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

Safe and Optimal Variable Impedance Control via Certified Reinforcement Learning

Shreyas Kumar and Ravi Prakash
shr-eyas.github.io/CGMS

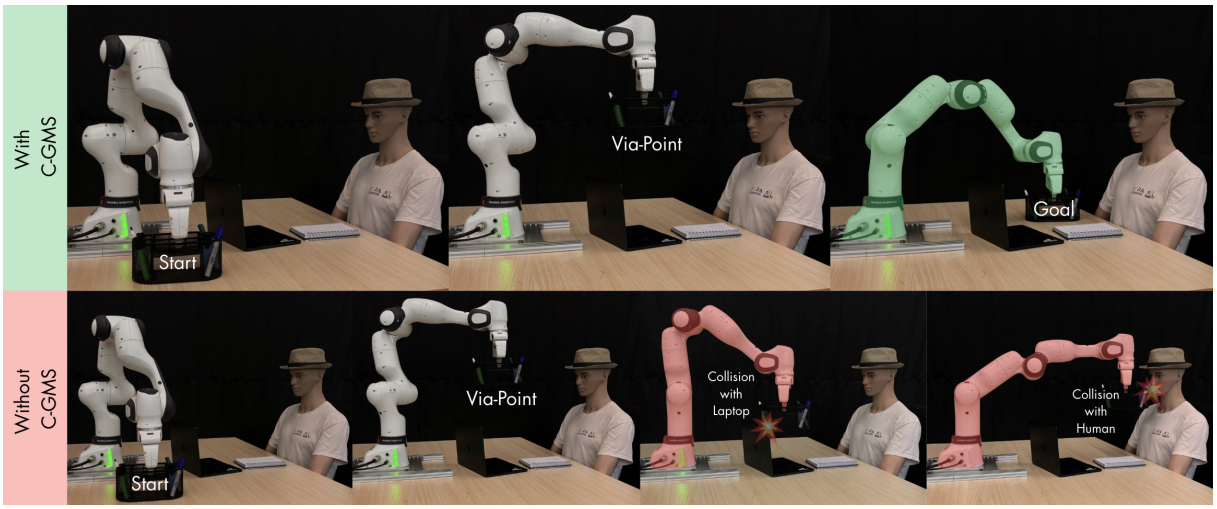


Fig. 1: Comparison of policy execution under certified vs. unconstrained learning. Top: With (proposed) C-GMS, policy sampling is restricted to a certified manifold, ensuring Lyapunov stability and safe execution through a via-point to the goal. Bottom: Without C-GMS, unconstrained sampling may violate the stability conditions, leading to unsafe behaviors, including collisions with the environment/human.

***Abstract*— Reinforcement learning (RL) offers a powerful approach for robots to learn complex, collaborative skills by combining Dynamic Movement Primitives (DMPs) for motion and Variable Impedance Control (VIC) for compliant interaction. However, this model-free paradigm often risks instability and unsafe exploration due to the time-varying nature of impedance gains. This work introduces Certified Gaussian-Manifold Sampling (C-GMS), a novel trajectory-centric RL framework that learns combined DMP and VIC policies while guaranteeing Lyapunov stability and actuator feasibility by construction. Our approach reframes policy exploration as sampling from a mathematically defined manifold of stable gain schedules. This ensures every policy rollout is guaranteed to be stable and physically realizable, thereby eliminating the need for reward penalties or post-hoc validation. Furthermore, we provide a theoretical guarantee that our approach ensures bounded tracking error even in the presence of bounded model errors and deployment-time uncertainties. We demonstrate the effectiveness of C-GMS in simulation and verify its efficacy on a real robot, paving the way for reliable autonomous interaction in complex environments.**

I. INTRODUCTION

The field of robotics is undergoing a fundamental shift, moving from static, repetitive industrial tasks to dynamic, unstructured environments where physical interaction is not only inevitable but essential. To navigate this new paradigm, robots require the ability to adapt learned behaviors to new settings, which is often achieved through task parameterization using representations like Dynamic Movement Primitives (DMPs) [1]. While DMPs excel at generalizing a movement to new endpoints, handling complex motions that require passing through a series of via-points is not trivial. For this, and for safe physical interaction, robots also need the ability to dynamically adjust their stiffness and damping, a capability provided by Variable Impedance Control (VIC).

The complexity of designing time-varying impedance profiles has motivated the adoption of data-driven methods, with Reinforcement Learning (RL) emerging as a powerful approach. In model-based families, such as Differential Dynamic Programming (DDP) [2], Iterative Linear Quadratic Regulator (iLQR) [3], and Model Predictive Control (MPC) [4], a robot's dynamics model is exploited to compute optimal sequences. These methods can incorporate actuator and stability constraints via Control Lyapunov Functions (CLFs), but they are computationally expensive, depend heavily on model accuracy, and typically pre-schedule gains rather than learning them adaptively.

Model-free approaches avoid explicit dynamics modeling, enabling the discovery of complex behaviors in unstructured environments. RL, particularly Path Integral (PI^2) [5], [6], has proven effective for learning intricate motor skills. Sem-

Both authors are with Human-Interactive Robotics Lab, IISc Bangalore.
This work was supported in part by ARTPARK, IISc Bangalore.
Corresponding author: shreyaskumar@iisc.ac.in

Method Family	Stability Aware (VIC)	Actuator-Limit Aware	Handles Non-linear Dynamics	Optimizes Task Cost	Adaptive Gains	Model-Free Learning
iLQR / DDP / Trajectory-MPC	⚠ (local CLF constraints)	⚠ (include as constraint)	⚠ (local linearization)	✓	⚠ (pre-schedule)	✗
CBF / RMP Safety Filters	⚠ (safety invariance, not VIC-specific)	⚠ (included as constraint)	✓	✗	–	–
Energy-Tank / PO-PC	✓	✗ (passivity/power aware)	✓	✗	✓	–
Model-based VIC learning	✗	✗	✓	✓	✓	✗
PI^2 / PI-BB VIC	✗	✗	✓	✓	✓	✓
Our method (PI^2 + C-GMS)	✓	✓	✓	✓	✓	✓

TABLE I: Comparison of method families for learning variable impedance control policies. Our method is the only one to combine model-free learning, gain scheduling, stability certification, and actuator feasibility within a unified optimization framework.

inal work by Buchli et al. [7] pioneered the use of PI^2 to simultaneously learn DMP trajectory parameters and time-varying impedance gains, laying the groundwork for data-driven compliant control. Later, Rey et al. [8] refined this approach by incorporating human demonstrations, improving efficiency. While highly successful at optimizing task costs, these methods, like many in the safe RL literature that use reward penalties [9] or post-hoc safety filters [10], do not explicitly enforce stability during learning, leaving Variable Impedance Control susceptible to instability, as rigorously proven by Kronander and Billard [11].

Alternative approaches have focused on safety through other principles. Passivity-based controllers, such as Energy-Tank methods [12]–[14], guarantee stability by maintaining passivity, but they cannot directly optimize general task costs. Safety Filters, such as Control Barrier Functions (CBFs) [10] and Reactive Motion Planning (RMP) [15], provide generic safety invariance layers, but they are not tailored to the time-varying nature of VIC and are not tightly integrated into policy search. Model-based VIC learning approaches [16] optimize gains directly but remain limited by modeling assumptions and lack guarantees during exploration. As summarized in Table I, current methods lack built-in stability and actuator-limit awareness, posing a significant risk of instability in VIC [11].

In this work, we introduce Certified Gaussian-Manifold Sampling (C-GMS), a novel reinforcement learning framework that unifies model-free policy search with rigorous stability analysis. Unlike prior methods that penalize instability or apply safety filters post hoc, our approach guarantees Lyapunov stability [11] and actuator feasibility by construction. By embedding the stability criterion directly into the RL exploration loop, we restrict policy sampling to a certified manifold of stable gain schedules. Each Gaussian perturbation is analytically sampled from this manifold, ensuring that every single rollout remains stable and physically realizable. This integration eliminates the need for separate safety critics or penalty terms and enables reliable policy optimization in complex tasks. Furthermore, we establish a formal theorem demonstrating that our method not only ensures internal stability but also guarantees uniform ultimate boundedness of the tracking error in presence of model and sensor inaccuracies, thus providing a strong foundation for its applicability in real-world scenarios.

II. PRELIMINARIES

A. Rigid Body Dynamics

We assume that the joint-space rigid-body dynamics of an n-DoF manipulator is given by $\mathbf{M}(\mathbf{q})\,\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q},\dot{\mathbf{q}})\,\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \boldsymbol{\tau}_c + \boldsymbol{\tau}_e$, where $\mathbf{q},\dot{\mathbf{q}},\ddot{\mathbf{q}} \in \mathbb{R}^n$ denote joint position, velocity, and acceleration, $\mathbf{M}(\mathbf{q}) \succ \mathbf{0}$ the inertia matrix, $\mathbf{C}(\mathbf{q},\dot{\mathbf{q}})\,\dot{\mathbf{q}}$ the Coriolis/centrifugal term, $\mathbf{g}(\mathbf{q})$ the torque due to gravity, and $\boldsymbol{\tau}_c, \boldsymbol{\tau}_e \in \mathbb{R}^n$ are the commanded and external torques. In task space with Jacobian $\mathbf{J}(\mathbf{q}) \in \mathbb{R}^{m\times n}$, the operational-space inertia [17] is defined as $\boldsymbol{\Lambda}(\mathbf{q}) = (\mathbf{J}(\mathbf{q})\,\mathbf{M}^{-1}(\mathbf{q})\,\mathbf{J}^\top(\mathbf{q}))^{-1} \in \mathbb{R}^{m\times m}$. Let $\boldsymbol{\mu}(\mathbf{q},\dot{\mathbf{q}}) \in \mathbb{R}^m$ collect Coriolis/centrifugal wrenches and $\mathbf{p}(\mathbf{q}) \in \mathbb{R}^m$ be the gravity wrench mapped to task space. Then the end-effector dynamics reads

$$\boldsymbol{\Lambda}(\mathbf{q})\,\ddot{\mathbf{x}} + \boldsymbol{\mu}(\mathbf{q},\dot{\mathbf{q}}) + \mathbf{p}(\mathbf{q}) = \mathbf{f}_c + \mathbf{f}_e, \tag{1}$$

where $\mathbf{x},\dot{\mathbf{x}},\ddot{\mathbf{x}} \in \mathbb{R}^m$ are the task-space position, velocity, and acceleration, and $\mathbf{f}_c, \mathbf{f}_e \in \mathbb{R}^m$ are the commanded and external wrenches. Torques are obtained via the wrench-torque map $\boldsymbol{\tau}_c = \mathbf{J}^\top(\mathbf{q})\,\mathbf{f}_c$. These relations follow the standard operational-space formulation and fix the notation used throughout.

B. Variable Impedance Control

Let $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}_d$ and $\dot{\tilde{\mathbf{x}}} = \dot{\mathbf{x}} - \dot{\mathbf{x}}_d$ denote the Cartesian error and its velocity, where the reference $\mathbf{x}_d(t)$ is assumed to be twice differentiable. We shape the interaction behavior using time-varying symmetric positive-definite gain schedules $\mathbf{K}(t) = \mathbf{K}^\top(t) \succ \mathbf{0}$, and $\mathbf{D}(t) = \mathbf{D}^\top(t) \succ \mathbf{0}$ The desired task-space inertia is fixed to a constant matrix $\mathbf{H} = \mathbf{H}^\top \succ \mathbf{0}$. Under operational-space inverse-dynamics (OSID), we command the wrench $\mathbf{f}_c$

$$\mathbf{f}_c = \boldsymbol{\Lambda}\ddot{\mathbf{x}}_{\text{cmd}} + \boldsymbol{\mu} + \mathbf{p} + \underbrace{\left(\boldsymbol{\Lambda}\mathbf{H}^{-1} - \mathbf{I}\right)\mathbf{f}_e}_{\text{feedforward term } \mathbf{f}_\text{f}}, \tag{2}$$

where the commanded acceleration is defined as

$$\ddot{\mathbf{x}}_{\text{cmd}} = \ddot{\mathbf{x}}_d - \mathbf{H}^{-1}\left(\mathbf{D}(t)\,\dot{\tilde{\mathbf{x}}} + \mathbf{K}(t)\,\tilde{\mathbf{x}}\right). \tag{3}$$

Substituting (2), (3) into operational-space dynamics (1) and simplifying, the closed-loop error dynamics become

$$\mathbf{H}\,\ddot{\tilde{\mathbf{x}}} + \mathbf{D}(t)\,\dot{\tilde{\mathbf{x}}} + \mathbf{K}(t)\,\tilde{\mathbf{x}} \;=\; \mathbf{f}_e. \tag{4}$$

This matches the classical form of a time-varying impedance behavior with desired inertia $\mathbf{H}$.

C. Stability in VIC

Under dynamic decoupling with a constant desired inertia $\mathbf{H} \succ \mathbf{0}$ and in free space ($\mathbf{f}_e = \mathbf{0}$), Kronander & Billard [11] show that the closed-loop impedance dynamics in (4) are globally uniformly stable if there exists $\alpha > 0$ such that, for all t,

$$\begin{aligned} &\alpha\,\mathbf{H} - \mathbf{D}(t) \;\preccurlyeq\; \mathbf{0} \quad \text{and,} \\ &\dot{\mathbf{K}}(t) + \alpha\,\dot{\mathbf{D}}(t) - 2\alpha\,\mathbf{K}(t) \;\preccurlyeq\; \mathbf{0}. \end{aligned} \tag{5}$$

These are state-independent, pointwise-in-time constraints on the gain schedules that can be constructed offline. In our setting, the external wrench $\mathbf{f}_e$ appears explicitly on the right-hand side of the dynamics, while Eq. (5) certifies the stability of the internal (unforced) system. This ensures that the impedance behavior defined by $(\mathbf{H}, \mathbf{D}(t), \mathbf{K}(t))$ is a dissipative and well-posed map throughout the motion.

D. Policy Improvement with Path Integrals

We consider the problem of minimizing the expected trajectory cost

$$J(\tau) = \Phi\big(\mathbf{x}(T)\big) + \int_0^T \ell\big(\mathbf{x}(t), \mathbf{u}(t), t\big)\,dt,$$

where $\mathbf{x}(t) \in \mathbb{R}^{n_x}$ is the system state, $\mathbf{u}(t) \in \mathbb{R}^{n_u}$ is the control input, $\ell(\cdot)$ is the running cost, and $\Phi(\cdot)$ is the terminal cost. The dynamics are assumed to be control-affine with additive noise:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t) + \mathbf{G}(\mathbf{x}, t)\,\mathbf{u}(t) + \mathbf{G}(\mathbf{x}, t)\,\boldsymbol{\xi}(t),$$

where $\boldsymbol{\xi}(t)$ is zero-mean Gaussian noise. When the noise covariance and control penalty satisfy $\boldsymbol{\Sigma}_\xi = \lambda\,\mathbf{R}^{-1}$, a logarithmic transformation of the value function yields a linear HJB equation. Following established results in stochastic optimal control [5], [6], the solution admits the form

$$\Psi(\mathbf{x}_i, t_i) = \mathbb{E}\left[\exp\left(-\tfrac{1}{\lambda}\left(\Phi(\mathbf{x}_T) + \int_{t_i}^T q(\mathbf{x}(t), t)\,dt\right)\right)\right],$$

where $q(\cdot)$ is the state cost, and the expectation is over stochastic trajectories initiated at $(\mathbf{x}_i, t_i)$.

Thus the stochastic optimal control problem becomes a path-integral estimation problem. Rather than computing the value function explicitly, the optimal control can be written as an expectation over trajectories [5], [6]:

$$\mathbf{u}_{t_i} \;=\; \int P(\tau_i)\,\mathbf{u}(\tau_i)\,d\tau_i, \tag{6}$$

with $\mathbf{u}(\tau_i) = \mathbf{R}^{-1}\mathbf{G}^\top\big(\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^\top\big)^{-1}\big(\mathbf{G}\,\boldsymbol{\xi}_{t_i} - \mathbf{b}_{t_i}\big)$.

Here $P(\tau_i)$ is the probability density of a trajectory segment τ_i starting at $(\mathbf{x}_i, t_i)$, and $\mathbf{b}_{t_i}$ collects drift and cost terms (see [6] for the explicit expression). PI^2 evaluates Eq. (6) by Monte Carlo rollouts, i.e., it approximates the path-integral expectation with a sample-weighted average over trajectories. (cf. Table 1 in [7] for the algorithm).

III. METHOD

We present *Certified Gaussian-Manifold Sampling* (C-GMS), a trajectory-centric reinforcement learning framework for variable impedance control (VIC) that guarantees Lyapunov stability and actuator feasibility during exploration. C-GMS samples policies via Gaussian perturbations in parameter space, but each sample is from a manifold where safety and stability conditions are enforced analytically. This manifold is defined by a time-varying Lyapunov certificate (cf. Eq. (5)), and gain schedules are synthesized using slack variables that satisfy these conditions by construction. As a result, every rollout, regardless of the sampled policy, remains certified–yielding stable, safe, and physically realizable interaction. C-GMS thus combines model-free policy learning with model-based guarantees, eliminating the need for penalty terms, barriers, or post-hoc projection.

A. Trajectory Parametrization via DMPs

PI^2 computes the optimal control update (cf. Eq. (6)) for a system with a parameterized policy $\mathbf{a}_t = \boldsymbol{\Phi}_t\,(\boldsymbol{\theta} + \boldsymbol{\xi}_t)$, where $\boldsymbol{\theta}$ is a learned parameter vector and $\boldsymbol{\xi}_t$ is exploration noise. In C-GMS, Gaussian perturbations are sampled once per episode, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, and applied consistently across time, i.e., $\boldsymbol{\xi}_t = \boldsymbol{\xi}$ for all t. The basis functions are defined by

$$[\Phi(s_t)]_j = \frac{\Psi_j(s_t)}{\sum_{k=1}^M \Psi_k(s_t)}, \quad \Psi_j(s_t) = \exp\Big(-\tfrac{(s_t - c_j)^2}{2\sigma_j^2}\Big),$$

with canonical phase $s_t = 1 - t/\tau$, where M is the number of radial basis functions (RBFs).

To ensure smooth, structured control signals compatible with PI^2, we parametrize the desired trajectory using Dynamic Movement Primitives (DMPs). Let $(\mathbf{x}(t), \dot{\mathbf{x}}(t), \ddot{\mathbf{x}}(t))$ denote the kinematic state in $\mathbb{R}^D$ over the horizon $t \in [0, T]$. The DMP transformation dynamics are defined as

$$\tau^2 m\ddot{\mathbf{x}}(t) = k(\mathbf{g}(t) - \mathbf{x}(t)) - \tau d\dot{\mathbf{x}}(t) + \gamma(t) f_{\text{forcing}}(t),$$

where τ is the temporal scaling factor, and k, d, m are the stiffness, damping, and mass parameters, respectively. The goal trajectory $\mathbf{g}(t)$ may be constant or time-varying, and $\gamma(t)$ is a phase-dependent scaling term. The nonlinear forcing term $f_{\text{forcing}}(t)$ modulates the trajectory to encode complex motion patterns, and is modeled as

$$f_{\text{forcing}}(t) = \Phi_{\text{traj}}(s_t)(\boldsymbol{\theta}_{\text{traj}} + \boldsymbol{\xi}_{\text{traj}}),$$

where $\Phi_{\text{traj}}(s_t)$ is the normalized RBF vector, and $\boldsymbol{\theta}_{\text{traj}} \in \mathbb{R}^{M \times D}$ parametrizes the forcing profile and thus the trajectory. We next describe how C-GMS ensures that all sampled gain schedules $(\mathbf{D}(t), \mathbf{K}(t))$ satisfy stability and torque feasibility throughout the policy search process.

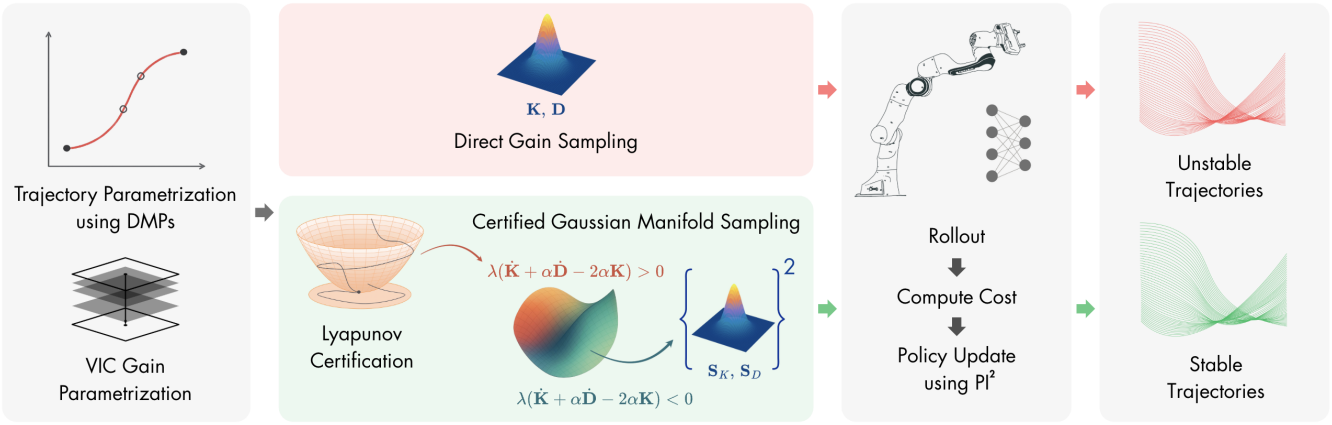


Fig. 2: Overview of the C-GMS framework. Trajectories are parameterized using DMPs, and time-varying VIC gains are parameterized using slacks. In a standard approach, gains are sampled directly from a Gaussian, which can violate Lyapunov stability conditions and lead to unstable rollouts. In contrast, C-GMS enforces stability by sampling from a certified manifold where the Lyapunov condition [11] holds resulting in stable trajectories throughout learning.

B. Gain Parametrization via C-GMS

To extend the PI^2 parametrization beyond trajectories, we introduce an analogous representation for time-varying impedance gains. The objective is to ensure that exploration remains confined to a certified manifold, where every sample satisfies Lyapunov stability and feasibility conditions by construction. Following the conditions for global uniform stability in [11], we enforce the inequalities structurally by introducing matrix-valued slack variables $\mathbf{S}_D(t)$ and $\mathbf{S}_K(t)$:

$$\alpha\,\mathbf{H} - \mathbf{D}(t) = -\mathbf{S}_D(t)\,\mathbf{S}_D^\top(t) \preccurlyeq 0, \tag{7}$$

$$\dot{\mathbf{K}}(t) + \alpha\,\dot{\mathbf{D}}(t) - 2\alpha\,\mathbf{K}(t) = -\mathbf{S}_K(t)\,\mathbf{S}_K^\top(t) \preccurlyeq 0. \tag{8}$$

These slacks are parametrized analogously to the trajectory forcing term using time-varying basis functions:

$$\begin{aligned}\mathrm{vec}_\triangle(\mathbf{S}_D(t)) &= \Phi_\mathrm{D}(s_t)\,(\boldsymbol{\theta}_\mathrm{D} + \boldsymbol{\xi}_\mathrm{K}),\\ \mathrm{vec}_\triangle(\mathbf{S}_K(t)) &= \Phi_\mathrm{K}(s_t)\,(\boldsymbol{\theta}_\mathrm{K} + \boldsymbol{\xi}_\mathrm{D}),\end{aligned} \tag{9}$$

where $\Phi_\mathrm{D}(s_t)$ and $\Phi_\mathrm{K}(s_t)$ share the normalized RBF structure, and the parameter vectors $\boldsymbol{\theta}_\mathrm{D}, \boldsymbol{\theta}_\mathrm{K} \in \mathbb{R}^{M\times D}$ define the gain profiles.

To ensure $\mathbf{K}(t)$ remains symmetric positive definite (SPD), we evolve its Cholesky factor $\mathbf{Q}(t)$:

$$\begin{aligned}\mathbf{B}(t) &:= -\alpha\dot{\mathbf{D}}(t) - \mathbf{S}_K(t)\,\mathbf{S}_K^\top(t),\\ \dot{\mathbf{Q}}(t) &= \alpha\,\mathbf{Q}(t) + \tfrac{1}{2}\,\mathbf{Q}^{-\top}(t)\,\mathbf{B}(t),\\ \mathbf{K}(t) &= \mathbf{Q}^\top(t)\,\mathbf{Q}(t),\end{aligned} \tag{10}$$

with $\mathbf{Q}(0)$ initialized such that $\mathbf{K}(0) \succ 0$. Differentiating $\mathbf{K} = \mathbf{Q}^\top\mathbf{Q}$ and substituting (10) yields $\dot{\mathbf{K}}(t) = 2\alpha\,\mathbf{K}(t) + \mathbf{B}(t)$, ensuring that $\dot{\mathbf{K}} + \alpha\dot{\mathbf{D}} - 2\alpha\mathbf{K} = -\mathbf{S}_K\mathbf{S}_K^\top \preccurlyeq 0$ holds identically while maintaining $\mathbf{K}(t) \succ 0$ for all t.

This construction confines policy search to a certified manifold of gain schedules that provably satisfy the Lyapunov conditions, eliminating the need for penalties, constraints, or post-hoc projections.

C. Certificate-Aware Gain Contraction

Leveraging C-GMS's slack-based gain synthesis, which admits a certificate preserving contraction, we impose actuator-limit awareness directly in gain space–without auxiliary constraints, projections, or loss of stability. This is achieved by introducing a torque governor that uniformly scales the slack variables by $\sqrt{\beta} \in [0,1]$: $\mathbf{S}_D^\beta(t) = \sqrt{\beta}\,\mathbf{S}_D(t)$, $\mathbf{S}_K^\beta(t) = \sqrt{\beta}\,\mathbf{S}_K(t)$, resulting in scaled gains

$$\begin{aligned}&\mathbf{D}^\beta(t) = \alpha\,\mathbf{H} + \beta\,\mathbf{S}_D(t)\,\mathbf{S}_D^\top(t),\\ &\dot{\mathbf{K}}^\beta(t) + \alpha\,\dot{\mathbf{D}}^\beta(t) - 2\alpha\,\mathbf{K}^\beta(t) = -\beta\,\mathbf{S}_K(t)\,\mathbf{S}_K^\top(t).\end{aligned} \tag{11}$$

The gains $\mathbf{K}^\beta(t)$ and $\mathbf{D}^\beta(t)$ remain certificate-compliant for all $\beta \in [0,1]$, as they preserve the structure required by the Lyapunov inequalities. Since the control law is affine in both $\mathbf{K}$ and $\mathbf{D}$, the overall torque command is also affine in β. Denoting the β-independent component as $\boldsymbol{\tau}_0$ and the β-dependent component as $\boldsymbol{\tau}_1$, the control input takes the form:

$$\boldsymbol{\tau}(\beta) = \boldsymbol{\tau}_0 + \beta\,\boldsymbol{\tau}_1,$$

the maximum admissible scaling factor $\beta^\star$ under actuator box constraints $\boldsymbol{\tau}_\mathrm{min} \le \boldsymbol{\tau}(\beta) \le \boldsymbol{\tau}_\mathrm{max}$ is

$$\beta^\star = \min_i \begin{cases} \frac{\tau_{\mathrm{max},i} - \tau_{0,i}}{\tau_{1,i}}, & \tau_{1,i} > 0,\\ \frac{\tau_{\mathrm{min},i} - \tau_{0,i}}{\tau_{1,i}}, & \tau_{1,i} < 0,\end{cases} \in [0,1].$$

The controller is executed with $\mathbf{S}_{(\cdot)}^{\beta^\star}$ and corresponding gains. This actuator-limit governor preserves the Lyapunov certificate pointwise, enabling safe execution under torque saturation.

D. Convergence Guarantee

We analyze the closed-loop stability of the VIC under time-varying gains synthesized via the C-GMS framework. Specifically, we aim to show that the tracking error remains bounded, even in the presence of model mismatch and other deployment-time uncertainties.

At each time step t, the desired trajectory $(\mathbf{x}_d(t), \dot{\mathbf{x}}_d(t), \ddot{\mathbf{x}}_d(t))$ is generated by evaluating the DMP forcing function $f_\mathrm{forcing}(t) = \Phi_\mathrm{traj}(s_t)\,\boldsymbol{\theta}_\mathrm{traj}$, followed by integration of the DMP dynamics. Simultaneously, the time-varying gains are synthesized via basis expansions $\Phi_D(s_t)$, $\Phi_K(s_t)$, which define slack variables

$\mathbf{S}_D^{\{\cdot\}}(t), \mathbf{S}_K^{\{\cdot\}}(t)$. These are converted into damping and stiffness gains $\mathbf{D}(t)$, $\mathbf{K}(t)$ using equations (7) and (10), and used in the OSID control law (2)-(3).

The deployment of a trained policy in real-world settings brings along challenges such as plant-model mismatch, feedforward wrench $\mathbf{f}_\text{f}$ errors, and sensor noise, all of which may induce a residual input $\mathbf{u}_\text{res}(t)$ into the closed-loop system. To this end, we establish that the gains generated by C-GMS ensure a bounded tracking error under these conditions.

Theorem. Consider the perturbed dynamics under VIC with model error and feedforward uncertainty aggregated as a bounded residual input $\|\mathbf{u}_\text{res}(t)\| \leq \bar{u} < \infty$. If the gains $\mathbf{D}(t)$, $\mathbf{K}(t)$ satisfy the differential Lyapunov conditions (17) with strict margins $(\varepsilon_D, \varepsilon_K) > 0$, then the tracking error $(\tilde{\mathbf{x}}(t), \dot{\tilde{\mathbf{x}}}(t))$ is uniformly ultimately bounded, with an ultimate error radius $O(\bar{u}_\text{res})$ that decreases as $(\varepsilon_D, \varepsilon_K)$ increase. Proof in Appendix A. □

IV. EXPERIMENTS

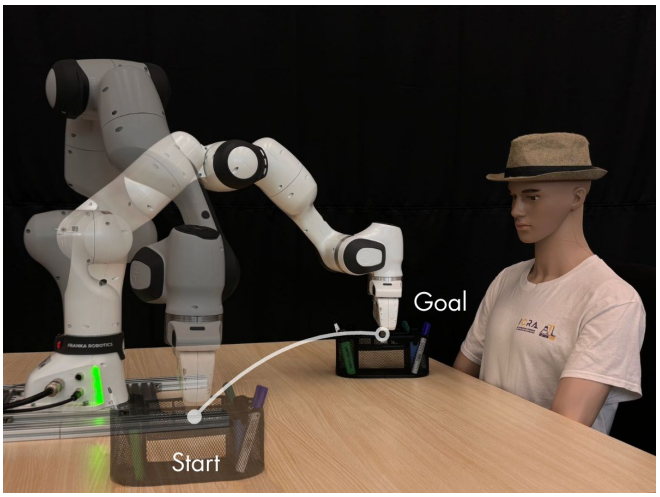
A. Experimental Setup

We consider a collaborative human-robot handover task in which a 7-DoF Franka Research 3 (FR3) manipulator transfers a stationery organizer to a seated human participant. The motivation for the task arises from an everyday scenario: the human is engaged with a notebook and requires a pen, prompting the robot to initiate a handover. The motion is executed under real-time VIC at 1 kHz in the task space, with feedback from the robot's internal force-torque sensors.

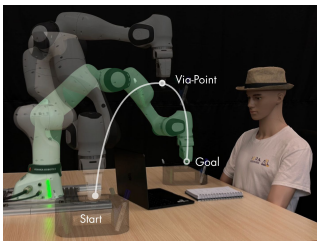
In the initial phase (Fig. 3a), the robot learns a smooth reaching motion from a starting pose $x_\text{start} = [0.55,\ 0.00,\ 0.11]\,\text{m}$ to a final handover location $x_\text{goal} = [0.05,\ 0.72,\ 0.11]\,\text{m}$, following a minimum-jerk trajectory in task space over a 10 second horizon. This motion serves as the nominal demonstration and establishes a baseline for downstream optimization. To evaluate the system's adaptability, an obstacle is introduced along the nominal path, partially occluding the direct line between the start and goal. The robot must now adjust both its trajectory and impedance behavior to complete the handover safely. The adapted motion bends around the obstruction by passing through an intermediate region centered near $[0.30,\ 0.48,\ 0.40]\,\text{m}$ (the obstacle's coarse pose and was specified a priori, in practice they can be obtained automatically via open-vocabulary perception pipelines [18]). Fig. 3b shows the physical configuration of the task space, including the obstacle location and handover geometry. We further evaluate generalization by executing five diverse free-space via-point policies under a consistent task setup, with quantitative results in Table II.

B. Policy Representation

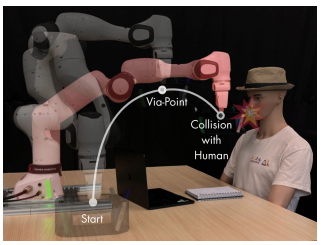
We adopt the PI^2 as policy improvement framework (cf. §II-D) to jointly optimize the task-space trajectory and the time-varying Cartesian impedance gains. The policy is parameterized by a vector $\boldsymbol{\theta}$, consisting of DMP weights for the reference trajectory as well as slack parameters that indirectly control the stiffness and damping schedules as detailed in §III.



(a) The robot learns to move from a start pose to a predefined goal near the human's hand, initially following a minimum-jerk trajectory.



(b) A learned policy under C-GMS introduces a via-point to avoid the obstacle while ensuring stable, compliant behavior resulting in a collision-free trajectory.



(c) A policy learned without C-GMS violates the Lyapunov condition, resulting in unsafe trajectories and potential collision with the environment or human.

Fig. 3: Experimental setup for the human-robot collaborative task.

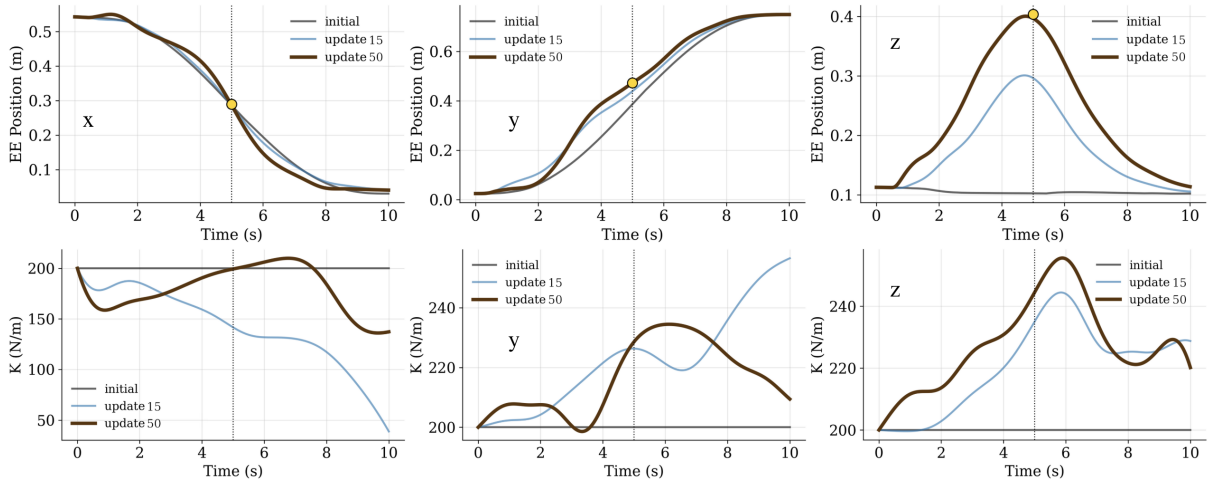
The initial parameters for the trajectory are obtained by training the DMP to a minimum-jerk trajectory connecting x_start and x_goal. The slack parameters are initialized such that the resulting stiffness matrix $\mathbf{K}(t)$ is constant and isotropic with magnitude $200\,\text{N/m}$ along each axis. Damping is initialized at a constant value of $30\,\text{Ns/m}$, also uniformly across axes. Complete parameter settings, including DMP bases, slack initializations, and cost weights are provided in Appendix B for reproducibility.

C. Optimization Objective

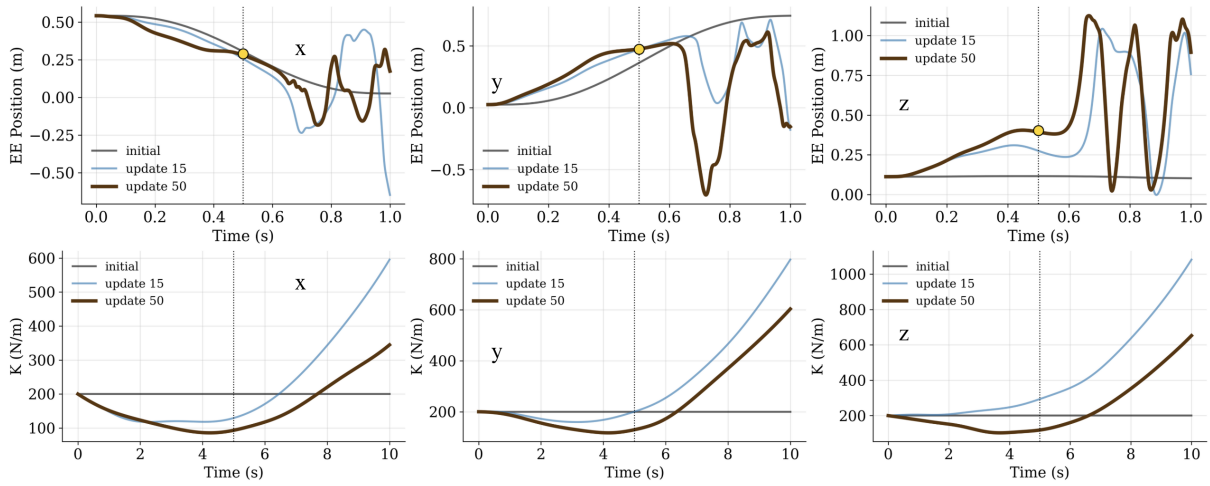
We formulated the cost function to primarily support via-point tracking. To achieve this, the objective function explicitly emphasizes tracking accuracy near the via-point, while also regularizing stiffness magnitude and motion smoothness. The total cost over a trajectory of length T is:

$$J = \sum_{t=1}^{T} \left[\lambda_K\ \text{tr}(\mathbf{K}_t) + \lambda_\text{acc}\|\ddot{\mathbf{x}}_t\|^2 + w_\text{via}(t)\,\|\mathbf{x}_t - \mathbf{x}_{\text{ref},\,t}\|^2\right],$$

- $\text{tr}(\mathbf{K}_t)$ penalizes the magnitude of the task-space stiffness matrix at time t, encouraging compliant behavior.
- $\|\ddot{\mathbf{x}}_t\|^2$ regularizes acceleration, promoting smooth motion.
- The tracking error term $\|\mathbf{x}_t - \mathbf{x}_{\text{ref},t}\|^2$ is scaled by a time-varying weight $w_\text{via}(t)$, which intensifies the penalty near the via-point.



(a) Under full C-GMS, policies remain stable and smooth throughout learning, with impedance gains adapting locally around the via-point



(b) When C-GMS is applied only until the via-point (i.e., stability constraints are disabled afterward), the policy continues to optimize task cost but exhibits unstable behavior beyond the via-point, including oscillations and unbounded gain growth.

Fig. 4: VIC gain schedules and corresponding end-effector trajectories of the robot initially, after 15 updates and after 50 updates. The policy obtained after the 50^{th} update was executed on hardware (cf. § III-D). Via-points are marked by circles.

The weighting function is given by $w_{\text{via}}(t) = w_0 + \gamma \cdot g(t)$ where $g(t)$ is a Gaussian kernel centered at the via-point time $\hat{t}$.

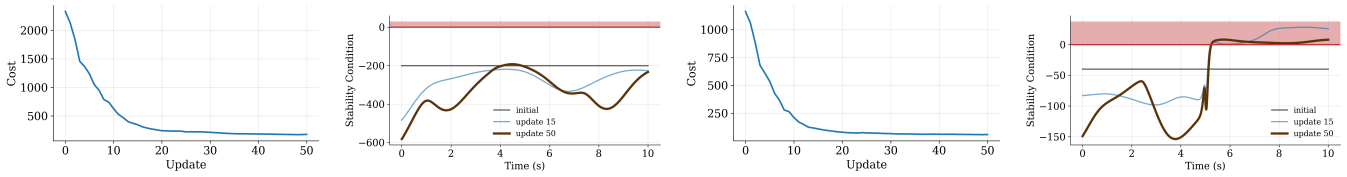
Learning is performed over 50 policy updates, each with 12 sampled rollouts in MuJoCo, a physics engine for simulation using the FR3 model. At every iteration, the PI^2 update rule is applied to the parameter vector $\boldsymbol{\theta}$. All sampling is constrained to a certified manifold that satisfies the time-varying stability condition (cf. Eq. (5)). This ensures that all explored policies yield stabilizing behavior under the VIC dynamics in Eqs. (1)-(3).

D. Results and Analysis

Figure 3 demonstrates the task execution on real hardware. With certification (Fig. 3b), the learned policy follows a stable trajectory through the via-point and completes the handover safely. Without certification (Fig. 3c), the learned policy violates the Lyapunov condition and produces unstable trajectories eading to unsafe behaviors, including collisions with the environment/human. Figure 4 compares the evolution of end-effector trajectories and stiffness profiles under the two regimes. With C-GMS (Fig. 4a), trajectories converge smoothly and stiffness rises near the via-point before relaxing, aligning naturally with task demands. Without certification (Fig. 4b), trajectories become oscillatory and gains

Scenario	RMSE x	RMSE y	RMSE z	Sat. w/ Gov	Sat. w/o Gov
S1	2e−2	53e−4	54e−4	✗	✗
S2	38e−4	35e−4	76e−4	✗	✓
S3	11e−3	22e−3	3e−2	✗	✗
S4	27e−4	19e−4	28e−4	✗	✓
S5	23e−3	51e−4	1e−2	✗	✓

TABLE II: Hardware metrics across five unique scenarios [Start-Via-End]:
S1: $[0.30, 0.00, 0.47] - [0.42, 0.30, 0.34] - [0.54, 0.43, 0.47]$,
S2: $[0.37, -0.34, 0.03] - [0.62, 0.00, 0.32] - [0.45, 0.27, 0.06]$,
S3: $[0.40, 0.00, 0.15] - [0.32, 0.50, 0.42] - [0.00, 0.40, 0.10]$,
S4: $[0.20, 0.17, 0.43] - [0.34, 0.20, 0.36] - [0.48, 0.34, 0.0.43]$,
S5: $[0.58, -0.35, 0.18] - [0.31, 0.00, 0.43] - [0.00, 0.56, 0.05]$.
RMSE-x, y, z denotes the end-effector tracking error (in m) in task space. The final two columns indicate whether torque saturation was observed during execution, with and without the actuator-limit governor. Note that with governor, saturation was not reached for any segment.
Note: To evaluate the efficacy of § III-C, we virtually reduced FR3's default torque limits ([87, 87, 87, 87, 12, 12, 12] Nm) by half, emulating deployment on lower-capacity hardware. The proposed governor maintains certification without exceeding these reduced limits.



(a) Learning Curve (C-GMS) (b) Stability Condition (C-GMS) (c) Learning Curve (no C-GMS) (d) Stability Condition (no C-GMS)

Fig. 5: Learning curve and eigenvalue evolution for Eq. (5). Under C-GMS based sampling, the eigenvalues remain negative, ensuring that the impedance profile guarantees stable control. However, when sampling shifts to an unsafe region, the cost function may still converge (since the via-point is reached prior to C-GMS being disabled), but the eigenvalues become positive, potentially leading to severe instability in the end-effector trajectory.

diverge, despite reductions in cost. Figure 5 summarizes optimization across both cases. Under C-GMS (Figs. 5a, 5b), the cost decreases steadily while the minimum eigenvalue of Eq. (5) remains strictly negative, certifying stability at every iteration. Without certification (Figs. 5c, 5d), even though the cost converges, the eigenvalue crosses into the positive domain, indicating loss of guaranteed stability. This explains the unsafe executions in Fig. 4b. Table II summarizes performance of trained policy on hardware across five unique segments apart from the aforementioned collaborative task.

Together, these results show that while unconstrained learning can converge in terms of cost, it may yield unstable and unsafe policies. By embedding stability constraints directly into the sampling process, not only policies are guaranteed to be stable, but are also physically realizable on hardware, enabling the safe discovery of variable impedance policies.

V. CONCLUSION AND LIMITATIONS

We have introduced Certified Gaussian-Manifold Sampling (C-GMS), a novel framework that demonstrates how stable and optimal policy optimization can be achieved for variable impedance control by constraining the reinforcement learning exploration to a certified manifold. By leveraging the analytically verifiable stability criterion, our approach guarantees that every policy rollout is Lyapunov stable and physically realizable. The results highlight a critical distinction: even with identical cost shaping and initialization, an unconstrained optimizer continues to reduce task cost but can produce unsafe gain schedules that lead to erratic and unstable robot execution. In contrast, C-GMS ensures the policy converges smoothly while preserving formal safety guarantees at each update. Our experiments further validate that C-GMS produces compliant trajectories that respect task constraints, showcasing its practical viability for safe autonomous interaction. The integration of certificate-aware actuator-limit governor presents a robust foundation for a physically realizable and safe learning system.

A primary limitation of our current framework is its reliance on the stability criterion [11], which is formulated for free-space dynamics and does not account for external contact. This prevents its direct application to contact-rich tasks, a critical domain for compliant robotics. Additionally, the current cost function focuses solely on via-point tracking and lacks terms for orientation control, which is necessary for more realistic manipulation tasks. Despite these limitations, our work opens several promising avenues for future research. We plan to extend this analysis to explore broader task families, including contact-rich and orientation-sensitive interactions. Future work will also explore how this governor can be extended to dynamically adapt to varying payload conditions.

APPENDIX

A. Robustness to Modeling Error

During execution with OSID and wrench feedforward, the tracking error $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}_d$ evolves as

$$\mathbf{H}\,\ddot{\tilde{\mathbf{x}}}(t) + \mathbf{D}(t)\,\dot{\tilde{\mathbf{x}}}(t) + \mathbf{K}(t)\,\tilde{\mathbf{x}}(t) = \mathbf{u}_{\text{res}}(t), \tag{12}$$

where the residual input $\mathbf{u}_{\text{res}}(t)$ collects plant–model mismatch and feedforward/sensing imperfections and is bounded: $\|\mathbf{u}_{\text{res}}(t)\| \le \bar{u} < \infty$. Gains are synthesized by C-GMS from slacks with strict margins $\varepsilon_D, \varepsilon_K > 0$ so that for all t,

$$\alpha\,\mathbf{H} - \mathbf{D}(t) \preccurlyeq -\varepsilon_D\,\mathbf{I}, \tag{13}$$

$$\dot{\mathbf{K}}(t) + \alpha\,\dot{\mathbf{D}}(t) - 2\alpha\,\mathbf{K}(t) \preccurlyeq -\varepsilon_K\,\mathbf{I}. \tag{14}$$

C-GMS guarantees $\mathbf{K}(t) = \mathbf{Q}^\top(t)\mathbf{Q}(t) \succ 0$ and $\mathbf{D}, \mathbf{K}$ are continuous. On any horizon $[0, T]$, continuity implies uniform bounds

$$h_{\min}\mathbf{I} \preccurlyeq \mathbf{H} \preccurlyeq h_{\max}\mathbf{I}, \qquad 0 < \underline{k}\mathbf{I} \preccurlyeq \mathbf{K}(t) \preccurlyeq \bar{k}\mathbf{I}.$$

Assumption: $\mathbf{D}(t)$ is uniformly bounded, $\|\mathbf{D}(t)\| \le \bar{d} < \infty$ (true if $\mathbf{D}$ is continuous and its generating slacks are bounded).

Consider the standard energy

$$V(t) = \tfrac{1}{2}\,\dot{\tilde{\mathbf{x}}}^\top\mathbf{H}\,\dot{\tilde{\mathbf{x}}} + \tfrac{1}{2}\,\tilde{\mathbf{x}}^\top\mathbf{K}(t)\,\tilde{\mathbf{x}}. \tag{15}$$

Because $\mathbf{H} \succ 0$ and $\mathbf{K}(t) \succ 0$, V is positive definite; there exist $m_1, m_2 > 0$ such that

$$m_1\|z\|^2 \le V(t) \le m_2\|z\|^2, \qquad z := \begin{bmatrix}\dot{\tilde{\mathbf{x}}} \\ \tilde{\mathbf{x}}\end{bmatrix}. \tag{16}$$

Differentiating (15) along (12) and using $\tilde{\mathbf{x}}^\top\mathbf{K}\dot{\tilde{\mathbf{x}}} = \dot{\tilde{\mathbf{x}}}^\top\mathbf{K}\tilde{\mathbf{x}}$ gives

$$\dot{V} = -\,\dot{\tilde{\mathbf{x}}}^\top\mathbf{D}\,\dot{\tilde{\mathbf{x}}} + \tfrac{1}{2}\,\tilde{\mathbf{x}}^\top\dot{\mathbf{K}}\,\tilde{\mathbf{x}} + \dot{\tilde{\mathbf{x}}}^\top\mathbf{u}_{\text{res}}. \tag{17}$$

From (13), $\mathbf{D} \succeq \alpha\mathbf{H} + \varepsilon_D\mathbf{I}$, hence

$$-\,\dot{\tilde{\mathbf{x}}}^\top\mathbf{D}\,\dot{\tilde{\mathbf{x}}} \le -\,\alpha\,\dot{\tilde{\mathbf{x}}}^\top\mathbf{H}\,\dot{\tilde{\mathbf{x}}} - \varepsilon_D\,\|\dot{\tilde{\mathbf{x}}}\|^2. \tag{18}$$

From (14),

$$\tfrac{1}{2}\,\tilde{\mathbf{x}}^\top\dot{\mathbf{K}}\,\tilde{\mathbf{x}} \le \alpha\,\tilde{\mathbf{x}}^\top\mathbf{K}\,\tilde{\mathbf{x}} - \tfrac{\alpha}{2}\,\tilde{\mathbf{x}}^\top\dot{\mathbf{D}}\,\tilde{\mathbf{x}} - \tfrac{\varepsilon_K}{2}\,\|\tilde{\mathbf{x}}\|^2. \tag{19}$$

Using $\frac{\alpha}{2}\tilde{\mathbf{x}}^\top\dot{\mathbf{D}}\,\tilde{\mathbf{x}} = \frac{d}{dt}\big(\frac{\alpha}{2}\tilde{\mathbf{x}}^\top\mathbf{D}\tilde{\mathbf{x}}\big) - \alpha\,\tilde{\mathbf{x}}^\top\mathbf{D}\dot{\tilde{\mathbf{x}}}$, combine (17)-(19) to obtain

$$\dot{V} \le -\,\alpha\,\dot{\tilde{\mathbf{x}}}^\top\mathbf{H}\,\dot{\tilde{\mathbf{x}}} - \varepsilon_D\,\|\dot{\tilde{\mathbf{x}}}\|^2 + \alpha\,\tilde{\mathbf{x}}^\top\mathbf{K}\,\tilde{\mathbf{x}} - \tfrac{d}{dt}\Big(\tfrac{\alpha}{2}\tilde{\mathbf{x}}^\top\mathbf{D}\tilde{\mathbf{x}}\Big) + \alpha\,\tilde{\mathbf{x}}^\top\mathbf{D}\,\dot{\tilde{\mathbf{x}}} - \tfrac{\varepsilon_K}{2}\,\|\tilde{\mathbf{x}}\|^2 + \dot{\tilde{\mathbf{x}}}^\top\mathbf{u}_{\text{res}}. \tag{20}$$

Define the augmented storage $\mathcal{V} := V + \frac{\alpha}{2}\,\tilde{\mathbf{x}}^\top\mathbf{D}\tilde{\mathbf{x}}$. Then

$$\dot{\mathcal{V}} \le -\,\alpha\,\dot{\tilde{\mathbf{x}}}^\top\mathbf{H}\,\dot{\tilde{\mathbf{x}}} - \varepsilon_D\,\|\dot{\tilde{\mathbf{x}}}\|^2 - \tfrac{\varepsilon_K}{2}\,\|\tilde{\mathbf{x}}\|^2 + \alpha\,\tilde{\mathbf{x}}^\top\mathbf{K}\,\tilde{\mathbf{x}} \tag{21}$$

$$+\,\alpha\,\tilde{\mathbf{x}}^\top\mathbf{D}\,\dot{\tilde{\mathbf{x}}} + \dot{\tilde{\mathbf{x}}}^\top\mathbf{u}_{\text{res}}. \tag{22}$$

Use $\tilde{\mathbf{x}}^\top\mathbf{K}\tilde{\mathbf{x}} \le \bar{k}\,\|\tilde{\mathbf{x}}\|^2$, $\dot{\tilde{\mathbf{x}}}^\top\mathbf{H}\dot{\tilde{\mathbf{x}}} \ge h_{\min}\|\dot{\tilde{\mathbf{x}}}\|^2$, $\|\mathbf{D}(t)\| \le \bar{d}$, Young's inequality $\alpha\,\tilde{\mathbf{x}}^\top\mathbf{D}\dot{\tilde{\mathbf{x}}} \le \frac{\gamma}{2}\|\dot{\tilde{\mathbf{x}}}\|^2 + \frac{\alpha^2\bar{d}^2}{2\gamma}\|\tilde{\mathbf{x}}\|^2$, and $\dot{\tilde{\mathbf{x}}}^\top\mathbf{u}_{\text{res}} \le \eta\|\dot{\tilde{\mathbf{x}}}\|^2 + \frac{1}{4\eta}\|\mathbf{u}_{\text{res}}\|^2$ for any $\gamma \in (0, \varepsilon_D)$ and $\eta \in (0, \varepsilon_D - \gamma)$. Then

$$\dot{\mathcal{V}} \le -\big(\alpha h_{\min} + \varepsilon_D - \gamma - \eta\big)\,\|\dot{\tilde{\mathbf{x}}}\|^2 - \Big(\tfrac{\varepsilon_K}{2} - \alpha\bar{k} - \tfrac{\alpha^2\bar{d}^2}{2\gamma}\Big)\,\|\tilde{\mathbf{x}}\|^2 + \tfrac{1}{4\eta}\,\|\mathbf{u}_{\text{res}}(t)\|^2. \tag{23}$$

Choose margins so that $\varepsilon_K > 2\alpha\bar{k} + \frac{\alpha^2\bar{d}^2}{\gamma}$, and fix any $\gamma \in (0, \varepsilon_D)$, $\eta \in (0, \varepsilon_D - \gamma)$. Let

$$c_1 := \min\Big\{\alpha h_{\min} + \varepsilon_D - \gamma - \eta,\ \tfrac{\varepsilon_K}{2} - \alpha\bar{k} - \tfrac{\alpha^2\bar{d}^2}{2\gamma}\Big\} > 0,$$

$$c_2 := \tfrac{1}{4\eta} > 0.$$

Then (23) becomes

$$\dot{\mathcal{V}} \le -\,c_1\,\|z\|^2 + c_2\,\|\mathbf{u}_{\text{res}}(t)\|^2. \tag{24}$$

Because $\mathbf{K} \succ 0$ and $\mathbf{D} \succeq \alpha\mathbf{H} + \varepsilon_D\mathbf{I}$, $\mathcal{V}$ is positive definite and quadratically bounded with respect to z. Explicitly,

$$m_1'\|z\|^2 \le \mathcal{V}(t) \le m_2'\|z\|^2,$$
$$m_1' = \tfrac{1}{2}\min\{h_{\min},\ \underline{k} + \alpha\varepsilon_D\}$$
$$m_2' = \tfrac{1}{2}\max\{h_{\max},\ \bar{k} + \alpha\bar{d}\}. \tag{25}$$

By a comparison lemma applied to (24), for all $t \ge t_0$,

$$\|z(t)\|^2 \le \exp\Big(-\tfrac{c_1}{m_2'}(t - t_0)\Big)\,\tfrac{\mathcal{V}(t_0)}{m_1'} + \tfrac{m_2'}{m_1'}\,\tfrac{c_2}{c_1}\,\|\mathbf{u}_{\text{res}}\|_\infty^2. \tag{26}$$

Thus, $(\tilde{\mathbf{x}}, \dot{\tilde{\mathbf{x}}})$ is uniformly ultimately bounded (input-to-state practically stable), with ultimate radius $O(\|\mathbf{u}_{\text{res}}\|_\infty)$ that shrinks as the strict margins $(\varepsilon_D, \varepsilon_K)$ grow. □

B. Hyperparameters and Reproducibility

Table III lists all key hyperparameters used in the experiments. All values were held constant across trials and across all variants. No manual tuning was performed post-deployment. Code and configuration files are available at: https://github.com/shr-eyas/safe-vic

TABLE III: Key hyperparameters used in all experiments.

Parameter	Value
Time step dt	0.001 s
Certificate scaling α	0.05
Task matrix H	$\mathbf{I}_{3\times3}$
RBF count (DMP)	51
RBF count (slacks)	7
RBF intersection height (DMP)	0.95
RBF intersection height (slacks)	0.7
RBF regularization	1e−6
PI^2 softmax sharpness β	20
Covariance decay (EMA)	0.98
Cost coefficient: λ_K	15e−7
Cost coefficient: λ_{acc}	1e−3
Cost coefficient: w_0	0.2
Cost coefficient: γ	5e4
Trajectory noise σ_{traj}	8.0
Stiffness noise σ_K	1.3
Damping noise σ_D	0.6

References

[1] M. Saveriano, F. J. Abu-Dakka, A. Kramberger, and L. Peternel, "Dynamic movement primitives in robotics: A tutorial survey," *The International Journal of Robotics Research*, vol. 42, no. 13, pp. 1133–1184, 2023.

[2] D. Q. Mayne, "Differential dynamic programming–a unified approach to the optimization of dynamic systems," in *Control and dynamic systems*, vol. 10, pp. 179–254, Elsevier, 1973.

[3] W. Li and E. Todorov, "Iterative linear quadratic regulator design for nonlinear biological movement systems," in *First International Conference on Informatics in Control, Automation and Robotics*, vol. 2, pp. 222–229, SciTePress, 2004.

[4] J. B. Rawlings, D. Q. Mayne, M. Diehl, *et al.*, *Model predictive control: theory, computation, and design*, vol. 2. Nob Hill Publishing Madison, WI, 2020.

[5] H. J. Kappen, "Linear theory for control of nonlinear stochastic systems," *Phys. Rev. Lett.*, vol. 95, p. 200201, Nov 2005.

[6] E. Theodorou, J. Buchli, and S. Schaal, "A generalized path integral control approach to reinforcement learning," *The Journal of Machine Learning Research*, vol. 11, pp. 3137–3181, 2010.

[7] J. Buchli, E. Theodorou, F. Stulp, and S. Schaal, "Variable impedance control a reinforcement learning approach," *Robotics: Science and Systems VI*, vol. 153, 2011.

[8] J. Rey, K. Kronander, F. Farshidian, J. Buchli, and A. Billard, "Learning motions from demonstrations and rewards with time-invariant dynamical systems based policies," *Autonomous Robots*, vol. 42, no. 1, pp. 45–64, 2018.

[9] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International conference on machine learning*, pp. 22–31, PMLR, 2017.

[10] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada, "Control barrier functions: Theory and applications," in *2019 18th European Control Conference (ECC)*, pp. 3420–3431, 2019.

[11] K. Kronander and A. Billard, "Stability considerations for variable impedance control," *IEEE Transactions on Robotics*, vol. 32, no. 5, pp. 1298–1305, 2016.

[12] B. Hannaford and J.-H. Ryu, "Time-domain passivity control of haptic interfaces," *IEEE Transactions on Robotics and Automation*, vol. 18, no. 1, pp. 1–10, 2002.

[13] F. Ferraguti, C. Secchi, and C. Fantuzzi, "A tank-based approach to impedance control with variable stiffness," in *2013 IEEE International Conference on Robotics and Automation*, pp. 4948–4953, 2013.

[14] C. Schindlbeck and S. Haddadin, "Unified passivity-based cartesian force/impedance control for rigid and flexible joint robots via task-energy tanks," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 440–447, 2015.

[15] C.-A. Cheng, M. Mukadam, J. Issac, S. Birchfield, D. Fox, B. Boots, and N. Ratliff, "Rmpflow: A computational graph for automatic motion policy generation," 2019.

[16] 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, 2023.

[17] O. Khatib, "A unified approach for motion and force control of robot manipulators: The operational space formulation," *IEEE Journal on Robotics and Automation*, vol. 3, no. 1, pp. 43–53, 2003.

[18] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.