

Data-driven control of hydraulic impact hammers under strict operational and control constraints

Francisco Leiva¹, Claudio Canales¹, Michelle Valenzuela¹ and Javier Ruiz-del-Solar¹

Abstract—This paper presents a data-driven methodology for the control of static hydraulic impact hammers, also known as rock breakers, which are commonly used in the mining industry. The task addressed in this work is that of controlling the rock-breaker so its end-effector reaches arbitrary target poses from any given initial configuration, which is required in normal operation to place the hammer on top of rocks that need to be fractured. The proposed approach considers several constraints, such as unobserved state variables due to limited sensing and the strict requirement of using a discrete control interface at the joint level. First, the proposed methodology addresses the problem of system identification in order to obtain an approximate dynamic model of the hydraulic arm. This is done via supervised learning, using only teleoperation data. The learned dynamic model is then exploited to obtain a controller capable of reaching target end-effector poses. For policy synthesis, both reinforcement learning (RL) and model predictive control (MPC) algorithms are utilized and contrasted. As a case study, we consider the automation of a Bobcat E10 mini-excavator arm with a hydraulic impact hammer attached as end-effector. Using this machine, both the system identification and policy synthesis stages are extensively studied in simulation and in the real world. The best RL-based policy consistently reaches target end-effector poses with position errors below 12 [cm] and pitch angle errors below 0.08 [rad] when deployed in the real world. Considering that the impact hammer has a 4 [cm] diameter chisel, this level of precision is sufficient for breaking rocks. Notably, this is accomplished by relying only on approximately 68 min of teleoperation data to train and 8 min to evaluate the dynamic model, and without performing any adjustments or fine-tuning for a successful policy Sim2Real transfer. A demonstration of policy execution in the real world can be found in <https://youtu.be/e-7tDhZ4ZgA>.

Keywords—Automation in mining, hydraulic impact hammers, data-driven control, discrete control, reinforcement learning.

I. INTRODUCTION

The increasing demand for automation comes with many challenges. This is particularly true when automating heavy-duty hydraulic machinery operating in unstructured environments. The automation of such systems is difficult due to their highly nonlinear dynamics, delayed responses, and the constraints associated with intervening manually operated machines that are already deployed in the field.

This work addresses the automation of hydraulic impact hammers (also known as rock-breakers) used in mining. These machines can be described as static hydraulic arms with an hydraulic impact hammer attached as an end-effector.

This work was supported by FONDECYT project 1251823, ANID-PIA project C1A250010, and ANID/Doctorado Nacional/2023-21232021.

¹Advanced Mining Technology Center (AMTC) and Department of Electrical Engineering, Universidad de Chile, Tupper 2007, Santiago, Chile.

francisco.leiva@ing.uchile.cl,
claudio.canales@amtc.uchile.cl,
michelle.valenzuela@amtc.uchile.cl,
jruizd@ing.uchile.cl

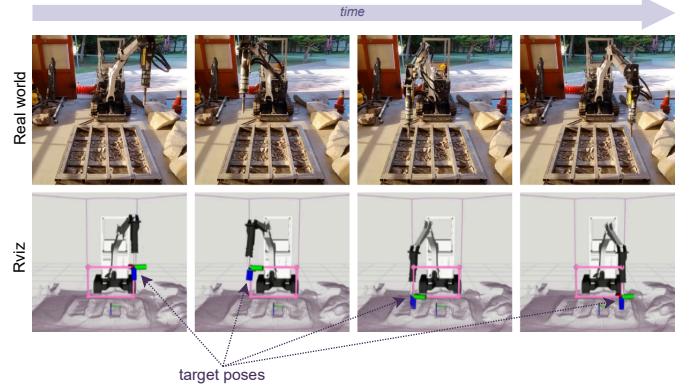


Fig. 1. Depiction of the reaching task performed by impact hammers in mining, using a Bobcat E10 mini-excavator. To perform this task, the hydraulic arm of the machine has to be controlled so its end-effector reaches target poses from arbitrary initial configurations, using minimal sensing and a discrete control interface at the joint level.

Operators remotely control these machines to break rocks that are too large to fall through steel grates installed on top of ore passes, allowing material to be transported to lower production levels (Correa et al., 2022). Particularly, we address the problem of synthesizing a controller capable of reaching end-effector target poses (for instance, rock breaking poses) for these machines. This task, commonly known as “reaching”, is constantly performed by human operators to place the impact hammer on top of rocks that need to be fractured, or to place it in safe areas whenever material is deposited on the steel grates by a load-haul-dump (LHD) machine. Fig. 1 illustrates the problem of reaching arbitrary target end-effector poses using a mini-excavator whose arm has the the same kinematic structure that full-sized rock breakers, and that has an hydraulic impact hammer attached as end-effector.

The main challenges associated with the automation of hydraulic impact hammers used in mining arise from the prohibition of any physical modification on the machines (which restricts sensing), the inability to accurately capture the system dynamics (as proprietary components often conceal their internal responses), and the requirement that control must be executed through discrete commands at the joint level (since this allows standardizing a common control interface for a fleet of impact hammers). These constraints originate from industrial requirements aiming to minimize intervention on machines that are already actively operating in production.

To achieve the automation goal given the above constraints, a two-stage data-driven methodology is adopted. In the first stage, the system’s dynamics is approximated by training a parameterized function using operational data, and then, in

the second stage, the obtained dynamic model is exploited to synthesize reaching controllers. The synthesis of policies is conducted by leveraging an efficient computational pipeline, in which the learned model is used to predict the real system's response to control commands. This allows generating experiences that are used to learn a controller using reinforcement learning (RL). In addition, a model predictive control (MPC) scheme is also implemented as a baseline. The obtained policies are then deployed and tested in the real world without performing any fine-tuning.

This methodology is implemented and validated to synthesize a reaching controller for a hydraulic mini-excavator. This mini-excavator is electro-hydraulically intervened to allow its automation, and in a manner that emulates the control challenges described for static rock-breakers used in mining. This results in unobserved state variables that makes classical dynamic modeling non-viable, and in a discrete control interface that makes accurately reaching target poses challenging.

With the above, the main contributions of this work are the following:

- A practical methodology for automating hydraulic arms with unobserved variables and a discrete control interface, covering the full pipeline from system identification to controller deployment. By relying on a minimal sensing setup, the approach can be transferred to other machines with minimal adaptation. The results demonstrate that effective hydraulic control can be achieved even in conditions of low observability and with limited ground truth data, thereby streamlining the deployment of industrial automation.
- A policy (obtained by exploiting a learned data-driven model) that perform the reaching task in a three-dimensional space. Unlike most existing controllers, the policy operates in a discrete action space, dictated by the hardware and operational requirements of the target application (secondary rock reduction in mining using hydraulic breakers). This design choice introduces additional complexity in both training and implementation. Despite these challenges, the controller achieves high accuracy, consistently reaching target poses with a position error below 12 [cm] and an absolute orientation error below 0.08 [rad]. Remarkably, the policy is transferred directly from simulation to the real machine without any parameter tuning.

II. RELATED WORK

The automation of industrial machinery has been extensively explored in recent years, driven by the demand for increased efficiency, safety, and productivity in complex work environments. In this domain, hydraulic machines are unique because of their power and versatility. These machines are widely used in construction, mining, forestry, and demolition, where they perform physically demanding tasks such as excavation, rock breaking, material handling, and infrastructure maintenance. Their broad range of applications has made them a focal point in the study of robotic automation. Consequently, hydraulic system automation has been an active area of research for several years, with numerous works that address the

unique challenges posed by their nonlinear dynamics, delayed responses, and the need for precise control under uncertain conditions (Mattila et al., 2017).

Traditional approaches to hydraulic machine automation are predominantly model-based. Early works such as Plummer and Vaughan (1990); Sohl and Bobrow (1997); Habibi and Richards (1991); Soroushpour and Salcudean (2001) aimed to achieve control by deriving dynamic models of the hydraulic system and applying different control techniques. For instance, in Sohl and Bobrow (1997), a non-linear tracking controller based on Lyapunov theory and backstepping is presented for a hydraulic servosystem, guaranteeing exponential stability for force tracking. However, the approach relies heavily on accurate parameter identification, such as valve gains and fluid bulk modulus, and requires access to signal derivatives, which are often noisy in practice. This class of requirements are difficult to fulfill in real-world industrial settings, often leading to complicated or limited implementations, compromising effectiveness, scalability, and robustness. This challenge is common across many classical model-based control strategies.

Even today, model-based approaches remain an active area of research in hydraulic control. Modern studies continue to explore analytical control strategies, often combining model-based formulations with advanced techniques such as adaptive control, sliding mode control (SMC) and model predictive control (MPC). These methods frequently incorporate online parameter estimation or disturbance observers to improve performance under uncertainty. For example, in Bender et al. (2017), an offset-free MPC controller is implemented in a hydraulic mini-excavator, allowing precise control of the arm movement while handling disturbances. However, these methods still inherently rely on the assumed model structure, which remains a critical weakness.

In parallel, model-free control strategies remain widely used in practice, particularly proportional-integral-derivative (PID) control (e.g., Egli et al. (2024)), due to their simplicity and ease of implementation. Nevertheless, PID controllers require extensive tuning for each machine, and their performance tends to degrade as the system moves away from the operating point around which the gains were calibrated.

Given the difficulty of deriving accurate analytical models for hydraulic machines, data-driven approaches have long been considered a promising alternative. In Song and Koivo (1995), it was proposed to learn the inverse dynamics of the plant (an hydraulic excavator) using operational data. More recently, this data-driven strategy has gained renewed interest, driven by advances in machine learning techniques. Works such as Park et al. (2017), Lee et al. (2022), Ma and Zhou (2024), Ma et al. (2024) and Greiser et al. (2024) use modern machine learning techniques to model complex system behaviors from sensor and control data. While these approaches offer clear advantages, such as adaptability, reduced modeling effort, and the ability to capture non-linearities, they are not without challenges. Issues such as data quality, generalization to unseen conditions, or limited interpretability remain active areas of concern.

An alternative strategy to control hydraulic machines and learn complex tasks is reinforcement learning (RL). One of the

main advantages of RL is its ability to learn control policies directly from interaction with the environment, without necessarily requiring an accurate model of the system dynamics. This is particularly appealing for hydraulic systems, which are notoriously difficult to model due to strong nonlinearities and complex internal dynamics, making classical control design a highly challenging task. Moreover, RL methods have shown promising results in simulation and even on real machines, successfully learning behaviors that are difficult to encode explicitly using conventional control techniques.

However, it is important to note that RL is not without drawbacks. The design of a reward function that effectively captures the objective of a task often requires extensive domain knowledge and entails a substantial trial-and-error process. Additionally, policies trained in simulation may fail to transfer to real-world machines due to discrepancies between simulated and real dynamics, a well known issue referred to as the sim-to-real gap. To mitigate this, it is crucial to ensure that the simulated dynamics closely approximate those of the real system. As seen in Egli and Hutter (2020), Egli and Hutter (2022), Spinelli et al. (2024) and Spinelli et al. (2025), a promising way to tackle this issue is the use of a data-driven approach to approximate the plant dynamics and then to rely on this learned model to synthesize an RL-based policy.

It is also worth mentioning that, even if many of these approaches offer promising results, they rely strongly on the observability of the system. For example, in works such as those of Jud et al. (2021), Egli and Hutter (2020), Egli and Hutter (2022), and Spinelli et al. (2024), precise measurements for the joint positions and velocities, often alongside pressures, torques, engine RPM measurements, among others, are utilized to characterize the instantaneous state of the machines being controlled.

Regarding the automation of hydraulic impact hammers (rock breakers) used in mining, prior work has reported the development of teleoperation, assisted teleoperation (Correa et al., 2022), controllers that solve a particular task of the full operating cycle of the rock breaker (e.g. rock breaking in Samtani et al. (2023)) and complete modular automation systems (Lampinen et al., 2021).

In Lampinen et al. (2021), for example, motion control is achieved by learning a mapping between input control signals and the resulting velocity for each valve-actuator pair, following the approach proposed by Nurmi and Mattila (2017). These learned models are then used to control the machine to track trajectories generated by a planner that takes into account the flow restrictions of the hammer's hydraulic unit, using the algorithm proposed by Lampinen et al. (2020). In Samtani et al. (2023) only the rock-breaking stage of the machine's operation cycle (which starts once the end-effector is positioned over a rock) is automated via model-free RL.

Given the state of the art, this work follows the data-driven approach adopted by many recent works (e.g., Egli and Hutter (2020, 2022)): learning a model of the system dynamics, and then obtaining a policy by exploiting the learned model. However, in contrast to the existing literature, this work addresses the challenge of learning a dynamic model by relying on a minimal set of measured variables, namely,

only joint positions, and performing the policy synthesis stage considering a discrete action space. Note that both restrictions arise from operational constraints on the target application and, while Lampinen et al. (2021) achieve good results regarding the motion control of a full-sized impact hammer, their method is not directly applicable in these settings, since their control law is designed for a continuous action space. Moreover, in contrast to works like those of Lampinen et al. (2020) and Egli and Hutter (2022), we address the problem of reaching an arbitrary target pose from a given initial configuration, without the need for a full plan or waypoints.

III. PROBLEM DESCRIPTION

A. Automation of impact hammers in underground mining

Impact hammers, also known as rock breakers, are hydraulic arms, typically with four degrees of freedom and a hydraulic impact hammer as end-effector. These machines are widely used in underground mining, where they perform the task of fracturing rocks that cannot pass through transfer steel grates installed on top of ore passes, which allows the material to be transported to a lower production level due to gravity (Correa et al., 2022).

Impact hammers used in these settings are installed adjacent to steel grates and are remotely controlled by expert human operators, who work in a safe control room, usually many kilometers away from the mine. The operators controlling the impact hammers do so by remotely actuating its electro-valves using a joystick, and by relying mainly on visual information of the environment, which is provided by CCTV cameras installed near the steel grates.

Since there are several ore passes in an underground mine, a human operator is responsible for the teleoperation of multiple impact hammers at once; however, a given operator can only control one machine at a time. The loaders that deposit material on the steel grates (e.g., load-haul-dump machines) cannot do so if the grates are completely obstructed due to large rocks, implying that the rock-breaking task is a potential bottleneck in the production chain.

The above poses the automation of impact hammers in underground mining as a relevant problem to solve due to the potential enhancements in terms of production that a fleet of efficient and autonomous impact hammers could provide.

1) *Challenges and constraints:* Automating a fleet of impact hammers comes with many challenges. Firstly, many different models of impact hammers may be present in a mine, which means that developing a solution that accounts for different dynamics and control interfaces is required. Secondly, most impact hammers in a mine do not come with proprioceptive sensors other than those required to provide diagnostics on its hydraulic unit (e.g., sensors to measure the temperature and pressure of the hydraulic fluid of its tank), which is insufficient to implement classical closed-loop control systems. Finally, these machines are subjected to deterioration due to the task they perform, making replacement of their end-effectors, actuators, and other mechanisms common during maintenance sessions.

The aforementioned challenges impose requirements on any control system that is to be deployed on a fleet of impact

hammers. The variability across impact hammers' models and the lack of sensor information on them motivate the selection of a minimal set of variables to characterize their state. This avoids relying on information that is not available or may not be obtained for all hammers in the fleet,¹ whilst standardizing and reducing implementation costs.

In addition, the variability across impact hammers' control interfaces has already resulted in the adoption of discrete control for their teleoperation in some underground mines. Although this design decision may prevent operators from having fine-grained control over the impact hammers, it comes with the benefit of only requiring characterizing the dead zones of the actuators of each impact hammer for its implementation at scale. This information is enough for the full standardization of the program that maps joystick commands to electro-valves setpoints for a whole fleet.²

B. The Bobcat E10 mini-excavator

As a testbed, we consider a Bobcat E10 mini-excavator equipped with an hydraulic impact hammer as end-effector. The hydraulic arm of this mini-excavator has the same kinematic chain structure as most impact hammers used in mining operations; in fact, this machine has previously been used to test an RL-based policy to control the impact hammer's boom and end-effector whenever the machine is in a suitable configuration to attempt breaking a rock (Samtani et al., 2023).

The Bobcat E10 used in this work has been modified by installing electrovalves to control its actuators. To control its hydraulic arm, four pairs of electrovalves send pilot signals to the original machine's valves, so as to extend or retract the hydraulic cylinders of the arm. Similarly, to control the impact hammer used as end-effector, an on-off electrovalve is used. All of these electrovalves are actuated using a programmable I/O module connected to a programmable logic controller (PLC), which, in turn, is connected to an on-board PC and to a set of rotary encoders, using a CAN bus interface. The above is illustrated in Figure 2.

An important feature of the Bobcat E10 mini-excavator is that its arm's valves are integrated into a single hydraulic valve block, which makes it difficult to get accurate pressure measurements to monitor the state of the valves. This imposes restrictions on the overall control design for the arm, as there is missing information to fully characterize its dynamics. A simplified hydraulic circuit diagram of the intervened mini-excavator is described in Appendix A.

To get the Bobcat E10's hydraulic arm configuration, four absolute rotary encoders measure the angular position of each joint. Custom mounts were fabricated to accommodate each encoder so that their shafts are aligned to their corresponding joint axis. Moreover, these encoders are connected to the same PLC used for actuation, using the CAN protocol.

Although it may seem unusual to use rotary encoders to get the angular positions of the arm's joints instead of

¹Examples on this matter include pressure measurements for a proper characterization of the hydraulic actuators' dynamics.

²Note that this is only true if the machines composing said fleet share the same kinematic chain structure, which is generally the case for impact hammers with four degrees of freedom.

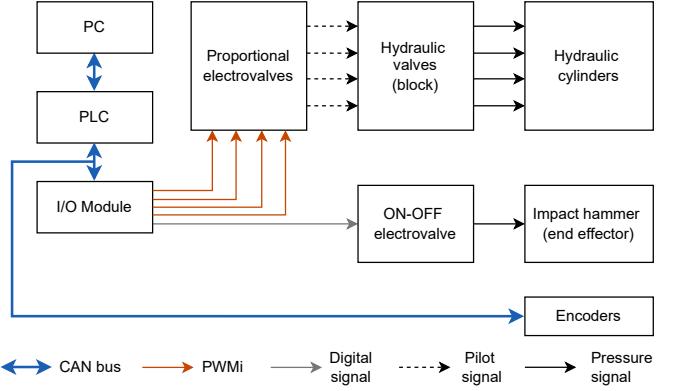


Fig. 2. Simplified electro-hydraulic diagram for the modified Bobcat E10's mini-excavator.

using draw-wire encoders to measure the displacement of each of its hydraulic cylinders, this design decision is due to two main factors related to impact hammers in underground mining. First, directly measuring joint positions using rotary encoders does not require a detailed modeling of the machine's kinematic chain (whereas draw-wire encoders generally do); moreover, for a given impact hammer in the fleet, the said kinematic chain may be unknown (for instance, because of structural modifications that occurred in maintenance sessions) and it will vary across different rock breaker models. Second, the use of properly protected draw-wire encoders, for instance those installed inside hydraulic cylinders, cannot be easily implemented in underground mines at scale, as it would require the disassembly of each intervened impact hammer and possibly the replacement of some actuators due to incompatibilities.

IV. PROPOSED APPROACH

To provide a solution to the problem of controlling a hydraulic impact hammer under limited observability and the requirement of using discrete control for its actuation, we follow an approach that can be described as a two-stage process. The first stage addresses the problem of modeling the impact hammers' dynamics (i.e., performing system identification), and the second stage addresses the problem of synthesizing a reaching controller for the machine, by leveraging the model obtained in the first stage.

For the first stage, the parameters θ of a function f_θ that approximate the impact hammer's dynamics are learned through supervised learning given teleoperation data. For a history of previous (discrete) actions (a_{t-k}, \dots, a_t) , and arm configurations (q_{t-k}, \dots, q_t) , the model f_θ has to predict the next joint configuration, q_{t+1} .

For the second stage, the model f_θ is used to generate experiences given some initial (approximate) state and a stream of actions, where the predicted configurations are fed back to the model itself to obtain rollouts over some temporal horizon. The generated data are used to obtain a reaching policy π_ϕ via RL or MPC. Note that this policy is conditioned on observations o_t constructed using only the rollout data obtained using f_θ to govern the underlying system dynamics. These observations

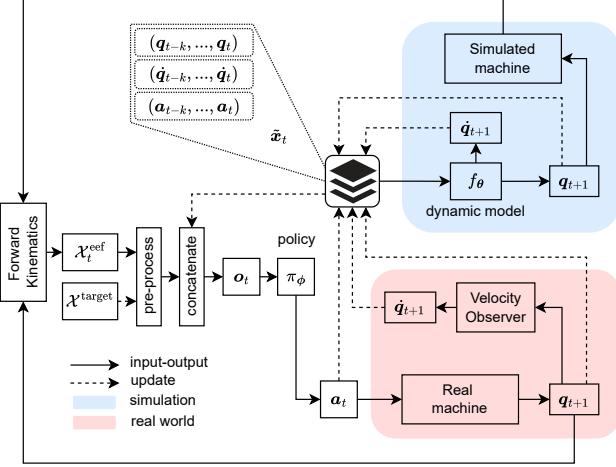


Fig. 3. Overview of the methodology used to obtain reaching controllers for hydraulic impact hammers, given a learned dynamic model f_θ . Note that estimating \dot{q}_t is only necessary when f_θ is trained to predict velocity residuals (see Section IV-A1).

contain both the current and target end-effector poses, $\mathcal{X}_t^{\text{eff}}$ and $\mathcal{X}_t^{\text{target}}$, respectively, where the current end-effector pose can be computed using forward kinematics. The synthesized policy can then be directly deployed in the real-world.

The whole described procedure is illustrated in Fig. 3, where optionally, besides configurations and actions, joint velocity estimates, \dot{q}_t , can be used to condition both the dynamic model f_θ and the policy π_ϕ . In what follows, each stage of the proposed approach is described in detail.

A. Data-driven system identification

We consider the problem of modeling the dynamics of a four-DoF hydraulic arm with an impact hammer attached as end-effector. To get a model of the machine kinematics, we must note that detailed 3D models of rock breakers, such as those used in mining, are rarely publicly available, and if they are, the real machines may not reflect the geometry documented by the manufacturer due to modifications performed during maintenance sessions. Therefore, the parameters of their kinematic chains, in general, must be obtained by taking measurements of the machines themselves (e.g., as in Lampinen et al. (2021)), and then using CAD software to obtain coarse models of their links. The outcome of this process can then be utilized to specify an URDF file that describes the impact hammer's kinematics in a format that is ROS compatible (Quigley et al., 2009). However, in this work the arm's hydraulic cylinders are omitted given the limitations of the standard URDF when attempting to represent kinematic loops, and also because we use rotational encoders to directly measure the joints' angular positions.

By applying the above on the Bobcat E10 mini-excavator, we get an URDF file that describes its kinematics. A rendering of the model obtained, alongside joint names, relevant frames, and a scaled steel grill similar to those used in real mining operations, is shown in Fig. 4.

To model the response of the impact hammer's arm actuators given control commands, we follow a data-driven

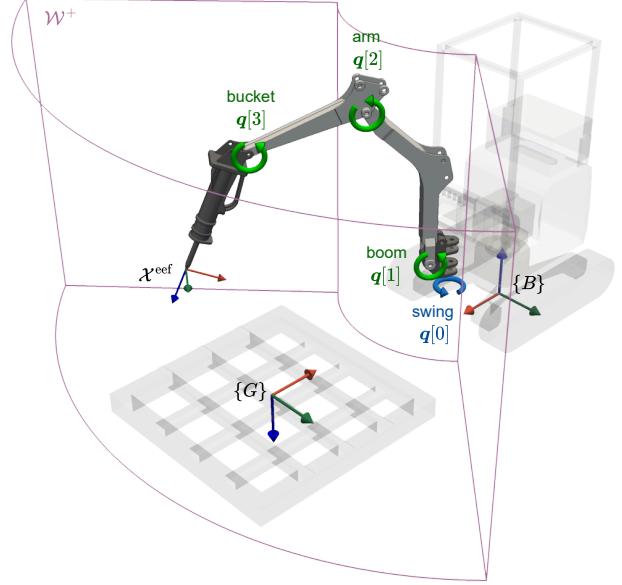


Fig. 4. Kinematic layout of the Bobcat E10 mini-excavator, alongside a scaled steel grill. The fixed frames $\{B\}$ and $\{G\}$ are attached to the mini-excavator base link and to the center of the grill, respectively. The cylindrical sector defines the boundaries for the position of end-effector poses that may be elements of the restricted workspace \mathcal{W}^+ .

approach. However, many variables that would be required for an accurate dynamic model are often unknown, and we abstain from relying on measurements that would not be available straightforwardly for a fleet of impact hammers used in an underground mine. For the Bobcat E10 mini-excavator, examples for variables in the first category include pressure differentials for the valves (see Section III-B), whilst measurements that, although available, are purposefully omitted, include the RPMs of its motor and the temperature of its hydraulic fluid.

Given the above, to learn a forward dynamics model for an impact hammer, and in particular, for the Bobcat E10's arm, we only consider the measurements provided by rotary encoders, and the control commands sent to electrovalves. Since not all rotary encoders provide speed estimations, we rely only on angular position measurements. These measurements are used to obtain the hydraulic arm configuration at time step t , $q_t \in \mathbb{R}^4$. The electro-valve setpoints, on the other hand, are fixed due to the control restrictions for impact hammer fleets discussed in Section III-A1. Thus, each hydraulic actuator contracts or expands given a fixed electrical signal, or may simply not act for a zero set point. Since these setpoints are fixed once tuned, a given (normalized) control command for time step t , a_t , can be defined as an element of the set $\{-1, 0, 1\}^4$.

1) Learning a model for the hydraulic arm actuators: We aim to learn a model such that, given available measurements to characterize the state of the impact hammer for a discrete time step t , predicts what its configuration will be for the next time step $t + 1$. We try to accomplish this by learning the parameters θ of a model f_θ that satisfies Eq. (1), where \tilde{x}_t is

a function of previous configurations and control commands.

$$\mathbf{q}_{t+1} = \mathbf{q}_t + f_{\theta}(\tilde{\mathbf{x}}_t) \quad (1)$$

To properly characterize the state of the arm and the delays of the hydraulic actuators, $\tilde{\mathbf{x}}_t$ is defined according to Eq. (2), where $\mathbf{q}_{t-k:t} = (\mathbf{q}_{t-k}, \mathbf{q}_{t-k+1}, \dots, \mathbf{q}_t) \in \mathbb{R}^{4 \times (k+1)}$, $\mathbf{a}_{t-k:t} = (\mathbf{a}_{t-k}, \mathbf{a}_{t-k+1}, \dots, \mathbf{a}_t) \in \mathbb{R}^{4 \times (k+1)}$, and g is the composition of functions that will be defined later.

$$\tilde{\mathbf{x}}_t = g(\mathbf{q}_{t-k:t}, \mathbf{a}_{t-k:t}) \quad (2)$$

We consider two different approaches for a practical instantiation of parameterized models that satisfy Eq. (1). For the first approach, the goal is to learn how to predict the residuals of the angular position ($\Delta \mathbf{q}_t$) to estimate \mathbf{q}_{t+1} . For the second approach, the goal is to learn how to predict angular velocity residuals ($\Delta \dot{\mathbf{q}}_t$) to obtain future velocity estimates and then compute \mathbf{q}_{t+1} by integrating the predicted velocities. In both cases, multilayer perceptrons (MLPs) or Kolmogorov-Arnold networks (KANs) (Liu et al., 2025) are utilized as function approximators. In what follows, each approach is described in further detail.

- Learning $\Delta \mathbf{q}_t$: This approach follows Eq. (1) by representing f_{θ} as an MLP or a KAN, and defining g as the composition of a min-max normalization for the arm configurations, a flattening operation over both configurations and actions, and the concatenation of the min-max normalized current configuration to the result. The min-max normalization for the angular positions of the arm is defined by Eq. (3), where the j -th components of \mathbf{q}_{\min} and \mathbf{q}_{\max} are set according to the joint limits of the arm, and $\mathbf{1}$ is a column vector in \mathbb{R}^4 with all its components equal to 1.

$$\text{minmax}(\mathbf{q}_t, \mathbf{q}_{\min}, \mathbf{q}_{\max}) = 2 \left(\frac{\mathbf{q}_t - \mathbf{q}_{\min}}{\mathbf{q}_{\max} - \mathbf{q}_{\min}} \right) - 1 = \bar{\mathbf{q}}_t \quad (3)$$

The above results in $\tilde{\mathbf{x}}_t$ being defined by Eq. (4)³, where $\tilde{\mathbf{q}}_{t-k:t}^T = (\bar{\mathbf{q}}_{t-k}^T, \bar{\mathbf{q}}_{t-k+1}^T, \dots, \bar{\mathbf{q}}_t^T) \in \mathbb{R}^{4(k+1)}$ and $\mathbf{a}_{t-k:t}^T = (\mathbf{a}_{t-k}^T, \mathbf{a}_{t-k+1}^T, \dots, \mathbf{a}_t^T) \in \mathbb{R}^{4(k+1)}$ are row vectors.

$$\tilde{\mathbf{x}}_t = (\tilde{\mathbf{q}}_{t-k:t}^T, \mathbf{a}_{t-k:t}^T, \bar{\mathbf{q}}_t^T) \in \mathbb{R}^{8(k+1)+4} \quad (4)$$

- Learning $\Delta \dot{\mathbf{q}}_t$: This approach follows Eq. (1) by defining $f_{\theta}(\tilde{\mathbf{x}}_t) := \Delta t \cdot (\dot{\mathbf{q}}_{t-1} + f'_{\theta}(\tilde{\mathbf{x}}_t))$, such that f'_{θ} (represented by an MLP or a KAN), predicts velocity residuals to get velocity estimations that are then integrated to get future angular positions, as in Eqs. (5) and (6).

$$\dot{\mathbf{q}}_t = \dot{\mathbf{q}}_{t-1} + f'_{\theta}(\tilde{\mathbf{x}}_t) \quad (5)$$

$$\mathbf{q}_{t+1} = \mathbf{q}_t + \Delta t \cdot \dot{\mathbf{q}}_t \quad (6)$$

Here, $f'_{\theta}(\tilde{\mathbf{x}}_t)$ is conditioned on velocities which are estimated given measured angular positions. To get these estimates, a loop tracker for the measured joint positions is implemented, such that a filtered angular velocity estimation can be obtained as an intermediate variable.

³Note that, although $\bar{\mathbf{q}}_t^T$ is already in $\tilde{\mathbf{q}}_{t-k:t}^T$, we include it as the last four components of $\tilde{\mathbf{x}}_t$ in Eq. (4) to get vectors of the same dimensions and overall structure than those defined by Eq. (7). In practice, this redundancy is introduced to simplify implementation.

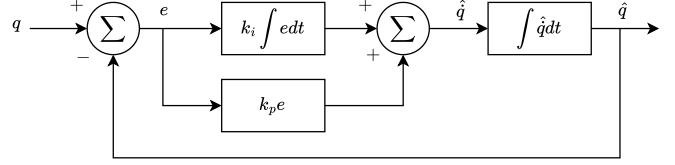


Fig. 5. Block diagram of the loop tracker implemented to get $\dot{\mathbf{q}}_t$ estimates.

For parameters k_i and k_p , the implemented loop tracker for a given joint is illustrated in Fig. 5. Note that to filter high frequencies, we do not use $\dot{\mathbf{q}}$ as the velocity estimate; instead, we use the integral term of the proportional-integral (PI) controller that tracks the joint position.

Using the loop tracker, we can get an estimate for $\dot{\mathbf{q}}_t$ given a history of joint positions up to time step t . Similarly to the case where we learn angular position residuals, here we define g as the composition of a velocity estimator (implemented as previously described), a min-max normalization of angular velocity estimations (as $\text{minmax}(\dot{\mathbf{q}}_t, \dot{\mathbf{q}}_{\min}, \dot{\mathbf{q}}_{\max})$, see. Eq. (3)), a flattening operation over velocity estimations and actions, and the concatenation of the min-max normalized current configuration. This results in $\tilde{\mathbf{x}}_t$ being defined by Eq. (7).

$$\tilde{\mathbf{x}}_t = (\bar{\mathbf{q}}_{t-k:t}^T, \mathbf{a}_{t-k:t}^T, \bar{\mathbf{q}}_t^T) \in \mathbb{R}^{8(k+1)+4} \quad (7)$$

The models that predict residuals are trained using back-propagation through time (Werbos, 1988). To do so, we rely on a dataset \mathcal{D} of temporally ordered tuples $(\mathbf{q}_t, \dot{\mathbf{q}}_t, \mathbf{a}_t)$ that allows the construction of short trajectories of the form $\tau_i = (\mathbf{q}_{t_i}, \dot{\mathbf{q}}_{t_i}, \mathbf{a}_{t_i}, \dots, \mathbf{q}_{t_i+H}, \dot{\mathbf{q}}_{t_i+H}, \mathbf{a}_{t_i+H})$ for a given initial time step t_i , along with a corresponding state approximation $\tilde{\mathbf{x}}_{t_i}$ associated with the first configuration action tuple of the trajectory. These trajectories are used to calculate a loss function that measures the approximation error of the models over a short time horizon H :

$$\mathcal{L}(\theta) = \frac{1}{H} \sum_{j=1}^H \|\hat{\mathbf{q}}_{t_i+j} - \mathbf{q}_{t_i+j}\|_2^2, \quad (8)$$

where the configurations $\hat{\mathbf{q}}_{t_i+j}$ are predicted in an “open-loop” manner by the model, that is, starting from $\tilde{\mathbf{x}}_{t_i}$:

$$\begin{aligned} \hat{\mathbf{q}}_{t_i+1} &= \mathbf{q}_{t_i} + f_{\theta}(\tilde{\mathbf{x}}_{t_i}), \\ \hat{\mathbf{q}}_{t_i+2} &= \hat{\mathbf{q}}_{t_i+1} + f_{\theta}(\hat{\mathbf{x}}_{t_i+1}), \\ &\vdots \\ \hat{\mathbf{q}}_{t_i+H} &= \hat{\mathbf{q}}_{t_i+H-1} + f_{\theta}(\hat{\mathbf{x}}_{t_i+H-1}). \end{aligned}$$

Note that $\tilde{\mathbf{x}}_{t_i+j}$, $j \in \{1, \dots, H-1\}$ is constructed by progressively updating the components of $\tilde{\mathbf{x}}_{t_i}$ with the ground-truth actions in τ_i and the predictions of the model (which may be angular positions or velocities, depending on the chosen model).

B. Controller synthesis for reaching

1) Problem formulation: In this work, we address the reaching task for hydraulic rock breakers, that is, we aim at synthesizing a controller capable of reaching end-effector target poses. In practice, performing this task would allow these machines to position their hydraulic hammer on top of rocks that need to be fractured.

The interaction between the agent and the environment is modeled as a Partially Observable Markov Decision Process (POMDP), which is defined by a set of states \mathcal{S} , a set of actions \mathcal{A} , a scalar reward function $\mathcal{R}(s, a)$, a stochastic transition function $T(s, a, s') = p(s'|s, a)$, an observation function $\mathcal{O}(s', a, o) = p(o|s', a)$, a set of observations Ω , and a discount factor $\gamma \in [0, 1]$. At each discrete time step t the agent observes o_t , executes an action a_t according to its policy $\pi(a_t | o_t)$, receives a scalar reward r_t , and transitions to a new state s_{t+1} .

We will assume that the rock breakers operate in a restricted region of their workspace and that this region is always free of obstacles (see Section IV-B3). This allows solving the reaching task without relying on exteroceptive information.

2) Modeling:

- **Dynamics:** The impact hammer dynamics is forward simulated using a parameterized function f_θ , which is trained as described in Section IV-A1. Depending on whether the learned model predicts the angular position residuals or the angular velocity residuals, this is accomplished by iteratively evaluating Eq. (1), or Eqs. (5) and (6).

- **Observations:** To provide the agent with enough information to make decisions (given the underlying learned dynamics), the policy is conditioned on \tilde{x}_t , on a representation of the current impact hammer’s end-effector pose, $\mathcal{X}_t^{\text{eef}} \in \text{SE}(3)$, and on a representation of the target pose, $\mathcal{X}^{\text{target}} \in \text{SE}(3)$. Note that \tilde{x}_t is given by Eq. (4) or by Eq. (7), depending on the chosen dynamics model.

We use two complementary representations for the end effector poses (instantaneous and target):

- As tuples (\mathbf{p}, \mathbf{r}) , where $\mathbf{p} \in \mathbb{R}^3$ denotes position and $\mathbf{r} \in \mathcal{S}^3$ denotes orientation.
- As configurations $\mathbf{q} \in \mathbb{R}^4$ that would be mapped to the poses via forward kinematics.

Thus, the observations o_t are given by the concatenation of \tilde{x}_t with representations for the instantaneous and target end-effector poses.

- **Actions:** Given the operational constraints described in Section III-A1, the control of the impact hammer is done using discrete control commands. Thus, using the same definition as in Section IV-A, a normalized action at time t , a_t , is an element of the set $\{-1, 0, 1\}^4$.

- **Reward function:** The reward function is designed to encourage the completion of the reaching task, while also regularizing the policy’s actions to prevent the exploitation of the underlying learned dynamics inaccuracies. We define a penalty function to measure the deviation of the current end-effector pose, $\mathcal{X}_t^{\text{eef}}$, from the target pose, $\mathcal{X}^{\text{target}}$, by weighting their positional and rotational differences. Representing the poses as tuples (\mathbf{r}, \mathbf{p}) , the

position error is quantified as the Euclidean distance between their positions:

$$d(\mathbf{p}^{\text{target}}, \mathbf{p}_t^{\text{eef}}) = \|\mathbf{p}^{\text{target}} - \mathbf{p}_t^{\text{eef}}\|_2, \quad (9)$$

whereas the rotational error is computed as the normalized geodesic distance between their unit quaternions:

$$d_g(\mathbf{r}^{\text{target}}, \mathbf{r}_t^{\text{eef}}) = \frac{1}{\pi} \cos^{-1} (2(\mathbf{r}^{\text{target}} \cdot \mathbf{r}_t^{\text{eef}})^2 - 1). \quad (10)$$

Thus, the described penalization function is given by:

$$r_t^{\mathcal{X}} = -\lambda_p d(\mathbf{p}^{\text{target}}, \mathbf{p}_t^{\text{eef}}) - \lambda_r d_g(\mathbf{r}^{\text{target}}, \mathbf{r}_t^{\text{eef}}), \quad (11)$$

where $\lambda_p, \lambda_r > 0$ are fixed scalars.

In addition, we include a penalty for deviations in joint-space from the desired target configuration, $\mathbf{q}^{\text{target}}$ (which can be mapped to $\mathcal{X}^{\text{target}}$ via forward kinematics):

$$r_t^{\mathbf{q}} = -\lambda_q \|\mathbf{q}_t - \mathbf{q}^{\text{target}}\|_2. \quad (12)$$

To encourage fine-grained control, we consider binary rewards that the agent receives only if it manages to meet certain accuracy criteria. These binary rewards are computed in terms of positional and rotational errors (Eqs. (13)⁴ and (14)⁵), and deviations in joint space (Eq. (15)), where $\epsilon_p, \epsilon_r, \epsilon_\alpha, \epsilon_q > 0$ are fixed scalar thresholds.

$$b_{\mathcal{X}} = [\![d(\mathbf{p}^{\text{target}}, \mathbf{p}_t^{\text{eef}}) < \epsilon_p]\!] \cdot [\!(d_g(\mathbf{r}^{\text{target}}, \mathbf{r}_t^{\text{eef}}) < \epsilon_r)\!] \quad (13)$$

$$b_\alpha = [\!(\alpha_{\text{yaw}} - \alpha_{\text{yaw}}^{\text{target}})^2 < \epsilon_\alpha]\!] \cdot [\!(\alpha_{\text{pitch}} - \alpha_{\text{pitch}}^{\text{target}})^2 < \epsilon_\alpha]\!] \quad (14)$$

$$b_{\mathbf{q}} = \prod_{j=0}^3 [\!(\mathbf{q}_t[j] - \mathbf{q}^{\text{target}}[j])^2 < \epsilon_q]\!] \quad (15)$$

The reward the agent receives depending on the fulfillment of the above conditions is given by Eq. (16), where $w_{\mathcal{X}}, w_\alpha, w_{\mathbf{q}} > 0$ are fixed scalars.

$$r_t^\epsilon = w_{\mathcal{X}} \left(b_{\mathcal{X}}|_{(\epsilon_p, \epsilon_r)} + b_{\mathcal{X}}|_{(\epsilon'_p, \epsilon'_r)} \right) + w_\alpha b_\alpha + w_{\mathbf{q}} b_{\mathbf{q}} \quad (16)$$

To encourage temporal consistency in the selection of actions, we penalize changes in action components between successive time steps:

$$r_t^{\mathbf{a}} = -\lambda_a \cdot \frac{1}{4} \sum_{i=0}^3 |\mathbf{a}_t[i] - \mathbf{a}_{t-1}[i]|, \quad (17)$$

where $\lambda_a > 0$ is the action regularization weight.

Finally, we also penalize the agent if the impact hammer’s end effector goes outside a restricted region of its workspace, $\mathcal{W}^+ \subset \mathcal{W} \subset \text{SE}(3)$ (see Fig. 4), where \mathcal{W} denotes the full reachable workspace, and $\lambda_{\mathcal{W}^+} > 0$ is a fixed scalar:

$$r_t^{\mathcal{W}^+} = -\lambda_{\mathcal{W}^+} \neg([\mathcal{X}_t^{\text{eef}} \in \mathcal{W}^+]). \quad (18)$$

Thus, the full reward is a weighted combination of all of the above terms:

$$r_t = r_t^{\mathcal{X}} + r_t^{\mathbf{q}} + r_t^\epsilon + r_t^{\mathbf{a}} + r_t^{\mathcal{W}}. \quad (19)$$

⁴Here $[\cdot]$ is the Iverson bracket, meaning $[\![P]\!] = 1$ if P is True, and $[\![P]\!] = 0$ otherwise.

⁵We only consider the yaw and pitch Euler angles, since given the impact hammer kinematics, the roll-angle is fixed. Although a similar argument could be used to also omit the yaw-angle, we use it because its value is fully determined by the “swing” joint, i.e., by $q_t[0]$ (see Fig. 4).

3) *Episodic conditions*: The reaching task, as formulated in this work, is episodic. In each episode, a random initial configuration for the impact hammer is selected. This initial configuration is such that the position of the end-effector of the hammer remains inside a restricted region of its workspace, which is denoted by \mathcal{W}^+ (see Fig. 4), and its orientation has a bounded pitch angle, $|\alpha_{\text{pitch}}^{\text{target}}| \leq \alpha_{\text{pitch}}^{\max}$. Afterwards, a target end effector pose, $\mathcal{X}_{\text{target}} \in \text{SE}(3)$, also within the restricted workspace and satisfying the pitch angle condition, is randomly selected.

For a given episode, if the end effector of the impact hammer reaches $\mathcal{X}_{\text{target}}$, then the episode is deemed successful. On the contrary, if the hammer’s end effector does not reach $\mathcal{X}_{\text{target}}$ after $T_{\text{reset}} = 500$ discrete time steps, then the episode is considered unsuccessful, and episodic conditions are reset. Note that these conditions are checked only to measure the performance of the reaching controller but do not imply getting to a terminal state.

Fig. 6a shows sampled end-effector poses that fulfill the position and pitch conditions described above for the Bobcat E10 mini-excavator, by setting $\alpha_{\text{pitch}}^{\max} = 1.05$ [rad]. Fig. 6b, on the other hand, shows the sampled end effector poses that, besides the above conditions, fulfill $|\alpha_{\text{pitch}}^{\text{target}}| \leq \alpha_{\text{pitch}}^{\text{ts}}$, with $\alpha_{\text{pitch}}^{\max} > \alpha_{\text{pitch}}^{\text{ts}} = 0.69$ [rad], and the z-axis component of their position being in $[z_{\min}^{\text{ts}}, z_{\max}^{\text{ts}}]$, with $z_{\min}^{\text{ts}} = 0.155$ [m] (which matches the grill height) and $z_{\max}^{\text{ts}} = 2.355$ [m], respectively.

During policy learning, initial and target poses (and their respective configurations) are sampled considering the conditions applied to generate Fig. 6a. For evaluation purposes the same applies for initial poses, however, the target poses are obtained by enforcing the conditions used to generate Fig. 6b. Note that the latter poses, besides being a subset of those used during learning, fulfill restrictions that are *ad-hoc* to the task space associated to rock-breaking, in which it is expected for the end-effector to reach poses that are quasi-orthogonal to the surfaces of the rocks that need to be fractured.

4) *RL-based controller*: Given the formulation described in Section IV-B1, the goal is to learn a policy that maximizes the expected discounted return that the agent receives, i.e., to maximize $J_{\text{RL}}(\pi) = \mathbb{E}_{\mathbf{a}_t \sim \pi(\mathbf{a}_t | \mathbf{o}_t)} \left[\sum_{t=1}^T \gamma^{t-1} r_t \right]$.

To accomplish the above, we train a policy using the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017). To adapt PPO to discrete actions, we follow an approach close to that proposed by Dulac-Arnold et al. (2015) for their “Wolpertinger architecture”: Given a continuous proto-action $\mathbf{a}_t^p \in [-1, 1]^4$ (output by the PPO policy), we construct a valid action \mathbf{a}_t by discretizing each component of the proto-action, $\mathbf{a}_t^p[j], j \in \{0, 1, 2, 3\}$, according to Eq. (20).

$$\mathbf{a}_t[j] = \begin{cases} -1 & \text{if } \mathbf{a}_t^p[j] < -0.5, \\ 0 & \text{if } |\mathbf{a}_t^p[j]| \leq 0.5, \\ 1 & \text{if } \mathbf{a}_t^p[j] > 0.5 \end{cases} \quad (20)$$

While the discretized action \mathbf{a}_t is used for control, the proto-action is used to update the policy given the learning objectives of PPO.

5) *MPC-based controller*: As an alternative to RL, we also synthesize a controller following an MPC approach. MPC

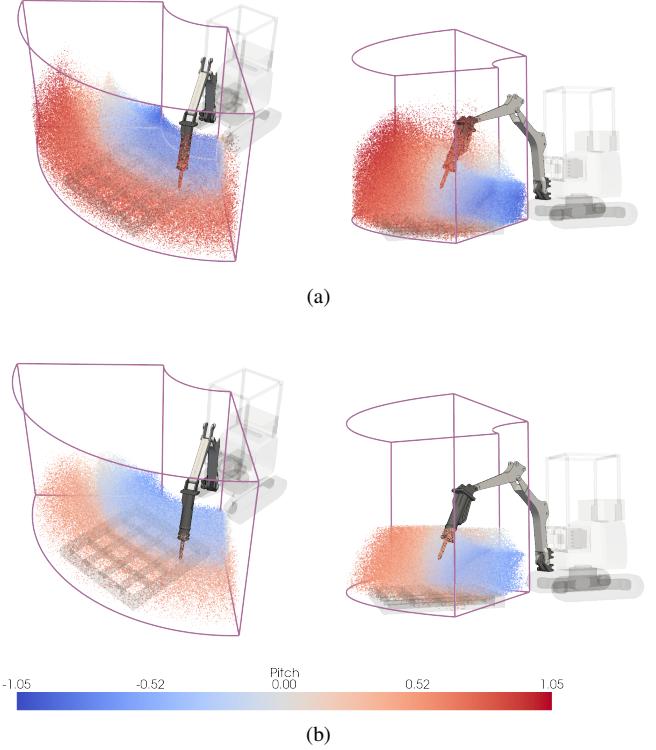


Fig. 6. Sampled end-effector poses used to set episodic conditions for the reaching task during (a) policy learning, and (b) policy evaluation.

formulates control as a receding horizon optimization problem that leverages the learned dynamic model to plan action sequences over a finite time horizon. At each time step t , the MPC controller solves the following optimization problem:

$$\mathbf{a}_{t:t+H_{\text{MPC}}-1}^* = \arg \max_{\mathbf{a}_{t:t+H_{\text{MPC}}-1}} \sum_{k=0}^{H_{\text{MPC}}-1} \gamma^k r_{t+k} \quad (21)$$

subject to:

$$\mathbf{q}_{t+k+1} = \mathbf{q}_{t+k} + f_{\theta}(\tilde{\mathbf{x}}_{t+k}), \quad k \in \{0, \dots, H_{\text{MPC}} - 1\} \quad (22)$$

$$\mathbf{a}_{t+k} \in \{-1, 0, 1\}^4, \quad k \in \{0, \dots, H_{\text{MPC}} - 1\} \quad (23)$$

where $\mathbf{a}_{t:t+H_{\text{MPC}}-1}^*$ denotes the optimal action sequence over the planning horizon H_{MPC} , γ is the discount factor, and r_{t+k} are the rewards computed using the function defined in Section IV-B2. Constraint (22) enforces that the predicted configurations evolve according to the learned dynamic model f_{θ} , and constraint (23) restricts the actions to be discrete. Note that only the first action \mathbf{a}_t^* of the optimal sequence is executed, and the optimization is repeated at the next time step with updated state information.

To solve the problem defined by Eqs. (21)–(23), we employ a variant of the Cross-Entropy Method (CEM) (De Boer et al., 2005), a sampling-based optimization algorithm that iteratively refines a distribution over candidate action sequences. At each iteration, CEM samples a population of N_{pop} continuous proto-action sequences $\{\mathbf{a}_{t:t+H_{\text{MPC}}-1}^{p,(i)}\}_{i=1}^{N_{\text{pop}}}$ from a parameterized Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with diagonal covariance matrix over the continuous action space $[-1, 1]^{4 \times H_{\text{MPC}}}$, discretizes

each proto-action to obtain valid discrete action sequences $\{\mathbf{a}_{t:t+H_{\text{MPC}}-1}^{(i)}\}_{i=1}^{N_{\text{pop}}}$ according to Eq. (20), evaluates their cumulative rewards using the learned model f_θ via forward simulation, selects the top N_{elite} sequences with highest returns, and updates the distribution parameters (μ, Σ) to concentrate probability mass around these elite samples. This process repeats for a fixed number of iterations or until convergence, yielding an approximately optimal action sequence.

The variant we use is referred to as iCEM (improved CEM), which incorporates the enhancements proposed by Pinneri et al. (2021) into the standard CEM formulation. These improvements include colored action noise to enforce temporal smoothness in the planned trajectories, importance mixing to leverage information from previous optimization rounds, and momentum-based updates to stabilize the distribution refinement process. These modifications have been shown to improve sample efficiency and solution quality, particularly in contact-rich manipulation tasks with complex dynamics (Sancaktar et al., 2022; Li et al., 2024; Jiang et al., 2024).

V. EXPERIMENTAL RESULTS

A. Computational pipelines

The models used for the data-driven identification of the system (see Section IV-A) are implemented and trained using Flax (Heek et al., 2024). The simulation environment is entirely built on top of JAX (Bradbury et al., 2018), taking advantage of XLA (accelerated linear algebra), just-in-time (JIT) compilation, and parallelization. Moreover, the simulation pipeline integrates Brax (Freeman et al., 2021) to compute forward kinematics. For visualization purposes, we leverage the data structures provided by Brax to render the simulation using the MuJoCo (Todorov et al., 2012) visualizer. For RL-based policy learning, we employ the PPO implementation provided by Brax, whereas for MPC-based control, iCEM is implemented in JAX.

Implementing simulation, modeling, and learning in JAX (and JAX-based libraries) allows for a highly efficient computational workflow. As a result, for instance, we can train RL-based policies using PPO for hundreds of millions of steps in just a few minutes on mid-range consumer desktops (e.g., equipped with an Intel i7-12700 CPU, an NVIDIA RTX 4060 GPU, and 32 GB of RAM).

To deploy the controllers on the real machine, a ROS-based pipeline (Quigley et al., 2009) was developed. Moreover, a pre-deployment stage using the Gazebo simulator (Koenig and Howard, 2004) was incorporated to allow testing in a safe environment, thereby minimizing the risk of accidents that could harm either staff or the machine. The implementation details for the ROS-based and ROS-Gazebo computational pipelines are provided in the Appendix B.

B. Data collection for system identification

We collect teleoperation data to construct a dataset \mathcal{D} of temporally ordered tuples $(\mathbf{q}_t, \dot{\mathbf{q}}_t, \mathbf{a}_t)$. This dataset is then used to learn the parameters of the dynamic model f_θ through the minimization of the loss function in Eq. (8) (see Sec. IV-A1 for a detailed explanation).

The teleoperation of the Bobcat E10 mini-excavator is performed by sending discrete commands $\mathbf{a}_t \in \{-1, 0, 1\}^4$ to its actuators. These control commands are constructed by discretizing the analog signals from the left and right sticks of a wireless Xbox 360 controller. The joystick interpreter is implemented in ROS, and the control commands are sent via CAN bus to the PLC installed in the mini-excavator. The PLC program performs a direct mapping of the discrete control commands to current set points for each electro-valve, so each stick axis of the Xbox 360 controller (four in total) can actuate one of the four hydraulic cylinders of the mini-excavator arm. It is important to note that this direct joint-level teleoperation resembles that of impact hammers in underground mining (see Section III-A1).

The instantaneous angular positions of the hydraulic arm joints, \mathbf{q}_t , are obtained directly through rotational encoder measurements (as explained in Section III-B), and their angular velocity, $\dot{\mathbf{q}}_t$, is estimated using the observer described in Section IV-A1.

Using the ROS-based teleoperation module, the mini-excavator arm is controlled at 20 Hz, and the tuples $(\mathbf{q}_t, \dot{\mathbf{q}}_t, \mathbf{a}_t)$ are stored at the same rate. To construct a dataset with representative samples, the mini-excavator arm is operated with the objective of covering the subset of its configuration space that would result in its end-effector position to lie within the restricted workspace, \mathcal{W}^+ , described in Section IV-B3. With this aim, while controlling the machine, the human operator can see a 3D representation of the restricted workspace limits, and has real-time visual feedback of the history of visited end-effector positions, thus, is instructed to “paint” the interior of the restricted workspace, with the end effector position acting as a 3D pencil.

Fig. 7 shows the visual feedback described, constructed from data acquired by teleoperating the mini-excavator arm for approximately 21, 23 and 32 minutes, respectively.

This data, which amounts to a total of approx. 76 min of teleoperation, is used to learn a model for the hydraulic arm’s actuators response, as will be discussed in the following section.

C. Training and evaluation of models for the arm dynamics

The teleoperation dataset \mathcal{D} contains temporally ordered tuples $(\mathbf{q}_t, \dot{\mathbf{q}}_t, \mathbf{a}_t)$. In principle, this dataset can be constituted by the aggregation of several trajectories (with a length of at least $H + 1$ elements each, plus the data to construct $\tilde{\mathbf{x}}_t$), as long as these are managed separately when sampling short sequences $\tau_i = (\mathbf{q}_{t_i}, \dot{\mathbf{q}}_{t_i}, \mathbf{a}_{t_i}, \dots, \mathbf{q}_{t_i+H}, \dot{\mathbf{q}}_{t_i+H}, \mathbf{a}_{t_i+H})$, for a given initial time step t_i . In practice, the dataset we use for training is composed by the three trajectories depicted in Fig. 7, and is divided into training and validation subsets. The training subset is constructed by using 90% of each trajectory, starting from the first sample, whereas the evaluation subset is constructed using the remaining portion of the trajectories.

Following the method described in Section IV-A1, to learn the parameters θ of a given dynamic model f_θ , we minimize the loss function defined in Eq. (8), which is computed by sampling the trajectories τ_i from \mathcal{D} . Note that evaluating

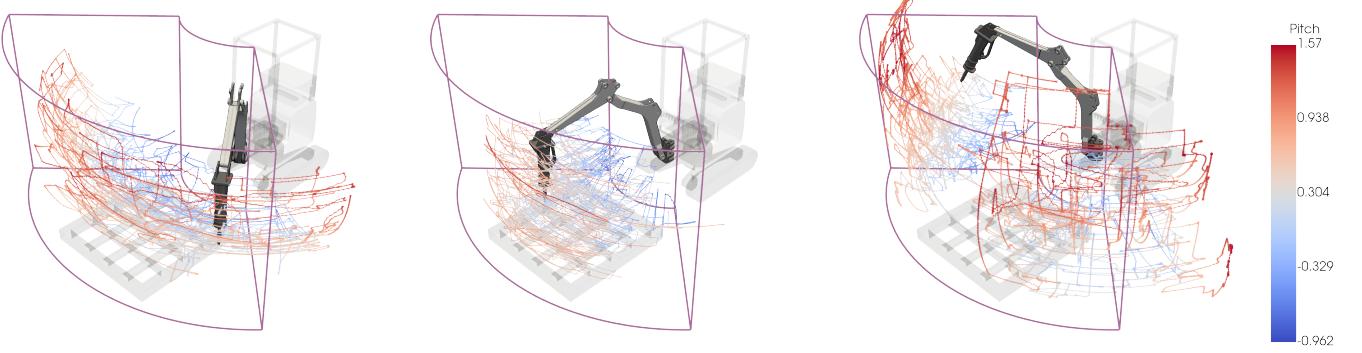


Fig. 7. End effector trajectories resulting from the teleoperation of the Bobcat E10 mini-excavator, for three data collection sessions (left, center, and right). The trajectories' color represent the instantaneous end effector pitch angle.

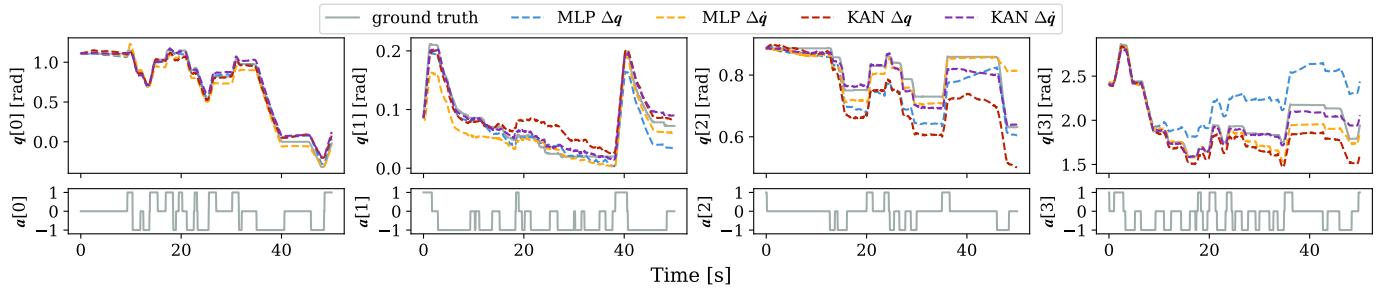


Fig. 8. Open-loop predictions performed by the best learned dynamic models over an evaluation subset. Given an initial ground-truth construction of \tilde{x}_t and a sequence of 10^3 ground-truth actions captured at 20 Hz, the dashed lines show the predicted angular positions for the Bobcat E10 arm's joints. Note that the terms \tilde{x}_{t+j} , $j \in \{1, \dots, H-1\}$, $H = 10^3$, are sequentially constructed by updating the components of \tilde{x}_t with ground-truth actions and model predictions.

$\mathcal{L}(\theta)$ requires setting a temporal horizon H , the number of temporal lags on angular positions/velocities and actions used to construct \tilde{x}_t , which is given by k , and an architecture for an MLP or KAN to represent f_θ . Moreover, as discussed in Section IV-A1, we can choose to predict angular position residuals, Δq , or angular velocity residuals, $\Delta \dot{q}$.

To manage the aforementioned possibilities, for each prediction type and for each type of neural network (MLP/KAN), we search for suitable hyperparameters through the optimization of the metric $\mathcal{L}(\theta)|_{H=80}^{\text{eval}}$, which has the same definition as the loss function used for training (see Eq. (8)), but is evaluated using 100 short trajectories sampled from the evaluation subset over a fixed time horizon, $H = 80$. This process is done using the Optuna library (Akiba et al., 2019). More details on this optimization process are documented in Appendix C.

The above produces a ranking of models with respect to the evaluation metric $\mathcal{L}(\theta)|_{H=80}^{\text{eval}}$. Since the top ranked models only differ by a slight margin with respect to $\mathcal{L}(\theta)|_{H=80}^{\text{eval}}$, we filter those with higher lags k and prefer smaller models due to their faster inference time (which is specially relevant for the policy obtained using iCEM). Under these criteria, the selected models are characterized in Table I.

Since these learned models are meant to serve as a surrogate for the real Bobcat E10 arm dynamics, their performance is further assessed in an open-loop prediction setting. Fig. 8 shows a ground-truth trajectory of configurations and normalized discrete commands sampled from the evaluation subset, alongside the configurations predicted by the learned

TABLE I
SELECTED PARAMETERIZATIONS AND TRAINING HYPERPARAMETERS FOR THE DYNAMIC MODELS.

Pred.	Parameterization	k	H	lr	n_B	$\mathcal{L}(\theta) _{H=80}^{\text{eval}}$
Δq	MLP(128, 128)	18	8	10^{-3}	512	$5.586 \cdot 10^{-3}$
	KAN(32, 16, 32)	18	8	10^{-5}	1024	$5.563 \cdot 10^{-3}$
$\Delta \dot{q}$	MLP(512, 128)	18	16	10^{-3}	256	$5.675 \cdot 10^{-3}$
	KAN(32, 32, 16)	18	16	10^{-4}	512	$5.555 \cdot 10^{-3}$

k : temporal lags; H : training loss horizon; lr : learning rate; n_B : batch size.

dynamic models. The ground-truth trajectory consists of 10^3 configuration-action pairs, which (as described in Section V-B) are captured at 20 Hz, so the open-loop predictions result in a rollout of 50 seconds. Note that the models only have access to a first ground-truth construction of \tilde{x}_t and to the sequence of ground truth actions, therefore, prediction errors compound.

The results obtained show that all the dynamic models can approximately capture the response of the actuators with low error (for short time horizons) in the evaluation subset. The dynamic models not being perfect is an expected result, given that their inputs lack relevant variables that influence the machine's responses, such as differential hydraulic pressures, the hydraulic fluid's temperature, and the motor's RPM. Moreover, the models are trained under a low data regime that does not contain samples from excitation signals purposefully designed for system identification, but only from the arm teleoperation.

Note that since the purpose of these models is to enable the synthesis of reaching controllers via RL or MPC, we want them not only to achieve low error in the teleoperation evaluation subset, but also to behave properly outside the training distribution. The above is crucial because we want the reaching controllers to perform appropriately for arbitrary initial and target poses sampled from the restricted workspace. Moreover, out-of-distribution (OOD) action sequences will likely be fed to the dynamic models during the RL training and the MPC optimization, so inconsistent dynamic model transitions for OOD inputs will result in policies with a high reality gap, which may render them useless.

Given the above, it is not clear whether there is an overall best dynamic model across those selected through hyper-parameter optimization. Therefore, using each of them to synthesize policies seems a necessary experimental step, as only through the comparison of their respective policies it would be possible to know which dynamic model serves its purpose the best. This is done in the next section.

D. Synthesis and evaluation of controllers for reaching

Using the best learned dynamic models, we can synthesize policies for reaching. To do so, the learned dynamic models f_θ are used as surrogates of the real dynamics of the mini-excavator arm. Given an action a_t and an initial history of angular positions/velocities and previous actions to construct \tilde{x}_t , we can get a next configuration for the hydraulic arm. Given a new control signal a_{t+1} and the predicted configuration to construct \tilde{x}_{t+1} , and repeating the above procedure, we can obtain a sequence $(\tilde{x}_t, a_t, \dots, \tilde{x}_T, a_T)$ for a given time horizon T . If we refer to Section IV-B2, we can see that this sequence can then be used to compute a sequence of observations, actions, and rewards, $(o_t, a_t, r_t, \dots o_T, a_T, r_T)$, which we can leverage to obtain policies via RL or MPC (see Sections IV-B4 and IV-B5).

It is important to note that the selection of a given dynamic model to train an RL-based policy conditions the policy's observations, because observations are constructed using (\tilde{x}_t, a_t) tuples, and the definition of \tilde{x}_t varies depending on the dynamic model's temporal lags and whether it predicts angular position or angular velocity residuals. However, since all selected dynamic models share the same number of temporal lags k (see Table I), the main difference between the observations used as input for the RL-based policies is simply whether they are constructed using normalized angular positions or angular velocity estimations (see Eqs. (4) and (7)).

To implement the above, we use the JAX-based computational pipeline described in Section V-A. Specifically, we implement a Brax-compliant environment that specifies the observations, actions, reward function, and episodic settings of the problem modeling, as described in Section IV-B2, and use the Bobcat E10 robot description illustrated in Fig. 4 to manage aspects such as the forward kinematics computation. This allows using the PPO implementation provided by Brax, and to easily parallelize the execution of rollouts based on the learned dynamic model predictions, which is crucial to achieve real-time control using MPC. The resulting environment can be rendered using the MuJoCo visualizer, as shown in Fig. 9.

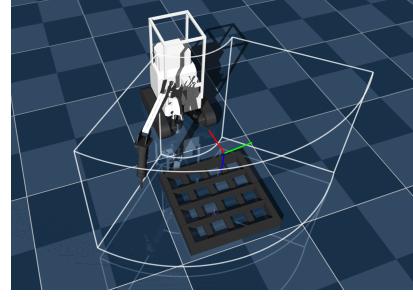


Fig. 9. Training environment renderization in the MuJoCo visualizer, with markers to represent the restricted workspace \mathcal{W}^+ and a randomly chosen target end-effector pose.

Using this environment, we can train RL-based policies using PPO for 400M steps, which only takes about 6 min. using a mid-range computer. Similarly, we can instantiate iCEM-based controllers that will provide action sequences that aim at maximizing cumulative rewards over short horizons (said rewards being obtained using the same reward function used for RL).

To evaluate the performance of the policies, we consider the following metrics:

- Avg. return: Undiscounted sum of instantaneous rewards obtained up until $T_{\max} = 500$ steps pass, averaged across evaluation episodes.
- Success rate (conditioned on precision requirements): Given a threshold for the Euclidean distance between the current and target end-effector poses, ε_p , and a threshold for the absolute difference between their pitch angles, ε_α , we consider the reaching task as successful if b_{SR} , defined by Eq. (24), equals one at least once during a given episode.

$$b_{\text{SR}}^t = \llbracket d(\mathbf{p}^{\text{target}}, \mathbf{p}_t^{\text{eef}}) < \varepsilon_p \rrbracket \cdot \llbracket |\alpha_{\text{pitch}} - \alpha_{\text{pitch}}^{\text{target}}| < \varepsilon_\alpha \rrbracket \quad (24)$$

Thus, the success rate (SR) of a reaching policy is computed as $\text{SR}(\varepsilon_p, \varepsilon_\alpha) := \frac{1}{N} \sum_{i=1}^N \llbracket \sum_{t=1}^{T_{\max}} b_{\text{SR}}^{t,(i)} \geq 1 \rrbracket$, where the dependence on threshold selection is highlighted, and the metric is computed for N episodes, each of them with a time horizon T_{\max} .

- Rate at which the end-effector leaves the restricted workspace \mathcal{W}^+ (OOW): Consider the variable $b_{\mathcal{W}}$ defined in Eq. (25), which equals one if the end effector is within the restricted workspace.

$$b_{\mathcal{W}}^t = \llbracket (\mathbf{x}_t^{\text{eef}} \in \mathcal{W}^+) \rrbracket \quad (25)$$

Similarly to the success rate, the out-of-workspace rate is defined as $\text{OOW} := \frac{1}{N} \sum_{i=1}^N \llbracket \sum_{t=1}^{T_{\max}} (1 - b_{\mathcal{W}}^{t,(i)}) \geq 1 \rrbracket$ for N evaluation episodes, where each of them has a time horizon T_{\max} .

- Rate at which the end-effector goes below the lower z -axis boundary of \mathcal{W}^+ (OOZ): Defined analogous to OOW. Note that $\text{OOZ} \leq \text{OOW}$.
- Minimum z -axis component of the end-effector position across all the evaluated trajectories: This metrics measures the worst possible end-effector breach through the lower z -axis plane that defines the restricted workspace.

TABLE II

EVALUATION RESULTS OBTAINED IN THE BRAX-BASED TRAINING ENVIRONMENT. THE PERFORMANCE METRICS ARE COMPUTED ACROSS 500 INDEPENDENT REACHING TRIALS (EPISODES). THE REPORTED AVERAGE AND STANDARD DEVIATION PER METRIC IS COMPUTED ACROSS FIVE INDEPENDENTLY SYNTHESIZED CONTROLLER PER POLICY TYPE.

Algo.	Dyn. model	Avg. return	SR(0.02, 0.02)	SR(0.04, 0.04)	OOW	OOZ	Min. z [m]
PPO	MLP Δq	644.45 \pm 10.31	0.97 \pm 0.03	1.00 \pm 0.00	0.13 \pm 0.01	0.08 \pm 0.01	-0.22 \pm 0.28
	MLP $\Delta \dot{q}$	683.36 \pm 16.94	0.99 \pm 0.01	1.00 \pm 0.00	0.15 \pm 0.01	0.11 \pm 0.01	-0.10 \pm 0.08
	KAN Δq	458.85 \pm 17.13	0.91 \pm 0.02	0.98 \pm 0.01	0.15 \pm 0.02	0.10 \pm 0.02	-1.17 \pm 0.04
	KAN $\Delta \dot{q}^*$	-220.78 \pm 604.06	0.26 \pm 0.21	0.42 \pm 0.35	0.36 \pm 0.25	0.24 \pm 0.25	-0.71 \pm 0.34
iCEM		15.96 \pm 5.95	0.20 \pm 0.01	0.28 \pm 0.01	0.03 \pm 0.01	0.02 \pm 0.01	0.14 \pm 0.01
iCEM [†]	MLP $\Delta \dot{q}$	-348.59 \pm 1.10	0.66 \pm 0.01	0.78 \pm 0.00	0.02 \pm 0.00	0.01 \pm 0.00	0.13 \pm 0.02
iCEM [‡]		-148.48 \pm 0.31	0.81 \pm 0.01	0.91 \pm 0.01	0.01 \pm 0.00	0.00 \pm 0.00	0.14 \pm 0.01

*Learning rate set to $0.8 \cdot 10^{-4}$ and evaluated at 800M training steps. [†] r_t^ϵ set to zero. [‡] r_t^ϵ and r_t^χ set to zero.

With the above, we train RL-based policies for each best learned dynamic model (which are characterized in Table I). We consider five independent training trials per policy and evaluate them over a set of $N = 500$ episodes, each of them with fixed but different episodic conditions. For MPC-based policies, we do the same, but only using the MLP $\Delta \dot{q}$ model, since it allows synthesizing the best RL-based controllers, as it will be explained later. Moreover, for all the policies we set a control frequency of 20 Hz (matching the time discretization of the dynamic models) and 4 action repeats. In addition, all the RL-based policies are evaluated after 270M training steps, except for the PPO KAN $\Delta \dot{q}$ whose learning rate and training steps were further tuned due to worse performance being obtained when training under the same conditions than the rest of the controllers. The results obtained are presented in Table II and all the training details and hyperparameters used to obtain policies are documented in Appendix D.

From training RL-based policies using PPO, we verified that using different dynamic models has an impact on the final controller's performance. The first indicator of this difference lies in the cumulative reward obtained at the end of the training process, where the policies trained using the best KAN $\Delta \dot{q}$ model achieve far lower avg. returns than the rest, which motivates filtering them from further analysis. We can also note a difference in average return (albeit less pronounced) between the policies trained using the KAN Δq model and those trained using MLPs as surrogate dynamic models, the latter achieving better overall performances. It is also observed that the breaches through the lower plane of the restricted workspace are much less pronounced for the PPO-MLP $\Delta \dot{q}$ model, as the “min. z ” metric is lower for this controller when compared to the rest of the RL-based policies.

From training MPC-based policies, we note that optimizing slight variations of the reward function (by setting $r_t^\epsilon = 0$ and $r_t^\epsilon = r_t^\chi = 0$) improves performance. This result can be explained considering that r_t^ϵ is a positive scalar only given to the agent when certain conditions are fulfilled, which makes the reward function highly non linear and may make its optimization over short horizons more susceptible to convergence to local optima. Moreover, also setting $r_t^\chi = 0$ further improves results for iCEM, which can be explained considering the reward ablation documented in Appendix D for an RL based policy (using the same dynamic model used

by iCEM), which suggests that r_t^ϵ and r_t^χ may be conflicting terms when optimized simultaneously. When comparing the overall performance of controllers based on RL and MPC, we can see that although RL policies in general achieve a higher success rate, iCEM controllers are better at avoiding breaching the restricted workspace boundaries. This result may be due to iCEM discarding the action sequences that would result in such breaches during optimization, and PPO policies, on the other hand, affording the cost associated with leaving \mathcal{W}^+ whenever a higher return (in the long run) can be achieved.

The results motivate studying the behavior of the best policies obtained in the real world. To do so, in the next section we conduct experiments to measure the performance of all the RL-based policies (except the PPO-KAN $\Delta \dot{q}$ policies) and the best iCEM controller when undergoing a Sim2Sim and a Sim2Real transfer.

E. Sim2Sim and Sim2Real transfer of the learned controllers

We deploy the synthesized policies in the real-world by leveraging a ROS-based computational pipeline. Nevertheless, a pre-deployment stage using the Gazebo simulator was utilized to test the different system components prior to using them to control the real mini-excavator. In both the pre-deployment (Sim2Sim) and real-world deployment (Sim2Real) stages, the reaching controllers are encapsulated in action servers that, once queried with a target pose, will execute actions (according to an RL or MPC-based policy) until successfully reaching the target or preemptively stopping when an undesired behavior is detected. We must note, however, that to isolate the performance of the policies, all safety measures are disabled during the Sim2Sim and Sim2Real transfer experiments. The computational pipeline that allows implementing the pre-deployment and real-world deployment stages is described in Appendix B. In what follows, the performed experiments and their results are documented.

1) *Sim2Sim transfer:* A first pre-deployment experiment is conducted using Gazebo with the reaching controllers encapsulated in ROS-based action servers. When evaluating a given policy, the learned dynamic model used to synthesize it is encapsulated in a module that allows bypassing the Gazebo simulator physics. This is done by using the inferences of these dynamic models, conditioned on previous joint positions/velocities and actions, to set the hydraulic arm

TABLE III
SIM2SIM TRANSFER EVALUATION RESULTS OBTAINED ACROSS 500 INDEPENDENT REACHING TRIALS.

	Algo.	Dyn. model	SR ^a	SR ^b	OOW	OOZ	Min. z [m]
Brax	PPO	MLP Δq	0.99	1.00	0.124	0.068	-0.066
		MLP $\Delta \dot{q}$	0.98	1.00	0.130	0.092	-0.103
		KAN Δq	0.92	1.00	0.136	0.088	-1.184
	iCEM [‡]	MLP $\Delta \dot{q}$	0.81	0.98	0.012	0.004	0.124
Gazebo	PPO	MLP Δq	0.02	0.93	0.124	0.066	-0.004
		MLP $\Delta \dot{q}$	0.11	1.00	0.118	0.068	0.000
		KAN Δq	0.12	0.90	0.142	0.086	-0.574
	iCEM [‡]	MLP $\Delta \dot{q}$	0.74	0.98	0.122	0.068	0.107

^aSR(0.02, 0.02). ^bSR(0.12, 0.08). [‡] r_t^e and r_t^x set to zero.

TABLE IV
SIM2REAL TRANSFER EVALUATION RESULTS OBTAINED ACROSS 100 SUCCESSIVE REACHING TRIALS.

	Algo.	Dyn. model	SR ^a	SR ^b	OOW	OOZ	Min. z [m]
Gazebo	PPO	MLP Δq	0.04	0.94	0.040	0.010	0.091
		MLP $\Delta \dot{q}$	0.08	1.00	0.090	0.020	0.005
		KAN Δq	0.13	0.87	0.060	0.030	0.000
	iCEM [‡]	MLP $\Delta \dot{q}$	0.77	0.98	0.070	0.020	0.151
R. World	PPO	MLP Δq	0.01	0.43	0.010	0.010	0.154
		MLP $\Delta \dot{q}$	0.02	0.99	0.070	0.040	0.050
		KAN Δq	0.00	0.69	0.020	0.020	0.099
	iCEM [‡]	MLP $\Delta \dot{q}$	0.08	0.71	0.200	0.050	0.120

^aSR(0.02, 0.02). ^bSR(0.12, 0.08). [‡] r_t^e and r_t^x set to zero.

configuration directly, similar to what is done in the Brax-based environment (see Appendix B).

The Sim2Sim transfer experiments use the exact same initial and target end-effector poses set during the experiments conducted in the Brax-based environment, for the same 500 distinct episodes (see Section V-D). Moreover, as in the Brax-based environment evaluation, the execution of the action servers for a given target pose is only terminated by timeout (after 500 time steps). In this case, however, only a single policy is evaluated (which corresponds to the first of the five synthesized controllers for each policy type). Under these conditions, this Sim2Sim transfer study allows measuring the effects of time delays due to the communication of messages to construct observations and set control commands in an isolated manner (as the learned dynamic models dictate the response of the impact hammer to actions). It is worth mentioning that such delays are not present in the Brax-based training environment, but exist during real world deployment.

To get a fine-grained performance measure of the controllers when evaluated under this new setting, we construct success rate matrices by computing the SR metric with thresholds $(\epsilon_p, \epsilon_\alpha) \in \{0.02, 0.04, \dots, 0.34\} \times \{0.02, 0.04, \dots, 0.16\}$, which can be visualized in Appendix E. Table III shows the performance metrics obtained, highlighting the success rates obtained at $(\epsilon_p, \epsilon_\alpha) = (0.02, 0.02)$, and $(\epsilon_p, \epsilon_\alpha) = (0.12, 0.08)$.

The results show a noticeable performance drop for the success rates at high precision requirements for the RL-based policies, while this is not the case for the iCEM controller. We attribute this to two complementary causes: the differences in the reward function optimized by RL-based policies and the iCEM controller, and the asynchronous nature of the Gazebo-ROS computational pipeline, which introduces delays that result in perturbations on the environment state transitions. Firstly, PPO policies are trained using the full definition of r_t (see Eq. (19)), while iCEM policies consider a slight variation of this function, where the terms r_t^x and r_t^e are set to zero. One can think of r_t^e as a bonus term encouraging the end-effector to stabilize on the target pose, since it is only non-zero if precision requirements over the end-effector pose and the arm configuration are fulfilled (see Eqs. (13)–(15)). This makes oscillations around the target pose less pronounced once it has been reached at a certain level of precision, however, when deploying the controllers in the Gazebo simulator, the

dynamic models, the reaching controller, and the simulator itself, all run asynchronously, and that induces time delays that are not present when training and evaluating the policies in Brax. These perturbations harm the performance of RL-based policies at low success rate thresholds. In contrast, the iCEM controller tends to oscillate around the target (since it does not optimize for r_t^e), which makes it more likely to eventually approach the target pose (during the whole execution of a given episode), regardless of the aforementioned perturbations. We must note, however, that iCEM is also affected by the asynchronous nature of Gazebo, as the OOW and OOZ metrics are higher in this environment compared to Brax, however, the Min. z coordinate of its end-effector during all the experiments is still higher than those of RL-based policies. Despite these observations, when deployed in Gazebo, one can see that the PPO MLP $\Delta \dot{q}$ controller nearly matches the SR of the iCEM controller at thresholds $(\epsilon_p, \epsilon_\alpha) = (0.06, 0.06)$, and surpasses it at $(\epsilon_p, \epsilon_\alpha) = (0.08, 0.04)$ and higher (see Fig. 14 in Appendix E). Moreover, it matches or surpasses the precision levels of the rest of RL-based policies.

2) *Sim2Real transfer*: The policies are deployed in the real Bobcat E10 mini-excavator using the same ROS-based action servers used for the Sim2Sim transfer experiments. Again, the execution of the action servers is only stopped by a timeout after 500 time steps pass. Since in this case it is not possible to easily set random initial configurations for each reaching episode, given an initial configuration for the hydraulic arm, the evaluation experiments consist of attempting to reach 100 poses, which are sequentially fed to the action servers. Given that episodes only finish by time-out, this translates in the target end effector pose changing every 500 time-steps, making the initial condition for episodes (except for the first one) dependent on the target pose of the previous episode and the performance of the policy itself. Moreover, the 100 target end-effector poses coincide with the first 100 targets that are used for the evaluation of policies in simulation, both for the Brax-based environment and when using the Gazebo-ROS pipeline.

The execution of these experiments for policies trained using PPO with dynamic models MLP $\Delta \dot{q}$, MLP Δq and KAN Δq , and for the iCEM controller alongside the MLP $\Delta \dot{q}$ model results in the success rate matrices shown in Fig. 10 and in the performance metrics summarized in Table IV. Moreover, these experiments translate into approximately 50 minutes of

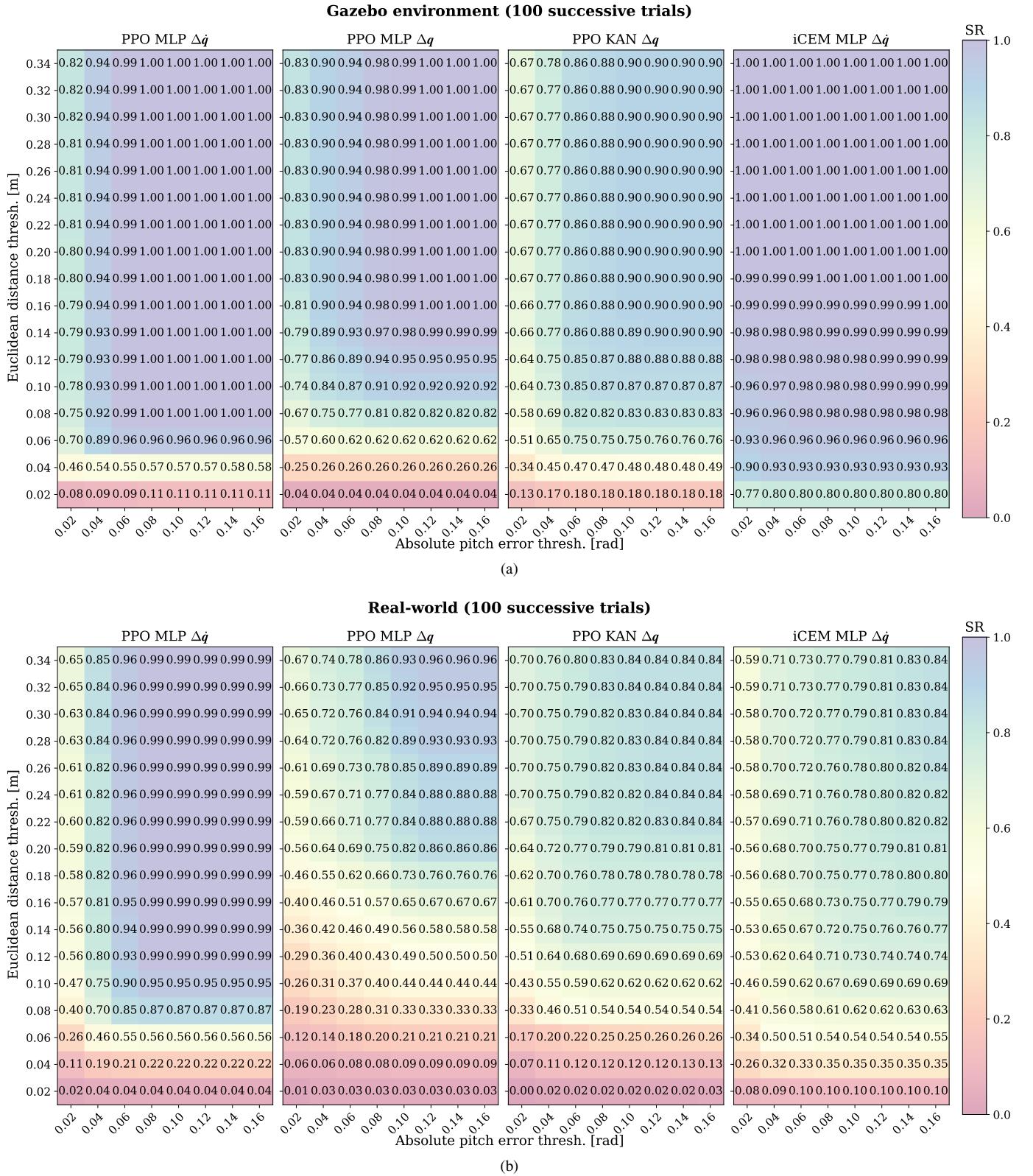


Fig. 10. Success rate matrices obtained when evaluating the reaching policies in (a) the Gazebo simulator and (b) the real-world, in both cases, by encapsulating the policies in action servers.

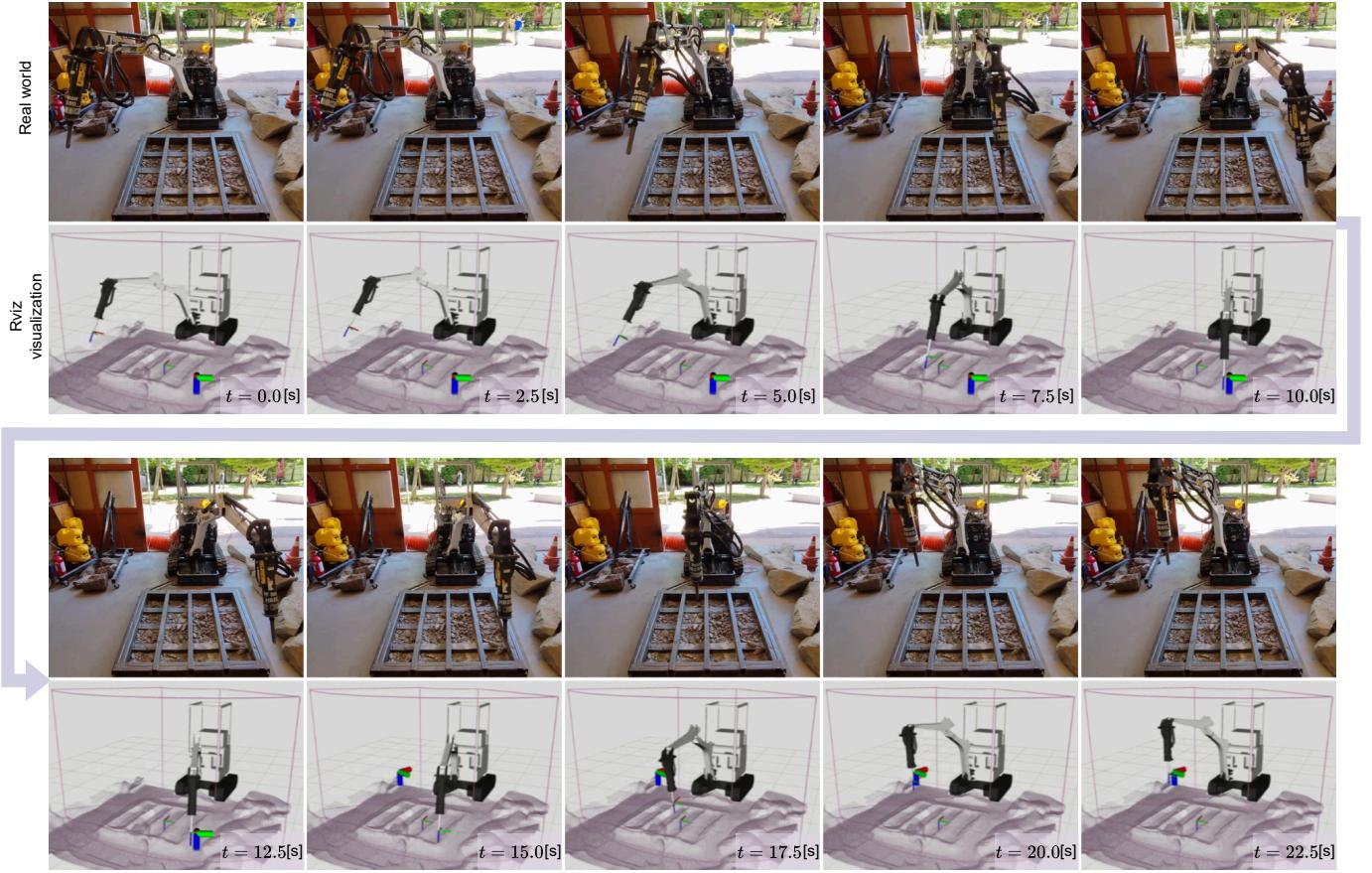


Fig. 11. Real-world PPO-MLP $\Delta\dot{q}$ policy execution for two sequentially queried target poses.

autonomous operation per evaluated controller, whose footage, for a qualitative evaluation of performance, can be seen in the following link: <https://youtu.be/lshZwPQMIFg>.

The results obtained show that RL-based controllers display a detriment on their success rate at low Euclidean distances and pitch angle error thresholds when comparing their performance in the real world and in the Gazebo-ROS pipeline. The reaching controller that is the least affected by the Sim2Real transfer and also achieves better overall performance remains the PPO-MLP $\Delta\dot{q}$ policy, achieving a near perfect success rate for the sequential queried target poses 100 at thresholds $(\epsilon_p, \epsilon_\alpha) = (0.12, 0.08)$. Moreover, contrary to what was observed in the Sim2Sim experiments for the iCEM controller, in the real world it drastically dropped its success rate for all thresholds while also achieving higher OOW and OOZ values.

We note that a common failure mode observed for the PPO KAN Δq and the iCEM controller during real world deployment was that sometimes they were unable to generate commands that resulted in an actual displacement of the hydraulic arm actuators. This happened whenever the mini-excavator's arm reached certain configurations and specific target poses were being queried. In such situations, the policies would often output commands set to zero for all joints or would rapidly alternate the values of certain action components, which, given the hydraulic coupling and high actuation delays of the arm's hydraulic cylinders, resulted in no movement, regardless of the enforced action regularization

via the reward function definition, and the action repeats used to synthesize and deploy the controllers. This behavior may be explained due to the learned dynamic models being imperfect and producing misleading transitions in certain regions of the state space, which translates into inadequate control strategies for certain situations and, therefore, poor performance when facing those situations in the real world.

The fact that the PPO MLP $\Delta\dot{q}$ policy achieves good performance and does not display the failure modes that iCEM shows (although both depend on the same dynamic model), may be due to the PPO MLP $\Delta\dot{q}$ policy, since parameterized as an MLP itself, has the capacity to “average” its behavior across poorly represented regions of the state space; moreover, given that its observations are mostly constructed using velocity estimations, it is likely that this policy puts less importance on the instantaneous configuration of the arm, which may help generalization.

Finally, in addition to the Sim2Real transfer experiment documented above, we further tested the PPO MLP $\Delta\dot{q}$ policy, setting fixed euclidean distance and absolute pitch error thresholds to $\epsilon_p = 0.12$ [m] and $\epsilon_\alpha = 0.08$ [rad] to decide whether a target pose was reached, so as to assess its performance in a setting more close to the practical application of the controller. The result of these experiments is qualitatively presented in Fig. 11 for two sequentially queried target poses, and the corresponding video with further trials can be found in <https://youtu.be/e-7tDhZ4ZgA>.

VI. CONCLUSION

In this work, we presented a practical, data-driven methodology to automate hydraulic impact hammers such as those used for secondary reduction in mining. This challenge is addressed considering operational constraints that may be present in real mining operations, such as unobserved state variables, and the requirement of relying on a discrete control interface. In particular, a methodology for learning dynamic models using teleoperation data and then exploiting said models to obtain policies for the task of reaching target end-effector poses is presented and validated. Said validation is performed both in simulation and the real world, using a Bobcat E10 mini-excavator with an impact hammer attached as end-effector.

The results obtained are promising given that the best obtained reaching policy is able to consistently reach target poses with relatively high precision in the real world, regardless of the mini-excavator arm being inherently challenging to control due to partially observed states, actuation delays, and hydraulic coupling, and using discrete commands at the joint level to perform the reaching task.

However, the developed reaching controller has some drawbacks. One of its limitations is that it does not fully comply with the restrictions over the end-effector position given by the restricted workspace, and enforced by the reward term r_t^W . This limitation could be alleviated by activating the safety measures included in the action server and using the control supervisor module included in the deployment computational pipeline (see Appendix B), since if the action server is stopped due to the end-effector leaving the workspace or due to the detection of oscillatory behaviors, a feedback signal reporting the unsuccessful completion of the reaching task could be sent to an hypothetical higher level system (e.g., a state machine) that in turn could trigger a recovery behavior.

However, the most notable limitation of the reaching policies is that they are not informed by exteroceptive information, so it would not be able for them to avoid collisions if the constrained workspace in which the impact hammer operates ceases to be obstacle free. This would be troublesome for the deployment of these controller in a real mining operation, since frontal loaders and LHDs depositing material above ore passes' grills should be regarded as potential obstacles, regardless of additional safety measures. Moreover, the grill itself and the rocks remaining above it should also be regarded as obstacles, because grill geometries vary across ore passes and moving from one target pose to another may induce an unintended collision with a rock. Therefore, developing a collision-aware reaching controller is left for future work.

Finally, the performance gap between iCEM and PPO-based policies in certain regions of the state space motivates investigating whether constraining the action search space to regions closer to the training distribution could improve MPC real-world performance.

APPENDIX A

SIMPLIFIED BOBCAT E10 HYDRAULIC CIRCUIT

Fig. 12 shows a simplified hydraulic circuit for the Bobcat E10 mini-excavator. From right to left, this diagram can be read as follows:

- The supply line provided by the pumps goes through a pressure control sub-circuit that generates a pilot signal, and also directly feeds the (simplified) hydraulic block displayed in the diagram.
- The pilot signal feeds eight proportional electro-valves that are controlled via discrete current signals sent to them using an I/O module (see Fig. 2). The electro-valves' output goes to the valves within the hydraulic block shown in the diagram.
- The valves within the hydraulic block control the four actuators of the mini-excavator arm, that is, the hydraulic cylinders associated to the “boom”, “swing”, “arm” and “bucket” joints (see Fig. 4).
- The hydraulic block also feeds the supply line to an on-off electro-valve that actuates the impact hammer end-effector. Note that this on-off valve is also controlled by using the I/O module mentioned previously.

Furthermore, this diagram shows some aspects of the Bobcat E10 mini-excavator that make controlling its arm challenging. Note, for instance, that controlling the arm's hydraulic cylinders depends on a pilot stage and a main hydraulic stage, which contributes to actuation delays. Also note that the valves within the hydraulic block share supply, which contributes to coupling across actuators.

APPENDIX B ROS-BASED COMPUTATIONAL PIPELINES

To deploy reaching controllers in the real-world, a ROS-based pipeline was developed. This computational pipeline, along with its components and their interactions, is illustrated in Fig. 13. To minimize discrepancies between components, the pipeline only bifurcates from the drivers (which are distinct for the simulation in Gazebo and the real world) until the end, resulting in two distinct execution branches.

For the pre-deployment stage, the learned dynamic models described in Section V-C were used to replicate the behavior of the real machine in the Gazebo simulator. To do so, these models are instantiated in the `bobcat_e10_sim_driver` component. This module allows bypassing Gazebo's built-in physics, enforcing the dynamics learned from the data obtained by operating the real machine, allowing the replication of hard-to-model effects, such as command-response delays.

The reaching controller is encapsulated within the `autonomous_server` module, which allows selecting among MPC and RL-based policies and provides them with the necessary data to populate the buffers they use to construct \hat{x}_t . This data is obtained from the `JointStates` messages, which specify the joint positions and velocities of the impact hammer. Through this server, the user can specify a target end-effector pose. The same server generates the corresponding commands, sends them to the subsequent components, maintains a fixed control frequency, and monitors terminal conditions to determine when to stop the hammer's operation, for instance, upon reaching the goal within a predefined tolerance, leaving the workspace, or exhibiting oscillatory behaviors. The two latter safety measures, however, are fully turned off during the experiments reported in this work, so as to measure the controllers' performance in an isolated manner.

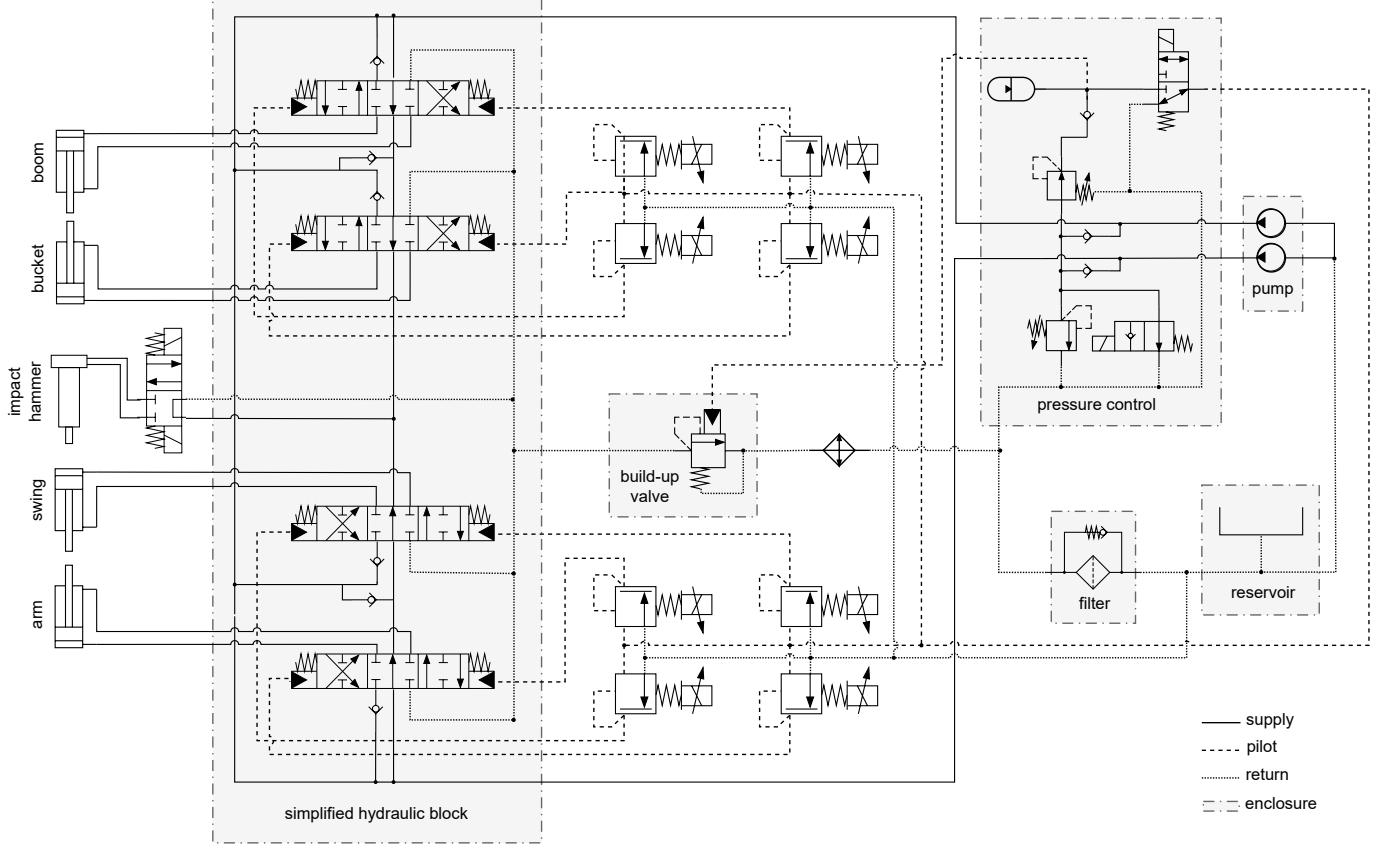


Fig. 12. Simplified hydraulic circuit of the intervened Bobcat E10 mini-excavator (based on the original Bobcat E10 documentation and the modifications implemented on the machine).

At this stage, it can be observed that the commands generated by both the `teleoperation_server` and the `autonomous_server` follow the same pipeline. Whenever a command is sent, it is received by the `control_supervisor` and the `control_interface` modules. The `control_supervisor` module is responsible for enforcing several safety constraints. It monitors factors such as control frequency, command delays, and potential self-collisions. If an abnormal condition is detected, it triggers an emergency stop signal to stop the hammer. This signal is received by the `control_interface`, which converts the commands into a different message type that can be consumed by the simulated or real machine drivers. When the `control_supervisor` issues an emergency signal, the `control_interface` immediately mutes all outgoing commands. However, we must note that to isolate the performance of the real policies' during evaluations in Gazebo and in the real world, the `control_supervisor` module remains disabled for all experiments conducted in this work.

Finally, in the case of the real machine, the `bobcat_e10_driver` translates the incoming ROS commands into CAN frames, which are received by the `emcb_200u_driver` and then written to the machine's PLC. This component also reads information from the PLC and shares it with the Bobcat driver through CAN frames, which are parsed, pre-processed, and published to the rest of the system; in particular, this allows the population of data to

construct the `JointState` messages of the real machine.

APPENDIX C TRAINING DETAILS FOR DYNAMIC MODELS

Table V shows the search space of all the hyperparameters that are jointly optimized when attempting to learn dynamic models via different parameterizations (MLPs or KANs). Note that for a given parameterization and hyperparameter, the search space does not make a distinction between the type of prediction of the dynamic model (Δq or $\Delta \dot{q}$). Also note that all hyperparameter search spaces are discrete and, except for k (the number of lags used to construct the approximate state \tilde{x}_t), are defined with values that are of common usage in standard machine learning practice. For k , the search space is defined considering the maximum measured actuation delay (around one second) when controlling each actuator of the mini-excavator arm with a constant joint command (-1 or 1 , prior to denormalization) at 20 Hz.

When training MLPs, we use sigmoid linear unit (SiLU) activation functions for each hidden layer and a linear activation function for the output layer. When training KANs, we use Legendre basis functions for each layer. The instantiation of KANs in JAX is done using the `jaxKAN` package (Rigas and Papachristou, 2025).

To adjust the hyperparameters shown in Table V, we use the Optuna package. To cover the range of possibilities for the type of model parameterization (MLP or KAN) and prediction

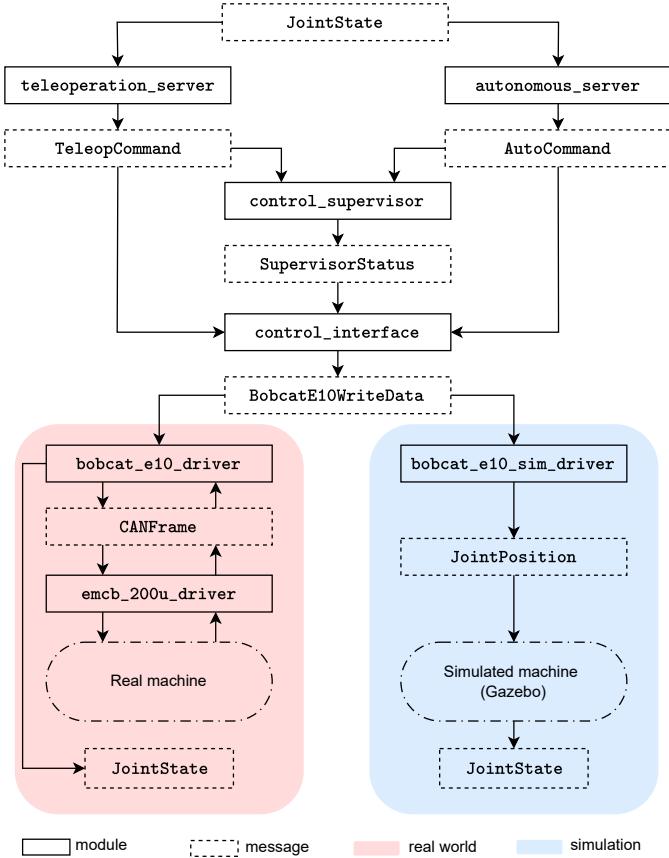


Fig. 13. ROS-based pipeline for the reaching controller when deployed in simulation (using Gazebo) and in the real world.

TABLE V
SEARCH SPACES FOR THE DYNAMIC MODELS HYPERPARAMETERS.

Parameterization	Hyperparameter	Search space
MLP & KAN	Nb. of lags k	{18, ..., 25}
	Learning rate	{ $10^{-3}, 10^{-3}, 10^{-5}$ }
	Batch size	{256, 512, 1024, 2048}
MLP	Loss time horizon H	{4, 8, 16, 32}
	Nb. of layers	{2, 3, 4, 5}
	Nb. of units per layer	{64, 128, 256, 512}
KAN	Loss time horizon H	{8, 16, 32}
	Nb. of layers	{3, 4, 5}
	Nb. of unites per layer	{4, 8, 16, 32}

$(\Delta q$ or $\Delta \dot{q}$), four independent optimization processes are conducted, each of them for models parameterized using an MLP or a KAN, and predicting angular position or angular velocity residuals. The optimization, in all cases, is performed with respect to $\mathcal{L}(\theta)|_{H=80}$, which, as stated in Section V-C, corresponds to the loss function used for training with a fixed time horizon, $H = 80$.

In all cases, we train each model for a maximum of 1000 and 1200 epochs for MLP and KAN parameterizations, respectively, and use the validation subset to end a given trial if $\mathcal{L}(\theta)|_{H=80}$ does not decrease for 50 consecutive training steps (early stopping). The optimization process allows us to obtain the four top dynamic models that are studied in Section V-C.

TABLE VI
MODELING AND ALGORITHMS HYPERPARAMETERS.

	Parameter	Value
Control	Control frequency [Hz]	20
	Action repeats	4
	Loop tracker k_p, k_i	10.0, 60.0
Episodic conditions	$\alpha_{pitch}^{max}, \alpha_{pitch}^{ts}$ [rad]	1.047, 0.698
	$z_{min}^{ts}, z_{max}^{ts}$ [m]	0.155, 2.355
	Max number of steps T_{reset}	500
Reward function	$\lambda_p, \lambda_r, \lambda_q, \lambda_a, \lambda_{W+}$	1.0, 1.0, 1.0, 0.5, 2.0
	$\epsilon_p, \epsilon_r, \epsilon'_p, \epsilon'_r$	0.1, 0.1, 0.05, 0.02
	$\epsilon_\alpha, \epsilon_q, w_\mathcal{X}, w_\alpha, w_q$	0.0001, 0.0025, 0.5, 1.0, 1.0
PPO	Training steps	400M
	Learning rate	$3 \cdot 10^{-4}$
	Entropy cost	10^{-2}
	Discount factor γ	0.97
	Unroll length	40
	Clipping ϵ	0.2
	Batch size	256
	Number of mini-batches	32
	Number of environments	128
	Updates per batch	4
iCEM	Policy parameterization	MLP(256,) × 4
	Critic parameterization	MLP(256,) × 5
iCEM	Horizon H_{MPC}	20
	Number of samples N_{pop}	500
	Number of elites N_e	50
	Initial std σ_0	0.5
	Smoothing coefficient α	0.05
	Optimization steps K	12
	Exponent	2.0
	Elite set fraction	0.5
	Action repeats	4

APPENDIX D FURTHER DETAILS REGARDING POLICY SYNTHESIS

Table VI shows the hyperparameters and the values they take for the control and modeling aspects associated with the reaching task. This table also presents the parameters set for policy synthesis using PPO and iCEM.

The policy learning process using PPO is conducted using a computer equipped with an Intel i7-12700 CP CPU, a NVIDIA RTX 4060 GPU, and 32 GB of RAM. Training the reaching policies for 400M of steps takes approximately 6 min. on average. The iCEM controller instantiated with the parameters reported in Table VI (running on the same computer) averages an optimization computation time of approx. 90 ms. This result makes this method suitable for controlling the machine at 20 Hz when using 4 action repeats, considering that the experiments conducted in Gazebo (Sim2Sim) show that the lag in the generation of control commands that the above induces does not affect performance severely (see Section V-E1).

For the PPO policy trained using the learned MLP $\Delta \dot{q}$ model, an ablation study on the components of the reward function is provided in Table VII. Besides the evaluation metrics used in the results presented in Sections V-D and V-E, we include the avg. traversed end-effector path length (EEF-PL), which is computed by summing the Euclidean distance between successive end-effector positions for a given episode up until the target is reached with a certain precision (0.02 [cm] and 0.02 [rad]), or until T_{max} steps pass. The distances obtained

TABLE VII
PERFORMANCE OF THE PPO MLP $\Delta\dot{q}$ POLICY WHEN TRAINED USING DIFFERENT REWARD FUNCTIONS.

Reward function	Avg. Return*	SR(0.02, 0.02)	SR(0.04, 0.04)	OOW	OOZ	Min. z [m]	EEF-PL [m]
r_t^X	-201.65 ± 98.81	0.30 ± 0.33	0.44 ± 0.38	0.22 ± 0.04	0.12 ± 0.03	-0.20 ± 0.12	4.50 ± 1.04
r_t^Q	-91.17 ± 0.49	0.97 ± 0.02	1.00 ± 0.00	0.21 ± 0.01	0.16 ± 0.01	-0.16 ± 0.01	2.41 ± 0.04
$r_t^X + r_t^\epsilon$	842.83 ± 20.26	1.00 ± 0.01	1.00 ± 0.00	0.16 ± 0.01	0.12 ± 0.01	-0.24 ± 0.06	2.12 ± 0.04
$r_t^Q + r_t^\epsilon$	870.74 ± 10.72	1.00 ± 0.00	1.00 ± 0.00	0.22 ± 0.01	0.17 ± 0.01	-0.19 ± 0.09	2.10 ± 0.02
$r_t^X + r_t^Q$	-290.18 ± 68.68	0.59 ± 0.35	0.77 ± 0.21	0.27 ± 0.10	0.20 ± 0.09	-0.30 ± 0.21	3.83 ± 1.37
$r_t^X + r_t^Q + r_t^\epsilon$	753.17 ± 12.60	1.00 ± 0.00	1.00 ± 0.00	0.18 ± 0.01	0.13 ± 0.01	-0.24 ± 0.06	2.10 ± 0.03
$r_t^X + r_t^Q + r_t^\epsilon + r_t^a$	719.32 ± 14.33	0.99 ± 0.01	1.00 ± 0.00	0.19 ± 0.02	0.14 ± 0.01	-0.24 ± 0.06	2.14 ± 0.03
$r_t^X + r_t^Q + r_t^a + r_t^W$	-356.61 ± 203.17	0.59 ± 0.30	0.78 ± 0.37	0.16 ± 0.04	0.10 ± 0.03	-0.12 ± 0.11	3.97 ± 1.77
$r_t^X + r_t^Q + r_t^\epsilon + r_t^a + r_t^W$	683.36 ± 16.94	0.99 ± 0.01	1.00 ± 0.00	0.15 ± 0.01	0.11 ± 0.01	-0.10 ± 0.08	2.15 ± 0.05

* For each row, the computation of returns depends on the reward function used for training.

are then averaged across the evaluation episodes.

The results suggest that including the r_t^ϵ term in the reward function definition is crucial for achieving high success rates at high precision for controllers trained using PPO.

The results also show that good performance in terms of success rate can be achieved by relying on r_t^ϵ alongside either r_t^Q , r_t^X or both terms. We must note that this is expected, since both r_t^Q and r_t^X serve the purpose of guiding the robot towards the target pose; however, combining r_t^Q and r_t^X without the bonus term results in high variance across training trials. We attribute this behavior to r_t^Q and r_t^X being potentially conflicting, and r_t^ϵ serving as a term that dominates the optimization landscape due to the high reward the agent receives if it reaches the target pose. Finally, we must also note that, regardless of the potential tension between terms, we choose to use both r_t^Q and r_t^X in the final reward function definition because the agent's observations are constructed using both position-quaternion tuples and the associated configurations to represent current and target end-effector poses.

APPENDIX E

FURTHER RESULTS REGARDING THE SIM2SIM TRANSFER EXPERIMENTS

Fig. 14 shows the success rate matrices obtained by evaluating the RL- and MPC-based reaching policies in the Brax-based environment in which they were synthesized (Fig. 14a) and the matrices obtained when evaluating the same policies in Gazebo (Fig. 14b). For a detailed analysis on these results, refer to Section V-E1.

ACKNOWLEDGMENTS

The authors thank Claudio Palacios for designing and implementing the electro-hydraulic interventions that were required to automate the Bobcat E10. The authors also acknowledge Pablo Alfessi for creating a CAD model of the Bobcat E10, and for designing, manufacturing, and installing the mountings of the rotary's encoders on the robot's arm. Finally, the authors also thank Cristian Rivera for implementing most of the additional instrumentation required for automating the machine, and for designing and implementing the program that runs on the PLC.

REFERENCES

- Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.
- Bender, F. A., Göltz, S., Bräunl, T., and Sawodny, O. (2017). Modeling and offset-free model predictive control of a hydraulic mini excavator. *IEEE Transactions on Automation Science and Engineering*, 14:1682–1694.
- Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2018). JAX: composable transformations of Python+NumPy programs.
- Correa, M., Cárdenas, D., Carvajal, D., and Ruiz-del Solar, J. (2022). Haptic teleoperation of impact hammers in underground mining. *Applied Sciences*, 12(3):1428.
- De Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67.
- Dulac-Arnold, G., Evans, R., van Hasselt, H., Sunehag, P., Lillicrap, T., Hunt, J., Mann, T., Weber, T., Degrif, T., and Coppin, B. (2015). Deep reinforcement learning in large discrete action spaces. *arXiv preprint arXiv:1512.07679*.
- Egli, P. and Hutter, M. (2020). Towards rl-based hydraulic excavator automation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2692–2697. IEEE.
- Egli, P. and Hutter, M. (2022). A general approach for the automation of hydraulic excavator arms using reinforcement learning. *IEEE Robotics and Automation Letters*, 7:5679–5686.
- Egli, P., Terenzi, L., and Hutter, M. (2024). Reinforcement learning-based bucket-filling for autonomous excavation. *IEEE Transactions on Field Robotics*.
- Freeman, C. D., Frey, E., Raichuk, A., Girgin, S., Mordatch, I., and Bachem, O. (2021). Brax - a differentiable physics engine for large scale rigid body simulation.
- Greiser, L., Demir, O., Hartmann, B., Hose, H., and Trimpe, S. (2024). Feedforward controllers from learned dynamic local model networks with application to excavator assistance functions. *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 4105–4112.
- Habibi, S. R. and Richards, R. J. (1991). Computed-torque and variable-structure multi-variable control of a hydraulic

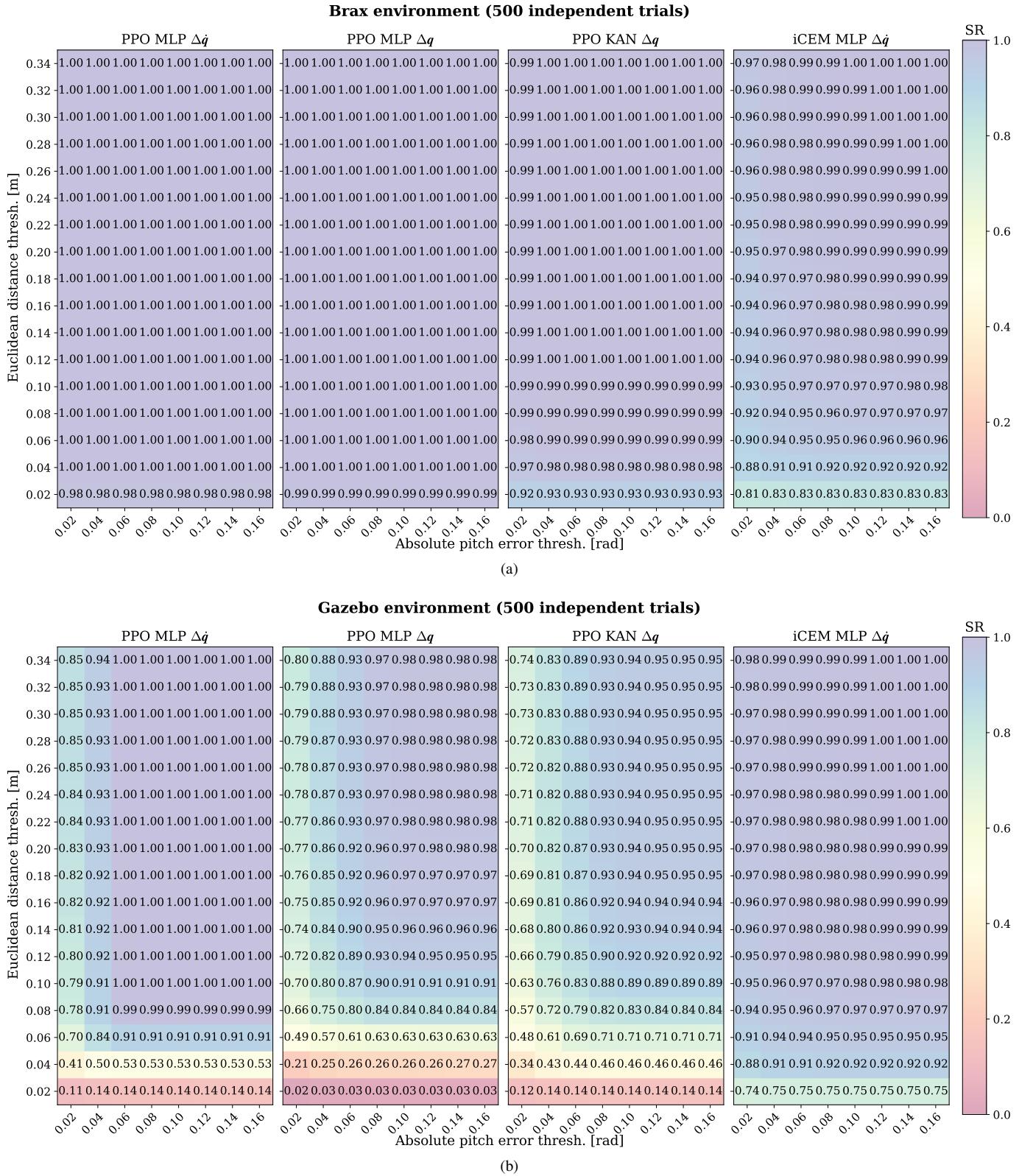


Fig. 14. Success rate matrices obtained when evaluating the reaching policies in (a) the Brax-based environment used for training, and (b) the Gazebo simulator, by encapsulating the policies in action servers.

- industrial robot. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 205:123 – 140.
- Heek, J., Levskaya, A., Oliver, A., Ritter, M., Rondepierre, B., Steiner, A., and van Zee, M. (2024). Flax: A neural network library and ecosystem for JAX.
- Jiang, T., Guan, Y., Ma, L., Xu, J., Meng, J., Chen, W., Zeng, Z., Li, L., Wu, D., and Chen, R. (2024). Dexsim2real²: Building explicit world model for precise articulated object dexterous manipulation. *CoRR*, abs/2409.08750. arXiv preprint arXiv:2409.08750.
- Jud, D., Kerscher, S., Wermelinger, M., Jelavic, E., Egli, P., Leemann, P., Hottiger, G., and Hutter, M. (2021). Heap - the autonomous walking excavator. *ArXiv*, abs/2106.05059.
- Koenig, N. and Howard, A. (2004). Design and use paradigms for gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, pages 2149–2154. Ieee.
- Lampinen, S., Niemi, J., and Mattila, J. (2020). Flow-bounded trajectory-scaling algorithm for hydraulic robotic manipulators. In *2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 619–624. IEEE.
- Lampinen, S., Niu, L., Hulttin, L., Niemi, J., and Mattila, J. (2021). Autonomous robotic rock breaking using a real-time 3d visual perception system. *Journal of Field Robotics*, 38(7):980–1006.
- Lee, M., Choi, H., Kim, C., Moon, J., Kim, D., and Lee, D. (2022). Precision motion control of robotized industrial hydraulic excavators via data-driven model inversion. *IEEE Robotics and Automation Letters*, 7(2):1912–1919.
- Li, C., Ai, Z., Wu, T., Li, X., Ding, W., and Xu, H. (2024). Deformnet: Latent space modeling and dynamics prediction for deformable object manipulation. *CoRR*, abs/2402.07648. arXiv preprint arXiv:2402.07648.
- Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljacic, M., Hou, T. Y., and Tegmark, M. (2025). KAN: Kolmogorov–arnold networks. In *The Thirteenth International Conference on Learning Representations*.
- Ma, D., Liu, Y., Liu, W., and Zhou, B. (2024). A data-driven modeling and motion control of heavy-load hydraulic manipulators via reversible transformation. *arXiv preprint arXiv:2411.13856*.
- Ma, D. and Zhou, B. (2024). Data-driven multi-step nonlinear model predictive control for industrial heavy load hydraulic robot. *ArXiv*, abs/2411.13859.
- Mattila, J., Koivumäki, J. E. M., Caldwell, D. G., and Semini, C. (2017). A survey on control of hydraulic robotic manipulators with projection to future trends. *IEEE/ASME Transactions on Mechatronics*, 22:669–680.
- Nurmi, J. and Mattila, J. (2017). Automated feed-forward learning for pressure-compensated mobile hydraulic valves with significant dead-zone. In *Fluid Power Systems Technology*, volume 58332, page V001T01A027. American Society of Mechanical Engineers.
- Park, J., ju Lee, B., Kang, S., Kim, P. Y., and Kim, H. J. (2017). Online learning control of hydraulic excavators based on echo-state networks. *IEEE Transactions on Automation Science and Engineering*, 14:249–259.
- Pinneri, C., Sawant, S., Blaes, S., Achterhold, J., Stueckler, J., Rolinek, M., and Martius, G. (2021). Sample-efficient cross-entropy method for real-time planning. In *Conference on Robot Learning*, pages 1049–1065. PMLR.
- Plummer, A. R. and Vaughan, N. D. (1990). Robust adaptive control for hydraulic servosystems. *Journal of Dynamic Systems Measurement and Control-transactions of The Asme*, 118:237–244.
- Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A. Y., et al. (2009). Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe.
- Rigas, S. and Papachristou, M. (2025). jaxkan: A unified jax framework for kolmogorov–arnold networks. *Journal of Open Source Software*, 10(108):7830.
- Samtani, P., Leiva, F., and Ruiz-del Solar, J. (2023). Learning to break rocks with deep reinforcement learning. *IEEE Robotics and Automation Letters*, 8(2):1077–1084.
- Sancaktar, C., Blaes, S., and Martius, G. (2022). Curious exploration via structured world models yields zero-shot object manipulation. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 24170–24183. Curran Associates, Inc.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Soroushpour, M. R. and Salcudean, S. E. (2001). Nonlinear control of hydraulic robots. *IEEE Trans. Robotics Autom.*, 17:173–182.
- Sohl, G. A. and Bobrow, J. E. (1997). Experiments and simulations on the nonlinear control of a hydraulic servosystem. *Proceedings of the 1997 American Control Conference (Cat. No.97CH36041)*, 1:631–635 vol.1.
- Song, B. and Koivo, A. J. (1995). Neural adaptive control of excavators. *Proceedings 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human Robot Interaction and Cooperative Robots*, 1:162–167 vol.1.
- Spinelli, F. A., Egli, P., Nubert, J., Nan, F., Bleumer, T., Goegler, P., Brookes, S., Hofmann, F., and Hutter, M. (2024). Reinforcement learning control for autonomous hydraulic material handling machines with underactuated tools. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12694–12701. IEEE.
- Spinelli, F. A., Zhai, Y., Nan, F., Egli, P., Nubert, J., Bleumer, T., Miller, L., Hofmann, F., and Hutter, M. (2025). Large scale robotic material handling: Learning, planning, and control. *arXiv preprint arXiv:2508.09003*.
- Todorov, E., Erez, T., and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE.
- Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1(4):339–356.