

CA-AC-MPC: CUDA-Accelerated Actor-Critic Model Predictive Control

Antonio Buo, Vittorio Cammarota, Michele Avagnale, Pierluigi Arpentì, Vincenzo Lippiello, Fabio Ruggiero

Abstract—In the literature, actor-critic model predictive control (AC-MPC) integrates MPC with reinforcement learning to enable high-performance control of complex dynamical systems. However, its differentiable MPC layer requires repeatedly solving an optimization problem in both the forward and backward passes, leading to substantial training and inference latency. This paper tackles this bottleneck introducing a CUDA-accelerated variant that significantly reduces end-to-end execution time while preserving the control performance of the baseline formulation. Simulation results on an agile drone racing task show that our approach achieves state-of-the-art lap times and near-limit dynamic behaviour with markedly reduced training and inference time.

SUPPLEMENTARY MATERIAL

Code can be found at <https://github.com/prisma-lab/CA-AC-MPC>.

LIST OF MOST USED ACRONYMS

Acronym	Meaning
AC-MPC	Actor-Critic Model Predictive Control
AC-MLP	Actor-Critic Multilayer Perceptron
CA-AC-MPC	CUDA-Accelerated Actor-Critic Model Predictive Control
CA-DiffMPC	CUDA-Accelerated Differentiable Model Predictive Control
CUDA	Compute Unified Device Architecture
DiffMPC	Differentiable Model Predictive Control
MLP	Multilayer Perceptron
MPC	Model Predictive Control

I. INTRODUCTION

Control design is especially difficult in high-performance robotic applications, where robots operate in dynamic conditions with unmodelled effects, parameter uncertainties, and rapid transients. The bottleneck of classical model-based control strategies—including online optimization methods such as MPC—is not the model-based paradigm per se, but the mismatch between model fidelity and computational tractability [1]. Pushing performance often requires dynamical models that are either too inaccurate to capture the relevant phenomena or too expensive to be used in real-time [2].

This gap motivated growing interest in reinforcement learning (RL) and other data-driven approaches, which can optimize control policies directly from task performance

without requiring an explicit high-fidelity model. While RL still requires significant design choices (e.g., reward shaping, curriculum design, hyperparameter tuning, etc.), it provides a flexible framework in which feedback policies are learned through interaction with the environment. In particular, RL can handle non-differentiable and structured objectives without relying on proxy references such as pre-defined trajectories or paths. This has enabled state-of-the-art results in highly dynamic robotic tasks where model-free RL has been shown to outperform classical optimal control in near-limit conditions [3], [4]. However, RL approaches typically require massive amounts of data—often millions of simulated interactions.

These challenges have led research to focus on hybrid control frameworks that combine the structure and reliability of model-based optimal control with the flexibility and task-driven optimization of RL. Among these, AC-MPC emerged as a promising paradigm, in which an MPC-based actor provides short-horizon optimal control actions while a learned critic captures long-term value information [5]. This integration yields policies that retain the interpretability and constraint-handling capabilities of MPC while benefiting from the adaptability and performance of RL [2]. Nevertheless, embedding an optimization-based control layer within a neural policy introduces a computational burden, since each policy evaluation requires solving a parameterized optimal control problem (OCP) rather than executing a simple feed-forward pass [6], [7]. Moreover, MPC typically assumes access to an accurate *differentiable* dynamic model [2], [7]. In the AC-MPC, the OCP must be solved during the forward pass to compute the control action, and an additional linear system must be solved during the backward pass to propagate gradients, increasing training time. During inference, gradients are not required and only the forward MPC pass is performed; however, the policy remains substantially slower than a standard AC-MLP [5].

This paper addresses this limitation by introducing a CUDA-accelerated actor-critic model predictive control framework, referred to as CA-AC-MPC. Unlike the original PyTorch-based implementation [2], [8], which relies on PyTorch’s GPU backend, our approach directly invokes CUDA kernels for the most computationally intensive operations. This provides finer-grained control over GPU parallelism, reduces overhead, and accelerates both the forward and backward passes of the differentiable MPC layer.

The specific contribution of this work is a fused-kernel CUDA implementation of iterative linear quadratic regulator (iLQR) for differentiable MPC. We present a C++/CUDA PyTorch extension that restructures the iLQR computation

This work was partially supported by the Italian Ministry of University and Research, under the complementary actions to the NRRP ‘Fit4MedRob - Fit for Medical Robotics’ Grant (# PNC0000007). This research was also partially supported by the INVITALIA Project NEMESI (CUP C67G22000420008).

¹Authors are with PRISMA Lab and CREATE Consortium, Department of Electrical Engineering and Information Technology, University of Naples Federico II, Naples, Italy. Corresponding author’s email: [fabio.ruggiero@unina.it].

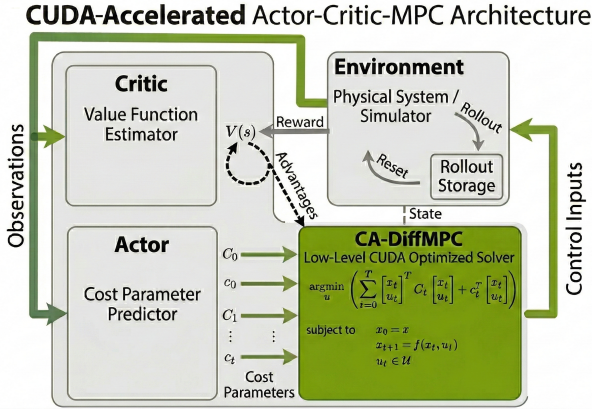


Fig. 1: Inspired from the initial picture in [5], we present the overview of the proposed CA-AC-MPC framework, highlighting the introduced low-level CUDA optimized solver.

into three fused stages (rollout and linearization, Riccati backward pass with box constraints, and line-search rollout). This replaces the Python-level recursion over the horizon with a fixed number of CUDA kernel launches per iteration. We refer to this implementation as CA-DiffMPC. When integrated into the AC-MPC framework, CA-DiffMPC gives rise to our CA-AC-MPC (see Fig. 1), yielding an overall speed-up of one to two orders of magnitude in our simulations, while also reducing runtime inference. The proposed framework is validated on an agile drone-racing task, where CA-AC-MPC achieves lap-time performance comparable to, and in some cases better than, an end-to-end AC-MLP baseline, consistent with the performance reported for AC-MPC in [9]. We release the source code to facilitate future research¹.

II. RELATED WORKS

The integration of MPC with RL has its roots in differential dynamic programming and dual control. Early approaches such as guided policy search [10] and MPC-guided policy learning [11] leveraged MPC as a supervisory signal or data generator for policy training. Recent developments in DiffMPC [7] have enabled the inclusion of OCPs as differentiable layers within neural networks, allowing gradient-based policy optimization [5].

Nevertheless, DiffMPC-based approaches remain computationally demanding. In particular, end-to-end training of the actor requires computing gradients of the MPC solution with respect to the cost parameters predicted by the neural network. Since the MPC solver does not define a closed-form mapping, naive automatic differentiation through the solver iterations (i.e., unrolling) is both computationally expensive and numerically unstable [12]. An alternative is implicit differentiation, which computes gradients by differentiating the Karush–Kuhn–Tucker (KKT) optimality conditions at convergence, thus avoiding the need to store the full computational graph [7]. A complementary direction to mitigate this computational burden is to exploit parallel hardware acceleration. For instance, a GPU-based MPC framework

for legged locomotion that enables both temporal and state-space parallelization via a parallel associative scan to solve the primal–dual KKT system is proposed in [13]. Recently, a GPU-accelerated DiffMPC solver based on sequential quadratic programming, combined with a custom preconditioned conjugate gradient routine and structure-exploiting tridiagonal preconditioning, was introduced in [14], enabling efficient parallelization of the backward pass. However, existing approaches predominantly address the computational bottleneck of DiffMPC at the algorithmic level [13], [14]. As acknowledged in those works, GPU acceleration can become ineffective—and in some cases slower than CPU-based implementations—for small- to medium-scale problems, where temporal parallelism does not amortize kernel launch overhead because the per-time step arithmetic is relatively cheap. This regime is highly relevant in robotics, where state and control dimensions are often moderate and dispatch latency can dominate end-to-end execution time. The proposed CA-AC-MPC targets this dispatch overhead directly.

III. METHODOLOGICAL BACKGROUND

In this section, the fundamental concepts underlying the AC-MPC framework are briefly summarized. The reader can refer to the cited works for a more detailed discussion.

A. Economic Model Predictive Control

Let $n_x \in \mathbb{N}$ and $n_u \in \mathbb{N}$ denote the numbers of states and control inputs, respectively, of a given discrete-time dynamical system with nonlinear state-transition map $f(\cdot, \cdot) : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \rightarrow \mathbb{R}^{n_x}$, i.e., $x_{t+1} = f(x_t, u_t)$, where $x_t \in \mathcal{X} \subseteq \mathbb{R}^{n_x}$ denotes the state of the system, belonging to the set $\mathcal{X}$, at time t , and $u_t \in \mathcal{U} \subseteq \mathbb{R}^{n_u}$ denotes the control input of the system, belonging to the set $\mathcal{U}$, at time t . Consider the discrete-time finite-horizon OCP [15]

$$\begin{aligned} \min_{u_0, \dots, u_{T-1}} \quad & \sum_{t=0}^{T-1} \ell_t(x_t, u_t) + V_f(x_T) \\ \text{s.t.} \quad & x_{t+1} = f(x_t, u_t), \quad t = 0, \dots, T-1 \\ & u_t \in \mathcal{U}, \quad t = 0, \dots, T-1 \\ & x_0 = x_{\text{init}} \end{aligned} \quad (1)$$

where $\ell(\cdot, \cdot) : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \rightarrow \mathbb{R}$ is the stage cost, $V_f(\cdot) : \mathbb{R}^{n_x} \rightarrow \mathbb{R}$ is the terminal cost, and $x_{\text{init}} \in \mathbb{R}^{n_x}$ and $x_T \in \mathbb{R}^{n_x}$ are the initial and terminal states of the system, respectively. In MPC, at each sampling instant, the finite-horizon optimization problem (1) is solved using the current state at time step t as the initial condition x_{init} [16]. Subsequently, only the first control input u_0^* of the resulting optimal control sequence is applied to the closed-loop system. The procedure is then repeated at the next time step using the newly measured state [15].

In standard tracking MPC, the stage cost $\ell_t(\cdot, \cdot)$ is typically chosen as a positive-definite quadratic function that penalizes deviations from a prescribed trajectory/setpoint. A broader class of formulations is referred to as economic MPC [17], in which the tracking cost is replaced by a more general

¹<https://github.com/prisma-lab/CA-AC-MPC>.

stage cost encoding a task-specific performance objective. This flexibility introduces significant challenges. In many tasks, the stage cost $\ell_t(\cdot, \cdot)$ is sparse, non-smooth, or non-differentiable, complicating both the formulation and the numerical solution of the OCP. Indeed, performance indicators of practical interest are often not smooth by nature [18]: they may exhibit discontinuities, piecewise definitions, or abrupt changes depending on the operating regime. Hence, constructing a globally smooth and informative cost function is often difficult. The lack of differentiability is particularly problematic for gradient-based optimization method. This often requires smoothing approximations, differentiable surrogate costs, or dedicated non-smooth optimization techniques. A common strategy to address these issues in economic MPC is to approximate the original stage cost with a quadratic surrogate that is more amenable to numerical optimization. Defining $z_t = [x_t^\top \ u_t^\top]^\top \in \mathbb{R}^{n_x+n_u}$, the stage cost can be approximated as

$$\ell(z_t) \approx \frac{1}{2} z_t^\top C_t z_t + c_t^\top z_t, \quad (2)$$

where $C_t \in \mathbb{R}^{(n_x+n_u) \times (n_x+n_u)}$ is a positive semidefinite cost matrix and $c_t \in \mathbb{R}^{n_x+n_u}$ collects the cost parameters. The quadratic surrogate in (2) yields a smooth objective and makes the OCP differentiable with respect to (C_t, c_t) , which is essential when embedding MPC as a differentiable layer within a learning pipeline. However, in economic control tasks, the relevant objective is often time-varying and context-dependent [17]; therefore, fixed hand-tuned parameters may fail to generalize across operating regimes.

B. Actor-Critic Policy Architecture

In RL, the interaction between an agent and its environment is commonly formalized as a Markov decision process (MDP), described by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$, where $\mathcal{S}$ is the state space (with states $s \in \mathcal{S}$), $\mathcal{A}$ is the action space (with actions $a \in \mathcal{A}$), $P(s' | s, a) = \Pr(s_{t+1} = s' | s_t = s, a_t = a)$ is the state-transition probability, $R(s, a) = \mathbb{E}[r_t | s_t = s, a_t = a]$ is the expected immediate reward function (with reward r), and $\gamma \in [0, 1]$ is the discount factor. One way to solve these MDP problems is by using temporal difference (TD) methods. The class of TD algorithms updates a value estimate $V(s)$ toward a target that combines the observed reward and the estimate at the next state. The update for a state-value function under policy π is given by $V(s_t) \leftarrow V(s_t) + \alpha [r_t + \gamma V(s_{t+1}) - V(s_t)]$ where $\alpha > 0 \in \mathbb{R}$ is a learning rate and the temporal-difference error $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$ quantifies the mismatch between successive value estimates [19]. The actor-critic class of algorithms integrates policy optimization with TD value estimation. The actor, with parameters $\theta \in \mathbb{R}^{n_\theta}$, represents a stochastic policy $\pi_\theta(a | s)$ and it is responsible for choosing actions. The critic, with parameters $\phi \in \mathbb{R}^{n_\phi}$, estimates a value function that evaluates the quality of the actor's actions. While the critic is typically trained using TD updates to reduce prediction error, the actor is updated via policy gradients that rely on the critic's evaluation as a performance signal. Let $N \in \mathbb{N}$ denote the number of

sampled trajectories and $L \in \mathbb{N}$ the episode length. The actor parameters θ are optimized to maximize the expected discounted return $J(\pi_\theta) \in \mathbb{R}$ via the policy-gradient estimate

$$\nabla_\theta J(\pi_\theta) = \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^L \nabla_\theta \log \pi_\theta(a_t^i | s_t^i) A_\phi(s_t^i, a_t^i), \quad (3)$$

where the advantage function A_ϕ is approximated using the TD error, $A_\phi(s_t, a_t) \approx r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$. Here, s_t^i and a_t^i denote the state and action of the i -th sampled trajectory at time t , respectively, and actions are sampled from the stochastic policy as $a_t \sim \pi_\theta(\cdot | s_t)$.

C. Actor-Critic Model Predictive Control

In the AC-MPC architecture [5], [9], the MPC module is embedded as the final layer of the actor network within a standard actor-critic pipeline trained with proximal policy optimization. By casting MPC as a differentiable layer, gradients can be propagated through the underlying optimal control problem, tightly coupling policy learning with model-based control. Instead of manually specifying task-dependent cost functions, the MPC objective is parameterized by a neural network that outputs quadratic cost coefficients at each time step. This learned cost representation allows the control objective to be shaped directly by the reinforcement signal while preserving key advantages of MPC, including constraint handling and generalization across the state space. To promote exploration during training, control actions are sampled around the MPC solution via a stochastic policy. At deployment time, the deterministic MPC solution is applied directly, retaining the structure of a model-based controller. Algorithm 1 summarizes the resulting AC-MPC workflow.

In detail, the AC-MPC framework combines two aspects: it is *economic* in the objective design and *differentiable* in the solver implementation. The cost structured adopted by the AC-MPC is (2) [5]. At each timestep, the actor network maps the current observation to the quadratic cost parameters (C_t, c_t) , defining the stage objective optimized by the MPC controller online. In parallel, the differentiable iLQR-based solver lets gradients of the RL loss propagate back through the MPC layer, enabling end-to-end actor optimization in an actor-critic setting. Thus, economic specifies what is optimized at runtime, while differentiable specifies how the policy is trained via the underlying optimization.

A key limitation of this architecture is computational overhead: each policy evaluation requires solving an MPC problem, and end-to-end training additionally requires differentiating through the solver—typically via the solution of an auxiliary linear system—to propagate gradients. As a result, AC-MPC incurs substantially higher training and inference latency than a standard feed-forward MLP. Although PyTorch provides highly optimized CUDA kernels for many tensor operations [8], differentiable MPC workloads are often composed of repeated horizon-wise recursions and numerous small-to-medium tensor operations. In this regime, end-to-end performance is frequently limited not by raw GPU throughput, but by Python-side orchestration and kernel launch overhead associated with many short-running kernels.

Algorithm 1 AC-MPC deployment.

Require: Trained actor π_θ , dynamics f , horizon T , iLQR iterations K , bounds $\mathcal{U}$

- 1: **for** $t = 0, 1, 2, \dots$ **do**
- 2: Observe state x_t
- 3: $\{(C_t, c_t)\}_{0:T-1} \leftarrow \pi_\theta(x_t)$ $\triangleright$ stage-cost parameters over the horizon
- 4: $u_t^* \leftarrow \text{DIFFMPC}(x_t, \{(C_t, c_t)\}_{0:T-1}, f, \mathcal{U}, K)$ $\triangleright$ DIFFMPC: iLQR forward solve [7]
- 5: Apply u_t^* to the system $\triangleright$ receding horizon
- 6: **end for**

IV. CUDA-ACCELERATED ACTOR-CRITIC MODEL PREDICTIVE CONTROL

Motivated by the computational bottlenecks highlighted in Section III-C, we propose the CA-AC-MPC, a CUDA-accelerated extension of the AC-MPC. Rather than relying solely on standard tensor-level dispatch, the proposed implementation leverages fused custom C++/CUDA operators that merge multiple computational steps into single kernel launches. Although graph-based batching (e.g., CUDA Graphs) can also mitigate these costs when tensor shapes and control flow are sufficiently static, our design targets the differentiable MPC workload directly and yields a practical reduction in end-to-end execution time for both training and inference [20], [21].

We consider the same box-constrained control problem adopted by the reference differentiable MPC implementation in [8], which we denote hereafter as the baseline backend DiffMPC². The resulting control problem is

$$\begin{aligned} \min_{x_t, u_t} \quad & \sum_{t=0}^{T-1} \frac{1}{2} z_t^\top C_t z_t + c_t^\top z_t \\ \text{s.t.} \quad & x_{t+1} = f(x_t, u_t), \quad x_0 = x_{\text{init}}, \\ & u_{\min} \leq u_t \leq u_{\max}, \quad t = 0, \dots, T-1, \end{aligned} \quad (4)$$

where (C_t, c_t) parametrize the quadratic surrogate cost predicted by the actor network. Gradients of an external loss $\mathcal{L}$ with respect to (C_t, c_t) and x_{init} are recovered via implicit differentiation of the KKT conditions at the converged solution [7], making the backward pass independent of the number of solver iterations. The forward solver follows the control-limited iLQR formulation [22]. During AC-MPC training with proximal policy optimization (PPO), the forward solve is invoked at every rollout step; for a single training run with horizon T and K iLQR iterations, this results in millions of MPC solves. In DiffMPC [8], each iLQR phase (linearization, Riccati backward sweep, and line-search rollout) is implemented via a Python-level `for` loop over the horizon T , triggering $\mathcal{O}(T)$ GPU kernel launches per phase and per iteration. When n_x and n_u are small to moderate (e.g., below 10^2), each operation completes in a few microseconds, so host-side dispatch and kernel launch overhead can dominate overall runtime. Our

contribution is a C++/CUDA extension that fuses each iLQR phase into a single kernel launch, eliminating this dispatch overhead while retaining the same mathematical formulation and implicit differentiation targets.

The underlying solver follows a standard iLQR structure and it is implemented as a three-stage loop, described below. These stages are mapped to three fused CUDA kernels, invoked sequentially at each iLQR iteration.

1) *Rollout and linearization*: Given the current nominal inputs, the kernel computes the state rollout $x_{t+1} = f(x_t, u_t)$ and the analytical Jacobians $A_t = \partial f / \partial x \in \mathbb{R}^{n_x \times n_x}$ and $B_t = \partial f / \partial u \in \mathbb{R}^{n_x \times n_u}$ for all batch elements and timesteps in a single pass. Analytical derivatives avoid storing large autograd graphs (i.e., the data structures employed by automatic differentiation to track operations for gradient computation), reducing memory usage and directly providing the linear time-varying model required for the iLQR backward pass.

2) *Backward pass with input bounds*: The second kernel runs the iLQR/DDP backward pass (Riccati recursion) over the horizon, yielding $\delta u_t = K_t \delta x_t + k_t$ (unconstrained) where $K_t \in \mathbb{R}^{n_u \times n_x}$ and $k_t \in \mathbb{R}^{n_u}$ are the feedback and feedforward gains, respectively. With box constraints $u_{\min} \leq u_t \leq u_{\max}$, the stagewise control update is obtained by solving the corresponding box-constrained quadratic subproblem [22] in CUDA. The recursion is sequential in time and parallelized across batch elements.

3) *Forward pass with line-search*: We denote by $\alpha \in (0, 1]$ the scalar line-search coefficient used to scale the feedforward term. In practice, the third kernel evaluates a small set of candidate values, $\alpha \in \{1.0, 0.5, 0.25, 0.1\}$, in parallel. For each candidate, the kernel applies the affine feedback update and rolls out the dynamics to obtain the new trajectory and cost. The best candidate is selected and used to update the nominal solution for the next iLQR iteration.

The three stages outlined above constitute our core contribution, the CA-DiffMPC. Compared with a Python-level recursion over time (which typically incurs $\mathcal{O}(T)$ GPU operator launches per phase), CA-DiffMPC requires only three kernel launches per iLQR iteration. Embedding CA-DiffMPC as the final differentiable layer of an actor-critic network yields the overall framework, CA-AC-MPC, whose workflow is summarized in Algorithm 2. The mathematical formulation and implicit differentiation strategy of the proposed approach are unchanged with respect to DiffMPC [7], [8]. Our contribution is thus primarily architectural: we replace per-timestep Python-level loops with three fused CUDA kernel launches per iLQR iteration.

V. NUMERICAL EVALUATION

In this section, we numerically evaluate the proposed CA-AC-MPC framework and compare it with state-of-the-art methods. We assess task performance in a simulated drone racing scenario following [2], [4], [9]. Since the exact SplitS track geometry used in prior studies is not publicly available, we train and evaluate on a reconstructed layout

²In [8], this backend is referred to as `mpc.pytorch`.

Algorithm 2 CA-AC-MPC deployment.

Require: Actor π_θ , dynamics f , horizon T , max iter K , bounds $\mathcal{U}$

- 1: Initialize CA-DIFFMPC backend
- 2: **for** $t = 0, 1, 2, \dots$ **do**
- 3: Observe state x_t
- 4: $(C, c)_{0:T-1} \leftarrow \pi_\theta(x_t)$ $\triangleright$ Cost parameters
- 5: $(X^0, U^0) \leftarrow \text{ROLLOUT}(x_t, f)$
- 6: **for** $k = 1 \dots K$ **do**
- 7: $(A, B, \ell)_{0:T-1} \leftarrow \text{LQ}(f, (C, c), X^{k-1}, U^{k-1})$ $\triangleright$ Quadraticize cost function
- 8: $(K, k)_{0:T-1} \leftarrow \text{FBP}((A, B, \ell), \mathcal{U})$ $\triangleright$ Fused backward pass
- 9: $(X^k, U^k) \leftarrow \text{FFP}(x_t, U^{k-1}, (K, k))$ $\triangleright$ Fused forward pass + line-search
- 10: **if** $\text{CONVERGED}(X^k, U^k)$ **then**
- 11: **break**
- 12: **end if**
- 13: **end for**
- 14: $u_t^* \leftarrow U^k[0]$ $\triangleright$ Extract current action
- 15: Apply u_t^* to system $\triangleright$ Receding horizon
- 16: **end for**

based on available descriptions³. For consistency, we report lap-time comparisons only within our environment and do not compare absolute lap times with those in the literature. The reconstructed track is available in our provided code.

The comparison evaluation is thus organized considering four metrics: ($\mathbb{M}_1$) standalone CA-DiffMPC forward- and backward-pass latencies benchmarked against DiffMPC; ($\mathbb{M}_2$) CA-AC-MPC policy training time; ($\mathbb{M}_3$) CA-AC-MPC policy inference time; and ($\mathbb{M}_4$) closed-loop task performance. Metrics $\mathbb{M}_2 - \mathbb{M}_4$ are evaluated across three MPC horizon lengths ($T \in \{2, 5, 10\}$) and three iLQR iteration counts ($K \in \{1, 5, 10\}$). Metrics $\mathbb{M}_1 - \mathbb{M}_3$ are reported to quantify the computational advantages of the proposed implementation. In $\mathbb{M}_1$, we compare solver latency in isolation by benchmarking the CA-DiffMPC against the baseline DiffMPC, without integrating the actor-critic components. Metrics $\mathbb{M}_2$ and $\mathbb{M}_3$ evaluate the end-to-end training and inference latency of the complete CA-AC-MPC stack against the original AC-MPC implementation of [9]. However, it is worth citing that, during these comparisons, we observed persistent memory leaks when running the public AC-MPC codebase [9] with the DiffMPC backend⁴. A practical consequence of this issue is that the baseline AC-MPC policy could not be trained to converge in our setup, preventing a direct, like-for-like, closed-loop comparison in metrics $\mathbb{M}_4$.

Regarding the learning setup, the actor predicts the stage-cost coefficients via a neural cost map implemented as a two-layer MLP—512 hidden units per layer, rectified linear

³Gates were reconstructed from top-down views and metric scales reported in the cited papers; gate heights are more uncertain.

⁴See: <https://github.com/uzh-rpg/acmpc-public/issues/1>

TABLE I: Training hyperparameters values.

Setting	Value
Discount factor γ	0.99
GAE λ	0.95
Steps per update	256
Mini-batch size	2048
SGD epochs	10
Clip range	0.2
Learning rate	$3 \times 10^{-4} \rightarrow 3 \times 10^{-5}$
Entropy coef.	0.0
Value-loss coef.	0.5
Grad. clip	0.5

TABLE II: Execution time comparison (in milliseconds) between CA-DiffMPC and DiffMPC.

B	T	Forward time		Backward time	
		CA-DiffMPC	DiffMPC	CA-DiffMPC	DiffMPC
1	2	0.93	9.92	0.32	4.81
	10	1.08	41.0	0.80	9.16
	50	1.84	492	2.35	30.1
256	2	1.21	25.4	0.38	9.71
	10	1.68	90.3	0.87	14.5
	50	4.00	674	2.60	37.3

unit (ReLU) activations—with a sigmoid output layer that enforces the same scaling and bounds as the reference implementation. The critic is a two-layer ReLU MLP with 512 hidden units per layer and outputs a scalar value. All other implementation details follow the original AC-MPC formulation and its public codebase. Training hyperparameters are reported in Table I. All simulations were conducted on a desktop workstation equipped with an Intel Core i9-13900K CPU and an NVIDIA RTX A6000 GPU.

A. Solver latency comparison ($\mathbb{M}_1$)

To evaluate $\mathbb{M}_1$ we did not consider the drone racing scenario. Here, the MPC aims at stabilizing the drone at random target hovering positions, and the time required to compute each control update is recorded. The task is performed for three prediction horizons ($T \in \{2, 10, 50\}$), both for a single environment ($B = 1$) and for a batch of $B = 256$ environments to characterize scaling under parallel execution. Table II reports the corresponding numerical values. The proposed CA-DiffMPC achieves consistently low forward-pass latency in the single-instance setting compared to the baseline DiffMPC and demonstrates effective scaling when evaluated on large batches.

B. Training time comparison ($\mathbb{M}_2$)

Here we performed shortened training runs under identical conditions for both solvers. In preliminary evaluations with AC-MPC (which uses DiffMPC as the differentiable MPC backend [7], [8]), random access memory usage increased steadily until the process terminated, preventing completion of the full 10^7 training steps. We attribute this behaviour to two likely causes: (i) the LQRStep closure factory instantiated a new `torch.autograd.Function`

TABLE III: Policy training times for the AC-MLP, the AC-MPC, and the CA-AC-MPC.

AC-MLP		16m:18s		
	T	$K = 1$	$K = 5$	$K = 10$
AC-MPC	2	–	2h:15m:03s	–
CA-AC-MPC	2	21m:22s	27m:32s	30m:30s
CA-AC-MPC	5	21m:55s	37m:03s	29m:01s
CA-AC-MPC	10	21m:34s	30m:29s	33m:22s

subclass on each call, causing unbounded accumulation of class-level metadata in PyTorch’s autograd engine; (ii) the iLQR iteration loop executed with gradient tracking enabled, constructing large autograd graphs during each of the ~ 80 MPC solves per PPO update that were never backpropagated through. We patched the library by replacing the closure factory with a single module-level `Function` class, executing the iteration loop within a `torch.no_grad()` context, and invoking `malloc.trim` after each rollout to reclaim fragmented heap memory. These modifications remove only unnecessary autograd bookkeeping and do not affect the solver’s numerical output. With these patches, memory usage plateaus, and the full training completes successfully. The resulting training times are reported in Table III. Across both horizons, our implementation (see Fig. 2 for CA-AC-MPC training throughput) consistently reduces training time relative to AC-MPC, decreasing runtime from more than two hours to about 27 minutes (about $5.4\times$ speedup). For reference, training a policy based solely on an AC-MLP (without the MPC in the loop) requires roughly 16 minutes under the same hardware setup. While this indicates that MPC-based training remains more computationally demanding, the overhead introduced by our solver is limited to a moderate increase (about $1.3\times$ with respect to the AC-MLP). For the sake of transparency, we note that, under our setup, the AC-MPC absolute training times differ significantly from those reported in [9]. This discrepancy is likely due to differences in stopping criteria, hardware, and other implementation details not stated in public repositories. Moreover, since the AC-MPC baseline did not reliably converge in our simulations, the only result reported is the one achieving convergence, also allowing for inference time comparison.

C. Policy Inference Times ($\mathbb{M}_3$)

The proposed solver achieves substantial improvements in policy inference latency across all tested configurations (Table IV). Relative to the inference times reported in [5], [9], we observe an approximately $10\times$ speedup in the baseline AC-MPC configuration ($T = 2$, five iLQR iteration). For longer horizons and multiple quadratic iterations, the performance improvements become even more pronounced. The improved scaling stems from the constant number of GPU kernel launches in our implementation, which avoids the horizon-dependent overhead present in previous approaches. As a result, inference time grows significantly more slowly with problem complexity. For the same reasons reported at

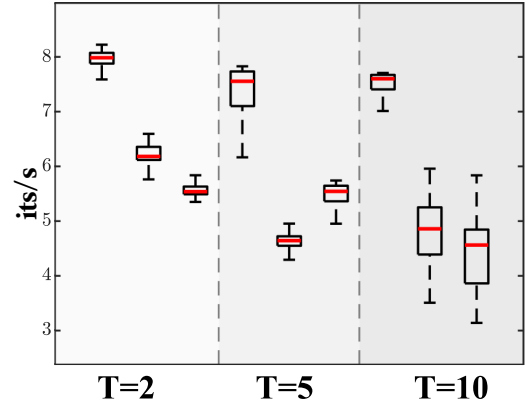


Fig. 2: CA-AC-MPC training throughput, measured in thousands of iterations per second. Results are grouped by prediction horizon, from left to right ($T \in \{2, 5, 10\}$). Within each group, bars represent varying numbers of iLQR iterations ($K \in \{1, 5, 10\}$, from left to right, respectively).

TABLE IV: Policy inference times (in milliseconds) for the AC-MLP, the AC-MPC, and the CA-AC-MPC.

AC-MLP		0.245 ± 0.01			
	T	$K = 1$	$K = 5$	$K = 10$	
AC-MPC	2	–	11.6 ± 0.9	–	
CA-AC-MPC	2	1.26 ± 0.06	1.26 ± 0.06	1.26 ± 0.07	
CA-AC-MPC	5	1.43 ± 0.08	2.39 ± 0.08	2.32 ± 0.17	
CA-AC-MPC	10	1.60 ± 0.01	2.83 ± 0.10	2.99 ± 0.30	

the end of Section V-B, comparison with AC-MPC baseline was possible only with the case achieving convergence.

D. Policy Performance ($\mathbb{M}_4$)

This section evaluates the performance of the trained policy across different prediction horizons. The resulting trajectories are illustrated in Fig. 3, while a top-down view featuring a velocity-based color map is provided in Fig. 4. Although the peak velocity remains comparable across the evaluated trajectories, the optimal policy maintains a consistently higher speed throughout the entire track. This is particularly evident between gates 2 and 4, before the split-S maneuver, whereas the alternative methods exhibit noticeable deceleration when approaching the waypoints.

As shown in Fig. 4, all trained policies achieve peak velocities of approximately 20 m/s, demonstrating the ability to operate the platform near its dynamic limits. The primary qualitative difference among the policies lies in the geometry of the resulting trajectories: increasing the number of iterations leads to progressively smoother paths.

Table V reports the achieved lap times to provide a comprehensive assessment of closed-loop performance. As said, direct comparisons with [5], [9] are intentionally omitted because the same SplitS track geometry was not publicly available. Consequently, the reported results are intended as controlled evaluation of the performance gains unlocked by the proposed solver, whose improved computational efficiency enables the use of longer MPC horizons and more

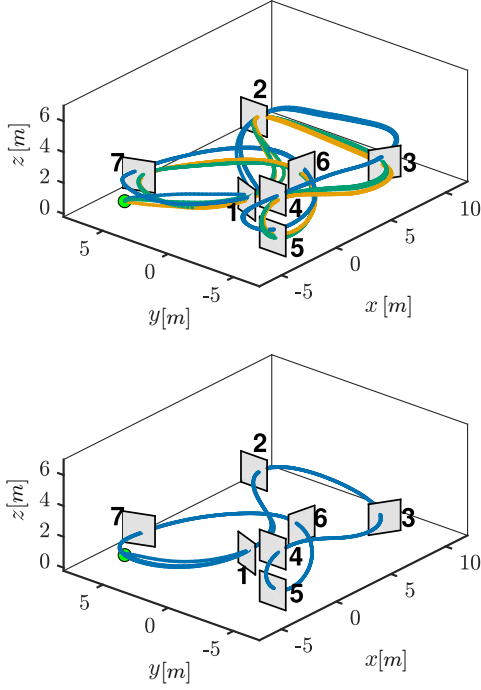


Fig. 3: Simulated agile drone flight trajectories for horizon $T = 2$ (top) under different iLQR iteration counts ($K \in \{1, 5, 10\}$, shown in blue, orange, and green, respectively) and $T = 5$ (bottom). The green circle indicates the starting position. Gates are numbered.

TABLE V: Lap times (in seconds) for the AC-MLP and the CA-AC-MPC.

	T	K	Lap time [s]
AC-MLP	—	—	5.32
CA-AC-MPC	2	5	5.26
CA-AC-MPC	2	10	5.10
CA-AC-MPC	5	5	4.98

solver iterations. The most stable and effective results were obtained for horizons $T = 2$ and $T = 5$. Increasing the horizon to $T = 10$, the resulting MPC problem requires the prediction of 140 optimization variables, increasing the problem complexity. Given the limited capacity of the adopted neural network architecture, this led to training instability and ultimately prevented to a policy convergence.

VI. DISCUSSION

Despite the substantial reduction in solver latency enabled by CA-AC-MPC, several limitations remain. Increasing the MPC horizon in the CA-AC-MPC framework tends to improve performance in agile tasks (see Table V) but it also increases the number of cost parameters that the actor must predict. As the horizon grows, a simple feed-forward AC-MLP struggles to output temporally consistent costs, leading to noisy or inconsistent objectives across timesteps. This can slow down training and reduce robustness.

A promising direction is to enforce structure in the cost parametrization: a move blocking cost could be a solution, or use temporal architectures that explicitly model time/progress dependencies [23]. Our solver enforces box constraints on the control inputs within the iLQR/DDP update; however, general state constraints are not directly supported by the underlying differentiable MPC layer and are currently handled through differentiable barrier/penalty terms. While this preserves differentiability, it does not provide hard constraint guarantees and may increase the learning burden, as the actor must learn both task optimality and implicit constraint satisfaction. Agile flight and racing objectives vary with track progress and maneuver phase. Thus, if the MPC cost is parameterized by a purely feed-forward network, coherent time-varying cost schedules generally require explicit progress/time inputs or a sequence model; otherwise, the policy may struggle to maintain consistent objectives over long horizons.

The three-kernel-per-iteration structure also makes CA-DiffMPC a promising candidate for embedded GPU deployment. By replacing horizon-wise PyTorch recursions with a compact, horizon-independent CUDA launch pattern, the solver reduces CPU-side dispatch overhead, a key concern for small-scale MPC workloads on platforms such as NVIDIA Jetson Orin. Although hardware validation remains future work, this architecture suggests that deployment on realistic robotic compute platforms should preserve much of the observed latency advantage.

VII. CONCLUSION

In this work, we extend AC-MPC with CUDA-based GPU acceleration. Our resulting CA-AC-MPC substantially reduces both training and inference time relative to the baseline implementation. The speedups come from faster differentiable MPC solves across horizon lengths and batch sizes, enabled by a fused-kernel CUDA iLQR implementation for differentiable MPC, which we call CA-DiffMPC. Overall, CA-DiffMPC yields approximately at least one order-of-magnitude speed-ups in forward and backward solve time for single instances, and at least $20\times$ for batched problems. As a result, the proposed CA-AC-MPC increases end-to-end training time by only about 30% relative to the AC-MLP baseline, whereas the reference AC-MPC implementation requires approximately $30\times$ the AC-MLP training time [5]. This substantially reduces the latency introduced by embedding a differentiable MPC layer within the actor-critic paradigm, which constitutes the main bottleneck of the original architecture. Crucially, these computational advantages are obtained without compromising closed-loop performance in a drone racing scenario. Although a direct, like-for-like comparison against the full baseline AC-MPC was not possible due to issues in the public repository, our simulations show that CA-AC-MPC enables near-limit dynamic behaviour consistent with [5].

REFERENCES

- [1] Z. Xie, Y. Wang, X. Luo, P. Arpent, F. Ruggiero, and B. Siciliano, “Three-dimensional variable center of mass height biped walking using

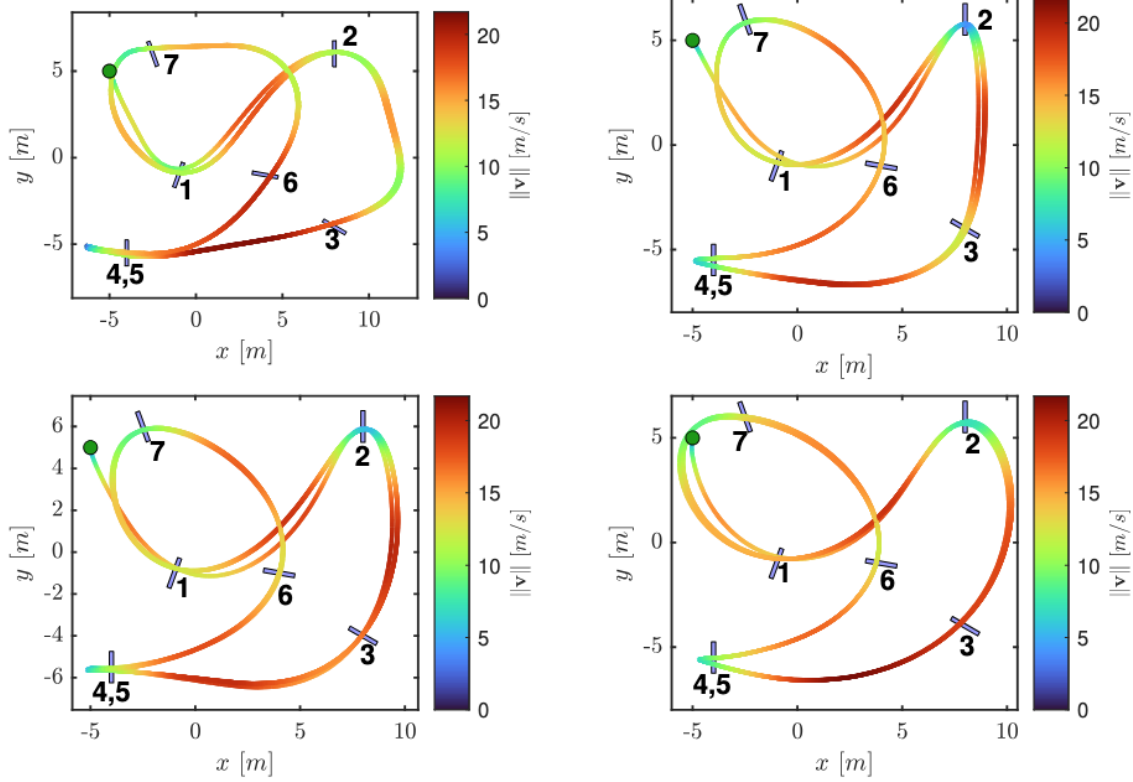


Fig. 4: Top view of simulated agile drone trajectories across distinct horizons and iLQR iterations: (a) $T = 2$, $K = 1$; (b) $T = 2$, $K = 5$; (c) $T = 2$, $K = 10$; (d) $T = 5$, $K = 5$. The green circle indicates the starting position. The colorbar indicates variations in the instantaneous velocity norm. Gates are numbered.

- a new model and nonlinear model predictive control,” *Mechanism and Machine Theory*, vol. 197, p. 105651, 2024.
- [2] A. Romero, S. Sun, P. Foehn, and D. Scaramuzza, “Model predictive contouring control for time-optimal quadrotor flight,” *IEEE Transactions on Robotics*, vol. 38, no. 6, pp. 3340–3356, 2022.
 - [3] Y. Song, A. Romero, M. Müller, V. Koltun, and D. Scaramuzza, “Reaching the limit in autonomous racing: Optimal control versus reinforcement learning,” *Science Robotics*, vol. 8, no. 82, p. eadg1462, 2023.
 - [4] E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza, “Champion-level drone racing using deep reinforcement learning,” *Nature*, vol. 620, no. 7976, pp. 982–987, 2023.
 - [5] A. Romero, Y. Song, and D. Scaramuzza, “Actor-critic model predictive control,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14777–14784, 2024.
 - [6] B. Amos and J. Z. Kolter, “OptNet: Differentiable optimization as a layer in neural networks,” in *International Conference on Machine Learning (ICML)*, pp. 136–145, 2017.
 - [7] B. Amos, I. D. J. Rodriguez, J. Sacks, B. Boots, and J. Z. Kolter, “Differentiable MPC for end-to-end planning and control,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.
 - [8] B. Amos, I. Jimenez, J. Sacks, B. Boots, and J. Z. Kolter, “mpc.pytorch: A fast and differentiable model predictive control solver for pytorch (v0.0.6).” <https://github.com/locuslab/mpc.pytorch/tree/v0.0.6>, 2024. GitHub repository, MIT License. Accessed: 2026-02-14.
 - [9] A. Romero, E. Aljalbout, Y. Song, and D. Scaramuzza, “Actor-critic model predictive control: Differentiable optimization meets reinforcement learning for agile flight,” *IEEE Transactions on Robotics*, vol. 42, p. 673–692, 2026.
 - [10] S. Levine, C. Finn, T. Darrell, and P. Abbeel, “End-to-end training of deep visuomotor policies,” *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.
 - [11] A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine, “Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.
 - [12] A. Agrawal, B. Amos, S. Barratt, S. Boyd, S. Diamond, and J. Z. Kolter, “Differentiable convex optimization layers,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.
 - [13] L. Amatucci, J. Sousa-Pinto, G. Turrissi, D. Orban, V. Barasuol, and C. Semini, “Primal-dual ilqr for gpu-accelerated learning and control in legged robots,” *IEEE Robotics and Automation Letters*, vol. 11, no. 1, pp. 1010–1017, 2026.
 - [14] E. Adabag, M. Greiff, J. Subosits, and T. Lew, “Differentiable model predictive control on the gpu,” *arXiv preprint arXiv:2510.06179*, 2025.
 - [15] F. Borrelli, A. Bemporad, and M. Morari, *Predictive Control for Linear and Hybrid Systems*. Cambridge University Press, 2017.
 - [16] L. Grne and J. Pannek, *Nonlinear model predictive control: theory and algorithms*. Springer Publishing Company, Incorporated, 2013.
 - [17] M. Ellis, J. Liu, and P. D. Christofides, *Economic Model Predictive Control: Theory, Formulations and Chemical Process Applications*. Advances in Industrial Control, Springer, 2017.
 - [18] R. Findeisen, L. Grüne, and M. A. Müller, eds., *Economic Model Predictive Control*. Cham: Springer, 2017.
 - [19] R. Sutton and A. Barto, “Reinforcement learning: An introduction,” *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 1054–1054, 1998.
 - [20] PyTorch Blog, “Accelerating pytorch with cuda graphs.” <https://pytorch.org/blog/accelerating-pytorch-with-cuda-graphs/>, 2021. Accessed: 2026-02-14.
 - [21] PyTorch Contributors, “Cuda semantics – pytorch documentation.” <https://docs.pytorch.org/docs/stable/notes/cuda.html>, 2026. Accessed: 2026-02-14.
 - [22] Y. Tassa, N. Mansard, and E. Todorov, “Control-limited differential dynamic programming,” in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1168–1175, 2014.
 - [23] W. Li, H. Luo, Z. Lin, C. Zhang, Z. Lu, and D. Ye, “A survey on transformers in reinforcement learning,” 2023.