Example-Driven Model-Based Reinforcement Learning for Solving Long-Horizon Visuomotor Tasks

Anonymous Author(s)
Affiliation
Address
email

Abstract: In this paper, we study the problem of solving long-horizon multi-stage manipulation tasks from raw image observations. Reinforcement learning (RL) is a promising approach for acquiring skills that can be sequenced to solve such tasks. However, the focus of RL algorithms has largely been on the success of those low-level skills, more so than satisfying the demands of extended multi-stage tasks that we ultimately want robots to complete. The latter demands robustness and persistence, as errors can compound over time, and may require the robot to have a number of primitive skills in its repertoire, rather than just one. To this end, we introduce EMBR, a model-based RL method that can learn primitive skills that are suitable for solving long-horizon visuomotor tasks. EMBR learns and plans using a learned model, critic, and success classifier, where the success classifier serves both as a reward function for RL and as a grounding mechanism to continuously detect if the robot should retry a skill when unsuccessful or under perturbations. Further, the learned model is task-agnostic and trained using data from all tasks, enabling the robot to efficiently learn a number of distinct primitives. These visuomotor primitive skills and their associated pre- and post-conditions can then be combined with off-the-shelf symbolic planners to solve long-horizon tasks. On a Franka Emika robot arm, we find that EMBR enables the robot to solve three long-horizon visuomotor tasks at 85% success rate, such as organizing an office desk, a file cabinet, and drawers, which require sequencing up to 12 skills, involve 14 unique learned primitives, and demand generalization to novel objects. Videos are at https://sites.google.com/view/anonymous-ember.

Keywords: model-based reinforcement learning, long-horizon planning

1 Introduction

We want robots to robustly perform a variety of long-horizon visuomotor manipulation tasks. Such a capability can enable robots to gradually become more useful in our physical world. Long-horizon manipulation tasks can be solved by composing a sequence of low-level primitive skills, otherwise known as "manipulation primitives" [1] or "options" [2]. For example, the long-horizon task of "organizing an office desk" often means performing a long sequence of shorter tasks, such as picking objects up, opening or closing drawers, placing objects into drawers, and putting pens into pen-holders. However, solving such a long-horizon task may require the robot to learn many low-level visuomotor skills, compose them sequentially, and generalize across novel objects. During task

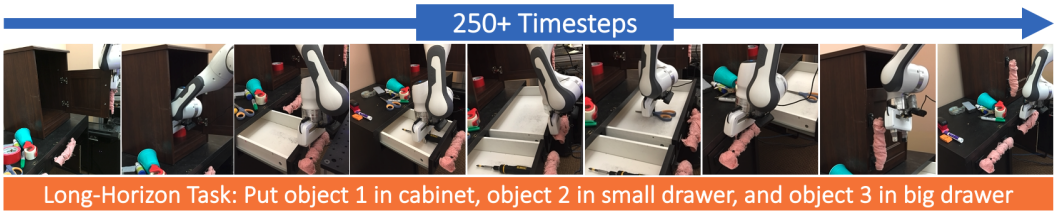

Figure 1: **Example-Driven Model-Based RL (EMBR).** EMBR enables a Franka Emika robot to solve long-horizon manipulation tasks involving up to 12 unique skills at 85% success with novel objects from raw image observations. Here, the robot organizes a desk, which involves a sequence of 10 unique skills.

Submitted to the 5th Conference on Robot Learning (CoRL 2021). Do not distribute.

execution, the robot also needs to detect its own failures (e.g. accidentally dropping an object) and correct them (e.g. picking the object up again) before continuing to the remaining parts of the task. We hope to endow robots with such generalization and robustness.

Approaches such as symbolic planning [3], hierarchical RL [4, 5, 6, 7, 8], and hierarchical planning [9] can accomplish compositional task generalization by reasoning in an abstract or symbolic space. While these approaches have shown promising results on tasks such as simulated tool use [10] and non-vision-based manipulation [9], they rely on the ability to acquire a repertoire of robust, low-level motor skills. These low-level skills can in principle be acquired through reinforcement learning, but a number of challenges remain. First, the acquisition of robust and persistent low-level visuomotor manipulation skills is still challenging, and the errors in completing any sub-sequence of a long-horizon task can compound over time and lead to overall task failure. Second, acquiring the requisite low-level visuomotor skills requires data-efficient learning and reward or task specification across a broad set of skills. Finally, composing the learned primitive skills to perform a long-horizon task requires persistent grounding of the robot's visual observation within the broader abstract plan.

To address these challenges, we introduce example-driven model-based RL (EMBR), an algorithm that learns a dynamics model and a set of Q-functions and success classifiers to efficiently acquire a repertoire of visuomotor skills that are both robust and can be grounded into symbols for symbolic planning of a long-horizon visuomotor task. More specifically, the success classifiers in EMBR enable both robustness and skill grounding, as they allow the robot to determine when a skill should be re-tried and can be used to determine the pre- and post-conditions of a skill. For example, a reward function and post-condition for the skill of opening a drawer can be obtained from learning a binary image classifier of "whether the drawer is open" given the robot's pixel observation. At the same time, this classifier also enables the robot to persistently try to open the drawer until it is finally open. Further, these success classifiers require only a modest number of example images, making it easy to specify a variety of skills. Finally, by learning a task-agnostic dynamics model, EMBR allows data to be shared across all tasks, leading to greater data-efficiency and generalization.

The main contribution of this work is a framework for learning grounded visuomotor skills from human-provided example images that can be sequenced to solve long-horizon tasks with off-the-shelf symbolic planners. Experiments on a Franka Emika Panda robot arm find that EMBR enables the robot to solve three long-horizon tasks from raw pixels with 85% success, such as organizing a desk, a cabinet, and two drawers, which require up to 14 learned unique primitives, sequencing up to 12 skills, and generalization to novel objects. Moreover, our ablation studies confirm that all components of our method including the dynamics model, success classifiers, Q-functions, and stages are essential to performance, particularly for tackling challenging long-horizon tasks.

2 Related Work

Reinforcement learning (RL) has a rich history of being used for robotic control problems on tasks from locomotion [11, 12, 13] to object manipulation [14, 15, 16]. In this work, we focus on learning *vision-based* robotic manipulation skills, a problem studied by many prior works [17, 18, 19, 20]. Model-free deep RL has been effective on short-horizon skills like object grasping [21, 20, 22], pushing and throwing [19, 23, 24], and multi-task learning [23, 25] from images. Alternatively, model-based approaches [26, 27], which explicitly learn the environment forward dynamics from images have also been employed for multi-task vision-based robotic manipulation tasks, either with planning algorithms [28, 29, 30, 31, 32, 33, 34, 35] or for optimizing a parametric policy [36]. Unlike these prior works, our algorithm uses a model in conjunction with task-specific success classifiers and Q-functions to learn a wide range of skills suitable for sequencing into long-horizon tasks. In Section 5, we find that both components of our approach are essential for robust performance.

Even with powerful reinforcement learning methods, task specification on real robots remains challenging, as engineering rewards on physical systems can be costly and time-consuming [37]. Motivated by this, many works have studied the reward specification, including inverse reinforcement learning [38] with robot demonstrations [39, 40, 41, 42, 43], learning rewards from user preferences [44, 45, 46, 47], and learning rewards from videos of humans [48, 49]. One common approach in visual RL is to learn to reach a goal image using a coarse measure of reward like negative ℓ_2 pixel distance [30, 50, 51] or temporal distance [52, 53, 34]. While these techniques have had some success on real robots, such rewards often provide a sparse and difficult to optimize reward signal. Alternatively, a number of recent works learn a classifier from a modest number of examples goal states [54, 55, 24, 56, 57, 25] and tries to learn agents which maximize the classifier score. In this

work, we extend this approach to multiple stages of human-provided goals, which we find enables more effective learning of hard exploration skills, and find that such reward specification can be used to learn 14 unique visuomotor skills on a real robot, without the need for any demonstrations.

Like many prior works, the goal of this work is to complete challenging long-horizon visuomotor tasks. One class of prior work is hierarchical RL, which aims to learn temporally-extended primitive skills and a high-level policy over them [4, 5, 6, 7, 8, 58]. Such approaches include jointly learning the primitive skills with the high-level policy through goal generation or an "options" framework [6, 8, 59], learning primitives through intrinsic motivation or other auxiliary objectives [60, 61, 62], from demonstration behavior [63, 64, 65, 66, 67], or from low-level, potentially goal conditioned, reward functions [7, 58, 68, 69]. EMBR is not a hierarchical RL algorithm, as it only learns low-level skills provided to the high-level planner and does not learn a model over options. Nevertheless, our work is similar to the last group, except that we learn visuomotor primitives on a real robot from rewards derived from human examples. Other works have also explored learning hierarchical policies on top of hand-designed skills [70, 71, 72]. These methods are complementary to EMBR, as skills learned via EMBR can be incorporated into most hierarchical policy learning algorithm.

Alternatively, a number of recent works have studied long horizon visual planning [52, 53, 73, 74, 75, 50, 76, 77, 78], often by combining a structured search approach with learned components like generative models [73, 50, 76] or distance functions [53, 34]. While these approaches have demonstrated impressive results on problems such as visual navigation [52, 53], unlike prior work we learn manipulation tasks on a *real robots* that consists of up to 12 unique skills and 250+ timesteps.

Lastly, the task and motion planning (TAMP) literature [79, 80] has extensively studied long-horizon robotic tasks. Unlike these works, we do not assume pre-defined state representations or accurate state estimators, making it possible to handle novel objects in cluttered scenes. Like our work, a number of works have also explored combining learned models and skills with hand-designed or symbolic planning [10, 9, 81] and learning to ground skills for planning [82]. Unlike these techniques, a key insight in this work is that we can jointly learn visuomotor skills and the associated grounding of their pre- and post-conditions, allowing us to leverage off-the-shelf symbolic planners to execute long horizon tasks in a closed-loop fashion directly from pixels on a real robot.

3 Preliminaries

Tasks and skills. We consider the problem of solving a long-horizon visuomotor task $\mathcal{M}$ by learning a repertoire of visuomotor skills. We model the robot's environment as a controlled Markov process $\mathcal{E} = \langle \mathcal{S}, \rho_0, \mathcal{A}, \mathcal{T}, \gamma, H \rangle$, with an image observation space $s \in \mathcal{S}$, an initial state distribution ρ_0, an action space $a \in \mathcal{A}$, a dynamics model $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, a discount factor $\gamma \in [0, 1)$, and a finite horizon H. We notice that many long-horizon tasks are composed of a sequence of lower-level skills. For example, to organize an office desk, the robot needs to acquire a number of skills such as grasping an object, opening a drawer, and placing an object in a drawer. Let K denote the total number of unique skills the robot has acquired, and $k \in [1, K]$ denote the kth skill in the robot's skill repertoire. Depending on the environment state, the order in which the skills should be executed to successfully complete a long-horizon task varies. For example, to organize a desk, the robot may not need to use the skill of closing a drawer if the drawer is already closed. Here, each skill is an MDP $\mathcal{M}^k = \langle \mathcal{E}, \mathcal{R}^k \rangle$, where the robot's environment $\mathcal{E}$ is shared across all tasks and skills, and $\mathcal{R}^k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function for the kth skill (e.g. the skill of opening a drawer).

Problem Definition. Formally, EMBR takes as input N human-provided example images $\Psi^k = s_1, \ldots, s_N$ for each skill k we intend to robot to learn. Our goal (i.e. EMBR's output) is to learn the symbolic grounding of the pre- and post-conditions $g^k_{pre}, g^k_{post} : \mathcal{S} \to \{0, 1\}$, and a policy $\pi^k : \mathcal{S} \to \Pi(\mathcal{A})$ for each skill k, where $\Pi(\cdot)$ defines a probability distribution over a set. Each policy has a Q-function, $Q^k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which estimates the discounted cumulative reward for taking action a from observation s and following π^k onward.

Variational Autoencoders (VAEs). VAEs compress high-dimensional observations s such as images into an embedding z. VAEs can be optimized by maximizing the evidence lower bound (ELBO): $\max_{p,q} \mathcal{L}_{\text{vae}}(s)$, where $\mathcal{L}_{\text{vae}}(s) = \mathbb{E}_{q(z|s)} \Big[\log p \left(s \mid z \right) \Big] - D_{\text{KL}} \Big(q(z \mid s) \parallel p(z) \Big)$ (1), where p denotes the generative model, and q denotes the variational distribution.

Symbolic planning. A symbolic planner performs task planning in the symbolic domain. Concretely, given a set of candidate actions, each action's associated pre- and post-conditions, the current

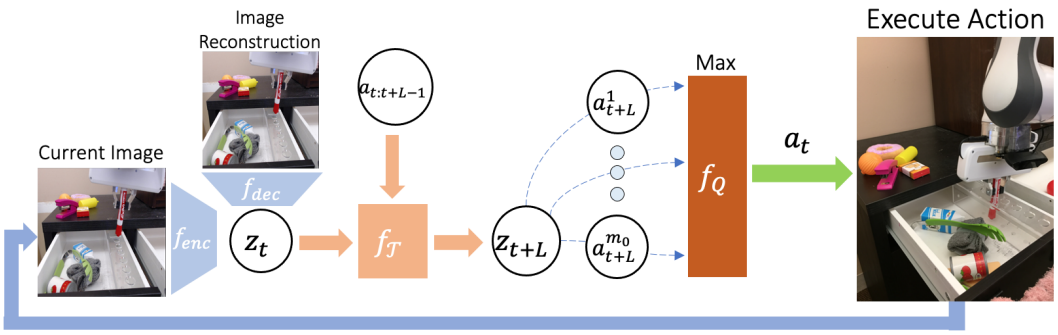

Figure 2: **Example-Driven Model-based RL (EMBR)**. During training, EMBR learns a VAE f_{vae} and a dynamics model $f_{\mathcal{T}}$ in the VAE latent space (**left**). For each skill, EMBR uses human provided example images to train a success-classifier-based reward function $f_{\mathcal{R}}$, and associated Q function f_Q. At test time actions are selected by model predictive control over the latent dynamics model using the Q-value as the planning cost (**middle**). The reward function further serves as a success classifier that encourages the robot to persistently try to reach a successful state, such as inserting the marker in the "Insert Marker" skill (**right**).

condition h and a goal condition g, a symbolic planner outputs the sequence of actions that allows the robot to reach from h to g. The pre-condition of an action is a set of predicates that defines the condition that needs to be satisfied in order for the action to be executable. The post-condition of a skill is a set of predicates that defines the effect of executing an action. For example, the pre-condition and post-condition of an "Open left drawer" action is "Drawer is closed" and "Drawer is open" respectively. In our context, the candidate actions are the primitive skills learned by EMBR.

4 Example-Driven Model-Based RL (EMBR)

To learn a large repertoire of skills that can be sequenced to solve long-horizon robotic manipulation, we propose Example-driven Model-based RL, or EMBR. Below we first describe how EMBR learns each visuomotor skill without any demonstration skills or trajectories during training, and then elaborate on how these learned skills can be composed to solve a long-horizon task at test time.

4.1 Learning a Repertoire of Visuomotor Skills Autonomously without Demonstrations

During training, EMBR learns a repertoire of low-level persistent visuomotor skills, such as opening a drawer or inserting a marker, using a set of example success images provided by a person for each task. The model learned by EMBR, denoted as f, is composed of four components: a global variational autoencoder (VAE) f_{vae}, a global latent dynamics model $f_{\mathcal{T}}$, stage-specific binary reward functions $f^k_{\mathcal{R}}$, and stage-specific Q-functions f^k_Q. The VAE and latent dynamics model jointly learn a low-dimensional representation of the image observations to address the challenge of efficiently learning many primitive skills; both are trained on data collected across all skills. The skill-specific binary classifiers allow us to learn a reward function for each skill from human-provided images. We use the latent dynamics model to perform model predictive control with the skill-specific Q-function as the terminal planning objective. The skill-specific Q-functions and reward functions are trained on skill-specific data. Next, we elaborate on each of these components.

Learning a low-dimensional latent space with a VAE and a latent dynamics model. Performing a long-horizon task using a large number of visuomotor skills demands the algorithm to be sample-efficient and fast at execution time. To address both of these demands, we learn a single VAE $f_{\text{vae}} = \{f_{\text{enc}}, f_{\text{dec}}\}$ to construct a latent space $\mathcal{Z}$. This VAE consists of an encoder $f_{\text{enc}} : \mathcal{S} \to \Pi(\mathcal{Z})$ and a decoder $f_{\text{dec}} : \mathcal{Z} \to \Pi(\mathcal{S})$. We train f_{vae} using data collected across all skills via Eq. 1, where $p(s \mid z) \equiv f_{\text{dec}}(s_t \mid z_t)$, $p(z) \sim \mathcal{N}(0, 1)$, and $q(z \mid s) \equiv f_{\text{enc}}(z_t \mid s_t)$. By learning from all skills, the VAE can improve data efficiency. In practice, we use the greedy hierarchical VAE [51] as f_{vae} with the current image observation as a context frame. Based on this latent space, EMBR learns a latent dynamics model $f_{\mathcal{T}}$ jointly with f_{vae} with data collected across all skills via l_2 loss:

$$\min_{f_{\text{vae}}, f_{\mathcal{T}}} \mathbb{E}_{s_{t:t+L}, a_{t:t+L} \sim \mathcal{D}} \left[\sum_{l=1}^{L} -\mathcal{L}_{\text{vae}}(s_{t+l-1}) + \left(f_{\mathcal{T}}(z_t, a_{t:t+l-1}) - z_{t+l} \right)^2 \right] \tag{2}$$

where $z_t \sim f_{\text{enc}}(\cdot \mid s_t)$, L is the rollout length of $f_{\mathcal{T}}$ and $\mathcal{D}$ is the dataset collected by the robot.

Obtaining rewards by learning image classifiers. Obtaining a reward function in RL typically requires some forms of human supervision such as reward engineering or demonstrations. In EMBR,

binary image classifiers serve not only as reward functions for learning primitive visuomotor skills, but also as success detectors as well as grounding of raw pixels for long-horizon symbolic planning. Concretely, we learn a reward function $f_{\mathcal{R}}^k$ for each skill $\mathcal{M}^k$ by training binary image classifiers to distinguish human-provided example images from robot-collected non-example images using the binary cross-entropy objective [55, 56]. The reward learning objective for skill k is:

$$\max_{f_{\mathcal{R}}^k} \mathbb{E}_{s^+\sim\mathcal{D}^+, z^+\sim f_{\text{enc}}(\cdot|s^+)}\left[\log\left(f_{\mathcal{R}}^k(z^+)\right)\right] + \mathbb{E}_{s^-\sim\mathcal{D}, z^+\sim f_{\text{enc}}(\cdot|s^+)}\left[\log\left(1 - f_{\mathcal{R}}^k(z^-)\right)\right] \tag{3}$$

Here, $\mathcal{D}^+$ is the dataset of images labeled as positive by a person (typically 100-200 images), s^+ are the images sampled from $\mathcal{D}^+$, and s^- are sampled from all images in $\mathcal{D}$, the dataset collected by the robot, and are considered negative by default. In our experiments, these positive examples can be collected in 15-30 minutes per visuomotor skill, and 345 minutes in total for all 14 skills. No demonstration trajectories are collected. For skills that are longer horizon, such as grasping objects, we break the skill down into two "stages" and learn a reward function for each stage. Details of such skill breakdown and the example-image collection process are in Appendix C and M respectively.

At test time, the learned image classifiers are re-purposed as post-condition grounding of each skill for long-horizon task execution. Some of the pre-conditions the task planner requires can be derived from existing post-conditions. For example, for a drawer-opening skill, the pre-condition of "drawer is not open" can be derived from the post-condition of "drawer is open" by negating this post-condition. For the remaining pre-conditions that cannot be derived from any existing post-conditions, we train additional image classifiers using Eq. 3.

Learning Q-functions for model-based control. While one can directly use the latent dynamics model for model-based planning, the difficulty of learning an accurate dynamics model for the full horizon of a primitive skill increases with the length of the skill itself. To alleviate this challenge and further accelerate skill learning, we learn a Q-function f_Q^k in the latent space. Concretely, we perform Q-learning via the following objective for skill k:

$$\min_{f_Q^k} \mathbb{E}_{s_t, a_t, s_{t+1}\sim\mathcal{D}}\left[f_Q^k\left(z_t, a_t\right) - \left(\overline{f_{\mathcal{R}}^k}\left(z_{t+1}\right) + \gamma\overline{f_{\mathcal{R}}^k}\left(z_{t+1}\right)\max_{a_{t+1}} f_Q^k(z_{t+1}, a_{t+1})\right)\right]^2 \tag{4}$$

where $\overline{f_{\mathcal{R}}^k}(z) \equiv \mathbb{1}\{f_{\mathcal{R}}^k(z) > 0.5\}$, $z_t \sim f_{\text{vae}}(\cdot \mid s_t)$, $z_{t+1} \sim f_{\text{vae}}(\cdot \mid s_{t+1})$, and the Q-value target is computed by maximizing over m_1 randomly and uniformly sampled actions $a_{t+1}^{1:m_1}$. In the case of a longer-horizon skill such as grasping objects, we break the skill down into two "stages" and learn a reward function for each stage (Appendix C).

Algorithm 1 EMBR at Training Time: Learn a Repertoire of Visuomotor Skills

Input: Example images per skill; **Output**: K primitive visuomotor skills
$\mathcal{D} \leftarrow$ to empty dataset, $f \leftarrow$ random weights, $J \leftarrow$ gradient update steps, $L \leftarrow$ rollout horizon
for Skill $k \in [1, K]$ **do**
  while Not reaching target success rate for skill k as determined by success classifier $f_{\mathcal{R}}^k$ **do**
    Collect trajectory $d = \{s_{1:H}, a_{1:H}\}$ using actions selected by Eq. 5 and append to dataset $\mathcal{D}$
    for $j = 1 : J$ **do**
      Sample a minibatch of trajectories of timesteps $L \leq H$: $\overline{\mathcal{D}} = \{s_{t:t+L}^{1:B}, a_{t:t+L}^{1:B}\} \sim \mathcal{D}$
      Update $f_{\text{vae}}, f_{\mathcal{T}}$ jointly via Eq. 2, $f_{\mathcal{R}}^k$ via Eq. 3, and f_Q^k via Eq. 4, all with $\overline{\mathcal{D}}$

Algorithm 2 EMBR at Test Time: Long-Horizon Task Execution via Skill Composition

Input: trained model f, symbolic goal condition g
Compute current symbolic state h using $f_{\mathcal{R}}$; $s \leftarrow$ new image observation
while $h \neq g$ **do**
  Compute the first skill to perform $k \in [1, K]$ using symbolic planner given goal g and current state h
  while $f_{\mathcal{R}}^k(f_{\text{enc}}(s)) < 0.5$ **do**
    Execute robot action a via Eq. 5 given s
    $s \leftarrow$ current image observation
  Compute current symbolic condition h from $f_{\mathcal{R}}$

At both training and test time, we perform model-based planning in the latent space by first rolling out $f_{\mathcal{T}}$ for L timesteps into the future and then evaluating the predicted future latent space observa-

(a) Robot Environment (b) Train Objects (c) Validation Objects (d) Test Objects

Figure 3: **Real-Robot Experimental Setup**. In Fig. **(a)**, a Franka Emika Panda robot is equipped with five RGB cameras (red circles) capturing 64×64 RGB images. Quantitative justification for having five cameras is in Appendix J. The robot interacts with a desk that has two drawers and a cabinet, as identified by three pink handles in Fig. **(a)**. Fig. **(b), (c) and (d)** visualize train, validation, and test objects used in all experiments.

tion at timestep $t + L$ using the learned Q-function, a process that is illustrated in Fig. 2:

$$a_t = \left(\underset{a_{t:t+T-1}, a_{t+L} \in \mathcal{A}}{\arg\max} f_Q^k \left(f_{\mathcal{T}} \left(z_t, a_{t:t+L-1} \right), a_{t+L} \right) \right)_1, \text{where } z_t \sim f_{\text{vae}}(\cdot \mid s_t) \tag{5}$$

Here, the outer and inner maximization operations are taken over m_0 and m_1 samples respectively.

Integrating these four components, EMBR is trained simultaneously across all components from scratch (i.e. without pre-training) for learning a repertoire of visuomotor primitive skills (Alg. 1). Concretely, the robot begins with an empty dataset $\mathcal{D}$ and can learn each skill in almost any order (Appendix D.2). During skill learning, the robot collects a new trajectory using actions selected from Eq. 5 and updates all model components using $\mathcal{D}$. This process repeats until the current skill reaches a desirable success rate as computed by the success classifiers, before the robot moves on to learn the next skill. As the robot acquires more skills, the dataset $\mathcal{D}$ accumulates and improves the generalization of the visual dynamics model, which in-turn accelerates future skill learning. Wall-clock training time per skill and hyperparameters are detailed in Appendix O and P respectively.

4.2 Composing Learned Visuomotor Primitive Skills for Long-Horizon Task Execution

At test time, EMBR provides the learned primitive skills and the symbolic grounding of each skill's post-conditions (in the form of success classifiers) to a high-level task planner (standard PDDL STRIPS symbolic planner [83] in our case) to perform long-horizon task. As elaborated in Alg. 2, the human first specifies a long-horizon task using a symbolic goal condition g. Next, EMBR computes the current symbolic condition of the environment from raw pixel observations by passing the current image observations to each of the success classifiers. Using both the symbolic goal condition and the current symbolic condition, the symbolic planner computes the sequence of skills to perform and executes the first skill of this sequence using actions selected from Eq. 5. This procedure repeats until the current condition matches the goal condition, after which task execution terminates. Note that the arm position of the robot is reset after executing each skill. The predicate definitions, the pre- and post-conditions and the full details of the symbolic planner used in Section 5 are in Appendix D.3.

5 Experiments

Our experiments aim to answer four key questions: 1) how does EMBR compare to prior RL algorithms on both performing primitive skills and when sequencing those skills to solve long-horizon tasks? 2) how necessary are stage-specific reward functions (as elaborated in Appendix C) compared to a single skill-specific reward function? 3) how important is replanning using the classifiers for the robustness of the algorithm? and 4) how important is each of the five camera observations to the performance of the primitive skills? To answer these questions, we conduct experiments across 14 primitive skills and three long-horizon tasks, using the experimental setup in Fig. 3, in which a Franka Emika Panda robot has access to five cameras, each capturing 64×64 RGB images. One of the five cameras is mounted on the wrist. To test object generalization, all objects used in all evaluations are novel and unseen during training, except for the desk, drawers, cabinet, and the transparent marker holder. Fig. 3 visualizes the train, validation, and test objects used in all experiments. See Appendix D for full experimental details. Videos are at https://sites.google.com/view/anonymous-ember.

Table 1: Successful trials (out of 20) and success rates of visuomotor primitive skills. Here, EMBR outperforms the three prior methods by 10-20%, 15-25%, and 50-65% respectively on high-precision tasks.

k	Primitive Skill	Precision	EMBR (Ours)	Prior Method 1: EMBR w/o $f_{\mathcal{T}}$	Prior Method 2: EMBR w/o f_{vae}, $f_{\mathcal{T}}$ (Qt-Opt) [20]	Prior Method 3: BEE [56]	Ablation 1: EMBR w/ Skill-specific Rewards
1	Insert marker into a marker holder	High	20 (100%)	16 (80%)	15 (75%)	8 (40%)	20 (100%)
2	Grasp object from left drawer (single)	High	18 (90%)	15 (75%)	14 (70%)	8 (40%)	15 (75%)
	Grasp object from left drawer (clutter)	High	18 (90%)	14 (70%)	15 (75%)	7 (35%)	14 (70%)
3	Grasp object from right drawer (single)	High	18 (90%)	16 (80%)	14 (70%)	5 (25%)	16 (80%)
	Grasp object from right drawer (clutter)	High	18 (90%)	15 (75%)	15 (75%)	5 (25%)	16 (80%)
4	Grasp object from desk (single)	High	18 (90%)	16 (80%)	15 (75%)	5 (25%)	12 (60%)
	Grasp object from desk (clutter)	High	18 (90%)	16 (80%)	15 (75%)	6 (30%)	12 (60%)
5	Open right drawer	Medium	20 (100%)	18 (90%)	16 (80%)	14 (70%)	17 (85%)
6	Open left drawer	Medium	20 (100%)	19 (95%)	16 (80%)	8 (40%)	16 (80%)
7	Open cabinet	Medium	20 (100%)	18 (90%)	17 (85%)	12 (60%)	18 (90%)
8	Place object on desk	Low	20 (100%)	20 (100%)	20 (100%)	20 (100%)	20 (100%)
9	Place object in cabinet	Low	20 (100%)	20 (100%)	20 (100%)	20 (100%)	20 (100%)
10	Place object into left drawer	Low	20 (100%)	20 (100%)	20 (100%)	20 (100%)	20 (100%)
11	Place object into right drawer	Low	20 (100%)	20 (100%)	20 (100%)	20 (100%)	20 (100%)
12	Close right drawer	Low	20 (100%)	20 (100%)	20 (100%)	20 (100%)	20 (100%)
13	Close left drawer	Low	20 (100%)	20 (100%)	20 (100%)	20 (100%)	20 (100%)
14	Close cabinet door	Low	20 (100%)	20 (100%)	20 (100%)	20 (100%)	20 (100%)

Table 2: Successful trials (out of 20) and success rates of long-horizon tasks. K^* and K refer to the minimum number of skills that need to be sequenced and the minimum number of *unique* skills required, respectively. Here, EMBR outperforms the three prior methods by 15-25%, 20-30%, and 85% respectively.

	Long-horizon Task	K^*	K	EMBR (Ours)	Prior Method 1: EMBR w/o $f_{\mathcal{T}}$	Prior Method 2: EMBR w/o f_{vae}, $f_{\mathcal{T}}$ (Qt-Opt) [20]	Prior Method 3: BEE [56]	Prior Method 4: Deep Skill Chaining [59]	Ablation 2: EMBR w/o Replanning
1	Organize Desk and Cabinet	8	4	17 (85%)	14 (70%)	11 (55%)	0 (0%)	0 (0%)	11 (55%)
2	Organize Markers	9	5	17 (85%)	13 (65%)	11 (55%)	0 (0%)	0 (0%)	10 (50%)
3	Rearrange Objects	12	12	17 (85%)	15 (75%)	13 (65%)	0 (0%)	0 (0%)	11 (55%)

Comparisons. We compare EMBR to three prior methods: **(1)** EMBR w/o $f_{\mathcal{T}}$, which corresponds to EMBR without the dynamics model, i.e. using only the Q-function and the VAE for action selection; **(2)** EMBR w/o f_{vae}, $f_{\mathcal{T}}$ (Qt-Opt) [20], which is our re-implemented version of Qt-Opt [20] with a VICE-like success classifier [55] that does not learn a latent space and uses only the Q-function for action selection; and **(3)** BEE [56], which uses the success classifier instead of the Q-function for model predictive control. Comparison to Prior Method 4, the Deep Skill Chaining [59] hierarchical RL method, is in Appendix G. All comparisons are representative of prior methods for vision-based robotic RL. For all experiments, the data used to train each method is the same dataset collected by EMBR using Alg. 1, which contains 300-400 trajectories collected in 16 hours per skill on average. This provides a more direct comparison and in effect makes the prior methods stronger since the challenge of online exploration is addressed by EMBR during data collection.

5.1 Short-horizon Primitive Skill Performance

To answer the first question, we compare all methods across 14 visuomotor skills. In Table 1, we observe that "EMBR (Ours)" can reliably solve all 14 skills with at least 90% success. Compared to "EMBR w/o $f_{\mathcal{T}}$" and "EMBR w/o f_{vae}, $f_{\mathcal{T}}$ (Qt-Opt)", "EMBR (Ours)" improves success rates by 10-20% and 15-25% across high-precision skills, 5-10% and 15-20% for medium-precision skills and 0% for low-precision skills. We hypothesize that EMBR's improvement over "EMBR w/o $f_{\mathcal{T}}$" is due to its ability to alleviate overestimation of Q-values in the presence of limited training trajectories (300-400 per skill), and the addition success rate improvement of EMBR over "EMBR w/o f_{vae}, $f_{\mathcal{T}}$" is due to less overfitting of the success classifiers and Q-functions in the VAE latent space. Finally, EMBR outperforms BEE [56] by 50-65% on high-precision tasks because BEE [56] cannot capture action consequences beyond its planning horizon and the model is not always accurate for the full horizon of a skill. We elaborate on these hypotheses in Appendix L.

5.2 Long-horizon Task Performance

Next, we compare EMBR to the three prior methods when sequencing the learned visuomotor skills with symbolic planning to solve three challenging long-horizon tasks (Fig. 4):

1. **Organize Desk and Cabinet (8 Skills)** - putting three novel objects cluttered on the desk into the cabinet.
2. **Organize Markers (9 Skills)** - picking up three markers and inserting them into a transparent holder.[1]

[1] In this task, we also equip all methods with a scripted skill that uses the edge of the desk as a supporting point to re-orient a marker.

Figure 4: **Long-horizon Task Experiments**. EMBR enables a Franka Emika robot to solve long-horizon manipulation tasks with novel objects from raw image observations. Here, the robot performs three long-horizon tasks with 85% success, which involve a sequence of up to 12 unique primitive visuomotor skills.

3. **Rearrange Objects (12 Skills)** - picking up three objects and placing them to the desk, cabinet or drawers.

Detailed task descriptions are in Appendix D.7. In Table 2, we find that "EMBR (Ours)" solves these long-horizon visuomotor tasks with 85% success, including the most challenging tasks that require up to 12 unique skills. Compared to "EMBR w/o $f_{\mathcal{T}}$" and "EMBR w/o $f_{\text{vae}}, f_{\mathcal{T}}$ (Qt-Opt)", EMBR leads to 10-20% and 20-30% improvement in success rates respectively because EMBR achieves higher success rate on individual primitive skills that require higher precision, such as inserting marker and grasping object from desk or drawers. Finally, BEE [56], is unable to perform any long-horizon task due to low performance on skills with medium or high precision. This reflects the importance of learning robust, high-success-rate primitive skills using EMBR.

5.3 Ablations

Next, we conduct six ablations to examine the importance of various components of EMBR. Ablation 3, which compares EMBR with vs. without any one camera observation, is in Appendix J. Comparisons to the original BEE and training without stages are in Appendix H and I.

Ablation 1: Stage-specific vs. skill-specific reward function. While EMBR accelerates skill acquisition by learning a reward function for each stage of a skill, one can in principle learn just one reward function for each skill. In this ablation, we answer Question 2 by comparing EMBR to an ablated EMBR that learns each skill with a single skill-specific reward function. This does not affect data collection and exploration, which stage-specific rewards can also help. In Table 1, we find that using stage-specific reward functions outperforms using a single skill-specific reward function by 10-30% for higher precision skills such as grasping an object in various locations, indicating the importance of using stage-specific reward functions in long-horizon or higher precision skills.

Ablation 2: Task performance with vs. without classifier-based replanning. While EMBR improves the robustness of long-horizon task execution by using success classifiers during planning, one can in principle perform model-based planning without these classifiers. In this ablation, we answer Question 3 by comparing EMBR to an ablation that executes each skill of a task sequentially without using the success classifiers to determine when the skill is successful. In Table 2, we find that long-horizon task execution with classifiers improves performance by 30-35% success rate, due to the robot's ability to detect success in a closed-loop manner until successful. These ablations and Table 2 suggest that *all* components of EMBR are critical to solving long-horizon tasks.

6 Conclusion and Future Work

In this work, we propose EMBR – an example-driven model-based RL algorithm for learning visuomotor skills that can be sequenced to perform long-horizon tasks from raw pixels. Across three long-horizon visuomotor tasks which require up to 14 unique primitives and sequencing up to 12 skills, we find that EMBR consistently achieves 85% success rates, outperforms prior methods, and generalizes to novel objects. Nonetheless, EMBR and our experiments also have a number of important limitations, including the limited scope of tasks, the amount of supervision required to solve these complex long-horizon tasks, and the limited degree of generalization especially across different scenes. We discuss these and other limitations in detail in Appendix Q, all of which leave exciting avenues for further pushing the capabilities of robot learning systems in future work.

References

[1] T. Kröger, B. Finkemeyer, and F. M. Wahl. Manipulation primitives—a universal interface between sensor-based motion control and robot programming. In *Robotic systems for handling and assembly*, pages 293–313. Springer, 2010.

[2] P.-L. Bacon, J. Harb, and D. Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[3] P. Smolensky. Connectionist ai, symbolic ai, and the brain. *Artificial Intelligence Review*, 1(2):95–109, 1987.

[4] R. Sutton, D. Precup, and S. Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112:181–211, 1999.

[5] A. G. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13(1-2):41–77, Jan. 2003. ISSN 0924-6703.

[6] P. Bacon, J. Harb, and D. Precup. The option-critic architecture. *CoRR*, abs/1609.05140, 2016.

[7] O. Nachum, S. Gu, H. Lee, and S. Levine. Data-efficient hierarchical reinforcement learning. *CoRR*, abs/1805.08296, 2018.

[8] A. Levy, R. Platt, and K. Saenko. Hierarchical reinforcement learning with hindsight. In *International Conference on Learning Representations*, 2019.

[9] G. Konidaris, L. Kaelbling, and T. Lozano-Perez. From skills to symbols: Learning symbolic representations for abstract high-level planning. *J. Artif. Intell. Res.*, 61:215–289, 2018.

[10] M. Toussaint, K. R. Allen, K. Smith, and J. Tenenbaum. Differentiable physics and stable modes for tool-use and manipulation planning. In *Robotics: Science and Systems*, 2018.

[11] H. Benbrahim and J. A. Franklin. Biped dynamic walking using reinforcement learning. *Robotics and Autonomous Systems*, 22(3):283–302, 1997. ISSN 0921-8890. doi:https://doi.org/10.1016/S0921-8890(97)00043-2. URL https://www.sciencedirect.com/science/article/pii/S0921889097000432. Robot Learning: The New Wave.

[12] N. Kohl and P. Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, 3:2619–2624 Vol.3, 2004.

[13] R. Tedrake, T. Zhang, and H. Seung. Stochastic policy gradient reinforcement learning on a simple 3d biped. *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, 3:2849–2854 vol.3, 2004.

[14] V. Gullapalli. Skillful control under uncertainty via direct reinforcement learning. *Robotics and Autonomous Systems*, 15(4):237–246, 1995. ISSN 0921-8890. doi:https://doi.org/10.1016/0921-8890(95)00006-2. URL https://www.sciencedirect.com/science/article/pii/0921889095000062. Reinforcement Learning and Robotics.

[15] J. Peters and S. Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–697, 2008. ISSN 0893-6080. doi:https://doi.org/10.1016/j.neunet.2008.02.003. URL https://www.sciencedirect.com/science/article/pii/S0893608008000701. Robotics and Neuroscience.

[16] M. Deisenroth, C. Rasmussen, and D. Fox. Learning to control a low-cost manipulator using data-efficient reinforcement learning. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2011. doi:10.15607/RSS.2011.VII.008.

[17] S. Lange, M. Riedmiller, and A. Voigtländer. Autonomous reinforcement learning on raw visual input data in a real world application. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2012. doi:10.1109/IJCNN.2012.6252823.

[18] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016. URL http://jmlr.org/papers/v17/15-522.html.

[19] A. Ghadirzadeh, A. Maki, D. Kragic, and M. Björkman. Deep predictive policy training using reinforcement learning. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2351–2358, 2017.

[20] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018.

[21] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3406–3413, 2016.

[22] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. 2018.

[23] A. V. Nair, V. Pong, M. Dalal, S. Bahl, S. Lin, and S. Levine. Visual reinforcement learning with imagined goals. *Advances in Neural Information Processing Systems*, 31:9191–9200, 2018.

[24] A. Singh, L. Yang, C. Finn, and S. Levine. End-to-end robotic reinforcement learning without reward engineering. In *Proceedings of Robotics: Science and Systems*, FreiburgimBreisgau, Germany, June 2019.

[25] D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, and K. Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *ArXiv*, abs/2104.08212, 2021.

[26] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. Learning latent dynamics for planning from pixels. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2555–2565. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/hafner19a.html.

[27] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2019.

[28] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in neural information processing systems*, pages 64–72, 2016.

[29] C. Finn and S. Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2786–2793. IEEE, 2017.

[30] F. Ebert, C. Finn, S. Dasari, A. Xie, A. Lee, and S. Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.

[31] F. Ebert, S. Dasari, A. X. Lee, S. Levine, and C. Finn. Robustness via retrying: Closed-loop robotic manipulation with self-supervised learning. In *Conference on Robot Learning (CoRL)*, 2018.

[32] Y.-C. Lin, M. Bauzá, and P. Isola. Experience-embedded visual foresight. *ArXiv*, abs/1911.05071, 2019.

[33] H. Suh and R. Tedrake. The surprising effectiveness of linear models for visual foresight in object pile manipulation. *ArXiv*, abs/2002.09093, 2020.

[34] S. Tian, S. Nair, F. Ebert, S. Dasari, B. Eysenbach, C. Finn, and S. Levine. Model-based visual planning with self-supervised functional distances. *International Conference on Learning Representations (ICLR)*, 2021.

[35] M. Bhardwaj, A. Handa, D. Fox, and B. Boots. Information theoretic model predictive q-learning. In *Learning for Dynamics and Control*, pages 840–850. PMLR, 2020.

[36] R. Rafailov, T. Yu, A. Rajeswaran, and C. Finn. Offline reinforcement learning from images with latent space models. *arXiv preprint arXiv:2012.11547*, 2020.

[37] H. Zhu, J. Yu, A. Gupta, D. Shah, K. Hartikainen, A. Singh, V. Kumar, and S. Levine. The ingredients of real-world robotic reinforcement learning. *ArXiv*, abs/2004.12570, 2020.

[38] P. Abbeel and A. Y. Ng. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 1, 2004.

[39] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 729–736, 2006.

[40] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proc. AAAI*, pages 1433–1438, 2008.

[41] M. Wulfmeier, P. Ondruska, and I. Posner. Maximum entropy deep inverse reinforcement learning, 2016.

[42] C. Finn, S. Levine, and P. Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*, pages 49–58. PMLR, 2016.

[43] J. Fu, K. Luo, and S. Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.

[44] D. Sadigh, A. D. Dragan, S. S. Sastry, and S. A. Seshia. Active preference-based learning of reward functions. In *Proceedings of Robotics: Science and Systems (RSS)*, 7 2017. doi:10.15607/RSS.2017.XIII.053.

[45] D. S. Brown and S. Niekum. Deep bayesian reward learning from preferences, 2019.

[46] M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh. Learning reward functions by integrating human demonstrations and preferences. In *Proceedings of Robotics: Science and Systems (RSS)*, 6 2019. doi:10.15607/rss.2019.xv.023.

[47] D. S. Brown, R. Coleman, R. Srinivasan, and S. Niekum. Safe imitation learning via fast bayesian reward inference from preferences, 2020.

[48] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020.

[49] A. S. Chen, S. Nair, and C. Finn. Learning generalizable robotic reward functions from "in-the-wild" human videos, 2021.

[50] S. Nair and C. Finn. Hierarchical foresight: Self-supervised learning of long-horizon tasks via visual subgoal generation. *ICLR*, 2020.

[51] B. Wu, S. Nair, R. Martin-Martin, L. Fei-Fei*, and C. Finn*. Greedy hierarchical variational autoencoders for large-scale video prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[52] N. Savinov, A. Dosovitskiy, and V. Koltun. Semi-parametric topological memory for navigation, 2018.

[53] B. Eysenbach, R. Salakhutdinov, and S. Levine. Search on the replay buffer: Bridging planning and reinforcement learning, 2019.

[54] A. Xie, A. Singh, S. Levine, and C. Finn. Few-shot goal inference for visuomotor learning and planning. In *CoRL*, 2018.

[55] J. Fu, A. Singh, D. Ghosh, L. Yang, and S. Levine. Variational inverse control with events: A general framework for data-driven reward definition. In *Advances in Neural Information Processing Systems*, 2018.

[56] A. S. Chen, H. Nam, S. Nair, and C. Finn. Batch exploration with examples for scalable robotic reinforcement learning. *arXiv preprint arXiv:2010.11917*, 2020.

[57] B. Eysenbach, S. Levine, and R. Salakhutdinov. Replacing rewards with examples: Example-based policy search via recursive classification, 2021.

[58] Y. Jiang, S. Gu, K. Murphy, and C. Finn. Language as an abstraction for hierarchical deep reinforcement learning. In *NeurIPS*, 2019.

[59] A. Bagaria and G. Konidaris. Option discovery using deep skill chaining. In *International Conference on Learning Representations*, 2019.

[60] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is all you need: Learning skills without a reward function. *CoRR*, abs/1802.06070, 2018.

[61] A. Sharma, S. Gu, S. Levine, V. Kumar, and K. Hausman. Dynamics-aware unsupervised discovery of skills, 2020.

[62] O. Biza, D. Wang, R. W. Platt, J.-W. van de Meent, and L. L. S. Wong. Action priors for large action spaces in robotics. *ArXiv*, abs/2101.04178, 2021.

[63] S. Krishnan, R. Fox, I. Stoica, and K. Goldberg. DDCO: discovery of deep continuous options forrobot learning from demonstrations. *CoRR*, abs/1710.05421, 2017.

[64] R. Fox, R. Shin, S. Krishnan, K. Goldberg, D. Song, and I. Stoica. Parametrized hierarchical procedures for neural programming. In *International Conference on Learning Representations*, 2018.

[65] T. Yu, P. Abbeel, S. Levine, and C. Finn. One-shot hierarchical imitation learning of compound visuomotor tasks. *ArXiv*, abs/1810.11043, 2018.

[66] A. Gupta, V. Kumar, C. Lynch, S. Levine, and K. Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. In *CoRL*, 2019.

[67] A. Mandlekar, D. Xu, R. Martín-Martín, S. Savarese, and L. Fei-Fei. Learning to generalize across long-horizon tasks from human demonstrations. *ArXiv*, abs/2003.06085, 2020.

[68] X. Chen, Y. Gao, A. Ghadirzadeh, M. Bjorkman, G. Castellano, and P. Jensfelt. Skew-explore: Learn faster in continuous spaces with sparse rewards, 2020. URL https://openreview.net/forum?id=HJewxlHFwH.

[69] R. Hafner, T. Hertweck, P. Kloppner, M. Bloesch, M. Neunert, M. Wulfmeier, S. Tunyasuvunakool, N. Heess, and M. A. Riedmiller. Towards general and autonomous learning of core skills: A case study in locomotion. *ArXiv*, abs/2008.12228, 2020.

[70] D. Xu, S. Nair, Y. Zhu, J. Gao, A. Garg, L. Fei-Fei, and S. Savarese. Neural task programming: Learning to generalize across hierarchical tasks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3795–3802. IEEE, 2018.

[71] D.-A. Huang, S. Nair, D. Xu, Y. Zhu, A. Garg, L. Fei-Fei, S. Savarese, and J. C. Niebles. Neural task graphs: Generalizing to unseen tasks from a single video demonstration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8565–8574, 2019.

[72] M. Janner, S. Levine, W. Freeman, J. Tenenbaum, C. Finn, and J. Wu. Reasoning about physical interactions with object-oriented prediction and planning. *ArXiv*, abs/1812.10972, 2019.

[73] T. Kurutach, A. Tamar, G. Yang, S. J. Russell, and P. Abbeel. Learning plannable representations with causal infogan. *CoRR*, abs/1807.09341, 2018.

[74] A. Wang, T. Kurutach, K. Liu, P. Abbeel, and A. Tamar. Learning robotic manipulation through visual planning and acting. *CoRR*, abs/1905.04411, 2019.

[75] D. Jayaraman, F. Ebert, A. Efros, and S. Levine. Time-agnostic prediction: Predicting predictable video frames. In *International Conference on Learning Representations*, 2019.

[76] K. Liu, T. Kurutach, C. Tung, P. Abbeel, and A. Tamar. Hallucinative topological memory for zero-shot visual planning, 2020.

[77] B. Ichter, P. Sermanet, and C. Lynch. Broadly-exploring, local-policy trees for long-horizon task planning, 2020.

[78] K. Pertsch, O. Rybkin, F. Ebert, C. Finn, D. Jayaraman, and S. Levine. Long-horizon visual planning with goal-conditioned hierarchical predictors, 2020.

[79] L. P. Kaelbling and T. Lozano-Perez. Hierarchical task and motion planning in the now. In *IEEE Conference on Robotics and Automation (ICRA)*, 2011. Finalist, Best Manipulation Paper Award.

[80] 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)*, May 2014.

[81] OpenAI, I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang. Solving rubik's cube with a robot hand. *ArXiv*, abs/1910.07113, 2019.

[82] L. Kaelbling and T. Lozano-Perez. Learning composable models of parameterized skills. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 886–893, 2017.

[83] D. McDermott, M. Ghallab, A. Howe, C. Knoblock, A. Ram, M. Veloso, D. Weld, and D. Wilkins. Pddl–the planning domain definition language–version 1.2. *Yale Center for Computational Vision and Control, Tech. Rep. CVC TR-98-003/DCS TR-1165*, 1998.

[84] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.