

# Beyond Tabula-Rasa: a Modular Reinforcement Learning Approach for Physically Embedded 3D Sokoban

Peter Karkus<sup>1,2,\*</sup>, Mehdi Mirza<sup>2</sup>, Arthur Guez<sup>2</sup>, Andrew Jaegle<sup>2</sup>, Timothy Lillicrap<sup>2</sup>,  
Lars Buesing<sup>2</sup>, Nicolas Heess<sup>2</sup>, Theophane Weber<sup>2</sup>

<sup>1</sup>National University of Singapore, <sup>2</sup>DeepMind  
karkus@comp.nus.edu.sg

**Abstract:** Intelligent robots need to achieve abstract objectives using concrete, spatiotemporally complex sensory information and motor control. *Tabula rasa* deep reinforcement learning (RL) has tackled demanding tasks in terms of either visual, abstract, or physical reasoning, but solving these jointly remains a formidable challenge. One recent, unsolved benchmark task that integrates these challenges is Mujoban, where a robot needs to arrange 3D warehouses generated from 2D Sokoban puzzles. We explore whether integrated tasks like Mujoban can be solved by composing RL modules together in a *sense-plan-act* hierarchy, where modules have well-defined roles similarly to classic robot architectures. Unlike classic architectures that are typically model-based, we use only model-free modules trained with RL or supervised learning. We find that our modular RL approach dramatically outperforms the state-of-the-art monolithic RL agent on Mujoban. Further, learned modules can be reused when, e.g., using a different robot platform to solve the same task. Together our results give strong evidence for the importance of research into modular RL designs. Project website: <https://sites.google.com/view/modular-rl/>

**Keywords:** Hierarchical reinforcement learning, planning, partial observability

## 1 Introduction

Deep RL has successfully tackled tasks involving complex motor control [2, 3, 4], long-horizon abstract reasoning [5, 6, 7], as well as partial and visual observations [8, 9, 10, 11, 12]. But these challenges have largely been tackled in isolation. Moving forward, an important question is how to best utilize deep RL techniques for building real world intelligent robot systems – such as, for instance, a robot that cooks and cleans up. All the above challenges now need to be addressed *jointly* in a single system. Several integrated benchmark domains have recently been proposed to facilitate research in this direction [13, 1]. Unfortunately, deep RL results in these domains are currently very poor. For example, in the Mujoban domain [1] a robot needs to solve physically embedded Sokoban puzzles by pushing boxes in a maze given partial, first-person visual observations (Fig. 1). State-of-the-art RL can solve less than 10% of the trials even after weeks of training on TPUs and with access to privileged information such as a top-down view (see [1]).

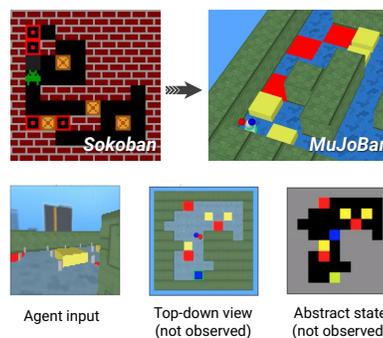


Figure 1: Mujoban is a challenging embodied task that embeds Sokoban puzzles in a physical 3D simulator [1]. The robot receives partial, first-person observations and has to push boxes (yellow) onto targets (red).

This paper investigates whether difficult embodied tasks can be approached by *composing* RL modules together similarly to classic robot architectures. Specifically, we focus on the Mujoban domain and

\*Work done at DeepMind.

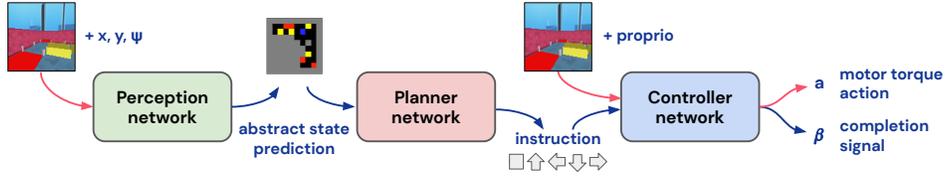


Figure 2: The proposed modular RL architecture composes separate modules for perception, planning, and control (each trained using model-free RL) to produce an overall policy.

explore a modular RL architecture with a sense-plan-act hierarchy (see Arkin 14, Ch. 4), where separate modules perform perception (“sense”), abstract reasoning (“plan”), and low-level motor control (“act”). In contrast to most hierarchical RL approaches [15] we carefully prescribe the role of each module in the hierarchy, aiming to build in structure specific to embodied tasks, as well as knowledge about useful abstractions. This modular design allows us to choose effective network structures and training regimes, and it also allows us to reuse trained modules, e.g., for different tasks or on different robot platforms. Modern robot architectures often combine model-based and learned components [4], but to the best of our knowledge, exclusively model-free RL modules in a sense-plan-act hierarchy have not been studied at the scale of integrated tasks like Mujoban.

Our modular RL architecture is shown in Fig. 2. The architecture consists of three modules, which are trained to optimize different objectives: a controller is trained for goal-oriented locomotion and pushing; a planner is trained for high-level abstract reasoning; and a perception network is trained to infer an abstract 2D state from sequences of first-person RGB inputs. We do not impose modular structure on the abstract reasoning problem itself, but learn it only from physical interactions, for which we introduce a simple time-abstracted RL algorithm based on MPO [16]. Modular RL improves the state-of-the-art in Mujoban by a large margin (from 9.4% to 78.7% success when using only first-person input). We also show that our learned abstract reasoning module transfers when the 2-DoF ball body is replaced with a 8-DoF ant robot. Our results reaffirm the benefit of prior knowledge in the form of modular structural biases and suggest the importance of research in this direction.

The specific contributions of this paper are as follows. We empirically demonstrate 1) that model-free RL modules can be composed in a sense-plan-act hierarchy and achieve state-of-the-art in Mujoban, an unsolved simulation domain that jointly exhibits important challenges of robot intelligence (partial visual observations, long-horizon planning, motor control); 2) that learned model-free RL modules for planning can be reused on different robot platforms. 3) We introduce a modular RL algorithm that trains hierarchical policies with time-abstraction.

## 2 Related work

The question of how to incorporate domain knowledge and structure into model-free deep RL has been the focus of much recent research. Domain knowledge may appear in various forms, including reward shaping [17, 18], training curricula [19, 20], auxiliary tasks [21], state representation [22], algorithmic structure [23, 24, 12] and memory [25, 26]. In this work we incorporate structure in the form of modularity, leveraging knowledge about embodied reasoning and abstractions in the target domain.

Hierarchical RL (HRL) has been the subject of research for many decades [27, 28, 29, 30, 31, 15, 32, 33]. Our modular RL approach is a form of hierarchical RL: we use a hierarchy of RL policies acting at different levels of abstractions and time scales. Importantly, while HRL typically aims to discover useful abstractions from environment interactions, we aim to utilize knowledge about task abstractions as a form of inductive bias, and thus define the role of each module and prescribe fixed interfaces. Learning abstractions provide more flexibility, which naturally comes at the cost of (often prohibitively) lower data efficiency. The wide-spread use of the sense-plan-act architecture in real-world robot system suggests that imposing such hierarchy might be a viable form of inductive bias for a large class of embodied problems.

Modern robot learning systems often combine hand-crafted and/or model-based modules with learned model-free modules. Recent examples include Andrychowicz et al. [4] on solving Rubik’s cubes with a fixed abstract planner and a learned RL controller; and Chaplot et al. [34] on navigation in

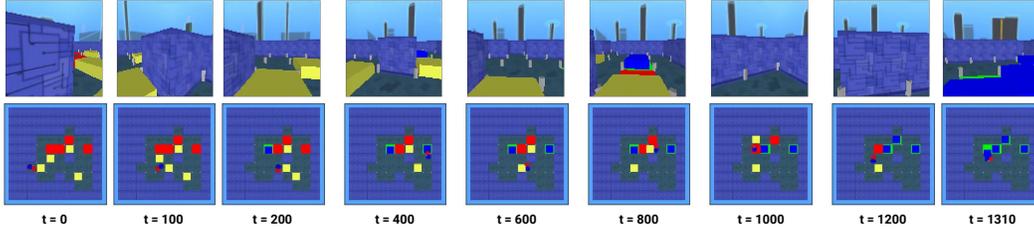


Figure 3: Modular RL trajectory in Mujoban. We show first-person inputs (top row) and unobserved top-down views (bottom row). Viewed from above, the robot appears as two red and blue discs. Boxes are yellow, target pads are red. When a box is pushed on a target its color changes to blue and the target pad turns green.

indoor spaces using a fixed path planner with learned mapping and control modules. In contrast, our modular RL approach uses model-free *learned* modules exclusively, and trains them with reward or supervised signals. Importantly, we learn abstract planning exclusively from physical interactions with the environment, without accessing a model or simulator of the abstract problem.

Finally, the Differentiable Algorithm Network (DAN) of Karkus et al. [35] combines a sense-plan-act hierarchical structure with end-to-end learning. The DAN composes differentiable structures for perception (vision + filtering), planning, and control, and trains them jointly for partially observed map-based navigation tasks. The use of modules with well-defined roles in a hierarchy is similar in our work, however, our modules do not encode task-specific algorithmic structure, and our modules are trained with rewards and supervised signals instead of end-to-end expert demonstrations.

We provide a more extensive discussion on modular structures and RL in Appendix D.

### 3 Mujoban

We are interested in integrated embodied tasks that jointly exhibit challenges of visual perception, abstract reasoning and motor control. We choose Mujoban [1], a recently proposed benchmark task that embeds Sokoban puzzles ([36, 37, 38]) in the Mujoco simulator [39].

Mujoban generates 3D mazes from 2D Sokoban puzzles with randomized visual appearance. A robot navigates the maze and pushes boxes onto target areas with its body. The default robot is a 2-DoF ball, but we also experiment with a 8-DoF ant. We use the harder "peg" version of Mujoban, where small pegs are inserted at grid points to enforce Sokoban rules (boxes cannot be pushed diagonally and cannot be recovered when next to a wall or another box). The robot receives partial observations: first-person camera images  $o^{\text{cam}}$ , proprioceptive signals  $o^{\text{PF}}$  (including touch, position, velocity, and acceleration sensors), and global pose  $o^{\text{pos}} = (x, y, \psi)$ . Rewards are +10 for solving the puzzle, and +1/-1 for pushes a box on/off a target area. The challenges of this task are three-fold:

- **Motor control.** The agent needs to learn locomotion ("crawling" in case of the ant) to reach targets in the maze. It must also push and carefully align boxes, such that they do not get stuck among grid pegs when pushed from another direction. We found this manipulation problem to be challenging on its own.
- **Abstract reasoning / planning.** Sokoban puzzles, by design, require long-horizon planning. Boxes can only be pushed but not pulled, so many moves are irreversible, making it important to plan ahead.
- **Perception.** Despite the simplistic visual appearance of Mujoban, perception involves non-trivial visual mapping: to infer the underlying 2D Sokoban state, information needs to be integrated from sequences of RGB images showing only parts of the 3D maze.

Importantly, different challenges compound: e.g., a poor box alignment may prevent executing a high-level strategy, and imperfect perception may lead to invalid plans with wrong irreversible moves.

In contrast to [1], and motivated by real-world applications, we target a more realistic partially observable setting of Mujoban. Most importantly, we only provide first-person observations as opposed to first-person and top-down inputs in [1]. As is common in robot learning we utilize the simulator more explicitly for training. That is, we access the full physical state to compute rewards, and abstract states  $S^*$  which serve as supervisory signals.  $S^*$  captures the underlying Sokoban state in  $N \times N \times 4$  binary image (the same  $S^*$  is input to some agents in [1]). Finally, we assume the robot

pose can be observed for the purpose of mapping, but importantly, we do not observe the environment layout and box poses directly at test time.

## 4 Modular Reinforcement Learning

### 4.1 Overview

We propose a modular RL approach for Mujoban that decouples layers of embodied reasoning into distinct RL modules: a perception module for spatial reasoning, a planner for abstract reasoning, and controller for motor control. We prescribe different objectives to each module, leveraging domain knowledge both in terms of generic properties of embodiment as well as task-specific abstractions.

Our modular RL architecture is shown in Fig. 2. The perception module integrates first-person visual information  $o$  into an abstract 2D state  $S$ , similar to a 2D map. The planner<sup>2</sup> takes in the abstract state and outputs an instruction  $A$ . We design instructions (move north, east, south, west, or stay) knowing that moving and pushing boxes between cells is sufficient to succeed in the task. Finally, the controller is a goal-oriented visuomotor policy for locomotion and object manipulation (pushing boxes). It takes in observations and an instruction and outputs motor torque actions  $a$ . We also predict if the last instruction was completed ( $\beta$ ) and use it for time-abstraction in the planner. Modules are trained in different regimes for different objectives, but they rely on each other to collect meaningful experience. Pseudo-code for training modules is given in Algorithm 1.

Next we introduce each module and the training algorithm. Additional details are in Appendix B.

### 4.2 Controller module

The controller is a goal-oriented RL policy for locomotion and object manipulation. The controller also predicts if the last instruction was completed, similarly to the option termination of HRL [28, 40]. We use separate networks for the control policy and the completion predictor.

The control policy takes an instruction  $A_t$ , visual observation  $o_t^{\text{cam}}$ , proprioceptive observations  $o_t^{\text{pr}}$ , and the (predicted) completion signal  $\beta_t$ , and outputs continuous motor torque actions  $a_t$ . Instructions are one-hot encodings of discrete high-level moves. We use a ResNet [41] connected to an LSTM [42] and train with the actor-critic MPO algorithm [16]. We use an asymmetric setup where the critic receives privileged side information during training ( $o_t^{\text{pos}}$  and box poses  $o_t^{\text{box}}$ ). This training setup is common for robot learning [43, 4].

We define the following rewards for the controller: +4 if the last instruction is completed and a small negative reward if the robot moves backwards. If the instruction is not completed within  $T_{instr} = 120$  steps the agent receives a reward of -5 and the episode terminates. An instruction is defined to be completed only if the robot is near the center of the grid cell for the instructed move, and if pushing a box the box is near the center of the adjacent cell.

The controller module also outputs a completion signal  $\beta_t$ , a prediction whether the last instruction was completed. We train a supervised sequential binary classifier, a ResNet-LSTM network independent from the policy network. We obtain training labels given the control rewards defined above. During evaluation a positive completion signal is added if there has been no completion predicted for  $T_{instr} = 120$  steps.

### 4.3 Planner module

The planner is a model-free RL policy responsible for solving the puzzle that underlies the physical environment. Inputs to the planner are abstract state estimates,  $S_t$ , i.e.,  $10 \times 10 \times 4$  images capturing the 2D maze (see Fig. 1 for an example). Outputs are discrete instructions,  $A_t$ , corresponding to moves between cells of the maze. The policy acts in abstract time, that is, in each step of the planner the robot may interact with the environment for a variable number of time steps until the instruction is (predicted to be) completed.

We use a policy network similar to the repeated ConvLSTM architecture of Guez et al. [38]. The network is built of generic components, but it has strong structural bias for 2D planning. We use a

<sup>2</sup>The term *planner* refers to the role of the model-free RL policy, not a model-based planning algorithm.

feed-forward variant of the architecture with an additional critic head that outputs Q-value estimates for each discrete instruction. The planner policy is trained with RL, but from the planner’s perspective it interacts with an abstract version of the environment (through the learned controller) that resembles the symbolic Sokoban domain. Rewards for the planner are task rewards accumulated during the execution of the instruction. The training algorithm is given in Algorithm 1 and it is further discussed in Section 4.5.

#### 4.4 Perception module

The perception module sequentially predicts the abstract state  $S_t$  from first-person visual observations  $o_t^{\text{cam}}$  and the agent poses  $o_t^{\text{pos}}$ . The task is similar to mapping, where  $S_t$  is equivalent to a 2D map of the 3D scene. We design a geometry-aware network architecture that builds in strong inductive bias for 2D mapping from first-person observations. The network is a ConvLSTM [44] with a novel spatial attention mechanism shown in Fig. 4. The attention mechanism was inspired by [45] but it is adapted to our 2D mapping task. Intuitively, the network works by attending to different elements of the visual input for each possible location of the 2D map ( $S_t$ ). Spatial attention produces features arranged in a 2D grid. The features are fed to a ConvLSTM, forming an implicit belief over abstract states. The LSTM output is fed to a classifier with 5 labels (None, Wall, Box, Target, Box-on-target), which predicts each spatial location of the abstract state separately. We train the network with independent cross-entropy losses, where labels are abstract state observations  $S_t^*$  from the simulator. The geometry-aware perception network currently relies on observing the agent’s pose, but it could be also predicted from visual and proprioceptive observations in the future. The network architecture could be also used for 2D mapping tasks beyond Mujoban in future work.

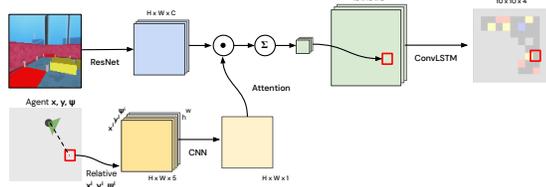


Figure 4: Geometry-aware perception network.

#### 4.5 Training algorithm

---

**Algorithm 1** Modular RL pseudo code for training modules with time abstraction

---

1:	Initialize <b>Perception</b> , <b>Planner</b> , <b>Controller</b> and replay buffers $B^{\text{abst}}$ , $B^{\text{phy}}$	
2:	<b>while</b> not converged <b>do</b>	
3:	$o_t \leftarrow \text{Reset}()$	▷ reset environment
4:	<b>while</b> not <i>terminal</i> <b>do</b>	▷ repeat until episode terminates
5:	$S \leftarrow \text{Perception}(o_t)$	▷ estimate abstract state
6:	$A \leftarrow \text{Planner}(S)$	▷ sample instruction from planner
7:	<i>// execute instruction through multiple environment steps</i>	
8:	$R \leftarrow 0; \beta_t \leftarrow \text{False}; t \leftarrow 0$	
9:	<b>while</b> not $\beta_t$ and $t \leq T_{\text{instr}}$ and not <i>terminal</i> <b>do</b>	
10:	$a_t, \beta_t \leftarrow \text{Controller}(o_t, A_T)$	
11:	$o_{t+1}, r_t, \text{terminal} \leftarrow \text{Execute}(a_t)$	▷ execute action
12:	$R \leftarrow R + r_t$	▷ accumulate task reward
13:	AddToReplay( $B^{\text{phy}}$ , ( $o_t, a_t, r_t, o_{t+1}, \text{terminal}$ ))	
14:	<i>terminal</i> $\leftarrow \text{terminal}$ or $t > T_{\text{instr}}$	▷ early termination
15:	AddToReplay( $B^{\text{abst}}$ , ( $S, A, R, \text{Perception}(o_{t+1}), \text{terminal}$ ))	
16:	<i>// update modules</i>	
17:	UpdateRL( <b>Controller</b> , $B^{\text{phy}}$ )	▷ uses MPO [16]
18:	UpdateRL( <b>Planner</b> , $B^{\text{abst}}$ )	▷ time-abstracted update, uses MPO [16]
19:	UpdateSupervised( <b>Perception</b> , $B^{\text{phy}}$ )	

---

Algorithm 1 provides a pseudo-code for training modules in our architecture. The algorithm addresses the general problem of learning abstract high-level reasoning from physical interactions by collecting time-abstracted experience with a low-level controller that predicts its completion ( $\beta$ ). We define rewards for an abstract step as the sum of environment rewards collected during the execution of the instruction (line 12). Failure to complete an instruction within a time limit is treated as an early-termination of the episode (line 14). The algorithm combines RL and supervised updates (lines

	Percep.	Planner	Contr.	Rewards		Success rates	
				hard	easy	med.	hard
Modular RL	-	<i>oracle</i>	learned	13.3 (0.1)	100%	97.9%	<b>94.3%</b>
Modular RL	<i>true</i>	learned	learned	11.8 (0.2)	100%	90.0%	<b>81.8%</b>
Modular RL	learned	learned	learned	11.4 (0.2)	90.2%	78.7%	<b>78.7%</b>

Table 1: Modular RL results with different set of modules.

17, 18 and 19), and thus allows for a pragmatic approach where supervising learning is used where a supervised signal is easy to derive, and RL is used everywhere else.

We omitted a number of task-specific implementation details from Algorithm 1 for clarity. We include side information ( $o^{\text{box}}, S^*$ ) in  $o$  during training which is used to compute controller rewards and supervised losses. We use distributed RL where experience collection and module updates are separated into different nodes. Further, while joint training would be possible, we only train one module at a time in a bottom-up sequence (controller, planner, perception), where untrained modules are replaced with ground-truth or random inputs. Further details are in Appendix B.4.

All models are implemented in Tensorflow [46] and trained with the Adam optimizer [47]. We use one Nvidia Tesla V100 GPU for learner nodes and CPU only for actor nodes. We train models until near convergence, which took up to 7, 21 and 5 days for modular RL modules, fully-observable baselines and partially-observable baselines, respectively.

## 5 Results

We present results for modular RL (Section 5.1) followed by comparison results with alternative learning methods (Section 5.2) and module transfer results for the Mujoco ant (Section 5.3).

### 5.1 Modular RL results

Table 1 evaluates trained modules in different configurations: controller only with instructions from an oracle (first row); planner and controller with perfect abstract state input (second row); using all learned modules (third row). We report success rates in 512 random episodes for different difficulty levels: easy ( $5 \times 5$  mazes with 1 box), medium ( $8 \times 8$  mazes with 3 boxes), and hard ( $10 \times 10$  mazes with 4 boxes). An episode is successful if the puzzle is solved within 4800 steps (240s). All modules are trained using the default hard level only. Rewards (with standard deviations) are reported for the default difficulty (hard). We observe strong performance, 78.7% success in the final setting using all modules (last row). The performance gap without perception is relatively small (78.7% to 81.8%), while the performance gap with and without planning is larger (81.8% to 94.3%).

We report learning curves for each module in Fig. 5. We plot the relevant objective metric (rewards or accuracy) against environment steps sampled from the replay buffer. We selected results for the best hyper-parameter and random seed (out of 2). We continued training for a large number of steps, but the rate of improvements indicate that good performance can be achieved with less training as well.

Fig. 3 shows a successful trajectory. Videos are available at <https://sites.google.com/view/modular-rl/>. We observe that generally the modular agent quickly builds good abstract representations, moves around and pushes boxes, and solves many levels that are hard (even for a human player with top-down view). Failures can be attributed to imperfect abstract states (leading to poor irreversible moves); the planner failing to solve the underlying Sokoban puzzle; and the controller failing an instruction, typically when grid pegs block the motion of a previously poorly aligned box.

Modular RL results are substantially stronger than the best known RL agents for Mujoco (Table 4). Note that success rates are not directly comparable because these agents receive additional inputs (top-down camera, abstract state, and expert instructions) and are evaluated with a different timeout (45s instead of 240s). The rows of Table 4 correspond to “Vanilla Agent”, “Random Planner” and “Expert Planner” of “with grid pegs” results of Mirza et al. [1]; and easy, medium and hard levels correspond to level 1, 4, and 5, respectively.

Method	Rewards		Success rates		
	hard	easy	med.	hard	
Modular RL (ours)	<b>11.4</b> (0.2)	<b>90.2%</b>	<b>78.7%</b>	<b>78.7%</b>	
Monolithic RL	0.3 (0.0)	25.0%	0.4%	0.0%	
Monolithic RL +shared	<b>1.6</b> (0.1)	51.6%	<b>6.8%</b>	<b>1.5%</b>	
Monolithic RL +curr	0.3 (0.0)	56.3%	0.0%	0.0%	
Monolithic RL +shared +curr	0.8 (0.0)	69.7%	3.9%	0.2%	
Monolithic RL +easy-only	0.6 (0.1)	<b>99.0%</b>	1.6%	0.4%	

Table 2: Comparison results with monolithic RL methods.

Method	Rewards		Success rates		
	hard	easy	med.	hard	
Monolithic RL +top-down	5.1 (0.2)	100%	51.0%	<b>27.9%</b>	
Structured RL +top-down + $S^*$	11.2 (0.2)	100%	88.7%	<b>77.0%</b>	

Table 3: RL results with top-down input.

Extra inputs	Success rates		
	easy	med.	hard
top-down	99.8%	31.5%	<b>9.4%</b>
top-down, $S^*$	75.2%	42.0%	<b>30.2%</b>
top-down, $S^*$ , expert $A$	100%	78.2%	<b>54.6%</b>

Table 4: Best known RL results from [1] (SOTA)

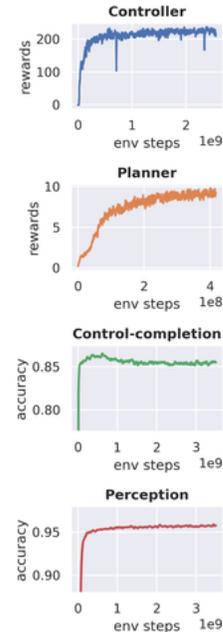


Figure 5: Module learning curves.

## 5.2 Comparison results

**Monolithic RL.** We first compare modular RL with monolithic RL methods in Table 2. We use standard ResNet-LSTM policy networks trained with MPO and the default task rewards. We use the same partially observable setting as for modular RL and the same asymmetric actor-critic setup where the critic receives privileged side information. In different rows of Table 2 we try a symmetric actor-critic with a shared torso that does not use side information (+shared); training with a curriculum (+curr), or training only on easy levels (+easy-only). We searched over network architectures of 3 significantly different sizes, learning rates, and MPO parameters. We report only the best performing combinations. Results indicate that monolithic RL can learn almost none but the easiest levels. Further, without modularity, asymmetric actor-critic cannot benefit from privileged observations; and curriculum helps on easy levels but not on harder ones.

**Structured and hierarchical RL** Prior work proposed various inductive structures for RL that could be useful in this domain, e.g., for mapping [25, 34, 48], or for state estimation from partial observations [49, 12]. Instead of direct comparisons we run experiments with additional top-down observations that *remove* partial observability and the need for mapping, thus results here can be expected to upper-bound RL approaches with specific structure for mapping or filtering. For these experiments we follow Mirza et al. [1] and use the same network structure, V-trace algorithm, and curriculum. While these choices are orthogonal, we did not try all combinations due to the substantial required computation time.

Table 3 reports results in the simplified top-down setting of Mujoban. Monolithic RL can only solve 27.9% of the hard episodes (first row), which suggests that without top-down input, RL with specific structure for mapping or filtering would not be effective. Next we further simplify the task (second row). We add direct observations of the abstract state and structured exploration, where extra pseudo rewards are given for reaching an adjacent abstract state, similarly to the “Random planner” in [1]. This structured agent performs significantly better, but still worse than our modular approach (77.0% with additional top-down input vs. 81.8% for modular RL). Interestingly, this agent learned to exploit shortcuts in Mujoban, such as recovering boxes pushed against a wall and push multiple boxes together. This suggests that in a stricter version of the task the benefit of the modular approach would be even larger.



Figure 6: Mujoco ant in Mujoban.

Method	Robot	Success rates		
		easy	med.	hard
Modular RL	ball (default)	100%	90.0%	<b>81.8%</b>
Modular RL	Mujoco ant	99.6%	83.1%	<b>73.7%</b>

Table 5: Module transfer results for the Mujoco ant.

To better understand the benefit of prescribed module objectives, future work may compare with general hierarchical RL methods, such as the Option-Critic [15]. We found that carefully designing objectives, e.g., for the controller to move to the center of cells had a large impact on overall performance. Learning this from reinforcements alone would likely be an additional significant challenge for general HRL methods. Further, knowing the abstract state of the puzzle is strictly necessary in many cases for Mujoban, but in early experiments we found that an unstructured LSTM was unable to learn the perception task (even in a supervised setting). This suggests that learning from much more indirect reinforcements, hierarchical RL alone (without task-specific structure or top-down observations) is unlikely to perform well in Mujoban.

**Model-based planning and control** A model-based equivalent of our approach could be considered for Mujoban with classic perception, planning and control modules. However, even with hard-coded rules and known abstract states, planning for Sokoban (e.g. with MCTS) is known to be computationally expensive [38]. Further, when the Sokoban state input to the planner is inferred from partial observations it can be (unavoidably) incomplete or unsolvable, for which planning is undefined. Finally, while building a classic controller for simple dynamics, such as for the ball robot, would be certainly possible, it would require significant effort for more complex dynamics, e.g. for the Mujoco ant, especially because of the need to push and align boxes with precision.

### 5.3 Module transfer: Mujoban with the Mujoco ant

Next, we want to validate if RL modules can be similarly reusable as modules of classic robot systems. We consider a scenario where a different robot needs to solve the same task. We replace the default ball robot of Mujoban with the Mujoco ant (Fig. 6). We scaled the physical size of the environment by a factor of three to account for the larger size of the ant, while keeping the weight of boxes and the size of grid pegs unchanged. To focus on the significantly increased control complexity we use the fully-observable Mujoban setting where abstract states are observed, and thus we only need the planner and controller modules. We also increase evaluation time limits to 480s to account for the larger environment and slower motion of the ant.

We train a new controller module for the ant, similarly as for the ball before. We then compose a policy from the ant controller and the old planner module that was trained with the ball robot. Results are in Table 5. Despite the significantly more challenging motor control problem, success rates with the ant are only slightly lower than with the ball body (73.7% vs. 81.8%). This suggests that RL modules can be indeed reusable. Such ability to generalize and reuse parts of the solution is an important benefit over monolithic approaches.

## 6 Discussion & Conclusions

Our results show that coupling RL with an appropriately designed modular learning architecture can lead to progress on domains that are off-puttingly difficult for monolithic approaches. Given enough capacity, computation and experience, a monolithic end-to-end architecture could of course learn to solve Mujoban. But as our baselines show, for tasks that combine perception, abstract reasoning, and motor control, these costs may be prohibitive and make research progress infeasible. Modular architectures can facilitate the search for the correct primitives, even in the face of intractable RL tasks. The interpretability of modules allow researchers to make more informed design interventions and incorporate learning signals both more flexibly and more directly.

Modularity is not without its pitfalls: a mis- or over-specified modular structure may lead to unnecessary overhead for learning and may even prevent the agent from finding the optimal solution. This pitfall can potentially be avoided by designing the modules to be as general as possible and by allowing joint refinement. For embodied tasks, we argue that inherent structural properties can be

duly exploited: the world is spatially organized, contains persistent objects, and the agent perceives and acts locally from a single physical location. In this spirit, future work should continue to probe the boundary between module engineering and monolithic RL to characterize the best trade-offs between domain-specific engineering and a tabula rasa design.

The modular approach was successful in Mujoban, perhaps because layers of reasoning can be well separated in this domain. Going forward, we want to investigate the same modular philosophy in scenarios with blurrier decompositions, such as assembling furniture [13], where independent training could be followed by joint fine-tuning with respect to the final task objective to compensate for potentially incorrect assumptions. Recent results in this direction are encouraging [35].

## 7 Acknowledgements

We thank Kenneth Chaney, Bernd Pfrommer, and Kostas Daniilidis for a helpful discussion of the classical robotics literature.

## References

- [1] M. Mirza, A. Jaegle, J. J. Hunt, A. Guez, S. Tunyasuvunakool, A. Muldal, T. Weber, P. Karkus, S. Racanière, L. Buesing, T. Lillicrap, and N. Heess. Physically embedded planning problems: New challenges for reinforcement learning. *arXiv preprint arXiv:2009.05524*, 2020.
- [2] N. Heess, D. TB, S. Sriram, J. Lemmon, J. Merel, G. Wayne, et al. Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*, 2017.
- [3] A. Gupta, V. Kumar, C. Lynch, S. Levine, and K. Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2019.
- [4] M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.
- [5] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
- [6] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dkebiak, C. Dennison, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.
- [7] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7587):350–354, 2019.
- [8] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.
- [9] F. Sadeghi and S. Levine. CAD<sup>2</sup>RL: Real single-image flight without a single real image. In *Robotics: Science and Systems (RSS)*, 2017.
- [10] K. Gregor, D. J. Rezende, F. Besse, Y. Wu, H. Merzic, and A. van den Oord. Shaping belief states with generative environment models for RL. In *Advances in Neural Information Processing Systems*, 2019.
- [11] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra. DD-PPO: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *International Conference on Learning Representations*, 2020.
- [12] X. Ma, P. Karkus, D. Hsu, W. S. Lee, and N. Ye. Discriminative particle filter reinforcement learning for complex partial observations. In *International Conference on Learning Representations ICLR*, 2020.
- [13] Y. Lee, E. S. Hu, Z. Yang, A. Yin, and J. J. Lim. IKEA furniture assembly environment for long-horizon complex manipulation tasks. *arXiv preprint arXiv:1911.07246*, 2019.
- [14] R. C. Arkin. *Behavior-Based Robotics*. MIT Press, 1998.
- [15] P.-L. Bacon, J. Harb, and D. Precup. The option-critic architecture. In *AAAI Conference on Artificial Intelligence*, 2017.
- [16] A. Abdolmaleki, J. T. Springenberg, Y. Tassa, R. Munos, N. Heess, and M. Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018.
- [17] A. Y. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning*, 1999.
- [18] I. Popov, N. Heess, T. Lillicrap, R. Hafner, G. Barth-Maron, M. Vecerik, T. Lampe, Y. Tassa, T. Erez, and M. Riedmiller. Data-efficient deep reinforcement learning for dexterous manipulation. *arXiv preprint arXiv:1704.03073*, 2017.

- [19] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *International Conference on Machine Learning*, 2009.
- [20] C. Florensa, D. Held, M. Wulfmeier, M. Zhang, and P. Abbeel. Reverse curriculum generation for reinforcement learning. In *Conference on Robot Learning*, pages 482–495, 2017.
- [21] M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017.
- [22] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [23] A. Tamar, Y. Wu, G. Thomas, S. Levine, and P. Abbeel. Value iteration networks. In *Advances in Neural Information Processing Systems*, 2016.
- [24] G. Farquhar, T. Rocktäschel, M. Igl, and S. Whiteson. TreeQN and ATreeC: Differentiable tree planning for deep reinforcement learning. In *International Conference on Learning Representations*, 2018.
- [25] E. Parisotto and R. Salakhutdinov. Neural map: Structured memory for deep reinforcement learning. In *International Conference on Learning Representations*, 2018.
- [26] M. Fortunato, M. Tan, R. Faulkner, S. Hansen, A. P. Badia, G. Buttimore, C. Deck, J. Z. Leibo, and C. Blundell. Generalization of reinforcement learners with working and episodic memory. In *Advances in Neural Information Processing Systems*, pages 12469–12478, 2019.
- [27] P. Dayan and G. E. Hinton. Feudal reinforcement learning. In *Advances in Neural Information Processing Systems*, 1993.
- [28] R. S. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181 – 211, 1999.
- [29] K. Gregor, D. J. Rezende, and D. Wierstra. Variational intrinsic control. In *International Conference on Learning Representations*, 2017.
- [30] N. Heess, G. Wayne, Y. Tassa, T. P. Lillicrap, M. A. Riedmiller, and D. Silver. Learning and transfer of modulated locomotor controllers. *arXiv preprint arXiv: 1610.05182*, 2016.
- [31] A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu. FeUdal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning*, 2017.
- [32] J. Merel, A. Ahuja, V. Pham, S. Tunyasuvunakool, S. Liu, D. Tirumala, N. Heess, and G. Wayne. Hierarchical visuomotor control of humanoids. In *International Conference on Learning Representations*, 2019.
- [33] J. Merel, S. Tunyasuvunakool, A. Ahuja, Y. Tassa, L. Hasenclever, V. Pham, et al. Catch & carry: Reusable neural controllers for vision-guided whole-body tasks. *arXiv preprint arXiv: 1911.06636*, 2019.
- [34] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov. Learning to explore using active neural slam. In *International Conference on Learning Representations*, 2020.
- [35] P. Karkus, X. Ma, D. Hsu, L. P. Kaelbling, W. S. Lee, and T. Lozano-Pérez. Differentiable algorithm networks for composable robot learning. In *Robotics: Science and Systems (RSS)*, 2019.
- [36] A. Junghanns and J. Schaeffer. Sokoban: A challenging single-agent search problem. In *IJCAI Workshop on Using Games as an Experimental Testbed for AI Research*, 1997.
- [37] S. Racanière, T. Weber, D. Reichert, L. Buesing, A. Guez, D. J. Rezende, A. P. Badia, O. Vinyals, N. Heess, Y. Li, et al. Imagination-augmented agents for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 2017.
- [38] A. Guez, M. Mirza, K. Gregor, R. Kabra, S. Racanière, T. Weber, et al. An investigation of model-free planning. In *International Conference on Machine Learning*, 2019.
- [39] E. Todorov, T. Erez, and Y. Tassa. MuJoCo: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems (IROS)*, 2012.
- [40] A. Barreto, D. Borsa, S. Hou, G. Comanici, E. Aygün, P. Hamel, D. Toyama, S. Mourad, D. Silver, D. Precup, et al. The option keyboard: Combining skills in reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 13052–13062, 2019.
- [41] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016.
- [42] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [43] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, and P. Abbeel. Asymmetric actor critic for image-based robot learning. In *Robotics: Science and Systems (RSS)*, 2017.

- [44] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, 2015.
- [45] P. Karkus, A. Angelova, V. Vanhoucke, and R. Jonschkowski. Differentiable mapping networks: Learning structured map representations for sparse visual localization. In *International Conference on Robotics and Automation (ICRA)*, 2020.
- [46] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.
- [47] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [48] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik. Cognitive mapping and planning for visual navigation. *arXiv preprint arXiv:1702.03920*, 2017.
- [49] M. Igl, L. Zintgraf, T. A. Le, F. Wood, and S. Whiteson. Deep variational reinforcement learning for pomdps. *arXiv preprint arXiv:1806.02426*, 2018.
- [50] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, et al. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, 2018.
- [51] N. J. Nilsson. *Principles of Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1980.
- [52] R. A. Brooks. A robust layered control system for a mobile robot. *AI Memo 864, Artificial Intelligence Lab MIT*, 1985.
- [53] E. Gat. On three-layer architectures. In *Artificial Intelligence and Mobile Robots*, 1997.
- [54] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, et al. Stanley: The robot that won the DARPA Grand Challenge. *Journal of Field Robotics*, 23(9):661–692, 2006.
- [55] J. Lim, I. Lee, I. Shim, H. Jung, H. M. Joe, H. Bae, et al. Robot system of DRC-HUBO+ and control strategy of team KAIST in DARPA robotics challenge finals. *Journal of Field Robotics*, 34(4), 2017.
- [56] K. Mohta, K. Sun, S. Liu, M. Watterson, B. Pfrommer, J. Svacha, Y. Mulgaonkar, C. J. Taylor, and V. Kumar. Experiments in fast, autonomous, GPS-denied quadrotor flight. In *International Conference on Robotics and Automation (ICRA)*, 2018.
- [57] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.
- [58] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, et al. A general reinforcement learning algorithm that masters Chess, Shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.
- [59] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, et al. Mastering Atari, Go, Chess and Shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019.
- [60] N. Heess, J. J. Hunt, T. P. Lillicrap, and D. Silver. Memory-based control with recurrent neural networks. In *NeurIPS Deep Reinforcement Learning Workshop*, 2015.
- [61] M. Jaderberg, V. Mnih, W. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017.
- [62] G. Wayne, C.-C. Hung, D. Amos, M. Mirza, A. Ahuja, A. Grabska-Barwinska, et al. Unsupervised predictive memory in a goal-directed agent. *arXiv preprint arXiv:1803.10760*, 2018.
- [63] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castaneda, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, 2019.
- [64] B. Amos, I. Jimenez, J. Sacks, B. Boots, and J. Z. Kolter. Differentiable MPC for end-to-end planning and control. In *Advances in Neural Information Processing Systems*, pages 8299–8310, 2018.
- [65] P. Bacon, J. Harb, and D. Precup. The option-critic architecture. In *AAAI Conference on Artificial Intelligence*, 2017.
- [66] D. Ha and J. Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, 2018.
- [67] L. Kaiser, M. Babaeizadeh, P. Milos, B. Osinski, R. H. Campbell, K. Czechowski, et al. Model-based reinforcement learning for Atari. In *International Conference on Learning Representations*, 2020.
- [68] J. B. Hamrick, A. J. Ballard, R. Pascanu, O. Vinyals, N. Heess, and P. W. Battaglia. Metacontrol for adaptive imagination-based optimization. In *International Conference on Learning Representations*, 2017.

- [69] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2019.
- [70] A. Faust, K. Oslund, O. Ramirez, A. Francis, L. Tapia, M. Fiser, and J. Davidson. PRM-RL: Long-range robotic navigation tasks by combining reinforcement learning and sampling-based planning. In *International Conference on Robotics and Automation (ICRA)*, 2018.
- [71] D. Silver, H. van Hasselt, M. Hessel, T. Schaul, A. Guez, T. Harley, et al. The predictron: End-to-end learning and planning. In *International Conference on Machine Learning*, 2017.
- [72] J. Oh, S. Singh, and H. Lee. Value prediction network. In *Advances in Neural Information Processing Systems*, 2017.
- [73] C. Finn and S. Levine. Deep visual foresight for planning robot motion. In *International Conference on Robotics and Automation (ICRA)*, 2017.
- [74] C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, and P. Abbeel. Deep spatial autoencoders for visuomotor learning. In *International Conference on Robotics and Automation (ICRA)*, 2016.
- [75] T. D. Kulkarni, A. Gupta, C. Ionescu, S. Borgeaud, M. Reynolds, A. Zisserman, and V. Mnih. Unsupervised learning of object keypoints for perception and control. In *Advances in Neural Information Processing Systems*, 2019.
- [76] N. Watters, L. Matthey, M. Bosnjak, C. P. Burgess, and A. Lerchner. COBRA: Data-efficient model-based rl through unsupervised object discovery and curiosity-driven exploration. *arXiv preprint arXiv:1905.09275*, 2019.
- [77] N. Fazeli, M. Oller, J. Wu, Z. Wu, J. B. Tenenbaum, and A. Rodriguez. See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion. *Science Robotics*, 4(26), 2019.
- [78] M. Janner, S. Levine, W. T. Freeman, J. B. Tenenbaum, C. Finn, and J. Wu. Reasoning about physical interactions with object-oriented prediction and planning. In *International Conference on Learning Representations*, 2019.
- [79] R. Veerapaneni, J. D. Co-Reyes, M. Chang, M. Janner, C. Finn, J. Wu, J. B. Tenenbaum, and S. Levine. Entity abstraction in visual model-based reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2019.
- [80] O. Nachum, S. S. Gu, H. Lee, and S. Levine. Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, 2018.
- [81] A. Zhang, A. Lerer, S. Sukhbaatar, R. Fergus, and A. Szlam. Composable planning with attributes. *arXiv preprint arXiv:1803.00512*, 2018.
- [82] F. Alet, T. Lozano-Pérez, and L. P. Kaelbling. Modular meta-learning. In *Conference on Robot Learning (CoRL)*, 2018.
- [83] B. Quack, F. Wörgötter, and A. Agostini. Simultaneously learning at different levels of abstraction. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 4600–4607, 2015.
- [84] G. Konidaris, L. P. Kaelbling, and T. Lozano-Perez. From skills to symbols: Learning symbolic representations for abstract high-level planning. *Journal of Artificial Intelligence Research*, 61:215–289, 2018.
- [85] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *Conference on Computer Vision and Pattern Recognition*, 2016.
- [86] N. J. Nilsson. Shakey the robot. Technical report, SRI AI Center Menlo Park CA, 1984.
- [87] L. P. Kaelbling and T. Lozano-Pérez. Integrated task and motion planning in belief space. *The International Journal of Robotics Research*, 32(9-10):1194–1227, 2013.
- [88] F. Lagriffoul, N. T. Dantam, C. Garrett, A. Akbari, S. Srivastava, and L. E. Kavraki. Platform-independent benchmarks for task and motion planning. *IEEE Robotics and Automation Letters*, 3(4):3765–3772, 2018.
- [89] M. A. Toussaint, K. R. Allen, K. A. Smith, and J. B. Tenenbaum. Differentiable physics and stable modes for tool-use and manipulation planning. In *Robotics: Science and Systems (RSS)*, 2018.
- [90] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, et al. Habitat: A platform for embodied AI research. In *International Conference on Computer Vision (ICCV)*, 2019.
- [91] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In *Conference on Robot Learning (CoRL)*, 2017.

## A Mujoban domain

**Overview.** Mujoban [1] is a simulation domain that embeds Sokoban puzzles in Mujoco simulation environment [39]. Sokoban is a popular RL benchmark with complex planning [36, 37, 38], where the agent pushes boxes onto target locations in procedurally generated 2D grids. Mujoban generates 3D maze equivalents of Sokoban levels with varied visual appearance. An embodied agent navigates the maze given partial visual observations, and manipulates boxes using its physical body.

**Simulator.** The Mujoban domain is visualized in Fig. 1. In the default configuration the agent has a 2-DoF ball body; and it receives partial observations in the form of first-person camera images  $o^{\text{cam}}$ , standard proprioceptive observations  $o^{\text{pr}}$ , and the agent’s absolute pose  $o^{\text{pos}} = (x, y, \psi)$ . The simulator also provides access to additional observations, e.g. for training: top-down camera images  $o^{\text{top}}$ , absolute pose of boxes  $o^{\text{box}}$ , and abstract state  $S$  that represents the underlying Sokoban state as a  $N \times N \times 4$  binary image, where image channel correspond to the presence of wall, target, agent, and box in each of the grid locations.

**Rewards.** Rewards in Mujoban are similar to Sokoban: +1 for pushing a box on a target pad, -1 for removing a box from a target pad, and +10 for solving the level.

**Difficulty levels.** The environment provides different levels of difficulty. In our experiments we aim for the hardest, partially observable configuration of the task:  $10 \times 10$  Sokoban levels with 4 boxes, first-person observations  $(o^{\text{cam}}, o^{\text{pr}}, o^{\text{pos}})$ . We use easier levels only for evaluation, and leverage extra observations only during training (and for baselines). Table 6 shows the configuration for each difficulty level. All grids are padded with walls so the final size of all grids are  $10 \times 10$ .

Difficulty level	Category	Grid size	Number of boxes
1	easy	$5 \times 5$	1
2		$7 \times 7$	1
3		$7 \times 7$	2
4	medium	$8 \times 8$	3
5	hard	$10 \times 10$	4

Table 6: Mujoban difficulty levels

**Grid pegs.** To make the underlying logic of Mujoban similar to Sokoban, pegs are inserted at each grid point, acting as physical barriers that confine the motion of boxes to horizontal and vertical axes (and make manipulating boxes harder). Yet the correspondence with Sokoban rules is not perfect, e.g., there is no built-in mechanism that prevents pushing multiple boxes, and although challenging, getting a box off a wall is also conceivable.

## B Modular RL details

### B.1 Perception

The perception network is a recurrent classifier mapping from  $(o_{t'}^{\text{cam}}, o_{t'}^{\text{pos}})_{t' \leq t}$  and to  $S_t$ . The perception network is composed of three components: the first one is a neural network with geometry-aware structural prior, shown in Fig. 4, which pre-processes each visual frame separately. The output of the network is fed into a ConvLSTM [44], which forms an implicit belief state over the abstract state. The output of the LSTM is input into a classifier with 5 labels (None, Wall, Box, Target, Box-on-target), which predicts each spatial location of the abstract state separately. An autoregressive classifier could also be used, but we found it not necessary.

The geometry-aware frame processor works intuitively by attending to different elements of the visual input for each possible location of the abstract state. More precisely, we first extract a  $H \times W \times C$  feature tensor from  $o_t^{\text{cam}}$  using a ResNet. Then, for each spatial location  $(x, y)$  of the abstract state  $S_t$ , we use the agent pose input  $o_t^{\text{pos}}$  to compute the coordinates of the spatial location relative to the pose. The relative spatial coordinates are tiled and concatenated with pixel coordinates to a  $H \times W \times C$  features tensor. We pass the resulting coordinate matrix through a CNN, which defines attention weights of size  $H \times W \times 1$ . We normalize weights to sum to one, and compute the weighted sum of image features over the  $H$  and  $W$  dimensions for each channel, which results in a single feature

column  $1 \times 1 \times C$ . This is done for each element  $(x, y)$  of the abstract state in parallel (using shared weights); the different outputs are then recombined into a  $10 \times 10 \times C$  feature map.

The geometry-aware perception network relies on knowing the agent’s pose. We used a setting where the pose is directly observed for simplicity, but it could be also predicted from visual and proprioceptive observations.

We collect data for training by executing the planner, controller and control-completion modules, and using true abstract states as labels. We treat the output of the perception network as independent binary classifiers for each cell, separately for the box, target pad and wall layers of the abstract state. For the agent state layer we treat the output as a single classifier over all cells. The loss is then given by the sum of independent cross-entropy losses for each classifier.

## B.2 Planner

The planner is a model-free RL policy for a time-abstracted RL task. Inputs are abstract state estimates,  $S_t$ , outputs are discrete instructions,  $A_t$ .

The planner policy is trained with MPO [16], a recent actor-critic off-policy RL algorithm. We adapt MPO to a time-abstracted discrete RL task. In each step of the time-abstracted task the controller module executes the abstract instruction through multiple real environment steps, as many as needed to complete the instruction. Rewards are defined as the sum of real environment rewards collected during the execution of the instruction, plus a  $-0.01$  reward for each planner step. In case the instruction is not completed within  $T_{instr} = 120$  steps, the experience is dropped and treated as an early-termination of the time-abstracted episode. During training we input ground-truth abstract states to the planner, which is replaced by predicted  $S_t$  during evaluation.

We use a policy network similar to the repeated ConvLSTM architecture of Guez et al. [38]. The network is built only of generic components, but it has strong structural bias for playing Sokoban. We use a feed-forward variant of the architecture, replacing the hidden states by a trainable variable that is not propagated through time. We add a separate critic head to the architecture with 5 output values, the Q-value estimates for each discrete instruction.

## B.3 Controller

**Control policy.** The controller is a goal-oriented model-free policy trained with MPO. Inputs are abstract instruction  $A_t$ , visual observation  $o_t^{cam}$ , proprioceptive observations  $o_t^{pr}$ , and completion signal  $\beta_t$ . Outputs are continuous motor torques  $a_t$ .

The policy network is a ResNet for processing the visual input connected to an LSTM [42] and a fully-connected layer that outputs parameters of a multivariate Normal distribution. We use an asymmetric actor-critic setup, where the critic network receives privileged information during training. The inputs to the critic are the agent pose  $o_t^{pos}$ , the pose of boxes  $o_t^{box}$ , proprioceptive observations  $o_t^{pr}$  and a control action  $a$ . The output is a Q-value estimate for  $a$ .

During training the controller receives multiple (random) instructions. We define rewards for training the controller as follows. The agent receives a positive reward (+4) when an instruction is completed, and a negative reward ( $-5$ ) if the instruction is not completed within  $T_{instr} = 120$  steps. Further, to encourage forward motion a small negative reward is given every step the agent is moving backwards. Rewards are computed using privileged state observations from the simulator.

**Control completion.** The controller also outputs a completion signal  $\beta_t$  using a separate learned classifier. Specifically, the control-completion network is a sequential binary classifier trained with supervised data. The inputs are  $o_t^{cam}, o_t^{pr}, A_{t-1}$  and  $\beta_{t-1}$ . The output is  $p_t(\beta)$ , the estimated probability that the controller has completed the last abstract instruction at time  $t$ . The network architecture is a convolutional ResNet connected to an LSTM. In our implementation ResNet weights are shared with the perception network. The completion-signal  $\beta_t$  is sampled according to  $\beta_t \sim p_t(\beta)$ . During evaluation a positive completion signal is added if there has been no completion predicted for 120 steps.

We train the control-completion component together with the perception module by executing the planner and controller modules. During execution the predicted  $\beta$  signal is fed to the controller. Supervised training labels are obtained according to a strict definition of completing an abstract

instruction (defined below), and computed using privileged observations available during training. We use a re-weighted cross-entropy loss that accounts for the imbalanced number of positive and negative samples. Specifically, we re-weight the loss with weight  $w_t = 1 + N_{missed}$ , where  $N_{missed}$  is the number of steps the prediction has been negative while the label has been positive.

The definition of completing an instruction is as follows. An instruction is completed if the agent is near the center of the target grid cell corresponding to the instructed move, with a  $d_{tol} = 0.1$  unit tolerance. If the target cell is occupied by a box the box needs to be pushed to the center of the adjacent cell with  $d_{tol} = 0.1$  tolerance. If the target cell is occupied by a wall the instruction is infeasible, and the agent must remain in its current grid cell. When training the controller we compute the completion criteria and rewards based on privileged state observations. When evaluating the system the completion signal is given by the  $\beta_t$  prediction.

#### B.4 Training and implementation

We train modules of modular RL on randomly generated Mujoban levels from the most difficult category (10×10 grid with 4 boxes). After modules are trained we evaluate the full system on a separate set of 512 random levels. We use privileged observations of the true environment state during training but not during evaluation. An evaluation episode is successful if the level is solved within 240s (4800 environment steps). We also run evaluations on easier levels of Mujoban, with smaller grid and less boxes, but we do not use the easier levels during training unless indicated.

Algorithm 1 describes a general algorithm for training a modular policy with time abstractions. We omitted the following domain-specific implementation details for clarity. We include side information ( $o^{box}, S^*$ ) in  $o$  during training which is used to compute controller rewards and supervised losses, as described under each module. We use distributed RL where experience collection and module updates are separated into different nodes. While joint training of all modules would be possible, for simplicity we only train one module at a time in a bottom-up sequence (controller, planner, perception). Untrained modules are replaced with ground-truth or random inputs. The perception module and the completion predictor of the controller are always trained together, and their ResNet weights are shared. Because the perception module is a recurrent network, even if predictions are not needed we have to unroll the network to update its hidden state in the innermost loop Algorithm 1. For terminal states to be meaningful in an abstract episode we drop the last step in case the controller fails to complete an instruction. Finally, we distinguish induced early-termination and environment termination and only do bootstrapping in MPO for the former.

In our distributed RL setup we use one learner and  $N_{actor}$  actors. We choose  $N_{actor}$  when training each module independently to keep the ratio of actor and learner steps for off-policy learning similar. We use  $N_{actor} = 200$  for the perception module,  $N_{actor} = 1000$  for the planner module, and  $N_{actor} = 256$  for the controller module. We train modules until near convergence, which took up to 7 days each. Baseline networks were trained in a similar distributed setup, while we continued training for 21 and 5 days, using  $N_{actor} = 1000$  and  $N_{actor} = 256$ , for fully-observable and partially-observable baselines, respectively. For partially observable baselines, following Mirza et al. [1], we terminated episodes early during training ( $T_{env} = 45s$ ) which we found to produced better results in early experiments. For evaluation we increase the time limit to  $T_{env} = 240s$ , same as for modular RL.

All models are implemented in Tensorflow [46] and trained with the Adam optimizer [47]. We use learner nodes equipped with an Nvidia Tesla V100 GPU. Actor nodes have access only to CPU cores.

## C Baselines

As an alternative to our modular design, we run MPO with standard monolithic network architectures and train end-to-end for the overall rewards. To gain better understanding of the task difficulty, we also consider a simplified setting with top-down input and train agents with monolithic and structured architectures using the V-Trace [50] algorithm. While the choice of RL algorithm, input observations, and network architecture are orthogonal, we did not try their combinations due to the substantial computation and time required to run these experiments (on the order of weeks).

### C.1 Baselines for the full, partially-observable Mujoban

**Monolithic RL.** We run MPO with standard network architectures that connect a ResNet to an LSTM. We search over network variants of three distinct sizes, as well as learning rates, discount factor, MPO parameters, and report the best setting. A small network variant is similar to the controller module except it does not receive instruction inputs. The medium network increases the LSTM hidden state size and the size of fully connected layers. The large network adds extra pre-processing layers for the non-visual inputs and extra layers after the LSTM output. We search over network variants and report best results.

**Asymmetric vs. symmetric actor-critic.** For a fair comparison, we first try an asymmetric actor-critic setting, where the critic network receives the same amount of privileged information we used for training modular RL, i.e., the true abstract state  $S_t^*$ , and poses  $o_t^{pos}$  and  $o_t^{box}$ . The policy network receives the same set of inputs as the modular policy: visual observation, agent pose, and proprioceptive observations. The critic network receives the same privileged information as we used for training modular RL, i.e., the true abstract state  $S_t^*$ , and physical state including the absolute pose of the agent and all boxes. We then also try a symmetric setup, where both actor and critic receive the same input. Here we use a single network torso with separate policy and Q-function network heads.

**Curriculum.** We train baselines only on hard levels, same as for modular RL, as well as using a training curriculum with levels randomly sampled from all difficulty categories ( $p_1 = 0.25, p_2 = 0.25, p_3 = 0.2, p_4 = 0.2, p_5 = 0.1$ ), and training only on easy levels ( $p_1 = 1.0$ ). After training we evaluate for all difficulty categories.

### C.2 Baselines for simplified, fully-observable Mujoban

**Fully-observable monolithic RL.** Here we provide the agent with access to top-down camera images of the full environment, removing (most of) the partial observability. We first use a monolithic LSTM architecture and the V-trace [50] algorithm. We train end-to-end for the overall RL task. We use curriculum with combination of all level difficulties.

The network details following Mirza et al. [1] are as follows. Proprioception inputs are all concatenated and passed to one layer MLP with 100 hidden units. The vision inputs are passed through 3 layer ResNet with channels sizes 16, 32, 32 and each layer consists of 2 blocks. The outputs flattened and passed through one layer MLP with size of 256. All the flat inputs finally concatenated together and passed to and LSTM agent with similar architecture as the controller of our modular RL approach.

**Fully-observable structured RL.** Finally, we also try an alternative approach with strong structured exploration priors for Mujoban. Here the agent observes both top-down images and the true abstract state. We add structured exploration similar to the “random planner” agent in [1]. That is we add an extra pseudo reward of 0.1 for reaching one of the randomly chosen 4 adjacent state in the abstract space in a given time frame. This subgoal is given to the agent by inputting the target abstract state along with the current abstract state. The agent has extra value head and computes a V-trace loss separately for the pseudo reward and the sub-task episode. The gradients from this auxiliary loss are added to those associated with the main task, using a weight of 0.5 for the auxiliary loss. The aim of the auxiliary loss is to help the agent to explore more meaningfully in the abstract space. If the subgoal is not reached within 50 steps a new random subgoal is sampled. The network architecture is similar to the fully-observable monolithic agent. We train using the same curriculum as for monolithic baselines.

## D Extended literature review

### Deep RL and end-to-end reasoning

Recent research in deep RL has led to dramatic progress in the components of embodied reasoning, including multi-step abstract reasoning [5, 6, 7], egocentric perception and state estimation, e.g., [8, 32, 33, 10], and embodied spatiotemporal control [2, 3, 4]. In spite of this, and though few in the community deny the importance of the full embodied reasoning problem, little recent work has attempted to tackle problems with all components of embodied intelligence. Notable exceptions include solving the Rubik’s cube using a robotic hand [4], and the IKEA RL domain [13]. To

contextualize our aims and approach, we review recent work in RL as it relates to solving full embodied reasoning tasks and contrast it to modular systems design in the wider control literature.

### **Modular systems in AI and robotics**

Although recent research in RL has largely focused on end-to-end solutions to behavioral problems, modular designs have a long history in artificial intelligence research. As early as the 1960s, researchers had designed systems with decomposable structure in order to solve challenging reasoning problems: see (Nilsson 51, §1.4) for a discussion. The later robotics literature is full of modular systems that encompass perception and motor control in addition to abstract problem solving, e.g. Brooks [52], Gat [53]. Modular design is one of the cornerstones of practical robotics systems because of the current intractability of end-to-end methods, the need for careful diagnosis of each component, and the desire for component reuse. The systems proposed for DARPA robotics challenge illustrate the general success of modular design patterns: current autonomous and semi-autonomous systems for controlling cars [54], humanoid robots [55], and aerial vehicles [56] consist of specialized modules for perception, reasoning, and actuator control coupled together with carefully designed interfaces.

### **Modular structure in deep RL**

Although methods in deep RL often attempt to avoid engineering structure in favor of developing general methods, research into system structure has been a key component of deep RL's success. For example, consider the line of work leading from AlphaGo to MuZero [5, 57, 58, 59]: the earliest models in this family included domain-specific engineering, such as data augmentation by board rotation and perfect knowledge of game rules. With the lessons learned by the success of earlier methods, later architectures progressively removed and modified structure - removing assumptions about game rules, refining the details of Monte Carlo tree search, and upgrading the convolutional architecture to residual networks - producing more general and more powerful methods. These later innovations were made possible by incorporating lessons learned by first engineering a more restrictively structured method for a very difficult problem of interest.

Compared to a monolithic system designed purely as a black box, a system that exploits modularity can exploit the system designer's intuition about how to structure a problem using specially designed architectures and losses. In this sense, many recently proposed methods have exploited modularity by engineering specialized architecture or losses: for example, by introducing structure exploiting memory [60], map-based reasoning [25], prediction [61, 10], or several such components [62, 63]. This additional structure is typically chosen by appealing to the information content of the environment or correlations with the behavioral task of interest and empirically verified on a difficult benchmark.

Other related work aims to use algorithmic structures as priors for DNN policies [23, 24, 37, 64]. One recent work that combines the benefits of internal system structure with end-to-end learning is the Differentiable Algorithm Network (DAN) of Karkus et al. [35]. This model composes differentiable structures for state estimation (vision + filtering), planning, and control and trains them jointly for partially observed map-based navigation tasks. The resulting model can make plans that take into account global map structure while avoiding visually identified obstacles.

Other methods propose to learn a model which is then used to simplify the task of another model in some way. For instance, in hierarchical reinforcement learning (HRL), this may correspond to learning motor primitives [30, 33], learning hierarchies of agents which work at different resolution in time and space [27, 28, 31, 65], or discovering structured options for exploration [29].

In model-based RL, this corresponds to learning a model of the environment, which can then be leveraged by a controller in various ways, such as providing synthetic training data [66, 67] or simulating trajectories to improve decision making at test time [68, 69, 70], improving learning of policies during training [71, 72, 73], or combining several of these aspects [59].

Yet another family of approach is to learn to predict structured representation of data and used for downstream reasoning and control; this includes methods trained to detect or segment visual structure and use the estimated structure for downstream reasoning [74, 75, 76] and methods that leverage knowledge of object and physics simulators [77, 78, 79].

Finally, a line of research closely related to our proposed approach has sought to partition or structure reasoning, for instance by using an architecture combining reasoning in structured goal spaces with low-level policies [80, 81, 82], combining abstract planning with learned symbolic representations and skills [83, 84], or introducing sub-modules corresponding to a natural task partition [85].

### **Integrated robot learning domains**

Integrated robotic tasks has been of interest for decades including indoor navigation [86], autonomous driving [54], performing household tasks [87, 88] and object manipulation with tools [89]. Recent works introduced a number of integrated robot learning domains suitable for large scale RL in simulation [1, 13, 90, 91]. In this paper we focus on the Mujoban task [1]. Mujoban exhibits key challenges of embodied reasoning: partial observability, long-horizon reasoning, continuous motor control and object manipulation. Compared to other, more realistic domains, an important benefit of Mujoban is the ability to control the difficulty of different layers of embodied reasoning. That is, we can control the complexity of abstract reasoning by changing the size of the underlying Sokoban level and the number of boxes; the complexity of perception by adding top-down or direct state observations; and the complexity of motor-control by replacing the robot body.