

Deliberate Practice: Learning Robot Skills under a Budget

Vats, Shivam; Harithas, Sudarshan; Akbulut, Mete; Raghunathan, Arvind; Konidaris, George

TR2026-147 October 03, 2026

Abstract

We consider the problem of autonomously learning robot skills under a limited practice budget for long-horizon tasks. We propose an active skill learning algorithm, Deliberate Practice (DP), that computes a provably budget-optimal allocation—practicing skills that maximize expected cumulative reward while being learnable within the budget. DP estimates both the time needed to master skills and the cumulative reward of the task plans that the skills unlock. Computing a budget-optimal allocation is challenging as it requires reasoning about combinatorially many skill plans over a large practice budget. Our key contribution is a bilinear program that can compute this exactly using off-the-shelf solvers. Through simulated and real-world experiments on long-horizon manipulation tasks, we show that our approach allows robots to optimally use limited practice time to acquire useful policies and improve long-horizon planning.

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) Workshop

© 2026 MERL. This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Deliberate Practice: Learning Robot Skills under a Budget

Shivam Vats¹, Sudarshan Harithas¹, Mete Tuluhan Akbulut¹, Arvind Raghunathan², George Konidaris¹
¹Brown University, ²Mitsubishi Electric Research Laboratories

Abstract—We consider the problem of autonomously learning robot skills under a limited practice budget for long-horizon tasks. We propose an active skill learning algorithm, *Deliberate Practice (DP)*, that computes a provably *budget-optimal* allocation—practicing skills that maximize expected cumulative reward while being learnable within the budget. DP estimates both the time needed to master skills and the cumulative reward of the task plans that the skills unlock. Computing a budget-optimal allocation is challenging as it requires reasoning about combinatorially many skill plans over a large practice budget. Our key contribution is a bilinear program that can compute this exactly using off-the-shelf solvers. Through simulated and real-world experiments on long-horizon manipulation tasks, we show that our approach allows robots to optimally use limited practice time to acquire useful policies and improve long-horizon planning.

I. INTRODUCTION

Recent progress in large-scale robot learning has produced general-purpose robot policies that can perform diverse skills and transfer to novel environments [1], [2], [3]. When combined with classical planning methods, such as Task and Motion Planning (TAMP) [4], [5], [6] and search [7], [8], these policies can allow robots to perform long-horizon tasks in homes and factories, charting a path for more autonomous robots. However, pretraining alone does not cover all the possible scenarios a robot may face in the real-world, often leading to unreliable task execution. Therefore, a promising solution is to allow robots to learn from experience through autonomous practice [9], [10], [11]. However, deployment offers limited time for practice, requiring robots to prioritize which skills to improve.

Our key insight is that downtime is often known in advance, and robots should leverage this information to adapt their learning process to the resulting practice budget. Intuitively, a large practice budget should encourage robots to practice more rewarding skills even if they are hard, while a small budget should lead to conservative learning behaviour. We formalize the problem of autonomous skill practice during deployment as budgeted skill learning, and propose an active learning algorithm, *Deliberate Practice (DP)*, that optimally uses the budget to maximize long-horizon task reward. We consider a setting in which a robot is given high-level skill specifications, including preconditions, termination conditions, and effects. For example, in the breakfast domain in Section VI, the `StartToaster` skill requires moving above the toaster lever to press it and has the effect of turning on the toaster. The robot must learn parameterized control policies that ground skills into low-level actions through practice in the environment.

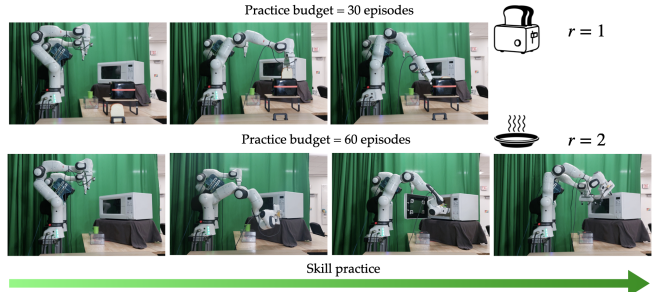


Fig. 1. *Deliberate Practice* allows robots to intelligently use a limited practice budget to actively learn skills to maximize long-horizon task reward. (top) Under a small budget, the robot practices an easy task plan that toasts bread and provides reward 1. (bottom) Under a larger budget, it practices a more difficult but higher-reward task plan that microwaves oatmeal and provides reward 2.

At decision time, a task planner computes a plan consisting of skills, e.g., (`Pick(bread)`, `Place(bread, toaster)`, `Start(toaster)`) to toast bread. Figure 1 illustrates how our approach allows the robot to decide which skills to practice to make breakfast. The robot can either practice one skill to toast bread or use additional practice time to learn two skills required to microwave oatmeal and achieve a higher reward. As shown in Section VI, with a budget of 30 episodes, DP correctly estimates that only one skill can be learned reliably and hence practices to toast bread. Under a budget of 60 episodes, it instead practices to microwave oatmeal to take advantage of the additional practice time.

Deliberate Practice first estimates the difficulty of learning each skill as a function of the practice budget. It then computes a budget allocation that provably maximizes expected planning performance, and practices its skills accordingly. We show in Section IV that computing an optimal allocation requires solving a challenging bilevel optimization problem. Existing active learning algorithms approximate this problem greedily, leading to myopic and suboptimal learning. By contrast, we derive an exact single-level reformulation of the bilevel problem using the linear programming (LP) formulation of Markov decision processes. This allows us to compute the optimal budget allocation using off-the-shelf solvers. Through experiments in simulated and real-world long-horizon manipulation tasks, we demonstrate that *Deliberate Practice* is uniquely budget-aware and enables robots to learn effectively under deployment constraints.

II. RELATED WORK

Active Skill Learning. Deliberate practice is a widely studied mechanism underlying expert human performance [12], characterized by structured, goal-directed training that allocates effort toward the most limiting components of a skill. Prior work includes methods for actively learning parameterized robot skills from practice [13], human demonstrations [14], and across multiple tasks [15], [16], [17]. However, these approaches focus on efficiently acquiring individual skills without considering the sequential nature of long-horizon problems. By contrast, our approach studies active skill learning embedded in a complex sequential decision process under explicit budget constraints.

Task and Motion Planning (TAMP). TAMP [4], [18], [5] addresses robotic planning problems that require joint reasoning over discrete task structure and continuous motion feasibility. Recent TAMP approaches leverage machine learning to improve scalability and performance, for example, by learning heuristics to guide symbolic search [19], or by learning symbolic operators and task representations [20], [21]. Our work is closely related to approaches that *plan to practice* skills for autonomous improvement in long-horizon tasks [22], [23]. However, these methods rely on greedy practice strategies and often allocate practice time suboptimally. By contrast, our method computes a budget-optimal allocation of practice time.

III. BACKGROUND

We model the environment as a Markov decision process (MDP) with continuous state and action spaces, defined by the tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the low-level state space, $\mathcal{A}$ is the low-level robot action space, $P(s'|s, a)$ is the transition function, $r(s, a)$ is the reward function, γ is the discount factor. We assume an abstract MDP $\mathcal{M}_h := (\mathcal{S}_h, \mathcal{A}_h, P_h, r_h, \gamma)$, where a state abstraction function typically $\alpha : \mathcal{S} \rightarrow \mathcal{S}_h$ projects raw sensory observations to abstract states.

Abstract Dynamics. We assume that the robot is provided with a library of skills $\mathcal{U}$, which are defined using an extension of the options framework, consisting of the following components: a) *Precondition* (PRE) defines the states in which a skill can be initiated, b) *Effect* (EFF) describes the outcome of executing a skill successfully, c) *Control policy* (π) computes the low-level actions to be executed by the robot, d) *Termination condition* (β) specifies when a skill should terminate, e) *Competence* (p) is the probability that the skill will successfully achieve the desired effect. The abstract action space $\mathcal{A}_h$ is defined by skill library $\mathcal{U}$ that the robot can use to transition between abstract states. Skill effect and competence together define the high-level transition function P_h : $P_h(s'|s, a) = p$, if s' corresponds to the effect of skill $u \in \mathcal{U}$ when executed in state s , otherwise transitioning to an absorbing failure state. We use a task planner to solve this abstract MDP $\mathcal{M}_h$ to compute a sequence of skills with the correct parameters to reach the goal, e.g., $(\text{Open}(\text{top-drawer}), \text{Pick}(\text{toy}), \text{Place}(\text{toy}, \text{top-drawer}))$. The skills then ground

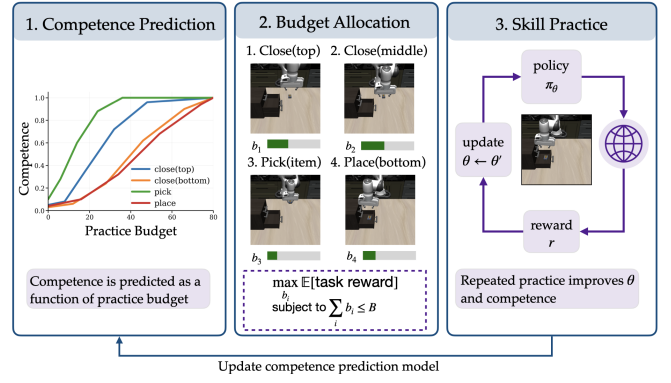


Fig. 2. **Overview:** Our approach predicts skill competence as a function of practice budget, computes a budget allocation for practice, and practices skills by interacting with the environment.

this abstract task plan by computing feasible robot trajectories.

We also define skill-specific reward functions to learn the policy parameters through practice.

IV. BUDGETED SKILL LEARNING FOR PLANNING

Given the high-level descriptions of a library of skills $\mathcal{U}$, consisting of their preconditions, termination conditions, and effects, the robot must learn the policy parameters of its skills to improve overall planning performance under a limited practice budget B , e.g., maximum number of trials. We adopt the standard episodic reinforcement learning setting without access to expert demonstrations. Our goal is to compute a budget allocation b_u across all skills $\mathcal{U}$ to maximize the expected planning performance. Formally, this can be expressed as the following optimization problem:

$$\begin{aligned} & \max_b \sum_{s \in \mathcal{S}} e_s v_s \\ & \text{subject to} \\ & \sum_{u \in \mathcal{U}} b_u \leq B, \\ & v_s \in \text{SolveMDP}(\mathcal{M}_h, b), \end{aligned} \quad (1)$$

where e_s is the initial state distribution, and v_s is the state value function. This is a bilevel optimization problem where the outer problem computes a budget allocation b for practicing policies, and the inner problem solves the resulting task MDP $\mathcal{M}_h$ to compute the expected planning performance after practicing with budget allocation b .

V. DELIBERATE PRACTICE

Our algorithm allocates the practice budget provably optimally by jointly reasoning about all robot skills across the entire practice budget (Figure 2). We describe the three steps of our approach below:

A. Competence Prediction

The competence prediction function $f_{\text{improv}}(p_u, b_u)$ predicts the expected skill competence after practicing a skill with a competence of p_u for b_u rounds. While our approach supports linear, piece-wise linear and exponential models of competence prediction, empirically a piece-wise linear model provides a good balance between performance and efficiency.

$$f_{\text{improv}}(u, b) = \min(1, p_u + \Delta_u b), \quad (2)$$

where Δ_u is the rate of improvement of a skill u estimated from practice. Δ_u is estimated online as a running average of competence improvement observed after practicing skill u over time: $\Delta_u^t = \epsilon \Delta_u^{t-1} + (1-\epsilon)(p_u^t - p_u^{t-1})$, where $\epsilon \in [0, 1]$ is a smoothing factor.

B. Budget Allocation

Next, we use the competence prediction model to compute the best task plan feasible within the practice budget. Long-horizon manipulation tasks require deliberately practicing multiple skills over many episodes to be reliably solved. As the bilevel optimization problem (Equation 1) shows, this requires jointly reasoning about all skills and the entire practice budget. However, this optimization is notoriously difficult due to its nested structure: the feasible set of the outer problem is defined implicitly by the solution of the inner optimization. This coupling leads to highly nonconvex and often non-smooth objectives, even when both levels are individually convex. We derive an **exact single-level reformulation** using duality, yielding a structured single-level bilinear program that enables efficient optimization using standard nonlinear programming solvers, eliminating the need for nested optimization, and greedy approximations. To our knowledge, this is the *first exact formulation of robot learning on a budget as an optimization problem*.

We first state the bilinear program and then outline the derivation. The lines in black in equation 3 correspond to the standard dual linear program of an MDP [24], and our proposed changes are highlighted in blue. μ_s^a are dual variables that correspond to state-action occupancy, e_s is the initial state distribution, $P_{s's}^a$ is the current transition function, and $\bar{P}_{s's}^a(b)$ is the predicted transition function after practicing with budget allocation b . Surprisingly, we show that practice budget constraints can be directly incorporated into the standard dual linear program.

$$\begin{aligned} & \max_{b, \mu} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r_s^a \mu_s^a \\ & \text{subject to} \\ & \sum_{a \in \mathcal{A}} \mu_s^a - \gamma \sum_{s'} \sum_a \bar{P}_{s's}^a(b) \mu_{s'}^a = e_s, \forall s \\ & \bar{P}_{s's}^a(b) = f_{\text{improv}}(P_{s's}^a, b) \\ & \sum b_{ij} \leq B \\ & \mu_s^a \leq \frac{1}{1-\gamma}, \forall s, a. \end{aligned} \quad (3)$$

This is a *nonsmooth bilinear mathematical program* because of the piece-wise linear model in equation 2 and

bilinear terms $\bar{P}_{s's}^a(b) \mu_{s'}^a$ in the constraints. While being a nonsmooth, nonconvex nonlinear problem, there exist powerful techniques to solve it to global optimality. In particular, *piecewise McCormick envelopes* can be used to develop convex relaxations of the bilinear constraints and are supported by popular solvers, such as Gurobi [25].

C. Skill Practice

Once the robot computes a budget allocation and a corresponding budget-optimal task plan, it constructs a curriculum to practice the selected skills. This requires the robot to plan with its existing skills to first reach the precondition of the skill being practiced [23]. Specifically, the robot sequentially masters each of the skills on the budget-optimal plan computed by *Deliberate Practice* and uses each newly acquired skill to reach and practice subsequent skills.

VI. EXPERIMENTAL EVALUATION

We evaluate our method in three long-horizon table-top manipulation tasks. Our simulated environments are implemented in MuJoCo [26] using LIBERO [27]:

- 1) *Cleanup (simulated)*. A Franka Panda robot must clear a table by placing an object into one of three drawers. The task MDP has 47 abstract states and 10 skills in total, with the longest task plan consisting of 4 skills. The robot is provided abstract specifications of `Pick`, `Place`, `OpenDrawer`, and `CloseDrawer` skills.
- 2) *Cleanup-Multi (simulated)*. This is a challenging multi-object version of the *Cleanup* task in which four objects must be placed into a drawer. The task MDP has 5000 abstract states and 22 skills in total, with the longest task plan consisting of 10 skills.
- 3) *Breakfast (real-robot)*. A Franka Panda robot can toast bread or make hot oatmeal using a microwave to achieve a higher reward. The former requires learning a `StartToaster` skill, while the latter requires learning `OpenMicrowave` and `CloseMicrowave` skills, and hence is harder. These skills require forceful interaction with novel articulated objects and hence must be learned by practicing. The robot has additional pick and place skills for transporting objects using motion planning that need not be learned.

Baselines. We compare *Deliberate Practice (DP)* against the following baselines: (1) *Estimate, Extrapolate, Situate (EES)* estimates the expected improvement of each skill and greedily practices the skill that most improves task performance [23]. (2) *Competence Improvement (CI)* practices the skill with the highest expected competence improvement [28], [29]. (3) *Least Competent First (LCF)* practices the skill with the lowest estimated competence. (4) *Round-robin (RR)* practices all reachable skills uniformly.

Skill Practice. We implement all learnable skills as Cartesian-space impedance controllers [30] parameterized by waypoints and impedance parameters. For each skill, we define a reward function and optimize the corresponding interaction policy parameters using CMA-ES [31]. We use

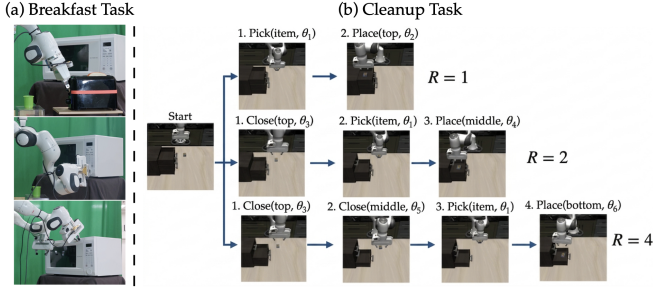


Fig. 3. (a) *Breakfast task* requires learning forceful manipulation skills to interact with novel articulated objects. (b) *Cleanup task* requires the robot to place objects in the top, middle, or bottom drawer with rewards 1, 2 and 4 respectively.

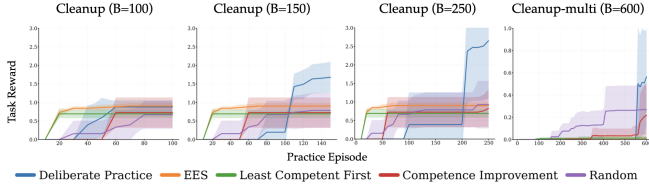


Fig. 4. We evaluate all methods on *Cleanup* with budgets (B) of 100 (low), 150 (medium), and 250 (high) episodes. While they perform similarly under a low budget, DP is significantly better under medium and high budgets. We report the mean and standard deviation averaged over 5 seeds.

CMA-ES because our objective is evaluated through rollouts and is not differentiable with respect to the policy parameters.

A. Experimental Results

DP is budget-aware. Figure 3 shows three possible task plans in the *Cleanup* task that place items in the top, middle and bottom drawers to achieve rewards of 1, 2, and 4, respectively. We compare the effect of different practice budgets on all methods in Figure 4. DP intelligently chooses which skills to learn based on the available budget: conservatively placing items in the top drawer when the budget is 100 episodes, placing items in the middle drawer under a budget of 150 episodes, and maximizing reward by placing items in the bottom drawer under a budget of 250 episodes. By contrast, baselines cannot adapt their behavior to the budget and greedily practice the easiest (and lowest-reward) task plan, irrespective of the budget.

DP significantly outperforms prior active learning methods. As shown in Figure 4, DP and EES perform similarly under a low budget, but DP significantly outperforms all baselines in medium (150 episodes) and high budget (250 episodes) settings. The greedy active learning baselines, EES and CI, reason myopically and, therefore, fail to discover higher-reward plans which require practicing skills that do not immediately improve task performance. This trend continues in the more complex *Cleanup-Multi* task, which requires sequencing 10 skills. Interestingly, *random* practice outperforms EES in this setting because EES lacks an exploration mechanism when the myopic task improvement

measure ΔJ_{task} is zero for all skills. By contrast, DP can look ahead and hence does not suffer from this limitation.

DP for larger problems. We evaluate the scalability of DP in *Cleanup-Multi*, which has 22 skills and 5000 abstract states resulting in a large bilinear program. Despite the size of this problem, DP consistently computes the optimal budget allocation within a maximum solve time of 6 minutes. In larger settings where the solver may not prove global optimality within the allotted time, the same optimization procedure can still return a bounded-suboptimal allocation as Gurobi maintains an incumbent feasible solution and a global bound from convex relaxations of the bilinear constraints, providing an optimality-gap certificate upon early termination.

Real-world Validation. We conduct experiments on a real Franka Panda robot to validate our design choices. To complete the *Breakfast* task, the robot must learn either to toast bread or microwave oatmeal. As shown in Figure 1, toasting provides a reward of 1 and requires practicing only the *StartToaster* skill, while microwaving oatmeal provides a reward of 2, but requires practicing two skills: *OpenMicrowave* and *CloseMicrowave*. The robot is provided with additional *Pick* and *Place* skills that can be used to transport items. We evaluate DP under two practice budgets: 30 episodes and 60 episodes. As shown in Figure 5, DP correctly chooses the conservative toast-bread plan under the smaller budget, but learns to microwave oatmeal under the larger budget, thereby achieving higher task reward.



Fig. 5. Real-world *Breakfast* task.

VII. CONCLUSION

We propose *Deliberate Practice*, an active skill learning algorithm that allocates a limited practice budget to many robot skills to provably maximize long-horizon planning performance. Our algorithm models each skill’s competence as a function of the practice budget, computes an optimal budget allocation, and sequentially practices the skills in the environment. Computing this allocation requires solving a challenging bilevel optimization problem, for which we derive an exact single-level reformulation that can be solved using off-the-shelf solvers. Simulated and real-robot experiments show that our method is uniquely budget-aware and allows robots to better leverage their practice budgets compared to existing methods to master long-horizon manipulation tasks.

Acknowledgments. This work was supported by the Office of Naval Research (ONR) under REPRISM MURI N000142412603 and ONR grant N00014-22-1-2592, as well as by the National Science Foundation (NSF) via grant 1955361. Sudarshan was supported by the Office of Naval Research (ONR) grant N00014-22-1-259.

REFERENCES

- [1] P. Pastor, M. Kalakrishnan, S. Chitta, E. Theodorou, and S. Schaal, "Skill learning and task outcome prediction for manipulation," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 3828–3834.
- [2] Y. Zhu, Z. Jiang, P. Stone, and Y. Zhu, "Learning generalizable manipulation policies with object-centric 3d representations," *arXiv preprint arXiv:2310.14386*, 2023.
- [3] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025.
- [4] S. Cambon, R. Alami, and F. Gravit, "A hybrid approach to intricate motion, manipulation and task planning," *The International Journal of Robotics Research*, vol. 28, no. 1, pp. 104–126, 2009.
- [5] C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez, "Integrated task and motion planning," *Annual review of control, robotics, and autonomous systems*, vol. 4, no. 1, pp. 265–293, 2021.
- [6] B. Hedegaard, Y. Wei, A. Jaafar, S. Tellex, G. Konidaris, and N. Shah, "Beyond task and motion planning: Hierarchical robot planning with general-purpose skills," *arXiv preprint arXiv:2504.17901*, 2025.
- [7] J. Liang, M. Sharma, A. LaGrassa, S. Vats, S. Saxena, and O. Kroemer, "Search-based task planning with learned skill effect models for lifelong robotic manipulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6351–6357.
- [8] I. Mishani, Y. Shaoul, and M. Likhachev, "Mosaic: A skill-centric algorithmic framework for long-horizon manipulation planning," *arXiv preprint arXiv:2504.16738*, 2025.
- [9] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.
- [10] M. P. Deisenroth, G. Neumann, and J. Peters, "A survey on policy search for robotics," *Foundations and Trends® in Robotics*, vol. 2, no. 1-2, pp. 1–142, 2013.
- [11] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 3389–3396.
- [12] K. Anders Ericsson, "Deliberate practice and acquisition of expert performance: a general overview," *Academic emergency medicine*, vol. 15, no. 11, pp. 988–994, 2008.
- [13] B. Da Silva, G. Konidaris, and A. Barto, "Active learning of parameterized skills," in *International Conference on Machine Learning*. PMLR, 2014, pp. 1737–1745.
- [14] S. Chernova and M. Veloso, "Interactive policy learning through confidence-based autonomy," *Journal of Artificial Intelligence Research*, vol. 34, pp. 1–25, 2009.
- [15] A. Fabisch and J. H. Metzen, "Active contextual policy search," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3371–3399, 2014.
- [16] S. Vats, O. Kroemer, and M. Likhachev, "Synergistic scheduling of learning and allocation of tasks in human-robot teams," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2789–2795.
- [17] S. Vats, M. Zhao, P. Callaghan, M. Jia, M. Likhachev, O. Kroemer, and G. Konidaris, "Optimal interactive learning on the job via facility location planning," *Robotics: Science and Systems (RSS)*, 2025.
- [18] S. Srivastava, E. Fang, L. Riano, R. Chitnis, S. Russell, and P. Abbeel, "Combined task and motion planning through an extensible planner-independent interface layer," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 639–646.
- [19] R. Chitnis, D. Hadfield-Menell, A. Gupta, S. Srivastava, E. Groshev, C. Lin, and P. Abbeel, "Guided search for task and motion plans using learned heuristics," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 447–454.
- [20] T. Silver, R. Chitnis, J. Tenenbaum, L. P. Kaelbling, and T. Lozano-Pérez, "Learning symbolic operators for task and motion planning," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3182–3189.
- [21] N. Shah and S. Srivastava, "Hierarchical planning and learning for robots in stochastic settings using zero-shot option invention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 9, 2024, pp. 10 358–10 367.
- [22] S. Vats, M. Likhachev, and O. Kroemer, "Efficient recovery learning using model predictive meta-reasoning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 7258–7264.
- [23] N. Kumar, T. Silver, W. McClinton, L. Zhao, S. Proulx, T. Lozano-Pérez, L. P. Kaelbling, and J. Barry, "Practice makes perfect: Planning to learn skill parameter policies," in *Robotics: Science and Systems (RSS)*, 2024.
- [24] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- [25] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2024. [Online]. Available: <https://www.gurobi.com>
- [26] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.
- [27] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *arXiv preprint arXiv:2306.03310*, 2023.
- [28] A. Stout and A. G. Barto, "Competence progress intrinsic motivation," in *2010 IEEE 9th international conference on development and learning*. IEEE, 2010, pp. 257–262.
- [29] C. Colas, P. Fournier, M. Chetouani, O. Sigaud, and P.-Y. Oudeyer, "Curious: intrinsically motivated modular multi-goal reinforcement learning," in *International conference on machine learning*. PMLR, 2019, pp. 1331–1340.
- [30] N. Hogan, "Impedance control: An approach to manipulation: Part i—theory," *Journal of Dynamic Systems, Measurement, and Control*, vol. 107, no. 1, pp. 1–7, 1985.
- [31] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary computation*, vol. 9, no. 2, pp. 159–195, 2001.