2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Contact-Safe Reinforcement Learning with ProMP Reparameterization and Energy Awareness

Bingkun Huang, Yuhe Gong, Zewen Yang, Tianyu Ren*, Luis Figueredo

***Abstract*— Reinforcement learning (RL) approaches based on Markov Decision Processes (MDPs) are predominantly applied in the robot joint space, often relying on limited task-specific information and partial awareness of the 3D environment. In contrast, episodic RL has demonstrated advantages over traditional MDP-based methods in terms of trajectory consistency, task awareness, and overall performance in complex robotic tasks. Moreover, traditional step-wise and episodic RL methods often neglect the contact-rich information inherent in task-space manipulation, especially considering the contact-safety and robustness. In this work, contact-rich manipulation tasks are tackled using a task-space, energy-safe framework, where reliable and safe task-space trajectories are generated through the combination of Proximal Policy Optimization (PPO) and movement primitives. Furthermore, an energy-aware Cartesian Impedance Controller objective is incorporated within the proposed framework to ensure safe interactions between the robot and the environment. Our experimental results demonstrate that the proposed framework outperforms existing methods in handling tasks on various types of surfaces in 3D environments, achieving high success rates as well as smooth trajectories and energy-safe interactions.**

I. Introduction

Contact-rich robotic manipulation imposes stringent requirements on safety, adaptability, and robustness due to discontinuous dynamics, transient contact forces, and complex energy exchanges. Uncontrolled interaction can lead to instability, excessive forces, or unintended motion, posing risks to both the robot and its environment. Ensuring safe physical interaction therefore requires not only effective regulation of energy flow, but also the ability to generate smooth and adaptable trajectories that remain robust under uncertainty. Traditional model-based motion-generation approaches (e.g., Movement Primitives) rely on accurate models and measurements [1], [2], yet such accuracy is difficult to guarantee in practice, particularly for physical interaction tasks. Reinforcement learning (RL), on the other hand, offers robustness through data-driven exploration and training (e.g., via domain randomization), but often produces non-smooth stepwise policies and lacks explicit safety guarantees.

To address these challenges, several key capabilities are required for a safe smooth manipulation framework: (1) contact-aware representations that capture task-space constraints and uncertainties; (2) trajectory-level planning strategies that ensure smooth, dynamically feasible motions, such as movement primitives; and (3) explicit mechanisms to regulate energy exchange, ensuring interactions remain within safe operating bounds through passivity-based and energy-aware control. Separately, each of these is well studied, yet no existing framework tightly integrates data-driven robustness, trajectory-level smoothness, and passivity-based safety for contact-rich manipulation. This work bridges that gap by embedding RL into a task-space, movement-primitive representation while explicitly enforcing energy-safe interaction during both learning and execution. We study two representative contact-rich tasks—box pushing and maze sliding—as illustrated in Fig. 1.

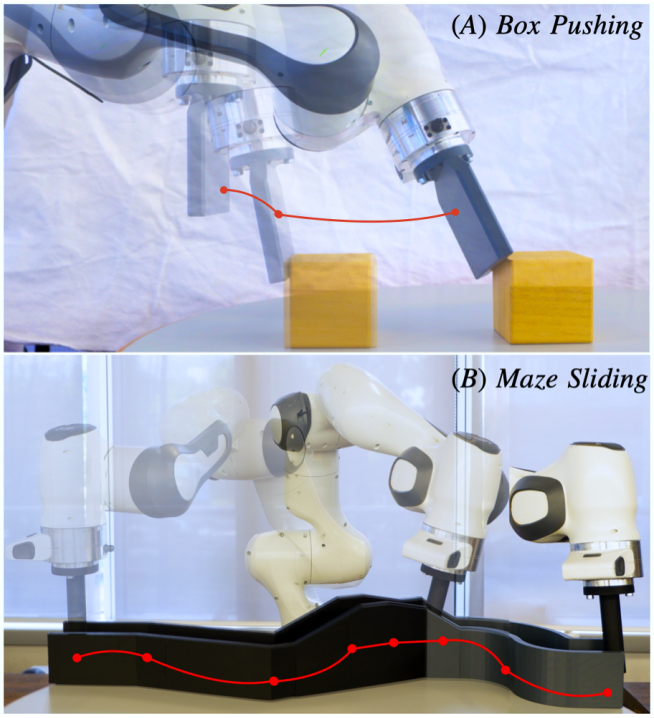

Fig. 1: Contact-rich manipulation tasks studied in this paper: (A) box pushing with sustained surface contact; (B) maze sliding where the tool navigates unseen turns relying on contact feedback. These tasks require regulating interaction energy to avoid unsafe force or power bursts during exploration and execution.

A. Related Work

Despite the growing success of RL for manipulation [3]–[5], its application to contact-rich tasks remains relatively limited [6]. Most RL methods are developed and evaluated in joint space or quasi-contact scenarios, where physical interactions with the environment are minimal and contact forces can often be neglected. Contact-rich manipulation (e.g., pushing, sliding, or assembly) introduces discontinuous dynamics and

This work was funded by the Lighthouse Initiative Geriatronics by StMWi Bayern (Project X, Grant No. 5140951). B. Huang and T. Ren are with Munich Institute of Robotics & Machine Intelligence, Technische Universität München (TUM), Germany. Y. Gong and L. Figueredo are with the School of Computer Science at The University of Nottingham. L. Figueredo is also an Associated Fellow at the MIRMI-TUM.
*Corresponding author: T. Ren <tianyu@robot-learning.de>

complex energy exchanges, which pose significant challenges for standard RL algorithms. Existing work has explored either model-based approaches that incorporate simplified contact dynamics or force feedback into policy learning [7], [8], or episodic RL combined with movement primitives to improve trajectory consistency under contact constraints [9], [10]. However, these approaches are still scarce and often struggle with generalization to unseen contacts and with explicit energy management on real hardware.

Safe reinforcement learning (SafeRL) [11] aims to enforce safety constraints during both learning and execution, such as limits on peak force, power, or energy. In contact-rich manipulation, transient forces, frictional uncertainties, and discontinuous dynamics make precise constraint modeling difficult, often limiting SafeRL's effectiveness. Complementarily, passivity-based control and energy-tank frameworks provide a principled safety layer by preventing the robot from injecting uncontrolled energy during physical interactions [12]–[14]. Yet, purely passivity-based approaches can be conservative and may not optimize task performance or trajectory smoothness.

Recent work has addressed complementary aspects of this problem. To improve trajectory consistency and handle sparse rewards, Otto et al. showed that policies parameterizing movement primitives can generate smoother trajectories than low-level stepwise policies [15]. For better generalization to unforeseen geometries, Zhou et al. proposed via-point movement primitives that adapt trajectories to novel constraints [16]. From a safety standpoint, passivity/energy-based methods explicitly regulate interaction energy; for example, Zhang et al. introduced energy constraints to guarantee safe interactions, though outside a trajectory-level policy-learning context [17]. Despite these advances, there remains no framework that jointly couples RL with trajectory-level movement primitives and energy-/passivity-aware execution for contact-rich manipulation.

B. Contribution

We propose **PPT**, a contact-safe RL framework that integrates Probabilistic Movement Primitives (ProMPs) for smooth task-space trajectory generation, Proximal Policy Optimization (PPO) for adaptive modulation, and an energy-tank passivity layer for energy-safe interaction in contact-rich manipulation. Our key contributions are: **(C1)** A task-space RL formulation that parameterizes actions in a low-dimensional ProMP weight space and executes them through Cartesian impedance control, enabling smooth, compliant trajectories for contact-rich tasks. **(C2)** A real-time energy-aware passivity controller (energy tank) that constrains interaction power/energy, providing safety guarantees during both learning and deployment under discontinuous contact dynamics. We validate the approach in simulation and on a Franka Panda robot on box-pushing and maze-sliding tasks, with ablations demonstrating the role of each component in safety, smoothness, and task success.

II. Preliminary

Robotic manipulation in contact-rich environments requires reasoning about trajectories, control forces, and safety simultaneously. In this section, we introduce the core concepts underlying our approach: trajectory representation, reinforcement learning for control adaptation, physical safety via passivity, and impedance-based execution.

A. Trajectory Representation with ProMPs

To model complex trajectories in a compact and adaptable way, we employ *Probabilistic Movement Primitives (ProMPs)* [18]. ProMPs encode a distribution over trajectories rather than a single deterministic path, allowing the robot to capture variability observed in demonstrations.

Let $\phi \in [0,1]$ denote a canonical phase variable, and $\boldsymbol{\Phi}(\phi) \in \mathbb{R}^{D\times K}$ a set of basis functions (e.g., radial basis functions). A trajectory $\boldsymbol{y}(\phi) \in \mathbb{R}^D$ is expressed as

$$\boldsymbol{y}(\phi) = \boldsymbol{\Phi}(\phi)\boldsymbol{w}, \quad \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w), \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^K$ are the trajectory weights. The Gaussian prior over $\boldsymbol{w}$ captures demonstration variability, while the linear combination ensures smoothness and computational tractability. In our method, we extend ProMPs with via-point conditioning and reinforcement-learning-based residual updates (Sec. IV).

B. Reinforcement Learning with PPO

While ProMPs provide structured priors, adapting them to new environments or tasks requires learning-based refinement. We model the control problem as a Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, P the transition dynamics, r the reward function, and γ the discount factor.

A stochastic policy $\pi_\theta(a|s)$ outputs a probability distribution over actions given a state. In the remainder, we denote the executed robot command as u_t; in generic RL notation it corresponds to the action a_t. We optimize the policy using *Proximal Policy Optimization (PPO)* [4], which stabilizes learning by limiting the magnitude of policy updates. The clipped surrogate objective is

$$\mathcal{L}_{\text{clip}}(\theta) = \mathbb{E}_t\Big[\min\left(r_t\hat{A}_t, \bar{r}_t\hat{A}_t\right)\Big], \tag{2}$$

$$\bar{r}_t \coloneqq \text{clip}(r_t, 1-\epsilon, 1+\epsilon), \quad r_t \coloneqq \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, \tag{3}$$

where $s_t \in \mathcal{S}$, $a_t \in \mathcal{A}$, $\hat{A}_t$ is the estimated advantage (e.g., via GAE), and ϵ is a small clipping parameter. PPO refines ProMP-generated references to improve task performance while maintaining stable learning.

C. Safety via Passivity and Energy-Tank Mechanisms

When interacting with the environment or humans, safety is critical. We enforce *passivity*, which ensures that the robot cannot inject unbounded energy [19], [20]. Let $E \in \mathbb{R}_{\geq 0}$ denote the robot's stored energy (kinetic + potential), and $p \in \mathbb{R}$ the instantaneous power exchanged with the environment. Passivity is expressed as

$$\dot{E} \leq p. \tag{4}$$

This constraint prevents high-impact forces, actuator saturation, and instability. Energy-tank mechanisms [21] explicitly

enforce this bound, ensuring safe execution of learned or planned trajectories.

D. Cartesian Impedance Control

Finally, to execute trajectories while interacting with the environment, we employ Cartesian impedance control. Let $(\boldsymbol{x}_d, \boldsymbol{R}_d)$ be the desired end-effector pose, with position error $\boldsymbol{e}_x = \boldsymbol{x}_d - \boldsymbol{x}$ and orientation error $\boldsymbol{e}_R = \frac{1}{2}(\boldsymbol{R}_d^\top \boldsymbol{R} - \boldsymbol{R}^\top \boldsymbol{R}_d)^\vee$, where $\boldsymbol{x} \in \mathbb{R}^3$, $\boldsymbol{R} \in SO(3)$, and $(\cdot)^\vee$ maps a skew-symmetric matrix to a vector. Let $\boldsymbol{\omega}$ and $\boldsymbol{\omega}_d$ denote the current and desired end-effector angular velocities, and define the angular-velocity error $\boldsymbol{\omega}_e := \boldsymbol{\omega}_d - \boldsymbol{\omega}$.

A standard Cartesian impedance law is

$$\boldsymbol{f} = \boldsymbol{K}_p\, \boldsymbol{e}_x + \boldsymbol{K}_d\, \dot{\boldsymbol{e}}_x, \qquad \boldsymbol{\tau} = \boldsymbol{K}_{p,R}\, \boldsymbol{e}_R + \boldsymbol{K}_{d,R}\, \boldsymbol{\omega}_e. \quad (5)$$

Here $\boldsymbol{f} \in \mathbb{R}^3$ and $\boldsymbol{\tau} \in \mathbb{R}^3$ are the commanded Cartesian force and torque at the end-effector. The resulting wrench and executed joint torques are

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{\tau} \end{bmatrix} \in \mathbb{R}^6, \qquad \boldsymbol{\tau}_{\text{imp}} = \boldsymbol{J}^\top(\boldsymbol{q}_t)\, \boldsymbol{\lambda},$$

where $\boldsymbol{J}(\boldsymbol{q}_t)$ is the $6 \times n$ geometric Jacobian at joint configuration $\boldsymbol{q}_t$, and $\boldsymbol{K}_p, \boldsymbol{K}_d, \boldsymbol{K}_{p,R}, \boldsymbol{K}_{d,R}$ are positive-definite gains. This controller executes ProMP- and PPO-generated references while maintaining compliance and safety.

III. Problem Definition

A. Overview of the System

Our framework, *PPT* (**P**roMP **P**PO Energy-**T**ank), integrates three complementary components: (i) **ProMPs** for structured, low-dimensional trajectory representation; (ii) **PPO** for adaptive, learning-based trajectory refinement; and (iii) **Energy-tank passivity control** for safe execution. An overview of the PPT framework is shown in Fig. 2. The system operates as follows:

1) **Trajectory Representation:** ProMPs encode robot trajectories as distributions over basis functions, allowing smooth and low-dimensional motion representation.
2) **Policy Refinement:** PPO updates the ProMP weights at each timestep,
$$\Delta \boldsymbol{w}_t = \pi(o_t), \quad (6)$$
to adapt trajectories based on observed performance.
3) **Safe Execution:** Trajectories generated from ProMP weights,
$$\boldsymbol{y}_t = \boldsymbol{\Phi}(\phi_t)\boldsymbol{w}_t, \quad (7)$$
are updated online according to energy feedback. The energy-tank mechanism ensures interaction forces remain within safe limits.
4) **Online Adaptation:** At test time, partial task-specific constraints can be provided via a small set of via-points
$$D_t = \{(\phi_i, y_i, \Sigma_i)\}_{i=1}^{N_t}, \quad (8)$$
which are incorporated by a conditioning operator $\mathcal{C}$:
$$\boldsymbol{w}_t^\star = \mathcal{C}(\boldsymbol{w}_t; D_t), \quad \boldsymbol{y}_t = \Phi(\phi_t)\boldsymbol{w}_t^\star. \quad (9)$$

This integration allows the system to generate *smooth, adaptive trajectories* while maintaining *safe energy interactions* and *generalization to unseen environments*.

B. Problem Statement.

We consider the problem of learning a policy $\pi_\theta(u_t \mid s_t)$ for a robot in a contact-rich environment, where $s_t \in \mathcal{S}$ denotes the robot state and $u_t \in \mathcal{U}$ the control input at time t. The robot interacts with dynamic and partially unknown objects, surfaces, or humans.

The goal is to maximize a cumulative task reward

$$R = \sum_{t=0}^{T} r(s_t, u_t),$$

while ensuring safety and adaptability. The reward encodes four key aspects:

1) **Goal / Path Shaping:** encourages progress toward the task goal and proper path following.
2) **Task Generalization:** encourages trajectory adaptation in free space.
3) **Contact-phase / Haptic Locomotion:** ensures safe and effective motion during contact.
4) **Power / Energy Safety:** penalizes excessive forces or energy usage.

Formally, the problem can be expressed as

$$\begin{aligned}
\max_{\pi_\theta} \quad & \mathbb{E}\left[\sum_{t=0}^{T} r(s_t, u_t)\right] \\
\text{s.t.} \quad & s_{t+1} = f(s_t, u_t), \quad s_0 \sim \mathcal{S}_0, \quad s_T \in \mathcal{S}_{\text{goal}}, \; u_t \in \mathcal{U}, \\
& \underbrace{P_t = \boldsymbol{\lambda}_t^{\text{nom}\top} \boldsymbol{\nu}_t, \quad p_t := |P_t|}_{\text{instantaneous power (6D)}}, \\
& 0 \le E_t \le E_{\max}, \quad \gamma_t \in [0, 1], \\
& \boxed{\gamma_t \le \frac{P_{\max}}{\max(\varepsilon,\, p_t)}, \qquad \gamma_t \le \frac{E_t}{\Delta t\, \max(\varepsilon,\, p_t)}} \\
& u_t = \gamma_t\, u_t^{\text{nom}}, \qquad u_t^{\text{nom}} = \pi_\theta(o_t), \\
& E_{t+1} = \min\{E_{\max},\; \max\{0,\; E_t - \gamma_t\, p_t\, \Delta t\}\}.
\end{aligned} \quad (10)$$

where $\pi_\theta(u_t \mid s_t)$ is the stochastic policy (parameters θ), $s_t \in \mathcal{S}$ the robot state, $u_t \in \mathcal{U}$ the executed control, o_t the observation, T the horizon, and $r(s_t, u_t)$ the step reward; the expectation is over rollouts induced by π_θ and dynamics f. $f : \mathcal{S} \times \mathcal{U} \to \mathcal{S}$ is the system dynamics, $\mathcal{S}_0$ the initial-state distribution, $\mathcal{S}_{\text{goal}}$ the goal set, and $\mathcal{U}$ the admissible control set. $\boldsymbol{\lambda}_t^{\text{nom}} = [\boldsymbol{f}_t^{\text{nom}}; \boldsymbol{\tau}_t^{\text{nom}}]$ is the nominal 6D wrench, $\boldsymbol{\nu}_t = [\dot{\boldsymbol{x}}_t; \boldsymbol{\omega}_t]$ the 6D twist, $P_t = \boldsymbol{\lambda}_t^{\text{nom}\top} \boldsymbol{\nu}_t$ the instantaneous power, $p_t := |P_t|$ its magnitude, Δt the control step, $\varepsilon > 0$ a small constant, E_t the energy-tank level (capacity $E_{\max}$), $P_{\max}$ the power limit, $\gamma_t \in [0, 1]$ the safety scaling factor, $u_t^{\text{nom}} = \pi_\theta(o_t)$ the nominal command, and $u_t = \gamma_t\, u_t^{\text{nom}}$ the safe command. (3D special case: replace $\boldsymbol{\lambda}, \boldsymbol{\nu}$ with $\boldsymbol{f}, \dot{\boldsymbol{x}}$.)

The objective is to learn a policy that can *generalize to unseen geometries*, produce *smooth, compliant trajectories*, and maintain *energy-safe interactions* in contact-rich environments.

IV. PPT: Contact-safe RL with ProMP

Building upon the preliminaries, our method introduces a task-space reinforcement learning framework that enables

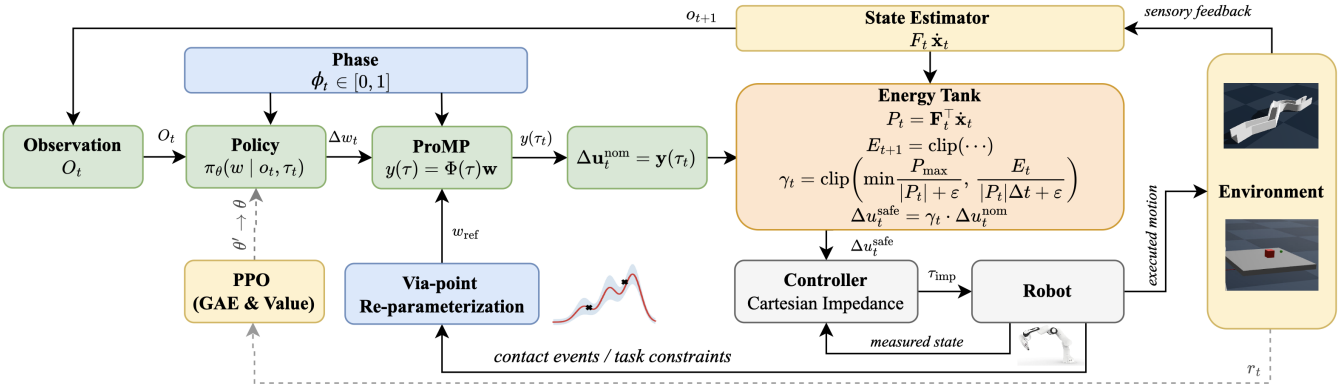

Fig. 2: Overview of the PPT framework. Observations and a phase variable are fed to the policy, which outputs residual ProMP weights. These are combined with via-point conditioned reference weights to form a ProMP trajectory. An energy-tank layer scales the command to ensure safe execution, which a Cartesian impedance controller tracks and converts to joint torques. Interaction feedback informs via-point reparameterization. PPO updates the policy using GAE and a value critic (dashed arrow).

online trajectory adaptation while explicitly enforcing *energy-safe interactions* in contact-rich tasks. The framework integrates three key components: structured trajectory priors, policy-driven adaptation, and passivity-based safety.

A. Trajectory Prior with ProMP

We represent a d-DoF task-space trajectory using a canonical phase variable $\phi \in [0,1]$ and K radial basis functions (RBFs):

$$y(\phi) = \mathbf{\Phi}(\phi)^\top \mathbf{W}, \qquad \Phi_k(\phi) = \exp\Big(-\frac{(\phi - c_k)^2}{2\sigma_k^2}\Big), \quad (11)$$

where $\mathbf{W} \in \mathbb{R}^{K\times d}$ is the basis-weight matrix (one column per task-space DoF), c_k and σ_k denote the centers and widths of RBF, respectively. We place a Gaussian prior over the vectorized weights $\boldsymbol{w} := \text{vec}(\mathbf{W})$:

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_w, \mathbf{\Sigma}_w), \quad (12)$$

with $(\boldsymbol{\mu}_w, \mathbf{\Sigma}_w)$ estimated from demonstrations or prior rollouts. This provides a smooth, low-dimensional, and probabilistic trajectory model that captures typical motion patterns while allowing variability.

B. Reinforcement Learning in ProMP Weight Space

Instead of acting directly in the raw control space, the policy refines the ProMP by outputting residual updates in weight space:

$$\boldsymbol{w}_t = \boldsymbol{w}_{\text{ref}} + \Delta \boldsymbol{w}_t, \qquad \Delta \boldsymbol{w}_t \sim \pi_\theta(\cdot \mid \tilde{o}_t), \quad (13)$$

where o_t denotes the robot observation and we augment it with the phase variable via $\tilde{o}_t := [o_t, \phi_t]$. The reference weights $\boldsymbol{w}_{\text{ref}}$ are given by the ProMP prior mean or the via-point posterior (Sec. IV-C). Let $\mathbf{W}_t := \text{unvec}(\boldsymbol{w}_t) \in \mathbb{R}^{K\times d}$. The adapted weights decode to references

$$\hat{y}_t = \mathbf{\Phi}(\phi_t)^\top \mathbf{W}_t$$

which are mapped to Cartesian impedance commands, yielding the nominal control u_t^{nom}. The policy π_θ is trained using PPO with the clipped surrogate objective:

$$\mathcal{L}(\theta) = \mathbb{E}_t\Big[\min\big(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\Big], \quad (14)$$

where $r_t(\theta)$ is the likelihood ratio and $\hat{A}_t$ is the advantage estimate. Operating in weight space leverages the structure and smoothness of ProMPs while enabling online trajectory adaptation.

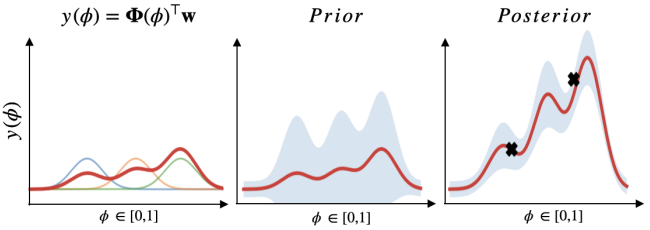

Fig. 3: **ProMP prior and via-point posterior (1D view).** Left: trajectory represented as $y(\phi) = \mathbf{\Phi}(\phi)^\top \mathbf{w}$ with K RBFs. Middle: prior $\text{vec}(\mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}_w, \mathbf{\Sigma}_w)$: mean (red) $\pm 2\sigma$ band (blue). Right: posterior conditioned on via-points $\mathcal{D}$ (black $\times$), showing tightened uncertainty near constraints while preserving smoothness.

C. Trajectory Posterior via Via-Point Conditioning

Fig. 3 illustrates the ProMP prior and the via-point posterior update. To incorporate partial geometric or contact constraints, we condition the prior on a set of via-points $\mathcal{D} = \{(\phi_j, \mathbf{y}_j)\}_{j=1}^M$ with observation covariance $\mathbf{\Sigma}_D$. Define

$$\mathbf{H} = \begin{bmatrix} \mathbf{\Phi}(\phi_1)^\top \\ \vdots \\ \mathbf{\Phi}(\phi_M)^\top \end{bmatrix}, \quad \mathbf{y}_\mathcal{D} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_M \end{bmatrix}.$$

The posterior distribution over weights is

$$\mathbf{\Sigma}_{w|\mathcal{D}} = (\mathbf{\Sigma}_w^{-1} + \mathbf{H}^\top \mathbf{\Sigma}_D^{-1} \mathbf{H})^{-1}, \quad (15)$$

$$\boldsymbol{\mu}_{w|\mathcal{D}} = \mathbf{\Sigma}_{w|\mathcal{D}}(\mathbf{\Sigma}_w^{-1}\boldsymbol{\mu}_w + \mathbf{H}^\top \mathbf{\Sigma}_D^{-1} \mathbf{y}_\mathcal{D}). \quad (16)$$

The posterior mean $\boldsymbol{\mu}_{w|\mathcal{D}}$ defines a trajectory that interpolates the via-points while preserving smoothness. We set $\boldsymbol{w}_{\text{ref}} = \boldsymbol{\mu}_{w|\mathcal{D}}$ and let PPO learn residual refinements $\Delta \boldsymbol{w}_t$ on top, effectively separating geometry-constrained adaptation from performance-driven learning.

D. Energy-Tank Layer for Safe Execution

To ensure contact safety and passivity, we integrate an *energy-tank mechanism* that monitors instantaneous power

$$p_t = |\boldsymbol{\lambda}_t^{\text{nom}\top} \boldsymbol{\nu}_t|$$

and scales the nominal command u_t^{nom} induced by the policy outputs (Sec. IV-B) by a factor $\gamma_t \in [0,1]$. According to the power and energy limits in Eq. (10), the executed command

$$u_t = \gamma_t\, u_t^{\text{nom}}$$

Variant	Policy type	Safety layer	Action
PP	Episode-level ProMP	–	ProMP parameters
PPT	**Episode-level ProMP**	**Energy tank ✓**	**ProMP parameters**
S	Step-wise PPO	–	Cartesian velocity
ST	Step-wise PPO	Energy tank ✓	Cartesian velocity

TABLE I: Method variants (PPT is our primary method).

is constrained by

$$\gamma_t \leq \min\left(\frac{P_{\max}}{\max(\varepsilon, p_t)},\ \frac{E_t}{\Delta t\ \max(\varepsilon, p_t)}\right),$$

with $\gamma_t \in [0,1]$ and $\varepsilon > 0$ a small constant.

Thus, the tank state is updated as

$$E_{t+1} = \min\{E_{\max},\ \max\{0,\ E_t - \gamma_t p_t \Delta t\}\}.$$

This layer directly scales Cartesian-force commands (extendable to velocity or other channels), ensuring passivity regardless of the policy and stabilizing contact-rich interactions.

V. Experiments

We design two complementary experiments to validate our method: (i) a pushing task to highlight the smoothness and stability benefits of ProMP-based trajectory generation, and (ii) a 3D maze sliding task to evaluate the generalization capability when facing unseen surface variations.

A. Common Experimental Setup

a) Platform and timing: All simulations are conducted in the Genesis physics simulator [22] using a 7-DoF Franka Emika Panda arm. Genesis performs dynamics integration at 2 kHz, while the high-level controller operates at 100 Hz with a fixed timestep of $\Delta t = 0.01$ s. The same controller frequencies are maintained for real-world experiments to ensure consistency between simulation and hardware.

b) Policy, actions, and tracking: We train policies using rsl_rl PPO with GAE [23], employing an actor–critic MLP with ReLU activations (256–256–128) and an adaptive learning rate, over 1000 episodes with parallel environments. Two policy parameterizations are considered: (i) *episode-level* ProMP, which outputs trajectory weights, and (ii) *step-wise* PPO, which directly outputs Cartesian velocity commands. For the ProMP variants, a Cartesian impedance controller tracks the generated reference trajectories. Method variants are summarized in Table I.

c) Safety constraint (energy tank): We enforce a power budget through the energy-tank mechanism described in Sec. IV. The instantaneous mechanical power is as $P_t = \boldsymbol{\lambda}_t^\top \boldsymbol{\nu}_t$ and is constrained by $P_t \leq P_{\max}$, where $\boldsymbol{\lambda}_t \in \mathbb{R}^6$ is the measured wrist wrench (force/torque) and $\boldsymbol{\nu}_t \in \mathbb{R}^6$ is the end-effector twist. Exceeding this limit incurs a penalty and may result in episode termination. The same power-safety mechanism is applied consistently in both simulation and real-world experiments.

d) Observations and sensing: Observations consist of joint positions and velocities, end-effector pose and twist, as well as wrist wrench measurements. On the hardware, we use the Franka external-wrench estimate, while in simulation we obtain the wrist wrench from Genesis contact reporting at the end-effector link; we denote it by $\boldsymbol{\lambda}_t$ and the corresponding end-effector twist by $\boldsymbol{\nu}_t$.

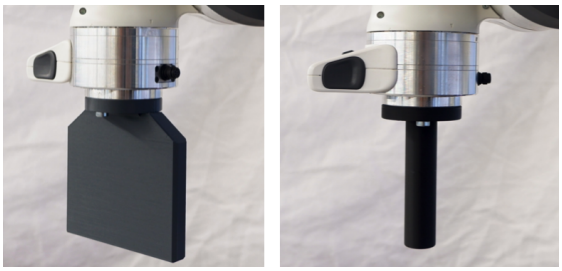

Fig. 4: End-effector tools for experiments.

Metric	Unit	Definition
Max Power	W	Maximum instantaneous power per episode.
Success rate	%	Percentage of episodes satisfying the task success criterion.
Jerk RMS	m/s^3	End-effector jerk root-mean-square.
Peak wrench (P95)	N	95th percentile of wrist-wrench norm $\mathrm{P95}(\|\boldsymbol{\lambda}_t\|)$.
Overload ratio	%	Time fraction with $P_t > P_{\max}$.
Contact continuity	0–1	Mean duration/ratio of contact segments.
Progress@T	0–1	Normalized progress at horizon T.

TABLE II: Evaluation metrics shared by all experiments; task-specific thresholds/horizons are defined in each subsection.

e) End-effector tools: Two rigid tools are employed (Fig. 4): a slim paddle ($12 \times 1 \times 1$ cm) for the pushing task, and a PLA cylinder (length 12 cm, diameter 2.5 cm) for the maze-sliding task. The physical maze is fabricated via PLA FDM, intentionally leaving a coarse surface finish to increase contact variability, while the simulation imports the exact STL geometry to match the real-world layout.

f) Training and reporting protocol: Unless otherwise specified, results are reported as mean ± SE over k random seeds, with curves smoothed using a moving-average window of w steps. The episode horizon and success criteria are task-specific and detailed in the corresponding subsections. Domain randomization is applied within task-dependent ranges (e.g., friction coefficients, object mass, and pose) to improve robustness. Shared evaluation metrics are listed in Table II.

B. Simulation Experiments

1) Box Pushing: A slim paddle is used to push a box across a planar tabletop from a designated start region to a goal. Task-specific domain randomization is applied on a per-episode basis: the kinetic friction coefficient is sampled as $\mu_k \sim \mathcal{U}(0.20, 0.60)$ (with static friction $\mu_s = 1.25\,\mu_k$), and the box mass is jittered by $\pm 15\%$ around two nominal sizes (6 cm / 50 g and 8 cm / 80 g). During training, the policy has access to privileged box pose information for faster credit assignment; at test time, this information is removed, leaving only start and goal positions (partial observability).

Across random seeds (Fig. 5), the ProMP-based trajectory policy with energy tank (PPT) exhibits rapid and steady learning, reaching a high success plateau while maintaining the lowest near-peak power. The energy tank effectively clamps force bursts during exploratory contact. The step-wise variant with the same safety layer (ST) converges more slowly and shows higher variance due to per-step action fluctuations. A step-wise policy without the tank (PP) learns

quickly but suffers occasional regressions, while the step-wise baseline without safety (S) is the least stable. Although heavier boxes increase absolute task difficulty, the relative ordering among methods is preserved, demonstrating that trajectory-level generation combined with energy shaping mitigates violent exploration and enhances reliability in contact-rich interactions.

2) Maze Sliding: Policies are trained exclusively in straight corridors to acquire a phase-aligned wall-following prior, and are subsequently deployed in unseen mazes (corridor width 5–6 cm, total length $\sim$1 m) featuring 20°–45° turns, a disc-shaped segment, and up to 4 cm vertical undulations. (Fig. 6) Observations are limited to proprioception and wrist wrench measurements (no map or vision). Domain randomization includes variations in start pose, heading, and bounded wall friction. The power budget follows the standard setup, and the episode horizon is $T = 20$ s.

To isolate the effect of trajectory parameterization, we compare only the deployable, safety-layered variants under the same power budget: PPT and ST. Reward components follow the definitions in Sec. V-A. While PPT requires no task-specific redesign, ST achieves consistent performance only after introducing stronger slip and heading regularization.

Figure 7 demonstrates that the wall-following prior learned in straight successfully transfers to curved and height-varying mazes. Waypoints concentrate near bends, and PPT produces a narrow posterior band that closely adheres to the wall, exhibiting smoother cornering and reduced lateral spread.

Overall, these results highlight that trajectory-level parameterization combined with energy-aware power gating is crucial for safe, contact-only navigation and effective generalization to novel geometries.

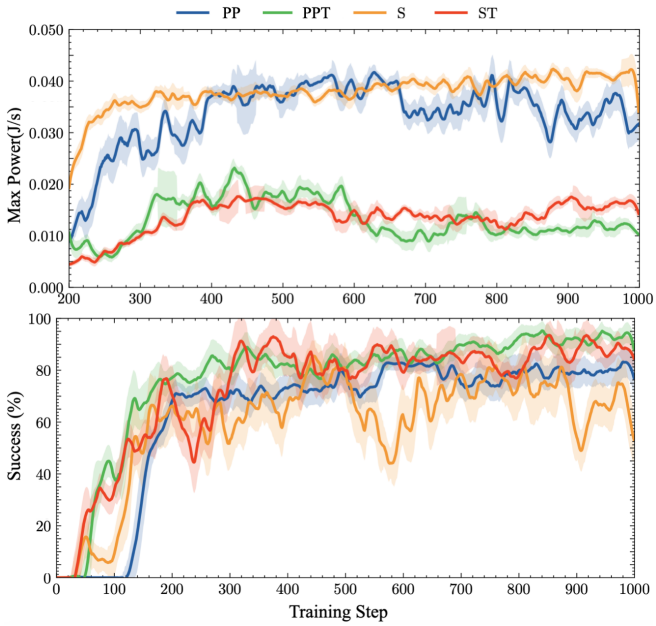

Fig. 5: **Training curves.** Success rate (%) and max instantaneous power (W) vs. training steps for **PP**, **PPT**, **S**, **ST**. Higher success and lower power are better.

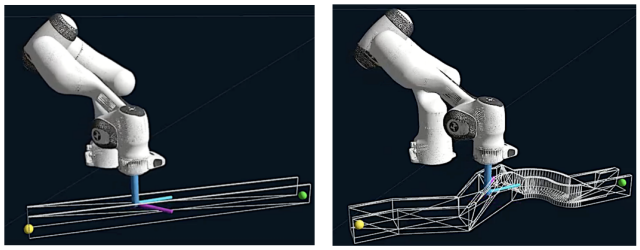

Fig. 6: **Maze sliding scenarios.** Left: training on straight corridors to learn a phase-aligned wall-following prior. Right: deployment in an unseen maze with turns, a disc-like segment, and height variations.

C. Real-World Experiments

1) Box Pushing: We evaluate two cubic boxes (edge lengths 6 cm and 8 cm; nominal masses 50 g and 80 g) on a flat laminate tabletop. Surface micro-roughness and dust naturally induce variability in friction and stick–slip behavior, which is *not* explicitly modeled. The energy tank remains active with the same budget as in simulation. Each trial begins from a randomized start pose within a fixed start region and terminates upon goal capture or safety violation.

PPT consistently yields the lowest jerk and near-peak wrench and the highest contact continuity, indicating smoother motions and steadier contact under the same power budget (Fig. 8). *S/ST* can reach slightly higher Progress@T on some trials, but at the cost of larger dispersion and more near-overload events, mirroring simulation. Absolute jerk/wrench are modestly higher than in sim due to sensing noise and real stick–slip, but the method ranking is preserved, supporting the robustness of trajectory-level parameterization plus power gating.

2) Maze Sliding: We deploy a 1 m-long PLA maze with corridor widths of 5–6 cm, printed with a deliberately coarse surface finish to enhance contact variability. The maze layout includes 20°–45° turns, a disc-like segment, and vertical undulations of up to 4 cm. The end-effector is a PLA cylinder, and sensing, filtering, and the energy-tank budget are identical to the simulation setup.

Figure 9 and Table III demonstrate that PPT successfully transfers the straight-corridor prior to complex maze geometries, producing trajectories with tighter dispersion and smoother cornering. It achieves higher task success and a stronger safety envelope, reflected in lower jerk, reduced near-peak wrench, fewer overload events, and higher contact continuity. ST attains slightly higher progress within a fixed horizon (Progress@T) but exhibits larger lateral spread and more frequent near-overload behaviors. Qualitative observations show that contact-inferred waypoints densify around

Metric	Unit	**PPT**	**ST**
Success rate	%	**89**	60
Jerk RMS	m/s^3	**1.85**	2.70
Peak wrench (P95)	N	**8.5**	11.2
Overload ratio	%	**2.3**	3.1
Contact continuity	0–1	**0.74**	0.48
Progress@T	0–1	0.70	**0.76**

TABLE III: Real-world maze sliding under the same power budget: PPT vs. ST (means).

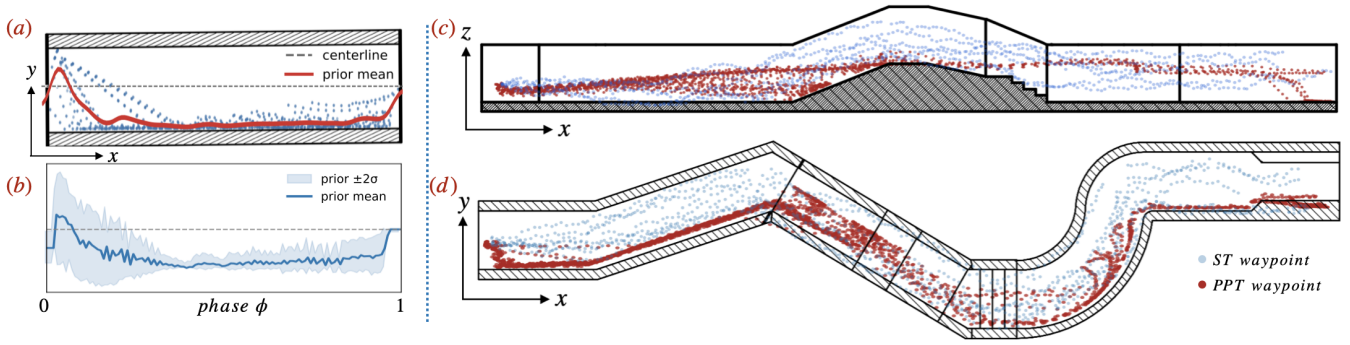

Fig. 7: **Maze sliding—training prior and test posterior rollouts (PPT vs. ST). (a)** Prior learned on straight corridors: dashed gray centerline and the learned prior mean (solid). **(b)** Prior dispersion over phase ϕ: mean with $\pm 2\sigma$ band. **(c)** Test maze in $(x-z)$ view: posterior rollouts with contact-inferred waypoints; **PPT** in red and **ST** in blue (dots mark detected waypoints). **(d)** Same as (c) in $(x-y)$ view. Waypoints cluster near bends/junctions; **PPT** shows a tighter band and smoother transitions than **ST**.

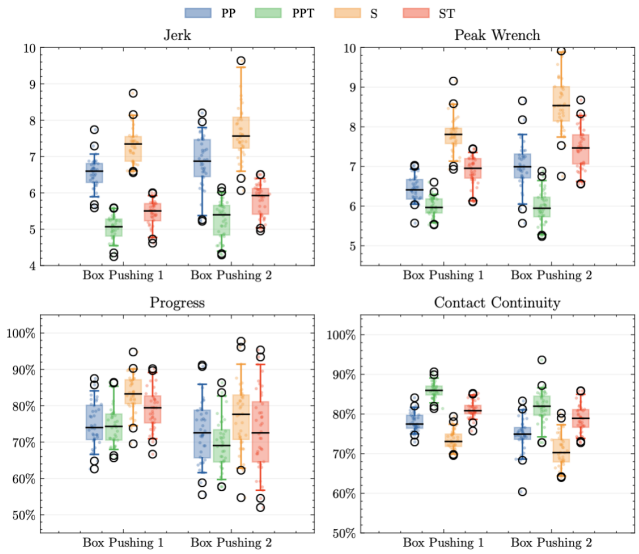

Fig. 8: **Real-world pushing metrics.** Jerk RMS, Peak Wrench P95 (N), Progress (%), and Contact Continuity (%) for two object conditions (*Box Pushing 1/2*). Boxes/whiskers show episode/seed distributions.

bends; PPT tracks these transitions with minimal overshoot, whereas ST often oscillates before settling. Overall, the sim→real trends are consistent: trajectory-level parameterization combined with power gating is key to safe, contact-only navigation, while step-wise control trades stability and safety for speed.

Across both tasks, real-world results corroborate the simulation findings: (i) the energy tank reliably enforces the power budget under unmodeled friction and sensor noise, and (ii) ProMP-based trajectory parameterization yields smoother, more stable behavior with higher success. Residual sim-to-real differences, such as slightly higher jerk or wrench, are attributable to measurement noise and surface stick–slip. Notably, PPT required no reward redesign or policy finetuning.

D. Discussion

Our experiments consistently show that the PPT outperforms step-wise policies ST in terms of smoothness and stability. The structured nature of ProMPs promotes globally coherent trajectories, resulting in lower jerk, reduced

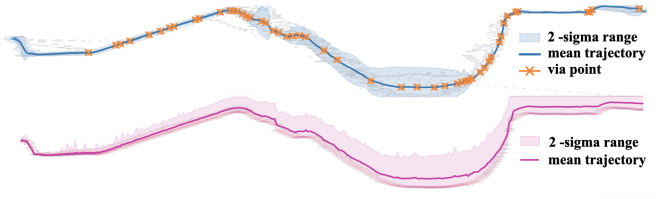

Fig. 9: **Real-world maze sliding.** Top: **PPT**; bottom: **ST**. Trajectory mean and $\pm 2\sigma$ bands are overlaid; dots denote contact-inferred waypoints. Waypoints cluster near bends/junctions; **PPT** shows tighter bands and smoother redirection through corners.

peak wrench, and higher contact continuity. This inherent smoothness mitigates high-frequency, reactive actions that often destabilize step-wise methods during contact.

The results further highlight the synergy between ProMP-based policies and the energy-tank safety layer. For PPT, the energy tank acts as a robust safety net for unexpected events, whereas for the more erratic ST policy, frequent tank interventions produce hesitant and inefficient motions, compromising performance despite ensuring safety.

Finally, the framework demonstrates strong practical utility, with robust sim-to-real transfer and generalization to unseen maze geometries without policy finetuning. The system effectively handles unmodeled friction and sensor noise, confirming that structured trajectory learning integrated with energy-aware safety is a powerful paradigm for reliable and safe contact-rich manipulation.

VI. Conclusion

We present a contact-safe reinforcement learning framework that integrates ProMP-based trajectory learning with a passivity-based energy tank. By combining the smooth, structured, and adaptive policies of ProMPs with the strong safety guarantees of the energy tank, our approach enables stable and efficient contact-rich manipulation across a variety of contact surfaces. Numerical simulations and real-world experiments demonstrate that our method outperforms step-wise RL baselines, achieving higher success rates and smoother, safer interactions with robust sim-to-real transfer. Despite these advantages, our method has limitations. The fixed-budget energy tank can be conservative, potentially limiting task performance, and effective generalization relies on a suitable trajectory prior. Future work will explore

adaptive energy management strategies to better balance safety and performance, as well as hierarchical trajectory priors to enhance generalization across a broader range of tasks.

REFERENCES

[1] B. Armstrong-Hélouvry, P. Dupont, and C. C. de Wit, "A survey of models, analysis tools and compensation methods for the control of machines with friction," *Automatica*, vol. 30, no. 7, 1994.

[2] M. Jean, "The non-smooth contact dynamics method," *Computer Methods in Applied Mechanics and Engineering*, vol. 177, no. 3-4, pp. 235–257, 1999.

[3] 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:1509.02971*, 2015.

[4] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[5] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *JMLR*, vol. 17, no. 39, pp. 1–40, 2016.

[6] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[7] P. Kormushev, S. Calinon, and D. G. Caldwell, "Reinforcement learning in robotics: Applications and real-world challenges," *Robotics*, vol. 2, no. 3, pp. 122–148, 2013.

[8] P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal, "From dynamic movement primitives to associative skill memories," *Robotics and Autonomous Systems*, vol. 61, no. 4, pp. 351–361, 2013.

[9] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in *RSS*, 2018.

[10] F. Otto, O. Celik, H. Zhou, H. Ziesche, V. A. Ngo, and G. Neumann, "Deep black-box reinforcement learning with movement primitives," in *CoRL*, 2023.

[11] J. García and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015. [Online]. Available: https://www.jmlr.org/papers/volume16/garcia15a/garcia15a.pdf

[12] R. Ortega, A. van der Schaft, I. Mareels, and B. Maschke, "Passivity-based control of euler–lagrange systems: A survey of some results," *Automatica*, vol. 38, no. 4, pp. 585–596, 2002.

[13] S. Haddadin and E. Croft, "Physical human–robot interaction," in *Springer Handbook of Robotics (2nd ed.)*, B. Siciliano and O. Khatib, Eds. Cham: Springer, 2016, pp. 1835–1874, often cited as a core reference for safety/pHRI.

[14] F. Ferraguti, N. Preda, A. Manurung, C. Secchi, and C. Fantuzzi, "An energy tank-based interactive control architecture for autonomous and teleoperated robotic surgery," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1073–1088, 2015.

[15] F. Otto, O. Celik, H. Zhou, H. Ziesche, V. A. Ngo, and G. Neumann, "Deep black-box reinforcement learning with movement primitives," in *Proceedings of The 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 14–18 Dec 2023, pp. 1244–1265. [Online]. Available: https://proceedings.mlr.press/v205/otto23a.html

[16] Y. Zhou, J. Gao, and T. Asfour, "Learning via-point movement primitives with inter- and extrapolation capabilities," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 4301–4308.

[17] H. Zhang, G. Solak, S. Hjorth, and A. Ajoudani, "Passivity-centric safe reinforcement learning for contact-rich robotic tasks," *arXiv preprint arXiv:2503.00287*, 2025. [Online]. Available: https://arxiv.org/abs/2503.00287

[18] A. Paraschos, C. Daniel, J. Peters, and G. Neumann, "Probabilistic movement primitives," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013. [Online]. Available: https://papers.nips.cc/paper/5177-probabilistic-movement-primitives

[19] N. Hogan, "Impedance control: An approach to manipulation, part i—theory," *Journal of Dynamic Systems, Measurement, and Control*, vol. 107, no. 1, pp. 1–7, 1985.

[20] R. Ortega, A. Loria, P. J. Nicklasson, and H. Sira-Ramírez, *Passivity-Based Control of Euler–Lagrange Systems: Mechanical, Electrical and Electromechanical Applications*, ser. Communications and Control Engineering. London: Springer, 2001.

[21] S. Haddadin, A. Albu-Schäffer, and G. Hirzinger, "Safety evaluation of physical human–robot interaction via energy-based, ergonomic and dynamic criteria," *The International Journal of Robotics Research*, vol. 31, no. 13, pp. 1578–1595, 2012.

[22] G. Authors, "Genesis: A generative and universal physics engine for robotics and beyond," December 2024. [Online]. Available: https://github.com/Genesis-Embodied-AI/Genesis

[23] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 164. PMLR, 2022, pp. 91–100. [Online]. Available: https://proceedings.mlr.press/v164/rudin22a.html