Full-Order Sampling-Based MPC for Torque-Level Locomotion Control via Diffusion-Style Annealing

Haoru Xue*, Chaoyi Pan*, Zeji Yi, Guannan Qu, Guanya Shi

Carnegie Mellon University Robotics Institute
{haorux, chaoyip, zejiy, gqu, guanyas}@andrew.cmu.edu

Abstract: Due to high dimensionality and non-convexity, real-time optimal control using full-order dynamics models for legged robots is challenging. Therefore, Nonlinear Model Predictive Control (NMPC) approaches are often limited to reduced-order models or local approximations. Sampling-based MPC has shown potential in nonconvex even discontinuous problems, but often yields suboptimal solutions with high variance, which limits its applications in high-dimensional locomotion. This work introduces DIAL-MPC (Diffusion-Inspired Annealing for Legged MPC), a sampling-based MPC framework with a novel diffusion-style annealing process. Such a process is supported by the theoretical landscape analysis of Model Predictive Path Integral Control (MPPI) and the connection between MPPI and single-step diffusion. Algorithmically, DIAL-MPC iteratively refines solutions online and achieves both global coverage and local convergence. In quadrupedal torque-level control tasks, DIAL-MPC reduces the tracking error of standard MPPI by 13.4 times and outperforms reinforcement learning (RL) policies by 50% in challenging climbing tasks *without any training*. In particular, DIAL-MPC enables precise real-world quadrupedal jumping with payload. To the best of our knowledge, DIAL-MPC is the first *training-free* method that optimizes over full-order quadruped dynamics in real-time. [1]

Keywords: Model Predictive Control, Sampling-based Optimization, Optimal Control, Reinforcement Learning, Whole-Body Control, Legged Locomotion

1 Introduction

Legged robots have demonstrated great potential in navigating through complex environments thanks to their agility and mobility [29, 20, 15, 19, 21, 27]. However, the online control of articulated legged systems remains challenging because of their high-dimensional, underactuated and contact-rich nature, leads to non-convex and non-smooth optimization landscapes.

Reinforcement learning (RL) has become a popular approach for learning control policies for legged robots, thanks to its ease of implementation and strong performance in contact-rich problems [31, 12, 5, 25, 32, 4, 9]. However, RL suffers from time-consuming training and tedious tuning, and the resulting policies heavily depend on the training setup, limiting their test-time generalization to unseen tasks and environments. In the meanwhile, NMPC [27, 22, 21] is often limited to reduced-order models due to the intractability of solving full-order problems involving contacts and nonlinear dynamics.

Sampling-based MPC [43, 45, 46, 44] has been applied to various nonlinear and hybrid dynamical systems because of its flexibility in handling arbitrary dynamics and constraints, as well as parallelizability. Nonetheless, these algorithms are sensitive to hyperparameters in high-dimensional and non-convex optimization problems, particularly the sampling kernel, leading to high variance and

[1] Paper website: https://anonymous-dial-mpc.github.io/dial-mpc/

CoRL 2024 Workshop on Whole-Body Control and Bimanual Manipulation (CoRL 2024 WCBM).

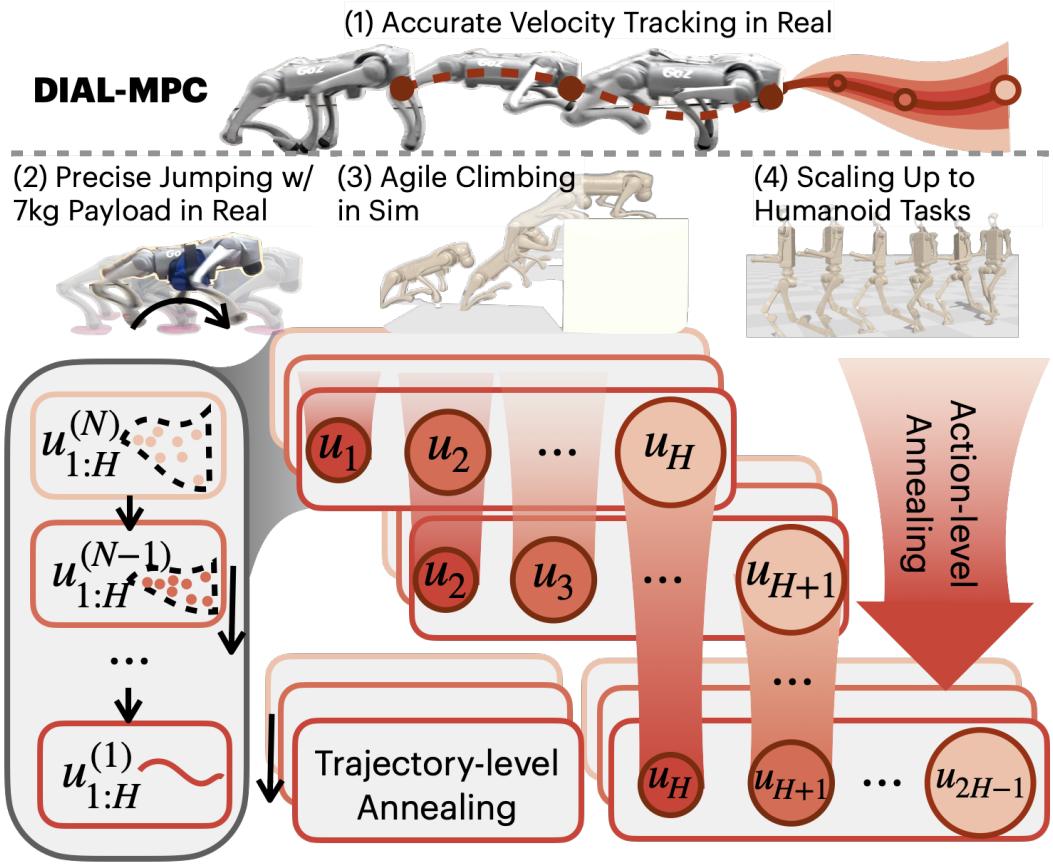

Figure 1: Diffusion-inspired annealing for legged MPC (DIAL-MPC). To achieve both global coverage and local convergence, DIAL-MPC involves a bi-level diffusion-inspired annealing process. Trajectory-wise annealing is performed with different sampling variance. Action-wise annealing is performed on control input at different horizion. Over time, u_H will be gradually refined by the two diffusion-inspired annealing processes, leading to a robust and efficient full-order online control.

suboptimal performance. Specifically, a large sampling range provides better global coverage but may result in solutions far from the optimum, while a small sampling range enhances local search ability but is more susceptible to local minima and initial guesses, leading to increased variance.

In this paper, we address these challenges by unveiling the intrinsic connection between sampling-based MPC and diffusion processes. Following the key iterative refinement idea of the diffusion model, we introduce a novel sampling-based MPC method, Diffusion-Inspired Annealing for Legged Model Predictive Control (DIAL-MPC). DIAL-MPC optimizes control sequences iteratively in a dual-loop manner. Specifically, DIAL-MPC starts optimizing the control sequence with smooth but inaccurate objectives and gradually shifts to more accurate local objectives.

Compared with MPPI, DIAL-MPC enables on-the-fly, full-order torque-level locomotion control by effectively balancing global coverage and local convergence. We evaluate DIAL-MPC on a quadrupedal robot with full-order dynamics and show that it can achieve real-time 50Hz control with both robustness and efficiency. As a training-free and purely online method, DIAL-MPC enables precise jumping with payload and outperforms RL methods. The contribution of this paper is threefold:

- **Novel Diffusion-Inspired Annealing Framework**: We propose a diffusion-inspired annealing framework for sampling-based MPC by revealing the connection between sampling-based MPC and diffusion processes.

- **Full-Order Torque-Level Control of Legged Robot**: We develop and implement DIAL-MPC for full-order torque-level control of legged robots based on the proposed annealing framework. To our knowledge, this is the first framework achieving both real-time flexibility and RL-level agility in legged locomotion.
- **Real-World Validation**: We validate the performance of DIAL-MPC for quadruped control, showing that it achieves real-time 50 Hz control with robustness and efficiency.

2 Related Work

2.1 Agility in Legged Locomotion

Nonlinear Model Predictive Control (NMPC), particularly gradient-based methods, has demonstrated significant success in real-time control of legged locomotion with closed-loop stability [29, 11, 27, 35]. These approaches typically employ reduced-order models to mitigate the burden of planning over full-order hybrid dynamics, necessitating lower-level whole-body controllers for motion execution [22, 15, 20, 21].

However, reduced-order models can lead to sub-optimal performance and constraint violations, particularly under high agility demands. For instance, [19] introduces a tailored solver for full-order NMPC online, yet remains constrained by predefined contact sequences, limiting motion agility and robustness. In contrast, our method enables full-order model MPC without redundant constraints, allowing adaptation to real-time feedback and environmental interactions.

Model-free Reinforcement Learning (RL) has also been explored to address full-order control by learning optimal policies directly from high-fidelity simulations [14, 8, 47, 13, 32, 9]. While these approaches eliminate the need for explicit modeling, they suffer from limited generalization to new tasks and dynamic environments.

Goal-conditioned RL [10, 2, 18] enhances task-level generalization but still struggles with unseen dynamics and novel task classes. Our approach overcomes these limitations by enabling rapid motion generation through training-free online optimization, thereby combining the agility of RL with the robustness and generalization capabilities of MPC.

2.2 Sampling-Based Optimization

Sampling-based or zeroth-order optimization methods, including Bayesian Optimization [6, 37], the Cross-Entropy Method [24], and Evolutionary Algorithms [42], are widely utilized for solving non-convex and non-smooth optimization problems. They differ from first-order methods as the gradient information is no longer required. These methods are particularly effective in applications such as hyperparameter tuning [36, 16] and generative modeling [38].

In the context of real-time robot control, Model Predictive Path Integral (MPPI) control [43] and its variants [45, 46, 44] have gained popularity for online motion planning [30, 33, 17], due to their inherent parallelizability and flexibility. Given a high-stiffness problem like legged locomotion, zeroth-order methods including policy gradient [41] have shown better convergence properties both empirically and theoretically [40] from their smoothing nature.

Despite their advantages, sampling-based methods are plagued by the curse of dimensionality and high variance, especially under constrained online sampling budgets. This is particularly problematic in high-dimensional, contact-rich environments.

2.3 Parallel Robot Simulation

Massively parallelizable simulation environments, such as Isaac Gym [23], Brax [7], and MuJoCo [26], have become essential tools in the development of zeroth-order optimization methods for robot control. These simulators facilitate the rapid generation of data required by sample-hungry algorithms like PPO [34], enabling efficient training of complex policies.

3 Method

In this section, we present our method by establishing the equivalence between MPPI and a single-stage diffusion process (section 3.1). The connection then explains why the annealing process in diffusion helps MPPI to optimize over a non-smooth landscape, as discussed in section 3.2. Leveraging this equivalence, we introduce a diffusion-inspired annealing technique for MPPI in section 3.3.

3.1 Sampling-Based MPC as Single-Stage Diffusion

Optimal control problem. Sampling-based MPC aims to solve the following optimization problem:

$$\begin{aligned} \min_{u_{t:t+H}} J(u_{t:t+H}) &= \sum_{h=0}^{H} c(x_{t+h}, u_{t+h}) + c_f(x_{t+H+1}), \\ \text{s.t.} \quad x_{t+h+1} &= f(x_{t+h}, u_{t+h}) \quad \forall h \in \{0, \dots, H\}, \\ x_{t:t+H+1} &\in \mathcal{X},\ u_{t:t+H} \in \mathcal{U} \end{aligned}$$

where x_{t+h} is the state at time $t+h$, u_{t+h} is the control input at time $t+h$, f is the system dynamics, c and c_f are the cost function and terminal cost function, $\mathcal{X}$ and $\mathcal{U}$ are the state and control constraints, respectively.

MPPI estimates the optimal control sequence through the following steps: First, draw N_W perturbations from a Gaussian distribution $[W]_i \sim \mathcal{N}(0, \Sigma_{t:t+H}), i = 1, \dots, N_W$ (which we collectively denote as $[W]_{1:N_W}$). Then the cost function $J(u_{t:t+H})$ is evaluated for each sampled control sequence by rolling out the system dynamics and cumulatively summing the cost. For each perturbed control sequence $U + [W]_i$, where $U = u_{t:t+H}$, evaluate the cost function $J(U + [W]_i)$ by simulating the system dynamics and accumulating the costs. Finally, update the control sequence using temperature λ:

$$U^+ = U + \frac{\sum_{i=1}^{N_W} \exp\left(-\frac{J(U+[W]_i)}{\lambda}\right)[W]_i}{\sum_{j=1}^{N_W} \exp\left(-\frac{J(U+[W]_j)}{\lambda}\right)}, \tag{1}$$

MPPI as a single-stage Diffusion. The optimization problem can be reframed in a sampling context. Define the target distribution $p_0(U) \propto \exp\left(-\frac{J(U)}{\lambda}\right)$. As $\lambda \to 0$, samples from $p_0(U)$ concentrate around the optimal control sequence U^* as illustrated in fig. 2. However, directly sampling from $p_0(U)$ is impractical due to its narrow support. To facilitate the sampling, we convolve $p_0(U)$ with a Gaussian noise kernel $\phi(\cdot)$, i.e., the density of $\mathcal{N}(0, \Sigma)$ to get the corrupted distribution $p_1(\cdot) \propto (p_0 * \phi)(\cdot)$ as depicted in fig. 4.

Proposition 1 (Adopted from [28])**.** *The MPPI update* (1) *can be viewed as a one-step ascent with the score function* $\nabla \log p_1(U)$ *with a learning rate* Σ*:*

$$U^+ = U + \Sigma \cdot \nabla \log p_1(U). \tag{2}$$

Proof. The diffused distribution p_1 is defined as $p_1(\cdot) \propto (p_0 * \phi)(\cdot)$ as in fig. 4. The score function $\nabla \log p_1(U)$ can be calculated in the following way:

$$\nabla \log p_1(U) = \frac{\nabla p_1(U)}{p_1(U)} = \frac{\nabla \left(p_0(U) * \phi(U)\right)}{p_1(U)} \tag{3a}$$

$$= \frac{p_0(U) * \nabla \phi(U)}{p_1(U)} = -\frac{p_0(U) * (\phi(U)\Sigma^{-1}U)}{p_1(U)} \tag{3b}$$

$$= -\Sigma^{-1} \frac{\int p_0(U-W)\phi(W)W dW}{\int p_0(U-W)\phi(W)dW} \tag{3c}$$

$$= -\Sigma^{-1} \frac{\mathbb{E}_{W\sim\phi(\cdot)}\left[p_0(U-W)W\right]}{\mathbb{E}_{W\sim\phi(\cdot)}\left[p_0(U-W)\right]} \tag{3d}$$

$$= \Sigma^{-1} \frac{\mathbb{E}_{W\sim\phi(\cdot)}\left[p_0(U+W)W\right]}{\mathbb{E}_{W\sim\phi(\cdot)}\left[p_0(U+W)\right]} \tag{3e}$$

$$\approx \Sigma^{-1} \frac{\sum_{i=1}^{N_W} \exp\left(-\frac{J(U+[W]_i)}{\lambda}\right)[W]_i}{\sum_{j=1}^{N_W} \exp\left(-\frac{J(U+[W]_j)}{\lambda}\right)}. \tag{3f}$$

From (3a) to (3b), we move the gradient into the convolusion. In (3b), the gradient of Gaussian kernal $\phi(W)$ is calculated. From (3b) to (3c), we use the definition of convolution. In (3d), we rewrite the integral as an expectation. In (3e), we flip the sign of W to match the form of MPPI since the Gaussian distribution is symmetric. From (3e) to (3f), Monte Carlo approximation is applied.

□

In other words, MPPI performs a "denoising" step with the score function $\nabla \log p_1(U)$ conditioned on a *fixed* noise level $\mathcal{N}(0, \Sigma)$ [39]. The key difference is that the score-based diffusion model [39] iteratively refines samples using different noise levels, which motivates our annealing design[2].

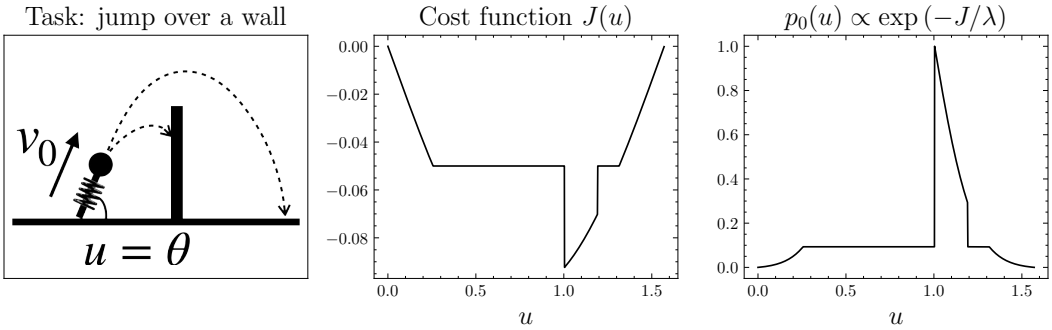

Figure 2: Cost function $J(U)$ and target distribution $p_0(U)$ for a task where robot need to jump over a wall. The cost function could be highly non-convex and non-smooth due to the contact constraint. The resulting distribution $p_0(U)$ is also non-convex and sparse, which is hard to sample from.

3.2 Diffusion-Inspired Annealing

Given that MPPI is a single-stage diffusion that moves particles toward the stationary point of $p_1(\cdot) = (p_0 * \phi)(\cdot)$ (proposition 1), a natural question arises: What are the pros and cons of optimizing p_1 compared to directly solving for the optimum of p_0?

Figure 3 illustrates that the convexity of the distribution increases with progressively larger kernels. Each successive density function from p_1 to p_4 is obtained by convolving the original distribution p_0 with a progressively larger kernel. As discussed in section 3.1, MPPI performs score ascent on

[2]There are some other differences between (2) and diffusion models: (2) does not have scaling factor or extra noise. See more discussion in [28].

p_i. Therefore, the primary **advantage** of conducting score ascent on p_3 and p_4 is that MPPI can more easily converge to the global optimum. The main **disadvantage** of optimizing a corrupted distribution is also clear from fig. 3. The optimum progressively shifts from U^* to a distorted optimum under a larger kernel, thereby compromising the optimality of the original problem.

Coverage and convergence trade-off. The advantages and disadvantages highlight a fundamental trade-off between exploration and exploitation in MPPI: a larger $\det \Sigma$ promotes greater exploration (i.e., coverage) by widening the sampling distribution, which helps to avoid suboptimal local minima. However, a larger $\det \Sigma$ may compromise optimality(i.e., convergence) as it will introduce a

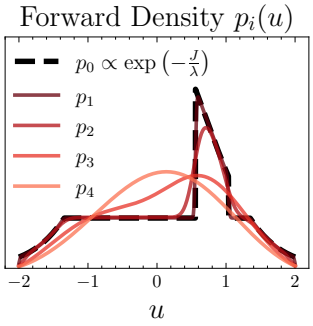

Figure 3: Forward density function in diffusion process.

larger optimality gap. In contrast, a smaller $\det \Sigma$ improves local optimality at the risk of getting trapped in local minima.

This trade-off becomes more pronounced in contact-rich tasks such as legged locomotion, where the optimization landscape is non-smooth, non-convex, and high-dimensional. The cost function J and the distribution p_0 often have sharp and asymmetric peaks. In such cases, even a small increase in $\det \Sigma$ can degrade performance, and global optimality cannot be guaranteed through score ascent on $p_1(U)$ due to the small convolution kernel.

By fixing the sampling kernel size, MPPI may either over-explore or over-exploit, failing to balance exploration and convergence effectively. This concern is illustrated in fig. 4, which shows how different kernel sizes affect the performance of MPPI. Therefore, designing an effective sampling strategy that balances coverage and convergence is crucial to unlocking the potential of MPPI in real-world legged locomotion tasks.

Fortunately, diffusion processes, known for their powerful sampling capabilities from complex distributions, offer a way to balance coverage and convergence through an annealing strategy. This strategy involves sampling iteratively at decreasing noise levels, effectively reversing the forward corruption process.

Annealing in diffusion process. Instead of sampling solely from $p_1(\cdot)$, the forward process in diffusion defines a sequence of distributions with increasing noise levels: $p_1(\cdot), \ldots, p_{N-1}(\cdot), p_N(\cdot)$, where N is the total number of diffusion stages. Each density is defined as $p_i(\cdot) = (p_0 * \phi_i)(\cdot)$ with $\phi_i(\cdot) \sim \mathcal{N}(0, \Sigma^i)$. Sampling proceeds in reverse order. Starting from a higher noise level Σ^N, the sampling distribution $p_N(\cdot)$ is highly spread out, ensuring good global exploration. As the noise level decreases, the sampling distributions $p_i(U)$ become more concentrated, refining the search towards the target distribution $p_0(U)$ and improving convergence to the optimal solution. We adopt an exponential schedule for the noise levels, inspired by diffusion processes, to design the sampling kernels:

$$\det\left(\Sigma^i\right) = \exp\left(-\frac{N-i}{\beta N} d\right), \quad \forall i \in \{N, \ldots, 1\}, \tag{4}$$

where β is the temperature parameter for the annealing process, and N is the number of iterations for the annealing process, d is the dimension of the sampling space.

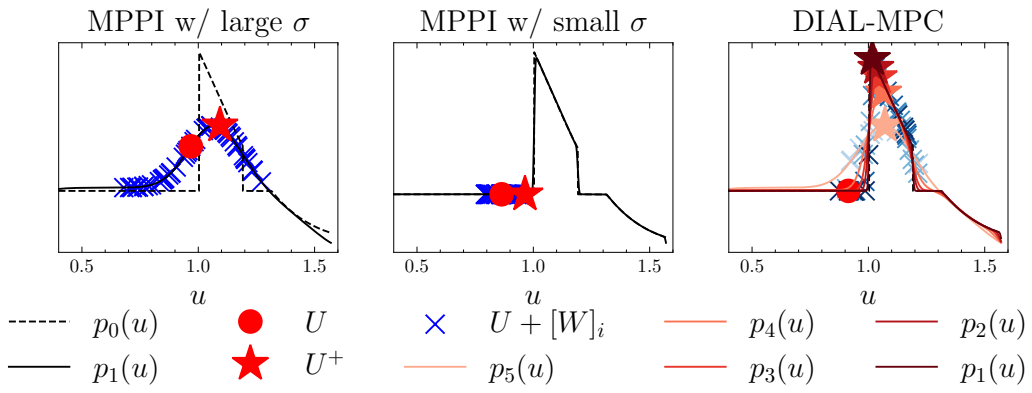

Figure 4: Coverage and convergence trade-off in sampling-based methods. Given the target distribution $p_0(U)$ and the same number of samples, MPPI either over-explores or over-exploits the solution, while our method balances the exploration and exploitation to converge to optima following a diffusion-inspired annealing process.

Given that MPPI corresponds to a single-stage diffusion process, we can naturally incorporate this multi-stage annealed diffusion approach to design an improved sampling strategy for MPPI. In section 3.3, we discuss how to implement this diffusion-inspired annealing process within the MPPI framework in a receding horizon manner.

3.3 Diffusion-Inspired Annealing for Sampling-Based MPC

Algorithm 1 Diffusion-Inspired Annealing for Legged MPC

Initialize $u_{0:H} \leftarrow \mathbf{0}$
for $t = 0$ to T **do**
  for $i = 1$ to N **do**
    Get diffusion noise kernel $\Sigma^i_{t:t+H}$ with (7).
    Sample $[W]_{1:N_W} \sim \mathcal{N}(0, \Sigma^i_{t:t+H})$.
    Rollout $u_{t:t+H} + [W]_{1:N_W}$ and evaluate the cost function $J(u_{t:t+H} + [W]_{1:N_W})$.
    Estimate Score $\nabla \log p_i(\cdot)$ with (3).
    Update $u^{(i)}_{t:t+H}$ with (2).
  end for
  Receding horizon $u_{t+1:t+H+1} \leftarrow \texttt{shift}(u_{t:t+H})$
end for

Combining MPPI with multi-stage annealing, we propose DIAL-MPC (algorithm 1) that leverages the receding horizon structure of MPC and introduces a dual-loop covariance design. This design comprises two annealing procedures: an outer-loop trajectory-level annealing and an inner-loop action-level annealing, which we detail below.

Dual-loop annealing. In a receding horizon MPC framework with a horizon length H and N update iterations for each control sequence at each timestep t, u_H (the last element of $u_{0:H}$) is first updated at time 0 by N times using convolution kernel $\Sigma^N_H \dots, \Sigma^1_H$. After u_0 is applied to the system, the control sequence shifts forward. At the next time step $t = 1$, u_H is updated again N times, now with kernel $\Sigma^N_{H-1}, \dots, \Sigma^1_{H-1}$ as u_H becomes the second-to-last element of the updated control sequence $u_{1:H+1}$. This procedure continues, and by the time u_H is applied to the system at $t = H$, it will have undergone NH updates with convolution kernels: $\Sigma^N_H \dots, \Sigma^1_H; \Sigma^N_{H-1}, \dots, \Sigma^1_{H-1}; \dots; \Sigma^N_0 \dots, \Sigma^1_0$ as shown in the last graph of fig. 1. Essentially, each control action u_h is updated in a dual-loop manner before being applied to the system. This motivates the design of two annealing procedures: the outer loop is a trajectory-level annealing schedule for all the control sequence at a certain stage i

(i.e. designing the overall size of $\Sigma^i_{t:t+H}$ for a given i) and the inner loop is an action-level annealing schedule for different control at different horizons (i.e. the size of Σ^i_{t+h} for $h = 0, \ldots, H$).

Trajectory-level annealing. For the outer-loop trajectory-level annealing, we define the schedule as:

$$\det\left(\Sigma^i_{t:t+H}\right) \propto \exp\left(-\frac{N-i}{\beta_1 N} H d_u\right), \quad \forall i \in \{N, \ldots, 1\}, \tag{5}$$

where β_1 is the temperature parameter for the trajectory-level annealing, and d_u is the dimension of a single control. The covariance matrix decreases over time as i decreases, progressively narrowing the sampling distribution towards the target distribution p_0.

Action-level annealing. For the inner-loop action-level annealing, the schedule is given by:

$$\det\left(\Sigma^i_{t+h}\right) \propto \exp\left(-\frac{H-h}{\beta_2 H} d_u\right), \quad \forall h \in \{0, \ldots, H\}, \tag{6}$$

where β_2 is the temperature parameter for action-level annealing. The covariance matrix increases with time as h increases, allowing for a larger sampling region for future control actions that have been updated fewer times compared to those at the front of the horizon.

Note that (5) and (6) specify only the overall size (i.e., the determinant) of the convolution kernels, leaving flexibility in designing the exact covariance matrices Σ^i_{t+h}. As a practical realization of (5) and (6), we assume that each Σ^i_{t+h} is isotropic and define it as:

$$\Sigma^i_{t+h} = \exp\left(-\frac{N-i}{\beta_1 N} - \frac{H-h}{\beta_2 H}\right) I. \tag{7}$$

where I is the identity matrix. This design combines both the outer-loop and inner-loop annealing schedules, adjusting the covariance matrices to ensure appropriate exploration and exploitation at each iteration and horizon step.

4 Experiment

In this section, we demonstrate the advantages of DIAL-MPC in terms of: (1) convergence and coverage, (2) efficiency in test-time generalizability, and (3) robustness to real-world model mismatch. We compare DIAL-MPC against MPPI and another sampling-based optimization method, CMA-ES [1]. In addition, we provide goal-conditioned reinforcement learning (GCRL) as a performance reference. Our results show that DIAL-MPC reduces the tracking error by 3.9 times in walking tasks compared to MPPI and outperforms GCRL on all tasks requiring both precision and agility, especially in the presence of significant model mismatch.

We design a set of agile locomotion tasks to showcase the performance of DIAL-MPC: (1) Walking Tracking: a quadruped robot[3] is tasked with tracking a desired linear velocity and a desired yaw rate, requiring precise control of the torso. (2) Sequential Jumping: a quadruped robot must jump onto a series of small circular platforms placed randomly, each with a radius of $10\,\mathrm{cm}$. (3) Crate Climbing: a quadruped robot is tasked with climbing a crate with a height of $60\,\mathrm{cm}$, which is more than twice the height of the robot. We leave the specific details of the tasks' implementation in Appendix A.2.

4.1 Convergence and Coverage

To answer the question of whether DIAL-MPC is a better solver for the legged MPC problems, we compare the performance of DIAL-MPC with vanilla MPPI, CMA-ES and NMPC. Since vanilla MPPI only uses a single kernel size, we tested both MPPI with large kernel size 0.2 (MPPI-explore) and small kernel 0.05 (MPPI-exploit). Compared with DIAL-MPC, CMA-ES optimizes the sampling covariance in an evolutionary way. For all sampling-based MPC methods, we use the same

[3]Detailed information of the hardware setup can be found in A.1.

number of samples $N_W = 2048$ and optimization steps $H = 20$ (0.4 s) to ensure a fair comparison. For NMPC, we use Mujoco MPC [17] as our baseline. As a reference, we also include the performance of GCRL, which is trained using PPO [34] over 500 million steps, which takes approximately 31 minutes. We share the same reward functions for all methods. Due to unstable exploration in GCRL, we added two additional reward terms to regularize the policy, whereas DIAL-MPC only requires six reward terms.

	Walk Track ↓	Seq Jump ↑	Crate Climb ↑
MPPI-explore	0.190	0.440	0.3
MPPI-exploit	0.230	0.450	0.2
CMA-ES	0.544	0.345	0.2
NMPC	0.055	-	-
GCRL*	0.119	0.855	0.6
DIAL-MPC	**0.024**	**0.885**	**0.9**

Table 1: Performance comparison of different methods over three tasks. For walk tracking, it represents the tracking error; For sequential jumping, it represents average total contact reward; For crate climbing, it represents success rate out of 10 trials of climbing onto random crates with heights ranging from 0.125 to 0.6 m. *GCRL requires offline training so it is not an apple-to-apple baseline.

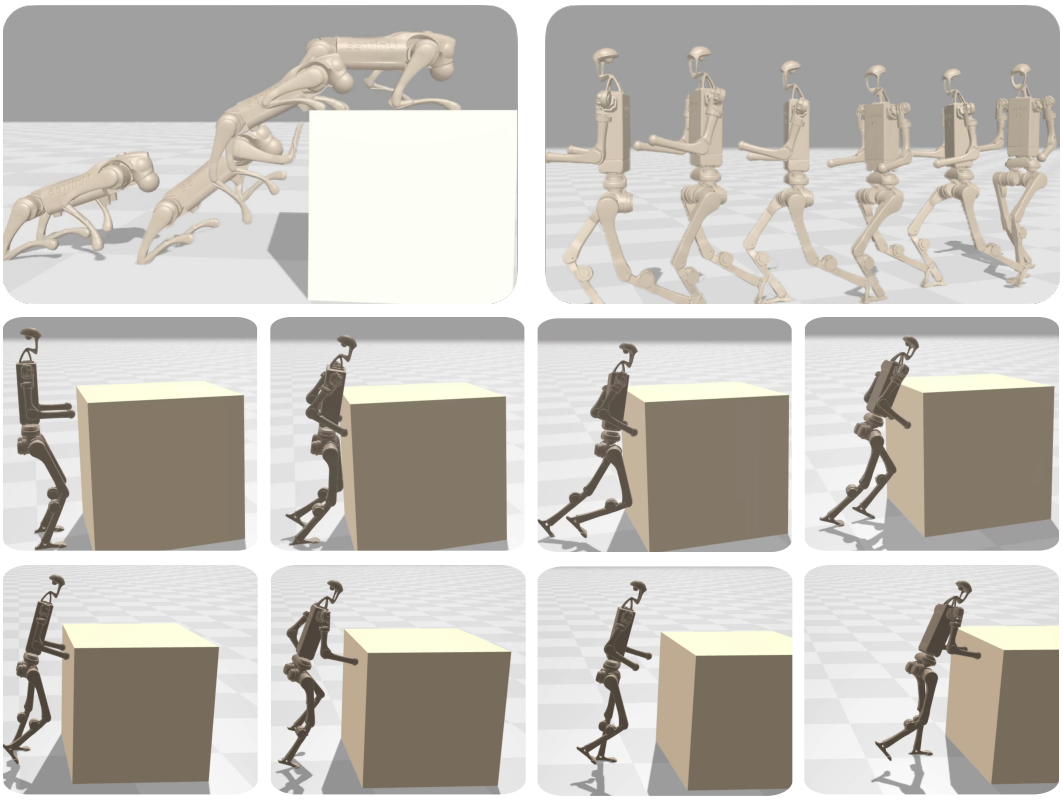

Figure 5: **Top**: the coverage of DIAL-MPC in crate-climbing and humanoid jogging task. The crate-climbing task requires the robot to climb up a crate more than two times higher than itself. The full-size humanoid jogging task demands DIAL-MPC to handle a higher-dimensional action space of 19. **Middle**: DIAL-MPC controlling a humanoid pushing a 30 kg crate. **Bottom**: DIAL-MPC generates a motion strategy with less effort when reducing the crate's weight to 15 kg.

Convergence of DIAL-MPC. As shown in table 1, DIAL-MPC achieves the best performance across all tasks. Compared with other sampling-based MPC methods (namely MPPI and CMA-ES), DIAL-MPC generates better solutions, with 13.4 times lower tracking error and 107.7% higher con-

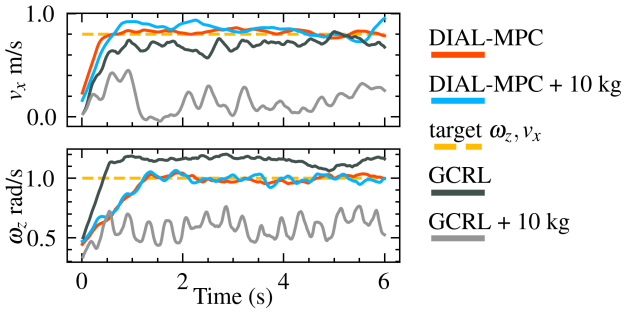

Figure 6: The linear and angular velocity tracking performance of DIAL-MPC and GCRL in the walking task in simulation. GCRL fails with a 10kg payload.

tact reward, thanks to its diffusion-inspired annealing process. For the crate climbing task, DIAL-MPC is the only sampling-based MPC that consistently generates feasible solutions, improving the success rate by 3 times compared to the best MPPI scheduling. This highlights the superior coverage of the solution space by DIAL-MPC. Compared with NMPC, DIAL-MPC is able to solve tasks requiring higher agility and non-smooth costs, such as sequential jumping and crate climbing, where NMPC is more likely to get stuck in local minima and fail to converge.

The superior convergence of DIAL-MPC is further demonstrated when compared with GCRL, where DIAL-MPC consistently outperforms GCRL in all tasks even if we use a lower-level leg position controller for RL and DIAL-MPC directly outputs the torque. Although DIAL-MPC is training-free, making a direct comparison potentially unfair, these results showcase the advantage of the annealing process in generating finer solutions compared to the Gaussian exploration in RL.

Coverage of DIAL-MPC. DIAL-MPC is also capable of searching for non-trivial solutions and generating diverse motions. As examples, we visualize the solutions in the crate climbing task and a humanoid jogging task in Figure 5. DIAL-MPC generates diverse solutions in high-dimensional spaces, thanks to the annealing process.

4.2 Test-Time Generalizability

As a training-free method, DIAL-MPC offers better test-time generalizability. Given a new task or model, DIAL-MPC can generate solutions in real time without finetuning.

Task-level generalizability. Compared with GCRL, DIAL-MPC reduces the tracking error by 3.9 times in the walking task and improves the contact reward by 3.5% given different goals. Figure 6 visualizes the tracking performance of both methods. DIAL-MPC achieves higher tracking accuracy thanks to its explicit conditioning on each task, whereas GCRL uses a universal policy for all tasks.

Dynamic-level generalizability. Another advantage of DIAL-MPC is its explicit conditioning on the dynamics model, enabling better adaptation to new dynamics given updated models. Figure 6 illustrates the tracking performance of both methods with a 10 kg payload attached to the robot's base. The crate-climbing task is deprecated due to the heavy payload leads to infeasible solutions. The GCRL policy is augmented with domain randomization of mass, actuator gain, and friction to improve robustness to model parameters. Table 2 shows the performance comparison of GCRL and DIAL-MPC under a 10 kg payload. Without explicit conditioning on the physical parameters, GCRL performs poorly in tracking and completely fails after adding the payload. In contrast, DIAL-MPC outperform GCRL with a larger margin in the jumping task, demonstrating the advantage of explicitly conditioning on the dynamics model. While RL can be enhanced by conditioning on physical parameters or history, this requires additional training time and engineering effort. Conversely, the training-free DIAL-MPC can be instantly deployed with an updated model.

	Walk Track ↓	Seq Jump ↑
GCRL	0.517	0.150
DIAL-MPC	**0.036**	**0.815**

Table 2: Performance comparison of GCRL and DIAL-MPC under 10 kg payload.

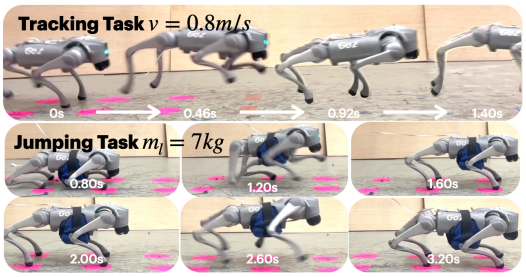

Figure 7: Velocity tracking task and sequential jumping task with a 7 kg payload on Go2 quadruped in real world with DIAL-MPC for direct torque control. DIAL-MPC enables both precise and agile motion.

4.3 Robustness to Model Mismatch

When dealing with unknown dynamics models, DIAL-MPC 's robustness stems from the noise injection process in the diffusion-inspired annealing. By controlling the final noise level, we can achieve a suboptimal but robust solution given a mismatched model. Since the sampling-based MPC baselines fail to operate effectively in the real world, we compare the performance of DIAL-MPC with the baseline in simulation with mismatched mass parameters. Table 3 shows the performance comparison of different methods under 2 kg base mass mismatch in simulation, where DIAL-MPC still outperforms the best sampling-based MPC baselines by 92% in the walking task and 126% in the jumping.

	Walk Track ↓	Seq Jump ↑
MPPI-explore	0.361	0.270
MPPI-exploit	0.407	0.200
CMA-ES	0.556	0.100
DIAL-MPC	**0.188**	**0.610**
DIAL-MPC in real	0.322	0.600

Table 3: Performance comparison of different methods under 2 kg base mass mismatch in simulation and real world.

Figure 7 shows the DIAL-MPC control a Unitree Go2 robot to walk and jump. The robot is able to walk smoothly and jump precisely with direct torque control.

5 Conclusion

This work presents DIAL-MPC, a sampling-based MPC method with a diffusion-inspired annealing process to balance coverage and convergence in real-world legged locomotion. DIAL-MPC can solve the full-order control problem efficiently with the help of diffusion process and is generalizable to various tasks and dynamics in a training-free manner. One limitation is that DIAL-MPC requires fast simulation to generate samples, which limits the application of DIAL-MPC in longer planning horizon tasks. In the future, we plan to further accelerate and improve the sample efficiency by learning a nominal policy, value function and model as what has been done in model-base RL.

References

[1] Youhei Akimoto et al. "Theoretical Foundation for CMA-ES from Information Geometry Perspective". In: *Algorithmica* 64.4 (Dec. 2012), pp. 698–716. ISSN: 0178-4617, 1432-0541. DOI: 10.1007/s00453-011-9564-8. (Visited on 06/25/2023).

[2] Vassil Atanassov et al. *Curriculum-Based Reinforcement Learning for Quadrupedal Jumping: A Reference-free Design*. Mar. 2024. DOI: 10.48550/arXiv.2401.16337. arXiv: 2401.16337 [cs]. (Visited on 09/12/2024).

[3] James Bradbury et al. *Google/Jax*. Google. Sept. 2024. (Visited on 09/10/2024).

[4] Shuxiao Chen et al. *Learning Torque Control for Quadrupedal Locomotion*. Mar. 2023. DOI: 10.48550/arXiv.2203.05194. arXiv: 2203.05194 [cs, eess]. (Visited on 09/05/2024).

[5] Xuxin Cheng et al. "Extreme Parkour with Legged Robots". In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. Yokohama, Japan: IEEE, May 2024, pp. 11443–11450. ISBN: 9798350384574. DOI: 10.1109/ICRA57147.2024.10610200. (Visited on 09/14/2024).

[6] Peter I. Frazier. *A Tutorial on Bayesian Optimization*. July 2018. arXiv: 1807.02811 [cs, math, stat]. (Visited on 05/17/2024).

[7] C. Daniel Freeman et al. *Brax – A Differentiable Physics Engine for Large Scale Rigid Body Simulation*. June 2021. DOI: 10.48550/arXiv.2106.13281. arXiv: 2106.13281 [cs]. (Visited on 09/03/2024).

[8] Zipeng Fu et al. "HumanPlus: Humanoid Shadowing and Imitation from Humans". In: ().

[9] Zipeng Fu et al. *Minimizing Energy Consumption Leads to the Emergence of Gaits in Legged Robots*. Oct. 2021. DOI: 10.48550/arXiv.2111.01674. arXiv: 2111.01674 [cs]. (Visited on 09/12/2024).

[10] Dibya Ghosh, Abhishek Gupta, and Sergey Levine. *Learning Actionable Representations with Goal-Conditioned Policies*. Jan. 2019. arXiv: 1811.07819 [cs, stat]. (Visited on 09/12/2024).

[11] Ruben Grandia et al. *Perceptive Locomotion through Nonlinear Model Predictive Control*. Aug. 2022. arXiv: 2208.08373 [cs]. (Visited on 09/12/2024).

[12] Tairan He et al. "Agile But Safe: Learning Collision-Free High-Speed Legged Locomotion". In: ().

[13] Tairan He et al. *Learning Human-to-Humanoid Real-Time Whole-Body Teleoperation*. Mar. 2024. arXiv: 2403.04436 [cs, eess]. (Visited on 09/09/2024).

[14] Tairan He et al. "OmniH2O: Universal and Dexterous Human-to- Humanoid Whole-Body Teleoperation and Learning". In: ().

[15] Andrei Herdt et al. "Online Walking Motion Generation with Automatic Footstep Placement". In: *Advanced Robotics* 24.5-6 (Jan. 2010), pp. 719–737. ISSN: 0169-1864, 1568-5535. DOI: 10.1163/016918610X493552. (Visited on 08/19/2024).

[16] José Miguel Hernández-Lobato et al. "Parallel and Distributed Thompson Sampling for Large-scale Accelerated Exploration of Chemical Space". In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, July 2017, pp. 1470–1479. (Visited on 05/17/2024).

[17] Taylor Howell et al. *Predictive Sampling: Real-time Behaviour Synthesis with MuJoCo*. Dec. 2022. DOI: 10.48550/arXiv.2212.00541. arXiv: 2212.00541 [cs, eess]. (Visited on 09/03/2024).

[18] Kevin Huang et al. "DATT: Deep Adaptive Trajectory Tracking for Quadrotor Control". In: *Conference on Robot Learning*. PMLR. 2023, pp. 326–340.

[19] Charles Khazoom et al. *Tailoring Solution Accuracy for Fast Whole-body Model Predictive Control of Legged Robots*. July 2024. arXiv: 2407.10789 [cs]. (Visited on 08/17/2024).

[20] Donghyun Kim et al. *Highly Dynamic Quadruped Locomotion via Whole-Body Impulse Control and Model Predictive Control*. Sept. 2019. arXiv: 1909.06586 [cs]. (Visited on 08/09/2024).

[21] J. Koenemann et al. "Whole-Body Model-Predictive Control Applied to the HRP-2 Humanoid". In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Sept. 2015, pp. 3346–3351. DOI: 10.1109/IROS.2015.7353843. (Visited on 09/12/2024).

[22] Scott Kuindersma, Frank Permenter, and Russ Tedrake. "An Efficiently Solvable Quadratic Program for Stabilizing Dynamic Locomotion". In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. May 2014, pp. 2589–2594. DOI: 10.1109/ICRA.2014.6907230. arXiv: 1311.1839 [cs]. (Visited on 08/17/2024).

[23] Viktor Makoviychuk et al. *Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning*. Aug. 2021. arXiv: 2108.10470 [cs]. (Visited on 09/12/2024).

[24] Shie Mannor, Reuven Rubinstein, and Yohai Gat. "The Cross Entropy Method for Fast Policy Search". In: ().

[25] Takahiro Miki et al. "Learning Robust Perceptive Locomotion for Quadrupedal Robots in the Wild". In: *Science Robotics* 7.62 (Jan. 2022), eabk2822. ISSN: 2470-9476. DOI: 10.1126/scirobotics.abk2822. arXiv: 2201.08117 [cs]. (Visited on 09/14/2024).

[26] *MuJoCo: A Physics Engine for Model-Based Control — IEEE Conference Publication — IEEE Xplore*. https://ieeexplore.ieee.org/document/6386109. (Visited on 09/12/2024).

[27] Michael Neunert et al. "Whole-Body Nonlinear Model Predictive Control Through Contacts for Quadrupeds". In: *IEEE Robotics and Automation Letters* 3.3 (July 2018), pp. 1458–1465. ISSN: 2377-3766, 2377-3774. DOI: 10.1109/LRA.2018.2800124. arXiv: 1712.02889 [cs]. (Visited on 09/12/2024).

[28] Chaoyi Pan et al. *Model-Based Diffusion for Trajectory Optimization*. May 2024. DOI: 10.48550/arXiv.2407.01573. arXiv: 2407.01573 [cs, eess, math]. (Visited on 09/01/2024).

[29] Hae-Won Park, Patrick M. Wensing, and Sangbae Kim. "High-Speed Bounding with the MIT Cheetah 2: Control Design and Experiments". In: *Other repository* (Feb. 2017). ISSN: 0278-3649. (Visited on 09/12/2024).

[30] Jintasit Pravitra et al. "$L1$-Adaptive MPPI Architecture for Robust and Agile Control of Multirotors". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Oct. 2020, pp. 7661–7666. DOI: 10.1109/IROS45743.2020.9341154. (Visited on 11/28/2023).

[31] Ilija Radosavovic et al. *Humanoid Locomotion as Next Token Prediction*. Feb. 2024. DOI: 10.48550/arXiv.2402.19469. arXiv: 2402.19469 [cs]. (Visited on 09/01/2024).

[32] Nikita Rudin et al. "Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning". In: ().

[33] Jacob Sacks et al. *Deep Model Predictive Optimization*. Oct. 2023. DOI: 10.48550/arXiv.2310.04590. arXiv: 2310.04590 [cs]. (Visited on 11/28/2023).

[34] John Schulman et al. *Proximal Policy Optimization Algorithms*. Aug. 2017. DOI: 10.48550/arXiv.1707.06347. arXiv: 1707.06347 [cs]. (Visited on 05/13/2023).

[35] Jean-Pierre Sleiman et al. *A Unified MPC Framework for Whole-Body Dynamic Locomotion and Manipulation*. Mar. 2021. arXiv: 2103.00946 [cs]. (Visited on 09/12/2024).

[36] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. (Visited on 05/17/2024).

[37] Jascha Sohl-Dickstein et al. *Deep Unsupervised Learning Using Nonequilibrium Thermodynamics*. Nov. 2015. DOI: 10.48550/arXiv.1503.03585. arXiv: 1503.03585 [cond-mat, q-bio, stat]. (Visited on 05/04/2024).

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising Diffusion Implicit Models". In: *International Conference on Learning Representations*. Oct. 2020. (Visited on 05/04/2024).

[39] Yang Song and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Advances in neural information processing systems* 32 (2019).

[40] H. J. Terry Suh et al. *Do Differentiable Simulators Give Better Policy Gradients?* Aug. 2022. arXiv: 2202.00817 [cs]. (Visited on 06/11/2024).

[41] Richard S Sutton et al. "Policy Gradient Methods for Reinforcement Learning with Function Approximation". In: *Advances in Neural Information Processing Systems*. Vol. 12. MIT Press, 1999. (Visited on 09/12/2024).

[42] Daan Wierstra et al. "Natural Evolution Strategies". In: *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*. Hong Kong, China: IEEE, June 2008, pp. 3381–3387. ISBN: 978-1-4244-1822-0. DOI: 10.1109/CEC.2008.4631255. (Visited on 11/23/2023).

[43] Grady Williams et al. "Aggressive Driving with Model Predictive Path Integral Control". In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. May 2016, pp. 1433–1440. DOI: 10.1109/ICRA.2016.7487277.

[44] Grady Williams et al. "Robust Sampling Based Model Predictive Control with Sparse Objective Information". In: *Robotics: Science and Systems XIV*. Robotics: Science and Systems Foundation, June 2018. ISBN: 978-0-9923747-4-7. DOI: 10.15607/RSS.2018.XIV.042. (Visited on 10/06/2023).

[45] Zeji Yi et al. *CoVO-MPC: Theoretical Analysis of Sampling-based MPC and Optimal Covariance Design*. Jan. 2024. DOI: 10.48550/arXiv.2401.07369. arXiv: 2401.07369 [cs]. (Visited on 02/19/2024).

[46] Ji Yin et al. "Trajectory Distribution Control for Model Predictive Path Integral Control Using Covariance Steering". In: *2022 International Conference on Robotics and Automation (ICRA)*. Philadelphia, PA, USA: IEEE, May 2022, pp. 1478–1484. ISBN: 978-1-72819-681-7. DOI: 10.1109/ICRA46639.2022.9811615. (Visited on 10/06/2023).

[47] Chong Zhang et al. *WoCoCo: Learning Whole-Body Humanoid Control with Sequential Contacts*. 2024. arXiv: 2406.06005 [cs.RO]. URL: https://arxiv.org/abs/2406.06005.

6 Acknowledgment

This research is sponsored by DOD Advanced Research Projects Agency: Design of Robustly Implementable Autonomous and Intelligent Machines: Award HR00112490425.

A Appendix

A.1 Hardware and Software Setup

We discuss various hardware platforms, including quadrupeds and humanoids, and provide details on the implementation of the algorithm.

Quadraped Configurations For quadrupedal tasks, we use a Unitree GO2 robot with 18 degrees of freedom (DoF) and 12 actuated joints. The abduction and hip joints can produce $24\,\mathrm{N}\cdot\mathrm{m}$ torque, and the knee joints can produce $45\,\mathrm{N}\cdot\mathrm{m}$ torque approximately. We perform basic system identification to match the joint response to that in the simulation as closely as possible. The low-level controller for each joint of the quadruped plays a crucial role across various tasks. To ensure fair comparisons, the PID controller parameters are kept consistent across different testbeds. In the GCRL reference implementation, which uses position control, we set the proportional gain to 30 and the derivative gain to 0.65. In our method, which utilizes torque control, we apply the same derivative gain of 0.65 to regulate excessively high motor speeds. This choice also helps stabilize the simulation, which uses the same derivative gain. Given a torque command τ_i, the final torque command sent to the i-th joint is

$$\hat{\tau}_i = \tau_i - d\omega_i, \tag{8}$$

where i is the index of the joint, $d = 0.65$ is the derivative gain and ω_i is the angular velocity of the i-th joint.

Algorithm Implementation: All algorithms are implemented in JAX [3], using the Brax simulator [7]. The GCRL model is trained using the reference PPO implementation provided by Brax. All inference runs are performed on a desktop equipped with an RTX 4090 GPU, i9-13900KF CPU, and 32 GB of RAM, with the computer communicating with the robot via an Ethernet connection. For real-time control, all algorithms operate at $50\,\mathrm{Hz}$, while a low-level controller repeats the last algorithm output and sends motor commands to the robot at $200\,\mathrm{Hz}$. Ground-truth state estimation is obtained via a motion capture system. All sampling-based MPCs utilize 2048 parallel environments and a 20-step horizon ($0.4\,\mathrm{s}$). To further speed up the simulations, contact forces are disabled for all robot parts except the feet.

A.2 Task Implementation Details

Quadruped Velocity Tracking We give all sampling-based MPCs the rewards specified in table 4. In particular, the gait is generated by a foot height generator outputting a trotting gait; energy is calculated as the product of joint torque τ_i and joint velocity ω_i. With only six reward terms, DIAL-MPC demonstrates robust walking behavior in this task. We then applied GCRL using the same reward terms; to regularize the training process and improve the behavior, two additional key terms are introduced. The first is an alive reward to counterbalance the negative rewards and promote meaningful exploration during the initial stages of training. The second is a control rate cost, which subsumes the spline reparameterization trick used in DIAL-MPC to produce smoother actions. Furthermore, we implemented standard RL domain randomizations, varying the base mass by up to 1 kg, actuator gains by up to 20%, and the friction coefficient between 0.6 and 1.0. For the goal, we randomized the linear velocity target to $\pm(1.5, 0.5, 0.0)\ \mathrm{m/s}$, and the yaw angular velocity target to $\pm 1.5\ \mathrm{rad/s}$. We trained GCRL using PPO for 500 million steps, which took 31 minutes.

Quadruped Sequential Jumping In this task, quadruped must rapidly jump onto a series of $10\,\mathrm{cm}$ plates. The task is depicted in Figure 8. At a regular time interval, a new contact target is given in terms of four-foot placements and a center-of-mass (CoM) location. The interval is so short ($1\,\mathrm{s}$) that the quadruped must transition smoothly and dynamically between the targets. We uniformly

	Walking and Tracking		Sequential Jumping	
	DIAL-MPC	**GCRL**	**DIAL-MPC**	**GCRL**
Upright	-1.0	-1.0	-1.0	-1.0
Base Height	-1.0	-1.0	-1.0	-1.0
Energy	-0.001	-0.001	-0.001	-0.001
Linear Velocity	-0.5	-1.0	-	-
Angular Velocity	-0.5	-1.0	-	-
Gait	-0.1	-0.1	-	-
Contact Reward	-	-	0.1	0.1
Contact Penalty	-	-	-0.1	-0.1
Alive	-	3.0	-	10.0
Control Rate	-	0.001	-	0.001

Table 4: Reward specifications for the quadruped walking-tracking and sequential jumping tasks are largely shared between our method and the GCRL baseline, underscoring the flexibility of DIAL-MPC in utilizing RL-style rewards.

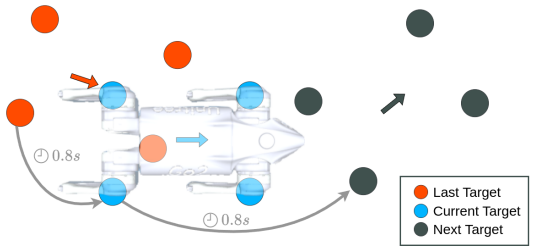

Figure 8: A visual illustration of the sequential jumping task. At every stage, the quadruped is given 1 s to jump onto the next target. Contacts outside the target are penalized.

sample the next contact target, whose CoM location is within $\pm(0.65, 0.65, 0.0)\mathrm{m}$ translation and $0.5\,\mathrm{rad}$ yaw rotation w.r.t. the current contact target.

Inspired by WoCoCo [47], we use a staged contact reward system at time t:

$$r_{\text{con}}^{(j)}(t) = w_{\text{correct}} n_{\text{correct}}^{(j)}(t) - w_{\text{wrong}} \left(n_{\text{wrong}}^{(j)}(t) - n_{\text{correct}}^{(j-1)}(t) \right), \tag{9}$$

where $r_{\text{con}}^{(j)}$ is the current-stage contact reward function; w_{correct} and w_{wrong} are the contact reward and penalty weights; $n_{\text{correct}}^{(j)}$ and $n_{\text{wrong}}^{(j)}$ compute the correct and wrong number of contacts w.r.t. the current stage target; $n_{\text{correct}}^{(j-1)}$ computes the correct number of contacts w.r.t. the last stage target; The last term offsets the penalty of wrong contacts that were valid last-stage contacts. This prevents the robot from being penalized or rewarded while preparing for its current jump.

We establish a performance measurement using the total contact reward over a sequence of 10 jumping stages in $8\,\mathrm{s}$, denoted as R_{con}:

$$R_{\text{con}} = \sum_{j=1}^{j_{\max}} \min_{t \in [T_{\min}(j), T_{\max}(j))} \left(r_{\text{con}}^{(j)}(t) \right), \tag{10}$$

where $j_{\max} = 10$; $T_{\min}(j), T_{\max}(j)$ maps the time interval of the j-th contact stage; the minimum function finds the worst-case contact score at every stage.

We let every algorithm perform 5 trials, and report the average total contact reward, $\bar{R}_{\text{con}}$. Same as in the walking-tracking experiment, we add a $10\,\mathrm{kg}$ payload to the robot, and update the model parameters in all sampling-based MPCs accordingly.

Quadruped Crate Climbing In fig. 9, we demonstrate DIAL-MPC's ability to achieve RL-level agility in 1/15 the speed of real world. The quadruped is tasked with climbing onto the $0.6\,\mathrm{m}$ high

	Crate Climbing
Target Position	0.5
Upright	0.01
Energy	0.001
Target Yaw	0.3
Contact Reward	0.2

Table 5: Reward weights of the crate climbing task.

crate, which is more than twice the standing height of the robot. Similar tasks have been demonstrated on several recent quadruped parkour works with RL [5], which report around 20 hours of training time for a single policy with perception. Although DIAL-MPC's integration with perception demands future work, it is currently comparable to a teacher policy with access to privileged environment information.

We use only five reward terms to guide DIAL-MPC to complete the task. They are shown in Table 5. Contact reward is given to feet on the top surface of the crate. Since this is a non-real-time planning task, we use 4096 parallel samples, 40-step (1 s) horizon, and 4 annealing steps. We also enable all contacts on the base and thighs. We then roll out the simulation synchronously after DIAL-MPC finishes computing at every step, which in total takes 30 seconds to generate a 2-second full-state motion plan with zero prior knowledge or training.

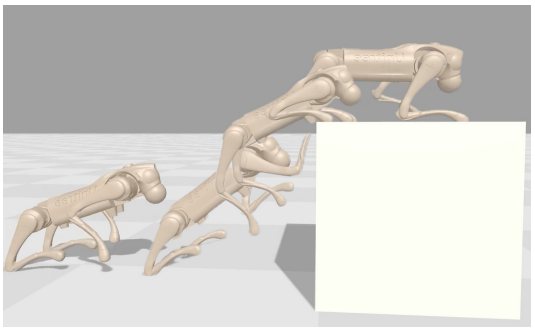

Figure 9: DIAL-MPC generates this well-planned agile motion in 30 seconds, training free. The quadruped leaps forward to catch the edge of the crate with its forelegs, launches off the ground with its hindlegs, and quickly retracts to a standing pose on the crate.

Humanoid Crate Pushing

Figure 5 shows DIAL-MPC performing whole-body control task on a Unitree H1 humanoid in simulation. The robot is asked to track a slow velocity of $0.8\,\mathrm{m/s}$ while pushing forward a crate of $30\,\mathrm{kg}$ and $15\,\mathrm{kg}$ respectively. The robot is rewarded for maintaining the correct speed and the correct contact with its two upper end effectors. The seven reward weights are listed in Table 6. This task shows that DIAL-MPC generalizes well to systems with higher degrees of freedom (25 for the humanoid and 18 for the quadruped) and is capable of handling whole-body control tasks with versatile parameters. Again we highlight that all reward terms in this experiment are straight forward, and that it takes no time to visualize the outcome of reward tuning since DIAL-MPC is a training-free algorithm.

	Humanoid Crate Pushing
Gait	5.0
Upright	0.01
Yaw	0.1
Velocity	1.0
Torso Height	0.5
Energy	0.01
Contact Reward	0.05

Table 6: Reward weights of the humanoid crate pushing task.