

Logic-Skill Programming: An Optimization-based Approach to Sequential Skill Planning

Teng Xue^{1,2}, Amirreza Razmjoo^{1,2}, Suhan Shetty^{1,2}, Sylvain Calinon^{1,2}
¹IIdiap Research Institute ²École Polytechnique Fédérale de Lausanne (EPFL)

Abstract—Recent advances in robot skill learning have unlocked the potential to construct task-agnostic skill libraries, facilitating the seamless sequencing of multiple simple manipulation primitives (aka. skills) to tackle significantly more complex tasks. Nevertheless, determining the optimal sequence for independently learned skills remains an open problem, particularly when the objective is given solely in terms of the final geometric configuration rather than a symbolic goal. To address this challenge, we propose Logic-Skill Programming (LSP), an optimization-based approach that sequences independently learned skills to solve long-horizon tasks. We formulate a first-order extension of a mathematical program to optimize the overall cumulative reward of all skills within a plan, abstracted by the sum of value functions. To solve such programs, we leverage the use of tensor train factorization to construct the value function space, and rely on alternations between symbolic search and skill value optimization to find the appropriate skill skeleton and optimal subgoal sequence. Experimental results indicate that the obtained value functions provide a superior approximation of cumulative rewards compared to state-of-the-art reinforcement learning methods. Furthermore, we validate LSP in three manipulation domains, encompassing both prehensile and non-prehensile primitives. The results demonstrate its capability to identify the optimal solution over the full logic and geometric path. The real-robot experiments showcase the effectiveness of our approach to cope with contact uncertainty and external disturbances in the real world.

I. INTRODUCTION

Consider the following task: "A large box is positioned on the table, next to a wall. The objective is to reorient the box to a new 6D pose using a single robot manipulator with minimal control efforts (or maximal rewards). The robot is allowed to have any interactions with the surroundings." A potential solution involves pushing the box until it reaches the wall, then pivoting against the wall, followed by pulling it to the target. It is noteworthy that the objective is only given in terms of the evaluation of the final geometric configuration, and potential control costs.

Such tasks are quite common in sequential manipulation scenarios, typically involving the sequencing of multiple manipulation primitives, such as push, pivot, and pull, to achieve a long-horizon target with sparse rewards. Solving these tasks requires reasoning about the appropriate sequence of primitives and corresponding motion trajectories. This hybrid structure results in combinatorial complexity, making it expensive to find a solution. To address this challenge, Mixed-Integer Programming (MIP) [15] is an intuitive approach that does not require a careful design of the system model but relies on branch-and-bound techniques to efficiently prune

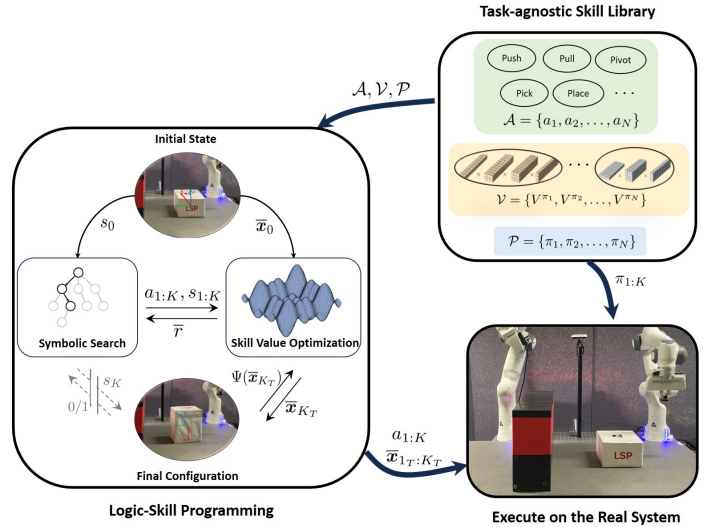


Fig. 1: Overview of the proposed approach: Given the evaluation function Ψ of the final configuration, along with the initial symbolic state s_0 and geometric state \bar{x}_0 , the objective of LSP is to find a solution that can accomplish the task with minimal control costs. A task-agnostic skill library is pretrained, consisting of N skill operators $\mathcal{A} = \{a_{1:N}\}$, along with corresponding value functions $\mathcal{V} = \{V^{\pi_{1:N}}\}$ and policies $\mathcal{P} = \{\pi_{1:N}\}$ in Tensor Train format. LSP solves this problem by alternating between symbolic search and skill value optimization for joint logic-geometric reasoning. Symbols $s_{1:K}$ are used as constraints for skill optimization, while skill optimization is used to check skeleton feasibility and final configuration performance, with a feedback reward \bar{r} informing the symbolic search. This results in the appropriate skill skeleton $a_{1:K}$ and subgoal sequence $\bar{x}_{1:T:K_T}$, which are then combined with the skill policies $\pi_{1:K}$ existing in the skill library to actuate the real robot. Notably, the gray channel with symbolic final state s_K is interrupted because our framework eliminates the need for a symbolic target goal s_T , while such information is typically required in existing sampling-based sequential skill planning methods.

solutions. While MIP works well with convex optimization formulations, such as graph of convex sets [16], manipulation tasks involving interactions with the surroundings usually violate this assumption. Another approach is to implicitly model the hybrid system with complementary constraints [23, 19]. By uncovering the internal structures of different manipulation primitives, this method eliminates the need for integer variables in problem formulation, allowing the use of continuous optimization techniques. However, it often leads to poor local optima.

A more general approach involves using logic as a combi-

natorial expression of possible manipulation primitives [36]. This approach follows the principles of MIP but extends the formulation to the first-order logic level, allowing the utilization of powerful classical AI techniques. This idea is related to a field known as Task and Motion Planning (TAMP) [9], where the objective is to find a feasible or optimal plan for a complicated task given full knowledge of the system and environments. Various approaches have been proposed to solve TAMP problems, including PDDLStream [8] and Logic-Geometric Programming [35]. TAMP methods have demonstrated high performance in diverse scenarios, such as construction assembly [11], table rearrangement [24], and mobile manipulation [43]. However, they often exhibit poor performance in many realistic scenarios involving model uncertainty and external disturbances. For instance, considering the task introduced at the beginning, although we can derive an offline feasible/optimal trajectory using predefined contact parameters like friction coefficients and damping parameters, this trajectory can be totally wrong because we cannot know exactly the contact model between the robot, objects, and the environment.

Thanks to the advances in learning-based techniques, it is now possible to learn a powerful skill policy for each specific manipulation primitive. The policy can be seen as a short-horizon model predictive controller (MPC) with a good terminal cost function, showing robust and reactive behaviors in the face of model uncertainty and external disturbances. When compared with methods that require online planning [41, 12], policy-based methods [33, 31, 4] usually exhibit good performance in physical manipulation tasks that need to cope with contact uncertainty.

We can construct a skill library composed of multiple task-agnostic policies trained independently. This allows for the iterative augmentation of the skill library in a lifelong manner. However, sequencing such task-agnostic policies remains an open problem. Two questions should be considered:

1) Within the library, multiple skills are available. To tackle an unseen, long-horizon manipulation task, the questions arise: which skills should be employed, and what is the correct order?

2) The acquired skills are task-agnostic. Each skill depends on a specific subgoal to generate the appropriate action given the current state. How to determine the subgoal sequence given a skill skeleton?

There has been prior work aimed at addressing these questions [2, 13, 39]. However, most existing approaches necessitate a well-defined task planning problem with an explicit symbolic goal description, followed by sampling-based methods to check whether the symbol-defined constraints can be satisfied. This limitation restricts the applicability of such methods to the tasks mentioned at the beginning. Moreover, it is worth noting that these tasks are, in fact, optimization problems rather than constraint satisfaction problems.

To address these issues, we introduce Logic-Skill Programming, an optimization-based approach designed to eliminate the need for explicit symbolic goal descriptions. It also

provides a notion of global optimality over the full logic-geometric path. This work draws inspiration from Logic-Geometric Programming [35] and shares the same philosophy, with a key distinction that we concentrate on skill policy planning, rather than motion planning in a fully-known environment.

Fig. 1 is an overview of our proposed approach. Given the initial state s_0 and \bar{x}_0 , along with the evaluation function Ψ for the final geometric configuration, LSP can generate the solutions to sequence multiple skills from the task-agnostic skill library. This problem is formulated as an extended first-order mathematical program, where first-order symbols are introduced to constrain the optimization problem. The objective is to optimize both the overall cumulative reward and the performance of the final geometric configuration. The overall cumulative reward is expressed as the sum of all value functions within the plan, forming a value function space. While Reinforcement Learning (RL) algorithms have excelled in tackling highly challenging problems, the approximations of the obtained value functions lack generalizability to the entire state space, which is critical in sequential skill planning for determining the optimal subgoal sequence with maximum cumulative reward. We therefore propose to use Tensor Train (TT) to approximate the value function space, which shows superior approximation capabilities compared to RL methods. The appropriate skill skeleton and optimal subgoal sequence are then obtained by alternating between symbolic search and skill value optimization over the value function space.

We summarize the contributions of this paper as follows:

- We propose to formulate the sequential skill planning problem as an extended first-order mathematical program, eliminating the need for symbolic goal description and enabling to find the optimal solutions rather than only feasible ones.
- We propose Logic-Skill Programming to address sequential skill planning tasks by alternating between symbolic search and skill value optimization.
- We propose to use Tensor Train to approximate the value function space, aiding in finding the optimal subgoal sequence with maximum cumulative reward respecting system dynamics/kinematics.

II. RELATED WORK

A. Skill Learning

Methods such as behavior cloning (BC) [7] involve deep neural networks learning state-action distributions from offline datasets. This class of methods heavily relies on the quality of the provided datasets, lacking exploration in complex state spaces. An alternative is to consider RL. Thanks to the exploration-exploitation mechanism, RL algorithms can actively explore the state space and find a feasible path given any initial states. However, RL primarily explores the local state space determined by the exploration metric. Once the final target is reached, RL usually loses motivation to explore further to find more optimal solutions.

In this paper, we aim to find global optimality over the entire (logic and geometric) path. To achieve this, we need to know the value function accurately over the entire state space. Notably, both BC and RL methods fall short in providing this information. This assertion aligns with the motivation of Approximate Dynamic Programming (ADP), where the goal is to approximate the optimal value function for the entire state space. However, this objective is challenging due to the expensive storage and computation involved in solving the dynamic programming algorithms such as value iteration. We believe this challenge explains why existing literature primarily focuses on feasibility rather than pursuing optimality in sequential skill planning. Recently, Shetty et al. [31] proposed the use of Tensor Train for ADP and performed well in several hybrid control benchmarks. We find it promising to extend this method to learning manipulation skills, leveraging its advantages for accurate value function approximation across the entire state space for optimal sequential skill planning.

B. Hybrid Long-horizon Planning

Robot manipulation is typically characterized by hybrid aspects, such as contact modes (sticking/sliding) and manipulation primitives (pushing/pivoting). This hybrid structure poses significant challenges for gradient-based optimization techniques. To address this issue, one class of methods uses MIP [15]. MIP involves both discrete and continuous variables in optimization problems and relies on Branch-and-Bound techniques to efficiently search for optimal solutions by pruning undesirable ones. It performs well if the original problem can be reformulated as combinations of several convex optimization sub-problems. However, manipulation tasks are usually challenging to convexify. Another well-studied method is Mathematical Program with Complementary Constraints (MPCC), which eliminates integer variables by adding complementary constraints on the decision variables. Subsequently, augmented Lagrangian techniques are employed to solve constrained optimization problems. This class of methods has demonstrated excellent performance in legged robot locomotion [23] and planar manipulation tasks [19]. However, the added constraints often render the problem more fragile and prone to getting stuck in poor local optima.

Another promising line of research is Task and Motion Planning (TAMP) [9]. It adopts the concept of MIP but replaces integers with symbols, enabling the utilization of powerful planning tools from the classical AI field. Existing methods in TAMP can be categorized into two branches: sampling-based and optimization-based methods. The sampling-based methods [32, 8] usually assume a symbolic goal description and rely on a high-level task planner to find a feasible action skeleton. These methods often introduce predicates to represent geometric properties at a symbolic level and aim to identify correct symbolic abstractions of geometries, allowing reasoning solely at the symbolic level. However, finding the right abstractions can be non-trivial. Conversely, optimization-based approaches [35, 42] aim to solve problems at the geometric level, treating symbolic logic as constraints in mathematical programming.

While these methods can be less efficient than sampling-based ones, they can handle arbitrary objectives specified only in terms of an evaluation function for the final geometric configuration. Moreover, they provide a notion of global optimality over the entire logic and geometric path. This philosophy has strongly influenced our work, inspiring the removal of symbolic goals and the pursuit of optimal solutions for sequential skill planning.

In general, TAMP performs well across various domains [11, 24, 43]. However, its reliance on full knowledge of the planning domain and dynamics model limits its applicability in realistic environments. For instance, modeling the contact or interaction between the robot and the environment is often impossible, leading to poor performance of offline planned trajectories in online setups. One potential solution to this issue could be the use of model predictive control for receding horizon planning [37]. Nevertheless, this approach may result in somewhat short-sighted behavior, and the intrinsic combinatorial structure also makes it expensive in terms of online planning. The weaknesses of TAMP highlight the importance of leveraging learned skill policies to address uncertainties and disturbances in the real world, particularly those involving physical contact.

C. Sequential Skill Planning

Prior works that focus on sequential skill planning typically adopt the options framework [34], where a high-level policy is trained to sequence low-level skills toward the final goals. For instance, Xu et al. [40] proposed learning a skill proposal network as the high-level policy and utilizes learned skill-centric affordances (value functions) to assess the feasibility of the proposed skill skeleton. Similarly, Shah et al. [29] suggested using the value functions of low-level skills as the state space to represent the symbolic skill affordance. This representation is then utilized to train an upper-level RL policy toward long-horizon goals. Although such methods have demonstrated the ability to solve long-horizon tasks, the resulting policies are typically task-specific and exhibit poor generalization on unseen tasks. Moreover, we believe the capabilities of value function has not been fully explored by [29, 40]. Instead of using the value function space symbolically, we regard it as an abstraction of cumulative reward, taking into account the system kinematics/dynamics in geometric level. Subsequently, we leverage it to identify the optimal subgoal sequence that results in the maximum cumulative reward of the full path.

To enhance generalization ability, several recent works rely on symbolic planning to sequence task-agnostic skills. In Agia et al. [2], the product of Q-functions is maximized to ensure the joint success of all skills sequenced in a plan, with the skill skeleton given by a high-level task planner. Similarly, in Huang et al. [13], a symbolic planner is used to sequence imitation learning policies, by leveraging continuous relaxation to improve symbolic grounding success. In Wu et al. [39], a repertoire of visuomotor skills is learned through human-provided example images. The pre-conditions and effects of each action are represented by images, enabling the verifi-

cation of the feasibility of the skill skeleton provided by a high-level symbolic planner. The key is the formulation of a constraint satisfaction problem where the effects of the parameterized skill primitive satisfy the preconditions of the next skill in the plan skeleton. The plan skeleton is either predefined [18] or obtained through PDDL planners [2], requiring an explicit symbolic goal description. None of these methods aims to optimize over a final configuration given only by an objective function, limiting their ability to solve the problem mentioned at the beginning.

Therefore, we are motivated to propose a method that eliminates the need for a symbolic goal description while ensuring the maximum cumulative reward of the obtained skill skeleton and motion trajectories. This work aligns with LGP [35], except that we plan over skills rather than motions, as we believe skills are more applicable in realistic environments.

III. BACKGROUND

A. Tensor Train for Function Approximation

A multivariate function $f(x_1, \dots, x_d)$ over a rectangular domain \mathcal{D} can be approximated by a tensor \mathcal{F} , where each element in the tensor represents the value of the function given the discretized inputs. The value of function f at any point in the domain can then be approximated by interpolating among the elements of the tensor \mathcal{F} .

Due to storage limitations, representing a high-dimensional tensor is challenging. Tensor Train (TT) decomposition [21] was proposed to solve this problem by representing the tensor using a set of third-order tensors called *cores*. TT-Cross [20, 27] is a widely used technique to approximate a function in TT format. For more introductions about these techniques, readers may refer to [31] for a detailed review.

B. Logic-Geometric Program

Logic-geometric Programming was proposed by [35] to integrate first-order logic into a mathematical program for addressing combined Task and Motion Planning (TAMP) problems. The general formulation is as follows: Given a logic \mathcal{L} , a knowledge base $\mathcal{K} \in \mathcal{L}$, and an objective function $l(x)$ over geometric configurations $x \in \mathcal{X}$, the logic-geometric program can be expressed as:

$$\min_{x, \kappa} l(x) \quad \text{s.t. } \kappa \models \mathcal{K}, \quad g(x, \kappa) \leq 0, \quad h(x, \kappa) = 0, \quad (1)$$

where \models represents logical implication, indicating that \mathcal{K} is satisfied once the logical statement κ is True. This defines the associated equality and inequality constraints: $g(x, \kappa) \leq 0$ and $h(x, \kappa) = 0$.

Our formulation can be viewed as a special case of logic-geometric program, where the objective function is the total cumulative reward and the evaluation of the final configuration, and the knowledge base is the selected skill skeleton.

IV. METHODOLOGY

A. Problem Formulation

Our objective is to address long-horizon manipulation tasks by sequentially executing a series of skills included in a skill

library $\mathcal{P} = \{\pi_1, \pi_2, \dots, \pi_N\}$. To ensure that the learned skills can be sequenced in an arbitrary manner and generalized to any long-horizon tasks, the skills should be task-agnostic, as also described in Agia et al. [2]. This implies that the learned policy is trained independently, with its own skill-centric parameters. Each skill domain can be modeled by a Markov Decision Process (MDP)

$$\mathcal{M}_n = (\mathcal{X}_n, \mathcal{U}_n, \mathcal{T}_n, \mathcal{R}_n), \quad (2)$$

where \mathcal{X}_n is the state space, \mathcal{U}_n is the action space, $\mathcal{T}_n(x'_n | x_n, u_n)$ is the transition model, $\mathcal{R}_n(x_n, u_n)$ is the reward function given current state and action.

The full K -length long-horizon domain is the union of $\{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_K\}$, where each time segment corresponds to a single skill policy. It is specified as

$$\overline{\mathcal{M}} = (\mathcal{M}_{1:K}, \overline{\mathcal{X}}, \Phi_{1:K}, \Gamma_{1:K}), \quad (3)$$

where $\overline{\mathcal{X}}$ is the full state space of the long-horizon domain, $\Phi_{1:K} : \mathcal{X}_k \rightarrow \overline{\mathcal{X}}$ is a function that maps the policy state space to the full state space, and $\Gamma_{1:K} : \overline{\mathcal{X}} \rightarrow \mathcal{X}_k$ is a function that extracts the policy state from the full state. Note that the state spaces of different skills usually have different dimensions and structures. Taking pushing and pivoting as examples, they are planar manipulation primitives defined in horizontal and vertical planes, respectively, affecting different elements of the object pose after execution. The long-horizon state space serves as a bridge, transmitting the local state changes from the previous skill to the subsequent one.

After acquiring these task-agnostic policies, our goal is to determine the optimal skill skeleton and subgoal sequence, given the evaluation function Ψ of the final configuration. Each skill $\pi_k(\mathbf{u} | \mathbf{x})$ is trained to maximize the cumulative reward given the current state \mathbf{x} . Therefore, finding the optimal skill sequence aims to maximize the overall cumulative reward for the complete skill sequence $a_{1:K}$ and the evaluation of the final configuration $\Psi(\overline{\mathbf{x}}_{K_T})$:

$$\max_{\mathbf{a}, a_{1:K}} \sum_{k=1}^K \mathbb{E}_{\pi_k} \left[\sum_{t=0}^{\infty} \gamma^t R_{k_t} \right] + \Psi(\overline{\mathbf{x}}_{K_T}). \quad (4)$$

Here, R_{k_t} denotes the reward of skill a_k at time t . However, directly optimizing over the infinite full horizon is impossible. Considering that the value function is defined to approximate the cumulative reward in dynamic programming:

$$V^{\pi_k}(\mathbf{x}) = \mathbb{E}_{\pi_k} \left[\sum_{t=0}^{\infty} \gamma^t R_{k_t}(\mathbf{x}_{k_t}, \pi_k(\mathbf{x}_{k_t})) \mid \mathbf{x}_{k_0} = \mathbf{x} \right], \quad (5)$$

and the objective of each policy is to find an action from the action space that can maximize the cumulative reward from current state:

$$Q^{\pi_k}(\mathbf{x}, \mathbf{u}) = \mathbb{E}_{\pi_k} \left[\sum_{t=0}^{\infty} \gamma^t R_{k_t}(\mathbf{x}_{k_t}, \mathbf{u}_{k_t}) \mid \mathbf{x}_{k_0} = \mathbf{x}, \mathbf{u}_{k_0} = \mathbf{u} \right],$$

$$\pi_k(\mathbf{x}) = \arg \max_{\mathbf{u}} Q^{\pi_k}(\mathbf{x}, \mathbf{u}). \quad (6)$$

Sequential skill planning problem can be formulated as maximizing the sum of value functions, along with the evaluation function Ψ of the final configuration, incorporating first-order logic as constraints:

$$\begin{aligned}
& \max_{\bar{\mathbf{x}}, a_{1:K}, s_{1:K}} \sum_{k=1}^K V^{\pi_k}(\mathbf{x}_{k_0}) + \Psi(\bar{\mathbf{x}}_{K_T}) \\
& \text{s.t. } s_k \in \text{succ}(s_{k-1}, a_k), \bar{\mathbf{x}}_{1_0} = \bar{\mathbf{x}}_0, \\
& h_{\text{path}}(\mathbf{x}_{k_t}, \pi_k(\mathbf{x}_{k_t}) | s_k) = 0, \\
& g_{\text{path}}(\mathbf{x}_{k_t}, \pi_k(\mathbf{x}_{k_t}) | s_k) \leq 0, \\
& h_{\text{switch}}(\bar{\mathbf{x}}_{k_T} | a_{k+1}, s_k) = 0, \\
& g_{\text{switch}}(\bar{\mathbf{x}}_{k_T} | a_{k+1}, s_k) \leq 0, \\
& \bar{\mathbf{x}}_{k_t} = \Phi_k(\mathbf{x}_{k_t} | a_k), \\
& \mathbf{x}_{k_0} = \Gamma_k(\bar{\mathbf{x}}_{k_0}, \bar{\mathbf{x}}_{k_T} | a_k),
\end{aligned} \tag{7}$$

where $\bar{\mathbf{x}}_{k_0}$ and $\bar{\mathbf{x}}_{k_T}$ are the initial and final configuration of skill operator a_k in the long-horizon domain $\bar{\mathcal{X}}$. The goal configuration of a_k is identical to the initial configuration of a_{k+1} , namely $\bar{\mathbf{x}}^{(k+1)_0} = \bar{\mathbf{x}}_{k_T}$. The initial state of a_k within the skill domain \mathcal{X}_k , denoted as \mathbf{x}_{k_0} , is computed using $\bar{\mathbf{x}}_{k_0}$ and $\bar{\mathbf{x}}_{k_T}$. $V^{\pi_k}(\mathbf{x}_{k_0})$ therefore represents the cumulative reward from $\bar{\mathbf{x}}_{k_0}$ to $\bar{\mathbf{x}}_{k_T}$ in \mathcal{X}_k . Φ_k and Γ_k are the mapping functions between long-horizon domain $\bar{\mathcal{X}}$ and skill domain \mathcal{X}_k . Thanks to Γ_k , given new initializations and targets, there is no need to retrain new value functions. The definition of Γ_k is outlined in Sec. V-A. s_0 and $\bar{\mathbf{x}}_0$ denote the initial symbolic state and geometric configuration, respectively. h_{path} and g_{path} indicate the constraints on the path \mathbf{x}_{k_t} given current symbolic state s_k . Such constraints are addressed by the skill policies π_k in each skill domain. h_{switch} and g_{switch} express the transition consistency of configuration $\bar{\mathbf{x}}_{k_T}$ with the following skill operator a_{k+1} . For example, to `pivot` an object against the wall, the object should be well-positioned in contact with the wall after the previous action. These constraints specify the degrees of freedom in the system configurations that can be actuated under the symbolic state s_k , thereby effectively reducing the number of decision variables in $\bar{\mathbf{x}}_{k_T}$ and streamlining the optimization process. For instance, an object marked as `non-graspable` and `onTable` can only be manipulated through `push` or `pull`, with actuation in the dimensions of `x`, `y` and `yall`. Conversely, if the object is marked as `atWall`, it can be actuated by `pivot`, with changes in `roll` or `pitch`.

The value functions of all skills $a_{1:K}$ collectively constitute a value function space, offering insights into optimal control for sequentially executed skills, while considering system dynamics and kinematics. By optimizing over it, we can identify the subgoal sequence that maximizes the cumulative reward while satisfying system constraints. For instance, as depicted in Fig. 4, the value function space can inform the planner which configuration along the table edge is more dynamically optimal, especially considering the under-actuated pushing dynamics.

We assume the initial geometric configuration $\bar{\mathbf{x}}_0$ is given, along with an initial symbolic state $s_0 \in \mathcal{L}$. The sym-

bolic states can be transited through the skill operator a_k as $s_k = \text{succ}(a_k, s_{k-1})$. The existing skill planning frameworks typically require an explicit symbolic goal target $s_T \models \mathcal{G}$, while our method eliminates this requirement by considering only the evaluation Ψ of the final geometric configuration.

B. Logic-Skill Programming

To solve Eq. (7), we propose an approach that alternates between searching for symbolic skills and optimizing the value functions. The role of symbolic search is to define the optimization constraints, while skill value optimization checks the skeleton feasibility and evaluates the final configuration. Additionally, our method, instead of only seeking feasible solutions, aims to find the optimal solution that maximizes the overall cumulative reward while achieving the final target.

1) *Level 1: Symbolic Search:* We use Planning Domain Definition Language (PDDL) [1] to describe the task domain and then utilize Monte Carlo Tree Search (MCTS) to search for the appropriate skill sequence $a_{1:K}$ from the skill library. The preconditions and effects of each skill form a rule-based representation of symbolic transitions $s_k = \text{succ}(s_{k-1}, a_k)$, serving as the forward model for symbolic search. In each iteration, MCTS relies on the Upper Confidence Bound (UCB1) [38] to select the node:

$$UCB1(s_k) = \frac{w_{s_k}}{v_{s_k}} + C_E \sqrt{\frac{2 \ln(v_{s_k}^p)}{v_{s_k}}}, \tag{8}$$

where v_{s_k} represents the number of visits on symbolic state s_k , and w_{s_k} indicates the total accumulated reward obtained through state s_k . $v_{s_k}^p$ denotes the number of visits on the parent node of s_k . C_E is a constant to balance exploitation and exploration.

If no child node is found in the current branch of the search tree, the branch will expand and simulate until it either reaches the target or surpasses the maximum length. Since our formulation does not have a symbolic goal, we rely on the skill value optimization process (Sec. IV-B2) to evaluate the performance of the final geometric configuration. A reward \bar{r} will be backpropagated through the branch, depending on the simulation result. This iterative process refines the search tree, focusing on promising branches and ultimately converging towards an optimal decision. It is important to note that the sequence length (K in Eq. (7)) is purely unknown before symbolic search, allowing diverse sequence lengths for the same target configuration. By applying a higher C_E in the UCB1 formula, MCTS can return multiple solutions. For example, as shown in Fig. 2b, if we want to grasp the cube that is only graspable from the lateral side, multiple symbolic sequences can be found by symbolic search. One can be `push-pick`, pushing the cube to the edge and then grasping it from the table side, while another solution can be `push-pivot-pull-pick`. Given multiple solutions, we can either choose the most robust one before execution using domain knowledge, or switch between solutions during execution. This will be further studied in the future to fully exploit the advantage of multiple solutions.

2) *Level 2: Skill Value Optimization:* The symbolic skill skeleton can only express the feasibility in symbolic level. It has to be verified by Eq. (7) to evaluate the final configuration while satisfying the constraints. This verification is particularly crucial in our work, since the objective is solely defined by an evaluation function Ψ of the final configuration. Additionally, since the sequenced skills are task-agnostic, finding the subgoal of each skill is crucial for the skill policy to reason about where to go given the current state. All of these considerations highlight the crucial necessity of solving Eq. (7).

The first step involves obtaining the value functions that can accurately approximate the cumulative reward in each skill domain. To achieve this, our approach utilizes Tensor Train for value function approximation and skill policy learning, based on an algorithm called Generalized Policy Iteration using Tensor Train (TTPI) [31]. It is an ADP method that aims to approximate the value function throughout the entire state space. TTPI has demonstrated great performance in hybrid control scenarios in [31], where the value function is used to compute the advantage function for policy retrieval. In this work, we apply TTPI to construct the value function space for skill sequencing.

After obtaining the optimal value functions for different manipulation skills, we can solve Eq. (7) conditioning on the skill skeleton generated from the symbolic search level. The resulting subgoals $\bar{\mathbf{x}}_{1:T:K_T}$ should satisfy both the path constraints and the switch constraints. The switch constraints h_{switch} and g_{switch} dictate that $\bar{\mathbf{x}}_{k_T}$ must lie within the intersection space of two adjacent skill domains, \mathcal{X}_k and \mathcal{X}_{k+1} , ensuring configuration consistency. The path constraints h_{path} and g_{path} are addressed by the skill policies in each skill domain. For example, to verify whether the subgoal $\bar{\mathbf{x}}_{k_T}$ can be achieved while respecting path constraints, the skill policy π_k should be executed in domain \mathcal{X}_k given initial state \mathbf{x}_{k_0} . We assume that the difference between the model parameters used for skill policy learning and those in the real world is modest. Thanks to the receding-horizon mechanism during policy execution, the skill policy can achieve the target despite model uncertainties and disturbances, thereby naturally satisfying h_{path} and g_{path} . This assumption is sensible, aligning with the objective of obtaining a powerful policy in the skill learning community and supported by the robust performance of learned task-specific policies in existing literature [17, 31, 4]. The focus of our paper is then more about how to sequence multiple skills to accomplish a much more complicated task. Additionally, if this assumption does not hold, we can still use the learned skill policies and value functions to find the feasible solution set and select the optimal one that maximizes the objective function in Eq. (7).

To design an appropriate optimization technique, it is essential to note that both discrete and continuous variables may be involved in this problem. For instance, in the task mentioned at the beginning, we have to rely on pivoting against the wall to change the `roll` or `pitch` angle of the box. This requires one face of the box to be parallel to the wall. In other words, the `yall` angle of the box has to be in $[-\pi, -\pi/2, 0, \pi/2]$

after pushing. Moreover, the objective function can be in any form, either convex or non-convex, depending on the value functions of selected skills. All of these factors pose significant challenges for optimization techniques. In this work, we employ the Cross-Entropy Method (CEM) [26, 6] with mixed distribution, namely CEM-MD, as the optimization technique. It can handle mixed-integer programming by using Gaussian distribution and categorical distribution for continuous and discrete variables, respectively. The distributions are iteratively updated towards the fraction of the population with higher objective scores until converging to the best solution. The pseudocode of CEM-MD is shown in Alg. 1. Note that the C samples in each iteration are generated in batch (lines 5-10) for fast computation.

Algorithm 1 CEM-MD: Cross-Entropy Method with Mixed Distribution

- 1: **Input:** Initial mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$ for continuous variables, initial probability vector \mathbf{p} for discrete variables, population size C , elite fraction p , maximum iterations H , number of categories K_c , skill sequence $a_{1:K}$, Initial configuration $\bar{\mathbf{x}}_0$, evaluation function Ψ
 - 2: **Output:** Subgoal sequence \mathbf{x}^* , cumulative reward: J^*
 - 3: Initialize $h \leftarrow 0$
 - 4: **while** $h < H$ **do**
 - 5: **for** $i = 1$ to C **do**
 - 6: Sample continuous variables: $\mathbf{x}_{c_i} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
 - 7: Sample discrete variables: $\mathbf{x}_{d_i} \sim \text{Categorical}(\mathbf{p})$
 - 8: Combine: $\mathbf{x}_i = (\mathbf{x}_{c_i}, \mathbf{x}_{d_i})$
 - 9: Evaluate the samples: $J_i \leftarrow l(\mathbf{x}_i, a_{1:K}, \bar{\mathbf{x}}_0, \Psi)$
 \triangleright objective function in Eq. (7)
 - 10: **end for**
 - 11: Select the top p solutions (elite set):
 $\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_{pC}\}$
 - 12: Update the mean and covariance for continuous variables:

$$\boldsymbol{\mu} \leftarrow \frac{1}{pC} \sum_{i=1}^{pC} \hat{\mathbf{x}}_{c,i}$$
 - 13: Update the probability vector for discrete variables:
 - 14: **for** $k_c = 1$ to K_c **do**

$$\mathbf{p}_{\text{elite},k_c} = \frac{\text{count}(\hat{\mathbf{x}}_{d,k_c})}{pC}$$
 - 15: **end for**

$$\mathbf{p} \leftarrow \mathbf{p}_{\text{elite}}$$
 - 16: $h \leftarrow h + 1$
 - 17: **end while**
 - 18: $\mathbf{x}_c^* \leftarrow \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{x}_d^* \leftarrow \text{Categorical}(\mathbf{p})$
 - 19: $\mathbf{x}^* = (\mathbf{x}_c^*, \mathbf{x}_d^*)$, $J^* \leftarrow l(\mathbf{x}^*, a_{1:K}, \bar{\mathbf{x}}_0, \Psi)$
 - 20: **Output** \mathbf{x}^* and J^*
-

Overall, our proposed approach is to alternate between symbolic search and skill value optimization. The symbolic search is achieved by MCTS, with the obtained skill skele-

ton as the constraints for skill value optimization. The skill value optimization is achieved by CEM-MD, with its results to inform symbolic search. The values of visited nodes in the search tree will be updated, initiating a new iteration. Given an evaluation function Ψ of the final configuration, this framework can return multiple solutions. The pseudocode of our Logic-Skill Programming method is shown in Alg. 2.

Algorithm 2 LSP: Logic-Skill Programming

```

1: Input: Initial configuration  $\bar{x}_0$ , evaluation function  $\Psi$ , initial symbolic state  $s_0$ , maximum iterations  $\tilde{H}$ , maximum number of solutions  $\tilde{N}_s$ 
2: Output: Skill skeleton  $a_{1:K}$ , Subgoal sequence  $\bar{x}_{1T:KT}$ 
3: Initialize  $h \leftarrow 0$ ,  $N_s \leftarrow 0$ , Value Set:  $\mathcal{O}_v \leftarrow \emptyset$ , Solution Set:  $\mathcal{O}_s \leftarrow \emptyset$ 
4: while  $h < \tilde{H}$  do
5:   Symbolic search:  $a_{1:K} = \text{MCTS}(s_0)$ 
6:   Skill value optimization:
7:      $\bar{x}_{1T:KT}, J = \text{CEM-MD}(a_{1:K}, \bar{x}_0, \Psi)$ 
8:   if solved then
9:      $N_s \leftarrow N_s + 1$ 
10:     $\mathcal{O}_v \leftarrow \mathcal{O}_v \cup (J)$ 
11:     $\mathcal{O}_s \leftarrow \mathcal{O}_s \cup (a_{1:K}, \bar{x}_{1T:KT})$ 
12:   end if
13:   if  $N_s \geq \tilde{N}_s$  then
14:     break
15:   end if
16:    $h \leftarrow h + 1$ 
17: end while
18: Output  $\mathcal{O}_s$ 

```

V. EXPERIMENTS

In this section, we compare our method with state-of-the-art baselines, in terms of policy learning, subgoal optimization, and sequential skill planning.

A. Evaluation on Skill Policy Learning

To evaluate the skill policies and their corresponding value functions, we initially construct a skill library using different skill learning methods. These methods include TTPI and two state-of-the-art RL methods: Soft Actor-Critic (SAC) [10] and Proximal Policy Optimization (PPO) [28].

The skill library encompasses five manipulation skills: pushing, pivoting, pulling, picking, and placing, each characterized by unique state and action spaces. Specifically, pushing, pivoting and pulling are planar manipulation primitives.

1) **Push:** The state is characterized by $(\mathbf{p}_o, \theta_o, \mathbf{p}_r, f_c)$, while the action is denoted by (\mathbf{v}_r, f_n) . Here, $(\mathbf{p}_o, \theta_o) \in \text{SE}(2)$ denotes the object’s pose in the world frame. \mathbf{p}_r and \mathbf{v}_r represent the position and velocity of the robot end-effector in the object frame. We assume that the object has a rectangular shape (common in industry), where $f_c \in 0, 1, 2, 3$ represents the current contact surface and f_n denotes the next contact surface. Thus, the system

encompasses a total of 6 states and 3 control variables, comprising both continuous and discrete variables.

- 2) **Pivot:** The pivoting domain features a goal-augmented state $(\beta, \tilde{\beta})$, where β is the current rotation angle of the object in the gravity plane, and $\tilde{\beta}$ is the desired angle. The control input is $\dot{\beta}$, which denotes the angular velocity of the robot end-effector.
- 3) **Pull:** The state is defined as $(\mathbf{p}_o, \theta_o) \in \text{SE}(2)$, representing the object’s position and orientation in a planar plane. The control input is $(\dot{\mathbf{p}}_o, \dot{\theta}_o)$, denoting the translational and angular velocities of the robot end-effector.
- 4) **Pick:** In this domain, the state is defined as $(x, y, z, \alpha, \beta, \theta) \in \text{SE}(3)$, representing the pose of the robot end-effector, while the control input is the Cartesian velocity of the end-effector, $(\dot{x}, \dot{y}, \dot{z}, \dot{\alpha}, \dot{\beta}, \dot{\theta}) \in \text{SE}(3)$. Without loss of generality, we assume that the object to be picked is located at $(0, 0, 0, 0, 0, 0)$.
- 5) **Place:** This domain shares the same state and action spaces as the **Pick** domain.

We define the general reward for skill learning as:

$$r = -1 \times (c_p + \rho \times c_o + 0.01 \times c_a + 0.1 \times c_f), \quad (9)$$

with

$$\begin{aligned} c_p &= \|\mathbf{x}_p - \mathbf{x}_p^{\text{des}}\|/l_p, & c_o &= \|\mathbf{x}_o - \mathbf{x}_o^{\text{des}}\|/l_o, \\ c_a &= \|\mathbf{u}\|, & c_f &= 1 - \delta(f_c - f_n), \end{aligned} \quad (10)$$

where \mathbf{x}_p and \mathbf{x}_o represent the current position and orientation of the system in each skill domain, while $\mathbf{x}_p^{\text{des}}$ and $\mathbf{x}_o^{\text{des}}$ denote the target position and orientation. We set the state space to be within $[-0.5m, 0.5m]$ for position elements, and $[-\pi, \pi]$ for orientation elements. l_p and l_o are therefore set to 0.5 and π respectively to normalize the position error and orientation error. ρ is a hyperparameter used to balance position and orientation errors. Due to the underactuated nature of pushing dynamics, we set $\rho = 0.5$ in practice. For other skills, ρ is set to 1. \mathbf{u} represents the control inputs of each skill domain. c_f is a specialized term unique to the pushing domain, used to penalize face switching during pushing. Here, $\delta(f_c - f_n)$ returns 1 when $f_c = f_n$ (indicating no face switching), otherwise, it returns 0.

Notably, the skill policies learned for push, pull, pick and place are in a regulated manner, with the target state \mathbf{x}^{des} set as $\mathbf{0}$. In terms of pivoting, the reward function does not include a positional term. In the orientation term, \mathbf{x}_o is defined as β , and $\mathbf{x}_o^{\text{des}}$ is defined as $\tilde{\beta}$. Given new initial and target states in the long-horizon domain, along with the skill operator a_k , the skill-specific state can be computed as $\mathbf{x}_{k_0} = \Gamma_k(\bar{\mathbf{x}}_{k_0}, \bar{\mathbf{x}}_{k_T} | a_k)$ (as shown in Eq. (7)). For push, pull, pick and place, the domain mapping function Γ_k is defined as:

$$\Gamma_k(\bar{\mathbf{x}}_{k_0}, \bar{\mathbf{x}}_{k_T}) = \varphi_k(\bar{\mathbf{x}}_{k_0}) - \varphi_k(\bar{\mathbf{x}}_{k_T}), \quad (11)$$

where φ_k is a dimension reduction function that selects skill-specific dimensions from the long-horizon state. In the case of pivot, the skill is acquired through goal-augmented policy learning, achieved by augmenting the state as $(\beta, \tilde{\beta})$. This

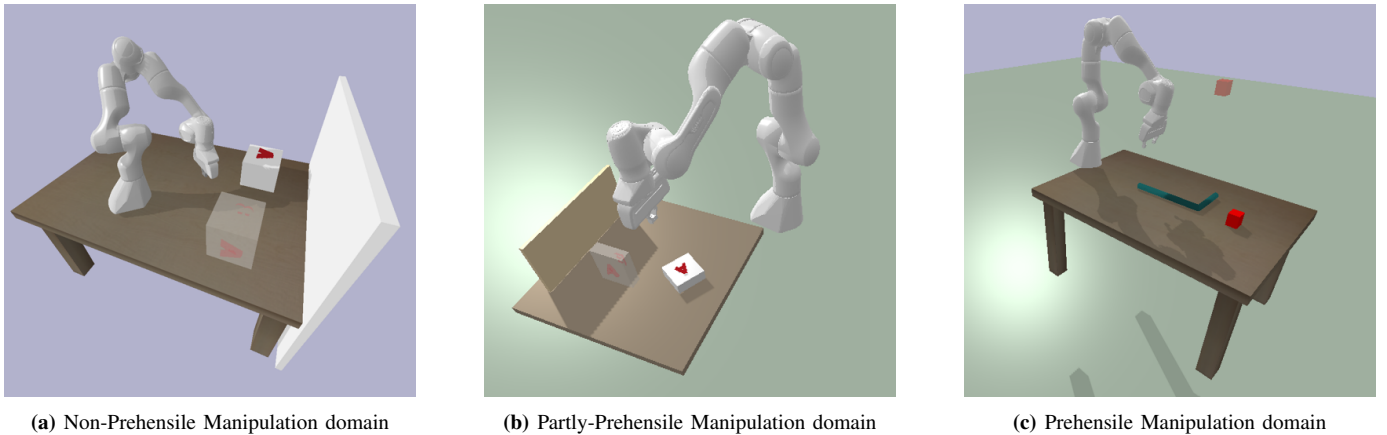


Fig. 2: Three sequential manipulation domains, including both prehensile and non-prehensile manipulation primitives. The transparent object represents the final target configuration in each domain.

augmented state is the concatenation of the current rotation angle $\varphi_{\text{pivot}}(\bar{\mathbf{x}}_{k_0})$ with the desired angle $\varphi_{\text{pivot}}(\bar{\mathbf{x}}_{k_T})$. Γ_{pivot} is therefore defined as:

$$\Gamma_{\text{pivot}}(\bar{\mathbf{x}}_{k_0}, \bar{\mathbf{x}}_{k_T}) = \text{Concat}(\varphi_{\text{pivot}}(\bar{\mathbf{x}}_{k_0}), \varphi_{\text{pivot}}(\bar{\mathbf{x}}_{k_T})). \quad (12)$$

The obtained \mathbf{x}_{k_0} can then be directly fed into V^{π_k} without the need to retrain new skill policies and value functions.

As for another domain mapping function $\Phi_k : \mathcal{X}_k \rightarrow \bar{\mathcal{X}}$ which maps skill-specific state to long-horizon state, it is defined as a function that updates the value of skill-specific dimensions in the long-horizon state based on the skill-specific state.

Notably, considering that SAC and PPO are not able to handle hybrid action space, we exclude f_c and f_n for a fair comparison of skill learning performance in this subsection. For both PPO and SAC, our primary implementation relies on Stable-Baselines3 [25], utilizing a Multilayer Perceptron (MLP) architecture with dimensions of 32×32 as the policy network. We set the discount factor to 0.99 and the learning rate to 0.001, while configuring the task horizon to 10^4 . In TTPI, we establish the accuracy threshold for TT-cross as $\epsilon = 10^{-3}$ and set the maximum rank r_{max} to 10^2 .

The obtained policies are then evaluated by comparing the success rates across 1000 initial states in skill domains. We define a successful task as achieving a position error of less than 0.03cm and an orientation error of less than 15° . Table I reveals that all the policies can nearly reach the final targets, demonstrating the proficiency of the learned policies in accomplishing individual manipulation tasks. However, rather than focusing solely on having skills for each specific task, we are also concerned with finding the optimal trajectory with the maximum cumulative reward for the sequenced skills. To achieve this, an accurate value function is required to inform which state in the intersection space between two adjacent skills leads to the maximum cumulative reward. Therefore, we examine whether the value functions can offer the same guidance as the cumulative reward given two states, shown as *value prediction* in Table I. The metric is to compare whether

$V^{\pi_k}(\mathbf{x}_1) - V^{\pi_k}(\mathbf{x}_2)$ has the same direction as $\mathcal{R}_c^{\pi_k}(\mathbf{x}_1) - \mathcal{R}_c^{\pi_k}(\mathbf{x}_2)$, where $\mathcal{R}_c^{\pi_k}(\mathbf{x})$ is the cumulative reward, defined as

$$\mathcal{R}_c^{\pi_k}(\mathbf{x}) = \sum_{t=0}^{\infty} \gamma^t R_{k_t}(\mathbf{x}_{k_t}, \pi_k(\mathbf{x}_{k_t})), \text{ s.t. } \mathbf{x}_{k_0} = \mathbf{x}. \quad (13)$$

We randomly selected 1000 state pairs in the state domain. The value function of TTPI demonstrates superior prediction performance compared to SAC and PPO. This indicates that value functions in TT format offer better accuracy in approximating the cumulative reward, which can inform which state in the state domain is more dynamically optimal given current policy. This observation aligns with our motivation, emphasizing that typical RL methods approximate the value function primarily within a local space and the policy retrieval is often sub-optimal due to gradient-based optimization. In contrast, TTPI, as an ADP method, aims to approximate the value function across the entire state space. This feature is advantageous in our sequential skill planning framework for identifying the subgoal for each skill, which can be anywhere within the skill-specific state space.

B. Evaluation on Skill Value Optimization

Table I illustrates that the acquired skills are successful in accomplishing each specific manipulation task. However, to achieve a long-horizon task where only the evaluation function Ψ of final configuration is given, finding the subgoal for each skill is important. This necessitates a good optimization technique capable of finding the optimal solution based on the given objective function.

Three different long-horizon domains are used to validate our method, involving both non-prehensile and prehensile manipulation primitives, as shown in Fig. 2. The simulation environments are built using PyBullet [5] and the AIRobot library [3].

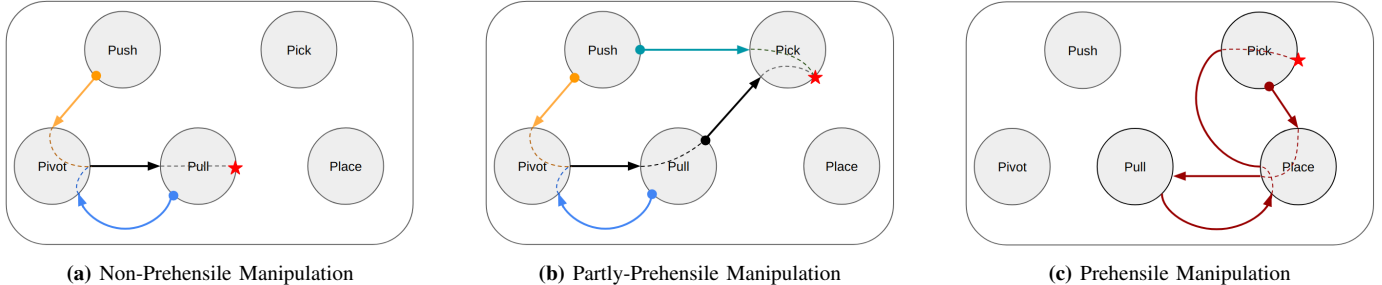
a) Non-Prehensile Manipulation (NPM): As depicted in Fig. 2a, this domain involves objects that cannot be grasped. The objective is to manipulate the box within the 3D world by leveraging multiple non-prehensile planar manipulation

Table I: Comparative Analysis of Skill Policy Performance and Value Function Precision

	TTPI			SAC			PPO		
	success rate	value prediction	time (min)	success rate	value prediction	time (min)	success rate	value prediction	time (min)
pushing	1.0	0.85	5.6	0.83	0.63	36.3	0.95	0.61	44.8
pivoting	1.0	0.94	0.9	1.0	0.51	16.0	1.0	0.68	8.7
pulling	1.0	0.97	1.1	1.0	0.84	16.5	1.0	0.72	10.7
pick/place	1.0	0.93	1.67	0.98	0.64	33.6	1.0	0.66	24.6

Table II: Comparison of Computation Error and Time for Skill Value Optimization

	TTGO			Shooting		CEM-MD	
	error	approximation time (s)	inference time (s)	error	time (s)	error	time (s)
NPM	0.03 ± 0.00	0.48 ± 0.21	0.003 ± 0.00	0.35 ± 0.01	0.01 ± 0.00	0.02 ± 0.01	0.06 ± 0.01
PPM	0.12 ± 0.01	3.90 ± 0.22	0.004 ± 0.00	0.59 ± 0.01	0.02 ± 0.00	0.05 ± 0.01	0.09 ± 0.01
PM	0.12 ± 0.01	20.28 ± 5.23	0.01 ± 0.01	0.75 ± 0.27	0.01 ± 0.00	0.03 ± 0.01	0.10 ± 0.01

**Fig. 3:** The action skeletons obtained by LSP for three domains. The dot point denotes the start of the skill sequence. Each color represents one solution, with black lines indicating the common shared tunnel. The red star illustrates the end of the skill skeleton.**Table III:** Comparison of Computation Time and Solution Quality for Sequential Skill Planning

	Computation Time (s) [w/ sym. goal]		Computation Time (s) [w/o sym. goal]		Cumulative Reward		Sequence Length
	LSP	STAP	LSP	STAP	LSP	STAP	
NPM	0.14 ± 0.23	0.06 ± 0.03	0.27 ± 0.15	NA	3.0 ± 0	2.4 ± 1.57	3.0 ± 0
PPM	0.17 ± 0.1	0.08 ± 0.02	0.22 ± 0.13	NA	2.9 ± 0.7	1.88 ± 1.01	2.9 ± 0.7
PM	0.25 ± 0.15	0.21 ± 0.02	0.41 ± 0.02	NA	5.0 ± 0	3.28 ± 0.62	5.0 ± 0

primitives and establishing contacts with the surroundings to achieve the final 6D pose. This task can also be seen as a special case of in-hand manipulation, where the robot and the wall act as active and passive "fingers", respectively.

b) Partly-Prehensile Manipulation (PPM): As shown in Fig. 2b, this domain involves objects that can only be grasped in specific directions. The goal in this domain is to manipulate the 6D pose of the cube, including the z direction. Therefore, the robot must strategically decide how to pick up the object in the end. This domain requires the robot to reason about physical contact and object geometry to unify both prehensile and non-prehensile primitives.

c) Prehensile Manipulation (PM): As illustrated in Fig. 2c, this domain involves objects that can be directly grasped. The goal is to alter the 6D pose of a block placed on the table, beyond the robot's reachability. To achieve this, the robot must pick up the block in the end. This requires the robot to figure out using a hook to extend the kinematic chain and pull the block back into the reachability region, and then grasp it. However, the way in which the robot picks up the hook will affect whether the block can be reached, and where to pull will also affect the entire trajectory and the total energy cost.

In this task, we want to show that our method can find an optimal trajectory which has the minimum energy cost, while successfully finishing the task.

We therefore define the evaluation function Ψ of the final configuration \bar{x}_{K_T} as follows:

$$\Psi(\bar{x}_{K_T}) = \lambda \|\bar{x}_{K_T} - \bar{x}_T\|, \quad (14)$$

where \bar{x}_T is the target configuration in each domain, and $\lambda = 10^2$. The initial configuration \bar{x}_0 and target configuration \bar{x}_T are randomly sampled from the long-horizon domain $\bar{\mathcal{X}}$, which is the configuration space of the entire environment. In NPM and PPM, $\bar{\mathcal{X}} = \mathcal{S}_R \times \mathcal{S}_O$, while in PM, $\bar{\mathcal{X}} = \mathcal{S}_R \times \mathcal{S}_O \times \mathcal{S}_T$. Here, $\mathcal{S}_O = SE(3)$ and $\mathcal{S}_R = SE(3)$ denote the pose of the object and robot end-effector, respectively, while $\mathcal{S}_T = SE(2)$ represents the pose of a tool positioned on the table.

It is worth noting that both PPM and NPM tasks involve optimizing both discrete and continuous variables, akin to mixed-integer programming. This complexity presents significant challenges in optimization, especially considering that the objective functions are arbitrary without any structure, depending on the value functions of skills. To this end, we compare CEM-MD with some other techniques capable of

handling mixed-integer variables, including TTGO [30] and the random shooting method. TTGO is an optimization technique specifically designed for functions in tensor train format, enabling it to find the optimal solution extremely fast once the TT model is provided. Random shooting is a naive technique that randomly samples variables from the domain and returns the one with the highest objective score. We apply these three optimization techniques across the three domains, conducting 10 random initializations for each domain. The maximum number of iterations for CEM-MD is set to $H = 300$, with an early stopping criterion of 10^{-3} to terminate the while loop if the objective value no longer decreases. The population size is set to $C = 1000$, and the elite fraction is set to $p = 0.3$. The number of categories K_c is set to 4. The initial parameters μ , Σ , and \mathbf{p} are computed using samples collected from a uniform distribution in the first iteration. The results are shown in Table II, with the computation error defined as the L2 norm between the final state \bar{x}_{K_T} and the target configuration \bar{x}_T .

Given an arbitrary function, TTGO needs to approximate it first using TT-cross and then optimize over the approximated TT model, corresponding to the approximation stage and inference stage, separately. Table II shows that the approximation stage usually takes a longer time, but once the TT model has been obtained, the inference stage will be very fast. This suits problems with static objective functions quite well [30]. However, in this work, if the high-level skill skeleton $a_{1:K}$ changes, the resulting objective function will change as well. This requires a new TT approximation, leading to expensive computation. Moreover, TTGO requires discretization on each dimension, resulting in a sub-optimal solution if no fine-tuning stage is used afterward. Meanwhile, the shooting method is much faster but can obtain poor solutions because of the random distribution. CEM-MD takes a bit longer time compared to random shooting but obtains good solutions by updating the mixed continuous and discrete distribution. The total computation time for these three domains is less than 0.1s, which is still fast enough in terms of long-horizon planning.

C. Evaluation on Overall Performance of LSP

We then run the complete LSP framework for these three domains to assess the overall performance. In MCTS, the exploration parameter C_E is set to 3, allowing for efficient tree search while exploring different skeleton solutions. The feedback reward \bar{r} is defined as a binary variable 0/1, determined by whether the geometric target is achieved. The maximum iteration is set to $\tilde{H} = 100$, with an early stopping criterion that terminates the while loop upon finding enough \tilde{N}_s solutions. We set $\tilde{N}_s = 5$ to explore multiple solutions. Note that the \tilde{N}_s solutions can be identical, depending on MCTS and the total number of feasible solutions. We derive 13 skill operators from the five skill policies existing in the skill library. The logic defining the preconditions and effects of these operators is presented in Table IV. Each operator's name, preceding the underscore, denotes the skill policy it utilizes. Elements highlighted in yellow are common across

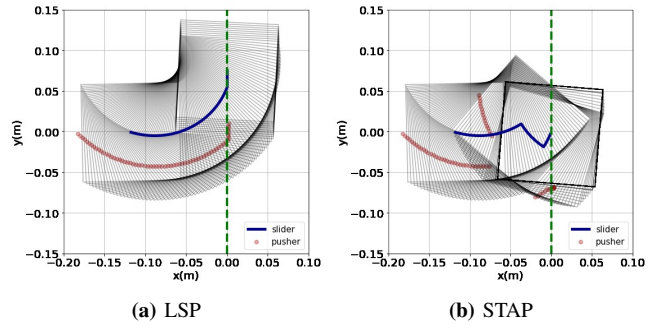


Fig. 4: The pushing subtask obtained for the PPM domain. Both 4a and 4b have the same initialization configuration. The objective is to employ the robot end effector (pusher) to move the slider to the green line ($x = 0$), representing the edge of the table. LSP provides a solution with the highest value in the space, requiring less control effort, while STAP outputs one that involves multiple face switches.

both NPM and PPM domains, those in gray are specific to the PPM domain, and those in green are specific to the PM domain. Throughout the table, we use the following symbols: o for an object, t for a tool, and r for a robot arm. Table V displays the initial symbolic state s_0 used in each domain, along with the symbolic goals used for comparison.

Given the same initial state s_0 and \bar{x}_0 in each domain, we randomly sample 10 different target configurations \bar{x}_T that require multi-step manipulation. Fig. 3 shows that LSP can actively find multiple solutions. For the Non-Prehensile domain, two solutions found by LSP are push-pivot-pull and pull-pivot-pull. For the Partly-Prehensile domain, four different skeletons are found: push-pick, pull-pick, push-pivot-pull-pick and pull-pivot-pull-pick. For the Prehensile domain, the solved skill skeleton is pick-place-pull-place-pick.

We then compare our method with another state-of-the-art sampling-based sequential skill planning method called STAP [2]. STAP focuses on constraints satisfaction. It requires an explicit symbolic goal for high-level task planning, followed by feasible solutions sampling. To ensure a fair comparison, we use MCTS with an explicit symbolic goal as the task planner in STAP and then employ CEM-MD for feasibility checking given the skill skeleton. Table III illustrates the time required to find one solution and the solution quality between LSP and STAP. We can observe that LSP does not require a symbolic target goal, whereas STAP relies on it. STAP can find the solution faster than LSP. The reason is that STAP focuses only on finding the feasible solution, while LSP aims to provide (global) optimality over the full logic-geometric path. This aligns with the comparison of cumulative rewards. Note that the cumulative reward of each skill in the sequence is normalized by the highest value in the corresponding value function. We can observe that the trajectory found by LSP leads to a higher cumulative reward compared with STAP, indicating better optimality. Moreover, in the PPM domain, the skill sequence varies with different lengths, showing that

Table IV: The Specification of Skill Operators, Preconditions and Effects. In the effects column, only the states that are changed after skill execution are listed. The remaining symbolic states in the preconditions will be directly inherited by the effects.

Skill Operator	Preconditions	Effects
push_wall	$(\neg (\text{AtEdge } o)) \wedge (\neg (\text{AtWall } o)) \wedge (\neg (\text{AfterFlip } o)) \wedge (\text{onTable } o)$	$(\text{AtWall } o)$
pivot	$(\text{AtWall } o)$	$(\text{AtWall } o) \wedge (\text{AfterFlip } o)$
pull_wall	$(\neg (\text{AtEdge } o)) \wedge (\neg (\text{AtWall } o)) \wedge (\neg (\text{AfterFlip } o)) \wedge (\text{onTable } o)$	$(\text{AtWall } o)$
pull_center	$(\text{AtWall } o) \wedge (\text{AfterFlip } o) \wedge (\text{onTable } o)$	$(\neg (\text{AtWall } o))$
pull_edge	$(\neg (\text{AtEdge } o)) \wedge (\neg (\text{AtWall } o)) \wedge (\neg (\text{AfterFlip } o)) \wedge (\text{onTable } o)$	$(\text{AtEdge } o)$
pick_edge	$(\text{AtEdge } o) \wedge (\text{PartGraspable } o) \wedge (\text{HandEmpty } r)$	$(\text{InHand } o) \wedge (\neg (\text{HandEmpty } r))$
pick_center	$(\neg (\text{AtWall } o)) \wedge (\text{AfterFlip } o) \wedge (\text{PartGraspable } o) \wedge (\text{HandEmpty } r)$	$(\text{InHand } o) \wedge (\neg (\text{HandEmpty } o))$
push_edge	$(\neg (\text{AtEdge } o)) \wedge (\neg (\text{AtWall } o)) \wedge (\neg (\text{AfterFlip } o)) \wedge (\text{onTable } o)$	$(\text{AtEdge } o)$
pick_tool	$(\neg (\text{ReadyPull } t \ o)) \wedge (\text{Graspable } o) \wedge (\text{Reachable } t) \wedge (\text{HandEmpty } r)$	$(\neg (\text{HandEmpty } r))$
pull_tool	$(\text{ReadyPull } t \ o) \wedge (\neg (\text{Reachable } o)) \wedge (\neg (\text{HandEmpty } r)) \wedge (\text{onTable } o)$	$(\text{Reachable } o)$
place_toolmove	$(\neg (\text{HandEmpty } r)) \wedge (\neg (\text{Reachable } o))$	$(\text{ReadyPull } t \ o)$
place_tool	$(\neg (\text{HandEmpty } r))$	$(\text{HandEmpty } r)$
pick_object	$(\text{Reachable } o) \wedge (\text{Graspable } o) \wedge (\text{HandEmpty } r)$	$(\text{InHand } o)$

Table V: Initial Symbolic State and Symbolic Goal of Each Domain. Note that LSP does not need symbolic goals. Such goals are needed by the sampling-based sequential skill planning methods for comparison.

Domain	Initial symbolic state	Symbolic goal
NPM	$(\neg (\text{AtEdge } o)) \wedge (\neg (\text{AtWall } o)) \wedge (\neg (\text{AfterFlip } o)) \wedge (\text{onTable } o)$	$(\text{AfterFlip } o) \wedge (\neg (\text{AtWall } o))$
PPM	$(\neg (\text{AtEdge } o)) \wedge (\neg (\text{AtWall } o)) \wedge (\neg (\text{AfterFlip } o)) \wedge (\text{HandEmpty } r) \wedge (\text{PartGraspable } o) \wedge (\text{onTable } o)$	$(\text{Inhand } o)$
PM	$(\neg (\text{Reachable } o)) \wedge (\text{Graspable } o) \wedge (\text{Reachable } t) \wedge (\text{HandEmpty } r) \wedge (\text{onTable } o)$	$(\text{Inhand } o)$

multiple solutions are found, as depicted in Fig. 3b.

Fig. 4 illustrates a toy example of the solutions found by LSP and STAP for the PPM domain, where the robot needs to push the block (slider) to the edge of the table (depicted as the green line) for grasping from the lateral side. LSP determines the subgoal with the highest cumulative reward by utilizing the value function, which represents the most optimal state considering the system dynamics. In contrast, STAP outputs a feasible state, which can be any configuration on the edge theoretically. Here, we pick up the one closest to the initialization in Euclidean space. However, to reach this target, the pusher exerts more effort, involving two face switches (visible in the discontinuous pusher trajectory in Fig. 4). This also underscores the utility of the value function space for finding the optimal path considering system dynamics, compared with Euclidean state space.

D. Real-robot Experiments

We conducted real-robot experiments in the Non-Prehensile Manipulation domain, employing a 7-axis Franka Emika robot and a RealSense D435 camera. A large box (21cm x 21cm x 16cm) was positioned on a flat plywood surface. The camera tracked the object motion at 30 Hz with ArUco markers, and the policy for each skill was updated at 100 Hz, with low-level controllers (1000 HZ) actuating the robot. We employed different controllers depending on the skill operator. Specifically, for pushing, we utilized the Cartesian velocity controller. For other skills, we utilized Cartesian impedance controller. These components are integrated into the Robot Operating System (ROS), enabling their operation at varying frequencies.

Fig. 5 displays the keyframes of the robot experiments. Given the initial and final configurations, the robot adeptly manipulated the box by utilizing three planar manipulation

primitives, establishing and breaking contact with the surroundings. It is worth noting that real-world contact-rich manipulation is quite challenging due to friction uncertainty and external disturbances [22]. This highlights the importance of introducing skills for online real-time control. Through skill sequencing, we demonstrate that the robot can actively engage with the physical world, accomplishing a much more complex long-horizon task. Additional results are presented in our accompanying video.

VI. CONCLUSION AND FUTURE WORK

In this paper, we introduced an optimization-based approach for sequential skill planning, namely Logic-Skill Programming (LSP). A first-order extension of a mathematical program is formulated to optimize the overall cumulative reward and the performance of the final configuration. The cumulative reward is abstracted as the sum of value function space. To address such problems, we employed Tensor Train to approximate the value functions and leveraged alternations between symbolic search and skill value optimization to find the optimal solutions.

We demonstrated that the value functions in TT format provide a better approximation of cumulative reward compared to state-of-the-art RL methods. Furthermore, the proposed LSP framework can generate multiple skill skeletons and their corresponding subgoal sequences, given only an evaluation function of the final geometric configuration. We validated this approach across three manipulation domains, highlighting its robust performance in sequencing both prehensile and non-prehensile manipulation primitives.

In this work, the skill learning method, TTPI, assumes a low-rank structure of the value functions for approximation.

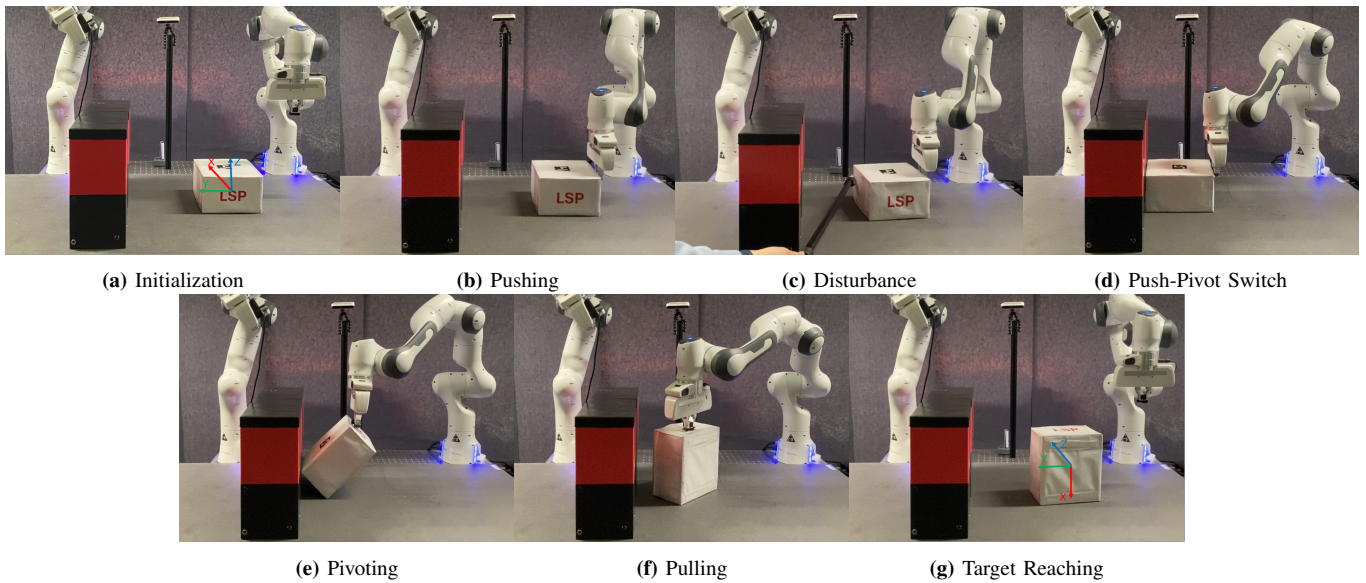


Fig. 5: Non-prehensile manipulation domain task. The system is initialized as (a), and the objective is to manipulate the box to achieve the target configuration as (g). The first stage involves pushing the box towards the wall with a 90° rotation. Additionally, we apply an external disturbance to test the skill policy (c). After the pushing stage, the robot switches to the pivoting skill (d, e), followed by pulling (f), until reaching the final geometric configuration.

This works well for skill domains with low-to-medium dimensionality but struggles with image-based policy learning. To address this issue, combining neural networks with TT decomposition could be an interesting direction.

Moreover, the proposed LSP formulation relies on powerful skills to address uncertainty and disturbances. This assumption might not scale well to complicated scenarios, such as pushing with moving obstacles. To address this, a promising approach could be combining global motion planners, such as Rapidly-Exploring Random Tree (RRT)[14], with the learned skill policies as local controllers.

Furthermore, while Large Language Models (LLMs) have demonstrated impressive results in task planning for robotics, there is still a gap between high-level task planning and low-level control due to the absence of geometric planning. We believe that our method could address this issue by sequencing multiple task-agnostic policies from a skill library. Given a skill skeleton provided by LLMs, our method could assess its feasibility and subsequently returns the optimal subgoal sequence. This subgoal sequence could then be utilized by the sequenced skill policies to actuate the robot in the real world.

ACKNOWLEDGMENTS

This work was supported by the China Scholarship Council (grant No.202106230104), and by the SWITCH project (<https://switch-project.github.io/>), funded by the Swiss National Science Foundation. We would like to thank Yan Zhang for providing helpful feedback on this manuscript, and Jiacheng Qiu for suggestions about the implementation of RL baselines.

REFERENCES

- [1] Constructions Aeronautiques, Adele Howe, Craig Knoblock, ISI Drew McDermott, Ashwin Ram, Manuela Veloso, Daniel Weld, David Wilkins SRI, Anthony Barrett, Dave Christianson, et al. Pddl— the planning domain definition language. *Technical Report, Tech. Rep.*, 1998.
- [2] Christopher Agia, Toki Migimatsu, Jiajun Wu, and Jeanette Bohg. Stap: Sequencing task-agnostic policies. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7951–7958. IEEE, 2023.
- [3] Tao Chen, Anthony Simeonov, and Pulkit Agrawal. AIRobot. <https://github.com/Improbable-AI/airobot>, 2019.
- [4] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.
- [5] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. <https://pybullet.org>, 2016–2019.
- [6] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinfeld. A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67, 2005.
- [7] Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on Robot Learning*, pages 158–168. PMLR, 2022.
- [8] Caelan Reed Garrett, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Pddlstream: Integrating symbolic planners and blackbox samplers via optimistic adaptive planning. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 30, pages 440–448, 2020.
- [9] Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, Beomjoon Kim, Tom Silver, Leslie Pack Kaelbling, and

- Tomás Lozano-Pérez. Integrated task and motion planning. *Annual review of control, robotics, and autonomous systems*, 4:265–293, 2021.
- [10] Tuomas Harnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.
- [11] Valentin N Hartmann, Andreas Orthey, Danny Driess, Ozgur S Oguz, and Marc Toussaint. Long-horizon multi-robot rearrangement planning for construction assembly. *IEEE Transactions on Robotics*, 39(1):239–252, 2022.
- [12] Francois R Hogan and Alberto Rodriguez. Reactive planar non-prehensile manipulation with hybrid model predictive control. *International Journal of Robotics Research (IJRR)*, 39(7):755–773, 2020.
- [13] De-An Huang, Danfei Xu, Yuke Zhu, Animesh Garg, Silvio Savarese, Li Fei-Fei, and Juan Carlos Nieves. Continuous relaxation of symbolic planner for one-shot imitation learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2635–2642. IEEE, 2019.
- [14] Steven LaValle. Rapidly-exploring random trees: A new tool for path planning. *Research Report 9811*, 1998.
- [15] Tobia Marcucci and Russ Tedrake. Mixed-integer formulations for optimal control of piecewise-affine systems. In *Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control*, pages 230–239, 2019.
- [16] Tobia Marcucci, Mark Petersen, David von Wrangel, and Russ Tedrake. Motion planning around obstacles with convex optimization. *Science robotics*, 8(84):eadf7843, 2023.
- [17] Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 7(62):eabk2822, 2022.
- [18] Utkarsh Aashu Mishra, Shangjie Xue, Yongxin Chen, and Danfei Xu. Generative skill chaining: Long-horizon skill planning with diffusion models. In *Conference on Robot Learning*, pages 2905–2925. PMLR, 2023.
- [19] João Moura, Theodoros Stouraitis, and Sethu Vijayakumar. Non-prehensile planar manipulation via trajectory optimization with complementarity constraints. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 970–976. IEEE, 2022.
- [20] Ivan Oseledets and Eugene Tyrtyshnikov. TT-cross approximation for multidimensional arrays. *Linear Algebra and its Applications*, 432(1):70–88, 2010.
- [21] Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.
- [22] Tao Pang, HJ Terry Suh, Lujie Yang, and Russ Tedrake. Global planning for contact-rich manipulation via local smoothing of quasi-dynamic contact models. *IEEE Transactions on Robotics*, 2023.
- [23] Michael Posa, Cecilia Cantu, and Russ Tedrake. A direct method for trajectory optimization of rigid bodies through contact. *International Journal of Robotics Research (IJRR)*, 33(1):69–81, 2014.
- [24] Carlos Quintero-Pena, Zachary Kingston, Tianyang Pan, Rahul Shome, Anastasios Kyriakidis, and Lydia E Kavraki. Optimal Grasps and Placements for Task and Motion Planning in Clutter. In *Proc. IEEE Intl Conf. on Robotics and Automation (ICRA)*, pages 3707–3713, 2023.
- [25] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL <http://jmlr.org/papers/v22/20-1364.html>.
- [26] Reuven Rubinfeld. The cross-entropy method for combinatorial and continuous optimization. *Methodology and computing in applied probability*, 1:127–190, 1999.
- [27] Dmitry V. Savostyanov and Ivan V. Oseledets. Fast adaptive interpolation of multi-dimensional arrays in tensor train format. *The 2011 International Workshop on Multidimensional (nD) Systems*, pages 1–8, 2011.
- [28] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [29] Dhruv Shah, Peng Xu, Yao Lu, Ted Xiao, Alexander T Toshev, Sergey Levine, et al. Value Function Spaces: Skill-Centric State Abstractions for Long-Horizon Reasoning. In *Proc. Intl Conf. on Learning Representations (ICLR)*, 2021.
- [30] S. Shetty, T. Lembono, T. Löw, and S. Calinon. Tensor Train for Global Optimization Problems in Robotics. *International Journal of Robotics Research (IJRR)*, 43(6):811–839, 2024. doi: 10.1177/02783649231217527.
- [31] S. Shetty, T. Xue, and S. Calinon. Generalized Policy Iteration using Tensor Approximation for Hybrid Control. In *Proc. Intl Conf. on Learning Representations (ICLR)*, 2024.
- [32] Siddharth Srivastava, Eugene Fang, Lorenzo Riano, Rohan Chitnis, Stuart Russell, and Pieter Abbeel. Combined task and motion planning through an extensible planner-independent interface layer. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 639–646. IEEE, 2014.
- [33] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [34] Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.
- [35] Marc Toussaint. Logic-geometric programming: an optimization-based approach to combined task and motion planning. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1930–1936, 2015.

- [36] Marc Toussaint, Kelsey R Allen, Kevin A Smith, and Josh B Tenenbaum. Differentiable Physics and Stable Modes for Tool-Use and Manipulation Planning. In *Proc. of Robotics: Science and Systems (R:SS)*, 2018. *Best Paper Award*.
- [37] Marc Toussaint, Jason Harris, Jung-Su Ha, Danny Driess, and Wolfgang Hönig. Sequence-of-Constraints MPC: Reactive Timing-Optimal Control of Sequential Manipulation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13753–13760. IEEE, 2022.
- [38] Brian C Williams. *Cognitive Robotics: Monte Carlo Tree Search*. Cambridge MA, 2016. MIT OpenCourseWare.
- [39] Bohan Wu, Suraj Nair, Li Fei-Fei, and Chelsea Finn. Example-Driven Model-Based Reinforcement Learning for Solving Long-Horizon Visuomotor Tasks. In *5th Annual Conference on Robot Learning*, 2021.
- [40] Danfei Xu, Ajay Mandlekar, Roberto Martín-Martín, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. Deep affordance foresight: Planning through what can be done in the future. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6206–6213. IEEE, 2021.
- [41] Teng Xue, Hakan Girgin, Teguh Santoso Lembono, and Sylvain Calinon. Demonstration-guided optimal control for long-term non-prehensile planar manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4999–5005. IEEE, 2023.
- [42] Teng Xue, Amirreza Razmjoo, and Sylvain Calinon. D-LGP: Dynamic Logic-Geometric Program for Reactive Task and Motion Planning. In *Proc. IEEE Intl Conf. on Robotics and Automation (ICRA)*, pages 14888–14894, 2024.
- [43] Zhutian Yang, Caelan R Garrett, Tomás Lozano-Pérez, Leslie Kaelbling, and Dieter Fox. Sequence-based plan feasibility prediction for efficient task and motion planning. In *Proc. Robotics: Science and Systems (R:SS)*, Daegu, Republic of Korea, July 2023.