictive shielding (MPS) [52] provides a model-predictive shield around any RL policy: they forecast violations and replace unsafe actions with safe, goal-directed recoveries, enabling zero-violation training and deployment. CBF-Guided MPC [53] embeds barrier

constraints in MPC and tunes them online for proactive, explainable hard-safety during tracking/avoidance at the cost of online optimisation and careful model or CBF design.

Constrained Markov Decision Processes

Constrained Markov Decision Processes (CMDPs) provide a formal framework for safe RL, where the agent optimises a reward signal subject to a cost threshold associated with unsafe states [54–57]. Lagrangian relaxation is commonly used to dynamically weight this cost, training the policy to satisfy safety constraints asymptotically. Alternatively, safety can be encoded into the value function itself [58, 59]. Safety critics or feasibility value functions estimate the long-term probability of constraint violation, guiding the policy to select actions that maximise the margin of safety [60].

2.4.2 Safe Execution Methods

Overview

While safe exploration methods prevent catastrophic failures before contact occurs, safe execution methods focus on maintaining stability, compliance, and constraint satisfaction during physical interaction. This is particularly critical in contact-rich tasks where the robot must actively exchange forces with the environment.

Compliant Control and Variable Impedance Learning

Learning-based approaches directly optimise policies, skills, and impedance from data, offering greater adaptability to unmodelled dynamics and complex contact phenomena with less hand-crafted modelling. Lyapunov actor-critic [61] learns policy weights under an explicit Lyapunov constraint, and executes through an adaptive impedance layer for compliant, stability-guaranteed motion. Safety-Aware Unsupervised Skill Discovery [62] learns a skill policy and a safety-critic that estimates failure probability, yielding intrinsically safe skills for downstream tasks. Variable-Impedance Skill Learning [63] learns a latent skill prior/decoder, a latent policy, and impedance as part of the action, achieving safety via compliant interaction and bounded contact forces.

A large family of methods couples force/torque sensing with variable impedance or admittance control. The policy input in these works typically concatenates joint or end-effector pose and velocity with measured interaction forces/torques, and sometimes task descriptors such as surface or tissue properties, environment stiffness, or task phase [64–67].

The action space is parameterised in terms of stiffness, damping, or target forces, and safety is enforced by bounding these impedance parameters, shaping them over time, or embedding stability and performance objectives into the learning problem [63, 68–70].

Passivity-Based Safety

Passivity is a fundamental property of physical systems that ensures stability during contact. A passive system does not generate energy internally and can only dissipate or store energy received from the environment. Energy-based safety methods enforce passivity constraints to ensure that the robot behaves as a passive system and does not inject unbounded energy into the environment [71, 72].

Energy tank methods are a prominent approach for enforcing passivity in variable impedance control. An energy tank stores the energy extracted from the environment and limits the energy that can be injected back, ensuring that the system remains passive at all times [24, 73]. The energy tank level T is governed by:

$$\dot{T} = -P_{\text{ext}} + P_{\text{diss}} \quad (2.2)$$

where P_{ext} is the power extracted from the environment and P_{diss} is the power dissipated by the system. When the tank level drops below a threshold, the impedance parameters are modified to prevent further energy injection.

2.4.3 Foundation Models in Contact-Rich Tasks

Recently, Vision-Language-Action (VLA) models have emerged as a powerful paradigm for end-to-end manipulation. These models bridge perception, reasoning, and control by learning from large-scale multimodal datasets that include language instructions, visual observations, and robot actions [18, 16].

Safety is enhanced in VLA systems through several mechanisms:

1. **Safety-aware prompting and grounding:** VLA models can be conditioned on safety-specific prompts or rules (e.g., “move slowly near the glass”) to generate compliant behaviours [74, 75].
2. **Safety critics and filtering:** Auxiliary safety modules or critics can be trained to evaluate the risk of generated actions before execution, filtering out potentially hazardous commands [76–78].

3. **Multimodal alignment:** By aligning visual and linguistic representations with physical safety constraints (e.g., force limits), VLA models can better generalise safety concepts to novel objects and environments [79–81].

Figure 2.6 illustrates how ChatGPT and other large language models can be used to generate task descriptions and safety constraints for contact-rich manipulation.

While these methods promise scalability and generalization, especially in contact-uncertain environments, they often suffer from limited interpretability and pose challenges in formal safety verification compared to modular architectures.

2.4.4 Safety Enforcement Spaces

In the context of learning for contact-rich manipulation, safety is not only a monolithic concept but is also enforced hierarchically across the robot’s control architecture. Constraints can be imposed at the level of physical interaction (task space), internal actuation (joint space), or within the decision-making abstraction (policy space) [13].

Task Space Safety Enforcement

Task-space mechanisms operate directly at the interface of physical interaction, regulating the kinematics and dynamics of the end-effector. By defining safety in Cartesian coordinates, these methods directly constrain the contact forces and motions exerted on the environment.

A predominant strategy for task-space safety is the use of compliance and stiffness modulation. Rather than tracking rigid trajectories, impedance and admittance control schemes allow the robot to deviate from reference paths in response to external forces, thereby absorbing impact energy and maintaining stable contact [82–85, 61]. The works in [86, 87] explicitly bound the forces and torques at the end-effector to prevent excessive contact pressures, integrating such constraints into the controller and learning process. Recent learning-based frameworks have extended this architecture by parameterising variable impedance policies, where the agent learns to output context-aware stiffness or virtual force targets to ensure gentle interaction [63, 68, 88]. To guarantee stability in these interactions, energy-based approaches often enforce passivity constraints, ensuring that the robot behaves as a passive system and does not inject unbounded energy into the environment [71, 72].

Beyond inherent compliance, formal constraints are frequently applied to bound the workspace. CBFs and safety filters are employed to strictly confine end-effector states within a safe set, overriding policy actions that threaten to violate positional or force limits [89–

91]. These safe manifolds can be learned from demonstrations or modelled via Gaussian processes, restricting the planner to regions with high probability of safety [92, 93].

In RL settings, task-space safety is often formulated via reward shaping to penalise excessive contact forces [94–97], or by restricting the action space to a safe subset of the workspace to prevent collisions during exploration [98, 62, 99, 100]. Before execution or learning, planning-based approaches compute contact-rich trajectories that satisfy contact stability, task-unrelated collision avoidance, and joint limits, so that any subsequent policy learning begins with inherently safe trajectories [74, 101].

Notably, in human-robot interaction scenarios, the constraints are adapted to dynamic environments, utilising safety fields to enforce separation distances, and controllers are designed to comply with industry safety standards, e.g., ISO force limits, during physical contact [102, 75, 103].

Joint Space Safety Enforcement

While task-space methods focus on extrinsic interaction, joint-space enforcement prioritises the robot’s intrinsic integrity. These methods operate in the configuration space, constraining joint angles, velocities, and torques to prevent self-collisions, singularity proximity, and actuator saturation.

Safety in this domain is often realised by augmenting low-level controllers with limiting functions. Joint-level CBFs and saturation filters ensure that commanded torques do not exceed hardware capabilities or result in unsafe kinetic energy, thereby indirectly limiting the force transfer to the environment [104]. Stability is further ensured through Lyapunov-based formulations, such as enforcing positive-definiteness in learned joint stiffness matrices to guarantee system passivity [105, 106].

To mitigate the risks of dynamic impacts, adaptive joint-torque controllers utilise proprioceptive feedback to dissipate shock through joint compliance [107, 108]. In the context of RL, policies operating in joint space are frequently constrained to prevent erratic high-frequency actuation. Strategies include penalising proximity to joint limits [109] or employing hierarchical schemes where the RL agent dictates high-level goals while a lower-level motion planner generates smooth, collision-free configuration trajectories [110].

Dual-Space Safety Enforcement

Exclusive reliance on either task or joint space constraints can create safety blind spots. Task space control may inadvertently command infeasible joint configurations, while joint-space

safety may overlook excessive contact pressures at the end-effector. Consequently, hybrid architectures have emerged to enforce safety concurrently across both domains.

Hierarchical control frameworks typically employ a task-space loop to manage environmental interaction forces, coupled with an inner joint-space loop to handle redundancy resolution and torque limits. A Cartesian impedance controller regulates contact stiffness, while a secondary joint-space reflex module monitors for collisions or internal limits, overriding commands to dissipate energy if discrepancies arise [111, 112]. Theoretical guarantees for such dual-loop architectures have been established via Lyapunov analysis, proving that errors in both force tracking and joint deviation remain bounded [113, 114].

Policy Space Safety Enforcement

Abstracting away from real-time control, safety is also enforced at the decision-making level. Policy-space methods integrate safety considerations directly into the learning objective or planning horizon, aiming to preemptively avoid states that would necessitate low-level emergency intervention.

A standard formalism for this is the CMDP, where the agent optimises a reward signal subject to a cost threshold associated with unsafe states [54–57]. Lagrangian relaxation is commonly used to dynamically weight this cost, training the policy to satisfy safety constraints asymptotically. Alternatively, safety can be encoded into the value function itself [58, 59]. Safety critics or feasibility value functions estimate the long-term probability of constraint violation, guiding the policy to select actions that maximise the margin of safety [60].

Hybrid Methods

Hybrid methods combine safety mechanisms across multiple abstraction levels to leverage the strengths of each [115]. For instance, a high-level planner might filter task options using a safety-aware graph while a low-level controller enforces contact force constraints through learned CBFs. Alternatively, end-to-end models may be guided by high-level logic rules or shielded using low-level safety filters during deployment.

Plan-learn-filter pipelines. A high-level planner (graph, sampling, optimisation, or semantic decomposition) proposes nominal contact strategies or intermediate sub-goals; a learned policy (often variable impedance/admittance or residual) generates parameterised actions (pose, force, stiffness); then a runtime safety layer (CBF / reachability / MPC shield or projection) minimally modifies unsafe commands while preserving progress [68, 116–

118, 15]. This layering reduces unsafe transients at approach, impact, and sustained contact phases by deferring exact constraint satisfaction to an online filter.

Compliant action abstraction and certified projection. Rather than outputting raw torques, hybrid methods learn action abstractions (desired Cartesian displacements, force targets, diagonal or structured stiffness gains) that preserve passivity margins and expose explicit hooks for safety modification [24, 21, 68, 116, 119]. Certified projection (QP with CBF constraints, shield veto, reachability pre-checks) enforces forward invariance or constraint satisfaction without retraining the core policy [117, 118, 15, 37].

Semantic or VLM/VLA hybrid safety. Vision-language(-action) and emerging foundation model components parse unstructured instructions and scene context into structured constraint objects [16, 18]. These are handed to the plan-learn-filter stack: semantic planner produces goal or sequence; learned variable impedance or residual policy outputs compliant parameters [76, 77]; runtime CBF, reachability, and MPC shield enforces limits; logging layer records violations for calibration [68, 116, 117, 15]. Multimodal fusion, especially vision and tactile and force, disambiguates occlusions and anticipates risky contact transitions, while structured prompting and action schema reduce hallucination-induced unsafe commands [17, 16, 18].

2.4.5 Sensing Modalities for Safe Contact-Rich Learning

The choice of sensing modality profoundly influences the safety and performance of contact-rich manipulation systems. Table 2.1 summarises the dominant sensing modalities used in safe learning for contact-rich tasks, along with their typical safety leverage and representative references.

Proprioceptive and Kinematic State

Several works rely predominantly on proprioceptive and kinematic state, joint configurations, end-effector poses and their derivatives, sometimes augmented with a few task-level geometric variables, while omitting explicit force/torque signals from the policy input. In these settings, safety emerges from conservative motion generation, collision-aware planning and geometric reasoning about contact, often supported by explicit constraints, recovery sets or safety-oriented value functions [92, 62, 28, 101, 120, 60, 156, 2, 99, 110, 124].

A first group of approaches couples a compact state representation (pose and velocity) with an explicit description of goals and safety constraints, and then uses model-predictive or safe-RL updates to keep trajectories inside a certified safe set [92, 62, 120, 60, 165, 90, 124].

Because distances to constraint boundaries are part of the observation, the algorithm can reason explicitly about safety margins without having to infer risk from raw contact forces.

A second family of methods organises behaviour around pose-based skills or motion primitives. Here, the observation typically consists of robot joint or end-effector states together with task-relevant geometric features, such as relative part poses or planner-generated waypoints, while the policy outputs bounded pose increments or selects from a library of safe trajectories [28, 101, 99, 110].

Force and Torque Sensing

For many contact-rich tasks, especially those involving insertion, surface following, or physical human-robot interaction, force/torque (F/T) measurements at the wrist or fingers act as the primary safety interface. Policies typically observe these signals together with joint and end-effector state and use them to cap contact loads, detect slip or jamming, and adapt impedance so that tracking performance is achieved without violating force or pressure limits [21, 86, 64, 82, 63].

F/T signals are also combined with explicit safety margins, constraints or stability certificates derived from control theory. Controllers additionally receive features derived from Lyapunov functions, control barrier functions, passivity errors or other safety indicators that are computed using the measured forces and state trajectories [89, 61, 117, 106]. Learning algorithms then operate directly on these structured safety variables, for example by penalising trajectories that approach constraint boundaries, adjusting admittance or impedance gains when passivity residuals grow [24, 73], or biasing exploration away from regions with high estimated risk [131, 104, 90, 91].

In physical human-robot interaction and teleoperation, F/T channels additionally encode human intent and comfort. Many frameworks treat human-applied forces at the end-effector, interaction torques, or teleoperated wrenches as part of the policy input together with robot joint or end-effector state and task phase [98, 126, 115, 135]. These signals are interpreted either as demonstrations of desired compliant behaviour or as real-time preferences over interaction forces, and are used to adapt impedance online, to learn tele-impedance mappings, or to construct task-oriented safety fields around the human partner [102, 166, 141, 146].

Vision Sensing

Vision is widely used to provide non-contact, global information about objects, surfaces and humans that is difficult or impossible to infer from proprioception and force alone [13].

Depending on the camera placement, visual inputs range from bird's-eye or third-person views that cover the entire workspace, through fixed eye-to-hand cameras that monitor a local cell, to wrist- or eye-in-hand cameras that track the interaction region from the robot's point of view [87, 153, 154, 102, 95, 129, 90, 157].

Global and external viewpoints are typically used to capture task geometry, object and fastener poses, and human motion. Assembly and disassembly systems exploit RGB or RGB-D images, often converted into scene graphs or object pose estimates, alongside joint and end-effector state to ensure that motion plans respect part tolerances and obstacle constraints before contact is made [95, 129, 139, 151, 100].

Tactile Sensing

Tactile sensing provides fine-grained information about contact geometry, pressure distribution, and slip that is not accessible through force/torque sensors or vision alone. High-resolution tactile arrays can detect incipient slip before it becomes a full slip event, enabling proactive corrective actions [161–164].

Recent advances in tactile sensor design, including optical tactile sensors such as GelSight and DIGIT, have made it possible to obtain rich contact information at low cost. These sensors have been used in a variety of contact-rich manipulation tasks, including in-hand manipulation, assembly, and surface following [167, 168].

2.4.6 Safety Evaluation Metrics

Overview of Evaluation Metrics

Safe learning works differ from traditional learning works in their primary goals. Traditionally, the learned models are evaluated for their predictive performance or their task completion success. Although the safe learning works are still concerned with these metrics, they add new ones related to safety, such as a statistic of constraint violations or a cost value. In the case of contact-rich problems, the safety metrics are usually related to the contact forces [13].

Figure 2.7 presents the prominent metrics categorised according to their goals. The identified metrics aim to measure the safety, task (success) and efficiency objectives.

Safety, Efficiency and Task Objectives

Safety is usually assessed through the violation of predefined constraints [154, 98] or the value of a cost function [157, 169]. In physical interaction tasks, it is often quantified through physical metrics such as force, acceleration and velocity [129, 104]. Control stability and human safety are important subcategories of the safety objective. Stability can be measured by the convergence rate [92, 64] or the energy flow of the system [73, 24, 170]. The human safety includes primarily the physical safety of the human in direct contact tasks like massaging or bathing assistance [65, 153], however, perceived safety concepts such as trust, predictability and comfort gain importance for the acceptance of robots in social spaces [166].

The task objective is quantified by the frequency of achieving a success [108] or the value of a task cost function (i.e., reward) [151]. In some works, the quality of a solution is assessed by the progress of task completion, the size of the explored space [86], the amount of produced output [85], or the tracking accuracy [153]. The efficiency objective is assessed by the energy consumption [163], and time requirements of the data collection (e.g., teaching) [171], learning a model [110], task execution [151, 87], or computation of an action or a plan [108, 102].

Analysis Levels

Figure 2.9 presents three levels of result analyses: global (high-level), variational (mid-level), and exemplar (low-level).

The safety aspect of a learning system can be further evaluated through added noise and disturbance. The noise can be injected in observation-space [94, 127, 108] or action-space [95]. Vision-based systems can be tested with changing backgrounds of cluttered scenes [172]. These tests help assess the limitations and sensing requirements of the systems.

2.4.7 Application Domains

A central motivation for contact-rich manipulation tasks is its potential to translate advances in algorithms into robust, real-world capabilities. In practice, these methods are deployed across a range of application domains that differ in task structure, safety constraints, and interaction modalities, which in turn shape the design and evaluation of systems. In this section, we provide an overview of the main domains where contact-rich tasks have been applied and discuss how their relative importance has evolved over time, before comparing the suitability of different learning-based methods for these settings.

Figure 2.10 summarizes the principal real-world environments in which contact-rich tasks have been studied, spanning industrial assembly, household manipulation, and healthcare assistance.

Beyond the breadth of domains, it is also important to understand how their prevalence has changed as the field has developed. Figure 2.11 depicts the temporal evolution of application domains in the contact-rich robotic tasks literature, highlighting an overall trend towards increasing diversity and a gradual shift from tightly controlled industrial scenarios to more unstructured household and healthcare environments.

Figure 2.11 shows the temporal evolution of application domains, showing the increasing diversity of applications over time.

These differing domains place distinct demands on learning algorithms, for example, in terms of data efficiency, generalisation to novel conditions, and robustness to complex contact interactions. To make these trade-offs explicit, Figure 2.12 provides a high-level comparison of several learning-based approaches along key performance dimensions that are particularly relevant for contact-rich tasks.

Figure 2.12 shows a comparison of different IL approaches across key performance dimensions.

Industrial Applications

Industrial robotics has been the primary driver of contact-rich manipulation research. Assembly tasks, including peg-in-hole insertion, nut-bolt threading, and connector mating, require precise force control and tight tolerances. RL and IL methods have been applied to automate these tasks, with significant improvements in success rate and robustness [1].

Surface interaction tasks, including polishing, sanding, and deburring, require continuous force control to maintain consistent contact quality. Variable impedance control and learning-based methods have been used to automate these tasks, with the robot learning to adapt its stiffness and damping to maintain consistent contact force [13].

Household and Service Applications

Household manipulation tasks, including door opening, drawer manipulation, and object handover, require the robot to interact with a wide variety of objects and environments. These tasks are challenging due to the variability of object properties and the need for generalisation to novel objects and environments.

Recent advances in foundation models and VLA models have enabled significant progress in household manipulation. RT-2 [173], OpenVLA [174], and π_0 [175] have demonstrated impressive generalisation to novel tasks and objects, opening the door to more capable household robots.

Healthcare Applications

Healthcare applications of contact-rich manipulation include surgical assistance, physical therapy, and personal care. These applications require extremely precise force control and safety guarantees, as any error could cause harm to the patient [1].

Physical therapy robots must apply precise forces to assist patients with movement rehabilitation, while maintaining safety constraints on force and velocity. Safe RL methods have been applied to learn adaptive therapy policies that can generalise to different patients and conditions [13].

2.5 Key Challenges

Despite rapid progress in learning-based methods for contact-rich robotic manipulation, several key challenges remain before these approaches can be deployed safely and reliably at scale [13].

2.5.1 Fragile Safety and Stability Guarantees

Learning-based controllers, including those using deep networks and VLAs, often lack formal guarantees of passivity, stability, and constraint satisfaction under hard contacts [24, 176, 72], which can lead to unsafe behaviours and hardware damage. While passivity-centric and energy-constrained formulations offer promising directions, integrating them seamlessly with large, expressive policies remains technically challenging, especially when contact conditions change rapidly or unpredictably.

2.5.2 Data Scarcity and Biased Supervision

Contact-rich skills require precise force control and tight tolerances, but high-quality demonstrations and interaction data are expensive, brittle, and difficult to scale compared to standard vision-language datasets. VLMs and VLAs promise better generalisation through pretraining, yet they are typically supervised by weak or indirect signals that do not capture fine-grained

contact events, leading to policies that may look competent visually but fail under subtle force or geometry variations [13].

2.5.3 Limited Robustness and Generalisation

Learning-based policies for contact-rich manipulation, especially ones trained on narrow task distributions, often overfit to specific layouts, materials, or sensing conditions and degrade sharply under distribution shifts. Although VLAs offer broader prior knowledge, current models typically have weak uncertainty estimation, making it difficult to detect when a contact-rich interaction is entering an unsafe regime and requiring fallback to safer behaviours or human intervention [13].

2.5.4 Safe Exploration Under Physical Risk

Exploration remains the most safety-critical phase: naive probing causes jamming, surface damage, or tool wear in insertion, polishing, and assembly contexts [15, 21]. Existing constrained and risk-sensitive RL frameworks provide expectation or probabilistic bounds, yet exhibit weakness under distribution shift. Shielding and action-projection layers offer hard online guarantees, but typically depend on conservative uncertainty sets, limiting data efficiency.

2.5.5 Safe Execution and Performance Trade-offs

Ensuring sustained constraint satisfaction post-training (safe execution) requires balancing responsiveness, task efficiency, and physical margin. Variable impedance or admittance policies learned from demonstrations or reinforcement signals modulate compliance adaptively [116, 68, 21], yet stability margins under rapid re-tuning are seldom quantified. Runtime safety filters can erode performance through excessive intervention, while soft filtering strategies risk cumulative force violations.

2.5.6 Benchmark and Dataset Limitations

Existing simulation suites and benchmarks provide limited coverage of deformable objects, multi-stage processes, tool-supported tasks, and tasks performed in close collaboration with humans [36, 19, 20]. Public datasets rarely include high-frequency wrench traces aligned with failure annotations, rich “near-miss” episodes essential for shaping recovery

policies, or semantic safety labels (forbidden regions, fragile surfaces). This scarcity impedes reproducible comparison of violation metrics, recovery latency, and generalisation [13].

2.5.7 Safety Risks Specific to VLMs and VLAs

Generalist VLM/VLA policies introduce new safety concerns, such as misinterpreting language instructions, hallucinating feasible contact strategies, or overconfidently executing plans that violate implicit safety constraints. Recent work on safety-aligned VLAs and failure detection for generalist policies underscores how difficult it is to systematically elicit unsafe behaviours, constrain them during learning, and verify that contact-rich deployments remain within acceptable risk bounds [13].

2.5.8 Human-Robot Collaboration Safety

Human proximity or direct physical collaboration introduces soft constraints (comfort, predictability) alongside hard limits (force, speed). For contact-rich tasks near or with humans, operators must be able to anticipate robot behaviour and intervene effectively, yet current learning-based controllers often lack transparency and interpretable failure modes. Existing safe learning pipelines predominantly encode physical thresholds while neglecting temporal predictability or trust metrics central to acceptance [177, 178]. Advanced models, including VLAs, can in principle explain intentions or request help, but designing interaction protocols, interfaces, and explanations that are both faithful and useful in high-contact scenarios remains an open challenge.

2.6 Positioning of This Thesis

This thesis addresses several of the key challenges and future directions identified in the two surveys. The contributions are organised into two main research thrusts:

Safe Reinforcement Learning for Contact-Rich Tasks (Chapters 3–5). The first research thrust develops novel safe RL methods for contact-rich manipulation that address the challenges of safe exploration, stability guarantees, and passivity enforcement. The SRL-VIC method (Chapter 3) introduces a safe RL framework with variable impedance control that enables safe exploration in contact-rich environments. The Bresa method (Chapter 4) introduces a bio-inspired reflexive safe RL framework that combines a primary RL policy with a fast reflex mechanism for immediate response to unsafe contact events. The Passivity-

Centric Safe RL method (Chapter 5) introduces an energy-tank-based passivity constraint that ensures the stability of the learned variable impedance policy.

Foundation Models for Contact-Rich Manipulation (Chapters 6–8). The second research thrust develops novel methods for integrating foundation models with compliant control and multimodal sensing for contact-rich manipulation. The OmniVIC method (Chapter 6) introduces a VLM-guided variable impedance control framework that enables language-conditioned compliant manipulation. The CompliantVLA method (Chapter 7) introduces a compliant control adaptor for VLA models that enables contact-rich manipulation with foundation models. The TacVLA method (Chapter 8) introduces a tactile-enhanced VLA framework that integrates tactile sensing into VLA models for contact-rich manipulation.

Together, these contributions form a coherent research programme that advances the state of the art in safe and effective learning for contact-rich robotic manipulation, from safe RL methods with formal safety guarantees to foundation model-based approaches with multimodal sensing.

2.7 Summary

This chapter has provided a comprehensive review of the state of the art in safe learning and imitation learning for contact-rich robotic manipulation. The key takeaways are as follows. Contact-rich manipulation is a fundamental challenge in robotics, requiring precise force control, compliance adaptation, and safety guarantees. Safe learning methods for contact-rich tasks can be categorised into constrained optimisation, safety filter, and safe exploration approaches, operating in task space, joint space, or policy space. Sensing modalities play a crucial role in safe contact-rich manipulation, with force/torque sensing, vision, tactile sensing, and multimodal fusion each offering distinct advantages and limitations. Imitation learning methods, including behaviour cloning, dynamic movement primitives, generative methods, and foundation models, provide complementary approaches to learning contact-rich manipulation skills from demonstrations. Key challenges include fragile safety guarantees, data scarcity, limited robustness, safe exploration, and the integration of foundation models with physical safety constraints. Future directions include standardised benchmarks, safety-aware learning frameworks, sim-to-real transfer, multimodal semantic grounding, and physically-aware foundation models.

The subsequent chapters translate these observations into concrete technical contributions. We first focus on safe reinforcement learning for contact-rich tasks, progressively moving from predictive safety and variable impedance adaptation to reflexive intervention and

passivity-based stability, before turning to foundation-model-based approaches that extend these safety insights to more general and multimodal manipulation settings.

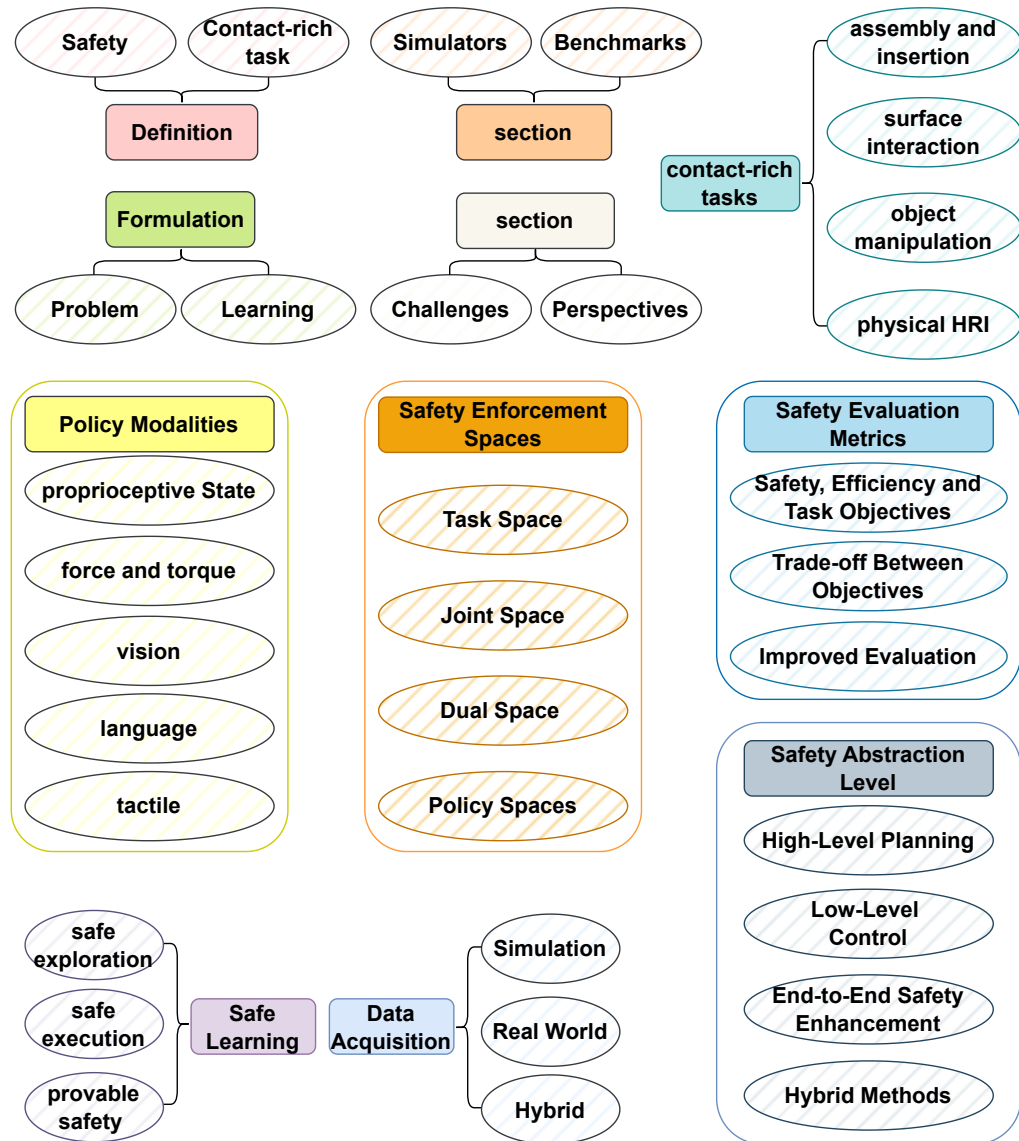


Figure 2.5 Structural overview of safe learning approaches for contact-rich robotic tasks, organised by safety mechanism, enforcement space, and learning paradigm

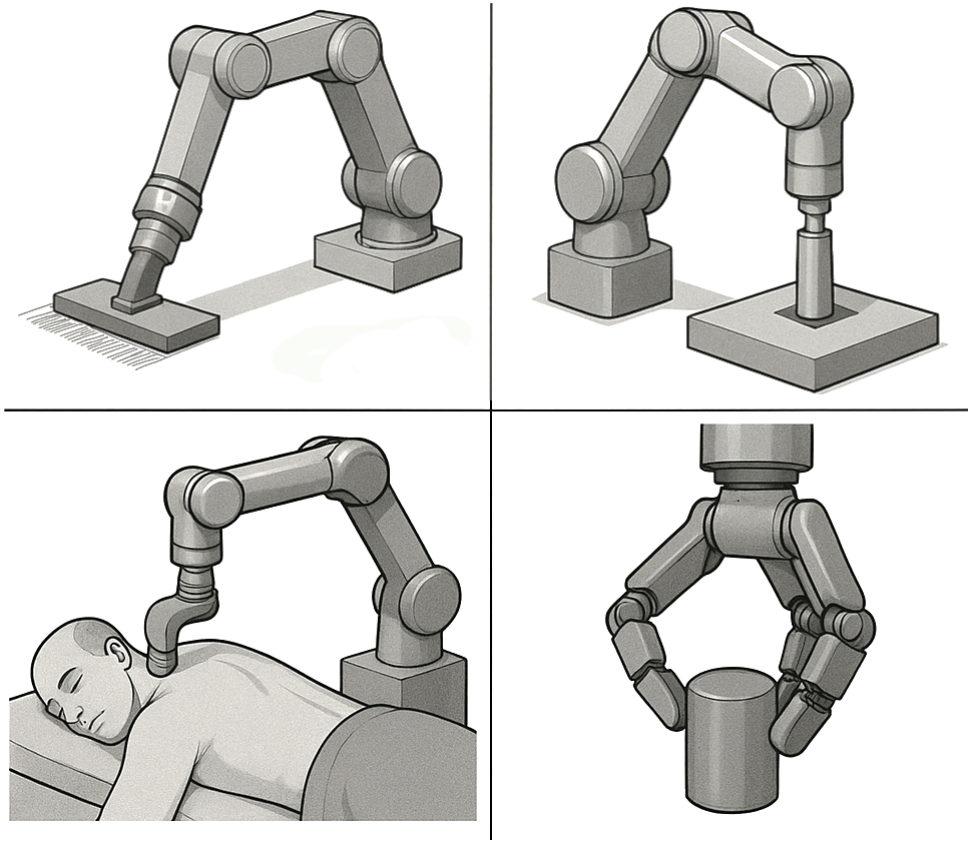


Figure 2.6 Illustration of how large language models can be used to generate task descriptions and safety constraints for contact-rich manipulation tasks.

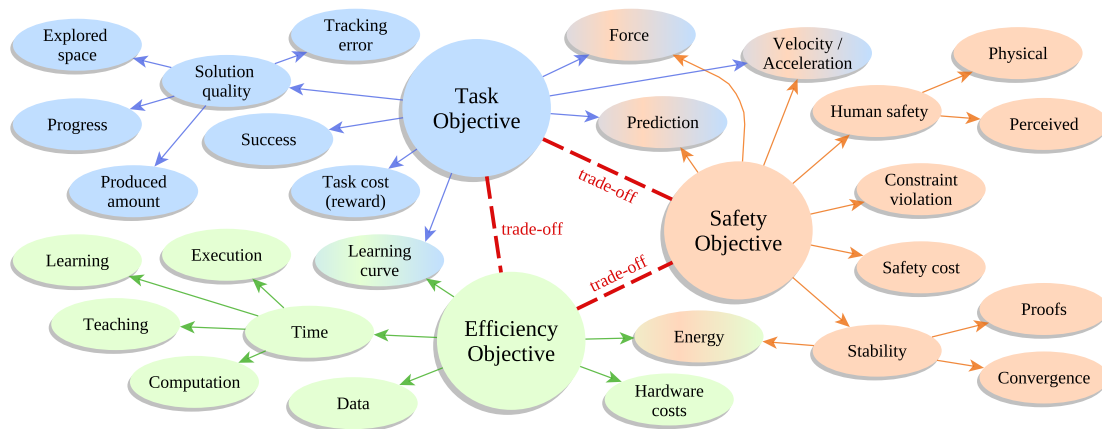


Figure 2.7 The metrics in the contact-rich safe learning literature can be categorised according to their objectives: safety, task, and efficiency. There are trade-offs between these objectives. The researchers aim to develop safe systems with a good trade-off for task success and efficiency. Some metrics, such as force, velocity and energy serve to measure multiple objectives.

Table 2.1 Sensing and policy modalities for safe learning in contact-rich tasks

Dominant Modality	Typical Safety Leverage	Key References
Proprioceptive / kinematic state	Contact reasoning from joint configurations and end-effector pose; conservative motion generation	[92, 28, 101, 120, 60, 121, 122, 99, 110, 123, 124]
Force/torque	Real-time enforcement of force and impedance limits, slip detection, compliant correction of model error	[86, 64, 87, 82–84, 125, 94, 98, 85, 89, 102, 61, 126, 127, 65, 95, 63, 68, 118, 117, 128, 66, 71, 129–132, 69, 104, 67, 133, 134, 113, 135, 106, 107, 136–138, 111, 139–145, 72, 146, 147, 70, 148, 91, 149, 150, 108, 151, 109, 100, 17, 152]
Vision	Contact evaluation, localising task features, and supplying global context	[86, 64, 87, 153, 83, 125, 85, 154, 102, 127, 95, 62, 116, 16, 129, 155, 156, 90, 136, 157, 139, 140, 158–160, 2, 108, 151, 100, 17, 88]
Natural-language instruction	Large multimodal models interpret human instructions and check internal low-level safety	[74, 102, 16, 75]
Tactile	High-resolution contact localisation, pressure distribution monitoring, and incipient slip detection	[161–164]
Multimodal fusion	Enables high precision under occlusion	[86, 64, 87, 83, 125, 85, 102, 127, 95, 16, 129, 136, 139, 140, 162, 108, 151, 100, 17, 164]

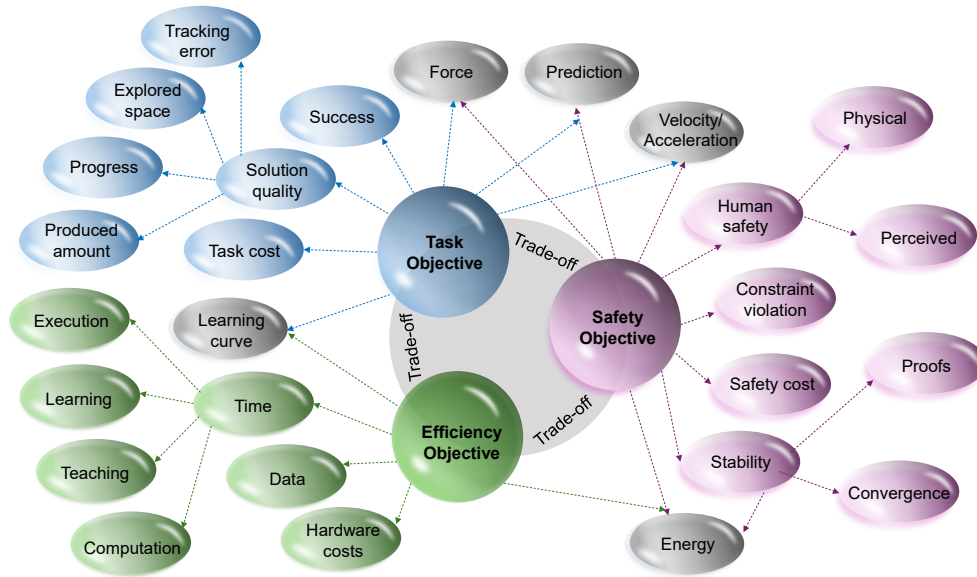


Figure 2.8 Three-dimensional bubble chart showing the distribution of safe learning works across safety, task, and efficiency metrics.

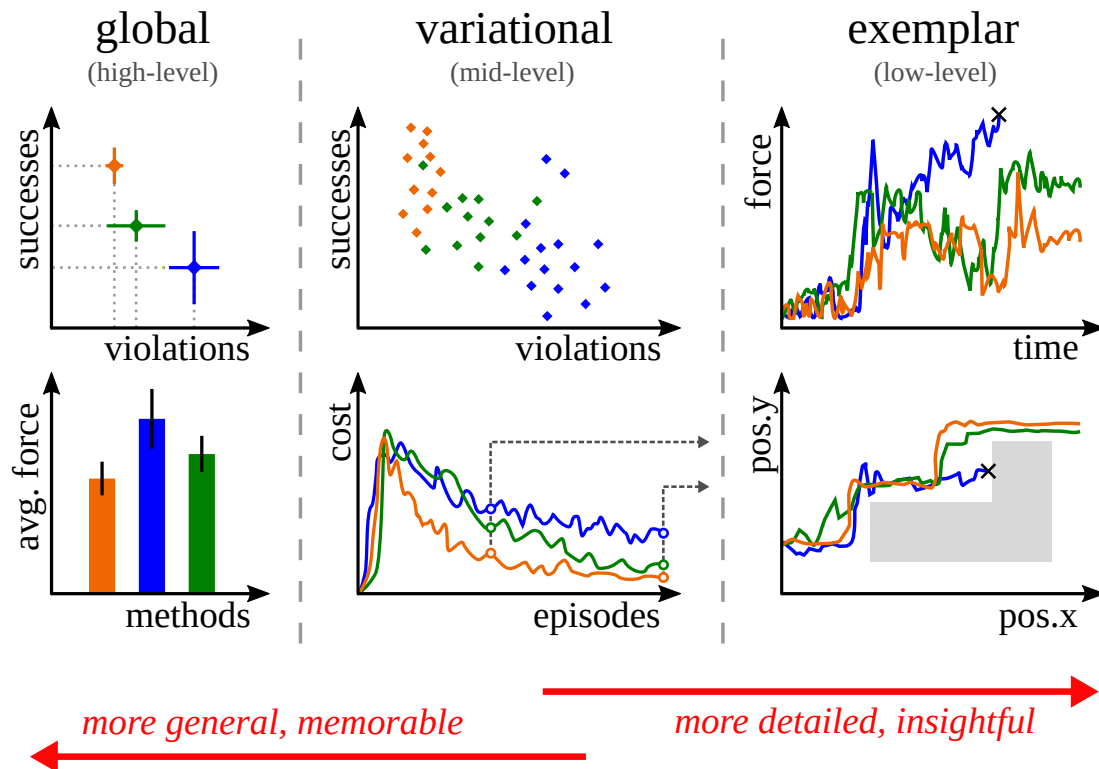


Figure 2.9 Three levels of result analyses in safe learning for contact-rich tasks: global (high-level), variational (mid-level), and exemplar (low-level).

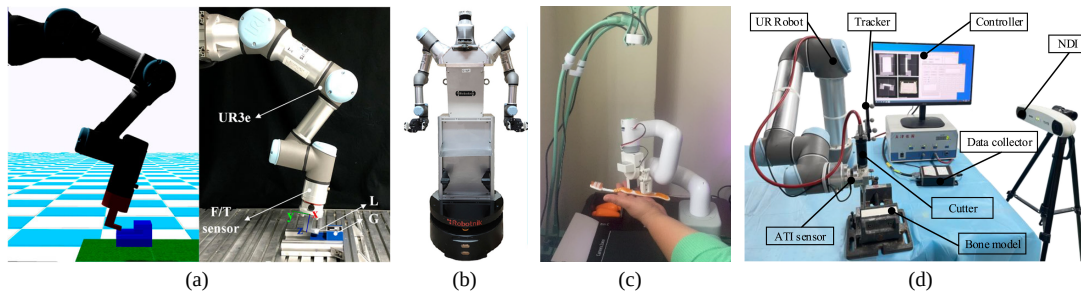


Figure 2.10 Application domains of contact-rich robotic tasks, spanning industrial assembly, household manipulation, and healthcare assistance.

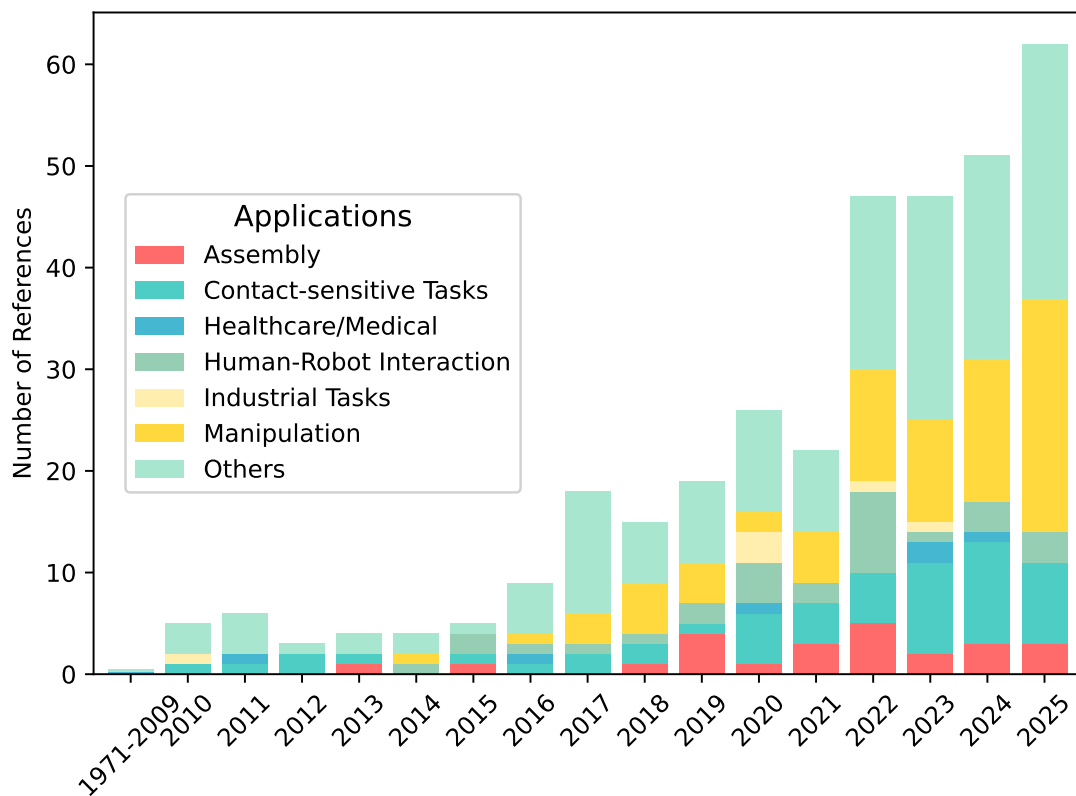


Figure 2.11 Temporal evolution of application domains in the contact-rich robotic tasks literature.

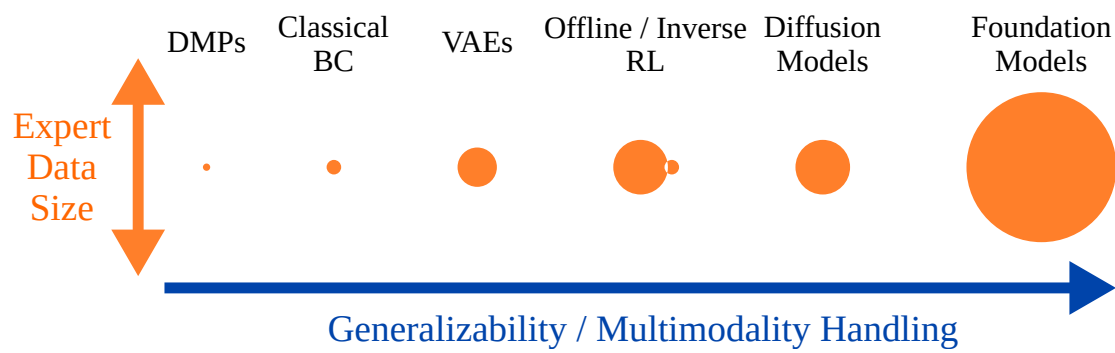


Figure 2.12 Comparison of different imitation learning approaches for contact-rich tasks across key performance dimensions, including data efficiency, generalisation, and contact handling.

Chapter 3

Safe Reinforcement Learning with Variable Impedance Control for Contact-Rich Robotic Tasks

3.1 Introduction

Exploring unstructured environments is a key capability for future robots, which are expected to operate in unknown and potentially hostile surroundings. The inherent physical interactions in such scenarios pose risks to robots, humans, and the environment. These risks become even more pronounced in the context of automated learning and exploration, for example in reinforcement learning (RL). Ensuring the safety of RL in tasks involving rich contact, therefore remains an ongoing challenge, especially in unstructured environments.

Following the background and problem framing in Chapter 2, this chapter presents the first technical contribution of the thesis. It addresses the most immediate challenge identified earlier, namely how to enable reinforcement learning in contact-rich tasks while preserving safety during both exploration and execution.

This chapter studies safe RL in a maze-exploration scenario as an illustrative example of a contact-rich task. This scenario is inspired by the common construction task of installing electricity cables within walls, as depicted in Fig. 3.1. Traditionally, these tasks are performed manually and are time-consuming and labor-intensive. These characteristics indicate strong potential for robotic automation, but they also raise several challenges: (1) the environment is unknown to the robot; (2) visual perception may be difficult to deploy, making physical contact essential; and (3) frequent contact with the environment creates safety concerns.

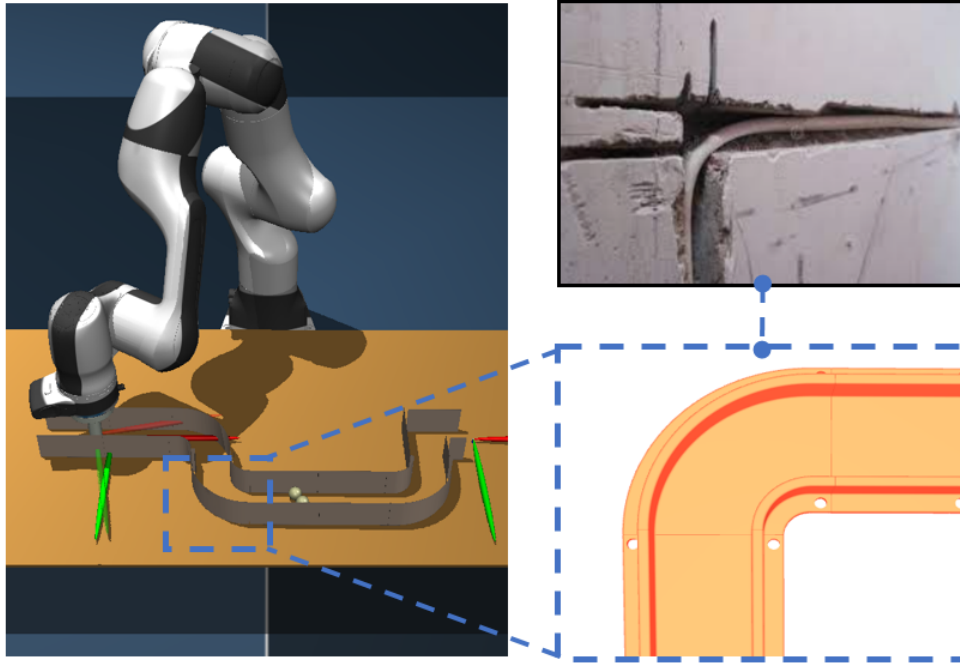


Figure 3.1 Maze exploration is an unstructured contact-rich task that has practical applications such as cable-laying in walls. The agent does not have access to vision, thus it navigates using only contact information.

In tasks involving rich contact, the robot must engage in continuous interaction with the environment, presenting a notable safety risk that could result in damage to either the robot itself or objects in its surroundings. Impedance control is widely used to alleviate such hazards during physical robot interaction [4]. Moreover, variable impedance control (VIC) methods enable adapting impedance parameters to the changing conditions of the task or the surrounding environment [179]. The robot can trade between compliance and accuracy to maximize task performance while staying safe. Nevertheless, determining the adjustment of these parameters in VIC, particularly in unstructured environments with dynamic obstacles, remains a challenging question.

Current state-of-the-art research on contact-rich tasks mainly follows two directions: (1) some studies employ VIC to ensure safe performance while using RL to learn impedance parameters [63, 88]; and (2) other works leverage the predictive capabilities of neural networks to estimate the safety value of future actions, allowing the robot to choose a safer course of action in advance [2, 180]. Each direction has its own advantage: the former adjusts the robot's behaviour reactively based on observed contact forces, while the latter proactively selects safer behaviour by predicting the safety of future actions and thereby avoiding actions that would generate dangerous contact forces.

This chapter takes a step toward combining the advantages of RL-based VIC with those of predictive safe RL. We introduce a new safe-RL framework that incorporates VIC for contact-rich tasks. Following the recovery RL (RRL) method [2], during the pre-training phase we train a *Safety-Critic* network using procedurally collected offline data to predict a risk value from the current state and the next action. Meanwhile, we pre-train a *recovery policy* network that can sample a safe action when the predicted risk exceeds the safety threshold. In the online training phase, a task policy is trained with the goal of maximizing the reward.

The main contributions of this chapter are:

1. A safe RL framework with VIC for contact-rich tasks without prior knowledge of the environment. Before action execution, a *safety critic* predicts the risk of the action and activates a *recovery policy* in risky situations, while VIC regulates contact force during execution to further enhance safety.
2. A combination of proactive safety (prediction and avoidance) with reactive safety (compliance adaptation), providing comprehensive risk management during physical interaction.
3. Demonstration that the policy trained in simulation can be deployed on a physical robot without fine-tuning, successfully completing the task across different obstacle types, maze sizes and shapes, and flange sizes.

3.2 Approach

In this section, we present the proposed framework for safe RL with VIC. We first introduce the impedance control method used in our framework, then we describe the safe RL approach and how it is integrated with VIC. A diagram of the proposed framework is shown in Fig. 3.2.

3.2.1 Impedance Control in Task Space

The robotic manipulator rigid body model in joint space may be written as:

$$M(q)\ddot{q} + C(q, \dot{q})\dot{q} + g(q) = \tau + \tau_{ext}, \quad (3.1)$$

where $q, \dot{q}, \ddot{q} \in \mathbb{R}^n$ are the position, velocity, and acceleration vectors in joint space, respectively, and n is the number of DoF. $M(q) \in \mathbb{R}^{n \times n}$ is the joint space inertia matrix, $C(q, \dot{q}) \in \mathbb{R}^n$

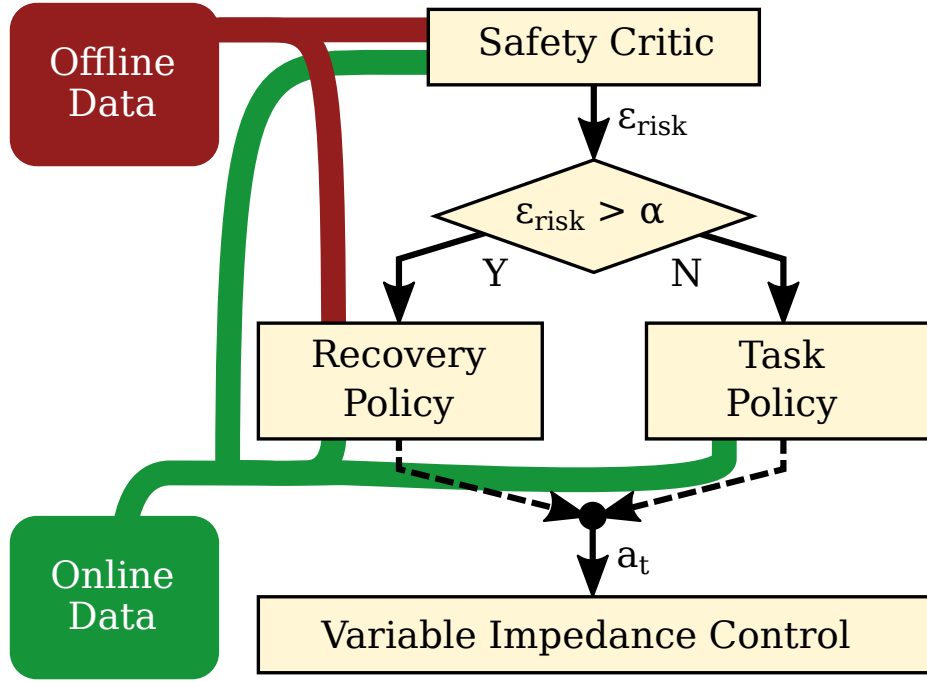


Figure 3.2 The proposed framework: we combine the recovery-based safe RL approach [2] with VIC to solve the contact-rich maze exploration task. We first use an automated procedure to collect the offline data and pre-train our safety critic and recovery policy. Then, we train all learning components using online data. The risk value ϵ_{risk} is used to activate either the task or recovery policy. The action a_t is chosen by the activated component, and it is fed to the VIC. Our action includes a relative position change and desired stiffness vector $\{K_x, K_y\}$.

the Coriolis term, $g(q) \in \mathbb{R}^n$ the gravity term, and $\tau \in \mathbb{R}^n$ is the actuation torques. The external torques generated by the interaction forces are given by $\tau_{ext} = J^T(q)F_{ext}$, being $J^T(q) \in \mathbb{R}^{n \times 6}$ the Jacobian transpose matrix.

The actuation torques are chosen to compensate for the Coriolis and gravitational vectors, plus to deal with the external forces risen by contact with the environment through an impedance model in task space:

$$F_{ext} = D^d \dot{\tilde{x}} + K^d \tilde{x}, \quad (3.2)$$

where $\tilde{x}$ is the Cartesian pose error given by the difference between the desired Cartesian pose $x_d \in \mathbb{R}^6$ and the actual pose $x \in \mathbb{R}^6$. The desired stiffness matrix $K^d \in \mathbb{R}^{6 \times 6}$ is a diagonal matrix with variable terms as $K^d = \text{diag}\{K_x, K_y, K_z, 0.15K_x, 0.15K_y, 0\}$, where the rotation elements for $[K_x, K_y]$ are set as 0.15 times of $[K_x, K_y]$ to reduce and simplify the parameters tuning. We set the rotational z-axis stiffness to zero, to keep it indifferent w.r.t. the end-effector roll. The desired Cartesian damping matrix $D^d \in \mathbb{R}^{6 \times 6}$ is also diagonal and its terms

are defined by the double diagonal rule, i.e., $D_i^d(t) = 2\zeta\sqrt{K_i^d}$ where i represents the linear or rotational components in Task Space. We use $\zeta=0.707$ in our experiments.

3.2.2 Safe RL for Contact-Rich Tasks

We adopt the RRL framework [2] to actively predict safety hazards and take corrective actions. In [2] contact was avoided; on the contrary, in our work continuous contact is required but possibly unsafe. Thus, we introduce impedance control in the task space to better handle the contact-rich aspect of the problem. Specifically, we add the Cartesian impedance parameters to the action space in RL.

As shown in Fig. 3.2, there are three components in our framework: Firstly, similar to RRL, a *safety critic* network and *recovery policy* are trained with the pre-collected offline data. Then, our task policy was trained online with the maze exploration experiments. The output of the task policy is given to the variable impedance controller which controls the robot to achieve the desired actions.

In this framework, the problem is defined as a constrained Markov decision process (CMDP) problem [181]. A tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma, \mu, \mathcal{C})$ denotes a CMDP, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function, P is the state transition probability, $\gamma \in (0, 1)$ is a reward discount factor, μ is the starting state distribution and $\mathcal{C} = \{(c_i: \mathcal{S} \rightarrow \{0, 1\}, \chi_i \in \mathbb{R}) \mid i \in \mathbb{Z}\}$ denote safety constraints that the agent must satisfy. We set γ as 0.9.

State Space, Action Space, and Reward

The state space is differentiated across different networks: it is a 6-dimensional vector for the *safety critic* network and *recovery policy* including 6 force/torque $[F_x, F_y, F_z, T_x, T_y, T_z]$ values measured from a F/T sensor. For the task policy, the position $[P_x, P_y, P_z]$ of the end-effector is added, obtaining a 9-dimensional vector.

We combine 2 impedance parameters $[K_x, K_y]$ with 2 end-effector displacement values $[\Delta P_x, \Delta P_y]$ which consist of a 4-dimensional vector as action space. The sample range of ΔP is $[-0.03, 0.03]$ m and the impedance values $[K_x, K_y]$ are within the sample range of $[300, 1000]$ N/m. We project the stiffness values onto the ΔP vector components as $[K_x^*, K_y^*]$ in the controller:

$$\begin{aligned} K_x^* &= |K_x \cdot \Delta P_x / \|\Delta P\| + |K_y \cdot -\Delta P_y / \|\Delta P\| \\ K_y^* &= |K_x \cdot \Delta P_y / \|\Delta P\| + |K_y \cdot \Delta P_x / \|\Delta P\| \end{aligned} \quad (3.3)$$

We use a negative Euclidean distance between the current position P_{cur} of the end-effector and the goal position P_{goal} multiplied by a constant $c_{pos}=100$ as reward. A penalty is introduced for high-force collisions.

Safety Critic and Recovery Policy

Based on [2], the *safety critic* learns a critic function Q_{risk}^π :

$$Q_{risk}^\pi(s_t, a_t) = \mathbb{E}_\pi \left[\sum_{t'=t}^{\infty} \gamma_{risk}'^{-t} c_{t'} \mid s_t, a_t \right] = c_t + (1 - c_t) \gamma_{risk} \mathbb{E}_\pi [Q_{risk}^\pi(s_{t+1}, a_{t+1}) \mid s_t, a_t] \quad (3.4)$$

where $c_t = 1$ denotes constraint violation, and $c_t = 0$ indicates safety in state s_t . Q_{risk}^π is trained by minimizing a MSE loss function:

$$J_{risk}(s_t, a_t, s_{t+1}; \phi) = \frac{1}{2} \left(\hat{Q}_{\phi, risk}^\pi(s_t, a_t) - \left(c_t + (1 - c_t) \gamma_{risk} \mathbb{E}_{a_{t+1} \sim \pi(\cdot \mid s_{t+1})} [\hat{Q}_{\phi, risk}^\pi(s_{t+1}, a_{t+1})] \right) \right)^2 \quad (3.5)$$

We use an off-policy DDPG algorithm to learn the *recovery* policy π_{rec} by performing gradient descent on the *safety critic* network $\hat{Q}_{\phi, risk}^\pi(s, \pi_{rec}(s))$.

Pre-training

By providing a set of offline data for the agent, we expect the agent to learn the relationship of safety with contact force and motion. It is possible to avoid catastrophic collisions caused by unexpected actions during online training by learning the implicit relationship between the two in advance. Unlike online training where the agent searches for actions that maximise the reward, in this stage, we run a scripted behavior to collect examples of collisions without any reward. This approach reduces the problem complexity by separating task learning and constraint learning which are otherwise interlaced.

We do not pre-train the task policy, because this requires recording complex expert demonstrations which are costly and difficult to obtain for a blind maze exploration task. We pre-train the *safety critic* network and *recovery* policy with the offline data which are collected in simulation. Specifically, we use 256 as batch size and update the *safety critic* network and *recovery* policy 10000 times in the pre-training stage.

Task Policy

The aim of the task policy is to finish the task and obtain maximum rewards within the constraints of safety. We train the task policy by the SAC algorithm [182], which is based on

the idea of maximum entropy. The main difference from other RL algorithms is that SAC maximizes the entropy while optimizing the policy for higher cumulative returns. For our task, the benefit of introducing entropy into the learning stage is the increased randomization, so the agent can sufficiently explore the state space to avoid getting stuck in the local optimum.

Our task policy works in complementarity with the recovery network. Action is selected as follows depending on the risk value:

$$a_t = \begin{cases} a_t^{\pi_{\text{task}}} & Q(s_t, a_t^{\pi_{\text{task}}}) \leq \epsilon_{\text{risk}} \\ a_t^{\pi_{\text{rec}}} & Q(s_t, a_t^{\pi_{\text{task}}}) > \epsilon_{\text{risk}} \end{cases} \quad (3.6)$$

where $a_t^{\pi_{\text{task}}} \sim \pi_{\text{task}}(\cdot | s_t)$ and $a_t^{\pi_{\text{rec}}} \sim \pi_{\text{rec}}(\cdot | s_t)$. Different from task policy, the *recovery* policy is trained to minimize the $\hat{Q}_{\phi, \text{risk}}^{\pi}(s_t, a_t)$ so that it can provide a safe action $a_t^{\pi_{\text{rec}}}$ with small risk value.

3.3 Evaluation

In the following evaluation experiments, we aim to study whether the proposed framework is: (1) safer; (2) more efficient, in comparison to our baselines. We label the baselines as Std_RL-VIC (standard RL with VIC, without the recovery mechanism), SRL-K300 (safe RL with constant stiffness $K=300$ N/m), and SRL-K1000 (safe RL with constant stiffness $K=1000$ N/m).

3.3.1 Experiment Setup

We evaluate the proposed framework on a maze exploration task using the 7-DoF Franka Emika Panda arm in MuJoCo simulation environment. We use a peg-shaped flange mounted on the end-effector of the robot with the diameter of 30 mm and the length of 55 mm. The maze channel is 50 mm in width and 70.35 cm in length. The goal for the agent is to reach from entrance to exit without constraint violations. We adhere to safety level 1 [183], in which the agent will receive a penalty when it violates the constraints, and the current training episode will be terminated.

This problem combines two challenges of finding the way to exit without vision, and staying safe by avoiding high contact forces. It creates a dilemma as the contacts are absolutely necessary to navigate while they are also risky. Furthermore, we add movable obstacles in the maze that requires the robot to push harder. Thus, resorting to a conservative

policy cannot solve the problem. The agent needs to learn both the safe behaviour and the task achievement simultaneously.

We implement the obstacles as three heavy balls in simulation, and a pile of bolts in real-world. We position the obstacles in the middle of the maze. We choose MuJoCo simulator (version 2.3.3) for its favorable multi-contact physics computation, suitable for contact-rich tasks. We used V-HACD library to decompose our maze into 128 convex primitive shapes and assembled them into the same maze.

3.3.2 Simulation Experiments

Offline Data Collection

In order to teach the notion of risk, we generate random states and actions that are likely to cause constraint violation. The robot moves to one of the six predefined points inside the maze, then moves in a random direction, which is likely to collide with the maze walls. The number of transitions are split equally between each point. We add noise to the coordinates of each point, so that we have many different starting positions. Then, we sample a *move direction* as an angle between 0° and 360° and a *move size* in the action range. Lastly, the robot is commanded to move in that direction and the transition is stored as a tuple $(s_t, a_t, c_t, s_{t+1}, done)$. Offline data for pretraining are given 40018 transitions of data, 185 of these transitions contain constraint-violating states.

Online Training

We train each setup online for 3 runs with 1500 episodes in each run. In each episode, the robot needs to move to the start point and then begin exploring the maze, until it meets termination conditions. We train the policy using 1 chunk of 20 cores and one GPU (NVIDIA Tesla V100 16Gb) from the cluster Franklin High Performance Computing (HPC). The average training time for 3 runs of each experiment SRL-VIC, Std_RL-VIC, SRL-K1000, SRL-K300 is 14.47, 7.32, 9.08, 68.28 hours respectively.

Fig. 3.3 shows the evolution of the robot’s behaviour with increasing episodes of online learning. In the beginning, the agent fails because it violates the constraints while interacting with the obstacles. The agent learns to complete the task in later episodes. The stiffness ellipsoids indicate that the agent learns to adapt the stiffness to different conditions such as obstacle or wall contact. We observe that the trajectory becomes smoother and the robot finishes the task faster with further training.

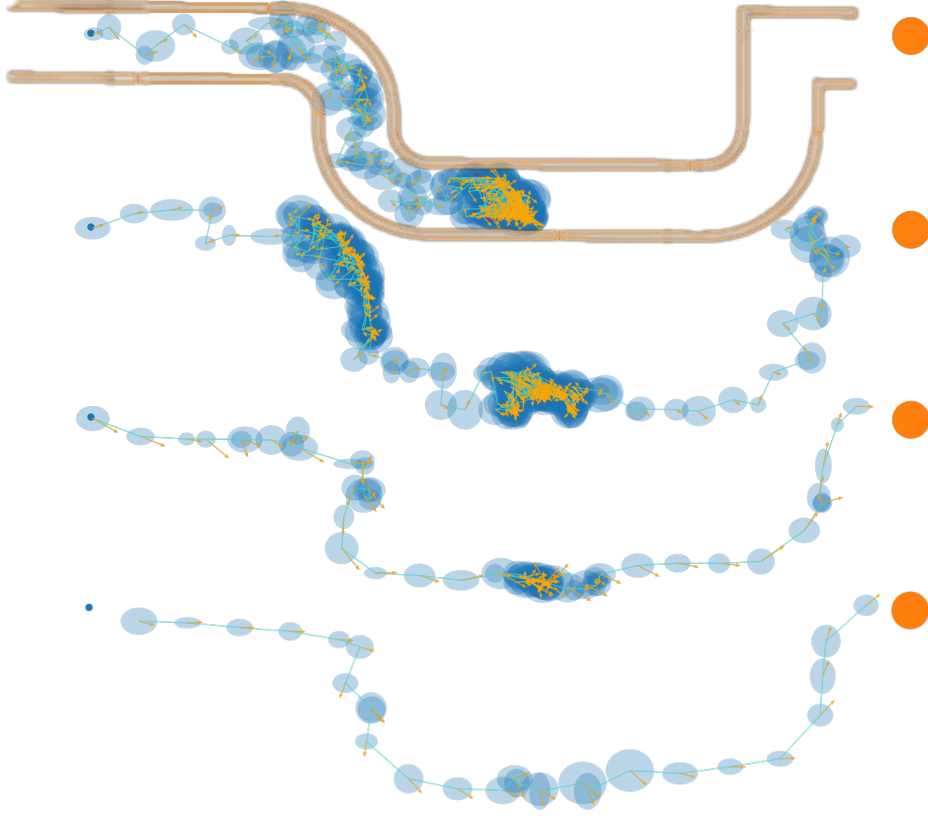


Figure 3.3 Trajectories of the end-effector during different stages of *training*, in top-down view. The orange circle denotes goal point of the task, the cyan circles and yellow arrows along the trajectory represent the stiffness ellipsoids and the motion directions respectively.

Results

We compare the proposed framework with baselines in four different perspectives: ratio of successes/violations, reward, cumulative task successes, and cumulative violations. The results of the simulation training are presented in Fig. 3.4.

In terms of safety, our method achieves minimal cumulative violations apart from SRL-K300. Although SRL-K300 has the minimum cumulative violations among all setups, it gets stuck in the obstacle area and cannot move forward. Thus, our method has an overwhelming advantage in the ratio of successes/violations. We also tested the final learned model of SRL-VIC 100 times in simulation and observed that it was successful in all cases without any violation.

Our method achieves the best results also in terms of task successes. Although Std_RL-VIC performed fairly well in terms of reward and cumulative task successes, our framework improves it further. The difference is even stronger in the success/violation ratio results. Mean-

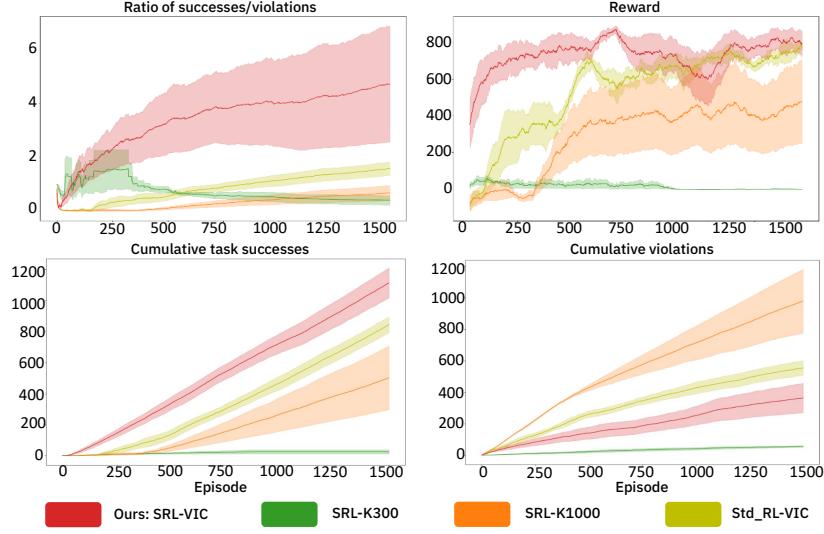


Figure 3.4 Learning curves for maze exploration. Our framework (SRL-VIC) outperforms others, particularly in terms of ratio of successes/violation and cumulative violations, i.e., our framework is safer than others. Furthermore, it learns faster, achieving success at an earlier stage.

while, our method significantly outperforms the other *recovery*-based methods (SRL-K300 and SRL-K1000). SRL-K300 cannot obtain high performance because it is not stiff enough to deal with obstacles. Fig. 3.5 shows the advantage of our method over SRL-K300 when encountering obstacles.

Safety Critic Behaviour

In order to understand the behaviour of our *safety critic* network, we visualise the relationship between the predicted risk value and the state/action values. The robot was commanded to repetitively apply a random action 200 times; starting on a random point (with 0.5 cm deviation), and moving in random direction (in $[0^\circ, 360^\circ]$) with random move size and stiffness. We recorded the risk value produced for each of these cases and plotted them in Fig. 3.6.

As seen in the figure, higher ΔP and K values are correlated with higher ϵ_{risk} . This matches the human intuition that stronger actions are riskier. We also see that the model is sensitive to the contact force, as the spikes in F are aligned with the spikes in ϵ_{risk} . However, it does not merely follow the force, as there is a more complex relationship between the state/action values and risk. These outcomes confirm that the system acts reasonably.

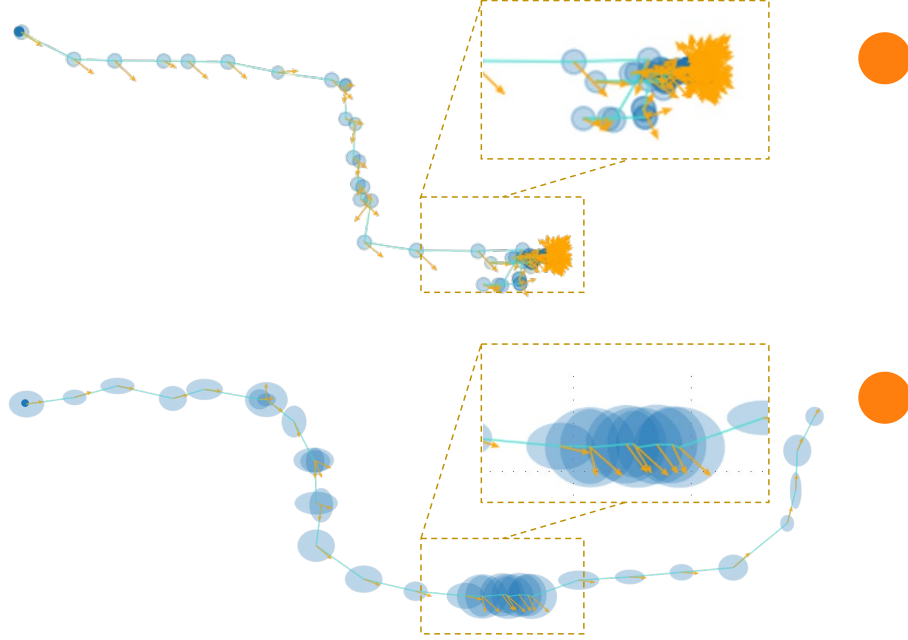


Figure 3.5 Different behaviours when encountering obstacles at training episode 1000. Top (SRL-K300): The robot cannot move forward and fails due to low stiffness. Bottom (SRL-VIC): VIC approach switches to high stiffness when there are some obstacles on the way. The method learns to safely push the obstacles without exceeding the force threshold.

3.3.3 Real-World Experiments

We trained the policy in MuJoCo for the contact-rich maze-exploration task and deployed it on a physical 7-DOF Franka robot arm without any fine-tuning. In the setup, we use Robot Operating System (ROS) where RL policy is a separate ROS node communicating with the physical robot.

Our policy succeeded in 4 out of 5 trials at first. In the failed case, the robot got stuck at a turning point. It was not highly robust to the real-world environment because our policy was trained with observations in simulation. This is because of the physical dynamics differences and the imprecision of the F/T sensing in simulation. For this reason, we retrained the policy by adding Ornstein-Uhlenbeck (OU) noise in state observations. The success rate increased to 6/6 after retraining.

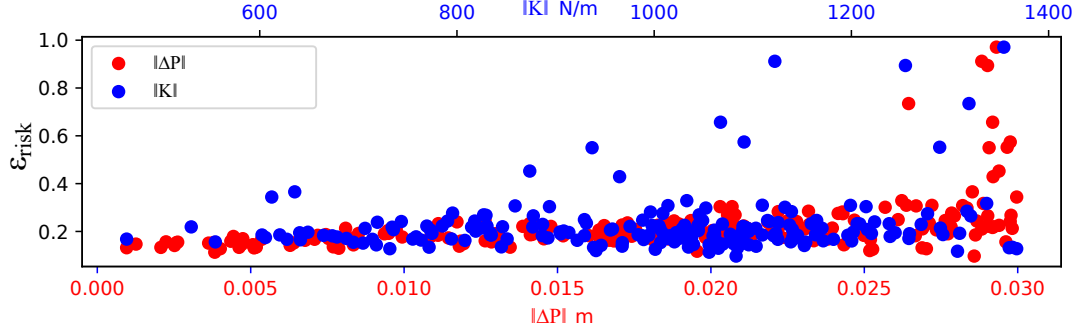


Figure 3.6 Relation between the risk value and state/action values in the offline test. Top: The relationship of the risk ϵ_{risk} to stiffness $\|K\|$ and desired move $\|\Delta P\|$ values. Bottom: Plot of ϵ_{risk} and contact force $\|F\|$. The force magnitude is scaled down by $1/100$.

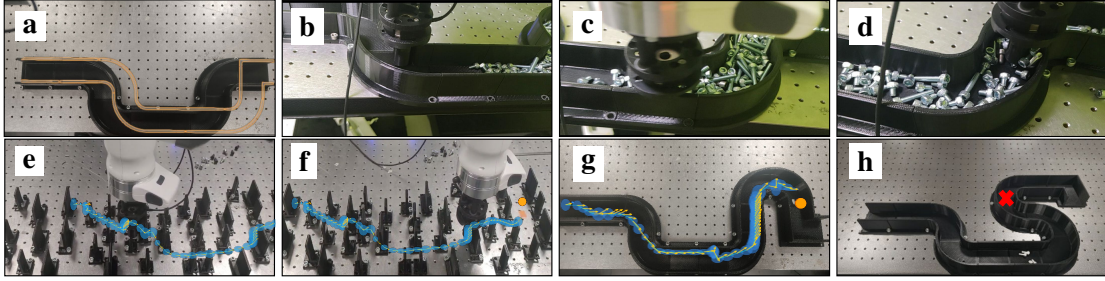


Figure 3.7 Real-world experiments. (a) *Maze shape-1*, without obstacles. The yellow overlay shows the original maze from the simulation. (b) The robot meets the pile of bolts halfway. (c) The obstacles create congestion at the curve. (d) The SRL-VIC policy overcomes the obstacles by increasing the stiffness. (e, f) The *complex task scenario* with small wall segments at new positions. (g) *Maze shape-2*. (h) *Maze shape-3*, failed at red X marker.

Fig. 3.7 shows the deployment on a physical robot. We put bolts of different size as dynamic obstacles (50 M6×30 for *small obstacles*, 100 M6×8 and 50 M6×30 for *mixed obstacles*).

The end-effector trajectories in real-robot experiments are shown in Fig. 3.8. The policy adapts the stiffness values effectively by becoming stiff in the goal direction and staying compliant laterally. It also employs distinct strategies for encounters with walls at curves and collisions with movable obstacles. For the latter, it adopts a larger, rounded stiffness ellipse to facilitate pushing through in all directions.

We also plot the measured force (F_y), stiffness K_y and move size (ΔP_x) values w.r.t. y -position in Fig. 3.9. The stiffness adaptation policy becomes more evident in these plots, comparing the *no obstacle* case to the others. The 2nd and 3rd cases have larger stiffness in

the obstacle area as they observe higher force, while the 1st case increases stiffness only at the curves, while touching the wall.

Generalization Results

The SRL-VIC policy transfers successfully from simulation to real world. The transfer success was increased by the addition of state randomization in the simulation training. In the real-world experiments, we tested the generalization capability of our safety model by changing the position of the maze curves and placing wall segments at new positions. The maze length varied between simulation (70.35 cm) and the real-world (84.62 cm). Additionally, the nature of the obstacles differed from those present in the simulation. The learned policy successfully completed most of these cases, suggesting that the safety mechanism generalizes to different wall positions. The agent can also complete the *Maze shape-1* when tested by changing the flange diameter (1, 2, 3 and 4 cm). Our task policy is trained on a single maze shape; thus, it could not generalize to the case shown in Fig. 3.7(h). It still did not violate a constraint, but it could not find the exit.

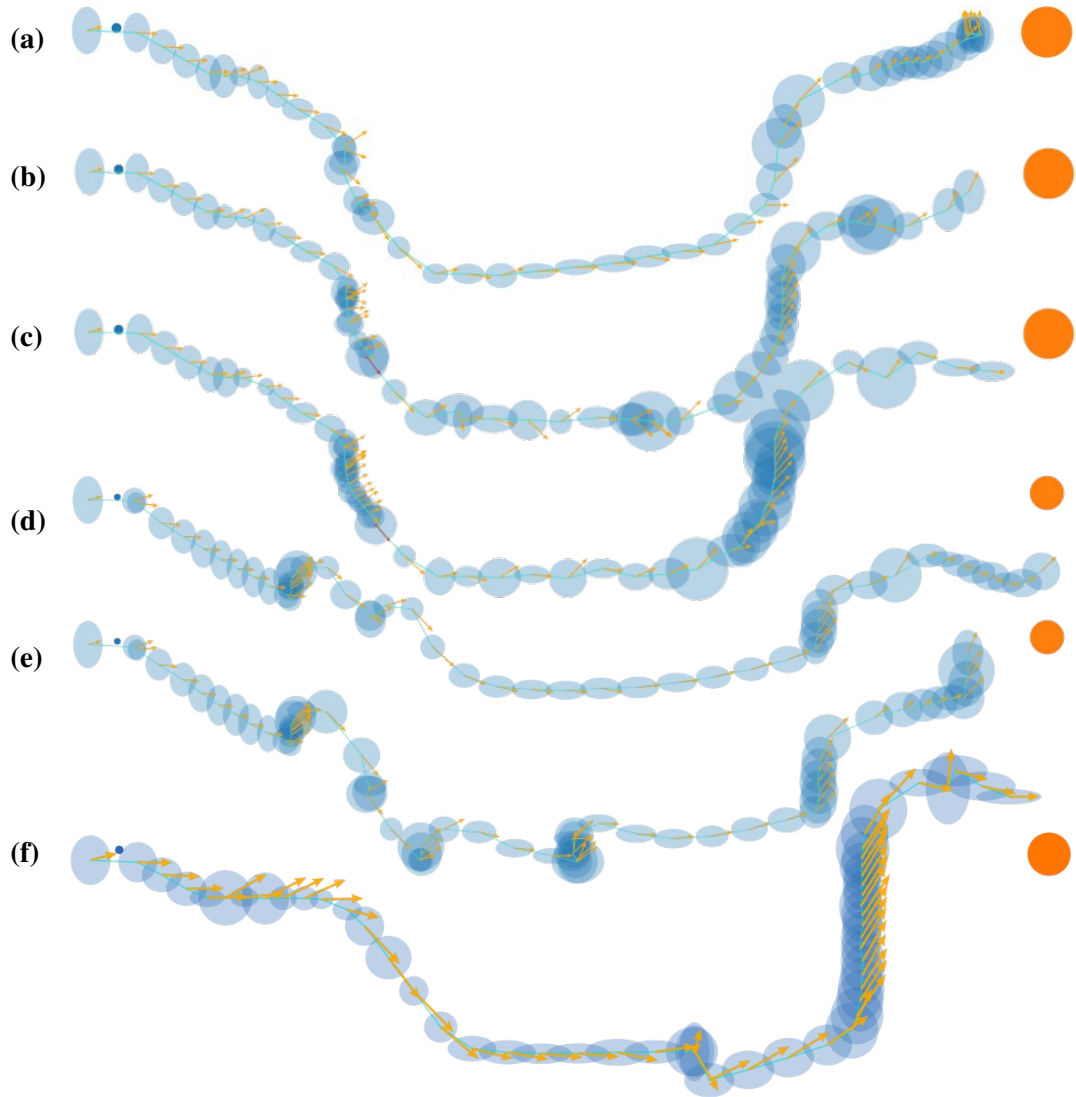


Figure 3.8 Trajectories of the end-effector during executing task in the real world. The orange circle denotes goal point of the task, the cyan circles and yellow arrows along the trajectory represent the stiffness ellipsoids and the motion directions respectively. (a-c) Maze shape-1 (Fig.3.7.a): (a) *no obstacle*, (b) *small obstacles*, (c) *mixed obstacles*. (d, e) *Complex task scenario with discrete walls*, (f) *Maze shape-2* (Fig.3.7.g).

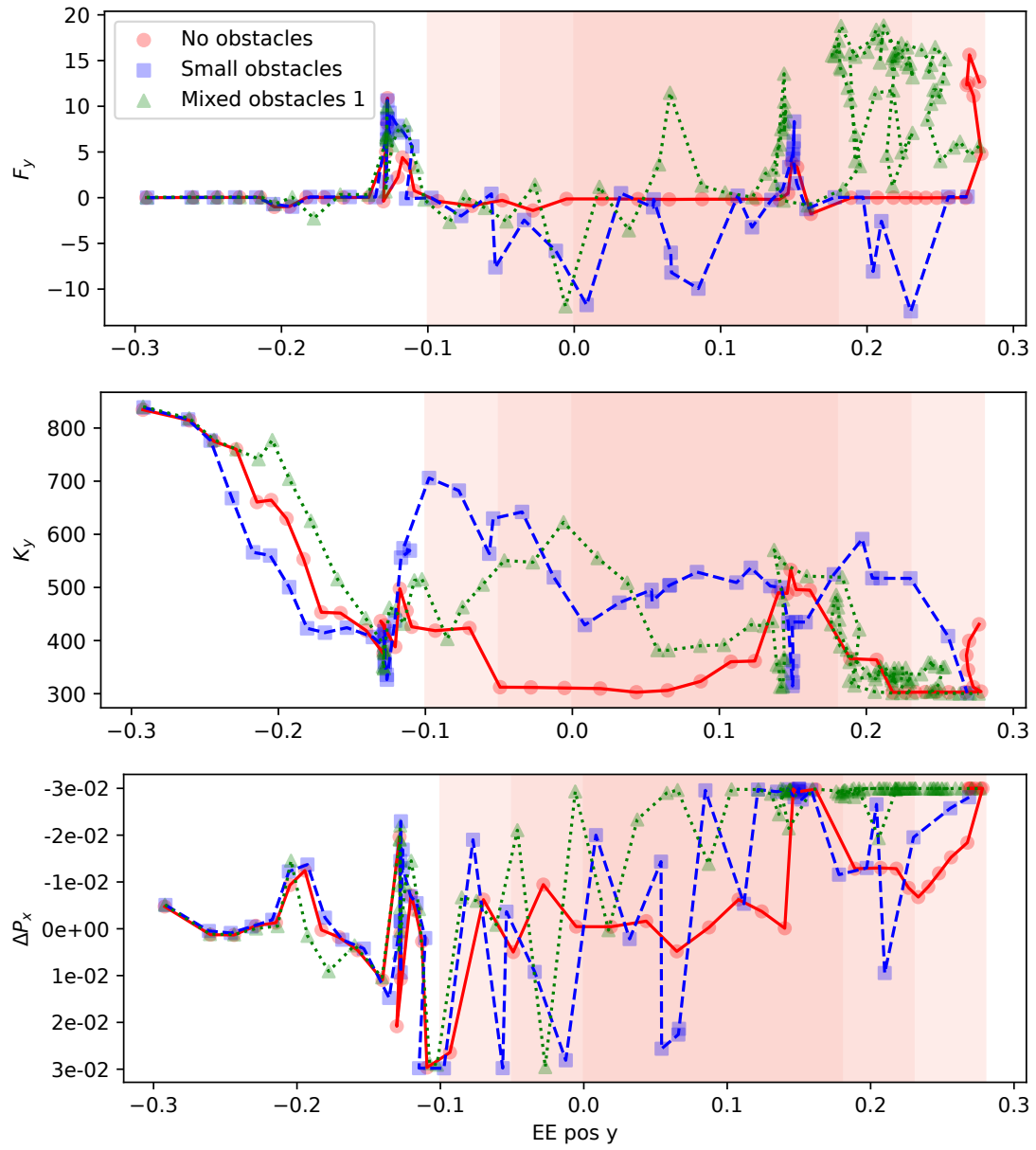


Figure 3.9 Force (F_y), move size (ΔP_x), and stiffness (K_y) values during the real-robot experiments plotted against the y - position. The pink highlight shows the approximate placement of obstacles.

3.4 Summary

We presented a framework that integrates the strengths of a safe RL method with VIC. This combination addresses the dual challenges of ensuring safety and maintaining continuous physical contact. Experiments in contact-rich maze tasks demonstrate that our framework achieved a good trade-off between task accomplishment and collision-based safety, and outperformed Std_RL-VIC. The inclusion of the VIC adds increased adaptability, creating a better trade-off between safety and success in comparison to the constant impedance approaches. The trained model can transfer to real-world, and also the safety model can generalize to different obstacles and wall positions.

This chapter has established the foundational safe RL framework for contact-rich tasks by combining predictive risk estimation with variable impedance adaptation. While this formulation improves safety at the action-selection stage, it still relies on intervention at the decision level. The next chapter builds on this foundation by introducing a bio-inspired reflexive mechanism that operates at a lower control level and can therefore respond more rapidly to emerging hazards during action execution.

Chapter 4

Reflexive Safe Reinforcement Learning for Contact-Rich Robotic Tasks

4.1 Introduction

Robotic actions in the real world present two major challenges: the complexity of unstructured environments and the safety hazards associated with physical interactions [12]. RL-based robotic systems have the potential to address both challenges to enable effective automated learning and exploration in such environments [184]. Traditionally, the complexity challenge has received significant attention, while the safety challenge has gained focus more recently, especially in contact-rich tasks. Drawing inspiration from the animal kingdom, where evolutionary processes have led to the development of solutions to these challenges, this chapter proposes a novel approach. Specifically, this chapter draws on the reflex mechanisms inherent to vertebrates to enhance the safety and robustness of RL systems.

The complexity challenge can be mitigated by imposing a hierarchy between long-term task-level actions and short-term motor-level actions. Early RL studies have shown the advantage of decomposing complex tasks into smaller subtasks [185]. Furthermore, it was recently shown that considering actions in task-space leads to more efficient learning, in comparison to using joint-space actions [88]. Thus, we assume a hierarchy between the high-level decision-making and low-level control loops as shown in Fig. 4.1.

The safety challenge has gained more attention recently due to the black-box nature of traditional learning-based RL models in safety-critical environments, under the umbrella of safe RL [183]. Specifically, we follow the hierarchical safe RL methods, which answer the task-solving and safety problems through separate learned components. A crucial component is a *safety critic* network that estimates the risk of failure to limit access to risky states [186].

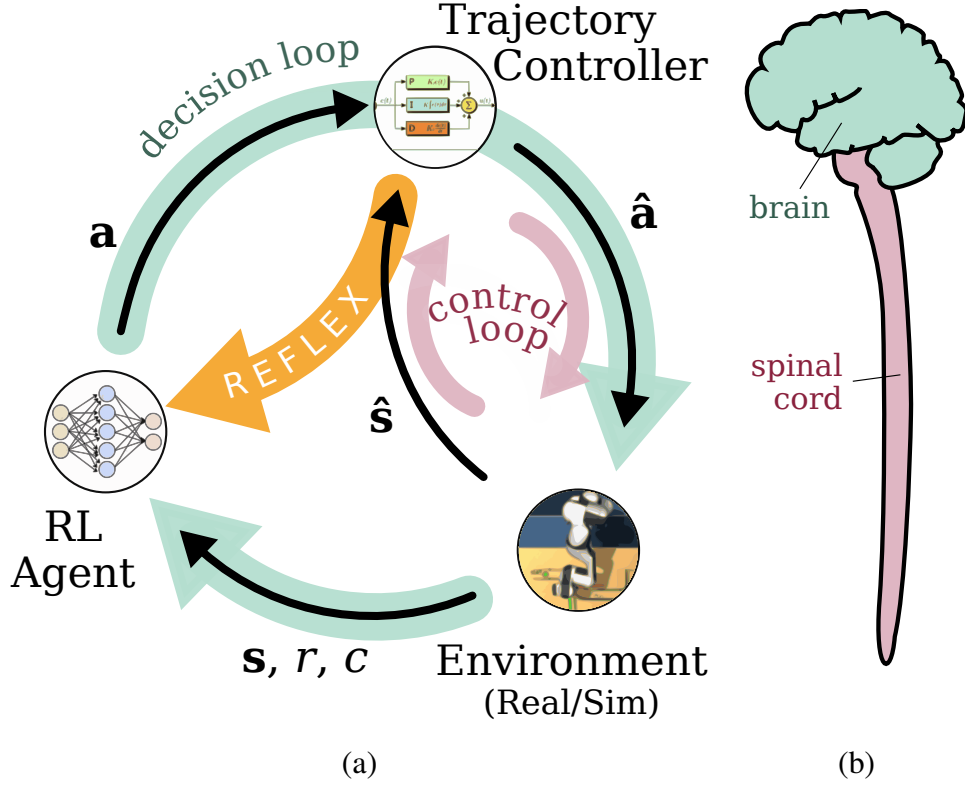


Figure 4.1 **a)** Bresa framework. The RL agent operates at the *decision loop*, planning the high-level action $\mathbf{a}$ that is executed by the trajectory controller. The controller operates at a high-frequency *control loop*, executing the low-level action $\hat{\mathbf{a}}$ based on the state feedback $\hat{\mathbf{s}}$ at each control step. The reflex mechanism gives the system a quick reaction capability by interrupting the control loop in the case of high risk. **b)** A simplified illustration of the human central nervous system. While high-level decisions are made in the brain, safety-related reflexes are managed by the spinal cord, allowing for faster responses that override slower, more complex decision-making processes.

Furthermore, a separate *recovery policy* can be deployed in addition to the usual *task policy* for learning more complex risk-evasive behavior [2]. Integrating variable impedance control (VIC) into hierarchical safe RL is shown to enhance contact-rich interaction safety further through adaptable compliant behavior (Chapter 3). However, existing safety methods primarily function at the decision-making stage, often overlooking risks that emerge during low-level action execution.

In this sense, Bresa can be viewed as a direct extension of the SRL-VIC framework developed in Chapter 3: rather than only improving which action is selected, we now address how safety can be enforced during the execution of that action at the control-loop level.

We propose **Bresa**, a **B**io-inspired **R**eflexive **H**ierarchical **S**afe RL method inspired by biological reflexes. Our method decouples task learning from safety learning, incorporating

a *safety critic* network that evaluates action risks and operates at a higher frequency than the task solver. Unlike existing recovery-based methods, our *safety critic* functions at a low-level control layer, allowing real-time intervention when unsafe conditions arise. The task-solving RL policy, running at a lower frequency, focuses on high-level planning (decision-making), while the *safety critic* ensures instantaneous safety corrections.

The main contributions of this chapter are:

1. A bio-inspired hierarchical safe RL method (Bresa) that decouples task learning from safety learning, incorporating a *safety critic* that operates at a higher frequency than the task solver.
2. A reflex mechanism that interrupts the control loop in real-time when unsafe conditions arise, enabling instantaneous safety corrections.
3. Validation on multiple tasks including a contact-rich robotic maze exploration task, demonstrating reflexive ability to enhance safety and adaptability in unforeseen dynamic environments.
4. Real-world deployment on a 7-DOF Franka robot arm achieving 100% success rate in 10 consecutive runs.

4.2 Method: Bio-inspired Hierarchical Reflexive Safe RL

The main concept of Bresa methodology is summarized in Fig. 4.1. A key element of our approach is the bio-inspired reflex mechanism, designed to rapidly interrupt the control loop and swiftly respond to predicted dangers. Our method follows the existing line of hierarchical safe RL works that separate the task learning and safety learning by learning a value function for assessing the risk of taking an action in a given state [186, 2].

4.2.1 Safe RL Problem Formulation

We follow the constrained Markov decision process (CMDP) formulation [181] to define the safe RL problem. A CMDP environment $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, P, \gamma_{\text{task}}, \mu, \mathcal{C})$ consists of the state space $\mathcal{S}$, the action space $\mathcal{A}$, reward function $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, the state transition probability P , reward discount factor $\gamma_{\text{task}} \in (0, 1)$, the starting state distribution μ and the safety constraints $\mathcal{C} = \{(c_i : \mathcal{S} \rightarrow \{0, 1\}, \chi_i \in \mathbb{R})\}$, where $c_i=1$ indicates the violation of i^{th} constraint.

4.2.2 Hierarchical Safe RL Baseline

We train two policies online, π_{task} for completing the task and π_{recovery} for evading unsafe states. We train also a safety critic Q_{risk} to estimate risk incurred by a state-action pair ($\mathbf{s} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$). Given a safety threshold $\epsilon_{\text{safe}} \in \mathbb{R}$, the estimated risk $\epsilon_{\text{risk}} \in \mathbb{R}$ of the task action determines which policy is sampled in the decision loop:

$$\begin{aligned} \epsilon_{\text{risk}} &= Q_{\text{risk}}(\mathbf{s}, \pi_{\text{task}}(\mathbf{s})), \\ \mathbf{a} &= \begin{cases} \pi_{\text{recovery}}(\mathbf{s}) & \text{if } \epsilon_{\text{risk}} > \epsilon_{\text{safe}} \\ \pi_{\text{task}}(\mathbf{s}) & \text{otherwise.} \end{cases} \end{aligned} \quad (4.1)$$

Q_{risk} is pre-trained before the task exploration phase using simple action samples that are procedurally collected offline. The training of Q_{risk} continues online during the RL exploration. We use a different discount factor (γ_{safe}) in critic training as the risk depends on shorter term effects than the task.

4.2.3 Reflex Mechanism

The complete flowchart and pseudo-code of the Bresa algorithm are shown in Fig. 4.2. The flowchart showcases the explicit relationships between the *decision* and *control loops*. The task-solving policy operates at the low-frequency *decision loop*, while the safety critic operates at the high-frequency *control loop*. In case a danger arises, the safety critic has the power to interrupt the current controller command quickly and activate the recovery policy.

Note that Q_{risk} is learned as a neural network that has predictive capabilities. Differently than a simple force check, it learns a complex multi-dimensional relationship that enables force contacts subject to other conditions.

We illustrate the benefit of the reflex mechanism on an obstacle avoidance scenario in Fig. 4.3. Given a high-level action $\mathbf{a}$, the low-level controller executes smaller intermediate actions that we call *minor action* $\hat{\mathbf{a}}$. As the intermediate state $\hat{\mathbf{s}}$ drifts towards high-risk area due to the action noise, Bresa triggers a reflex to establish safety.

When the reflex happens, the intended action is interrupted, and consequently only a proportion of $\mathbf{a}$ gets executed. Training the models using this action decreases the transparency, and consequently, undermines the learning performance. Thus, we add the executed action $\mathbf{a}'$, rather than the intended action $\mathbf{a}$ into the training dataset.

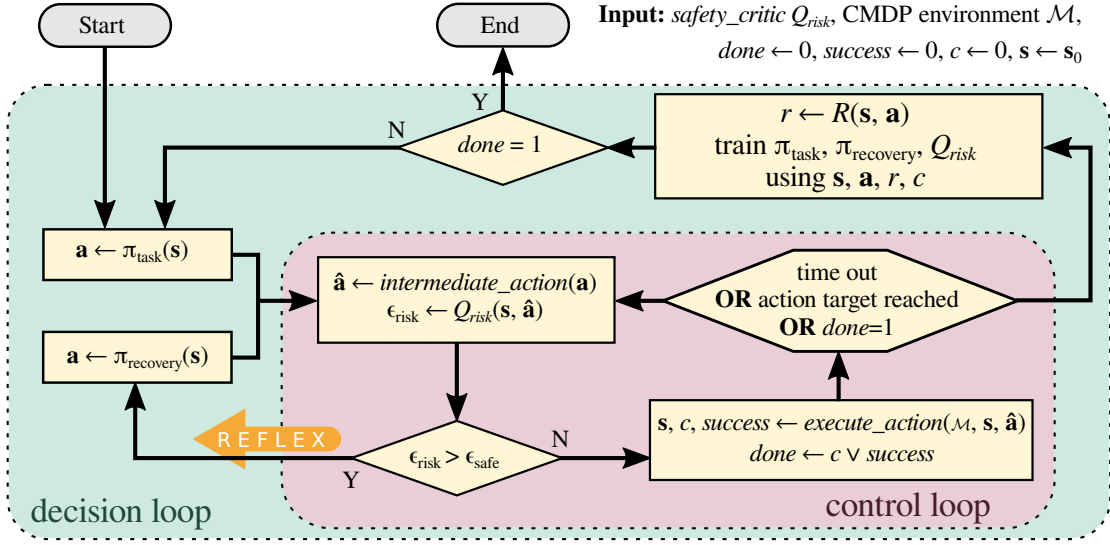


Figure 4.2 Flowchart of the Bresa algorithm. We color-coded the *decision loop*, *control loop* and *reflex* for comparison to Fig. 4.1. The safety critic operates at the control loop frequency, enabling real-time reflexive responses. The maze exploration environment in the MuJoCo simulator is shown on the right: the robot physically interacts with the maze walls and the obstacles through an end-effector flange equipped with F/T sensor.

4.2.4 Trajectory Control and VIC

The low-level trajectory controller in our framework is tasked to execute the high-level action $\mathbf{a}$ in the control loop. It is abstracted to be any controller that achieves subtasks of $\mathbf{a}$. In this work, we use linear interpolation between the state $\mathbf{s}$ at the beginning of the action and the desired target state.

In the contact-rich maze exploration task, we employ a VIC to compliantly execute the obtained linear trajectory. The stiffness actions K_1 and K_2 represent the stiffness along the major and minor axes of motion. We form the translational stiffness matrix in the world frame as $\mathbf{K}_t = \mathbf{R}_p^\top \mathbf{K}_a \mathbf{R}_p$, where:

$$\mathbf{R}_p = \begin{bmatrix} \Delta p_x & -\Delta p_y & 0 \\ \Delta p_y & \Delta p_x & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{K}_a = \begin{bmatrix} K_1 & 0 & 0 \\ 0 & K_2 & 0 \\ 0 & 0 & K_z \end{bmatrix}. \quad (4.2)$$

The effective stiffness matrix is formed as $\mathbf{K} = \text{diag}(\mathbf{K}_t, \mathbf{K}_r)$ with the fixed rotational stiffness $\mathbf{K}_r = \text{diag}(100, 100, 0)$. Translational stiffness K_z is fixed at 750 N/m. $\mathbf{K}$ is used in the VIC control as described in Chapter 3.

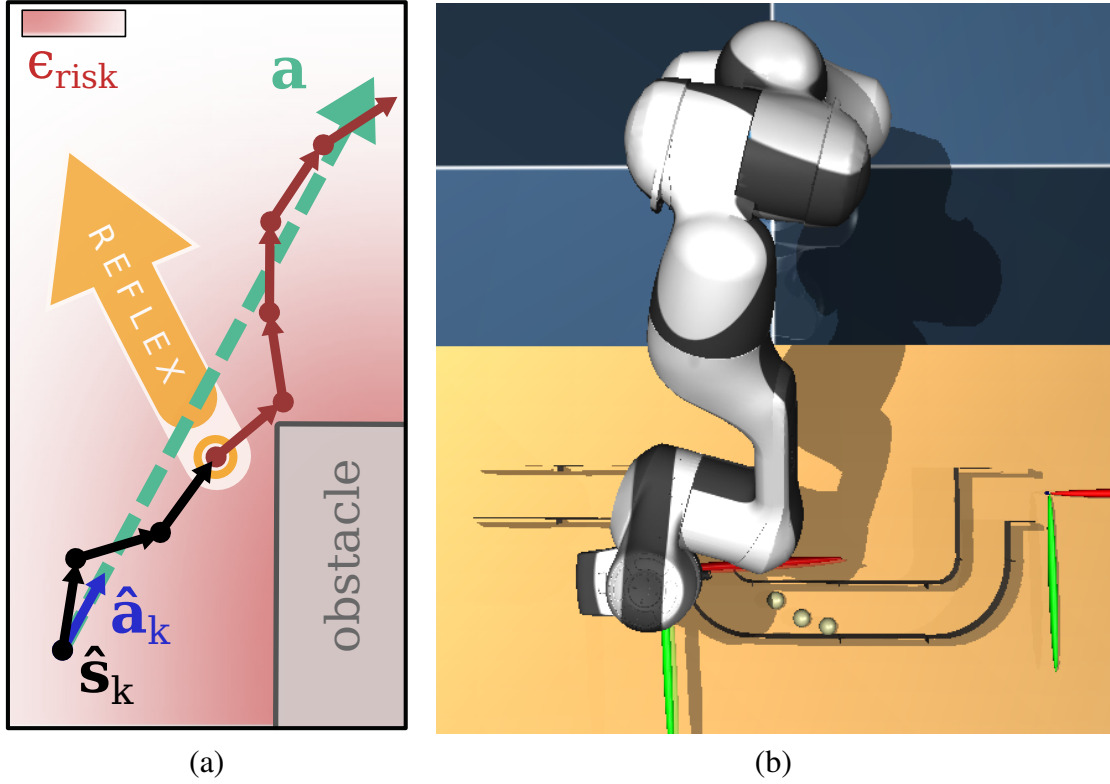


Figure 4.3 **a)** Reflex mechanism on an obstacle avoidance scenario. Even when the high-level state-action pair (s, a) is evaluated to be safe, an intermediate state-action pair $(\hat{s}, \hat{a})$ may entail high risk ($\epsilon_{\text{risk}} > \epsilon_{\text{safe}}$) and trigger the reflex mechanism. The stochasticity of the environment leads to a drift in the outcomes of minor actions $\hat{a}$. **c)** Maze exploration environment in the Mujoco simulator. The robot physically interacts with the maze walls and the obstacles through an end-effector flange equipped with F/T sensor.

4.3 Evaluation

4.3.1 Navigation Task

We first validate Bresa on a 2D navigation task where a point robot must navigate to a goal while avoiding obstacles. The state space is 4-dimensional (position and velocity), and the action space is 2-dimensional (velocity commands). The safety constraint is the distance to obstacles.

Fig. 4.4 shows the trajectories in the navigation task. The colormap shows the risk value predicted by the safety critic along the trajectory. Bresa shows fine-grained risk prediction, while the baseline shows coarse-grained risk prediction. This demonstrates that

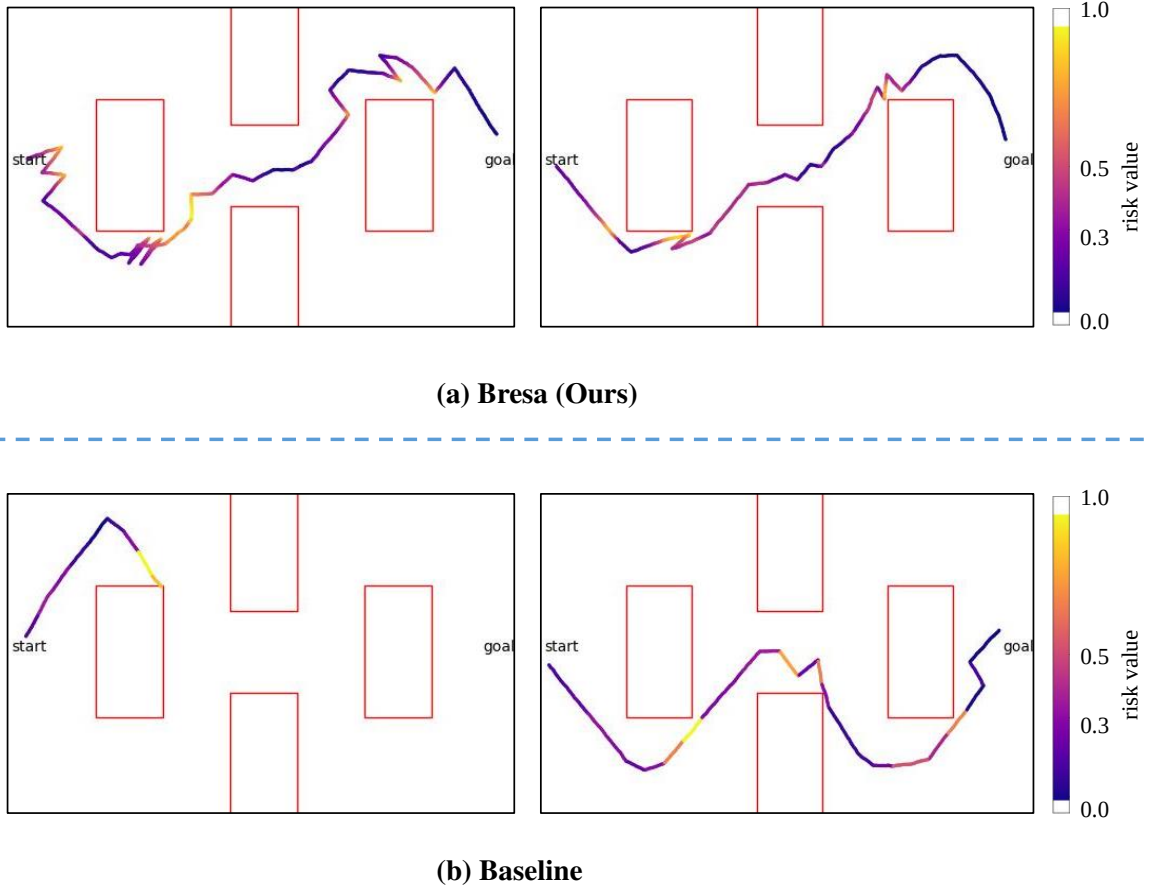


Figure 4.4 Reflexive mechanism in Navigation task during training. The risk value along with exploration is plotted in a colormap showing fine-grained risk prediction in our method while it is coarse-grained in the baseline. Blue annotations show single high-level actions.

the reflex mechanism enables the robot to react to sudden changes in the environment, such as unexpected obstacles, without violating safety constraints.

4.3.2 Maze Exploration Task

In this task, the 12-dimensional state-space consists of the position, linear velocity, measured force and measured torque of the end-effector. The 4-dimensional action-space is formulated as the end-effector position change $\Delta \mathbf{p}$ in the range of $[-a_{\max}, a_{\max}]$ and stiffness K_1, K_2 in the range of $[300, 1000]$. We use $a_{\max}=0.03$ m in this study. The safety constraint c in this task is the contact force threshold that the robot cannot exceed when interacting with

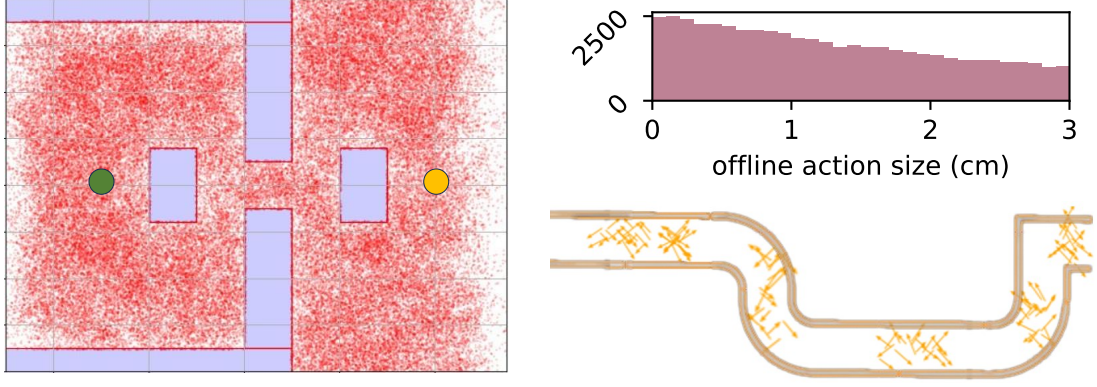


Figure 4.5 Offline data collection locations in both tasks. Left: navigation task. The green and yellow circles indicate start and goal points, and red dots indicate the sampled start positions. Upper right: histogram of exponentially sampled action sizes in the maze exploration task. Lower right: sampled action locations on the maze.

the environment. We set the force threshold as 30 N as a trade-off between safety and task success.

Offline Data Collection

Randomly sampled short-term action data was collected in five areas in the maze as shown in Fig. 4.5, where orange arrows indicate random actions with uniformly random stiffness $K_1, K_2 \sim \mathcal{U}(300, 1000)$ (N/m) and uniformly random direction $\angle \Delta \mathbf{p} \sim \mathcal{U}(-\pi, \pi)$ (rad). The magnitude of the position change is sampled from an exponential distribution to favor smaller actions for *minorization* as $\|\Delta \mathbf{p}\| \sim \exp(\lambda) a_{\max}$. We used $\lambda=1$ in our experiments and discarded values larger than $a_{\max}$. The offline data consist of 50000 tuples of $[\mathbf{s}, \mathbf{a}, \mathbf{s}', c]$ with 6034 constraint violations.

Training in Simulation

We train the policy in a contact-rich maze task. SAC [182] is used to train task policy while DDPG [187] is used for recovery policy and safety critic model. We run each experiment for 500 episodes and repeat the training for 12 seeds. We use the optimal discount factor parameters for both methods as $\gamma_{\text{safe}}=0.675$ and $\gamma_{\text{task}}=0.90$. The other important parameters are SAC temperature $\alpha=0.2$, safety threshold $\epsilon_{\text{safe}}=0.45$, and horizon $H=300$.

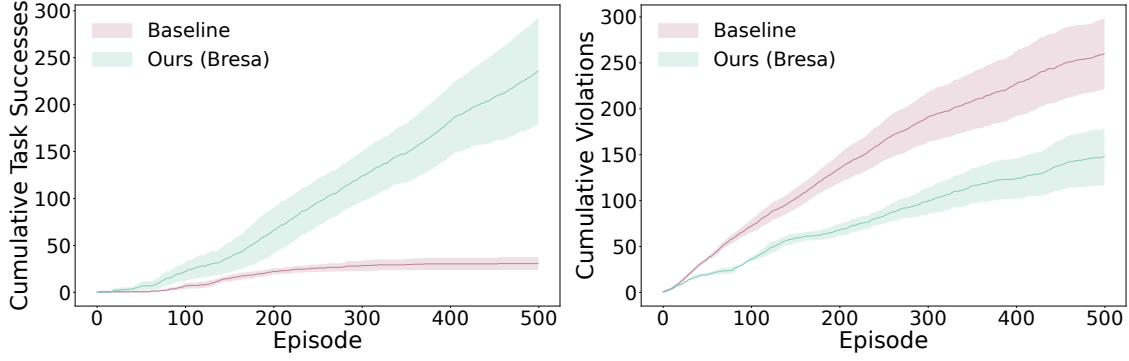


Figure 4.6 Overall performance of Bresa in contact-rich maze exploration task in terms of cumulative successes and violations. Bresa significantly outperforms the baseline in both metrics.

Table 4.1 Model performance in maze exploration task

Total runs	Success	Violation	Ratio	Success rate
200	196	2	98	98%

Results

On the average of 12 seeds, Bresa achieves 81.6 successes and 165.6 violations (ratio: 0.49) while the baseline achieves 10.5 successes and 271.1 violations (ratio: 0.04), showing a significant improvement with the help of the reflexes. Furthermore, we also report the results of best 3 seeds: Among 500 episodes of training, Bresa achieves 236.3 successes and 148.0 violations on average, while the baseline has only 30.7 in success but 260.7 in violation on average. The best 3 seed results are shown in Fig. 4.6, demonstrating that our method outperforms the baseline in terms of success, violation and the ratio between them.

Moreover, we investigate how the reflex mechanism works in Bresa during training, showing the risk prediction in Fig. 4.7. We can see that the *safety critic* predicts the risk value according to the current state and the next action. The trajectory of the end-effector is shown in the lower part of Fig. 4.7, where the ellipsoids indicate stiffness and the red color shows the *recovery action* in high-risk situations. The reflex appears in the key location where a significant change occurs, such as the turns and obstacle contact.

4.3.3 Model Test

We take the trained model of the best performing seed and run it 200 times to test the model performance in the contact-rich maze exploration task. The model achieves 98% success rate and only 2 violations, as shown in Table 4.1.

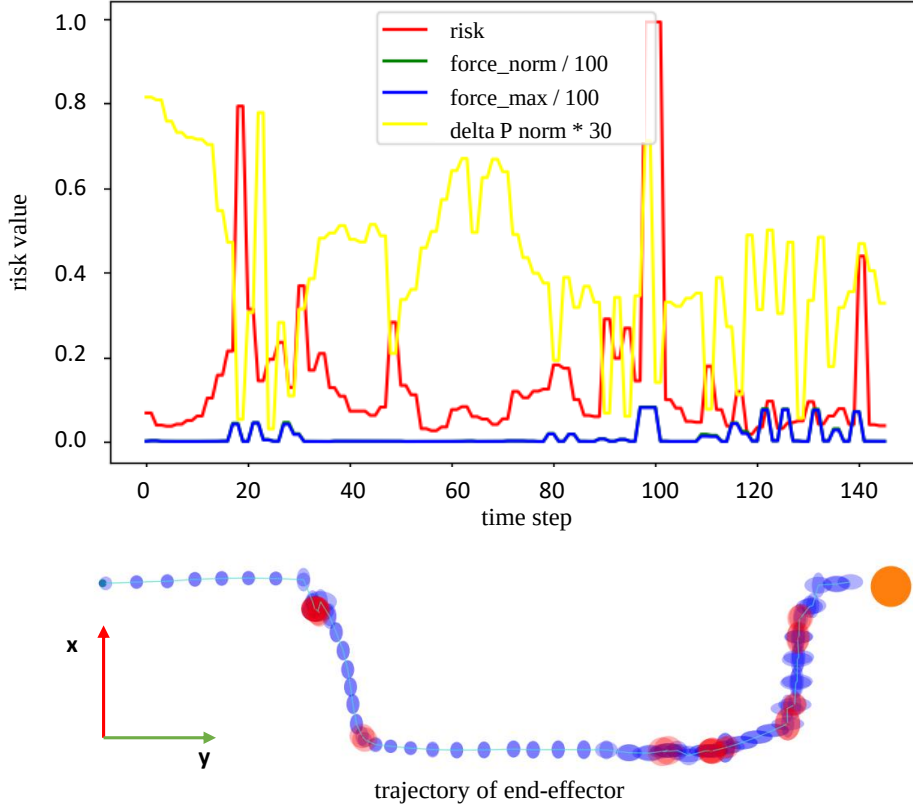


Figure 4.7 Illustration of model performance from a specific episode. Upper: reflexive response of *risk critic* according to the current state and action. Lower: trajectory of end-effector where the ellipsoids indicate stiffness and red color shows high-risk value.

4.3.4 Parameter Studies

We performed parameter studies on the navigation task to evaluate the impact of different action size $a_{\max}$, discount factors γ_{task} , and noise scalar σ , for a comprehensive assessment of our method's performance. These experiments are essential for gaining a deeper understanding of the method's behavior, showcasing its advantages, and pinpointing opportunities for refinement.

1. **Action size $a_{\max}$:** We compare the results of different $a_{\max}$ values (1,2,3,4). The results show that the action size of 3 achieves the best performance, as shown in Fig. 4.8. As discussed in the introduction, high-level actions increase the sample efficiency by decreasing the decision-making frequency. However, larger actions also have higher risk to violate constraints.

Table 4.2 Performance in real world

Total runs	Success	Violation	Ratio	Success rate
10	10	0	–	100%

2. **Action noise σ** : By default we use a noise scalar of $\sigma=0.02$, however, we also tested the effect of different σ values (0.02, 0.04, 0.08). The results in Fig. 4.8 show that our method is more robust to higher noise values. Comparatively, our method improved the success-violation ratio over the baseline by 292% for $\sigma=0.02$, 750% for $\sigma=0.04$, and 656% for $\sigma=0.08$, in average over 20 seeds.
3. **Discount factors $\gamma_{\text{safe}}, \gamma_{\text{task}}$** : We experimented with varying values of these discount factors. Each experiment is run 120 episodes with 24 seeds. As seen in the results of Fig. 4.8, Bresa obtains better success-violation trade-off consistently over different parameters. Both methods maintain a similar number of task successes during the training, however, our method decreases the number of constraint violations significantly. The best γ_{safe} (0.75) is smaller than the best γ_{task} (0.95), because the safety is more immediate (short-term) than the exploration task.

4.3.5 Real-World Experiments

We deployed the policy on a physical 7-DOF Franka robot arm without any fine-tuning. In the setup, the RL policy is run on a separate ROS node that communicates with the physical robot running in a real-time frequency 1000 Hz. We trained the best performing seed from the maze task further in 1500 episodes instead of 500. We tested the performance with 10 consecutive runs in a comprehensive set of scenarios, including added obstacles and human perturbation. The results are presented in Table 4.2 and Fig. 4.9.

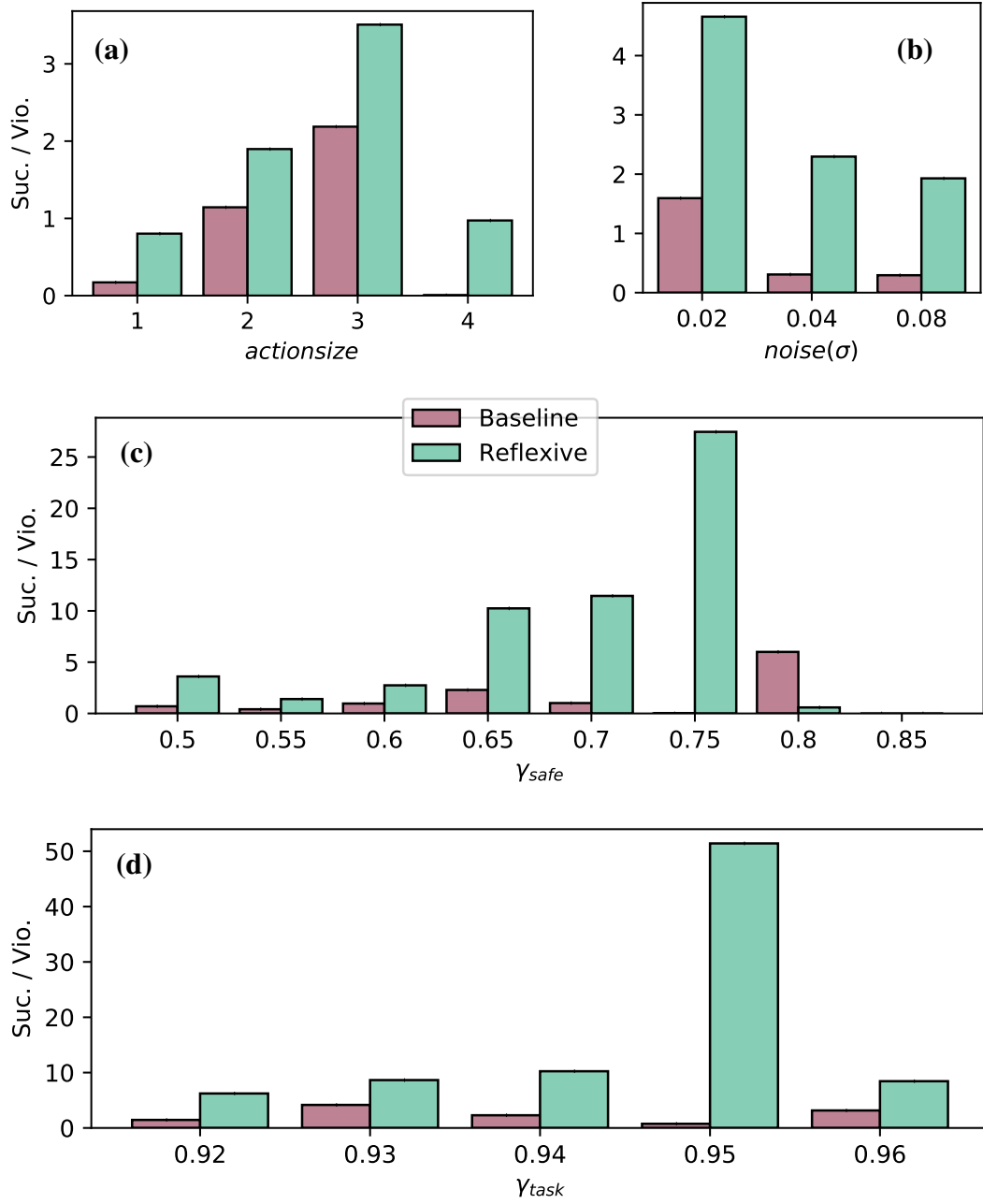


Figure 4.8 Success-violation ratio results over navigation task training with different *action size* $a_{\max}$, *noise* σ , γ_{safe} and γ_{task} parameters shown in subfigure (a), (b), (c) and (d) respectively.

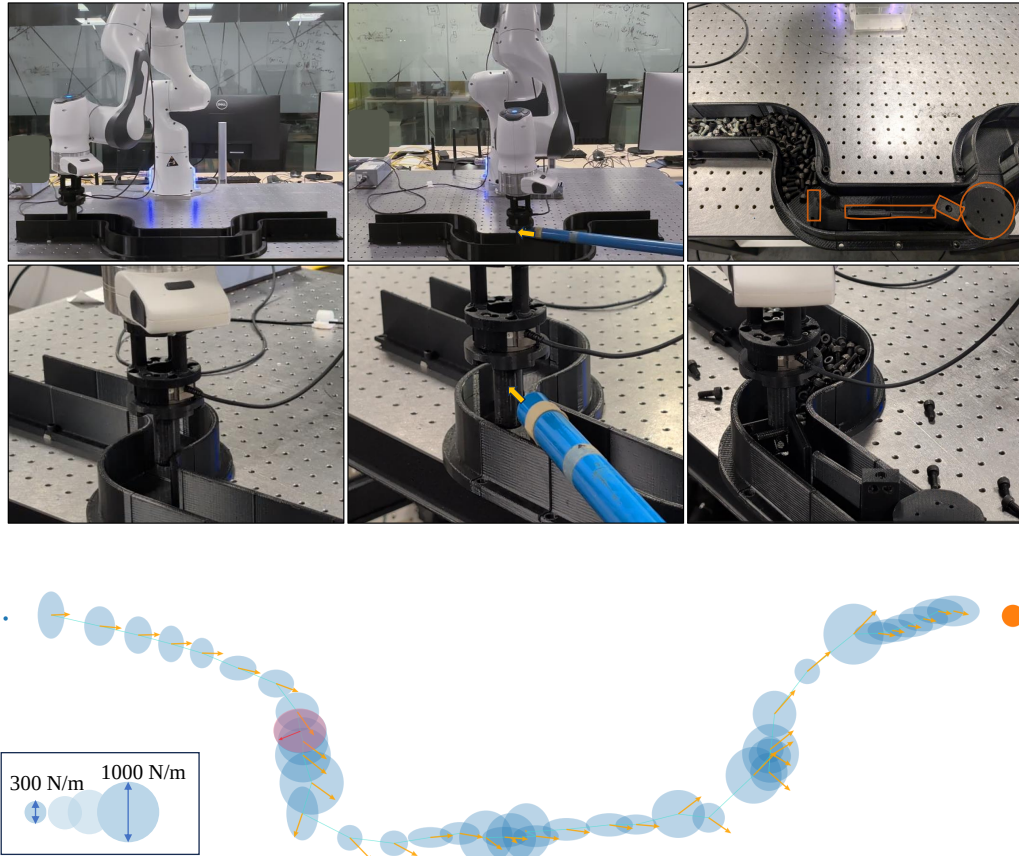


Figure 4.9 Real-world setup. Top: Overview of the entire setup. Bottom: Closeups of the robot in action. Left: The robot makes contact with a wall. Center: A stick applies random external forces during motion. Right: The robot interacts with large, stationary obstacles. Bottom: the trajectory of the external-force run shows robot behavior, where the ellipses indicate the stiffness. The red ellipse indicates the recovery action.

4.4 Summary

We presented Bresa, a novel hierarchical reinforcement learning method designed to enhance safety in contact-rich robotic tasks. Inspired by biological reflexes, our approach decouples task learning and safety learning, allowing a risk critic to operate at a higher frequency than the task-solving policy. By integrating low-level risk-aware control with VIC, our method ensures real-time intervention in unsafe situations while maintaining adaptability in dynamic and unstructured environments.

Experimental results demonstrate that Bresa outperforms the baseline, providing a robust and reflexive safety mechanism that bridges the gap between high-level planning and low-level execution. On the average of 12 seeds in the maze task, Bresa achieves a success-violation ratio of 0.49 versus 0.04 for the baseline. The real-world experiments achieve 100% success rate in 10 consecutive runs.

This chapter has introduced the bio-inspired reflexive safety mechanism as a complementary layer to predictive safe RL, improving responsiveness to hazards that arise during execution. However, fast intervention alone does not fully address the question of closed-loop physical stability during sustained interaction. The next chapter therefore builds upon this work by investigating the role of passivity in safe RL for contact-rich tasks and by introducing energy-based control for stronger stability guarantees.

Chapter 5

Passivity-Centric Safe Reinforcement Learning for Contact-Rich Robotic Tasks

5.1 Introduction

After addressing predictive safety in Chapter 3 and reflexive low-level intervention in Chapter 4, we now turn to a complementary but more fundamental question: how to ensure that the learned interaction behaviour remains physically stable during contact.

Reinforcement learning has earned increasing attention and success in addressing complex decision-making and control problems, especially in robotic applications [188]. From manipulation tasks to autonomous navigation, RL offers the potential to achieve unprecedented performance by learning optimal control policies. However, while RL excels in simulated environments, deploying these policies in real-world robotic systems remains a significant challenge due to safety concerns [183]. This is particularly important in contact-rich tasks, which require the robots to interact with their environment safely, particularly in terms of contact force regulation, which is necessary to avoid damaging the environment, the robot, or the objects involved in the task.

Although there has been significant effort to address the contact-rich safety problem, the role of control stability is often neglected within RL. Stability is crucial for maintaining predictable and controllable robot behavior, especially when interacting with complex and unstructured surroundings, which cannot be ignored in real-world robotic deployment [188]. For instance, an RL policy trained without considering system stability might generate control commands that cause high-frequency vibrations or unbounded energy accumulation. Such behaviors can damage the robot, compromise its performance, or pose safety risks to humans.

Despite these risks, a substantial portion of RL research focuses primarily on maximizing task success rates or improving learning efficiency, with limited attention given to the stability of policy. This gap presents a significant barrier to real-world deployment, where stability and safety are not only desirable but essential [189].

Passivity-based control (PBC) is a fundamental method to prove the stability of physical interactions with the environment [190]. It is even more critical for kinematically redundant robots, as internal motions within the null space of the task Jacobian can result in an unintended increase in system energy, potentially leading to instability. Moreover, it is vital to monitor and restrict the amount of energy allowed to be stored and injected by the controller into the robot to ensure the passivity and safety of the system [191]. Energy tanks are widely used to achieve passivity-based control by restricting the amount of energy allowed to be stored in the system [192]. However, as pointed out in [193], the bounding of the energy alone is not sufficient to ensure the passivity of the system, but also the energy flow between the controller and robot has to be limited. Most of the research mentioned above focuses on the design of the energy tank and its parameters, while few works have investigated learning-based passive system design. Especially its inclusion in RL-based control learning is not sufficiently explored. Although some works [114, 194] leverage imitation learning method to obtain a stable VIC, labor-collected demonstrations are needed. We follow recent works [192, 195, 196] with an augmented energy tank approach that also restricts the energy flow.

In this work, we aim to explore different modes of combination between RL and PBC, and study their implications. We term the standard RL approaches that do not include any passivity treatment as *passivity-agnostic* RL. Then, we distinguish between the *passivity-filtered* approach, that imposes a passivity filter at the RL output; and the *passivity-aware* approach which aims to learn the passivity in the RL policy. Subsequently, we propose the *passivity-ensured* RL that includes both passivity guarantee and energy efficiency by combining the two approaches, as summarized in Table 5.1.

The main contributions of this chapter are:

1. An extensive comparison of different RL-based passive control learning approaches (Table 5.1), and their effects on both training and deployment phases, evaluated from safety, stability, energy efficiency and task completion perspectives.
2. Design and evaluation of the augmented energy tank-based passivity formulation in RL, with varying energy tank and energy flow constraints.

Table 5.1 Comparison of passivity modes in RL

	passivity-agnostic	passivity-filtered	passivity-aware	passivity-ensured
passivity guarantee	×	✓	×	✓
energy efficiency	×	×	✓	✓

3. Introduction of a passivity-ensured safe RL framework that simultaneously answers both control passivity and contact-rich safety problems.

Accordingly, this chapter shifts the focus from avoiding unsafe states alone to regulating the energy exchanged during interaction, thereby connecting safe reinforcement learning with passivity-based control theory.

5.2 Method: Passive Safe Reinforcement Learning

The presented approach draws upon passive control and safe reinforcement learning fields to learn safe and stable policies for solving contact-rich tasks. It uses a passivity-based approach to observe and limit the energy stored in the system such that it can both: (1) impose constraints on the safe RL to train passivity-aware policies, and (2) achieve passivity-ensured control by a passivity layer running on the policy output.

The former approach by itself does not guarantee passivity due to the incompleteness of data and the stochasticity of learning methods. The latter approach provides this guarantee, but can significantly reduce task performance by disrupting the task-oriented policy behavior. We combine both approaches to achieve good task performance while ensuring stability. A diagram of our proposed method is shown in Fig. 5.1.

5.2.1 Safe RL for Contact-Rich Tasks

The baseline safe RL framework (*passivity-agnostic* in our terminology) is shown in Fig. 5.2. The RL agent consists of different components such as the *safety critic* model and *recovery policy*, in addition to the classic RL task policy. The *safety critic* is a separate network that predicts the potential risk of state-action pairs [186]. In the case of high risk, a *recovery policy* is activated to escape from risky areas [2].

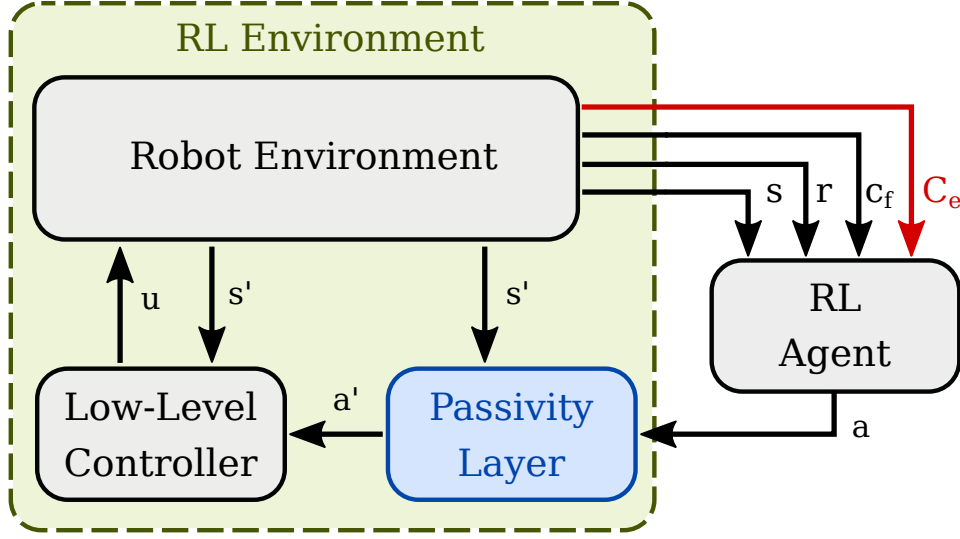


Figure 5.1 *Passivity-ensured Safe RL*: The proposed method adds both *energy-based constraints* C_e to enable learning the passivity-awareness and an active *passivity-filtering layer* between the RL agent and the low-level controller for guaranteed passive control. The signals are the state s , reward r , action a , force constraint c_f , energy-based constraints C_e , and control command u .

Safety can be formulated by introducing constraints to the RL policy in the form of a constrained Markov decision process (CMDP) [181]. A CMDP is expressed as a tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma, \mu, \mathcal{C})$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function, P is the state transition probability, $\gamma \in (0, 1)$ is the reward discount factor, μ is the starting state distribution and $\mathcal{C} = \{(c_i: \mathcal{S} \rightarrow \{0, 1\}, \chi_i \in \mathbb{R})\}$ denotes the constraints that the agent must satisfy.

For this specific formalism, we define the state space, action space, reward function, and safety constraints as follows:

- **State Space $\mathcal{S}$** : A 6-dimensional vector for *safety critic* network and *recovery policy* including the wrench $[f_x, f_y, f_z, t_x, t_y, t_z] \in se^*(3)$ measured from an F/T sensor. Additionally, a 9-dimensional state vector extends this representation by incorporating the end-effector position $[p_x, p_y, p_z] \in \mathbb{R}^3$.
- **Action Space $\mathcal{A}$** : A 4-dimensional vector consisting of two end-effector displacement values $[\Delta p_x, \Delta p_y] \in \mathbb{R}^2$ and 2 stiffness parameters $[k_1, k_2]$. The stiffness values are defined w.r.t. the motion vector $\Delta \mathbf{p}$ with $k_1 \in \mathbb{R}$ along the motion vector, $k_2 \in \mathbb{R}$ orthogonal to the motion vector and parallel to the ground plane, and k_z normal to the ground plane.

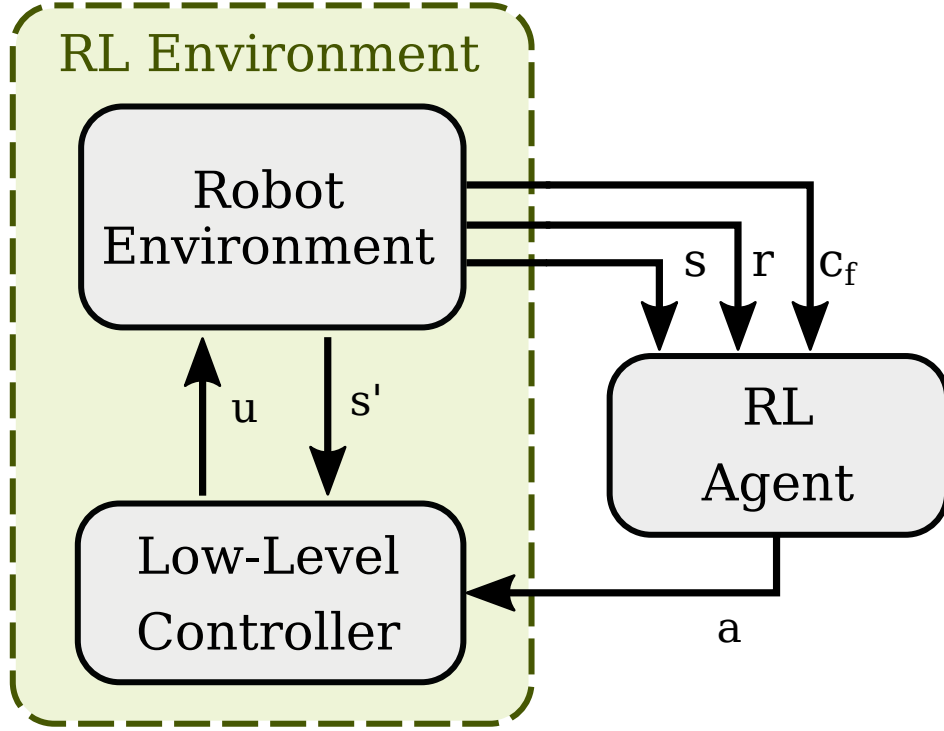


Figure 5.2 *Passivity-agnostic Safe RL*: the baseline RL framework that we extend. The baseline includes the force-based safety constraints c_f in addition to the usual state s , action a and reward r observations of an RL framework.

- **Reward Function R** : The agent is penalized for the Euclidean distance between the current position of the end-effector and the goal position multiplied by a constant $r_{pos} = -400$. High-force collisions are penalized with $r_{col} = -250$. A bonus is given for reaching the goal $r_{goal} = 1000$ to increase convergence speed.
- **Constraints $\mathcal{C}$** : The force constraint c_f requires the measured force magnitude to stay below a threshold of 40 N. We also propose two new constraints for passive control: energy budget constraint c_{eb} and energy flow constraint c_{ef} .

5.2.2 Cartesian Impedance Control

The control scheme utilized in this work is a Cartesian VIC, for which the control torques of the impedance term $\tau_{IC}^\top \in \mathbb{R}^n$ in a quasi-static condition consists of the torques generated by the cartesian spring $\tau_K^\top \in \mathbb{R}^n$ and the damping term $\tau_D^\top \in \mathbb{R}^n$,

$$\tau_{IC}^\top = \tau_K^\top - \tau_D^\top. \quad (5.1)$$

The torques generated by the cartesian spring $\tau_K^\top$ are generated by the elastic wrench $\mathbf{w}_K^{0,EE^\top} \in se^*(3)$, which can be expressed as

$$\mathbf{w}_K^{0,EE^\top} = \begin{bmatrix} \mathbf{K}_t & \mathbf{K}_c \\ \mathbf{K}_c^\top & \mathbf{K}_r \end{bmatrix} \Delta \eta. \quad (5.2)$$

Here, the infinitesimal displacement of the twist $\eta \in se(3)$ is described by $\Delta \eta \in se(3)$. The positive definite matrix $\mathbf{K}_c \in \mathbb{R}^{3 \times 3}$ describes the coupling between the stiffness values of the rotational ($\mathbf{K}_r \in \mathbb{R}^{3 \times 3}$), and translational ($\mathbf{K}_t \in \mathbb{R}^{3 \times 3}$) axes. In this work the values of translational stiffness matrix $\mathbf{K}_t$ are obtained by mapping the RL action stiffness matrix $\mathbf{K}_a = \text{diag}(k_1, k_2, k_z)$ to the world frame as follows $\mathbf{K}_t = \mathbf{R}_p^\top \mathbf{K}_a \mathbf{R}_p$, given

$$\mathbf{R}_p = \begin{bmatrix} \Delta p_x & -\Delta p_y & 0 \\ \Delta p_y & \Delta p_x & 0 \\ 0 & 0 & 1 \end{bmatrix} \in SO(3).$$

The damping related torques τ_D are generated by Cartesian damping matrix $\mathbf{D} \in \mathbb{R}^{6 \times 6}$, which is computed directly from the stiffness matrix using a modal decomposition. We compute the eigen decomposition $\mathbf{K}_t = \mathbf{Q} \Lambda \mathbf{Q}^\top$, where Λ contains the eigenvalues and $\mathbf{Q}$ contains the eigenvectors of $\mathbf{K}_t$. The damping matrix is then defined as $\mathbf{D} = 2\mathbf{Q}(\xi\sqrt{\Lambda})\mathbf{Q}^\top$, where ξ is a scalar coefficient. This approach ensures that $\mathbf{D}$ remains symmetric and positive definite while being consistent with the variable stiffness updates.

We calculate the joint torques using the spatial Jacobian matrix $\mathbf{J} \in \mathbb{R}^{6 \times n}$, rewriting (5.1) as

$$\tau_{IC}^\top = \mathbf{J}^\top (\mathbf{w}_K^{0,EE^\top} - \mathbf{D}\eta). \quad (5.3)$$

In the following subsection, we describe the passivity-based approach for guaranteeing the stability of our controller.

5.2.3 Passivity-Based Approach for Variable Stiffness Control

Passivity is a necessity to facilitate physical interactions between a robot and its environment. For an impedance-controlled robot, the energy stored in the system can be expressed as the sum of the storage functions of the controller and robot $H = H_{\text{Ctrl}} + H_{\text{Rob}} \in \mathbb{R}$ [197]. It was shown in [190] that since H_{Rob} is physically bounded, it is also passive. Therefore, the system is passive if and only if the controller described by H_{Ctrl} is passive. Hence, it is a necessity to observe and bound the energy injected by the controller at the power port

$(\tau_{\text{Ctrl}}, \dot{\mathbf{q}})$, which is expressed by:

$$\dot{H}_{\text{Ctrl}} + P_{\text{IC}} = 0. \quad (5.4)$$

Here, P_{IC} is expressed as the sum of the energy flows resulting from the spring $P_K = \mathbf{w}_K^{0, \text{EE}} \dot{\mathbf{x}} \in \mathbb{R}$ and damping $P_D = \dot{\mathbf{x}}^\top \mathbf{D} \dot{\mathbf{x}} \in \mathbb{R}$.

One can ensure the passivity of the overall system via the augmentation of the storage function H_{Ctrl} by an energy tank $E_T \in \mathbb{R}$ that is bounded by the upper and lower limits $E_T^{\text{max}}/E_T^{\text{min}} \in \mathbb{R}$. Thus, the energy flow at the power port can be rewritten as:

$$\dot{H}_{\text{Ctrl}} + \dot{E}_T + P_D \leq 0, \quad (5.5)$$

where the energy flow at the tank in this work is defined as $\dot{E}_T = -P_K$. The system passivity requires that the *tank condition* $E_T^{\text{min}} \leq E_T + \dot{E}_T \leq E_T^{\text{max}}$ to always hold.

However, as pointed out in [193], satisfying the *tank condition* is not enough to ensure the passivity of a system, but it is also a necessity to restrict the energy flow $\dot{E}_T$ between the controller and robot. This can be achieved by introducing a flow limit $\dot{E}_T^{\text{min}} \in \mathbb{R}$, and ensuring that the *flow condition* $\dot{E}_T^{\text{min}} \leq \dot{E}_T$ always holds.

5.2.4 Passivity-Aware Safe RL

This section describes how we add passivity awareness to the RL policy. The safe RL framework achieves safety through constraints in the CMDP. We extend this framework by adding passivity-related constraints as part of the RL problem formulation. Specifically, we introduce two new constraints:

- **Energy budget constraint** c_{eb} : This constraint requires the energy tank to stay within its bounds, i.e., $E_T^{\text{min}} \leq E_T \leq E_T^{\text{max}}$. The constraint is violated when the energy tank is depleted or overflows.
- **Energy flow constraint** c_{ef} : This constraint requires the energy flow to stay above the flow limit, i.e., $\dot{E}_T \geq \dot{E}_T^{\text{min}}$. The constraint is violated when the energy flow exceeds the limit.

The passivity-aware safe RL framework is shown in Fig. 5.3. The energy constraints $C_e = \{c_{eb}, c_{ef}\}$ are added to the CMDP to enable the RL agent to learn passive behavior.

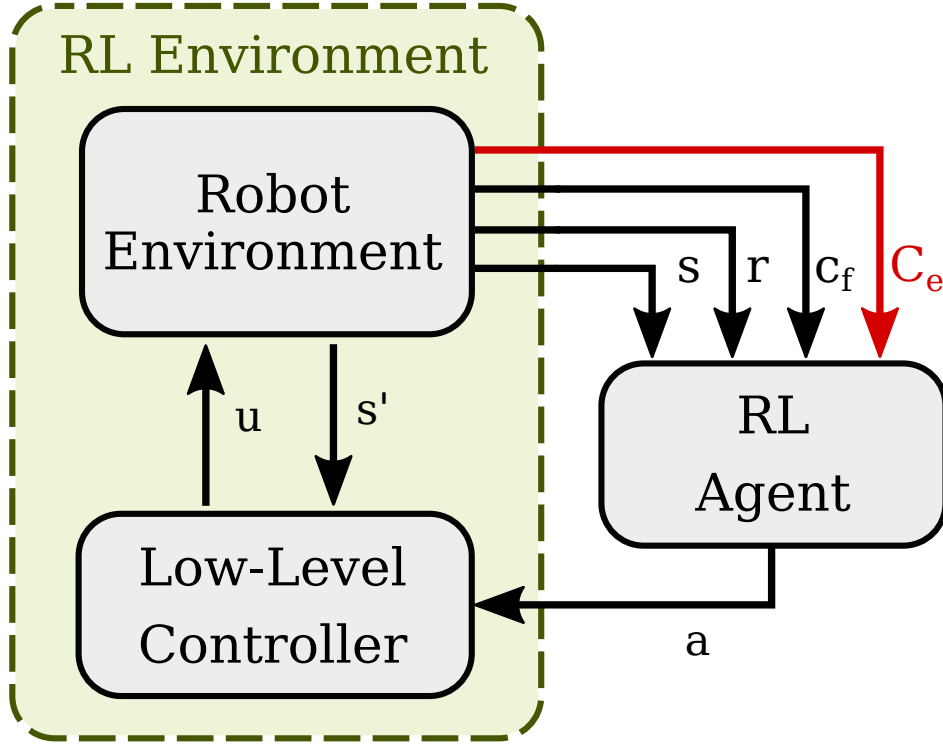


Figure 5.3 *Passivity-aware Safe RL*: We introduce the energy constraints ($c_{eb}, c_{ef} \in C_e$) to enable the RL agent to learn passive behavior.

5.2.5 Passivity-Ensured Control

The passivity layer is an analytical filter that modifies the RL action to ensure the passivity of the control. It is applied between the RL agent and the low-level controller, as shown in Fig. 5.1. The passivity layer calculates the system energy and modifies the RL actions to stay passive in deployment.

Integrating the described energy tank dynamics for the VIC, the change of the Cartesian stiffness can be defined as:

$$\mathbf{K}(t) = \begin{cases} \mathbf{K}(t-1) & \text{if } E_T + \dot{E}'_T \leq E_T^{\min} \\ \mathbf{K}(t) & \text{otherwise.} \end{cases} \quad (5.6)$$

In the event that the energy tank becomes depleted, $\mathbf{K}$ is hindered from increasing further, thereby keeping $\mathbf{K}$ constant. This results in a standard Cartesian impedance controller with constant gains.

In order to enforce the energy flow constraint, α is incorporated in the torques generated by the impedance control term:

$$\tau_{IC}^\top = \alpha \mathbf{J}^\top(\mathbf{q}) (\mathbf{w}_K^{0,EE^\top} + \mathbf{D}\eta). \quad (5.7)$$

These filtering actions are necessary to guarantee passive behavior. However, the RL agent perceives the low-level control and the passivity layer as part of the environment, because these are not part of the RL-based decision-making. Thus, any post-processing applied to the RL action decreases the transparency of the system, i.e., the outcome of the same action may be different depending on the filter variables. For this reason, we combine the passivity layer with passivity-aware policies and evaluate the effects of each component in our experiments.

5.3 Evaluation

We evaluate our method through the contact-rich maze exploration task. The experimental task is a maze exploration task where the robot has to explore an unknown maze blindly. Thus, it relies on the force sensing for finding its way out, making it a contact-rich task, i.e., the robot does not avoid the contact and it has to regulate the exerted wrenches for safety. During this process, the controller dynamically adjusts Cartesian stiffness and pose in response to disturbances, such as maze walls and movable obstacles represented by three simulated sphere balls in the middle of the maze.

We train the policy in simulation using a MuJoCo (version 2.3.3) maze environment, as shown in Fig. 5.4. First, 40,000 tuples with 1363 constraint violations were collected as offline data. We increased the force threshold to 40 N to better focus on the passivity-based constraints. The policies were trained through SAC [182] under different constraint configurations. The following notation for the experiments is introduced: $\text{Eb}\theta_b$ denotes experiments subjected to the energy tank constraint with size θ_b [J], and $\text{Eb}\theta_b\text{-Ef}\theta_f$ for experiments that include both budget and flow constraints, where θ_f [W] indicates the imposed energy flow limit.

5.3.1 Passivity-Aware Policy Training

The results of the training experiments are summarized as means and standard deviations of 10 seeds in Fig. 5.5. The figure shows the amount of success in terms of task completion, the type and number of constraint violations, as well as the ratio between the amount of successfully completed runs and violations.

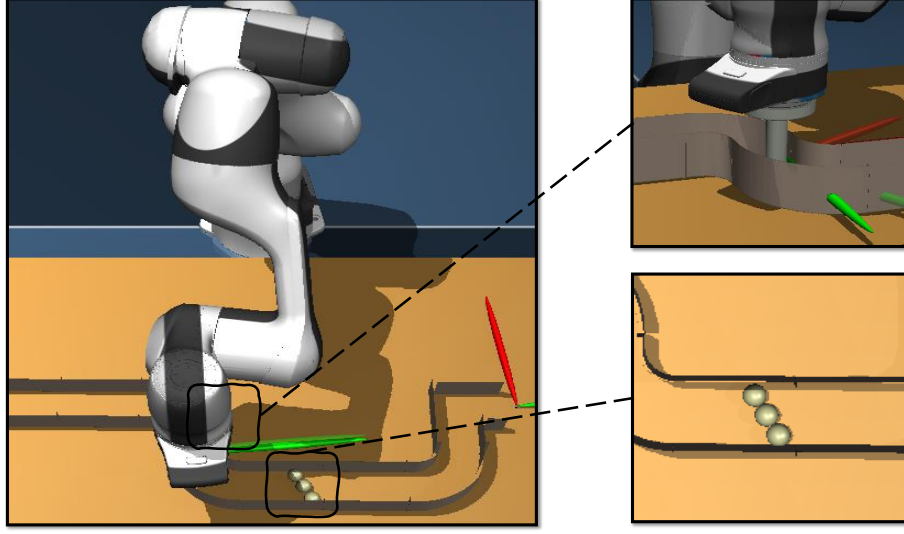


Figure 5.4 Experiments setup in MuJoCo. Right top: the robot explores the turn area with contact. Right bottom: three sphere obstacles.

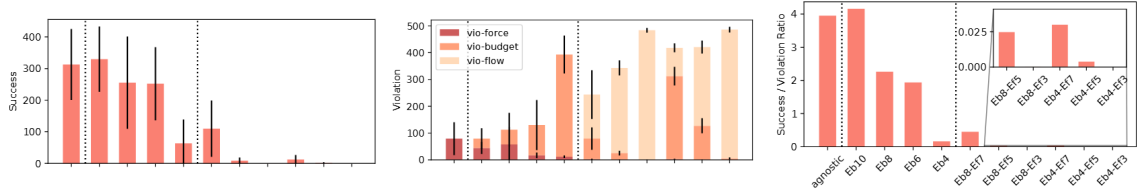


Figure 5.5 Passivity-aware policy training experiment results. We present the total numbers of successes and violations in the first two plots. The third plot shows the ratio of mean success to mean violation.

Firstly, the task's success is directly correlated with the amount of constraints introduced to the CMDP. Secondly, the introduction of c_{eb} has a direct effect on the amount of force violation, which overall decreases as the energy budget gets smaller. Furthermore, the introduction of c_{ef} results in the force constraint no longer being violated. Thirdly, the introduction of c_{ef} has a more aggressive effect on the task success than the introduction of only c_{eb} . This is due to the fact that the constraint on the flow hinders the agent from making aggressive changes in its stiffness, resulting in an overly conservative approach.

Fig. 5.6 shows how the changes of the stiffnesses $k_{tx}, k_{ty} \in \mathbf{K}_T$ during parts of the training affect the energy flow $\dot{E}_T$ and consequently the energy stored in the energy tank E_T .

Fig. 5.7 shows the impact that C_e has on passivity-aware policy training, showing the agent's energy expenditure at two different stages during training. When comparing both

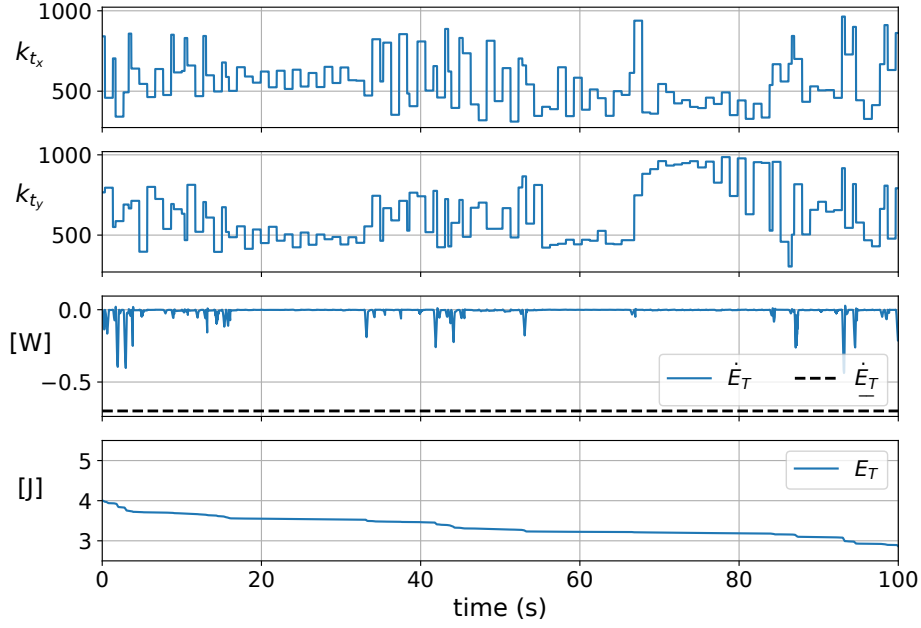


Figure 5.6 These plots visualize the effect that the changes of the stiffnesses $k_{t_x}, k_{t_y} \in \mathbf{K}_T$ during parts of the training have on the energy flow $\dot{E}_T$ and consequently on the energy stored in the energy tank E_T .

graphs it is evident that the agent has learned across 100 episodes to act more conservatively, resulting in reduced and more evenly distributed energy expenditure.

5.3.2 Model Tests

After training, we selected the policies for `agnostic`, `Eb4`, `Eb8`, `Eb8-Ef7` to test the model in representative experiments. All policies are evaluated over 100 episodes under the same constraint settings used during training. The results are shown in Table 5.2.

Table 5.2 Results of model tests. Performance evaluation over 100 episodes for each of three randomly selected seeds.

Run config.	Violations			Successes
	Force	Tank	Flow	$\mu \pm \sigma$
agnostic	0	0	0	100 ± 0
Eb4	0	26	0	91.3 ± 3.9
Eb8	2	0	0	99.3 ± 0.9
Eb8-Ef7	0	7	0	62 ± 40.1

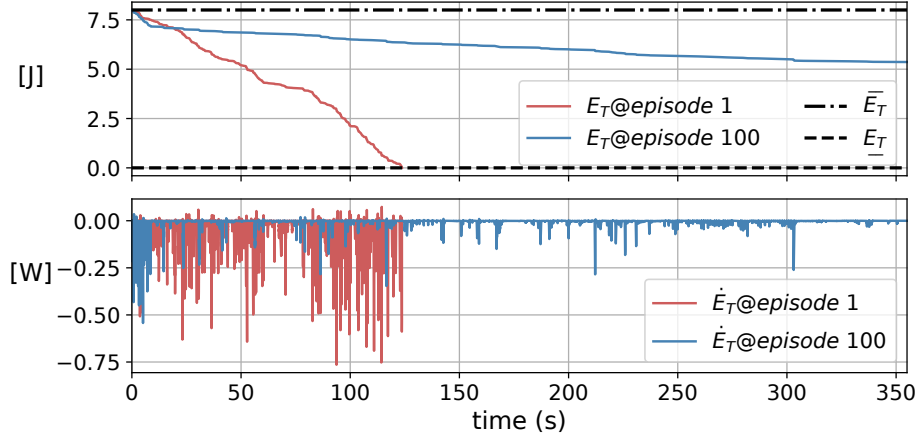


Figure 5.7 Plots illustrate the effect that the introduction of the *energy tank constraint* has on the agent’s energy expenditure at different stages during training.

Similar to the training results, the agent achieves the highest success rate when it is agnostic to C_e and can therefore operate more exploratorily during its runs. Looking at Eb8 and Eb4, a more conservative energy tank size results in more tank violations. Eb4 obtains high success, even though it had fewer successes than Eb8–Ef7 in training, showing that it learned a better task policy, although it was delayed by its limited energy budget.

5.3.3 Passivity-Ensured Deployment

In the deployment experiments, we evaluate the effect of the passivity layer on the trained models, by comparing the performances of the *passivity-agnostic* baseline and various *passivity-aware* approaches in combination with and without the passivity layer. We label the passivity-agnostic approach with the passivity layer as *passivity-filtered* and the passivity-aware approaches with the layer as *passivity-ensured*. The passivity layer is applied with uniform parameter settings: an initial energy budget of $E_T^{\max}=6$ J and an energy flow limit of $\dot{E}_T^{\min}=-0.5$ W.

The results in Fig. 5.8 showcase the task successes and violations. Overall, the results labeled as “with passivity layer” outperform the baseline labeled as “no passivity layer”. These results reveal the risks of deploying models without passivity guarantees, which potentially leads to system instabilities and unsafe behaviors despite good performance in model tests shown in Table 5.2.

Specifically: (1) The agnostic policy performs the worst when no passivity filter is applied, as it consistently violates the flow constraint. Although it achieves a 100% success rate when evaluated without considering energy constraints, this highlights the necessity of

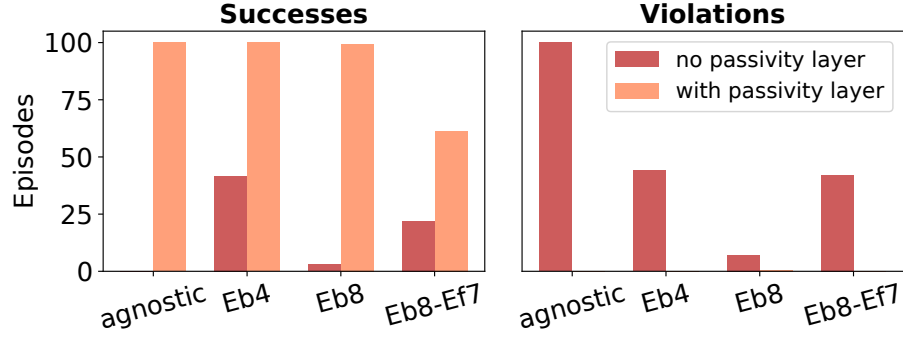


Figure 5.8 Policy performance in deployment experiments. The bars show how the models (agnostic, Eb4, Eb8, Eb8-Ef7) perform in terms of successes and violations with and without the passivity layer in simulation for 100 episodes.

incorporating passivity during training. (2) The success rate improves significantly when the passivity filter is applied, underscoring the dual benefits of passive-aware policies and the passivity layer, ensuring safer and more stable deployments. (3) Violations dropped significantly with the passivity layer, with only Eb8 having two force violations.

Fig. 5.9 presents the energy usage of each method to further demonstrate the benefit of passivity-aware training. Even though all of agnostic, Eb8 and Eb4 show full success in the tests, analysis of the energy budget progress in these runs reveals that the agnostic model depletes its budget more rapidly. On the other hand, Eb4 demonstrates the best success and energy efficiency, despite its delayed learning. This figure shows that the Eb8 and agnostic policies achieve comparable completion times, with Eb8 being slightly faster on average despite its lower energy consumption. This indicates that the agnostic policy expends more energy per unit progress, rather than simply completing the task faster. In contrast, more conservative policies (Eb4, Eb8-Ef7) exhibit a trade-off between efficiency and completion time.

5.3.4 Real-World Experiments

The trained models are deployed in the real world to complete our validation on a Franka Panda robot arm with a flange mounted at the end effector with an ATI45 FT sensor to perceive external force and torque, with our passivity-ensured impedance controller (frequency 1000 Hz). We run the best models of each setup 3 times in the real world without fine-tuning.

The robot can finish the task without violating force constraints and energy constraints in each experiment, which validates our method’s effectiveness. However, a failure case happened in the baseline (passivity-agnostic) model. All passivity-ensured configurations

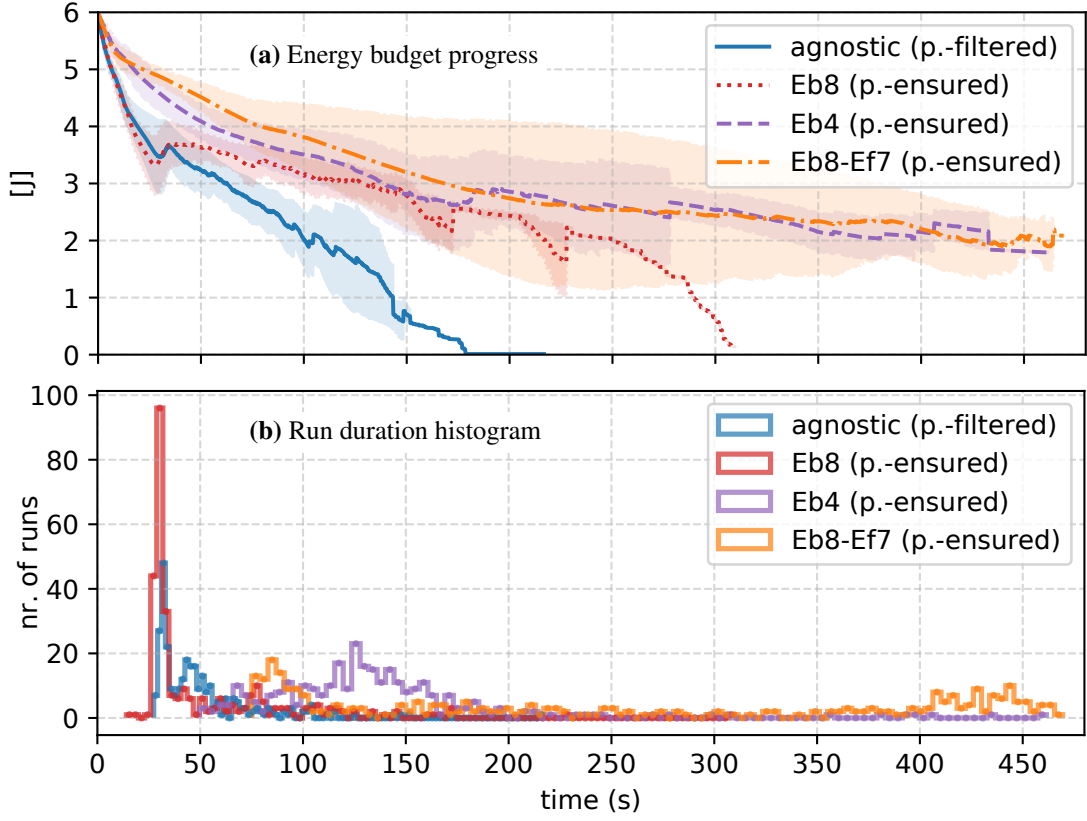


Figure 5.9 (a) Energy budget progress during passivity-ensured deployment experiments. The curves indicate the mean of 300 runs (100 per seed) and the shades indicate the standard deviation. (b) Histogram of the run times, i.e., the number of runs that finished in time duration intervals.

(Eb4, Eb8 Eb8-Ef7) achieved a 100% success rate over three trials each, with no constraint violations. In contrast, the passivity-agnostic policy failed in one out of three trials, largely due to its more aggressive energy usage, applying excessively high force at the turning point (Fig. 5.10.c). Specifically, Fig. 5.10 shows all models' energy expenditure over time steps. First, we can observe that the *agnostic* consumes energy most aggressively although it keeps passive due to the passivity layer, while passivity-aware models maintain lower energy consumption rates. Notably, Eb4 achieves the most efficient energy usage among all models. This suggests that training with tighter energy constraints leads to more energy-efficient behaviors that transfer well to real-world deployment.

5.4 Discussion

This section summarizes and discusses the performance of the different agents during training and deployment. Firstly, the training results in Sec. 5.3.1 show that the agent learns to be

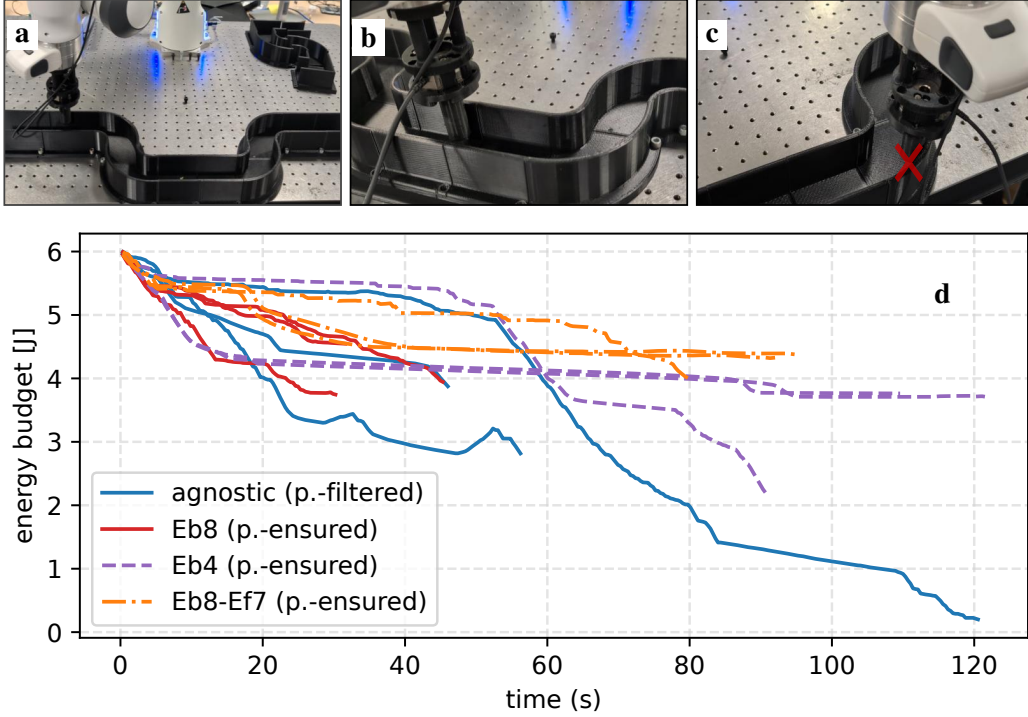


Figure 5.10 Snapshots and result analysis in real-world experiments. **(a)** the overview of the maze exploration task. **(b)** closeup when the robot moves along with contact force. **(c)** failure case agnostic but non-passive behavior at the turnaround area. **(d)** shows the energy budget consumption among different setups, demonstrating our method outperforms in terms of passivity and energy efficiency.

more conservative in its energy expenditure when subjected to the passivity-based constraints C_e as directly observed in Fig. 5.7. Furthermore, Fig. 5.5 showcase that the inclusion of these constraints directly affects the success rate of task completion during training. This is because these constraints restrict the agent's exploration, which is crucial, especially at the beginning of the learning phase. However, although they learn the task later (Fig. 5.5) constrained cases can still achieve high success rates in the model test (Table 5.2), suggesting that the passivity-constrained cases would benefit from further training compared to the agnostic case. On the other hand, overly restrictive constraints can prevent learning entirely if the task becomes physically infeasible.

Importantly, the differences between the passivity-agnostic and constrained policies in Fig. 5.5 and 5.9 indicate that the observed changes in energy usage are not merely a byproduct of task progress but arise from the explicit inclusion of C_e during training.

Another important observation can be made in the deployment results shown in Figure 5.8. It shows that passivity cannot be guaranteed without a passivity layer, regardless of whether the agent was trained with passivity awareness or not. Furthermore, it showcases that while

the passivity layer improves task completion and reduces violations for all agents, those trained with energy constraints (C_e) demonstrate better energy efficiency.

The correlation between the energy usage trends in simulation (Fig. 5.9.a) and real-world experiments (Fig. 5.10.d) demonstrates the successful transfer of learned energy-efficient behaviors. In both cases, the agnostic model shows the steepest decline in energy budget, while Eb4 maintains the most conservative energy usage. This consistent pattern validates that the energy-aware policies learned in simulation maintain their characteristics when deployed on physical hardware. The real-world results also confirm that tighter energy constraints during training (Eb4) lead to more efficient policies that better manage the energy budget in both simulated and real environments. However, the absolute rate of energy consumption differs between simulation and reality due to modeling discrepancies and real-world friction effects.

5.5 Summary

In this chapter, we investigated the safety and stability of the robotic system under multiple constraints for a contact-rich task, demonstrating that the traditional RL approach cannot guarantee passivity. Furthermore, we combined an augmented energy tank-based passivity formalism in a safe RL framework. We employed the passivity constraints in both training and deployment of the RL model. The proposed method completed the task without any force violation while ensuring system passivity. Additionally, passivity-aware policies showed improved energy efficiency in deployment, even though they converged later in the training. A set of experiments both in simulation and in the real world on a 7-DoF Franka robot arm validated the effectiveness of the proposed method.

The first three technical chapters of the thesis (Chapters 3–5) together establish a progressively richer safe RL framework for contact-rich robotic tasks, addressing safety from predictive, reflexive, and passivity-based perspectives. Having developed these safety principles in the RL setting, the next chapter transitions to foundation-model-based approaches and explores how vision-language models can guide variable impedance control for more generalizable and language-instructable manipulation.

Chapter 6

A Self-Improving Variable Impedance Controller with Vision-Language In-Context Learning

6.1 Introduction

The previous chapters focused on safe reinforcement learning formulations for contact-rich manipulation, showing how safety can be improved through predictive risk estimation, reflexive intervention, and passivity-aware control. In this chapter, we begin the second part of the thesis by asking how these contact-aware control principles can be combined with the semantic reasoning capabilities of foundation models.

Robotic manipulation in contact-rich environments presents significant challenges, particularly in ensuring safe and effective interactions with objects and the environment [198, 199]. Tasks such as wiping a surface, inserting a plug into a socket, or pushing a drawer require not only precise motion planning but also adaptive compliance to handle uncertainties and variations in contact dynamics.

Traditional variable impedance control (VIC) can improve safety and adaptability in specific scenarios by modulating the robot's stiffness and damping to achieve compliance. However, such approaches typically require manual parameter tuning and do not generalize well across different tasks or environments [200]. Even learning-based VIC approaches [21] often struggle with generalization due to the limited scope of training data and the complexity of contact dynamics.

The emergence of vision-language models (VLMs) has opened new avenues for enhancing robotic manipulation by providing rich semantic understanding and reasoning capabilities [201, 202]. By leveraging the contextual information provided by VLMs, robots can better understand the nuances of a task, such as the required level of force or compliance, and adjust their behavior accordingly. However, most existing VLM-based approaches focus on high-level task planning or trajectory generation, often neglecting the critical aspect of force-aware compliant control in contact-rich scenarios.

While VLMs alone struggle to inform compliant control, retrieval-augmented generation (RAG) can fill this gap by recalling relevant past experiences [203]. In robotics, RAG can enable robots to retrieve relevant prior experiences or demonstrations that inform their current actions, improving adaptability and generalization [204]. Moreover, in-context learning (ICL) enables robots to adapt behaviors from contextual examples without retraining [205].

Rather than replacing compliant control with foundation models, our goal is to use VLMs to inform and improve impedance adaptation, thereby preserving the physical advantages of VIC while expanding generalization across tasks and contexts.

To address these challenges, we propose **OmniVIC**, a universal variable impedance controller enhanced by a VLM and empowered by RAG and ICL. OmniVIC integrates high-level semantic reasoning from multimodal inputs with low-level compliant control, enabling robots to adapt their impedance parameters for safe and effective manipulation in contact-rich scenarios.

The key contributions of this chapter are:

- A new VIC paradigm that enables context-aware impedance adaptation through a VLM for safer physical interaction.
- A self-improving RAG mechanism that retrieves relevant prior experiences from a structured memory bank, facilitating impedance adaptation based on similar past tasks.
- An ICL strategy that conditions the VLM on retrieved examples and current task context, allowing for dynamic generation of impedance parameters without retraining.
- Comprehensive simulation and real-world experiments demonstrating that OmniVIC improves average success rates from 27% (baseline) to 61.4%.

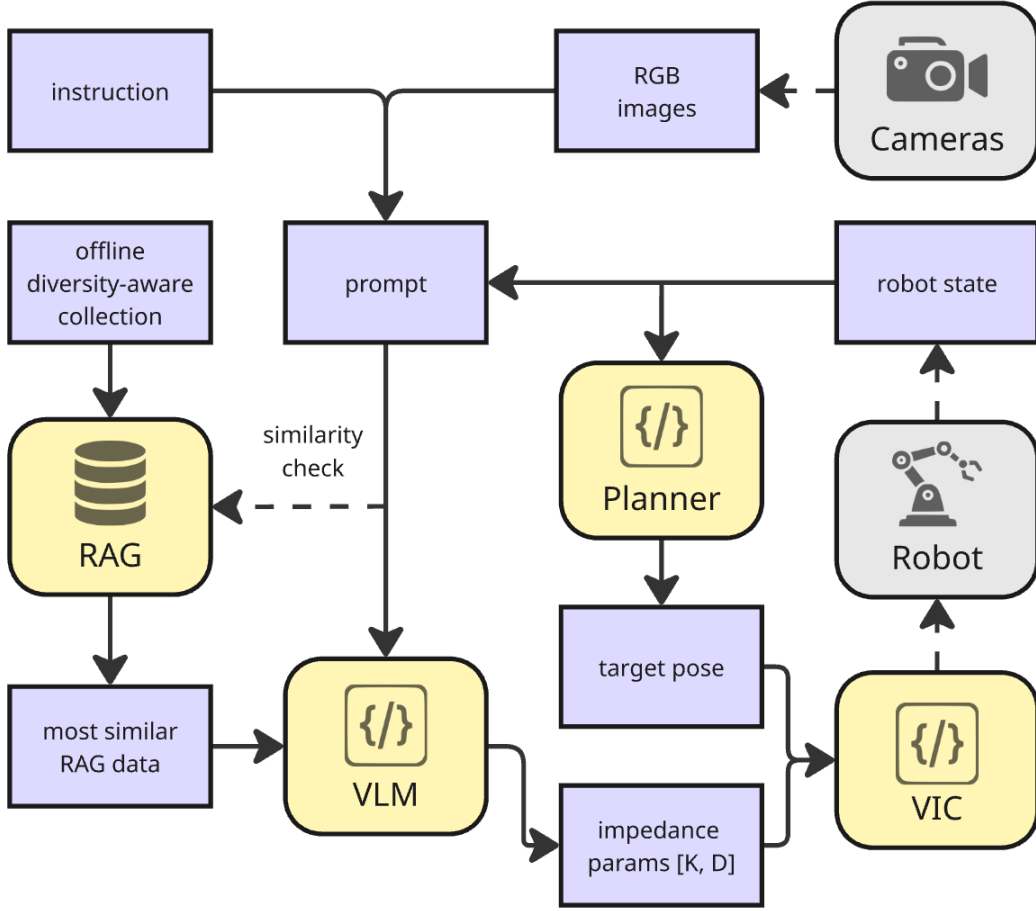


Figure 6.1 Overview of the proposed OmniVIC. The system integrates a VLM with a VIC to enable safe and adaptive manipulation in contact-rich tasks. The VLM processes multimodal inputs, including visual observations, natural language instructions, and real-time force/torque feedback, to generate context-aware impedance parameters (stiffness and damping) for the VIC.

6.2 Methodology

In this section, we describe the overall architecture and key components of OmniVIC (Fig. 6.1), a universal VIC empowered by RAG+ICL and assisted by VLM.

6.2.1 VLM-Informed Variable Impedance Control

A classical Cartesian VIC modulates the robot’s stiffness K and damping D to achieve compliant interactions with the environment. The control law is typically defined in task space as:

$$F_{ext} = D\dot{\tilde{x}} + K\tilde{x}, \quad (6.1)$$

where $\tilde{x}$ is the Cartesian pose error between the desired and actual pose $x_d, x \in \mathbb{R}^6$. The desired stiffness matrix $K \in \mathbb{R}^{6 \times 6}$ is a diagonal matrix with variable terms as:

$$K = \text{diag}\{K_x, K_y, K_z, \varepsilon K_x, \varepsilon K_y, \varepsilon K_z\} \quad (6.2)$$

where the ε is a coefficient for rotation elements. The desired damping matrix $D \in \mathbb{R}^{6 \times 6}$ is also a diagonal matrix with variable terms as:

$$D = \text{diag}\{D_x, D_y, D_z, \zeta D_x, \zeta D_y, \zeta D_z\} \quad (6.3)$$

where ζ is the coefficient for rotation elements.

In our method, the VLM processes visual and language inputs to derive the task context, which is then used to generate adaptive impedance parameters for the VIC controller. To simplify the problem, the VLM outputs diagonal translational stiffness and translational damping elements, reducing the parameter space to three stiffness values (K_x, K_y, K_z) and three damping values (D_x, D_y, D_z).

6.2.2 RAG Data Acquisition and Storage

RAG data are acquired by deploying any robot policy and applying two success criteria: a maximum allowable contact force $F_{\max}$ and a time budget $T_{\max}$. A trial is labeled *failure* if the measured force exceeds $F_{\max}$ at any time or if the task time exceeds $T_{\max}$; otherwise, it is labeled *success*.

For each *successful* execution, we append the following to the RAG database: (i) the instruction T , (ii) the end-effector twist $v_t \in \mathbb{R}^6$ expressed in the world frame, (iii) the gravity-compensated wrench $w_t = [F_t, \tau_t] \in \mathbb{R}^6$, (iv) the phase label p_t , and (v) the VLM-predicted controller impedance parameters (K, D).

Instruction: We store the instruction in two complementary forms: natural-language instruction text T_{text} (used later for in-context learning), and a precomputed sentence embedding T_{emb} (used for retrieval). The embedding is produced once at collection time by a text encoder.

Phase label: By querying a VLM with the instruction T , the current image I_t , and the world-frame twist v_t and wrench w_t , the model outputs one of `Free_motion`, `Approaching`, `Contact`, `Retreat` as the phase label at each step.

When a new successful record arrives and the bank is full, we temporarily pool the new record with the records belonging to this instruction, compute pairwise similarities, identify

the closest pair, and randomly discard one of the two. This closest-pair replacement prevents near-duplicate motions from accumulating and preserves diversity under a fixed memory budget.

6.2.3 RAG Retrieval

The RAG retrieval process identifies the most relevant prior experiences from the database to inform impedance parameter generation for the current request. The pre-processed items are used in a four-step progressive retrieval process:

Step 1 – Instruction filtering: We compare the query’s instruction embedding T_{emb} to all stored instruction embeddings using cosine similarity, and retain the top- $M\%$ most similar instructions.

Step 2 – Phase filtering: We refine the results of Step 1 by keeping only those records whose phase matches the query’s phase.

Step 3 – Similarity computation: We compute four modality-specific similarity scores between the query and each candidate record: (1) force similarity, (2) torque similarity, (3) linear velocity similarity, and (4) angular velocity similarity. All four similarities use the cosine similarity:

$$\cos_sim(a, b) = \frac{a \cdot b}{\|a\| \|b\|}, \quad (6.4)$$

where a and b are the two vectors being compared.

Step 4 – Final score: We sum the four similarity scores to obtain a single aggregate measure for each candidate, and keep the top five highest-scoring candidates as exemplars for in-context learning.

6.2.4 ICL-Enhanced Impedance Parameter Prediction

Given a new request, we run RAG retrieval to obtain the top- N most similar records. We then provide the top five similar records to the VLM as examples. Each example includes the instruction text (T_{ext}), phase label, controller gains (K, D), world-frame twist and wrench, and the similarity score. The VLM outputs (K, D) for the current step using a structured prompt that instructs the VLM to apply phase-based impedance principles (highest stiffness for free motion, lowest for contact) and to consider motion direction adaptation.

Table 6.1 Tasks in simulation experiments

Task No.	Description
Task 1	put the bowl on top of the cabinet
Task 2	pick up the salad dressing and place it in the basket
Task 3	pick up the tomato sauce and place it in the basket
Task 4	pick up the milk and place it in the basket
Task 5	pick up the orange juice and place it in the basket
Task 6	put the wine bottle on top of the cabinet
Task 7	push the plate to the front of the stove
Task 8	put the cream cheese in the bowl
Task 9	open the middle drawer of the cabinet
Task 10	pick up the butter and place it in the basket

6.3 Experiments

We evaluate OmniVIC on a suite of contact-rich manipulation tasks in both simulation and real-robot experiments. The evaluation metrics include task success rates, force violations, and overall interaction safety.

6.3.1 Simulation Experiments

Baselines: As OmniVIC is a general-purpose variable impedance controller rather than a specific policy, we adapt robot policies from a pretrained VLA model Pi0 [206] to execute tasks. The default controller used in Pi0 is our baseline, which is a low-level position controller with fixed translational stiffness $k_p = 150$ N/m and damping $K_d = 2\sqrt{k_p} = 24.494$ N·s/m.

Evaluation task and protocol: We select 20 related manipulation tasks from the LIBERO benchmark, 10 from *LIBERO-Object* and 10 from *LIBERO-Goal*. We design the split by stratified selection based on outcome, forming two disjoint sets: a Knowledge-Base Set of 10 tasks (retained solely for building the RAG store) and a Query Set of 10 tasks (held out solely for issuing queries). The 10 evaluation tasks in the Query Set are listed in Table 6.1.

Safety termination rule: An episode is marked as *failure* if the measured interaction force exceeds 30 N on three consecutive steps.

RAG data construction: We collect RAG data exclusively from the Knowledge-Base Set. The instruction embeddings are computed with the BGE-M3 text encoder [207]. The VLM used throughout simulation is GPT-4o-mini. The maximum capacity limit of our RAG

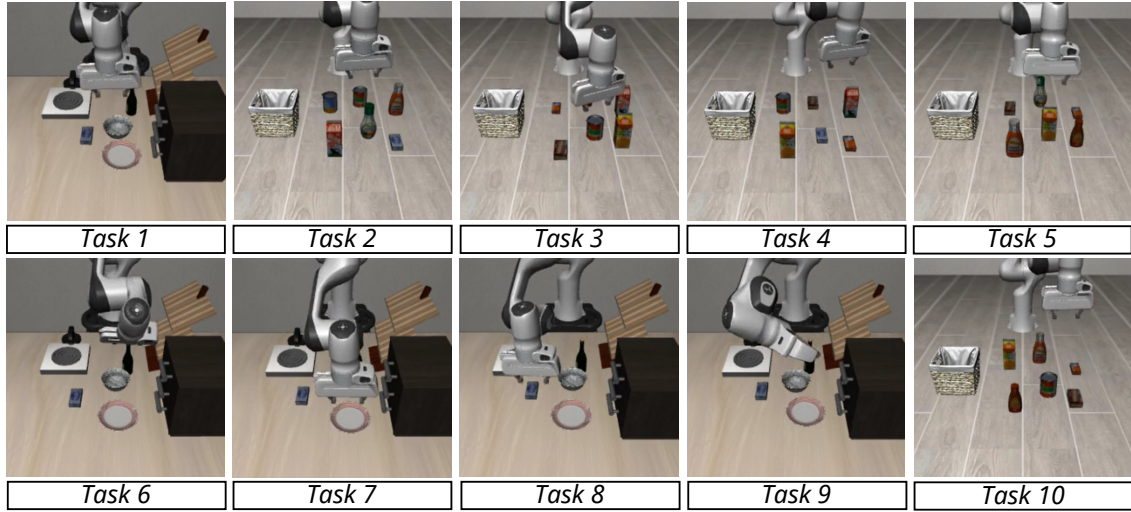


Figure 6.2 Tasks used in simulation experiments.

database is $B = 20K$. In our experiments, we built the RAG dataset by running 10 tasks, with 20 records per task instruction-phase pair, resulting in a total capacity of 200 records.

6.3.2 Simulation Results

Force Safety Analysis

Fig. 6.2 illustrates the tasks used in simulation experiments. The deep dive into force and impedance profiles during task execution shows that OmniVIC dynamically adjusts stiffness and damping based on task phases, reducing them during contact to enhance compliance. OmniVIC maintains interaction forces within safe limits, avoiding spikes that violate constraints, while the baseline controller frequently violates the force constraint and terminates the episode as a failure.

Task Performance Evaluation

As shown in Fig. 6.3, the integration of OmniVIC yields substantial improvements across all evaluated tasks. The baseline achieves an average success rate of 27.0% across tasks, with frequent force violations compromising both safety and task completion. In contrast, the OmniVIC-enhanced system achieves an average success rate of 61.4%, a 2.27-fold improvement over the baseline.

The success of OmniVIC stems from its ability to intelligently adapt impedance parameters based on task context. Through the RAG component, the system retrieves relevant

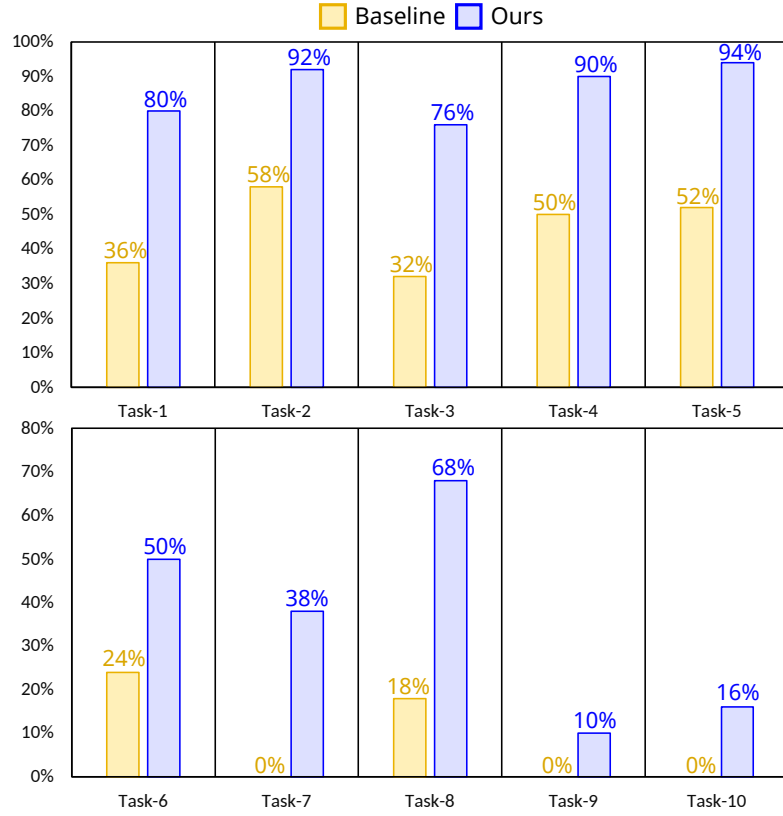


Figure 6.3 Simulation results: Task success rates of baseline (π_0 with position control) vs. OmniVIC (with VLM-informed variable impedance control).

prior experiences from similar manipulation scenarios, while the ICL mechanism enables real-time parameter generation tailored to the current task state. OmniVIC still maintains success rates even when the baseline fails due to force violations (Tasks 7–10 in Fig. 6.3), highlighting its robustness in handling complex contact-rich tasks.

Ablation Study: RAG vs. RAG+ICL (OmniVIC)

To quantify the contribution of ICL on top of RAG, we compare two configurations under identical settings: **RAG-only**, which applies the retrieved parameters without ICL, and **RAG+ICL (OmniVIC)**, which feeds the retrieved exemplars to the VLM for in-context learning. Across the ten tasks, **RAG+ICL** achieves a *1.59-fold* higher task success rate than **RAG-only**, indicating a substantial benefit from letting the VLM adapt the retrieved priors to the instantaneous scene and interaction state. This ablation supports our design choice to pair RAG with ICL for robust, fine-grained modulation of (K, D) in unseen and complex scenarios.

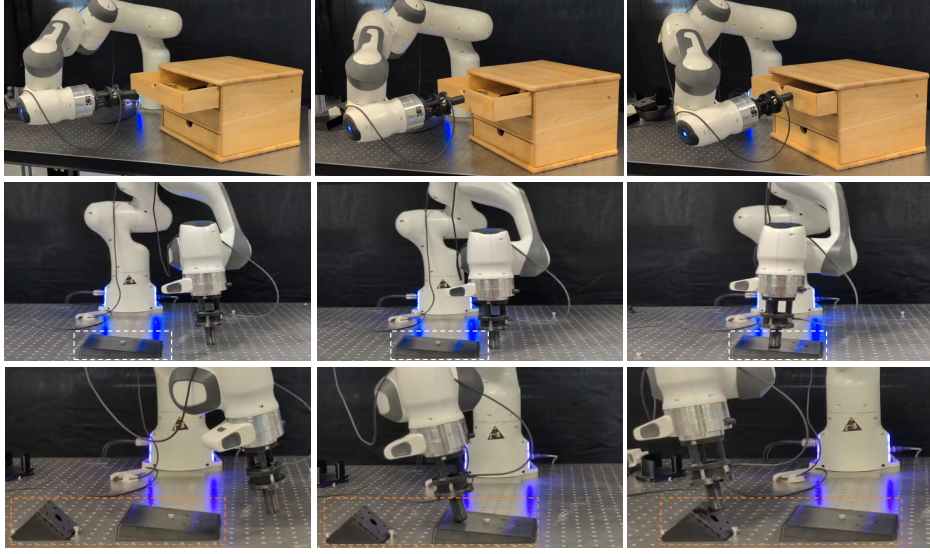


Figure 6.4 Real-world experiments.

6.3.3 Real-World Experiments

We validate our approach on a physical robot platform. We implement OmniVIC on a Franka Emika Panda robot arm (at 1000 Hz), equipped with a Logitech USB camera offering a global view of the robot arm and its workspace, and a 6-axis ATI Mini45 force/torque sensor for real-time interaction feedback. The VLM processes visual (~ 1 Hz) from the camera and natural language instructions to generate adaptive impedance parameters (stiffness range in $[300, 1000]$ N/m).

The robot executes a series of contact-rich manipulation tasks as shown in Fig. 6.4: (1.) Close the top drawer gently; (2.) Move along the negative Y-axis while keeping the same height in the Z-axis (one ramp); (3.) The same task setup as Task 2, but with two ramps along the way.

For Task 1, OmniVIC effectively regulates interaction forces while closing the drawer gently without safety constraint violation. The force stays within safe limits throughout task execution. In contrast, constant high impedance causes excessive force, while constant low impedance fails due to insufficient force to overcome friction.

For Tasks 2 and 3 (Fig. 6.6), the robot successfully navigates along a path with a ramp, maintaining safe interaction forces while adapting its impedance parameters to the varying contact conditions. The shaded regions indicate the contact phase as identified by OmniVIC.

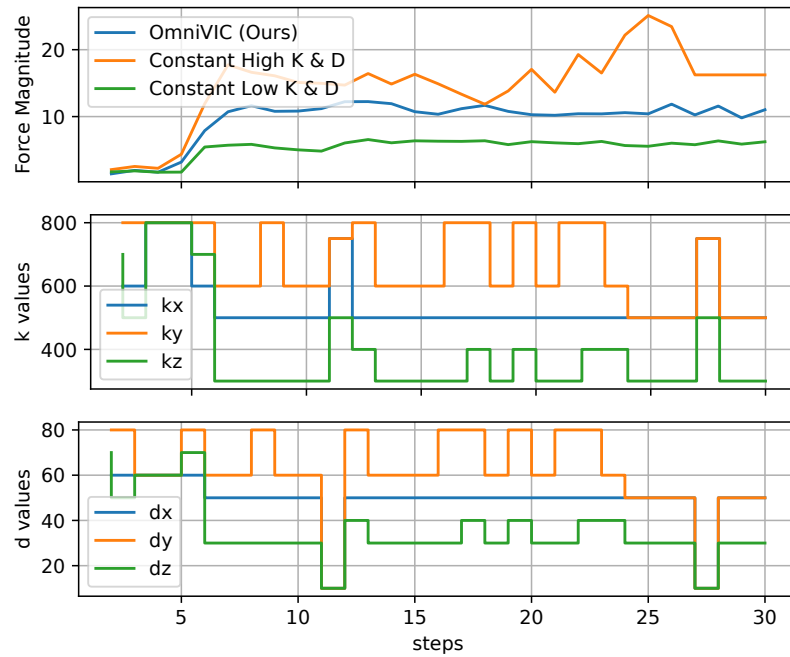


Figure 6.5 Task-1: The robot gently closes the top drawer using OmniVIC variable impedance control. The system adapts stiffness and damping parameters based on visual and language context to ensure safe interaction and avoid excessive force.

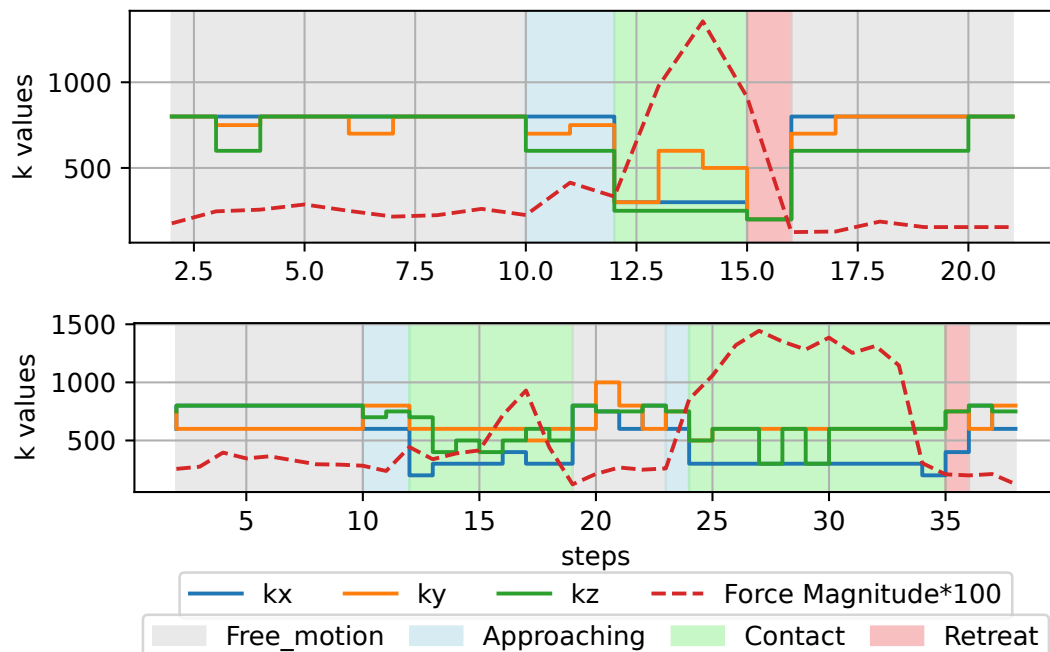


Figure 6.6 Task 2 (top) and Task 3 (bottom): The robot moves along the negative Y-axis while maintaining the same height in the Z-axis. The shaded regions indicate the contact phase as identified by OmniVIC.

6.4 Summary

In this chapter, we presented OmniVIC, a systematic variable impedance controller that integrates RAG and ICL with VLMs to enhance VIC for contact-rich robotic manipulation tasks. OmniVIC leverages the strengths of VLMs in understanding visual and language context to dynamically generate impedance parameters, enabling robots to adapt their interactions based on task requirements and environmental conditions. Through comprehensive simulation and real-world experiments, we demonstrated that OmniVIC significantly improves task success rates and reduces force violations compared to baselines, with the average success rate increasing from 27% (baseline) to 61.4% (OmniVIC). The integration of RAG allows the system to retrieve relevant prior experiences, while ICL enables the VLM to generate context-aware impedance parameters in real-time.

The next chapter extends this direction by moving from VLM-guided impedance generation as a separate controller module to a tighter integration with vision-language-action (VLA) models, exploring how compliant control can be incorporated directly into general-purpose action-generation pipelines.

Chapter 7

VLM-Guided Variable Impedance Action for Safe Contact-Rich Manipulation

7.1 Introduction

Chapter 6 showed that vision-language models can effectively guide impedance adaptation as a high-level controller for contact-rich interaction. Building on that insight, this chapter takes a further step toward tighter integration by asking how compliant control can be attached directly to modern vision-language-action policies.

Recent advances in Vision-Language-Action (VLA) models have enabled robots to understand and execute complex tasks described in natural language, such as RDT [208], Pi0 [206], DiffusionVLA [209], Pi0.5 [210], OpenVLA [174], and OpenVLA-oft [211]. These models leverage large-scale pretraining on diverse datasets to learn rich visual and linguistic representations, allowing them to generalize across a wide range of manipulation tasks.

However, current state-of-the-art VLA models fundamentally operate through position or trajectory control, treating the robot as a rigid position-tracking system that lacks consideration of the physical interaction dynamics involved in contact-rich tasks [212]. The contact force can become dangerously high when the robot executes raw VLA output in contact-rich tasks without force regulation. As illustrated in Fig. 7.1, existing VLA systems lack force-awareness, leading to unsafe interactions in physical tasks involving contact or uncertainty.

This rigid execution paradigm leads to critical failures when interacting physically. Consider a robot assembling a delicate electronic device (“insert the USB cable”): it must insert a fragile connector with just enough force to engage the latch, yet gentle enough to

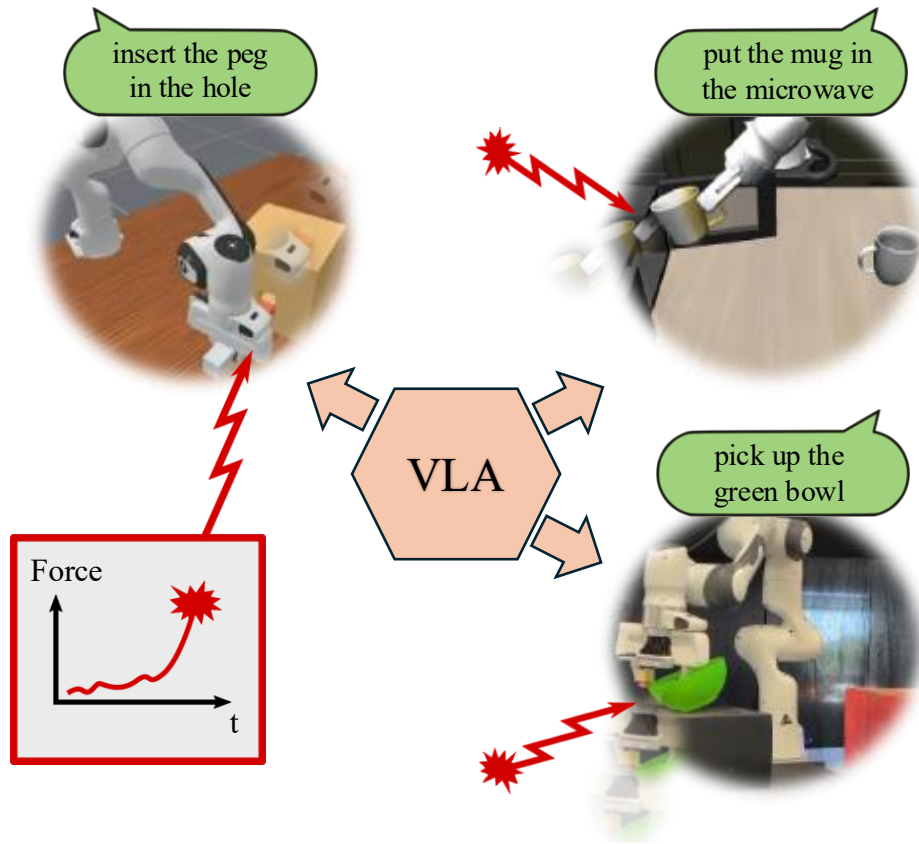


Figure 7.1 Existing VLA systems lack force-awareness, leading to unsafe interactions in physical tasks involving contact or uncertainty. CompliantVLA-adaptor addresses this challenge by augmenting VLA models with VLM-guided variable impedance control.

avoid damage. Or imagine a robot closing a drawer, where excessive force could damage the drawer or its contents, while insufficient force might leave the drawer ajar. The contact-rich manipulation tasks require not just visual understanding, but also precise force regulation and adaptive compliance.

In contrast, variable impedance control (VIC) provides compliant behavior in response to external forces through spring-damper dynamics, offering a complementary solution by modulating robot stiffness and damping to tackle interaction forces, enabling safe and adaptive physical interaction [4]. However, existing VIC approaches require manual tuning or task-specific parameter scheduling, lacking the semantic understanding to automatically adapt impedance based on visual context and task requirements.

We present **CompliantVLA-adaptor**, a plug-and-play modular solution that augments VLA models with VLM-guided variable impedance control for safe contact-rich manipulation. Our experimental findings demonstrate that CompliantVLA-adaptor not only enhances task

success rates but also fundamentally transforms the nature of failure modes. While baseline VLA models exhibit catastrophic failures characterized by contact-force threshold violations and toppled objects, our approach achieves graceful degradation, with failures primarily due to minor misalignments or slippage rather than unsafe forces.

In this way, the chapter bridges the gap between semantically informed impedance control and end-to-end generalist action models, while retaining a clear safety role for compliance during physical interaction.

The key contributions of this chapter are:

- The proposed CompliantVLA-adaptor leverages a VLM-enhanced VIC to endow VLA models with compliant capabilities for physical interaction, with the VLM generating context-aware impedance parameters that translate high-level semantic understanding into low-level control parameters (stiffness and damping).
- A control system that seamlessly integrates VLM parameter generation with VIC execution, maintaining the generalization benefits of VLM’s reasoning while adding the safety improvement of compliant control.
- Comprehensive simulation and real-world experiments showing improvements over the state-of-the-art VLA baselines in both success rates and safety metrics.

7.2 Methodology

We present CompliantVLA-adaptor, augmenting VLA models with VLM-enhanced context-aware variable impedance control. Our approach consists of three key components: (1) a hybrid VLA-VIC control architecture, (2) VLM-based impedance parameter generation from visual-linguistic context, and (3) a real-time force-regulated safety layer. Fig. 7.2 illustrates the complete system architecture.

7.2.1 Problem Formulation

Consider a robotic manipulation task defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{C})$, where $\mathcal{S}$ represents the state space including visual observations, $\mathcal{A}$ denotes the action space, $\mathcal{T}$ is the task specification in natural language, and $\mathcal{C}$ represents physical constraints including force limits.

For a single 7 DoF robotic arm, the traditional VLA model maps observations and language instructions to actions, typically end-effector displacement values $\mathbf{x}_d \in \mathbb{R}^6$ and

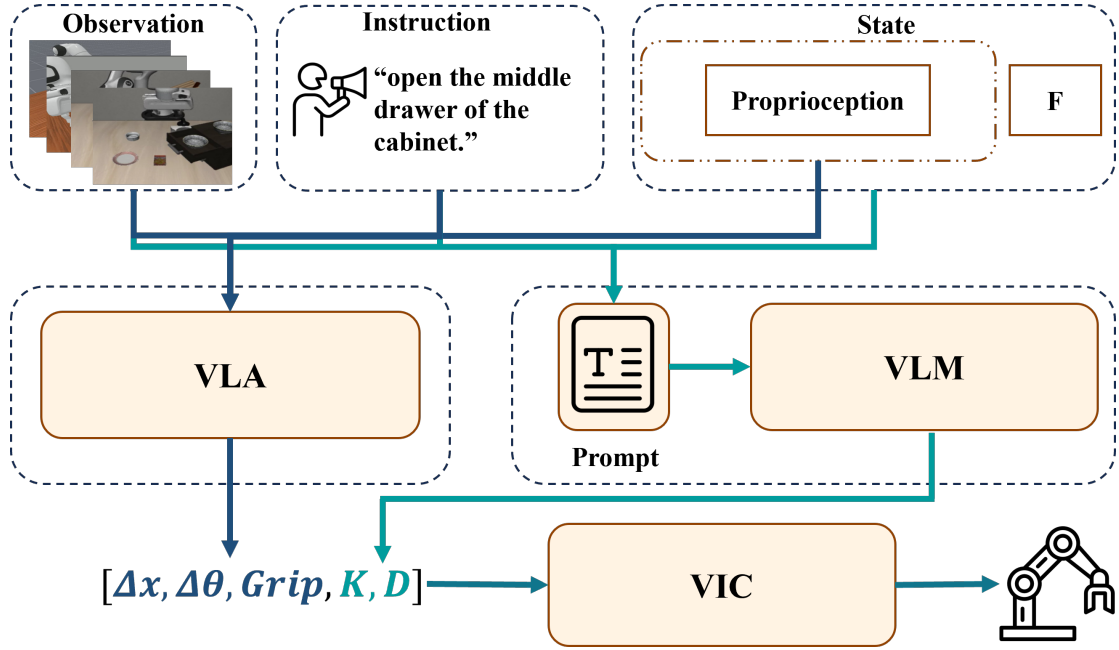


Figure 7.2 Overview of the CompliantVLA-adaptor. A VLM processes visual observations, language instructions, and real-time force feedback $\mathcal{F}$ to generate context-aware impedance parameters $\mathcal{K}, \mathcal{D}$. These parameters modulate a variable impedance controller (VIC) that executes actions generated by a VLA model, ensuring safe and adaptive contact-rich manipulation.

action for the gripper. However, this formulation ignores interaction forces $\mathbf{F} \in \mathbb{R}^6$, leading to unsafe contact interactions.

We augment an impedance-modulated control adaptor:

$$\text{CompliantVLA} : \text{VLM}(\mathcal{S} \times \mathcal{T} \times \mathcal{F}) \rightarrow (\mathcal{K}, \mathcal{D}) \quad (7.1)$$

to VLA model:

$$\pi_{\text{VLA}} : \mathcal{S} \times \mathcal{T} \rightarrow \mathcal{A} \quad (7.2)$$

where a 6D vector $\mathcal{F} \in se^*(3)$ denotes the external contact force and torque, $\mathcal{K} \in \mathbb{R}^3$ and $\mathcal{D} \in \mathbb{R}^3$ represent the translational stiffness and damping matrices, respectively:

$$\mathcal{K} = \text{diag}(k_x, k_y, k_z), \quad \mathcal{D} = \text{diag}(D_x, D_y, D_z) \quad (7.3)$$

We set orientation stiffness matrices K_o and orientation damping matrices D_o with proportional coefficients:

$$K_o^i = \varepsilon k^i, \quad D_o^i = 2\zeta \sqrt{K_o^i} \quad (7.4)$$

where $\varepsilon = 0.15$ and $\zeta = 0.707$. The robot controller takes desired action $\mathcal{A}$, stiffness $\mathcal{K}$ and damping $\mathcal{D}$ as inputs to enhance the safety along the physical interaction.

7.2.2 CompliantVLA-Adaptor

Our impedance generation module leverages a frozen VLM to extract multimodal features from external force, visual observations, and physics-injected language instructions, mapping them to context-aware impedance parameters. To ensure safe interaction forces, we implement a dual-layer safety system that combines VLM-based parameter modulation with contact phase recognition.

Given an RGB image of wrist $\mathbf{I}_w$, image of full overview $\mathbf{I}_f$, language instruction $\mathcal{T}$, and external force $\mathcal{F}$:

1. We query a pretrained VLM (e.g., ChatGPT-4o-mini) to reason about the current task execution phase *current phase*;
2. Integrating the suggested *current phase* and prompt, the VLM maps these features to impedance parameters.

Contact Phase Recognition through Multimodal Information

Vision-only phase detection presents inherent limitations, as VLMs cannot reliably detect contact states from visual information alone, particularly in scenarios with occlusions or subtle contact transitions. While force sensors provide direct contact information, relying solely on force feedback neglects valuable semantic context about expected interaction patterns. Therefore, we implement a hybrid approach that combines VLM-based visual understanding with force sensor feedback for robust phase detection.

The VLM can distinguish between Free-motion, Approaching, Contact, and Retreat phases. This semantic phase recognition eliminates the need for contact dynamics models while providing appropriate compliance for each interaction stage.

Multimodal-Informed Impedance Parameter Generation

The final impedance parameters fed to the VIC controller combine VLM-generated values with real-time force feedback:

$$\mathbf{K}_p^{\text{final}} = \mathbf{K}_p^{\text{VLM}} \cdot \alpha_{\text{force}} \quad (7.5)$$

$$\mathbf{D}_p^{\text{final}} = 2\sqrt{\mathbf{K}_p^{\text{final}} \cdot M_{\text{eff}} \cdot \zeta} \quad (7.6)$$

where $\alpha_{\text{force}} \in [0.2, 1]$ is a force-based scaling factor that reduces stiffness when measured forces exceed safe thresholds. This ensures that even if the VLM suggests high stiffness based on visual context, the controller remains compliant in response to unexpected contact forces. $\mathbf{D}_p^{\text{final}}$ is computed to ensure critical damping, where M_{eff} is the effective mass and $\zeta = 0.7$.

The VLM applies phase-based impedance hierarchy: highest impedance for Free_motion (precise position control), medium for Approaching (transitioning to compliance), lowest for Contact (maximum compliance), and medium for Retreat (controlled withdrawal). The VLM also considers motion direction adaptation: reduced impedance along the intended motion direction and elevated impedance perpendicular to motion for alignment maintenance.

7.2.3 Hybrid VLA-VIC Control Architecture

Our control architecture seamlessly integrates VLA action generation with VIC execution enhanced by VLM, maintaining the benefits of VLA generalization while adding compliant safety. The system operates at three temporal scales:

- **VLM-informed impedance generation (~1 Hz):** The VLM processes visual-linguistic context and force feedback to generate impedance parameters.
- **VLA action chunk (~3 Hz):** VLA generates desired end-effector displacements $\mathbf{x}_d$ based on current observations and instructions.
- **Low-level controller (1000 Hz):** VIC controller tracks desired poses with adaptive compliance to maintain safe contact.

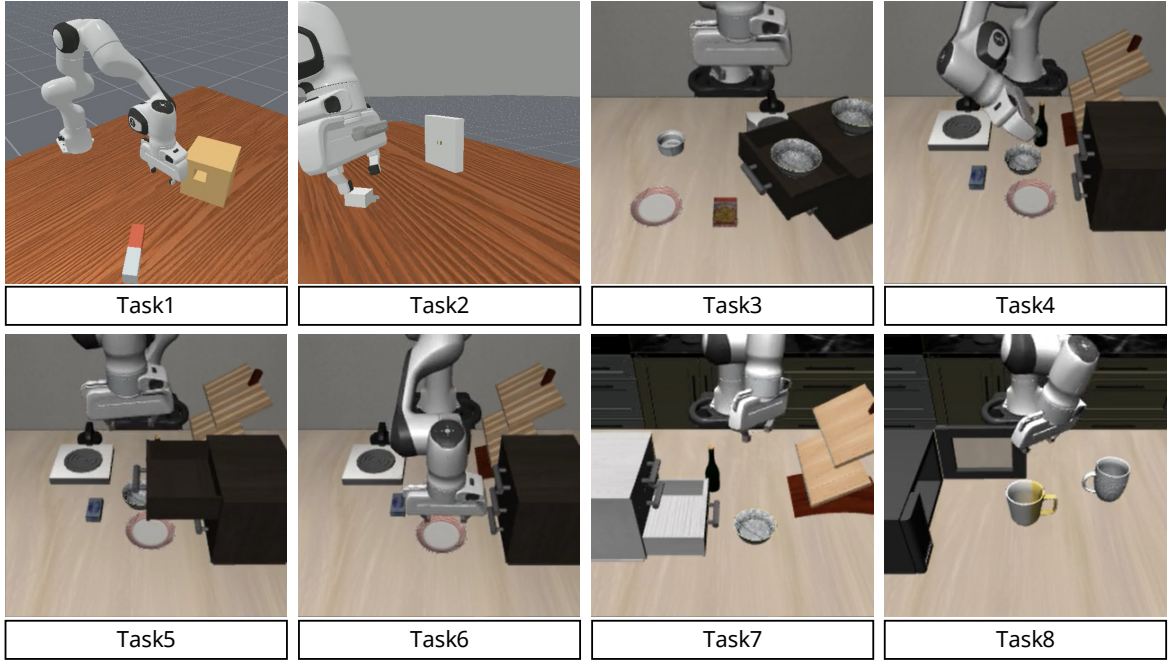


Figure 7.3 Contact-rich tasks in simulation. Task 1 and Task 2 are from ManiSkill for RDT, while the remaining tasks come from LIBERO for Pi0 and OpenVLA. Together, they form a contact-centric task suite.

7.3 Experiments

We evaluate CompliantVLA-adaptor on a suite of contact-rich manipulation tasks, such as object insertion and contact-sensitive pick-and-place. We compare our approach against the SOTA VLA models, which rely on position control without force adaptation.

7.3.1 Simulation Tasks

All simulations are conducted in the same environment as the baselines to ensure fair comparison. We compare our approach against three VLA models: Pi0 [206], RDT-1B [208], and OpenVLA-oft [211]. Focusing on contact-rich tasks, we selected 8 representative tasks from the LIBERO [213] and ManiSkill benchmarks [214]. All experiments were performed on 4 NVIDIA RTX A6000 GPUs.

The eight tasks are summarized in Table 7.1. These tasks are chosen to evaluate the CompliantVLA-adaptor’s ability to handle complex contact dynamics that require both semantic understanding and safe physical interaction.

Experiment protocol: We adopt a two-stage evaluation protocol to isolate the contribution of CompliantVLA-adaptor under the same contact force safe constraint:

Table 7.1 List of tasks in simulation

Task No.	Descriptions
Task 1	(Peg Insertion) Pick up an orange-white peg and insert the orange end into the box with a hole in it
Task 2	(Plug Charger) Pick up one of the misplaced shapes on the board/kit and insert it into the correct empty slot
Task 3	(Drawer Env) pick up the black bowl in the top drawer of the wooden cabinet and place it on the plate
Task 4	(Open Drawer) open the middle drawer of the cabinet
Task 5	(Open Drawer) open the top drawer and put the bowl inside
Task 6	(Push Plate) push the plate to the front of the stove
Task 7	(Close Drawer) put the black bowl in the bottom drawer of the cabinet and close it
Task 8	(Microwave Env) put the yellow and white mug in the microwave and close it

- *Stage 1 - Baseline VLA Performance:* We evaluate each baseline VLA model using its default control strategy. Since all three models employ position- or trajectory-based control without explicit force feedback, there is no compliance adaptation during contact. We set a strict force threshold of 30 N to ensure safety during execution. A task is terminated after three consecutive threshold violations or if not completed within the time limit.
- *Stage 2 - CompliantVLA-adaptor Integration:* We augment each baseline VLA model with the proposed CompliantVLA-adaptor, replacing its low-level controllers with our VLM-enhanced VIC controller for compliant execution.

Contact force measurement and safety criteria: Contact forces are measured using a simulated wrist-mounted force/torque sensor. For all selected contact-sensitive tasks, the force threshold is set to 30 N. If the contact force exceeds this threshold three times consecutively during a single trial, the task is immediately terminated and marked as failed.

7.3.2 Simulation Results

We present quantitative results from our simulation tasks shown in Fig. 7.4, comparing the performance of our CompliantVLA-adaptor with the SOTA VLA baselines. The results show that our approach achieves significantly higher task success rates and reduces force violations across most of the contact-rich manipulation tasks (7/8).

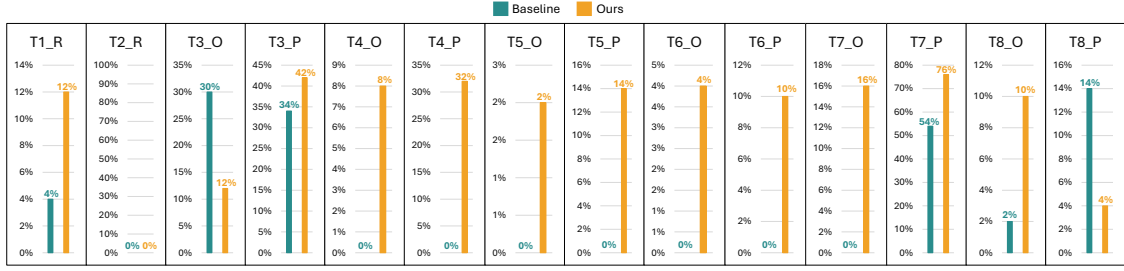


Figure 7.4 Evaluation results of task success rates under contact force constraint across 8 related tasks in simulation. The title of each bar diagram indicates a different setup, where “T” with a value denotes task number, and the suffix “-R”, “-P” and “-O” means using RDT, Pi0 and OpenVLA-of model, respectively. Different colors indicate the run with and without our CompliantVLA-adaptor.

The baseline VLA models exhibit highly unstable performance across the task suite, with a maximum success rate of only 54% and some tasks always failing if considering contact force during execution. After integrating the CompliantVLA-adaptor, most of the evaluated tasks showed consistent improvements while staying safe, with a maximum success rate of 76% and significant improvement among tasks that always failed in baselines. The average success rate across all tasks increases from 9.86% to 17.29%, representing a $1.74\times$ increase.

Tasks involving mechanical constraints (drawers, stove knobs) show the most dramatic relative improvements, as these scenarios particularly benefit from the adaptive compliance provided by our VLM-guided impedance control. Conversely, tasks with higher baseline performance, such as simple pick-and-place operations, show more modest but still meaningful improvements, primarily through enhanced force safety during object grasping and release phases.

In terms of failure cases, baseline VLA models frequently encounter force violations leading to task termination, particularly in RDT plug charger task (Task 2) and both Pi0 and OpenVLA-of in drawer manipulation tasks (Tasks 3–5). In contrast, our CompliantVLA-adaptor significantly reduces these violations by dynamically adjusting impedance parameters based on real-time force feedback and task context.

7.3.3 Real-World Experiments

Our experiment setup consists of a 7-DoF Franka Emika Panda robot arm with a Franka Panda parallel gripper. For visual perception, we employed a dual-camera configuration: a Logitech USB camera as the main camera providing a global view, and an Intel RealSense

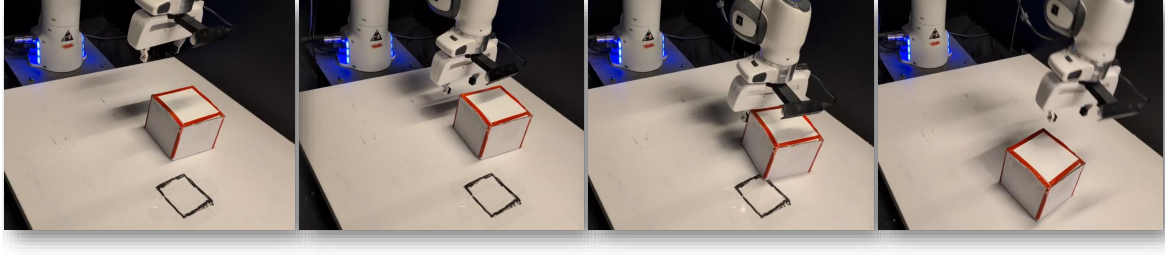


Figure 7.5 Real-world experiment setups. Left: box-pushing setup, where the robot pushes a 3 kg box toward a marked target region. Right: whiteboard-wiping setup, where the robot removes marker traces on the board.

D435i depth camera mounted on the end-effector as the wrist camera. Images from both cameras are uniformly resized to 224×224 pixels before being fed into the model.

We completed two real-world tasks: “push the heavy box to the target position” and “wipe the whiteboard and clean the mark” (Fig. 7.5).

7.3.4 Real-World Results



Figure 7.6 Real-world wiping task execution.

As shown in Fig. 7.6 and Fig. 7.8, the adaptor consistently avoids unsafe forces, demonstrating its effectiveness as a safety module. The stiffness and force profiles indicate that the robot maintains compliant interactions, adjusting its impedance parameters in response to contact conditions, as shown in Fig. 7.7 and Fig. 7.9, respectively. The interaction force remains within safe limits throughout the task execution, validating the safety benefits of our approach in real-world scenarios.

In the wiping task, the robot successfully modulates its stiffness in range of $[400, 600]$ N/m to maintain gentle contact with the whiteboard, while in the pushing task, it adapts its impedance to safely push the box with higher stiffness in $[800, 1000]$ N/m. The force profiles show smooth transitions without abrupt spikes, indicating that the robot is effectively regulating contact forces to prevent damage to the environment and itself.

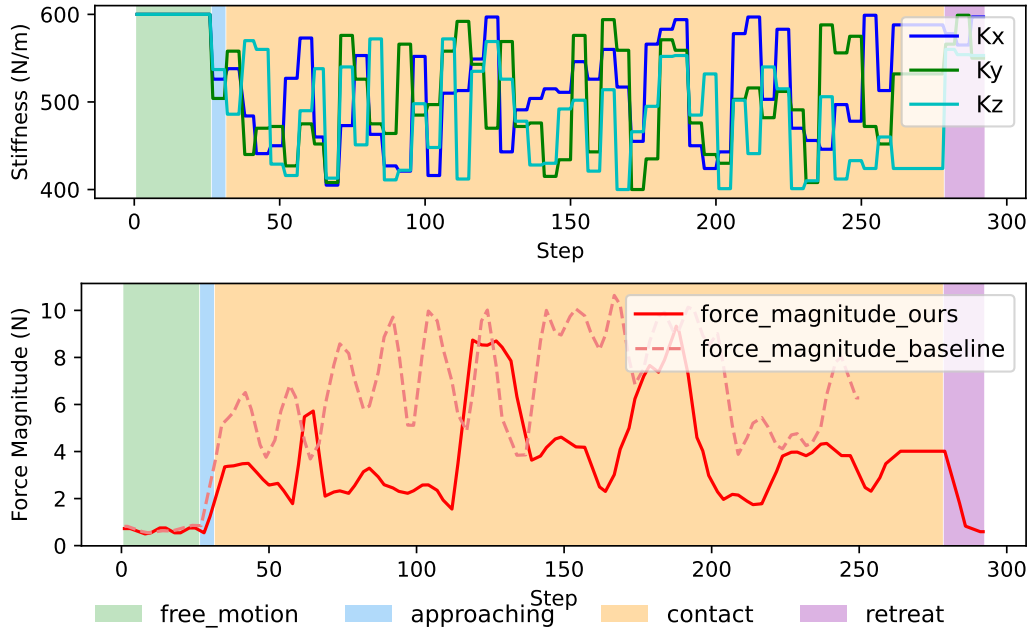


Figure 7.7 Wiping task: adaptive stiffness regulation and measured contact force. Background colors indicate task phases.

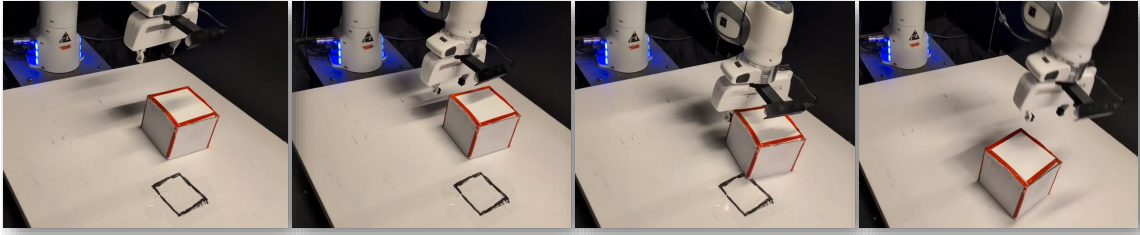


Figure 7.8 Real-world pushing task execution.

7.4 Discussion and Limitations

While our CompliantVLA-adaptor demonstrates significant improvements in safe contact-rich manipulation, several limitations remain:

1. VLM inference is slower than the servo loop; it is not on the safety-critical path: abnormal-contact handling remains in the 1000 Hz force-regulated VIC, while VLM latency only limits the refresh rate of the semantic impedance prior.
2. Using API calls is expensive and not eco-friendly. On the other hand, using onboard VLM is computationally heavy.

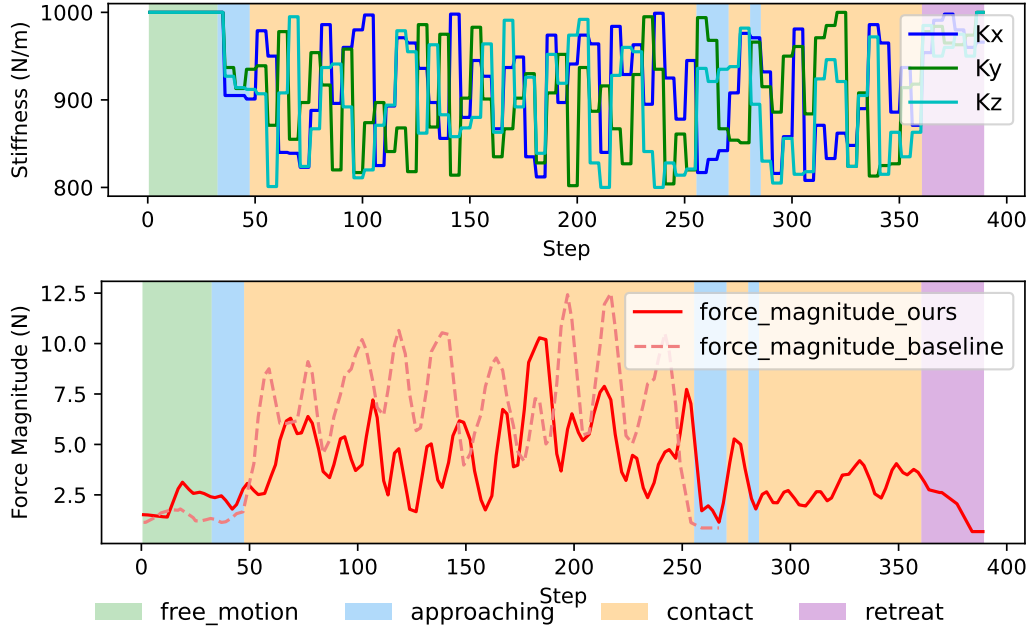


Figure 7.9 Pushing task: adaptive stiffness regulation and measured contact force. Background colors indicate task phases.

3. VLMs may not generalize well to unseen objects or environments, leading to suboptimal performance in novel situations.

7.5 Summary

In this chapter, we proposed a CompliantVLA-adaptor that leverages physical context-informed VLMs with VIC to enhance the performance of VLA in safe and effective manipulation in contact-rich tasks. By leveraging the semantic grounding of VLMs and the physical robustness of VIC, we enable robots to adapt their impedance parameters based on task context derived from visual and language inputs. Our experimental results validate the effectiveness of the proposed approach in both simulation and real-world experiments, achieving higher task success rates and reducing force violations. This adaptor provides a lightweight, plug-and-play safety layer for existing VLA models, offering a practical step toward safe and generalizable contact-rich manipulation, bridging high-level semantic reasoning and low-level compliant control.

This adaptor therefore serves as an intermediate step between semantically guided compliance and richer physically grounded foundation-model policies. While it equips VLA systems with safer compliant execution, it still relies primarily on vision and language for

perception. The next chapter extends this line of work by incorporating tactile sensing into the VLA framework, enabling robots to perceive and respond to fine-grained contact information during manipulation.

Chapter 8

Contact-Aware Tactile Fusion for Robust Vision-Language-Action Manipulation

8.1 Introduction

The previous chapter enhanced VLA models from the control side by introducing compliant execution through impedance adaptation. In this chapter, we complement that direction from the perception side by investigating how tactile sensing can provide the missing physical feedback required for robust contact-rich manipulation.

Vision-Language-Action (VLA) models have demonstrated remarkable potential in achieving generalized robotic manipulation [215, 174, 216]. By integrating pretrained Vision Language Models (VLMs), these models can effectively interpret complex natural language instructions and perceive unstructured environments. Their ability to reason causally and plan sequentially allows them to bridge the gap between high-level semantic understanding and low-level control policies, providing strong general-purpose backbones that support deployment across diverse robot embodiments and real-world scenarios.

Despite these advancements, a significant limitation of current VLA architectures is their predominant reliance on vision and language as the sole modalities for perceiving the physical world [217]. This visual dependency becomes a critical bottleneck in tasks involving severe visual occlusions, where the robot’s end-effector or the manipulated object itself obstructs the camera’s view. Moreover, for fine-grained contact-rich manipulation tasks such as disassembly or precise assembly, visual signals alone often fail to capture essential physical properties like contact forces, friction, and surface textures [218, 29]. Without direct feedback from physical contact, robots struggle to dynamically adjust their actions in response to unexpected resistance or slippage, leading to suboptimal performance. Therefore,

integrating complementary modalities that can sense physical interactions is essential for enhancing the robustness and precision of VLA models in contact-rich environments.

Recent studies have begun incorporating tactile sensing into VLA-style policies [219–225], showing that touch complements vision and language when visual information is limited or occluded. Tactile feedback provides direct signals of contact events, slippage, and local geometry, enabling more reliable grasp stabilization and corrective adjustments. However, most existing evaluations have limited emphasis on contact-intensive fine-grained manipulation tasks that require sustained and precise contact control, such as disassembly. Moreover, many existing approaches treat tactile observations as image-like inputs, relying on dense pixel representations that increase token length and computational cost in transformer architectures. Simply concatenating tactile tokens with visual and language tokens can also lead to inefficient cross-modal interaction, particularly during non-contact phases.

This raises a key question: **how should tactile modality be effectively integrated into VLA architectures to support robust fine-grained manipulation?**

This progression is natural within the overall thesis: once compliance has been introduced to make VLA execution safer, the next step is to improve the model’s awareness of contact itself, especially in situations where vision becomes unreliable.

To address these challenges, we propose **TacVLA**, an extension of VLA that integrates a compact tactile array into the transformer-based policy for contact-rich manipulation. Instead of treating tactile signals as image-like inputs, we represent them as low-dimensional tactile tokens, enabling efficient multimodal processing without substantially increasing token length. We further introduce a **contact-aware gating mechanism** that selectively activates tactile tokens only when physical contact is detected. This design allows the model to leverage tactile feedback during interaction while avoiding unnecessary cross-modal interference during non-contact phases.

The contributions of this chapter are threefold:

- We introduce TacVLA, a tactile-enhanced extension of VLA that enables efficient multimodal integration for contact-rich fine-grained manipulation.
- We propose a compact tactile tokenization scheme and a contact-aware gating mechanism that adaptively regulates cross-modal interaction within the transformer architecture.
- We validate our approach on contact-rich constraint-locked disassembly, in-box picking and robustness benchmarks, demonstrating consistent improvements particularly under visual occlusion and dynamic physical interactions.

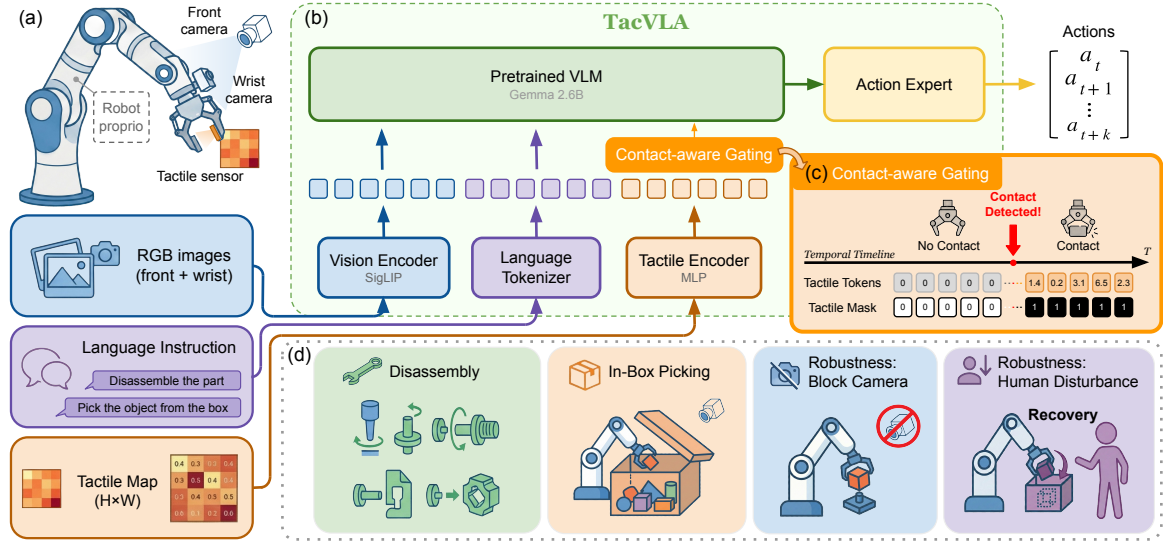


Figure 8.1 **Overview of TacVLA.** (a) Input modalities including visual observations, language instructions, and tactile measurements. (b) TacVLA architecture, consisting of modality-specific encoders and tokenizer, a pretrained VLM backbone, an action expert, and the contact-aware gating module. (c) The proposed contact-aware gating module that selectively activates tactile tokens based on the contact state, enabling adaptive multimodal fusion during contact-rich manipulation. (d) Experimental evaluation on contact-rich constraint-locked disassembly and in-box picking tasks, together with robustness tests under camera occlusion and human disturbance.

8.2 Methodology

We present the architecture and training procedure of our proposed TacVLA model. The core idea behind TacVLA is to integrate tactile sensing into a transformer-based VLA framework, enabling efficient multimodal fusion and adaptive grounding for contact-rich manipulation tasks. To achieve this, we design a unified tokenization scheme, and introduce a contact-aware gating mechanism that selectively activates tactile tokens based on contact state. See Fig. 8.1 for an overview of the model architecture.

8.2.1 Overall Architecture

Our TacVLA architecture follows a transformer-based VLA framework, extended with tactile perception for contact-rich manipulation. As shown in Fig. 8.1(b), the model consists of four components: modality tokenizers, a pretrained VLM backbone, an action expert, and a contact-aware gating module.

Visual observations from both the front and wrist-mounted cameras are encoded using a SigLIP [226] visual encoder, producing a sequence of visual tokens per image. Language

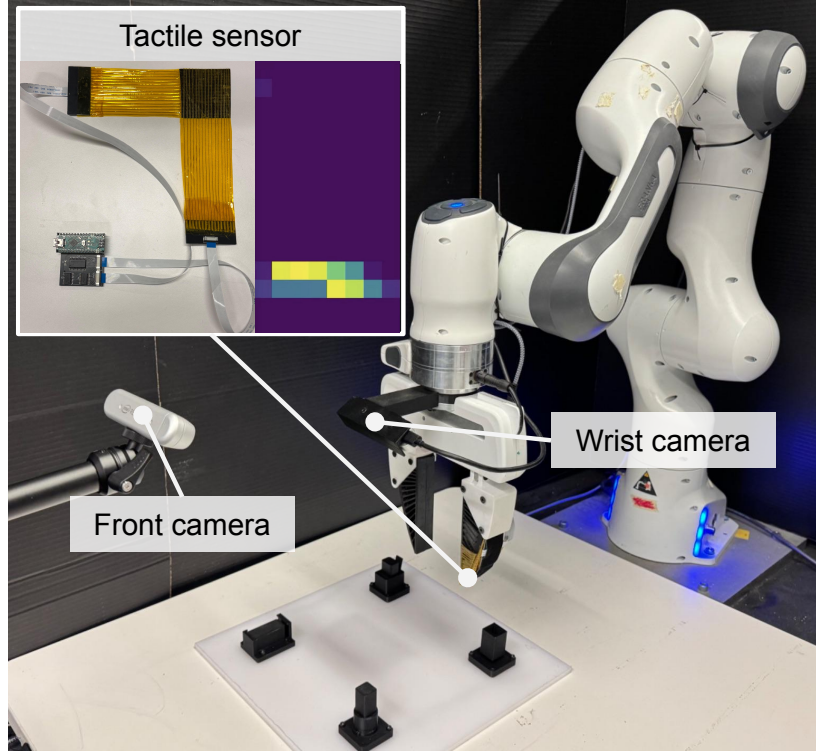


Figure 8.2 Hardware setup: a 7 DoF Franka robotic platform equipped with tactile sensors and two cameras for visual input to evaluate TacVLA on contact-rich manipulation tasks.

instructions along with robot proprioception are tokenized using the PaliGemma tokenizer. For tactile sensing, the 15×8 tactile array (Fig. 8.2) is embedded using a lightweight MLP-based encoder, which projects the tactile map into 36 tactile tokens. We further add fixed 2D sine-cosine positional embeddings to preserve spatial structure. This design yields a compact yet expressive representation that captures local contact patterns, global pressure distribution, and contact geometry.

Let the multimodal token sequence at time t be

$$\tilde{\mathbf{z}}_t = [\mathbf{z}_t^{vis}, \mathbf{z}_t^{lan+pro}, \tilde{\mathbf{z}}_t^{tac}], \quad (8.1)$$

where $\mathbf{z}_t^{vis}$, $\mathbf{z}_t^{lan+pro}$ denote visual tokens, language plus proprioceptive tokens, and $\tilde{\mathbf{z}}_t^{tac}$ denotes tactile tokens processed by the contact-aware gating module. All modality tokens are concatenated into a shared token sequence and fed into a pretrained VLM backbone. A non-causal attention mechanism over this prefix allows vision, language, and tactile tokens to freely cross-attend, producing a deeply integrated contextual representation for downstream control.

The fused representation is then provided as a prefix to an action expert module, following the Pi0.5 [215] design, which is trained with a flow-matching objective and predicts continuous action sequences. This policy predicts an action sequence conditioned on the multimodal tokens,

$$\mathbf{a}_{t:t+H} \sim \pi_{\theta}(\mathbf{a}_{t:t+H} \mid \tilde{\mathbf{z}}_t). \quad (8.2)$$

8.2.2 Contact-Aware Gating Tactile Fusion

Naively concatenating tactile tokens throughout the trajectory can be suboptimal, as tactile signals are typically uninformative during non-contact phases and may introduce unnecessary attention tokens.

To address this, we introduce a contact-aware token gating module as demonstrated in Fig. 8.1(c). Physical contact is detected using a threshold-based criterion: contact is declared when the number of taxels exceeding a predefined pressure threshold surpasses a fixed count. Rather than dynamically inserting or removing tokens, we maintain a fixed token structure and treat tactile tokens as conditional inputs.

Let $c_t \in \{0, 1\}$ denote the contact flag at time t , where $c_t = 1$ only when physical contact is detected. We apply a contact-dependent attention mask M_t^{tac} and embedding gating to tactile tokens $\mathbf{z}_t^{tac}$:

$$M_t^{tac} = c_t \cdot \mathbf{1}, \quad (8.3)$$

$$\tilde{\mathbf{z}}_t^{tac} = c_t \cdot \mathbf{z}_t^{tac}, \quad (8.4)$$

where $\mathbf{1}$ denotes a mask vector with the same length as the tactile token sequence. When $c_t = 0$, tactile tokens are excluded from the attention computation and cannot participate in cross-modal interaction. In addition, zeroing tactile tokens suppresses offsets from embedding layers or positional encodings and reduces the influence of pre-contact sensor noise.

By preserving a fixed token topology while enabling state-dependent modality routing, the proposed module allows tactile information to influence policy decisions only when physically informative.

8.2.3 Training Procedure

The training dataset consists of synchronized visual observations, language instructions, tactile maps, robot proprioception, and ground-truth action trajectories. We fine-tune TacVLA using Low-Rank Adaptation (LoRA) on top of an OpenPI backbone checkpoint (pi05_base).

All models are fine-tuned for 10,000 gradient steps under consistent optimization settings, while the parameters of the tactile encoder are frozen during fine-tuning.

8.3 Experiments

We evaluate the performance of our TacVLA model on a set of disassembly and in-box picking tasks, then compare it against baseline models. We also conduct robustness evaluations to assess the model’s performance under various conditions, such as visual occlusion and human disturbance.

8.3.1 Experimental Setup

Our experimental platform, shown in Fig. 8.2, consists of a 7-DoF Franka Emika Panda robotic arm equipped with parallel grippers. For visual perception, we employ two RGB cameras: a fixed front-facing camera providing a global scene view, and a wrist-mounted camera offering close-range observations. To capture physical interaction signals, we integrate a tactile array sensor, adapted from [227], mounted on the surface of one of the fingers. The sensor has a spatial resolution of 15×8 , providing 120 tactile measurements per frame. These tactile readings encode contact distribution across the sensor surface and are used as additional input to the policy during manipulation.

Disassembly Tasks

We design a set of contact-rich constraint-locked disassembly tasks (Fig. 8.3) where we structure scenarios that require fine-grained adjustments based on tactile feedback. In these tasks, components are locked by different geometric constraints and must be separated through specific motions such as pressing, twisting, or sliding. The four tasks are:

- **Task 1: Tight Shaft** — pull a shaft tightly fitted into a housing;
- **Task 2: Press Clip** — press and release a clip-locked component;
- **Task 3: Shaft Rotation** — rotate and extract a component with a rotational constraint;
- **Task 4: Slide Pull** — slide and pull a component constrained by a rail.

These tasks require precise manipulation and the ability to adapt to changing contact conditions, making them ideal for evaluating the benefits of tactile sensing in a VLA framework.

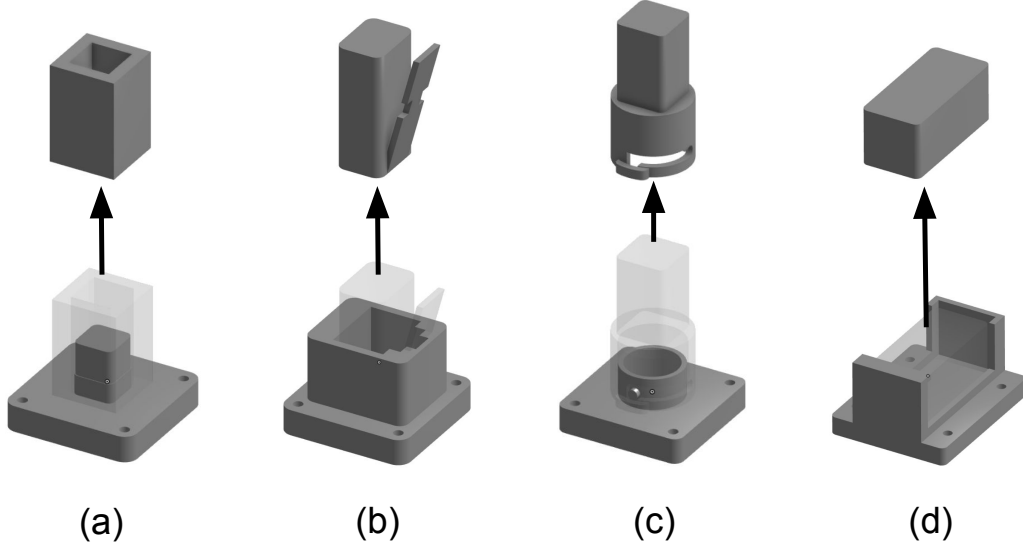


Figure 8.3 Four contact-rich constraint-locked disassembly tasks with diverse geometric constraints: (a) Task 1: tight shaft; (b) Task 2: press clip; (c) Task 3: shaft rotation; (d) Task 4: slide pull.

In-Box Picking

We also conduct an in-box manipulation task (Fig. 8.4), where the robot must retrieve objects from a confined space with limited visual access: (1) the front camera has no access to the box inside, and (2) the wrist camera has limited illumination and occlusion inside the box. This task further emphasizes the importance of tactile sensing for handling scenarios with severe visual occlusion and complex contact interactions.

Baselines

We compare TacVLA with two diffusion-based baselines under comparable training settings: image-based Diffusion Policy [228] and 3D Diffusion Policy [229]. Both baselines take multi-view images and the same tactile measurements as input, encode each modality with standard visual backbones, and concatenate the resulting features to condition the diffusion model for action generation. All policies are trained for 200 epochs, and the final checkpoints are used for evaluation.

8.3.2 Data Collection

To train and evaluate TacVLA, we collect a real-world dataset of contact-rich manipulation tasks. The dataset includes four disassembly tasks and one in-box picking task, each

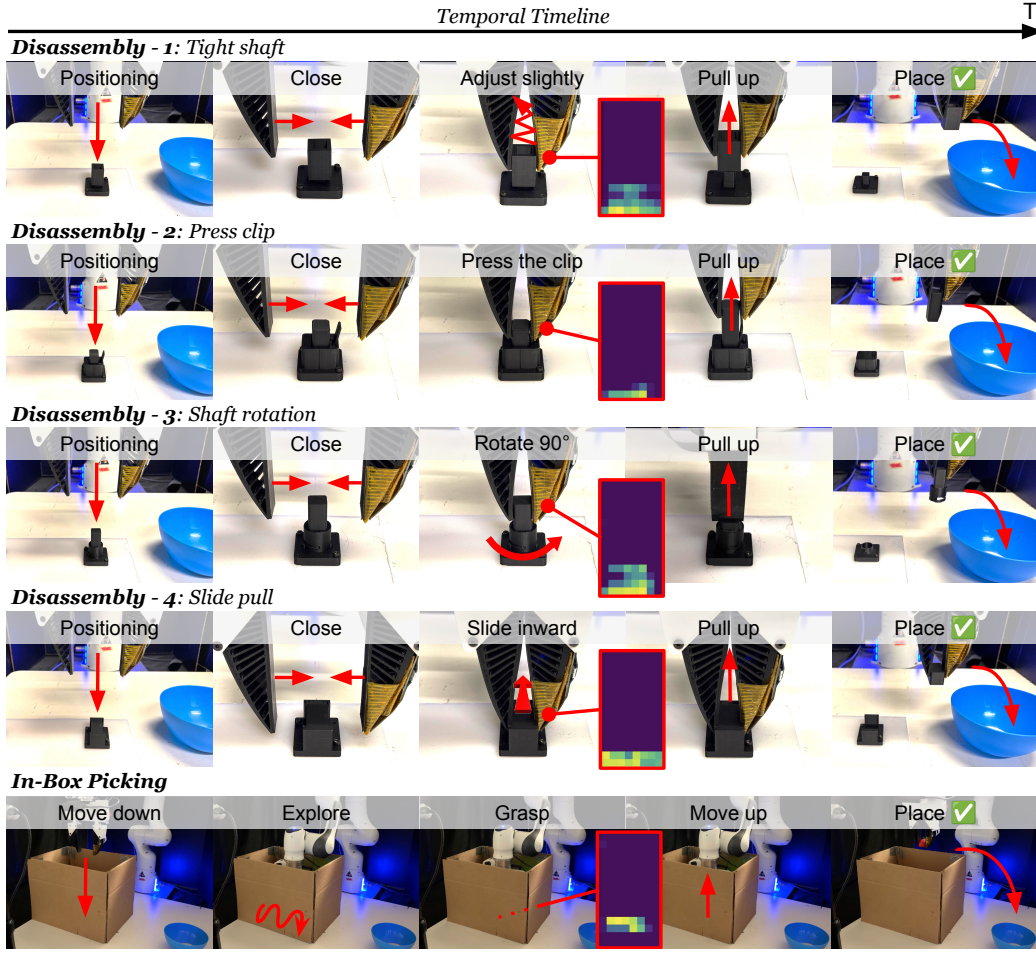


Figure 8.4 The real-world experimental setup and procedures for four constraint-locked disassembly and in-box picking tasks. The experiments demonstrate the capability of TacVLA in contact-rich fine-grained manipulation, as well as its robustness to visual occlusion in the in-box picking scenario.

containing 50 demonstrations collected by human teleoperation. Visual data consists of RGB images captured from a fixed front camera and a wrist-mounted camera. All modality streams are recorded at 10 Hz and temporally aligned to ensure synchronized observations at each timestep.

8.4 Results and Analysis

8.4.1 Main Results

Table 8.1 reports the success rates across four disassembly tasks and the in-box picking task among our model, Pi0.5 baseline and diffusion policy baselines.

Table 8.1 Performance comparison across Constraint-locked Disassembly and In-Box Picking tasks

Method	Constraint-locked Disassembly					In-Box Picking
	Task 1	Task 2	Task 3	Task 4	Average	
3D Diffusion Policy + Tactile	8/20 (40%)	0/20 (0%)	9/20 (45%)	0/20 (0%)	31.25%	1/20 (5%)
Diffusion Policy + Tactile	14/20 (70%)	14/20 (70%)	6/20 (30%)	5/20 (25%)	48.75%	0/20 (0%)
Finetuned Pi0.5	16/20 (80%)	16/20 (80%)	13/20 (65%)	6/20 (30%)	63.75%	2/20 (10%)
TacVLA (Ours)	20/20 (100%)	18/20 (90%)	14/20 (70%)	15/20 (75%)	83.75%	14/20 (70%)

Disassembly Results

Our TacVLA achieves the best overall performance, with an average success rate of 83.75%, significantly outperforming the finetuned Pi0.5 baseline (63.75%) and diffusion-based policies. While the Pi0.5 baseline performs reasonably well on Task 1 and Task 2 (both 80%), its performance drops substantially on Task 3 (65%) and Task 4 (30%), where contact-dependent adjustments are critical. In contrast, TacVLA maintains consistently strong performance across all tasks, achieving 100% and 90% success rates on Task 1 and Task 2, and improving to 70% and 75% on Task 3 and Task 4, respectively.

Notably, the largest gain is observed in Task 4 with a 45% improvement, where it involves fine-grained sliding interactions with partial occlusion. In Task 4, the Pi0.5 baseline frequently becomes stuck in intermediate states and attempts repeated re-grasping, indicating uncertainty in contact dynamics. In contrast, TacVLA demonstrates more stable behavior during contact transitions, suggesting that incorporating tactile sensing with state-dependent gating significantly enhances performance in fine-grained disassembly tasks.

In-Box Picking Results

The in-box picking task presents severe visual challenges: the front camera has no visibility inside the box, and the wrist camera suffers from limited illumination and frequent occlusion. Under these conditions, the finetuned Pi0.5 baseline achieves only 10% success rate. The baseline often attempts to execute the lifting motion without successfully grasping the object, indicating a lack of reliable in-hand state awareness in the absence of tactile feedback.

In contrast, TacVLA achieves a 70% success rate, representing a substantial improvement under heavy occlusion. Despite poor lighting and occasional object shifts due to imperfect initial grasps, TacVLA actively explores within the box and performs multiple re-grasp attempts. Importantly, it proceeds to the lifting and placing stage only after contact is physically detected.

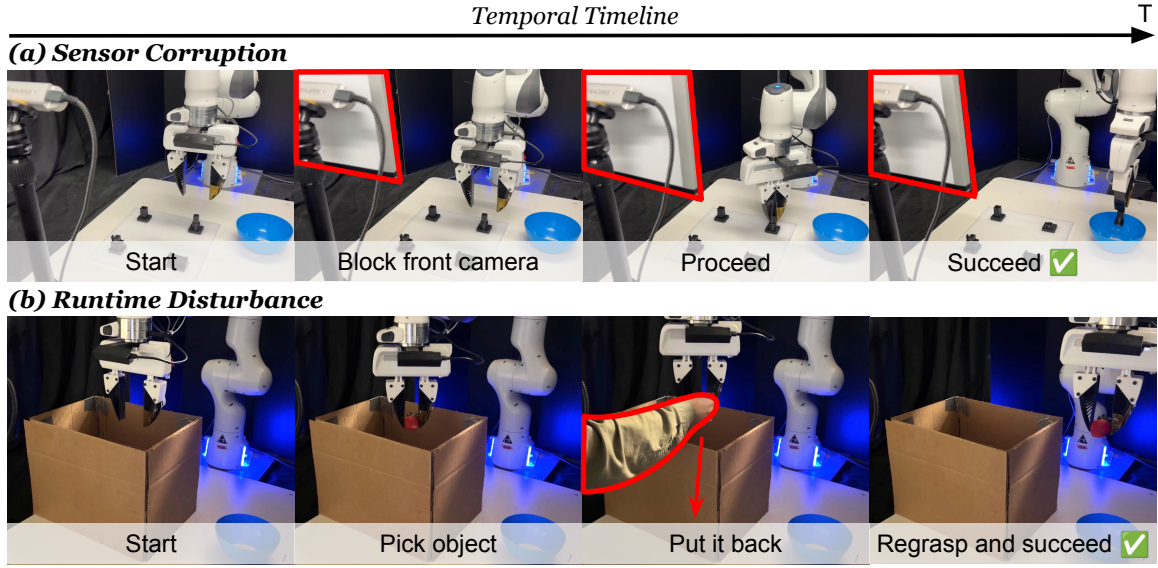


Figure 8.5 Robustness Evaluation: the performance of TacVLA under conditions of visual occlusion and runtime disturbance, demonstrating its ability to adapt and maintain performance in challenging scenarios.

Benchmark with Diffusion Policy

We further compare TacVLA with diffusion-based policies augmented with tactile input. On disassembly, the diffusion policy achieves an average success rate of 48.75%, while the 3D diffusion policy achieves 31.25%, both substantially lower than TacVLA’s 83.75%. Under severe visual occlusion in the in-box picking task, diffusion policies perform poorly (0% and 5%, respectively). We attribute this difference primarily to the architectural and training paradigms: TacVLA builds upon a pretrained multimodal backbone and is fine-tuned via LoRA, enabling it to leverage rich visual-language priors and structured cross-modal representations.

8.4.2 Robustness Evaluation

Block Camera

As a critical test of the model’s reliance on visual information, we evaluate performance under severe visual occlusion by physically blocking the front camera during task execution, as shown in Fig. 8.5(a). The vision-only baseline (finetuned Pi0.5) experiences a substantial performance drop, achieving 40%, 40%, 5%, and 35% success rates across Tasks 1–4, respectively. In contrast, TacVLA maintains significantly higher performance, reaching 70%, 65%, 45%, and 70%. Notably, Task 3 exhibits the largest improvement (+40%). Overall,

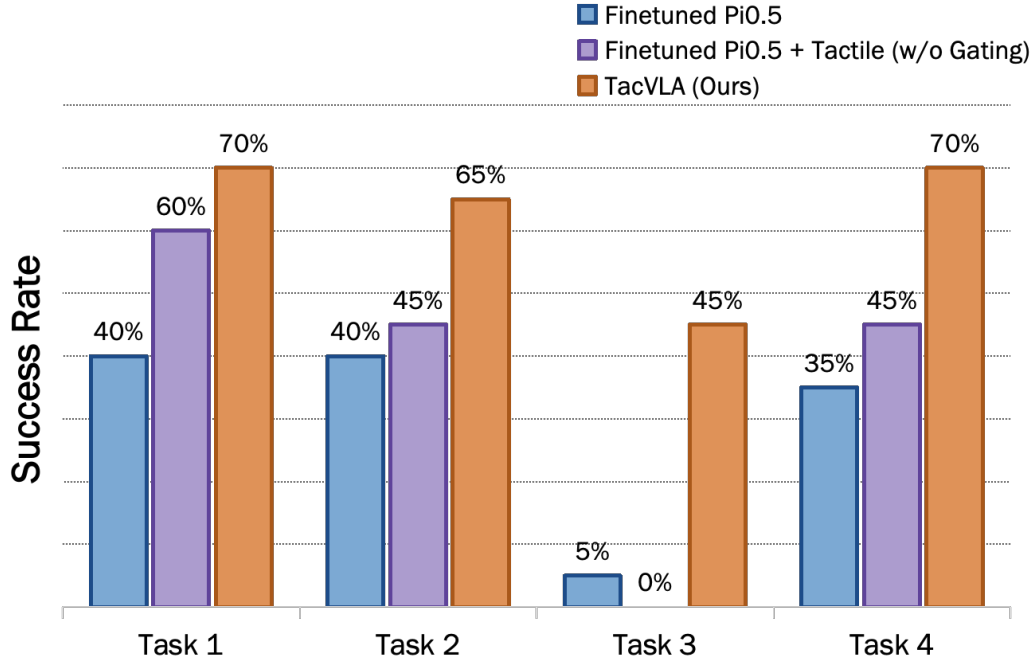


Figure 8.6 Success rates under block camera disturbance across four constraint-locked disassembly tasks. We compare finetuned Pi0.5, finetuned Pi0.5 with tactile input (w/o gating), and TacVLA (ours). Our method consistently improves performance under visual occlusion.

TacVLA improves the average success rate from approximately 30% to over 60% under visual occlusion, demonstrating enhanced robustness through state-dependent tactile integration.

Human Disturbance

To further assess robustness under dynamic environmental changes, we introduce human disturbances during task execution, as illustrated in Fig. 8.5(b). Specifically, while the robot is manipulating the object, a human operator unexpectedly places the object back into the box. As a result, TacVLA successfully detects the unexpected change, returns to the box, and re-grasps the object before continuing task execution. In contrast, the finetuned Pi0.5 baseline fails to recover from the disturbance and is unable to retrieve the object, indicating limited adaptability when unexpected environmental disturbance occurs.

8.4.3 Ablation Study

To isolate the effect of contact-aware gating, we compare TacVLA with a baseline that directly concatenates tactile tokens without applying the proposed contact-aware gating module, denoted as *Finetuned Pi0.5 + Tactile (w/o Gating)*.

Table 8.2 Ablation study on the effect of the contact-aware gating module

Method	Constraint-locked Disassembly					In-Box Picking
	Task 1	Task 2	Task 3	Task 4	Average	
Finetuned Pi0.5 + Tactile (w/o Gating)	18/20 (90%)	16/20 (80%)	12/20 (60%)	11/20 (55%)	71.25%	8/20 (40%)
TacVLA (Ours)	20/20 (100%)	18/20 (90%)	14/20 (70%)	15/20 (75%)	83.75%	14/20 (70%)

On the disassembly tasks, removing gating reduces the average success rate from 83.75% to 71.25%. Notably, on Task 3, the no-gating variant performs worse than the Pi0.5 baseline without tactile input (60% vs 65%), indicating that naive tactile fusion can even degrade performance.

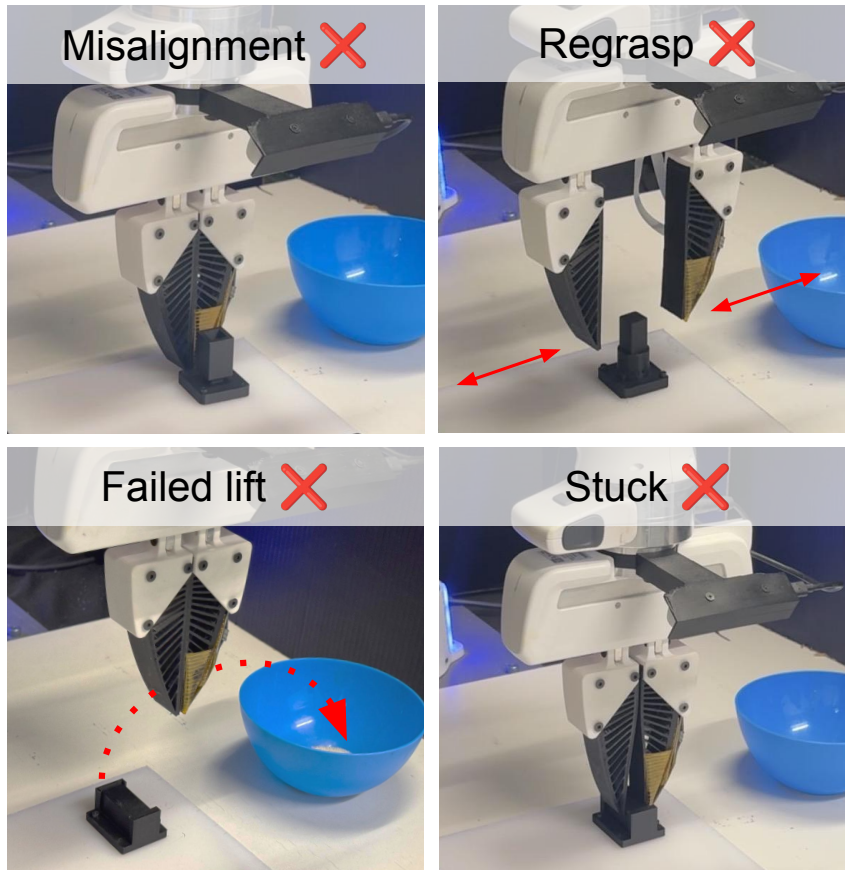


Figure 8.7 Failure cases of *Finetuned Pi0.5 + Tactile (w/o Gating)* during constraint-locked disassembly. Without contact-aware gating, tactile tokens remain active even in non-contact phases, leading to unstable behaviors.

Qualitatively, as shown in Fig. 8.7, the no-gating variant frequently exhibits misalignment during object approach, repeated re-grasp attempts, stalled intermediate states, and failed lift

attempts. These behaviors suggest that unconditional tactile token injection may interfere with visual localization before stable contact is established.

On the in-box picking task, performance drops from 70% to 40% without gating. Under severe visual occlusion (front camera blocked), removing the gating mechanism results in a substantial performance drop, reducing the average success rate by approximately 25% compared to TacVLA. This degradation indicates that contact-aware gating plays a critical role in leveraging tactile feedback when visual information is unreliable.

8.5 Discussion and Limitations

Although TacVLA improves performance in contact-rich manipulation, several limitations remain. First, the contact-aware gating relies on a binary threshold heuristic, which does not allow gradual or learnable adjustment of modality importance. More adaptive modality weighting strategies could further improve cross-modal coordination. Second, the tactile sensor provides efficient contact-state signals but has limited spatial resolution, restricting fine-grained contact geometry reasoning. Finally, our evaluation focuses on short-horizon, contact-centric tasks; extending state-dependent modality routing to longer-horizon and more complex task settings remains future work.

8.6 Summary

In this chapter, we proposed TacVLA, a vision-language-action model augmented with tactile sensing for contact-rich manipulation. Through unified tokenization and contact-aware gating, TacVLA enables state-dependent multimodal fusion and robust cross-modal grounding. Experiments on disassembly, in-box picking, and robustness evaluations show clear improvements over pure VLA and diffusion-based policies, with an average 20% improvement in disassembly success rates, 60% improvement in in-box picking, and a $2.1\times$ improvement in scenarios with visual occlusion.

Together, the CompliantVLA-adaptor (Chapter 7) and TacVLA demonstrate two complementary routes for extending foundation models toward contact-rich manipulation: the former through compliance-aware control, and the latter through contact-aware multimodal perception. The next chapter closes the thesis by synthesizing how these foundation-model-based developments relate back to the safe RL contributions presented earlier, and by outlining the broader research directions that emerge from their combination.

Chapter 9

Conclusions and Future Work

9.1 Summary of Contributions

The preceding chapters have developed the thesis argument in a staged manner, moving from problem formulation and literature grounding, to safe reinforcement learning methods for contact-rich tasks, and finally to foundation-model-based approaches enriched with compliant control and tactile sensing. This concluding chapter synthesizes these contributions into a unified perspective.

This thesis has addressed the challenge of safe and effective learning for contact-rich robotic manipulation, spanning from safe reinforcement learning to foundation models with multimodal sensing. The work has been organized around two main themes: (1) safe learning for contact-rich manipulation, and (2) foundation model-based manipulation with physical intelligence. We summarize the key contributions below.

9.1.1 Safe Reinforcement Learning with Variable Impedance Control (SRL-VIC)

In Chapter 3, we presented SRL-VIC, a safe RL framework that integrates variable impedance control with constrained policy optimization for contact-rich manipulation. The key insight was to use a maze-like exploration strategy that safely explores the task space while enforcing impedance-based safety constraints. The framework demonstrated that safe exploration can be achieved without sacrificing learning efficiency, and that variable impedance control provides a natural interface between the RL policy and the physical robot.

The experimental results on contact-rich manipulation tasks in both simulation and real-world environments demonstrated the effectiveness of the approach. SRL-VIC achieved higher task success rates than baseline methods while maintaining safety constraints.

9.1.2 Bio-Inspired Reflexive Safe RL (Bresa)

In Chapter 4, we presented Bresa, a bio-inspired safe RL method that incorporates a reflexive safety mechanism inspired by human reflexes. The method uses a two-level control architecture: a high-level RL policy that learns task-level behavior, and a low-level reflex controller that ensures safety during contact. The reflex controller is activated when the robot detects unexpected contact forces, providing rapid and reliable safety responses.

The experimental results demonstrated that Bresa achieves robust performance in the presence of unexpected disturbances, significantly outperforming baseline methods in terms of safety and task success rate. The bio-inspired reflex mechanism provides a principled approach to safety that is both effective and interpretable.

9.1.3 Passivity-Centric Safe RL

In Chapter 5, we presented a passivity-centric safe RL framework that enforces passivity constraints during both learning and execution. The framework uses energy tanks to monitor and control the energy flow during interaction, ensuring that the robot remains passive and safe. The passivity constraints are integrated into the RL optimization problem, enabling the robot to learn safe policies that are robust to uncertainty and disturbances.

The experimental results demonstrated that the passivity-centric approach achieves safe and stable interaction in contact-rich tasks, even in the presence of modeling errors and environmental variability. The energy tank formulation provides a principled approach to passivity enforcement that is compatible with modern RL methods.

9.1.4 OmniVIC: VLM-Guided Variable Impedance Control

In Chapter 6, we presented OmniVIC, a framework that integrates Vision-Language Models (VLMs) with variable impedance control for contact-rich manipulation. The framework uses a VLM to generate task-specific impedance parameters from natural language instructions and visual observations, enabling flexible and generalizable contact-rich manipulation. The key innovation was the design of a structured impedance parameter space that can be effectively controlled by a VLM.

The experimental results demonstrated that OmniVIC achieves effective contact-rich manipulation across a variety of tasks, outperforming baseline methods in terms of task success rate and force regulation. The VLM-guided impedance control enables the robot to adapt its compliance to different tasks and contact conditions, providing a flexible and generalizable approach to contact-rich manipulation.

9.1.5 CompliantVLA: VLA with Compliant Control

In Chapter 7, we presented CompliantVLA, a framework that extends VLA models with a compliance adaptor for contact-rich manipulation. The framework uses a lightweight compliance adaptor that modifies the VLA’s action outputs to incorporate impedance control, enabling safe and effective contact-rich manipulation. The key innovation was the design of a compliance adaptor that can be efficiently integrated into existing VLA architectures without requiring retraining.

The experimental results demonstrated that CompliantVLA achieves significant improvements in contact-rich manipulation performance compared to baseline VLA models, particularly in tasks that require precise force regulation. The compliance adaptor provides a practical and efficient approach to extending VLA models with physical intelligence.

9.1.6 TacVLA: Tactile-Enhanced VLA

In Chapter 8, we presented TacVLA, a framework that extends VLA models with tactile sensing for contact-rich manipulation. The framework uses a compact tactile tokenization scheme and a contact-aware gating mechanism to integrate tactile feedback into the VLA’s decision-making process. The key innovation was the contact-aware gating mechanism, which selectively activates tactile tokens only when physical contact is detected, avoiding unnecessary cross-modal interference during non-contact phases.

The experimental results on contact-rich constraint-locked disassembly and in-box picking tasks demonstrated clear improvements over pure VLA and diffusion-based policies, with an average 20% improvement in disassembly success rates, 60% improvement in in-box picking, and a $2.1\times$ improvement in scenarios with visual occlusion. The contact-aware gating mechanism provides a principled approach to tactile integration that is both effective and computationally efficient.

9.2 Connections Between Contributions

The contributions of this thesis form a coherent research program that progressively addresses the challenges of safe and effective contact-rich manipulation. The progression can be understood along two dimensions:

Safety mechanisms: The thesis progresses from constraint-based safety (SRL-VIC) to bio-inspired reflexive safety (Bresa) to passivity-based safety (Passivity-Centric Safe RL). Each approach addresses different aspects of safety in contact-rich manipulation, providing complementary perspectives on the problem.

Foundation model integration: The thesis progresses from VLM-guided impedance control (OmniVIC) to VLA with compliance control (CompliantVLA) to VLA with tactile sensing (TacVLA). Each approach extends foundation models with different aspects of physical intelligence, enabling increasingly robust and generalizable contact-rich manipulation.

The two themes are also connected: the safe RL methods provide the foundation for understanding the safety requirements of contact-rich manipulation, while the foundation model methods leverage these insights to develop more capable and generalizable manipulation systems.

9.3 Limitations

Despite the significant contributions of this thesis, several limitations remain:

1. **Generalization:** The foundation model methods demonstrate promising generalization capabilities, but they are still limited to specific task categories and environments. Extending these methods to more diverse tasks and environments remains an open challenge.
2. **Tactile sensor limitations:** The tactile sensor used in TacVLA has limited spatial resolution, restricting fine-grained contact geometry reasoning. More advanced tactile sensors with higher resolution and richer contact information could further improve the performance of tactile-enhanced VLA models.
3. **Contact-aware gating:** The contact-aware gating mechanism in TacVLA relies on a binary threshold heuristic, which does not allow gradual or learnable adjustment of modality importance. More adaptive modality weighting strategies could further improve cross-modal coordination.

9.4 Future Research Directions

Based on the contributions and limitations of this thesis, we identify the following promising future research directions:

9.4.1 Scalable Safe Learning for Contact-Rich Tasks

Future work should focus on developing scalable safe learning methods that can handle the complexity and diversity of real-world contact-rich tasks. This includes developing more efficient exploration strategies, better safety constraints, and more robust safety mechanisms. In particular, the integration of formal safety guarantees with data-driven learning methods remains an open challenge.

9.4.2 Foundation Models with Physical Intelligence

Future work should focus on developing foundation models that are natively aware of physical constraints and safety requirements. This includes developing new architectures that can effectively integrate force, tactile, and proprioceptive sensing with visual and language understanding, and developing training methods that can leverage diverse contact-rich manipulation data.

9.4.3 Multimodal Contact Sensing

Future work should focus on developing richer multimodal contact sensing capabilities, including higher-resolution tactile sensors, force/torque sensing, and proprioceptive feedback. The integration of these diverse sensing modalities into a unified representation that can be effectively used by foundation models remains an open challenge.

9.4.4 Safe Deployment in Human-Robot Collaboration

Future work should focus on developing safe deployment strategies for contact-rich manipulation systems in human-robot collaboration settings. This includes developing methods for real-time safety monitoring, adaptive safety constraints, and human-aware manipulation strategies. The integration of human intent recognition and prediction into the safety framework is also an important direction.

9.4.5 Benchmarks and Evaluation

Future work should focus on developing standardized benchmarks and evaluation metrics for safe contact-rich manipulation. This includes developing simulation environments that accurately capture the dynamics of contact-rich tasks, and developing evaluation metrics that can quantify both task performance and safety.

9.5 Concluding Remarks

Viewed as a whole, the thesis shows that safe and effective contact-rich manipulation cannot be solved at a single level alone; instead, it requires a coordinated treatment of learning, control, physical interaction, and multimodal perception.

This thesis has presented a comprehensive research program on safe and effective learning for contact-rich robotic manipulation. The contributions span from safe reinforcement learning with variable impedance control to foundation model-based manipulation with multimodal sensing. The key insights are:

1. Safety in contact-rich manipulation requires a multi-faceted approach that addresses both exploration safety and execution safety.
2. Variable impedance control provides a natural interface between learning-based policies and physical robots, enabling safe and effective contact-rich manipulation.
3. Foundation models can be effectively extended with physical intelligence, including compliance control and tactile sensing, to enable robust contact-rich manipulation.
4. Contact-aware gating is a principled approach to integrating tactile feedback into transformer-based architectures, enabling state-dependent multimodal fusion.

The research presented in this thesis contributes to the broader goal of developing robots that can safely and effectively interact with the physical world, enabling a wide range of applications in manufacturing, healthcare, and human-robot collaboration. As robots become increasingly capable and autonomous, the importance of safe and effective contact-rich manipulation will only grow, making the research directions outlined in this thesis increasingly relevant and impactful.

References

- [1] T. Tsuji, Y. Kato, G. Solak, H. Zhang, T. Petrič, F. Nori, and A. Ajoudani, “A survey on imitation learning for contact-rich tasks in robotics,” *The International Journal of Robotics Research*, p. 02783649261417694, 2025.
- [2] B. Thananjeyan, A. Balakrishna, S. Nair, M. Luo, K. Srinivasan, M. Hwang, J. E. Gonzalez, J. Ibarz, C. Finn, and K. Goldberg, “Recovery rl: Safe reinforcement learning with learned recovery zones,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4915–4922, 2021.
- [3] A. Billard and D. Kragic, “Trends and challenges in robot manipulation,” *Science*, vol. 364, no. 6446, p. eaat8414, 2019.
- [4] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kosuge, and O. Khatib, “Progress and prospects of the human–robot collaboration,” *Autonomous robots*, vol. 42, pp. 957–975, 2018.
- [5] Í. Elguea-Aguinaco, A. Serrano-Muñoz, D. Chrysostomou, I. Inziarte-Hidalgo, S. Bøgh, and N. Arana-Arexolaleiba, “A review on reinforcement learning for contact-rich robotic manipulation tasks,” *Robotics and Computer-Integrated Manufacturing*, vol. 81, p. 102517, 2023.
- [6] R. Cruz-Oliver, “Learning for contact-rich tasks with cobots,” Master’s thesis, Master’s thesis, 2024.
- [7] O. Khatib, “A unified approach for motion and force control of robot manipulators: The operational space formulation,” *IEEE Journal on Robotics and Automation*, pp. 43–53, 1987.
- [8] M. H. Raibert and J. J. Craig, “Hybrid position/force control of manipulators,” *Journal of Dynamic Systems, Measurement, and Control*, vol. 103, no. 2, pp. 126–133, 1981.
- [9] H. Neville, “Impedance control: An approach to manipulation: Part i-iii,” *Trans. of the ASME Journal of Dynamic System, Measurement, and Control*, p. 1, 1985.
- [10] D. E. Whitney, “Quasi-static assembly of compliantly supported rigid parts,” *Journal of Dynamic Systems, Measurement, and Control*, vol. 104, no. 1, pp. 65–77, 1982.
- [11] M. T. Mason, “Mechanics and planning of manipulator pushing operations,” *The International Journal of Robotics Research*, vol. 5, no. 3, pp. 53–71, 1986.

- [12] M. Suomalainen, Y. Karayiannidis, and V. Kyrki, “A survey of robot manipulation in contact,” *Robotics and Autonomous Systems*, vol. 156, p. 104224, 2022.
- [13] H. Zhang, G. Solak, S. Hjorth, and A. Ajoudani, “Towards passive safe reinforcement learning: A comparative study on contact-rich robotic manipulation,” *arXiv preprint arXiv:2503.00287*, 2025.
- [14] A. Calanca, R. Muradore, and P. Fiorini, “A review of algorithms for compliant control of stiff and fixed-compliance robots,” *IEEE/ASME transactions on mechatronics*, vol. 21, pp. 613–624, 2015.
- [15] Z. Yuan, A. W. Hall, S. Zhou, L. Brunke, M. Greeff, J. Panerati, and A. P. Schoellig, “Safe-control-gym: A unified benchmark suite for safe learning-based control and reinforcement learning in robotics,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 142–11 149, 2022.
- [16] B. Zhang, Y. Zhang, J. Ji, Y. Lei, J. Dai, Y. Chen, and Y. Yang, “Safevla: Towards safety alignment of vision-language-action model via constrained learning,” *arXiv preprint arXiv:2503.03480*, 2025.
- [17] S. Bahl, M. Mukadam, A. Gupta, and D. Pathak, “Neural dynamic policies for end-to-end sensorimotor learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 5058–5069, 2020.
- [18] M. Tölle, T. Gruner, D. Palenicek, J. Günster, P. Liu, J. Watson, D. Tateo, and J. Peters, “Towards safe robot foundation models,” *arXiv preprint arXiv:2503.07404*, 2025.
- [19] H. Geng, F. Wang, S. Wei, Y. Li, B. Wang, B. An, C. T. Cheng, H. Lou, P. Li, Y.-J. Wang *et al.*, “Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning,” *arXiv preprint arXiv:2504.18904*, 2025.
- [20] S. Gu, L. Shi, M. Wen, M. Jin, E. Mazumdar, Y. Chi, A. Wierman, and C. Spanos, “Robust gymnasium: A unified modular benchmark for robust reinforcement learning,” *arXiv preprint arXiv:2502.19652*, 2025.
- [21] H. Zhang, G. Solak, G. J. G. Lahr, and A. Ajoudani, “Srl-vic: A variable stiffness-based safe reinforcement learning for contact-rich robotic tasks,” *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5631–5638, 2024.
- [22] L. Roveda, J. Maskani, P. Franceschi, A. Abdi, F. Braghin, L. Molinari Tosatti, and N. Pedrocchi, “Model-based reinforcement learning variable impedance control for human-robot collaboration,” *Journal of Intelligent & Robotic Systems*, vol. 100, pp. 417–433, 2020.
- [23] L. Stan, A. F. Nicolescu, and C. Pupăză, “Reinforcement learning for assembly robots: A review,” *Proceedings in Manufacturing Systems*, vol. 15, pp. 135–146, 2020.
- [24] H. Zhang, G. Solak, S. Hjorth, and A. Ajoudani, “Passivity-centric safe reinforcement learning for contact-rich robotic tasks,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.00287>

- [25] N. Lin, L. Zhang, Y. Chen, Y. Zhu, R. Chen, P. Wu, and X. Chen, “Reinforcement learning for robotic safe control with force sensing,” in *2019 WRC Symposium on Advanced Robotics and Automation (WRC SARA)*, 2019, pp. 148–153.
- [26] A. Li, R. Liu, X. Yang, and Y. Lou, *Reinforcement Learning Strategy Based on Multimodal Representations for High-Precision Assembly Tasks*. Springer International Publishing, 2021, p. 56–66. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-89095-7_6
- [27] J. Luo, P. Dong, Y. Zhai, Y. Ma, and S. Levine, “Rlif: Interactive imitation learning as reinforcement learning,” *arXiv preprint arXiv:2311.12996*, 2023.
- [28] Y. Wang, C. C. Beltran-Hernandez, W. Wan, and K. Harada, “Robotic imitation of human assembly skills using hybrid trajectory and force learning,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 11 278–11 284.
- [29] Q. K. Luu, P. Zhou, Z. Xu, Z. Zhang, Q. Qiu, and Y. She, “Manifeel: Benchmarking and understanding visuotactile manipulation policy learning,” 2025. [Online]. Available: <https://arxiv.org/abs/2505.18472>
- [30] A. S. Anand, J. T. Gravdahl, and F. J. Abu-Dakka, “Model-based variable impedance learning control for robotic manipulation,” *Robotics and Autonomous Systems*, vol. 170, p. 104531, December 2023. [Online]. Available: <http://dx.doi.org/10.1016/j.robot.2023.104531>
- [31] H. Liang, X. Wang, J. Mao, and J. Yang, “Control barrier function-based force constrained safety compliance control for manipulator,” in *2023 28th International Conference on Automation and Computing (ICAC)*. IEEE, August 2023, p. 1–7. [Online]. Available: <http://dx.doi.org/10.1109/icac57885.2023.10275266>
- [32] H. Zhou, S. Ma, G. Wang, Y. Deng, and Z. Liu, “A hybrid control strategy for grinding and polishing robot based on adaptive impedance control,” *Advances in Mechanical Engineering*, vol. 13, no. 3, March 2021. [Online]. Available: <http://dx.doi.org/10.1177/16878140211004034>
- [33] T. Pang, *Planning, sensing, and control for contact-rich robotic manipulation with quasi-static contact models*. Massachusetts Institute of Technology, 2023.
- [34] Y. Xu, Z. Liu, G. Duan, J. Zhu, X. Bai, and J. Tan, “Look Before You Leap: Safe Model-Based Reinforcement Learning with Human Intervention,” in *Inflammatory Bowel Diseases*, June 2021. [Online]. Available: <https://openreview.net/forum?id=oqZrUx-PRqb>
- [35] F. Xu, G. Zhai, X. Kong, T. Fu, D. F. Gordon, X. An, and B. Busam, “Stare-vla: Progressive stage-aware reinforcement for fine-tuning vision-language-action models,” *arXiv preprint arXiv:2512.05107*, 2025.
- [36] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, M. Goulão, A. Kallinteris, M. Krimmel, A. KG *et al.*, “Gymnasium: A standard interface for reinforcement learning environments,” *arXiv preprint arXiv:2407.17032*, 2024.

- [37] J. Ji, B. Zhang, J. Zhou, X. Pan, W. Huang, R. Sun, Y. Geng, Y. Zhong, J. Dai, and Y. Yang, “Safety gymnasium: A unified safe reinforcement learning benchmark,” *Advances in Neural Information Processing Systems* 36, pp. 18 964–18 993, 2023.
- [38] Z. Zhao, S. Cheng, Y. Ding, Z. Zhou, S. Zhang, D. Xu, and Y. Zhao, “A survey of optimization-based task and motion planning: From classical to learning approaches,” *IEEE/ASME Transactions on Mechatronics*, pp. 1–27, 2025.
- [39] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, “Stable-baselines3: Reliable reinforcement learning implementations,” *Journal of machine learning research*, vol. 22, no. 268, 2021.
- [40] W. Chi and M. Q.-H. Meng, “Risk-rrt*: A robot motion planning algorithm for the human robot coexisting environment,” in *Proc. 18th Int. Conf. Adv. Robot. (ICAR)*. IEEE, 2017, pp. 583–588.
- [41] D. Rodríguez-Guerra, A. Mosca, A. Valente, I. Cabanes, and E. Carpanzano, “An advanced dual apf-based controller for efficient simultaneous collision and singularity avoidance for human-robot collaborative assembly processes,” *CIRP Annals*, vol. 72, no. 1, pp. 5–8, 2023.
- [42] S. A. Bowyer, B. L. Davies, and F. R. y Baena, “Active constraints/virtual fixtures: A survey,” *IEEE Trans. Robot.*, vol. 30, pp. 138–157, 2013.
- [43] K.-C. Hsu, H. Hu, and J. F. Fisac, “The safety filter: A unified view of safety-critical control in autonomous systems,” *Annu. Rev. Control Robot. Auton. Syst.*, vol. 7, 2024.
- [44] F. Ferraguti, C. T. Landi, S. Costi, M. Bonfè, S. Farsoni, C. Secchi, and C. Fantuzzi, “Safety barrier functions and multi-camera tracking for human–robot shared environment,” *Rob. Auton. Syst.*, vol. 124, p. 103388, 2020.
- [45] B. Dai, R. Khorrambakht, P. Krishnamurthy, V. Goncalves, A. Tzes, and F. Khorrami, “Safe navigation and obstacle avoidance using differentiable optimization based control barrier functions,” *IEEE Robot. Autom. Lett.*, 2023.
- [46] E. Das and J. W. Burdick, “Robust control barrier functions using uncertainty estimation with application to mobile robots,” *arXiv preprint arXiv:2401.01881*, 2024.
- [47] J. Li, Q. Liu, W. Jin, J. Qin, and S. Hirche, “Robust safe learning and control in an unknown environment: An uncertainty-separated control barrier function approach,” *IEEE Robot. Autom. Lett.*, 2023.
- [48] D. Wang and J. Zhang, “Flow shop scheduling with human–robot collaboration: A joint chance-constrained programming approach,” *Int. J. Prod. Res.*, vol. 62, no. 4, pp. 1297–1317, 2024.
- [49] N. E. D. Toit and J. W. Burdick, “Probabilistic collision checking with chance constraints,” *IEEE Trans. Robot.*, vol. 27, no. 4, pp. 809–815, 2011.
- [50] S. Ahmed and A. Shapiro, “Solving chance-constrained stochastic programs via sampling and integer programming,” in *State-of-the-Art Decision-Making Tools in the Information-Intensive Age*. INFORMS, 2008, pp. 261–269.

- [51] Y. S. Shao, C. Chen, S. Kousik, and R. Vasudevan, “Reachability-based trajectory safeguard (rts): A safe and fast reinforcement learning safety layer for continuous control,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3663–3670, 2021.
- [52] A. Banerjee, K. Rahmani, J. Biswas, and I. Dillig, “Dynamic model predictive shielding for provably safe reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 100 131–100 159, 2024.
- [53] J. Liu, J. Yang, J. Mao, T. Zhu, Q. Xie, Y. Li, X. Wang, and S. Li, “Flexible active safety motion control for robotic obstacle avoidance: A cbf-guided mpc approach,” *IEEE Robotics and Automation Letters*, 2025.
- [54] A. Agnihotri, R. Jain, D. Ramachandran, and S. Singla, “e-cop: Episodic constrained optimization of policies,” *Expert Systems with Applications*, pp. 110 909–110 934, 2024.
- [55] Y. Ding and J. Lavaei, “Provably efficient primal-dual reinforcement learning for cmdps with non-stationary objectives and constraints,” in *Advances in Neural Information Processing Systems*, 2023, pp. 7396–7404.
- [56] Y. Wang, S. S. Zhan, R. Jiao, Z. Wang, W. Jin, Z. Yang, Z. Wang, C. Huang, and Q. Zhu, “Enforcing hard constraints with soft barriers: Safe reinforcement learning in unknown stochastic environments,” in *Advances in Neural Information Processing Systems*, 2023, pp. 36 593–36 604.
- [57] H. Yu, W. Xu, and H. Zhang, “Towards safe reinforcement learning with a safety editor policy,” *Engineering Applications of Artificial Intelligence*, vol. 115, pp. 2608–2621, 2022.
- [58] P.-F. Massiani, S. Heim, F. Solowjow, and S. Trimpe, “Safe value functions,” *IEEE Transactions on Automatic Control*, vol. 68, no. 5, p. 2743–2757, May 2023. [Online]. Available: <http://dx.doi.org/10.1109/tac.2022.3200948>
- [59] D. C. Tan, R. McCarthy, F. Acero, A. M. Delfaki, Z. Li, and D. Kanoulas, “Safe value functions: Learned critics as hard safety constraints,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 2441–2448.
- [60] J. Günster, P. Liu, J. Peters, and D. Tateo, “Handling long-term safety and uncertainty in safe reinforcement learning,” *arXiv preprint arXiv:2409.12045*, 2024.
- [61] X. Li, J. Xiao, Y. Cheng, and H. Liu, “An actor-critic learning framework based on lyapunov stability for automatic assembly,” *Applied Intelligence*, vol. 53, pp. 4801–4812, 2023.
- [62] S. Kim, J. Kwon, T. Lee, Y. Park, and J. Perez, “Safety-aware unsupervised skill discovery,” in *ICRA*, 2023.
- [63] Q. Yang, A. Dürr, E. A. Topp, J. A. Stork, and T. Stoyanov, “Variable impedance skill learning for contact-rich manipulation,” *IEEE robotics and automation letters*, vol. 7, no. 3, pp. 8391–8398, 2022.

- [64] Y. Li, Y. Wang, Z. Li, L. Yingxiang, J. Chai, and E. Dong, “Deep reinforcement learning-based variable impedance control for grinding workpieces with complex geometry,” *Robotic Intelligence and Automation*, 2025.
- [65] M. Xiao, T. Zhang, Y. Zou, S. Chen, and W. Wu, “Research on robot massage force control based on residual reinforcement learning,” *IEEE Access*, vol. 12, pp. 18 270–18 279, 2024.
- [66] Y. Li, L. Zheng, Y. Wang, E. Dong, and S. Zhang, “Impedance learning-based adaptive force tracking for robot on unknown terrains,” *IEEE Transactions on Robotics*, vol. 41, pp. 1404–1420, 2025.
- [67] Z. Li, C. Zeng, Z. Deng, Q. Xu, B. He, and J. Zhang, “Learning variable impedance control for robotic massage with deep reinforcement learning: A novel learning framework,” *IEEE Systems, Man, and Cybernetics Magazine*, vol. 10, no. 1, pp. 17–27, 2024.
- [68] Q. Yang, A. Dürr, E. A. Topp, J. A. Stork, and T. Stoyanov, “Learning impedance actions for safe reinforcement learning in contact-rich tasks,” in *NeurIPS 2021 Workshop on Deployable Decision Making in Embodied Systems (DDM)*, (Online conference), Sydney, Australia, December 6-14, 2021, 2021.
- [69] Z. Ji, G. Liu, W. Xu, B. Yao, X. Liu, and Z. Zhou, “Deep reinforcement learning on variable stiffness compliant control for programming-free robotic assembly in smart manufacturing,” *International Journal of Production Research*, vol. 62, no. 19, p. 7073–7095, February 2024. [Online]. Available: <http://dx.doi.org/10.1080/00207543.2024.2318488>
- [70] X. Feng, T. Shi, W. Li, P. Lu, and Y. Pan, “Reinforcement learning-based impedance learning for robot admittance control in industrial assembly,” in *2022 International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, July 2022, p. 1092–1097. [Online]. Available: <http://dx.doi.org/10.1109/icarm54641.2022.9959152>
- [71] M. Dalle Vedove, E. Lamon, D. Fontanelli, L. Palopoli, and M. Saveriano, “A passivity-based variable impedance controller for incremental learning of periodic interactive tasks,” in *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, 2024, pp. 2604–2609.
- [72] K. Karacan, H. Sadeghian, R. Kirschner, and S. Haddadin, “Passivity-based skill motion learning in stiffness-adaptive unified force-impedance control,” in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 9604–9611.
- [73] R. Zanella, G. Palli, S. Stramigioli, and F. Califano, “Learning passive policies with virtual energy tanks in robotics,” *IET Control Theory & Applications*, vol. 18, no. 5, pp. 541–550, 2024.
- [74] M. Ni, L. Zhang, Z. Chen, K. Bai, Z. Chen, J. Zhang, and W. Zuo, “Don’t let your robot be harmful: Responsible robotic manipulation via safety-as-policy,” *arXiv preprint arXiv:2411.18289*, 2024.

- [75] J. Thumm, C. Agia, M. Pavone, and M. Althoff, “Text2interaction: Establishing safe and preferable human-robot interaction,” in *Proceedings of The 8th Conference on Robot Learning*, 2025, pp. 1250–1267.
- [76] H. Zhang, W.-H. Huang, G. Solak, and A. Ajoudani, “Omnivic: A self-improving variable impedance controller with vision-language in-context learning for safe robotic manipulation,” *arXiv preprint arXiv:2510.17150*, 2025.
- [77] F. Batool, Y. Yaqoot, M. Zafar, R. A. Khan, M. H. Khan, A. Fedoseev, and D. Tsetserukou, “Impedancegpt: Vlm-driven impedance control of swarm of mini-drones for intelligent navigation in dynamic environment,” *arXiv preprint arXiv:2503.02723*, 2025.
- [78] S. Yin, X. Pang, Y. Ding, M. Chen, Y. Bi, Y. Xiong, W. Huang, Z. Xiang, J. Shao, and S. Chen, “Safeagentbench: A benchmark for safe task planning of embodied llm agents,” *arXiv preprint arXiv:2412.13178*, 2024.
- [79] Y. Wi, M. Van der Merwe, P. Florence, A. Zeng, and N. Fazeli, “Calamari: Contact-aware and language conditioned spatial action mapping for contact-rich manipulation,” in *7th Annual Conference on Robot Learning*, 2023.
- [80] A. Macaluso, N. Cote, and S. Chitta, “Toward automated programming for robotic assembly using chatgpt,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2024, p. 17687–17693. [Online]. Available: <http://dx.doi.org/10.1109/icra57147.2024.10610554>
- [81] Z. Feng, H. Luan, P. Goyal, and H. Soh, “Ltlldog: Satisfying temporally-extended symbolic constraints for safe diffusion-based planning,” *IEEE Robotics and Automation Letters*, vol. 9, no. 10, p. 8571–8578, October 2024. [Online]. Available: <http://dx.doi.org/10.1109/lra.2024.3443501>
- [82] M. Okada, M. Komatsu, and T. Taniguchi, “A contact model based on denoising diffusion to learn variable impedance control for contact-rich manipulation,” in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7286–7293.
- [83] Y. Hou, Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, S. Feng, B. Burchfiel, and S. Song, “Adaptive compliance policy: Learning approximate compliance for diffusion guided control,” *arXiv preprint arXiv:2410.09309*, 2024.
- [84] B. Abbatematteo, E. Rosen, S. Thompson, T. Akbulut, S. Rammohan, and G. Konidaris, “Composable interaction primitives: A structured policy class for efficiently learning sustained-contact manipulation skills,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 7522–7529.
- [85] M. Aburub, C. C. Beltran-Hernandez, T. Kamijo, and M. Hamaya, “Learning diffusion policies from demonstrations for compliant contact-rich manipulation,” *arXiv preprint arXiv:2410.19235*, 2024.
- [86] E. Cramer, L. Jäschke, and S. Trimpe, “Cheq-ing the box: Safe variable impedance learning for robotic polishing,” *arXiv preprint arXiv:2501.07985*, 2025.

- [87] Y. Zang, Z. Wang, M. Pan, Z. Hou, Z. Ding, and M. Zhao, “Safe peg-in-hole automatic assembly using virtual guiding force: A deep reinforcement learning solution,” *Robotics and Autonomous Systems*, vol. 185, p. 104894, 2025.
- [88] R. Martin-Martin, M. A. Lee, R. Gardner, S. Savarese, J. Bohg, and A. Garg, “Variable impedance control in end-effector space: An action space for reinforcement learning in contact-rich tasks,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1010–1017.
- [89] F. Vinter-Hviid, C. Sloth, T. R. Savarimuthu, and I. Iturrate, “Safe contact-based robot active search using bayesian optimization and control barrier functions,” *Frontiers in Robotics and AI*, vol. 11, p. 1344367, 2024.
- [90] D. Zhang, M. Van, S. McIlvanna, Y. Sun, and S. McLoone, “Adaptive safety-critical control with uncertainty estimation for human–robot collaboration,” *IEEE Transactions on Automation Science and Engineering*, vol. 21, pp. 5983–5996, 2024.
- [91] C. Wang, X. Zhang, Z. Kuang, and M. Tomizuka, “Safe online gain optimization for cartesian space variable impedance control,” in *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, 2022, pp. 751–757.
- [92] I. Mitsioni, P. Tajvar, D. Kragic, J. Tumova, and C. Pek, “Safe data-driven model predictive control of systems with complex dynamics,” *IEEE Transactions on Robotics*, vol. 39, pp. 3242–3258, 2023.
- [93] R. Soni, D. Harnack, H. Isermann, S. Fushimi, S. Kumar, and F. Kirchner, “End-to-end reinforcement learning for torque based variable height hopping,” in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7531–7538.
- [94] M. Noseworthy, B. Tang, B. Wen, A. Handa, C. Kessens, N. Roy, D. Fox, F. Ramos, Y. Narang, and I. Akinola, “Forge: Force-guided exploration for robust contact-rich manipulation under uncertainty,” *arXiv preprint arXiv:2408.04587*, 2024.
- [95] M. Qu, Y. Wang, and D. T. Pham, “Robotic disassembly task training and skill transfer using reinforcement learning,” *IEEE Transactions on Industrial Informatics*, pp. 10 934–10 943, 2023.
- [96] A. Serrano-Munoz, N. Arana-Arexolaleiba, D. Chrysostomou, and S. Bøgh, “Learning and generalising object extraction skill for contact-rich disassembly tasks: an introductory study,” *The International Journal of Advanced Manufacturing Technology*, vol. 124, no. 9, pp. 3171–3183, 2023.
- [97] J. Wang, X. Zhu, H. Wang, X. Wang, and H. Ding, “Hybrid trajectory and force learning of complex assembly tasks: A combined learning framework,” *IEEE Access*, pp. 60 175–60 186, 2021.
- [98] A. Padalkar, G. Quere, A. Raffin, J. Silvério, and F. Stulp, “Guiding real-world reinforcement learning for in-contact manipulation tasks with shared control templates,” *Autonomous Robots*, vol. 48, no. 4-5, p. 12, 2024.

- [99] C. C. Beltran-Hernandez, D. Petit, I. G. Ramirez-Alpizar, T. Nishi, S. Kikuchi, T. Matsubara, and K. Harada, “Learning force control for contact-rich manipulation tasks with rigid position-controlled robots,” *IEEE Robotics and Automation Letters*, vol. 5, pp. 5709–5716, 2020.
- [100] X. Zhao, H. Zhao, P. Chen, and H. Ding, “Model accelerated reinforcement learning for high precision robotic assembly,” *International Journal of Intelligent Robotics and Applications*, vol. 4, no. 2, p. 202–216, June 2020. [Online]. Available: <http://dx.doi.org/10.1007/s41315-020-00138-z>
- [101] H. Zhu, T. Zhao, X. Ni, J. Wang, K. Fang, L. Righetti, and T. Pang, “Should we learn contact-rich manipulation policies from sampling-based planners?” *IEEE Robotics and Automation Letters*, vol. 10, no. 6, p. 6248–6255, June 2025. [Online]. Available: <http://dx.doi.org/10.1109/lra.2025.3564701>
- [102] C. Zhu, T. Yu, and Q. Chang, “Task-oriented safety field for robot control in human-robot collaborative assembly based on residual learning,” *Expert Systems with Applications*, vol. 238, p. 121946, 2024.
- [103] Z. Erickson, V. Gangaram, A. Kapusta, C. K. Liu, and C. C. Kemp, “Assistive gym: A physics simulation framework for assistive robotics,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2020, p. 10169–10176. [Online]. Available: <http://dx.doi.org/10.1109/icra40945.2020.9197411>
- [104] X. Wang, J. Yang, J. Mao, J. Liang, S. Li, and Y. Yan, “Guarding force: Safety-critical compliant control for robot-environment interaction,” *arXiv preprint arXiv:2405.04859*, 2024.
- [105] P. De Risi, F. Amadio, G. Garofalo, F. Ficuciello, and P. Falco, “Variable impedance control combining reinforcement learning and gaussian process regression,” in *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*, 2024, pp. 343–350.
- [106] Y. Pan, Z. Li, T. Shi, and C. Wen, “Composite learning variable impedance robot control with stability and passivity guarantees,” *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 119–126, 2023.
- [107] Z. Yao, F. Xu, G.-P. Jiang, and J. Yao, “Data-driven control of hydraulic manipulators by reinforcement learning,” *IEEE/ASME Transactions on Mechatronics*, vol. 29, no. 4, p. 2673–2684, August 2024. [Online]. Available: <http://dx.doi.org/10.1109/tmech.2023.3336070>
- [108] F. Wirmshofer, P. S. Schmitt, G. von Wichert, and W. Burgard, “Controlling contact-rich manipulation under partial observability,” in *Robotics: Science and Systems XVI*, 2020.
- [109] M. Bogdanovic, M. Khadiv, and L. Righetti, “Learning variable impedance control for contact sensitive tasks,” *IEEE Robotics and Automation Letters*, vol. 5, no. 4, p. 6129–6136, October 2020. [Online]. Available: <http://dx.doi.org/10.1109/lra.2020.3011379>

- [110] J. Yamada, Y. Lee, G. Salhotra, K. Pertsch, M. Pflueger, G. Sukhatme, J. Lim, and P. Englert, “Motion planner augmented reinforcement learning for robot manipulation in obstructed environments,” in *Conference on Robot Learning*. PMLR, 2021, pp. 589–603.
- [111] X. Zhu, S. Kang, and J. Chen, “A contact-safe reinforcement learning framework for contact-rich robot manipulation,” in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 2476–2482.
- [112] H. Wang, J. Peng, F. Zhang, H. Zhang, and Y. Wang, “High-order control barrier functions-based impedance control of a robotic manipulator with time-varying output constraints,” *ISA transactions*, vol. 129, pp. 361–369, 2022.
- [113] X. Yu, S. Liu, S. Zhang, W. He, and H. Huang, “Adaptive neural network force tracking control of flexible joint robot with an uncertain environment,” *IEEE Transactions on Industrial Electronics*, vol. 71, pp. 5941–5949, 2024.
- [114] Y. Michel, R. Rahal, C. Pacchierotti, P. R. Giordano, and D. Lee, “Bilateral teleoperation with adaptive impedance control for contact tasks,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5429–5436, 2021.
- [115] M. Sharifi, V. Azimi, V. K. Mushahwar, and M. Tavakoli, “Impedance learning-based adaptive control for human–robot interaction,” *IEEE Transactions on Control Systems Technology*, vol. 30, pp. 1345–1358, 2022.
- [116] Y. Li and D. Xu, “Skill learning for robotic insertion based on one-shot demonstration and reinforcement learning,” *International Journal of Automation and Computing*, pp. 457–467, 2021.
- [117] S. A. Khader, H. Yin, P. Falco, and D. Kragic, “Stability-guaranteed reinforcement learning for contact-rich manipulation,” *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 1–8, 2021.
- [118] C. Yan, J. Wu, and Q. Zhu, “Learning-based contact status recognition for peg-in-hole assembly,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 6003–6009.
- [119] M. A. Murtaza, S. Aguilera, M. Waqas, and S. Hutchinson, “Safety compliant control for robotic manipulator with task and input constraints,” *IEEE Robotics and Automation Letters*, pp. 10 659–10 664, 2022.
- [120] B. Chen, D. Martí Monsó, Y. Du, M. Simchowitz, R. Tedrake, and V. Sitzmann, “Diffusion forcing: Next-token prediction meets full-sequence diffusion,” *Advances in Neural Information Processing Systems* 37, pp. 24 081–24 125, 2024.
- [121] M. Kim, S. Niekum, and A. D. Deshpande, “Scape: Learning stiffness control from augmented position control experiences,” in *Proceedings of the 5th Conference on Robot Learning*, 2022, pp. 1512–1521.
- [122] G. Peng, C. P. Chen, and C. Yang, “Neural networks enhanced optimal admittance control of robot–environment interaction using reinforcement learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, pp. 4551–4561, 2022.

- [123] R. Soni, D. Harnack, H. Isermann, S. Fushimi, S. Kumar, and F. Kirchner, “End-to-end reinforcement learning for torque based variable height hopping,” in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7531–7538.
- [124] B. Thananjeyan, A. Balakrishna, U. Rosolia, F. Li, R. McAllister, J. E. Gonzalez, S. Levine, F. Borrelli, and K. Goldberg, “Safety augmented value estimation from demonstrations (saved): Safe deep model-based rl for sparse cost robotic tasks,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, p. 3612–3619, April 2020. [Online]. Available: <http://dx.doi.org/10.1109/lra.2020.2976272>
- [125] Z. Sun, Y. Wang, D. Held, and Z. Erickson, “Force-constrained visual policy: Safe robot-assisted dressing via multi-modal sensing,” *IEEE Robotics and Automation Letters*, 2024.
- [126] X. Xing, E. Burdet, W. Si, C. Yang, and Y. Li, “Impedance learning for human-guided robots in contact with unknown environments,” *IEEE Transactions on Robotics*, vol. 39, pp. 3705–3721, 2023.
- [127] Q. Liu, Z. Ji, W. Xu, Z. Liu, B. Yao, and Z. Zhou, “Knowledge-guided robot learning on compliance control for robotic assembly task with predictive model,” *Expert Systems with Applications*, vol. 234, p. 121037, 2023.
- [128] Z. Hou, Z. Li, C. Hsu, K. Zhang, and J. Xu, “Fuzzy logic-driven variable time-scale prediction-based reinforcement learning for robotic multiple peg-in-hole assembly,” *IEEE Transactions on Automation Science and Engineering*, pp. 218–229, 2022.
- [129] G. Kang, H. Seong, D. Lee, and D. H. Shim, “A versatile door opening system with mobile manipulator through adaptive position-force control and reinforcement learning,” *Robotics and Autonomous Systems*, vol. 180, p. 104760, 2024.
- [130] X. Zhang, M. Tomizuka, and H. Li, “Bridging the sim-to-real gap with dynamic compliance tuning for industrial insertion,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2024, p. 4356–4363. [Online]. Available: <http://dx.doi.org/10.1109/icra57147.2024.10610707>
- [131] S. Jadav and H. J. Palanthandalam-Madapusi, “Configuration and force-field aware variable impedance control with faster re-learning,” *Journal of Intelligent ; Robotic Systems*, vol. 110, no. 1, December 2023. [Online]. Available: <http://dx.doi.org/10.1007/s10846-023-02022-x>
- [132] A. S. Anand, R. Kaushik, J. T. Gravdahl, and F. J. Abu-Dakka, “Data-efficient reinforcement learning for variable impedance control,” *IEEE Access*, vol. 12, p. 15631–15641, 2024. [Online]. Available: <http://dx.doi.org/10.1109/access.2024.3355311>
- [133] Y. Gai, B. Wang, J. Zhang, D. Wu, and K. Chen, “Robotic assembly control reconfiguration based on transfer reinforcement learning for objects with different geometric features,” *Engineering Applications of Artificial Intelligence*, vol. 129, p. 107576, March 2024. [Online]. Available: <http://dx.doi.org/10.1016/j.engappai.2023.107576>

- [134] C. Chen, C. Zhang, and Y. Pan, “Active compliance control of robot peg-in-hole assembly based on combined reinforcement learning,” *Applied Intelligence*, vol. 53, no. 24, p. 30677–30690, November 2023. [Online]. Available: <http://dx.doi.org/10.1007/s10489-023-05156-5>
- [135] A. Giammarino, J. M. Gandarias, and A. Ajoudani, “An open tele-impedance framework to generate data for contact-rich tasks in robotic manipulation,” in *2023 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. IEEE, June 2023, p. 140–146. [Online]. Available: <http://dx.doi.org/10.1109/arso56563.2023.10187436>
- [136] Y. Ding, J. Zhao, and X. Min, “Impedance control and parameter optimization of surface polishing robot based on reinforcement learning,” *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 237, no. 1–2, p. 216–228, May 2022. [Online]. Available: <http://dx.doi.org/10.1177/09544054221100004>
- [137] P. Xiang, J. Zhang, D. Sun, K. Qiu, Q. Fang, X. Mi, Y. Wang, R. Xiong, and H. Lu, “Learning-based high-precision force estimation and compliant control for small-scale continuum robot,” *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 4, p. 5389–5401, October 2024. [Online]. Available: <http://dx.doi.org/10.1109/tase.2023.3311179>
- [138] H. Wu, Q. Hu, Y. Shi, J. Zheng, K. Sun, and J. Wang, “Space manipulator optimal impedance control using integral reinforcement learning,” *Aerospace Science and Technology*, vol. 139, p. 108388, August 2023. [Online]. Available: <http://dx.doi.org/10.1016/j.ast.2023.108388>
- [139] J. Li, D. Pang, Y. Zheng, X. Guan, and X. Le, “A flexible manufacturing assembly system with deep reinforcement learning,” *Control Engineering Practice*, vol. 118, p. 104957, 2022.
- [140] Y. Wang, C. C. Beltran-Hernandez, W. Wan, and K. Harada, “An adaptive imitation learning framework for robotic complex contact-rich insertion tasks,” *Frontiers in Robotics and AI*, vol. 8, January 2022. [Online]. Available: <http://dx.doi.org/10.3389/frobt.2021.777363>
- [141] A. S. Anand, M. Hagen Myrestrand, and J. T. Gravdahl, “Evaluation of variable impedance- and hybrid force/motion controllers for learning force tracking skills,” in *2022 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, January 2022, p. 83–89. [Online]. Available: <http://dx.doi.org/10.1109/sii52469.2022.9708826>
- [142] J. Xu, Y. Lin, H. Zhou, and H. Min, “Generating manipulation sequences using reinforcement learning and behavior trees for peg-in-hole task,” in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, October 2022, p. 2715–2720. [Online]. Available: <http://dx.doi.org/10.1109/smc53654.2022.9945269>
- [143] X. Li, J. Xiao, W. Zhao, H. Liu, and G. Wang, “Multiple peg-in-hole compliant assembly based on a learning-accelerated deep deterministic policy gradient strategy,”

- Industrial Robot: the international journal of robotics research and application*, vol. 49, pp. 54–64, 2022.
- [144] T. Zhang, C. Yuan, and Y. Zou, “Online optimization method of controller parameters for robot constant force grinding based on deep reinforcement learning rainbow,” *Journal of Intelligent & Robotic Systems*, vol. 105, p. 85, 2022.
- [145] G. Liu, B. Yao, W. Xu, and X. Liu, “Optimizing non-diagonal stiffness matrix of compliance control for robotic assembly using deep reinforcement learning,” *Journal of Physics: Conference Series*, vol. 2402, no. 1, p. 012013, December 2022. [Online]. Available: <http://dx.doi.org/10.1088/1742-6596/2402/1/012013>
- [146] L. Roveda, A. Testa, A. A. Shahid, F. Braghin, and D. Piga, “Q-learning-based model predictive variable impedance control for physical human-robot collaboration,” *Artificial Intelligence*, vol. 312, p. 103771, 2022.
- [147] C. C. Beltran-Hernandez, “Reinforcement learning for contact-rich assembly,” Ph.D. dissertation, The University of Osaka, 2021.
- [148] S. Li, X. Yuan, and J. Niu, “Robotic peg-in-hole assembly strategy research based on reinforcement learning algorithm,” *Applied Sciences*, vol. 12, no. 21, p. 11149, November 2022. [Online]. Available: <http://dx.doi.org/10.3390/app122111149>
- [149] X. Zhao, S. Han, B. Tao, Z. Yin, and H. Ding, “Model-based actor-critic learning of robotic impedance control in complex interactive environment,” *IEEE Transactions on Industrial Electronics*, vol. 69, pp. 13 225–13 235, 2022.
- [150] C.-Y. Kuo, A. Schaarschmidt, Y. Cui, T. Asfour, and T. Matsubara, “Uncertainty-aware contact-safe model-based reinforcement learning,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3918–3925, 2021.
- [151] Z. Hou, J. Fei, Y. Deng, and J. Xu, “Data-efficient hierarchical reinforcement learning for robotic assembly control applications,” *IEEE Transactions on Industrial Electronics*, vol. 68, no. 11, pp. 11 565–11 575, 2021.
- [152] C. C. Beltran-Hernandez, D. Petit, I. G. Ramirez-Alpizar, and K. Harada, “Variable compliance control for robotic peg-in-hole assembly: A deep-reinforcement-learning approach,” *Applied Sciences*, vol. 10, no. 19, p. 6923, 2020.
- [153] Y. Xu, X. Guo, G. Zhang, J. Li, X. Huo, B. Xuan, Z. Gu, and H. Sun, “A learning control strategy for robot-assisted bathing via impedance sliding mode technique with non-repetitive tasks,” *International Journal of Control, Automation and Systems*, vol. 22, pp. 946–962, 2024.
- [154] P. Lancaster, N. Hansen, A. Rajeswaran, and V. Kumar, “Modem-v2: Visuo-motor world models for real-world robot manipulation,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 7530–7537.
- [155] M. Mayr, *Learning with skill-based robot systems: Combining planning & knowledge representation with reinforcement learning*. Lund University, 2020.

- [156] C. Chuck, C. Qi, M. J. Munje, S. Li, M. Rudolph, C. Shi, S. Agarwal, H. Sikchi, A. Peri, S. Dayal *et al.*, “Robot air hockey: A manipulation testbed for robot learning with reinforcement learning,” *arXiv preprint arXiv:2405.03113*, 2024.
- [157] S. Zhan, Y. Wang, Q. Wu, R. Jiao, C. Huang, and Q. Zhu, “State-wise safe reinforcement learning with pixel observations,” in *6th Annual Learning for Dynamics & Control Conference*, 2023, pp. 1187–1201.
- [158] O. Spector and D. Di Castro, “Insertionnet-a scalable solution for insertion,” *IEEE Robotics and Automation Letters*, vol. 6, pp. 5509–5516, 2021.
- [159] O. Spector and M. Zacksenhouse, “Learning contact-rich assembly skills using residual admittance policy,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 6023–6030.
- [160] Y. Huo, P. Li, D. Chen, Y.-H. Liu, and X. Li, “Model-free adaptive impedance control for autonomous robotic sanding,” *IEEE Transactions on Automation Science and Engineering*, vol. 19, pp. 3601–3611, 2022.
- [161] H. Lin, R. Corcodel, and D. Zhao, “Generalize by touching: Tactile ensemble skill transfer for robotic furniture assembly,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 9227–9233.
- [162] W. Zhang, L. Ott, M. Tognon, and R. Siegwart, “Learning variable impedance control for aerial sliding on uneven heterogeneous surfaces by proprioceptive and tactile sensing,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, p. 11275–11282, October 2022. [Online]. Available: <http://dx.doi.org/10.1109/lra.2022.3194315>
- [163] M. Ulmer, E. Aljalbout, S. Schwarz, and S. Haddadin, “Learning robotic manipulation skills using an adaptive force-impedance action space,” *arXiv preprint arXiv:2110.09904*, 2021.
- [164] S. H. Huang, M. Zambelli, J. Kay, M. F. Martins, Y. Tassa, P. M. Pilarski, and R. Hadsell, “Learning gentle object manipulation with curiosity-driven deep reinforcement learning,” *arXiv preprint arXiv:1903.08542*, 2019.
- [165] S. Gu, A. Kshirsagar, Y. Du, G. Chen, J. Peters, and A. Knoll, “A human-centered safe robot reinforcement learning framework with interactive behaviors,” *Frontiers in Neurorobotics*, vol. 17, p. 1280341, 2023.
- [166] M. Farajtabar and M. Charbonneau, “The path towards contact-based physical human–robot interaction,” *Robotics and Autonomous Systems*, p. 104829, 2024.
- [167] X. Li, W. Shang, W. Rao, and S. Cong, “Hierarchical learning based on visual-haptic perception for robotic variable impedance control,” *IEEE Transactions on Industrial Electronics*, 2023.
- [168] K. Zhang, D.-G. Kim, E. T. Chang, H.-H. Liang, Z. He, K. Lampo, P. Wu, I. Kyminsis, and M. Ciocarlie, “Vibecheck: Using active acoustic tactile sensing for contact-rich manipulation,” *arXiv preprint arXiv:2504.15535*, 2025.

- [169] S. Gu, L. Shi, Y. Ding, A. Knoll, C. J. Spanos, A. Wierman, and M. Jin, “Enhancing efficiency of safe reinforcement learning via sample manipulation,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 17 247–17 285, 2024.
- [170] B. Huang, Y. Gong, Z. Yang, T. Ren, and L. Figueredo, “Contact-Safe Reinforcement Learning with ProMP Reparameterization and Energy Awareness,” 2025, arXiv:2511.13459 [cs] keywords = Computer Science - Robotics,. [Online]. Available: <http://arxiv.org/abs/2511.13459>
- [171] Y. Shi, Z. Chen, Y. Wu, D. Henkel, S. Riedel, H. Liu, Q. Feng, and J. Zhang, “Combining learning from demonstration with learning by exploration to facilitate contact-rich tasks,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 1062–1069.
- [172] Z. Zhao, S. Haldar, J. Cui, L. Pinto, and R. Bhirangi, “Touch begins where vision ends: Generalizable policies for contact-rich manipulation,” *arXiv preprint arXiv:2506.13762*, 2025.
- [173] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, “RT-2: Vision-language-action models transfer web knowledge to robotic control,” *arXiv preprint arXiv:2307.15818*, 2023.
- [174] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, “Openvla: An open-source vision-language-action model,” *arXiv preprint arXiv:2406.09246*, 2024.
- [175] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, “ π_0 : A vision-language-action flow model for general robot control,” *arXiv preprint arXiv:2410.24164*, 2024.
- [176] E. Spyrakos-Papastavridis, P. R. N. Childs, and J. S. Dai, “Passivity preservation for variable impedance control of compliant robots,” *IEEE/ASME Transactions on Mechatronics*, vol. 25, pp. 2342–2353, 2020.
- [177] S. Lyu and C. C. Cheah, “Human–robot interaction control based on a general energy shaping method,” *IEEE Transactions on Control Systems Technology*, vol. 28, no. 6, p. 2445–2460, November 2020. [Online]. Available: <http://dx.doi.org/10.1109/tcst.2019.2949525>
- [178] A.-N. Sharkawy and P. N. Koustoumpardis, “Human–robot interaction: A review and analysis on variable admittance control, safety, and perspectives,” *Machines*, vol. 10, no. 7, p. 591, July 2022. [Online]. Available: <http://dx.doi.org/10.3390/machines10070591>
- [179] F. J. Abu-Dakka and M. Saveriano, “Variable impedance control and learning—a review,” *Frontiers in Robotics and AI*, vol. 7, p. 590681, 2020.
- [180] S. Kim, J. Kwon, T. Lee, Y. Park, and J. Perez, “Safety-aware unsupervised skill discovery,” in *IEEE International Conference on Robotics and Automation*, 2023, pp. 894–900.

- [181] E. Altman, *Constrained Markov Decision Processes*. CRC Press, 1999.
- [182] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 1861–1870.
- [183] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, “Safe learning in robotics: From learning-based control to safe reinforcement learning,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, no. 1, pp. 411–444, 2022.
- [184] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll, “A review of safe reinforcement learning: Methods, theories and applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [185] T. G. Dietterich, “Hierarchical reinforcement learning with the maxq value function decomposition,” *Journal of artificial intelligence research*, vol. 13, pp. 227–303, 2000.
- [186] H. Bharadhwaj, A. Kumar, N. Rhinehart, S. Levine, F. Shkurti, and A. Garg, “Conservative safety critics for exploration,” *arXiv preprint arXiv:2010.14497*, 2020.
- [187] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” *arXiv preprint arXiv:1509.02971*, 2015.
- [188] Y. Tang *et al.*, “Deep reinforcement learning for robotic manipulation: A survey,” *IEEE Transactions on Robotics*, 2024.
- [189] H. Bharadhwaj, “Auditing robot learning for safety and compliance during deployment,” in *Conference on Robot Learning*. PMLR, 2022, pp. 1801–1806.
- [190] S. Stramigioli, “Energy-aware robotics,” in *Mathematical Control Theory I*, M. K. Camlibel, A. A. Julius, R. Pasumathy, and J. M. Scherpen, Eds. Cham: Springer International Publishing, 2015, pp. 37–50.
- [191] J. Lachner, F. Allmendinger, S. Stramigioli, and N. Hogan, “Shaping Impedances to Comply With Constrained Task Dynamics,” *IEEE Transactions on Robotics*, vol. 38, no. 5, pp. 2750–2767, 2022.
- [192] Y. Michel, C. Ott, and D. Lee, “Safety-Aware Hierarchical Passivity-Based Variable Compliance Control for Redundant Manipulators,” *IEEE Transactions on Robotics*, vol. 38, no. 6, pp. 3899–3916, 2022.
- [193] E. Shahriari, L. Johansmeier, E. Jensen, and S. Haddadin, “Power flow regulation, adaptation, and learning for intrinsically robust virtual energy tanks,” *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 211–218, 2020.
- [194] Z. Jin, A. Liu, W.-a. Zhang, L. Yu, and C. Yang, “Learning stable state-dependent variable impedance control for compliant manipulation,” *IEEE/ASME Transactions on Mechatronics*, 2024.

- [195] Y. Michel, M. Saveriano, and D. Lee, “A Novel Safety-Aware Energy Tank Formulation Based on Control Barrier Functions,” *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5206–5213, Jun. 2024.
- [196] J. Sebastian *et al.*, “Safe reinforcement learning for contact-rich manipulation,” in *IEEE International Conference on Robotics and Automation*, 2023.
- [197] J. Lachner, “A geometric approach to robotic manipulation in physical human-robot interaction,” PhD Thesis - Research UT, graduation UT, University of Twente, Netherlands, July 2022.
- [198] B. Tang, M. A. Lin, I. Akinola, A. Handa, G. S. Sukhatme, F. Ramos, D. Fox, and Y. Narang, “Industreal: Transferring contact-rich assembly tasks from simulation to reality,” *arXiv preprint arXiv:2305.17110*, 2023.
- [199] T. Tsuji, Y. Kato, G. Solak, H. Zhang, T. Petrič, F. Nori, and A. Ajoudani, “A survey on imitation learning for contact-rich tasks in robotics,” *arXiv preprint arXiv:2506.13498*, 2025.
- [200] Y. Narang, K. Storey, I. Akinola, M. Macklin, P. Reist, L. Wawrzyniak, Y. Guo, A. Moravanszky, G. State, M. Lu *et al.*, “Factory: Fast contact for robotic assembly,” *arXiv preprint arXiv:2205.03532*, 2022.
- [201] Z. Zhou, S. Veeramani, H. Fakhruddin, S. Uyanik, and A. I. Cooper, “Genco: A dual vlm generate-correct framework for adaptive peg-in-hole robotics,” in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 16 744–16 751.
- [202] H. Li, S. Zhang, and D. Guo, “Robocleaner: Robotic tabletop cleaning via vlm-powered multi-agent collaboration,” *IEEE Transactions on Automation Science and Engineering*, vol. 23, pp. 3651–3660, 2026.
- [203] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [204] Q. Xie, S. Y. Min, P. Ji, Y. Yang, T. Zhang, K. Xu, A. Bajaj, R. Salakhutdinov, M. Johnson-Roberson, and Y. Bisk, “Embodied-rag: General non-parametric embodied memory for retrieval and generation,” *arXiv preprint arXiv:2409.18313*, 2024.
- [205] X. Zhang, S. Liu, P. Huang, W. J. Han, Y. Lyu, M. Xu, and D. Zhao, “Dynamics as prompts: In-context learning for sim-to-real system identifications,” *IEEE Robotics and Automation Letters*, vol. 10, no. 4, pp. 3190–3197, 2025.
- [206] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, “\pi_0 : A vision-language-action flow model for general robot control,” *arXiv preprint arXiv:2410.24164*, 2024.
- [207] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, “Bge m3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation,” *arXiv preprint arXiv:2402.03216*, 2024.

- [208] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, “Rdt-1b: a diffusion foundation model for bimanual manipulation,” *arXiv preprint arXiv:2410.07864*, 2024.
- [209] J. Wen, Y. Zhu, M. Zhu, Z. Tang, J. Li, Z. Zhou, X. Liu, C. Shen, Y. Peng, and F. Feng, “Diffusionvla: Scaling robot foundation models via unified diffusion and autoregression,” in *Forty-second International Conference on Machine Learning*, 2025.
- [210] Physical Intelligence, “ $\pi_0.5$: A vision-language-action model with open-world generalization,” *arXiv preprint*, 2025.
- [211] M. J. Kim, C. Finn, and P. Liang, “Fine-tuning vision-language-action models: Optimizing speed and success,” *arXiv preprint arXiv:2502.19645*, 2025.
- [212] Y. Zhong, F. Bai, S. Cai, X. Huang, Z. Chen, X. Zhang, Y. Wang, S. Guo, T. Guan, K. N. Lui *et al.*, “A survey on vision-language-action models: An action tokenization perspective,” *arXiv preprint arXiv:2507.01925*, 2025.
- [213] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, “Libero: Benchmarking knowledge transfer for lifelong robot learning,” *Advances in Neural Information Processing Systems 36*, pp. 44 776–44 791, 2023.
- [214] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T.-k. Chan *et al.*, “Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai,” *arXiv preprint arXiv:2410.00425*, 2024.
- [215] K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. R. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, b. ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, “ $\pi_0.5$: a vision-language-action model with open-world generalization,” in *Proceedings of The 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Lim, S. Song, and H.-W. Park, Eds., vol. 305. PMLR, 27–30 Sep 2025, pp. 17–40. [Online]. Available: <https://proceedings.mlr.press/v305/black25a.html>
- [216] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski *et al.*, “Rt-2: Vision-language-action models transfer web knowledge to robotic control,” 2023. [Online]. Available: <https://arxiv.org/abs/2307.15818>
- [217] B. Bianchini, M. Zhu, M. Sun, B. Jiang, C. J. Taylor, and M. Posa, “Vysics: Object reconstruction under occlusion by fusing vision and contact-rich physics,” in *Robotics: Science and Systems XXI*, 2025.
- [218] H. Zhang, R. Dai, G. Solak, P. Zhou, Y. She, and A. Ajoudani, “Safe learning for contact-rich robot tasks: A survey from classical learning-based methods to safe foundation models,” *arXiv preprint arXiv:2512.11908*, 2025.
- [219] P. Hao, C. Zhang, D. Li, X. Cao, X. Hao, S. Cui, and S. Wang, “Tla: Tactile-language-action model for contact-rich manipulation,” *arXiv preprint arXiv:2503.08548*, 2025.

- [220] C. Zhang, P. Hao, X. Cao, X. Hao, S. Cui, and S. Wang, “Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation,” *arXiv preprint arXiv:2505.09577*, 2025.
- [221] S. Yu, K. Lin, and H. Soh, “Demonstrating the octopi-1.5 visual-tactile-language model,” *arXiv preprint arXiv:2507.09985*, 2025.
- [222] Z. Cheng, Y. Zhang, W. Zhang, H. Li, K. Wang, L. Song, and H. Zhang, “Omnivtla: Vision-tactile-language-action model with semantic-aligned tactile sensing,” *arXiv preprint arXiv:2508.08706*, 2025.
- [223] J. Bi, K. Y. Ma, C. Hao, M. Z. Shou, and H. Soh, “Vla-touch: Enhancing vision-language-action models with dual-level tactile feedback,” *arXiv preprint arXiv:2507.17294*, 2025.
- [224] Z. Liu, J. Liu, J. Xu, N. Han, C. Gu, H. Chen *et al.*, “Mla: A multisensory language-action model for multimodal understanding and forecasting in robotic manipulation,” *arXiv preprint arXiv:2509.26642*, 2025.
- [225] J. Huang, S. Wang, F. Lin, Y. Hu, C. Wen, and Y. Gao, “Tactile-vla: unlocking vision-language-action model’s physical knowledge for tactile generalization,” *arXiv preprint arXiv:2507.09160*, 2025.
- [226] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” *arXiv preprint arXiv:2303.15343*, 2023.
- [227] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, “3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing,” *arXiv preprint arXiv:2410.24091*, 2025.
- [228] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, “Diffusion policy: Visuomotor policy learning via action diffusion,” in *Proceedings of Robotics: Science and Systems*, 2023.
- [229] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, “3D Diffusion Policy: Generalizable visuomotor policy learning via simple 3d representations,” in *Proceedings of Robotics: Science and Systems*, 2024.

Appendix A

Simulation Environment and Implementation Details

A.1 MuJoCo Simulation Setup

All reinforcement learning experiments presented in this thesis were conducted using the MuJoCo physics simulator (version 2.3.3). MuJoCo (Multi-Joint dynamics with Contact) is a physics engine designed for simulating articulated rigid body systems with contact dynamics. It provides accurate simulation of contact forces, friction, and deformation, making it well-suited for contact-rich manipulation tasks.

A.1.1 Robot Model

The Franka Emika Panda robot arm was used in all simulation experiments. The Panda is a 7-DoF collaborative robot arm with a payload of 3 kg and a reach of 855 mm. The robot model in MuJoCo includes:

- Joint position, velocity, and acceleration limits.
- Joint torque limits.
- Collision geometry for all links.
- End-effector force/torque sensor.

The simulation runs at 1 kHz for the physics simulation, with the RL policy running at 10 Hz (SRL-VIC, Bresa) or 100 Hz (Passivity-Centric Safe RL).

A.1.2 Maze Environment

The maze environment was custom-built for the contact-rich maze exploration task. The maze is a 2D environment with:

- Channel width: 50 mm
- Channel length: 70.35 cm
- Wall material: Rigid (infinite stiffness)
- Movable obstacles: Cylindrical objects with mass 0.1 kg

The maze is designed to create a dilemma: contacts are absolutely necessary to navigate while they are also risky. The movable obstacles add an additional challenge, requiring the robot to push harder to overcome them.

A.2 Simulation Parameters

The key simulation parameters used in the experiments are summarized in Table A.1:

Table A.1 Simulation parameters for maze exploration experiments

Parameter	Value
Physics simulation frequency	1000 Hz
RL policy frequency	10 Hz
Episode horizon	500 steps
Force threshold F_{max}	20 N
Stiffness range $[K_{min}, K_{max}]$	[300, 1000] N/m
Displacement range $[\Delta P_{min}, \Delta P_{max}]$	[-0.03, 0.03] m

A.3 Real-World Robot Setup

The real-world experiments were conducted using a physical Franka Emika Panda robot. The robot was controlled using the Franka Control Interface (FCI), which provides access to the robot's joint positions, velocities, and torques at 1 kHz.

The real-world setup includes:

- Franka Emika Panda robot arm

- ATI Mini45 force/torque sensor at the end-effector
- Custom peg-shaped flange for maze exploration
- Custom maze environment with rigid walls and movable obstacles

The policy trained in simulation was deployed on the physical robot without fine-tuning. The sim-to-real transfer was successful due to the force-based state representation, which is consistent between simulation and real-world.

A.4 Neural Network Architectures

The neural network architectures used in the RL experiments are summarized in Table A.2:

Table A.2 Neural network architectures for RL experiments

Network	Input	Hidden Layers	Output
Task policy	9D state	2×256 (ReLU)	4D action
Safety critic	6D state + 4D action	2×256 (ReLU)	Scalar risk
Recovery policy	6D state	2×256 (ReLU)	4D action
Value function	9D state	2×256 (ReLU)	Scalar value

All networks use ReLU activations in the hidden layers and linear activations in the output layer. The networks are trained using the Adam optimizer with a learning rate of 3×10^{-4} .

A.5 Hyperparameter Sensitivity Analysis

A.5.1 SRL-VIC Hyperparameters

The sensitivity of SRL-VIC to key hyperparameters is analyzed:

Table A.3 SRL-VIC hyperparameter sensitivity analysis

Parameter	Default	Low	High	Sensitivity
Safety threshold ε_s	0.3	0.1	0.5	Medium
Recovery weight λ_r	0.5	0.1	0.9	High
Force threshold F_{max}	20 N	10 N	30 N	High
Stiffness range K_{max}	1000 N/m	500 N/m	2000 N/m	Medium
Discount factor γ	0.99	0.95	0.999	Low

The recovery weight λ_r and force threshold F_{max} are the most sensitive hyperparameters. A higher recovery weight leads to more conservative behavior, while a lower force threshold leads to more frequent safety interventions.

A.5.2 Bresa Hyperparameters

The sensitivity of Bresa to key hyperparameters is analyzed:

Table A.4 Bresa hyperparameter sensitivity analysis

Parameter	Default	Low	High	Sensitivity
Reflex threshold ε_r	0.7	0.5	0.9	High
Control loop frequency	100 Hz	50 Hz	200 Hz	Medium
Stiffness reduction α_K	0.5	0.2	0.8	Medium
Recovery duration T_r	0.5 s	0.2 s	1.0 s	Low

A.5.3 Passivity-Centric Safe RL Hyperparameters

Table A.5 Passivity-Centric Safe RL hyperparameter sensitivity

Parameter	Default	Low	High	Sensitivity
Energy tank capacity E_{max}	5.0 J	2.0 J	10.0 J	High
Passivity weight λ_p	0.1	0.01	0.5	High
Tank refill rate $\dot{E}_{refill}$	0.5 W	0.1 W	2.0 W	Medium

A.6 Training Curves and Convergence Analysis

A.6.1 SRL-VIC Training Curves

The training curves for SRL-VIC show:

- **Task reward:** Converges after approximately 200k training steps.
- **Safety violations:** Decreases monotonically from 100% to below 5% over 200k steps.
- **Safety critic loss:** Converges after approximately 100k training steps.
- **Recovery policy loss:** Converges after approximately 150k training steps.

A.6.2 Bresa Training Curves

The training curves for Bresa show:

- **Task reward:** Converges after approximately 150k training steps, faster than SRL-VIC due to the hierarchical structure.
- **Safety violations:** Decreases to near-zero after approximately 100k training steps.
- **Reflex activation rate:** Stabilizes at approximately 15% of control steps after convergence.

A.7 Real-World Experiment Details

A.7.1 Cable Routing Task Setup

The cable routing task was designed to simulate the installation of cables in industrial environments. The setup includes:

- A 3D-printed cable routing structure with multiple waypoints.
- A flexible cable with a diameter of 5 mm and a length of 50 cm.
- A custom end-effector designed to grip and route the cable.
- A force/torque sensor to measure contact forces during routing.

The task requires the robot to route the cable through a series of clips while maintaining safe contact forces. The cable's flexibility introduces additional compliance that must be accounted for by the impedance controller.

A.7.2 Peg-in-Hole Task Setup

The peg-in-hole task was designed to test the robot's ability to perform precise assembly operations. The setup includes:

- A custom peg with a diameter of 20 mm and a length of 100 mm.
- A hole with a diameter of 21 mm (1 mm clearance).
- A force/torque sensor to measure contact forces during insertion.

- A vision system to detect the initial peg and hole positions.

The 1 mm clearance makes this a challenging task that requires both precise positioning and compliant contact.

A.7.3 Disassembly Task Setup

The disassembly task was designed to test the robot's ability to perform constraint-locked disassembly operations. The setup includes:

- A custom fixture with four different fasteners: nut, bolt, screw, and pin.
- A custom end-effector designed to grip and unscrew/remove the fasteners.
- A force/torque sensor to measure contact forces during disassembly.
- A tactile sensor array on the gripper fingertips.

The disassembly task requires the robot to apply precise torques and forces to remove the fasteners while maintaining safe contact forces.

Appendix B

Foundation Model Details and Prompts

B.1 OmniVIC Prompt Design

This appendix provides the detailed prompt designs used in OmniVIC (Chapter 6) and CompliantVLA (Chapter 7).

B.1.1 Task Description Prompt

The task description prompt is used to provide the VLM with information about the current task. The prompt template is:

You are a robot controller expert. Given the following task description and visual observation, generate appropriate impedance parameters for safe contact-rich manipulation.

Task: {task_description}
Current phase: {phase}
Contact state: {contact_state}

Based on the task requirements and current state, provide:

1. Stiffness values $[K_x, K_y, K_z]$ in N/m
2. Damping values $[D_x, D_y, D_z]$ in Ns/m
3. Safety threshold $[F_{\max}]$ in N

Respond in JSON format.

B.1.2 RAG Knowledge Base

The RAG knowledge base contains examples of successful impedance parameter configurations for various contact-rich tasks. Each entry includes:

- Task description
- Contact state
- Impedance parameters
- Task outcome

The knowledge base is populated with expert demonstrations and successful task executions. At inference time, the most relevant entries are retrieved based on the current task description and contact state.

B.1.3 In-Context Learning Examples

The in-context learning examples are selected from the RAG knowledge base and provided to the VLM as context. The examples are formatted as:

Example 1:

Task: Peg-in-hole insertion (tight tolerance)

Phase: Approach

Contact state: No contact

Impedance: $K_x=800$, $K_y=800$, $K_z=200$, $D_x=40$, $D_y=40$, $D_z=20$

$F_{\max}$: 15 N

Example 2:

Task: Peg-in-hole insertion (tight tolerance)

Phase: Insertion

Contact state: Light contact

Impedance: $K_x=300$, $K_y=300$, $K_z=500$, $D_x=20$, $D_y=20$, $D_z=30$

$F_{\max}$: 20 N

B.2 Extended Prompt Examples for OmniVIC

B.2.1 Peg-in-Hole Insertion Prompt

System: You are an expert robot controller specializing in variable impedance control for contact-rich manipulation.

User: I need to perform a peg-in-hole insertion task.
The peg diameter is 20mm and the hole diameter is 21mm (1mm clearance). The current phase is "search" where the peg is near the hole but not yet aligned.

Please provide impedance parameters in JSON format with:

- stiffness: [Kx, Ky, Kz] in N/m
- damping: [Dx, Dy, Dz] in Ns/m
- force_threshold: F_max in N

The expected response from the VLM is:

```
{  
  "stiffness": [400, 400, 150],  
  "damping": [25, 25, 15],  
  "force_threshold": 12  
}
```

B.2.2 Cable Routing Prompt

System: You are an expert robot controller specializing in variable impedance control for contact-rich manipulation.

User: I need to route a flexible cable through a series of clips. The cable has a diameter of 5mm. The current phase is "routing" where the cable is being guided through a clip with a 6mm opening.

Please provide impedance parameters in JSON format with:

- stiffness: $[K_x, K_y, K_z]$ in N/m
- damping: $[D_x, D_y, D_z]$ in Ns/m
- force_threshold: $F_{\max}$ in N

The expected response from the VLM is:

```
{
  "stiffness": [200, 200, 300],
  "damping": [15, 15, 20],
  "force_threshold": 8
}
```

B.3 CompliantVLA Adaptor Architecture

The CompliantVLA adaptor consists of two main components:

B.3.1 VLM-based Impedance Generator

The VLM-based impedance generator takes the current visual observation and language instruction as input and generates impedance parameters. The generator uses a pre-trained VLM (GPT-4V or similar) with the OmniVIC prompt design described above.

The impedance generator operates at 1 Hz, generating new impedance parameters at each second. The parameters are then passed to the VIC controller, which executes the desired motion with the specified impedance.

B.3.2 Force Feedback Module

The force feedback module monitors the contact forces during execution and adjusts the impedance parameters if necessary. The module uses a simple threshold-based approach:

- If the contact force exceeds $F_{\max}$, the stiffness is reduced by 20%.
- If the contact force is below $0.5 \times F_{\max}$, the stiffness is increased by 10%.

This feedback mechanism ensures that the impedance parameters are adapted to the actual contact conditions, rather than relying solely on the VLM's predictions.

B.4 TacVLA Implementation Details

B.4.1 Tactile Sensor Specification

The tactile sensor used in TacVLA is a custom-designed compact tactile array with the following specifications:

- Array size: 8×8 taxels
- Taxel size: $2 \text{ mm} \times 2 \text{ mm}$
- Sampling frequency: 100 Hz
- Force range: 0–10 N
- Force resolution: 0.1 N

The tactile sensor is mounted on the fingertip of the robot gripper, providing contact force measurements at the point of contact.

B.4.2 Tactile Tokenization

The tactile signals are tokenized using a lightweight convolutional neural network (CNN) encoder. The encoder takes the 8×8 tactile array as input and outputs a 64-dimensional feature vector. The feature vector is then projected to the token dimension of the VLA model using a linear layer.

The tokenization procedure is:

1. Normalize the tactile signals to $[0, 1]$.
2. Apply the CNN encoder to obtain a 64-dimensional feature vector.
3. Project to the VLA token dimension using a linear layer.
4. Add positional encoding to indicate the tactile modality.

B.4.3 Contact-Aware Gating

The contact-aware gating mechanism selectively activates tactile tokens based on the contact state. The contact state is estimated from the tactile signals using a threshold:

$$\text{contact} = \mathbf{1} \left[\max_{i,j} T_{i,j} > T_{\text{threshold}} \right] \quad (\text{B.1})$$

where $T_{i,j}$ is the tactile signal at taxel (i, j) and $T_{\text{threshold}} = 0.1$ N is the contact threshold.

When contact is detected, the tactile tokens are included in the VLA model’s input. When no contact is detected, the tactile tokens are masked out, preventing irrelevant tactile information from interfering with the model’s predictions.

B.5 Training Details for Foundation Model Experiments

B.5.1 OmniVIC Training

OmniVIC does not require training, as it uses a pre-trained VLM with RAG and ICL. The knowledge base is populated offline with expert demonstrations.

B.5.2 CompliantVLA Training

The CompliantVLA adaptor is trained using the following procedure:

1. Collect demonstrations with synchronized visual, language, force, and action data.
2. Train the VIC controller using the collected demonstrations.
3. Fine-tune the VLA model with the VIC controller in the loop.

The training uses the AdamW optimizer with a learning rate of 1×10^{-4} and a batch size of 32.

B.5.3 TacVLA Training

TacVLA is trained using the following procedure:

1. Collect demonstrations with synchronized visual, language, tactile, and action data.
2. Pre-train the tactile encoder using contrastive learning.
3. Fine-tune the VLA model with the tactile encoder using LoRA.

The LoRA fine-tuning uses rank $r = 16$ and scaling factor $\alpha = 32$. The training uses the AdamW optimizer with a learning rate of 2×10^{-4} and a batch size of 16.