

Few-Shot Bayesian Imitation Learning with Logical Program Policies

Tom Silver, Kelsey R. Allen, Alex K. Lew, Leslie Kaelbling, Josh Tenenbaum

Massachusetts Institute of Technology
 {tslvr, krallen, alexlew, lpk, jbt}@mit.edu

Abstract

Humans can learn many novel tasks from a very small number (1–5) of demonstrations, in stark contrast to the data requirements of nearly tabula rasa deep learning methods. We propose an expressive class of policies, a strong but general prior, and a learning algorithm that, together, can learn interesting policies from very few examples. We represent policies as logical combinations of programs drawn from a domain-specific language (DSL), define a prior over policies with a probabilistic grammar, and derive an approximate Bayesian inference algorithm to learn policies from demonstrations. In experiments, we study six strategy games played on a 2D grid with one shared DSL. After a few demonstrations of each game, the inferred policies generalize to new game instances that differ substantially from the demonstrations. Our policy learning is 20–1,000x more data efficient than convolutional and fully convolutional policy learning and many orders of magnitude more computationally efficient than vanilla program induction. We argue that the proposed method is an apt choice for tasks that have scarce training data and feature significant, structured variation between task instances.

Introduction

People are remarkably good at learning and generalizing strategies for everyday tasks, like ironing a shirt or brewing a cup of coffee, from one or a few demonstrations. Websites like WikiHow.com and LifeHacker.com are filled with thousands of “how-to” guides for tasks that are hard to solve by pure reasoning or trial and error alone, but easy to learn and generalize from just one illustrated demo (Figure 1). We are interested in designing artificial agents with the same few-shot imitation learning capabilities.

A common approach to imitation learning is behavior cloning (BC), in which demonstrations are used as supervision to directly train a policy. BC is often thought to be too prone to overfitting to generalize from very little data. Indeed, we find that neural network policies trained with BC are susceptible to severe overfitting in our experiments. However, we argue that this failure is due not to BC in general, but rather, to an underconstrained policy class and a weak prior.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

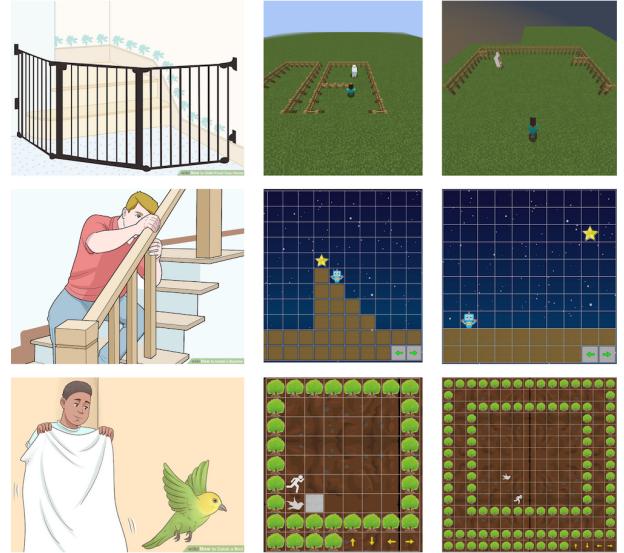


Figure 1: People can learn strategies for an enormous variety of tasks from one or a few demonstrations, e.g., “gate off an area,” “build stairs,” or “catch a bird” (left). We propose a policy class and learning algorithm for similarly data-efficient imitation learning. Given 1–5 demos of tasks like “Fence In,” “Reach for the Star,” and “Chase” (middle), we learn policies that generalize substantially (right).

More structured policies with strong Occam’s razor priors can be found in two lines of work: logical and relational (policy) learning (Džeroski, De Raedt, and Blockeel 1998; Natarajan et al. 2011), and program (policy) synthesis (Wingate et al. 2013; Sun et al. 2018). Policies expressed in predicate logic are easy to learn, but difficult to scale, since each possible predicate must be hand-engineered by the researcher. Programmatic policies can be automatically generated by searching a small domain-specific language (DSL), but learning even moderately sophisticated policies can require an untenably large search in program space.

We propose Logical Program Policies (LPP): an expressive, structured, and efficiently learnable policy class that

combines the strengths of logical and programmatic policies. Our first main idea is to consider policies that have logical “top level” structure and programmatic feature detectors (predicates) at the “bottom level.” The feature detectors are expressions in a domain-specific language (DSL). By logically combining feature detectors, we can derive an infinitely large, rich policy class from a small DSL. This “infinite use of finite means” is in contrast to prior work in relational RL where each feature is individually engineered, making it labor-intensive to apply in complex settings.

Our second main idea is to exploit the logical structure of LPP to obtain an efficient imitation learning algorithm, overcoming the intractability of general program synthesis. To find policies in LPP, we incrementally enumerate feature detectors, apply them to the demonstrations, invoke an off-the-shelf Boolean learning method, and score each candidate policy with a likelihood and prior. What would be an intractable search over full policies is effectively reduced to a manageable search over feature detectors. We thus have an efficient approximate Bayesian inference method for $p(\pi|\mathcal{D})$, the posterior distribution of policies π given demonstrations \mathcal{D} .

While LPP and the proposed learning method are agnostic to the particular choice of the DSL and application domain, we focus here on six strategy games which are played on 2D grids of arbitrary size (see Figure 3). In these games, a state consists of an assignment of discrete values to each grid cell and an action is a single grid cell (a “click” on the grid). The games are diverse in their transition rules and in the tactics required to win, but the common state and action spaces allow us to build our policies for all six games from one shared, small DSL. In experiments, we find that policies learned from five or fewer demonstrations can generalize perfectly in all six games. In contrast, policies learned as convolutional neural networks fail to generalize, even when domain-specific locality structure is built into the architecture (as in fully convolutional networks (Long, Shelhamer, and Darrell 2015)). Overall, our experiments suggest that LPP offers an efficient, flexible framework for learning rich, generalizable policies from very little data.

Problem Statement

In imitation learning, we are given a dataset \mathcal{D} of expert trajectories $(s_0, a_0, \dots, s_{T-1}, a_{T-1}, s_T)$ where $s_t \in \mathcal{S}$ are states and $a_t \in \mathcal{A}$ are actions. We suppose that the trajectories are sampled from a Markov process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, \mathcal{G})$, with transition distribution $T(s' | s, a)$ and goal states $\mathcal{G} \subset \mathcal{S}$, and that actions are sampled from an expert policy $\pi^* : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, where $\pi^*(a | s)$ is a state-conditional distribution over actions. For imitation learning, we must specify (1) a hypothesis class of policies Π and (2) an algorithm for learning a policy $\pi \in \Pi$ from \mathcal{D} that matches the expert π^* . We assume that the expert π^* is optimal with respect to \mathcal{M} , so we report the fraction of trials in which a learned policy π reaches goal states in \mathcal{G} from held-out initial states in \mathcal{M} to evaluate performance.

The LPP Policy Class

We seek a policy class Π with a concise parameterization that can be reasonably specified by a human programmer, and for which there is a tractable learning algorithm for recovering $\pi^* \in \Pi$ from demonstrations.

We consider policies that are parameterized by state-action classifiers $h : \mathcal{S} \times \mathcal{A} \rightarrow \{0, 1\}$. When $h(s, a) = 0$, action a will never be taken in state s ; when $h(s, a) = 1$, a may be taken. This parameterization allows us to handle arbitrarily large action spaces (variable grid sizes). Given $h(s, a)$, we can derive a corresponding policy $\pi(a | s)$ that samples a uniformly at random among those a such that $h(s, a) = 1$. In other words, $\pi(a | s) \propto h(s, a)$. This stochastic policy formulation reflects the fact that the demonstrator may randomly select among several optimal actions. For completeness, we define $\pi(a | s) \propto 1$ if $\forall a, h(s, a) = 0$. Specifying a policy class Π thus reduces to specifying a class of functions \mathcal{H} from which to learn an h .

One option for \mathcal{H} is to consider *logical* rules that compute Boolean expressions combining binary features derived from (s, a) . Although this enables fast inference using well-understood Boolean learning algorithms, it requires the AI programmer to hand-engineer informative binary features, which will necessarily vary from task to task. Another option is to consider *programmatic* rules: rules that are expressions in some general-purpose DSL for predicates on state-action pairs. In this case, the AI programmer need only specify a small core of primitives for the DSL, from which task-specific policies can be derived during inference. The challenge here is that finding a good policy in the infinitely large class of programs in the DSL is difficult; simple methods like enumeration are much too slow to be useful.

We combine the complementary strengths of logical and program-based policies to define the Logical Program Policies (LPP) class. Policies in LPP have a logical “top level” and a programmatic “bottom level.” The bottom level is comprised of feature detector programs $f : \mathcal{S} \times \mathcal{A} \rightarrow \{0, 1\}$. These programs are expressions in a DSL and can include, for example, loops and conditional statements. A feature detector program takes a state s and an action a as input and returns a binary output, which provides one bit of information about whether a should be taken in s . The top level is comprised of a logical formula h over the outputs of the bottom level. Without loss of generality, we can express the formula in disjunctive normal form:

$$h(s, a) \triangleq (f_{1,1}(s, a) \wedge \dots \wedge f_{1,n_1}(s, a)) \vee \dots \vee (f_{m,1}(s, a) \wedge \dots \wedge f_{m,n_m}(s, a)) \quad (1)$$

where the f ’s are possibly negated. LPP thus includes all policies that correspond to logical formulae over finite subsets of feature detector programs expressed in the DSL.

Imitation Learning as Bayesian Inference

We now address the imitation learning problem of finding a policy π that fits the expert demonstrations \mathcal{D} . Rather than finding a single LPP policy, we will infer a full posterior distribution over policies $p(\pi | \mathcal{D})$. From a Bayesian perspective, maintaining the full posterior is principled; from a

practical perspective, the full posterior leads to modest performance gains over a single MAP policy. Once we have inferred $p(\pi | \mathcal{D})$, we will ultimately take MAP actions according to $\arg \max_{a \in \mathcal{A}} \mathbb{E}_{p(\pi | \mathcal{D})}[\pi(a | s)]$.

Probabilistic Model $p(\pi, \mathcal{D})$

We begin by specifying a probabilistic model over policies and demonstrations $p(\pi, \mathcal{D})$, which factors into a prior distribution $p(\pi)$ over policies in LPP, and a likelihood $p(\mathcal{D} | \pi)$ giving the probability that an expert generates demonstrations \mathcal{D} by following the policy π .

We choose the prior distribution $p(\pi)$ to encode a preference for those policies which use fewer, simpler feature detector programs. Recall that a policy $\pi \in \text{LPP}$ is parameterized by a logical formula $h(s, a) = \bigvee_{i=1 \dots M} \left(\bigwedge_{j=1 \dots N_i} f_{i,j}(s, a)^{b_{ij}} (1 - f_{i,j}(s, a))^{1-b_{ij}} \right)$, in which each of the $f_{i,j}$ is a binary feature detector expressed in a simple DSL and the b_{ij} are binary parameters that determine whether a given feature detector is negated. We set the prior probability of such a policy to depend only on the number and sizes of the programmatic components $f_{i,j}$: namely, $p(\pi) \propto \prod_{i=1}^M \prod_{j=1}^{N_i} p(f_{i,j})$, the probability of generating each of the $f_{i,j}$ independently from a probabilistic context-free grammar $p(f)$ (Manning and Schütze 1999).¹ The grammar we use in this work is shown in the appendix.

The likelihood of a dataset \mathcal{D} given a policy π is $p(\mathcal{D} | \pi) \propto \prod_{i=1}^N \prod_{j=1}^{T_i} \pi(a_{ij} | s_{ij})$. In the appendix, we also consider the case where demonstrations are corrupted by noise.

Approximating the Posterior $p(\pi | \mathcal{D})$

Algorithm 1: LPP imitation learning

```

input: Demos  $\mathcal{D}$ , ensemble size  $K$ , max iters  $L$ 
Create anti-demos  $\bar{\mathcal{D}} = \{(s, a') : (s, a) \in \mathcal{D}, a' \neq a\}$ ;
Set labels  $y[(s, a)] = 1$  if  $(s, a) \in \mathcal{D}$  else 0;
Initialize approximate posterior  $q$ ;
for  $i$  in  $1, \dots, L$  do
     $f_i = \text{generate\_next\_feature}();$ 
     $X = \{(f_1(s, a), \dots, f_i(s, a))^T : (s, a) \in \mathcal{D} \cup \bar{\mathcal{D}}\}$ 
     $\mu_i, w_i = \text{logical.inference}(X, y, p(f), K);$ 
     $\text{update\_posterior}(q, \mu_i, w_i);$ 
end
return  $q$ ;

```

We now have a prior $p(\pi)$ and likelihood $p(\mathcal{D} | \pi)$, and we wish to compute an approximate posterior $q(\pi) \approx p(\pi | \mathcal{D})$. We take q to be a weighted mixture of K policies μ_1, \dots, μ_K (in our experiments, $K = 25$) and initialize it so that each μ_i is equally weighted and equal to the uniform policy, $\mu_i(a | s) \propto 1$. Our core insight is a way to

¹Note that without some maximum limit on $\sum_{i=1}^M N_i$, this is an improper prior, and for this technical reason, we introduce a uniform prior on $\sum_{i=1}^M N_i$, between 1 and a very high maximum value α ; the resulting factor of $\frac{1}{\alpha}$ does not depend on π at all, and can be folded into the proportionality constant.

exploit the structure of LPP to efficiently search the space of policies and update the mixture q to better match the posterior. In the appendix, we show that this scheme is formally a variational inference algorithm that iteratively minimizes the KL divergence from q to the true posterior.

Our algorithm is given a set of demonstrations \mathcal{D} . The state-action pairs (s, a) in \mathcal{D} comprise positive examples — inputs for which $h(s, a) = 1$. We start by computing a set of “anti-demonstrations” $\bar{\mathcal{D}} = \{(s, a') \mid (s, a) \in \mathcal{D}, a' \neq a\}$, which serve as approximate negative examples. ($\bar{\mathcal{D}}$ is approximate because it may contain false negatives, but they will generally constitute only a small fraction of the set.)

We now have a binary classification problem with positive examples \mathcal{D} and negative examples $\bar{\mathcal{D}}$. The main loop of our algorithm considers progressively larger feature representations of these examples. At iteration i , we use only the simplest feature detectors f_1, \dots, f_i , where “simplest” here means “of highest probability under the probabilistic grammar $p(f)$.” We can enumerate features in this order by performing a best-first search through the grammar.

Given a finite set of feature detectors f_1, \dots, f_i , we can convert any state-action pair (s, a) into a length- i binary feature vector $x \in \{0, 1\}^i = (f_1(s, a), \dots, f_i(s, a))^T$. We do this conversion on \mathcal{D} and $\bar{\mathcal{D}}$ to obtain a design matrix $X_i \in \{0, 1\}^{|\mathcal{D} \cup \bar{\mathcal{D}}| \times i}$. The remaining problem of learning a binary classifier as a logical combination of binary features is very well understood (Mitchell 1978; Valiant 1985; Quinlan 1986; Dietterich and Michalski 1986). In this work, we use an off-the-shelf stochastic greedy decision-tree learner (Pedregosa et al. 2011).

Given a learned decision tree, we can easily read off a logical formula $h(s, a) = \bigvee_{j=1 \dots M} \left(\bigwedge_{l=1 \dots N_j} f_{j,l}(s, a)^{b_{jl}} (1 - f_{j,l}(s, a))^{1-b_{jl}} \right)$, in which each of the $f_{j,l}$ is one of the i feature detectors under consideration at iteration i . This induces a candidate policy $\mu_*(a | s) \propto h(s, a)$. We can evaluate its prior probability $p(\mu_*)$ and its likelihood $p(\mathcal{D} | \mu_*)$, then decide whether to include μ_* in our mixture q , based on whether its unnormalized posterior probability is greater than that of the lowest-scoring existing mixture component. The mixture is always weighted according to our model over π and \mathcal{D} , so that $q(\mu_j) = \frac{p(\mu_j | \mathcal{D})}{\sum_{i=1}^K p(\mu_i | \mathcal{D})}$. In practice, we run the decision-tree learner several times (5 in experiments) with different random seeds to generate several distinct candidate policies at each iteration of the algorithm. We can stop the process after a fixed number of iterations, or when the prior probabilities of the enumerated programs f_i fall below a threshold: any policy that uses a feature detector f_i with prior probability $p(f_i) < p(\mu_j, \mathcal{D})$ for all μ_j in q ’s support has no chance of meriting inclusion in our mixture.

Once we have an approximation q to the posterior, we can use it to derive a final policy for use at test time:

$$\pi_*(s) = \arg \max_{a \in \mathcal{A}} \mathbb{E}_q[\pi(a | s)] = \arg \max_{a \in \mathcal{A}} \sum_{\mu \in q} q(\mu) \mu(a | s).$$

We could alternatively use the full distribution over actions to guide exploration, e.g., in combination with reinforce-

Method	Type	Description
cell_is_value	$V \rightarrow C$	Check whether the attended cell has a given value
shifted	$O \times C \rightarrow C$	Shift attention by an offset, then check a condition
scanning	$O \times C \times C \rightarrow C$	Repeatedly shift attention by the given offset, and check which of two conditions is satisfied first
at_action_cell	$C \rightarrow P$	Attend to the action cell and check a condition
at_cell_with_value	$V \times C \rightarrow P$	Attend to a cell with the value and check condition

Table 1: Methods of the domain-specific language (DSL) used in this work. A *program* (P) in the DSL implements a predicate on state-action pairs (i.e., $P = \mathcal{S} \times \mathcal{A} \rightarrow \{0, 1\}$), by attending to a certain cell, then running a *condition* (C). Conditions check that some property holds of the current state *relative* to an implicit attention pointer. V ranges over possible grid cell values and an “off-screen” token, and O over “offsets,” which are pairs (x, y) of integers specifying horizontal and vertical displacements.

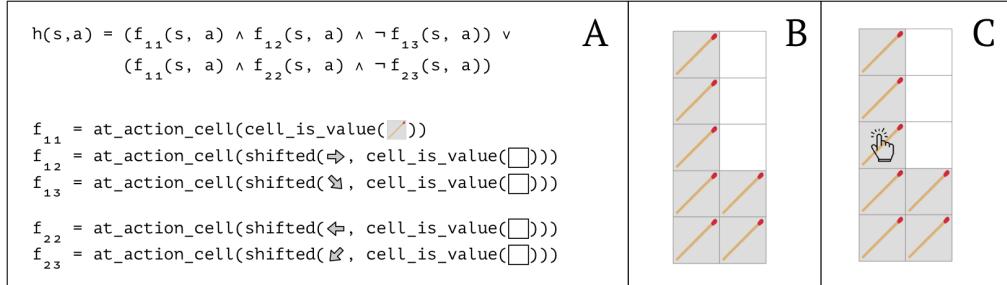


Figure 2: Example of a policy in LPP for the “Nim” game. (A) $h(s, a)$ is a logical combination of programs from a DSL. For example, f_{12} returns True if the cell to the right of the action a has value \square . The induced policy is $\pi(a | s) \propto h(s, a)$. (B) Given state s , (C) there is one action selected by h . This policy encodes the “leveling” tactic, which wins the game.

ment learning (Hester et al. 2018). In this work, we focus on exploitation and therefore require only the maximum *a posteriori* actions, for use with a deterministic final policy.

Experiments and Results

We now present experiments to evaluate the data efficiency, computational complexity, and generalization of LPP versus several baselines. We also analyze the learned policies, examine qualitative performance (see Figure 4 and the appendix), and conduct ablation studies to measure the contributions of the components of LPP. All experiments were performed on a single laptop running macOS Mojave with a 2.9 GHz Intel Core i9 processor and 32 GB of memory.

Tasks

We consider six diverse strategy games (Figure 3) that share a common state space ($\mathcal{S} = \bigcup_{h,w \in \mathbb{N}} V^{hw}$; variable-sized grids with discrete-valued cells) and action space $\mathcal{A} = \mathbb{N} \times \mathbb{N}$; single “clicks” on any cell). For a grid of dimension $h \times w$, we only consider clicks on the grid, i.e., $\{1, \dots, h\} \times \{1, \dots, w\}$. Grid sizes vary within tasks. These tasks feature high variability between different task instances; learning a robust policy requires substantial generalization. The tasks are also very challenging due to the unbounded action space, the absence of shaping or auxiliary rewards, and the arbitrarily long horizons that may be required to solve a task instance. Each task has a maximum episode length of 60 and counts as a failure if the episode terminates with-

out a success. There are 11 training and 9 test instances per task. Instances of Nim, Checkmate Tactic, and Reach for the Star are procedurally generated; instances of Stop the Fall, Chase, and Fence In are manually generated, as the variation between instances is not trivially parameterizable. We provide descriptions of the tasks in the appendix.

Domain-Specific Language

Recall that each feature detector program takes a state and action as input and returns a Boolean value. In our tasks, states are full grid layouts and actions are single grid cells (“clicks”). The specific DSL of feature detectors that we use in this work (Table 1) is inspired by early work in visual routines (Ullman 1987; Hay et al. 2018). Each program implements a procedure for attending to some grid cell and checking that a local condition holds nearby. Given input (s, a) , a program begins by initializing an implicit attention pointer either to the grid cell in s associated with action a (at_action_cell), or to an arbitrary grid cell containing a certain value (at_cell_with_value). Next, the program will check a condition at or near the attended cell. The simplest condition is cell_is_value, which checks whether the attended cell has a certain value. More complex conditions, which look not just *at* but *near* the attended cell, can be built up using the shifted and scanning methods. The shifted method builds a condition that first shifts the attention pointer by some offset, then applies another condition. The scanning method starts at the currently attended cell and “scans” along some direction, re-

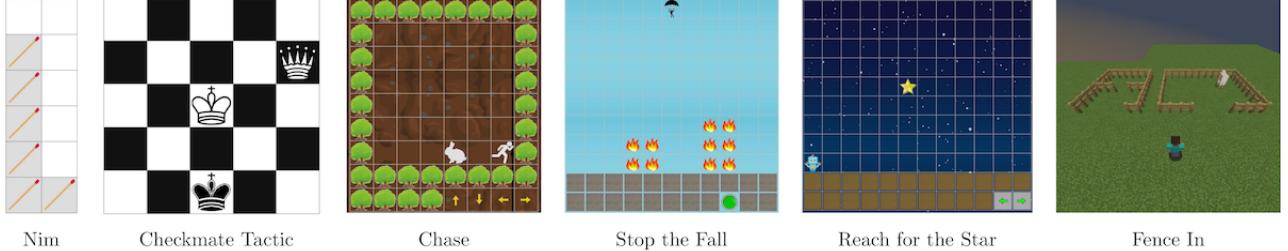


Figure 3: The strategy games studied in this work. See the appendix for descriptions and additional illustrations.

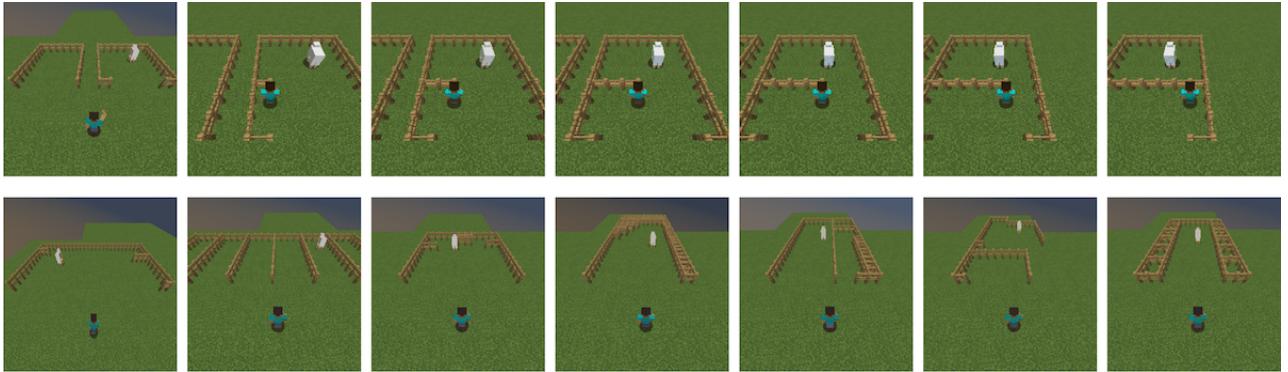


Figure 4: One-shot imitation learning in “Fence In.” From a single demonstration (top), we learn an enclosing strategy that generalizes to many new task instances (bottom).

peatedly shifting the attention pointer by a specified offset and checking whether either of two specified conditions hold. If, while scanning, the first condition becomes satisfied before the second, the scanning condition returns 1. Otherwise, it returns 0. Thus the overall DSL contains five methods, which are summarized in Table 1. See Figure 2 for a complete example of a policy in LPP using this DSL.

Baselines

Local Linear Network (LLN): A single 3×3 convolutional filter is trained to classify whether each cell in s should be “clicked,” based only on the 8 surrounding cells. **FCN:** A deep fully convolutional network (Long, Shelhamer, and Darrell 2015) is trained with the same inputs and outputs as “Local Linear.” The network has 8 convolutional layers with kernel size 3, stride 1, padding 1, 4 channels (8 in the input layer), and ReLU nonlinearities. This architecture was chosen to reflect the receptive field sizes we expect are necessary for the tasks. **CNN:** A standard convolutional neural network is trained with full grid inputs and discrete action outputs. Grids are padded so that all have the same maximal height and width. The architecture is: 64-channel convolution; max pooling; 64-channel fully-connected layer; $|\mathcal{A}|$ -channel fully-connected layer. All kernels have size 3 and all strides and paddings are 1. **Vanilla Program Induction (VPI):** Full policies are enumerated from a DSL grammar that includes logical disjunctions, conjunctions, and negations over the feature detector DSL. The number of disjunctions and conjunctions each follow a geometric distribution

($p = 0.5$). (Several other values of p were also tried without improvement.) Policies are then enumerated and mixed as in LPP learning; this baseline is thus identical to LPP learning but with the greedy Boolean learning removed.

Effect of Number of Demonstrations

We first evaluate the test-time performance of LPP and baselines as the number of training demonstrations varies from 1 to 10. For each number of demonstrations, we run leave-one-out cross validation: 10 trials, each featuring a distinct set of demonstrations drawn from the overall pool of 11 training demonstrations. LPP learning is run for 10,000 iterations for each task. The mean and maximum trial performance offer complementary insight: the mean reflects the expected performance if demonstrations were selected at random; the maximum reflects the expected performance if the most useful demonstrations were selected, perhaps by an expert teacher. Results are shown in Figure 5. On the whole, LPP markedly outperforms all baselines, especially on the more difficult tasks. The baselines are limited for different reasons. The highly parameterized CNN baseline is able to perfectly fit the training data and win all *training* games (not shown), but given the limited training data and high variation from training to task, it severely overfits and fails to generalize. The FCN baseline is also able to fit the training data almost perfectly. Its additional structure permits better generalization in Nim, Checkmate Tactic, Reach for the Star, and Fence In than the CNN, but overall its performance is still far behind LPP. In contrast, the LLN baseline is unable

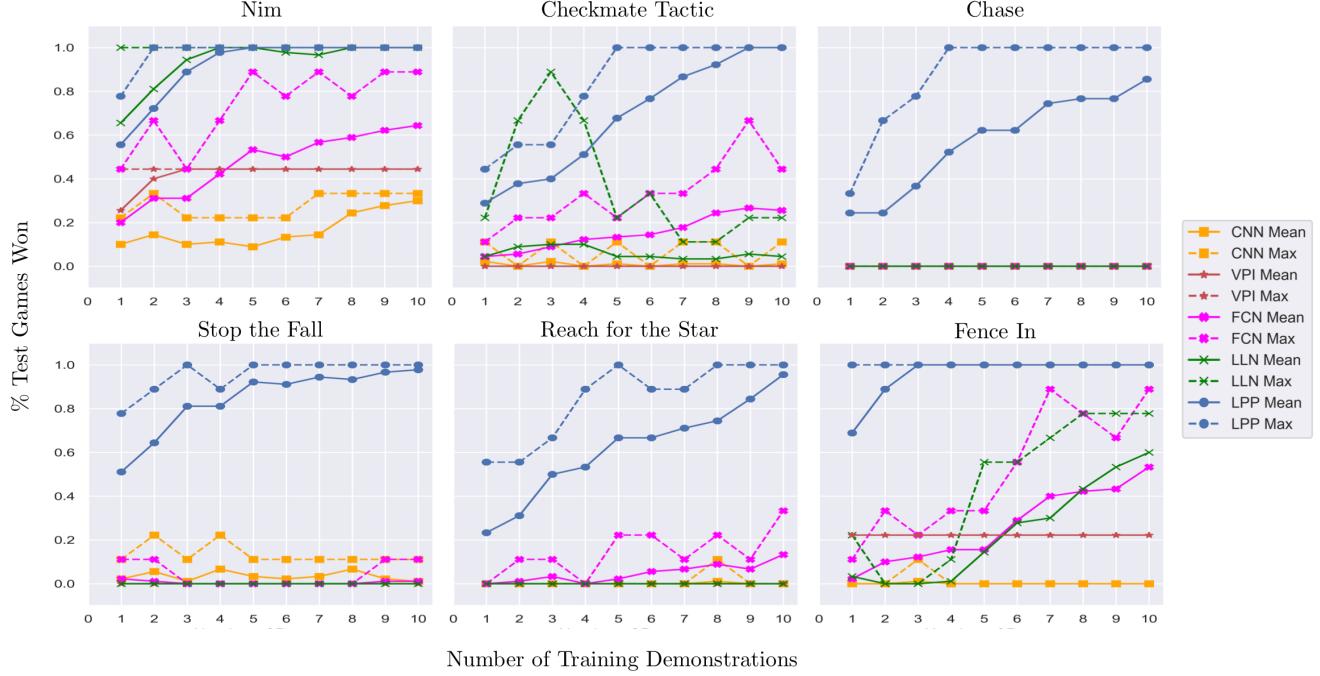


Figure 5: Performance on held-out test task instances as a function of the number of training demonstrations for LPP (ours) and four baselines. Maximums and means are over 10 training sets.

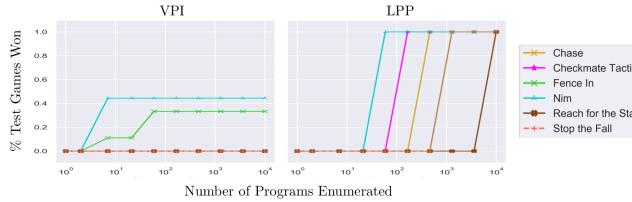


Figure 6: Performance on held-out test task instances as a function of the number of programs enumerated for the Vanilla Program Induction (VPI) baseline and LPP (ours).

to fit the training data; with the exception of Nim, its training performance is close to zero. Similarly, the training performance of the VPI baseline is near or at zero for all tasks beyond Nim. In Nim, there is evidently a low complexity program that works roughly half the time, but an optimal policy is more difficult to find.

Effect of Number of Programs Searched

We now examine test-time performance of LPP and VPI as a function of the number of programs searched. For this experiment, we give both methods all 11 training demonstrations for each task. Results are shown in Figure 6. LPP requires fewer than 100 programs to learn a winning policy for Nim, fewer than 1,000 for Checkmate Tactic and Chase, and fewer than 10,000 for Stop the Fall, Reach for the Star, and Fence In. In contrast, VPI is unable to achieve nonzero per-

	Nim	CT	Chase	STF	RFTS	Fence
LPP	1.0	1.0	1.0	1.0	1.0	1.0
Features + NN	1.0	0.67	0.0	0.0	0.22	0.67
Features + NN + L_1 Reg	1.0	0.11	0.0	0.0	0.0	0.0
No Prior	1.0	0.44	0.78	1.0	1.0	1.0
Sparsity Prior	1.0	0.78	1.0	0.78	1.0	1.0

Figure 7: Performance on held-out test task instances for LPP and four ablation models.

formance for any task other than Nim, for which it achieves roughly 45% performance after 100 programs enumerated. The lackluster performance of VPI is unsurprising given the combinatorial explosion of programs. For example, the optimal policy for Nim shown in Figure 2 involves six constituent programs, each with a parse tree depth of three or four. There are 108 unique constituent programs with parse tree depth three and therefore more than 13,506,156,000 full policies with six or fewer constituent programs. VPI would have to search roughly so many programs before arriving at a winning policy for Nim, which is by far the simplest task. In contrast, a winning LPP policy is learnable after fewer than 100 enumerations. In practical terms, LPP learning for Nim takes on the order of 1 second on a laptop without highly optimized code; after running VPI for six hours in the same setup, a winning policy is still not found.

Ablation Studies

We now perform ablation studies to explore which aspects of the LPP class and learning algorithm contribute to the

strong performance. We consider four ablated models. The “Features + NN” model learns a neural network state-action binary classifier on the first 10,000 feature detectors enumerated from the DSL. This model addresses the possibility that the features alone are powerful enough to solve the task when combined with a simple classifier. The NN is a multilayer perceptron with two layers of dimension 100 and ReLU activations. The “Features + NN + L_1 Regularization” model is identical to the previous baseline except that an L_1 regularization term is added to the loss to encourage sparsity. This model addresses the possibility that the features alone suffice when we incorporate an Occam’s razor bias similar to the one that exists in LPP learning. The “No Prior” model is identical to LPP learning, except that the grammatical prior is replaced with a uniform prior. Similarly, the “Sparsity Prior” model uses a prior that penalizes the number of top-level programs involved in the policy, without regard for the relative priors of the individual programs. Results are presented in Figure 7. They confirm that the each component — the feature detectors, the sparsity regularization, and the grammatical prior — adds value to the overall framework.

Related Work

Sample efficiency and generalization are two of the main concerns in imitation learning (Schaal 1997; Abbeel and Ng 2004). To cope with limited data, demonstrations can be used in combination with additional RL (Hester et al. 2018; Nair et al. 2018). Alternatively, a mapping from demonstrations to policies can be learned from a background set of tasks (Duan et al. 2017; Finn et al. 2017). A third option is to introduce a prior over a structured policy class (Andre and Russell 2002; Doshi-Velez et al. 2010; Wingate et al. 2011), e.g., hierarchical policies (Niekum 2013). Our work fits into the third tradition; our contribution is a new policy class with a structured prior that enables efficient learning.

LPP policies are logical at the “top level” and programmatic at the “bottom level.” Logical representations for RL problems have been considered in many previous works, particularly in *relational RL* (Džeroski, De Raedt, and Blocquel 1998; Natarajan et al. 2011). Also notable is work by Shah et al. (2018), who learn linear temporal logic specifications from demonstration using a finite grammatical prior. While some LPP programs may be seen as fixed-arity logical relations in the classical sense, others are importantly more general and powerful, involving loops and a potentially arbitrary number of atoms. For example, one program suffices to check whether a Queen’s diagonal path is clear in Chess; no relation over a fixed number of squares can capture the same feature. Furthermore, relational RL assumes that relations are fixed, finite, and given, typically hand-designed by the programmer. In LPP, the programmer instead supplies a DSL describing infinitely many features.

LPP learning is a particular type of *program synthesis*, which more broadly refers to a search over programs, including but not limited to the case where we have a grammar over programs, the programs are mappings, and input-output examples are available. When a grammar is given, the problem is sometimes called *syntax-guided synthesis* (Alur et al. 2013). Most relevant is work by Alur, Radhakrishna,

and Udupa (2017), who propose a “divide and conquer” approach that uses greedy decision tree learning in combination with enumeration from a grammar of “conditions”, similar to our “No Prior” baseline.

Recent work has also examined *neural program synthesis* (NPS) wherein a large dataset of (input, output, program) examples is used to train a guidance function for program enumeration (Parisotto et al. 2016; Devlin et al. 2017; Bunel et al. 2018; Huang et al. 2019). In practice, NPS methods are still limited to programs involving ~ 10 primitives. The neural guidance can delay, but not completely avoid, the combinatorial explosion of search in program space. LPP learning is not a generic program induction method, but rather, an algorithm that exploits the logical structure of LPP programs to dramatically speed up search, sometimes finding programs with ~ 250 primitives (Appendix Table 2).

The interpretation of policy learning as an instance of program synthesis is explored in prior work (Wingate et al. 2011; Sun et al. 2018; Verma et al. 2018). In particular, Lázaro-Gredilla et al. (2019) learn object manipulation concepts from before/after image pairs that can be transferred between 2D simulation and a real robot. In this work, we focus on the problem of *efficient inference* and compare against vanilla program induction in experiments.

Discussion and Conclusion

In an effort to efficiently learn policies from very few demonstrations that generalize substantially, we have introduced the LPP policy class and an approximate Bayesian inference algorithm for imitation learning. We have seen that the LPP policy class includes winning policies for a diverse set of strategy games, and moreover, that those policies can be efficiently learned from five or fewer demonstrations. In ongoing work we are studying how to scale our approach to a wider range of tasks, starting with more sophisticated DSLs that include counting or simple data structures. However, even our current DSL is surprisingly general. For instance, in preliminary experiments (see appendix), we find that our current algorithm can learn a generalizing policy for Atari Breakout from just one demonstration.

Beyond policy learning, this work contributes to the long and ongoing discussion about the role of prior knowledge in AI. In the common historical narrative, early attempts to incorporate prior knowledge via feature engineering failed to scale, leading to the modern shift towards domain-agnostic deep learning methods (Sutton 2019). Now there is renewed interest in incorporating inductive bias into contemporary methods, especially for problems where data is scarce. We argue that encoding prior knowledge via a probabilistic grammar over feature detectors and learning to combine these feature detectors with Boolean logic is a promising path forward. More generally, we submit that “meta-feature engineering” of the sort exemplified here strikes an appropriate balance between the strong inductive bias of classical AI and the flexibility and scalability of modern methods.

Acknowledgements

We gratefully acknowledge support from NSF grants 1523767 and 1723381; from ONR grant N00014-13-1-0333; from AFOSR grant FA9550-17-1-0165; from ONR grant N00014-18-1-2847; from Honda Research; and from the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. KA acknowledges support from NSERC. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsors.

References

- [Abbeel and Ng 2004] Abbeel, P., and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. *International Conference on Machine Learning*.
- [Alur et al. 2013] Alur, R.; Bodik, R.; Jiniwal, G.; Martin, M. M.; Raghothaman, M.; Seshia, S. A.; Singh, R.; Solar-Lezama, A.; Torlak, E.; and Udupa, A. 2013. Syntax-guided synthesis. In *Formal Methods in Computer-Aided Design*.
- [Alur, Radhakrishna, and Udupa 2017] Alur, R.; Radhakrishna, A.; and Udupa, A. 2017. Scaling enumerative program synthesis via divide and conquer. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, 319–336. Springer.
- [Andre and Russell 2002] Andre, D., and Russell, S. J. 2002. State abstraction for programmable reinforcement learning agents. *AAAI Conference on Artificial Intelligence*.
- [Bunel et al. 2018] Bunel, R.; Hausknecht, M.; Devlin, J.; Singh, R.; and Kohli, P. 2018. Leveraging grammar and reinforcement learning for neural program synthesis. *International Conference on Learning Representations*.
- [Devlin et al. 2017] Devlin, J.; Uesato, J.; Bhupatiraju, S.; Singh, R.; Mohamed, A.-r.; and Kohli, P. 2017. Robustfill: Neural program learning under noisy i/o. *International Conference on Machine Learning*.
- [Dietterich and Michalski 1986] Dietterich, T. G., and Michalski, R. S. 1986. Learning to predict sequences. *Machine learning: An artificial intelligence approach*.
- [Doshi-Velez et al. 2010] Doshi-Velez, F.; Wingate, D.; Roy, N.; and Tenenbaum, J. B. 2010. Nonparametric Bayesian policy priors for reinforcement learning. *Advances in Neural Information Processing Systems*.
- [Duan et al. 2017] Duan, Y.; Andrychowicz, M.; Stadie, B.; Ho, J.; Schneider, J.; Sutskever, I.; Abbeel, P.; and Zaremba, W. 2017. One-shot imitation learning. *Advances in Neural Information Processing Systems*.
- [Džeroski, De Raedt, and Blockeel 1998] Džeroski, S.; De Raedt, L.; and Blockeel, H. 1998. Relational reinforcement learning. In *International Conference on Inductive Logic Programming*, 11–22. Springer.
- [Finn et al. 2017] Finn, C.; Yu, T.; Zhang, T.; Abbeel, P.; and Levine, S. 2017. One-shot visual imitation learning via meta-learning. *Conference on Robot Learning*.
- [Hay et al. 2018] Hay, N.; Stark, M.; Schlegel, A.; Wendelken, C.; Park, D.; Purdy, E.; Silver, T.; Phoenix, D. S.; and George, D. 2018. Behavior is everything—towards representing concepts with sensorimotor contingencies. *AAAI Conference on Artificial Intelligence*.
- [Hester et al. 2018] Hester, T.; Vecerik, M.; Pietquin, O.; Lanctot, M.; Schaul, T.; Piot, B.; Horgan, D.; Quan, J.; Sendonaris, A.; Osband, I.; et al. 2018. Deep Q-learning from demonstrations. *AAAI Conference on Artificial Intelligence*.
- [Huang et al. 2019] Huang, D.-A.; Nair, S.; Xu, D.; Zhu, Y.; Garg, A.; Fei-Fei, L.; Savarese, S.; and Niebles, J. C. 2019. Neural task graphs: Generalizing to unseen tasks from a single video demonstration. In *Computer Vision and Pattern Recognition*.
- [Johnson et al. 2016] Johnson, M.; Hofmann, K.; Hutton, T.; and Bignell, D. 2016. The malmo platform for artificial intelligence experimentation. In *IJCAI*, 4246–4247.
- [Lázaro-Gredilla et al. 2019] Lázaro-Gredilla, M.; Lin, D.; Guntupalli, J. S.; and George, D. 2019. Beyond imitation: Zero-shot task transfer on robots by learning concepts as cognitive programs. *Science Robotics* 4(26).
- [Long, Shelhamer, and Darrell 2015] Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*.
- [Manning and Schütze 1999] Manning, C. D., and Schütze, H. 1999. *Foundations of statistical natural language processing*. MIT press.
- [Mitchell 1978] Mitchell, T. M. 1978. Version spaces: an approach to concept learning. Technical report, Stanford University.
- [Nair et al. 2018] Nair, A.; McGrew, B.; Andrychowicz, M.; Zaremba, W.; and Abbeel, P. 2018. Overcoming exploration in reinforcement learning with demonstrations. *International Conference on Robotics and Automation*.
- [Natarajan et al. 2011] Natarajan, S.; Joshi, S.; Tadepalli, P.; Kersting, K.; and Shavlik, J. 2011. Imitation learning in relational domains: A functional-gradient boosting approach. In *Twenty-Second International Joint Conference on Artificial Intelligence*.
- [Niekum 2013] Niekum, S. D. 2013. *Semantically grounded learning from unstructured demonstrations*. Ph.D. Dissertation, University of Massachusetts, Amherst.
- [Parisotto et al. 2016] Parisotto, E.; Mohamed, A.-r.; Singh, R.; Li, L.; Zhou, D.; and Kohli, P. 2016. Neuro-symbolic program synthesis. *arXiv preprint arXiv:1611.01855*.
- [Pedregosa et al. 2011] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.
- [Quinlan 1986] Quinlan, J. R. 1986. Induction of decision trees. *Machine learning* 1(1):81–106.
- [Schaal 1997] Schaal, S. 1997. Learning from demonstration. *Advances in Neural Information Processing Systems*.

[Shah et al. 2018] Shah, A.; Kamath, P.; Shah, J. A.; and Li, S. 2018. Bayesian inference of temporal task specifications from demonstrations. In *Advances in Neural Information Processing Systems*, 3804–3813.

[Sun et al. 2018] Sun, S.-H.; Noh, H.; Somasundaram, S.; and Lim, J. 2018. Neural program synthesis from diverse demonstration videos. In *International Conference on Machine Learning*.

[Sutton 2019] Sutton, R. 2019. The bitter lesson.

[Ullman 1987] Ullman, S. 1987. Visual routines. *Readings in Computer Vision* 298–328.

[Valiant 1985] Valiant, L. G. 1985. Learning disjunction of conjunctions. *International Joint Conference on Artificial Intelligence*.

[Verma et al. 2018] Verma, A.; Murali, V.; Singh, R.; Kohli, P.; and Chaudhuri, S. 2018. Programmatically interpretable reinforcement learning. *International Conference on Machine Learning*.

[Wingate et al. 2011] Wingate, D.; Goodman, N. D.; Roy, D. M.; Kaelbling, L. P.; and Tenenbaum, J. B. 2011. Bayesian policy search with policy priors. *International Joint Conference on Artificial Intelligence*.

[Wingate et al. 2013] Wingate, D.; Diuk, C.; O’Donnell, T.; Tenenbaum, J.; and Gershman, S. 2013. Compositional policy priors. Technical report, Massachusetts Institute of Technology.

	Production rule	Probability
Programs		
$P \rightarrow \text{at_cell_with_value}(V, C)$	0.5	
$P \rightarrow \text{at_action_cell}(C)$	0.5	
Conditions		
$C \rightarrow \text{shifted}(O, B)$	0.5	
$C \rightarrow B$	0.5	
Base conditions		
$B \rightarrow \text{cell_is_value}(V)$	0.5	
$B \rightarrow \text{scanning}(O, C, C)$	0.5	
Offsets		
$O \rightarrow (N, 0)$	0.25	
$O \rightarrow (0, N)$	0.25	
$O \rightarrow (N, N)$	0.5	
Numbers		
$N \rightarrow \mathbb{N}$	0.5	
$N \rightarrow -\mathbb{N}$	0.5	
Natural numbers (for $i = 1, 2, \dots$)		
$\mathbb{N} \rightarrow i$	$(0.99)(0.01)^{i-1}$	
Values (for each value v in this game)		
$V \rightarrow v$	$1/ V $	

Table 2: The prior $p(f)$ over programs, specified as a probabilistic context-free grammar (PCFG).

The algorithm for LPP learning presented in the main text can be understood as performing variational inference, iteratively minimizing the KL divergence from q to the true posterior. We make two observations to this effect. First, fixing the K component policies in the support of the mixture q , the component weights that minimize the KL divergence $D_{KL}(q(\pi) \parallel p(\pi \mid \mathcal{D}))$ are $q(\mu_j) = \frac{p(\mu_j \mid \mathcal{D})}{\sum_{i=1}^K p(\mu_i \mid \mathcal{D})}$, that is, each policy should be weighted according to its relative posterior probability. Second, if some policy μ_* not in the support of q has higher posterior probability than some existing mixture component μ_i , a lower KL is always achievable by replacing μ_i with μ_* and reweighting the mixture accordingly. Our algorithm maintains a list μ_1, \dots, μ_K of the K best policies seen so far, and iteratively searches an increasingly large space for new policies μ_* of high posterior probability. Whenever a μ_* is found that has higher (unnormalized) posterior probability than the lowest-scoring policy μ_i in our “best- K ” list, replace μ_i with μ_* . At each iteration, our variational approximation is simply $q(\mu_j) = \frac{p(\mu_j \mid \mathcal{D})}{\sum_{i=1}^K p(\mu_i \mid \mathcal{D})}$ for $j \in \{1, \dots, K\}$.

Environment Details

Nim

Task Description There are two columns (piles) of matchsticks and empty cells. Clicking on a matchstick cell changes

all cells above and including the clicked cell to empty; clicking on an empty cell has no effect. After each matchstick cell click, a second player takes a turn, selecting another matchstick cell. The second player is modeled as part of the environment transition and plays optimally. When there are multiple optimal moves, one is selected randomly. The objective is to remove the *last* matchstick cell.

Task Instance Distribution Instances are generated procedurally. The height of the grid is selected randomly between 2 and 20. The initial number of matchsticks in each column is selected randomly between 1 and the height with the constraint that the two columns cannot be equal. All other grid cells are empty.

Expert Policy Description The winning tactic is to “level” the columns by selecting the matchstick cell next to an empty cell and diagonally up from another matchstick cell. Winning the game requires perfect play.

Checkmate Tactic

Task Description This task is inspired by a common checkmating pattern in Chess. Note that only three pieces are involved in this game (two kings and a white queen) and that the board size may be $H \times W$ for any H, W , rather

than the standard 8×8 . Initial states in this game feature the black king somewhere on the boundary of the board, the white king two cells adjacent in the direction away from the boundary, and the white queen attacking the cell in between the two kings. Clicking on a white piece (queen or king) *selects* that piece for movement on the next action. Note that a selected piece is a distinct value from a non-selected piece. Subsequently clicking on an empty cell moves the selected piece to that cell if that move is legal. All other actions have no effect. If the action results in a checkmate, the game is over and won; otherwise, the black king makes a random legal move.

Task Instance Distribution Instances are generated procedurally. The height and width are randomly selected between 5 and 20. A column for the two kings is randomly selected among those not adjacent to the left or right sides. A position for the queen is randomly selected among all those spaces for which the queen is threatening checkmate between the kings. The board is randomly rotated among the four possible orientations.

Expert Policy Description The winning tactic selects the white queen and moves it to the cell between the kings.

Chase

Task Description This task features a stick figure agent, a rabbit adversary, walls, and four arrow keys. At each time step, the adversary randomly chooses a move (up, down, left, or right) that increases its distance from the agent. Clicking an arrow key moves the agent in the corresponding direction. Clicking a gray wall has no effect other than advancing time. Clicking an empty cell creates a new (blue) wall. The agent and adversary cannot move through gray or blue walls. The objective is to “catch” the adversary, that is, move the agent into the same cell. It is not possible to catch the adversary without creating a new wall; the adversary will always be able to move away before capture.

Task Instance Distribution There is not a trivial parameterization to procedurally generate these task instances, so they are manually generated.

Expert Policy Description The winning tactic advances time until the adversary reaches a corner, then builds a new wall next to the adversary so that it is trapped on three sides, then moves the agent to the adversary.

Stop the Fall

Task Description This task involves a parachuter, gray static blocks, red “fire”, and a green button that turns on gravity and causes the parachuter and blocks to fall. Clicking an empty cell creates a blue static block. The game is won when gravity is turned on and the parachuter falls to rest without touching (being immediately adjacent to) fire.

Task Instance Distribution There is not a trivial parameterization to procedurally generate these task instances, so they are manually generated.

Expert Policy Description The winning tactic requires building a stack of blue blocks below the parachuter that is high enough to prevent contact with fire, and then clicking the green button.

Reach for the Star

Task Description In this task, a robot must move to a cell with a yellow star. Left and right arrow keys cause the robot to move. Clicking on an empty cell creates a dynamic brown block. Gravity is always on, so brown objects fall until they are supported by another brown block. If the robot is adjacent to a brown block and the cell above the brown block is empty, the robot will move on top of the brown block when the corresponding arrow key is clicked. (In other words, the robot can only climb one block, not two or more.)

Task Instance Distribution Instances are generated procedurally. A height for the star is randomly selected between 2 and 11 above the initial robot position. A column for the star is also randomly selected so that it is between 0 and 5 spaces from the right border. Between 0 and 5 padding rows are added above the star. The robot position is selected between 0 and 5 spaces away from where the start of the minimal stairs would be. Between 0 and 5 padding columns are added to the left of the agent. The grid is flipped along the vertical axis with probability 0.5.

Expert Policy Description The winning tactic requires building stairs between the star and robot and then moving the robot up them.

Fence In

Task Description This task uses the Malmo interface to Minecraft (Johnson et al. 2016). As in the other tasks, observations are 2D grids with discrete values and actions are “clicks” on the grid. This task features fences, one sheep, and open space in the grid. Clicking on a grid cell has the effect of building a fence at that location if it is currently open. The objective of the task is to build fences in such a way that the sheep is completely enclosed by fences. (Enclosure is calculated by finding the connected component that contains the sheep and checking whether it reaches the boundary of the grid.)

Task Instance Distribution There is not a trivial parameterization to procedurally generate these task instances, so they are manually generated.

Expert Policy Description The expert policy builds fences from left to right one cell below the sheep, starting at the nearest fence on the left and finishing at the nearest fence on the right.

Additional Results

Data Efficiency of Baselines

We saw in our main experiments that generalizable convolutional and fully convolutional policies cannot be learned in our regime of very low data and very high variation between task instances. In an effort to better understand the gap in data efficiency between the deep policies (CNN and FCN)

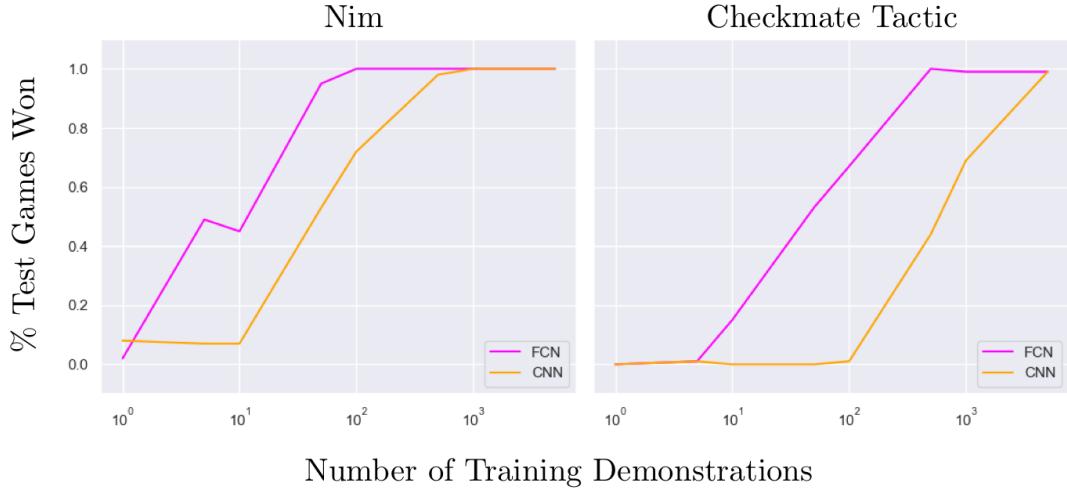


Figure 8: Performance on held-out test task instances as a function of the number of training demonstrations for CNN and FCN baselines.

and LPP, we again measure the test performance of the baselines as a function of the number of demonstrations, but with many more demonstrations than we previously considered. In the main paper, we report results for up to 10 demonstrations; here we report results for up to 5,000. We focus on Nim and Checkmate Tactic for this experiment. Recall that at most 5 demonstrations were necessary to learn a winning (test performance 1.0) policy in LPP. In Figure 8, we see that FCNs require 100 demonstrations in Nim and 500 demonstrations in Checkmate Tactic to achieve test performance 1.0. We also find that CNNs require 1,000 demonstrations in Nim and 5,000 demonstrations in Checkmate Tactic. Thus LPP achieves 20 – 1,000x data efficiency gains over these baselines.

Learned Policy Analysis

Here we analyze the learned LPP policies in an effort to understand their representational complexity and qualitative behavior. We enumerate 10,000 programs and train with 11 task instances for each task. Table 3 reports three statistics for the MAP policies learned for each task: the number of top-level programs (at_action_cell and at_cell_with_value calls), the total number of method calls (shifted, scanning, and cell_is_value as well); and the maximum parse tree depth among all programs in the policy. The latter term is the primary driver of learning complexity in practice. The number of top-level programs ranges from 11 to 97; the number of method calls ranges from 32 to 244; and the maximum parse tree depth is 3 or 4 in all tasks. These numbers suggest that LPP learning is capable of discovering policies of considerable complexity, far more than what would be feasible with full policy enumeration. We also note that the learned policies are very sparse, given that the maximum number of available top-level programs is 10,000. Visualizing the policies offers further confirmation of their strong performance (Figures 5–9).

An example of a clause in the learned MAP policy for Stop the Fall is shown in Figure 9. This clause is used to select the first action: the cell below the parachuter and immediately above the floor is clicked. Other clauses govern the policy following this first action, either continuing to build blocks above the first one, or clicking the green button to turn on gravity. From the example clause, we see that the learned policy may include some redundant programs, which unnecessarily decrease the prior without increasing the likelihood. Such redundancy is likely due to the approximate nature of our Boolean learning algorithm (greedy decision-tree learning). Posthoc pruning of the policies or alternative Boolean learning methods could address this issue. The example clause also reveals that the learned policy may take unnecessary (albeit harmless) actions in the case where there are no “fire” objects to avoid. In examining the behavior of the learned policies for Stop the Fall and the other games, we do not find any unnecessary actions taken, but this does not preclude the possibility in general. Despite these two qualifications, the clause and others like it are a positive testament to the interpretability and intuitiveness of the learned policies.

Noisy Demonstrations

Our main experiments used demonstrations that were sampled following the expert policy exactly. In realistic settings, it is often the case that demonstrations are corrupted by noise, e.g., due to human error. We can accommodate noise in learning by modifying the likelihood $p(\mathcal{D} | \pi)$ to include a noise model. A simple noise model that we consider here assumes that the expert policy is followed with probability $(1 - \epsilon)$, and with probability ϵ , a random action is taken. (This is akin to the expert following an ϵ -greedy policy.) The likelihood then becomes $p(\mathcal{D} | \pi) \propto \prod_{i=1}^N \prod_{j=1}^{T_i} (1 - \epsilon)\pi(a_{ij} | s_{ij}) + \frac{\epsilon}{|\mathcal{A}|}$.

In this set of experiments, we investigate how the noise

```

at_action_cell(scanning(↑, cell_is_value(↖)), cell_is_value(█)) ∧
at_action_cell(scanning(↑, cell_is_value(↖), cell_is_value(█)), cell_is_value(█)) ∧
at_action_cell(cell_is_value(█)) ∧
not at_action_cell(shifted(↓, cell_is_value(█))) ∧
not at_action_cell(cell_is_value(█)) ∧
at_action_cell(shifted(↓, cell_is_value(█)))

```

Figure 9: One of the clauses of the learned MAP policy for “Stop the Fall”. The clause suggests clicking a cell if scanning up we find a parachuter before a drawn or static block; if the cell is empty; if the cell above is not drawn; if the cell is not a green button; and if a static block is immediately below. Note that this clause is slightly redundant and may trigger an unnecessary action for a task instance where there are no “fire” cells.

Task	Number of Programs	Method Calls	Max Depth
Nim	11	32	4
Checkmate Tactic	23	60	3
Chase	97	244	3
Stop the Fall	17	50	4
Reach for the Star	34	151	4
Fence In	9	40	3

Table 3: Analysis of the MAP policies learned in LPP with 10,000 programs enumerated and all 11 training demonstrations. The number of top-level programs, the number of constituent method calls, and the maximum parse depth of a program in the policy is reported for each of the five tasks.

model affects learning and generalization if demonstrations are (1) perfect, (2) corrupted by independent random noise, and (3) corrupted by correlated random noise. There will be mismatch between the noise model and noise source except for the case where the independent random noise probability of the demonstrations is equal to ϵ in the noise model.

Demonstrations We use the Nim task for all experiments. The perfect demonstrations (1) are identical to the main experiments. Independent random noise (2) is introduced by taking a random action with probability 0.2. Correlated random noise (3) is introduced by clicking on a random *token* cell with probability 0.2. Recall that if a single incorrect token is clicked in Nim, the game is inevitably lost.

Methods and Results We modify the likelihood as described above and consider $\epsilon \in \{0.0, 1e-4, 1e-3, 1e-2, 1e-1, 1.0\}$. All hyperparameters are otherwise unchanged from the main experiments. For each value of ϵ , for each of the three sets of demonstrations, we perform LPP learning with demonstrations ranging from 2 to 16.

Results are shown in Figure 10. Across all demonstration types, we first note that too high of an ϵ results in a severe drop in performance. By inspecting the learned policies, we see that a high ϵ results in overly simplified policies that are favored by the prior; all actions not consistent with this simple policy are treated as noise. An ϵ of 0.0 also results in a severe drop in performance for the demos with noise, but for

a different reason: when no noise is assumed by the model, no policy consistent with the noisy demos is found within the allotted enumeration budget. Interestingly, a very small but nonzero ϵ seems to be consistently effective, even though the actual probability of a random action in the noisy demos is 0.2. This result suggests that the primary benefit of modifying the likelihood is to avoid discarding policy candidates that nearly, but not perfectly, match the demos.

Atari Breakout

As our principal focus in this work is few-shot generalization, the experiments described thus far have been carried out in a suite of grid game tasks that were designed to exhibit substantial structured variation. Popular reinforcement learning benchmarks such as Atari 2600 games involve relatively little variation between task instances and were therefore not selected for our main experiments. Nonetheless, the question of whether LPP policy learning and the particular DSL used for our grid games can be extended to Atari games and other familiar benchmarks is interesting, as it sheds light on the generality of the approach. In this section, we begin to answer this question with some preliminary experiments and results on Atari Breakout.

Task In Breakout (Figure 11), an agent must move a paddle left or right so that it hits a ball into bricks above. When a brick is hit, it disappears, and a reward of 1 is given. If the ball falls below the paddle, the episode is over. The ball’s initial position and velocity vary between task instances.

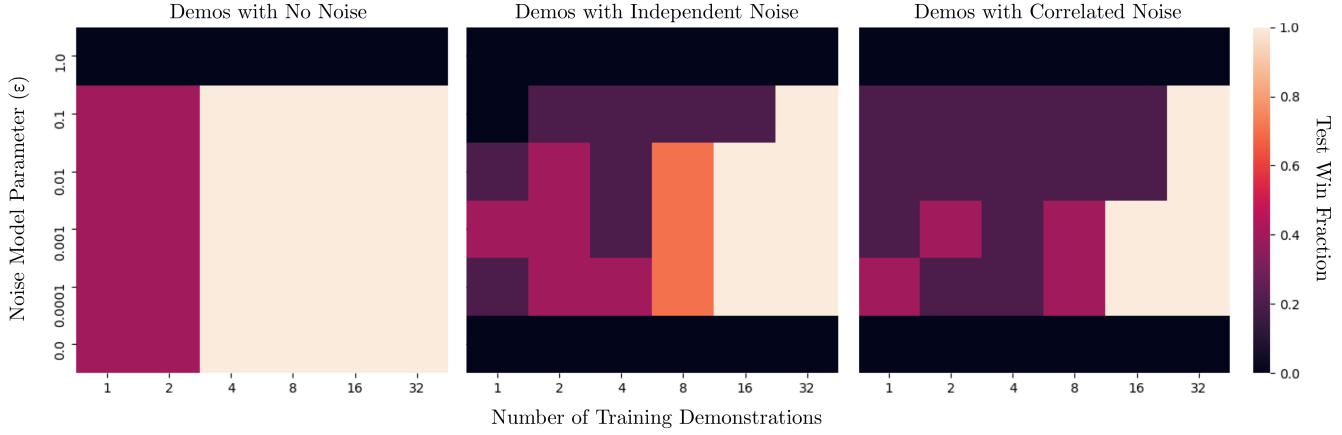


Figure 10: Effect of noisy demonstrations on learning performance and generalization in Nim. See text for details.

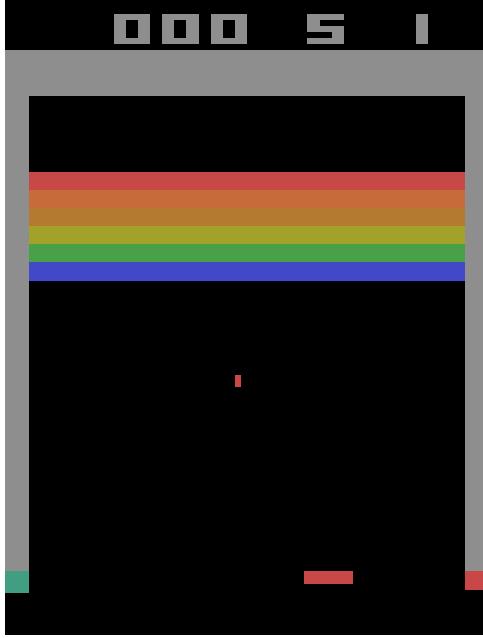


Figure 11: Atari Breakout.

We use the Breakout environment made available through OpenAI Gym (“BreakoutNoFrameskip-v4”) with three standard wrappers²: “EpisodicLifeEnv”, which ensures that the episode ends when the ball falls below the paddle; “FireResetEnv”, which executes the “fire” action at the beginning of the game to make the ball appear; and “NoopResetEnv”, which randomizes the initial position of the ball between task instances. Raw observations are 210×160 RGB images. We stack two consecutive observations to capture velocity

²These wrappers are implemented in the “stable baselines” package (<https://github.com/hill-a/stable-baselines>). Equivalent preprocessing was done in the original DQN paper and downstream work.

information, resulting in size $210 \times 160 \times 3 \times 2$ tensors³. We convert these tensors into the familiar 2D grid representation by flattening the latter two dimensions so that each 3×2 tensor is converted to a single float value. In the data, there are 15 unique flattened values across all pixels in all observations. The final observation space is thus isomorphic to a 2D grid with 15 discrete values, that is, $\{0, 1, \dots, 14\}^{210 \times 160}$.

Expert Policy We implement a simple reactive policy from which we can draw expert demonstrations. Given an observation, we first determine the velocity of the ball based on the two consecutive frames. We further determine whether the paddle is to the left or right of the ball. If the ball is moving towards the left and the paddle is not on the left, then we take the “left” action, and the same for the opposite direction. We take the noop action otherwise. This policy is not necessarily optimal, but it successfully keeps the ball in play and continues to reap rewards for several thousand frames.

Methods and Results We use the same DSL described in the main paper for this experiment, with one change: the `at_action_cell` method of the DSL is designed under the assumption that actions are clicks on a grid cell, but the action space in Atari Breakout is simply four discrete values (“left”, “right”, “fire”, “noop”); thus we remove this method from the DSL and replace it with four simple identity checks (`action_is_left`, `action_is_right`, `action_is_fire`, `action_is_noop`). The probabilistic grammatical prior is modified accordingly, giving uniform probability to these four action methods and to the `at_cell_with_value` method in the first substitution.

We sample a single demonstration for 500 frames using the expert policy described above. This demonstration shows 3 bricks breaking for a total reward of 3. We then run LPP learning with 120,000 programs enumerated. (Fewer programs were found to be insufficient.) To ease the computational burden, we enumerate programs involving only 4 of

³Stacking four frames is standard for Atari but redundant for Breakout

the 15 total discrete values, namely those involving the color of the paddle and ball. This is a problem-specific hack that injects additional prior knowledge. However, we do expect that results would be similar if all discrete values were included and more programs (on the order of 5 million) were enumerated.

To evaluate the learned policy, we test on 10 held-out task instances for 500 frames. (Recall that variation between task instances is parameterized by the initial position and velocity of the ball.) In all 10 cases, the learned policy accrues a reward of 3. This matches the expert policy and suggests perfect generalization from only a single relatively short demonstration, at least for the limited task horizon of 500. To further assess generalization, we continue past the horizon of 500 and record the total reward accumulated before the ball falls below the paddle. Across the 10 held-out task instances, we find an average reward of 8.80 with a standard deviation of 7.47. These results indicate that the learned policy can generalize immediately to observations with more bricks missing in different locations than seen during the single demonstration.

These preliminary results suggest that the policy learning method and the specific DSL proposed in this work may extend to other tasks with minor modification, including games that are not typically thought of as grid-based. We look forward to continuing to evaluating the scalability and applicability of the DSL and general LPP policy learning in future work.

Training and Test Environments

In the following figures, we illustrate each task with representative training demonstrations and LPP test performance. (See the main text for Fence In.)

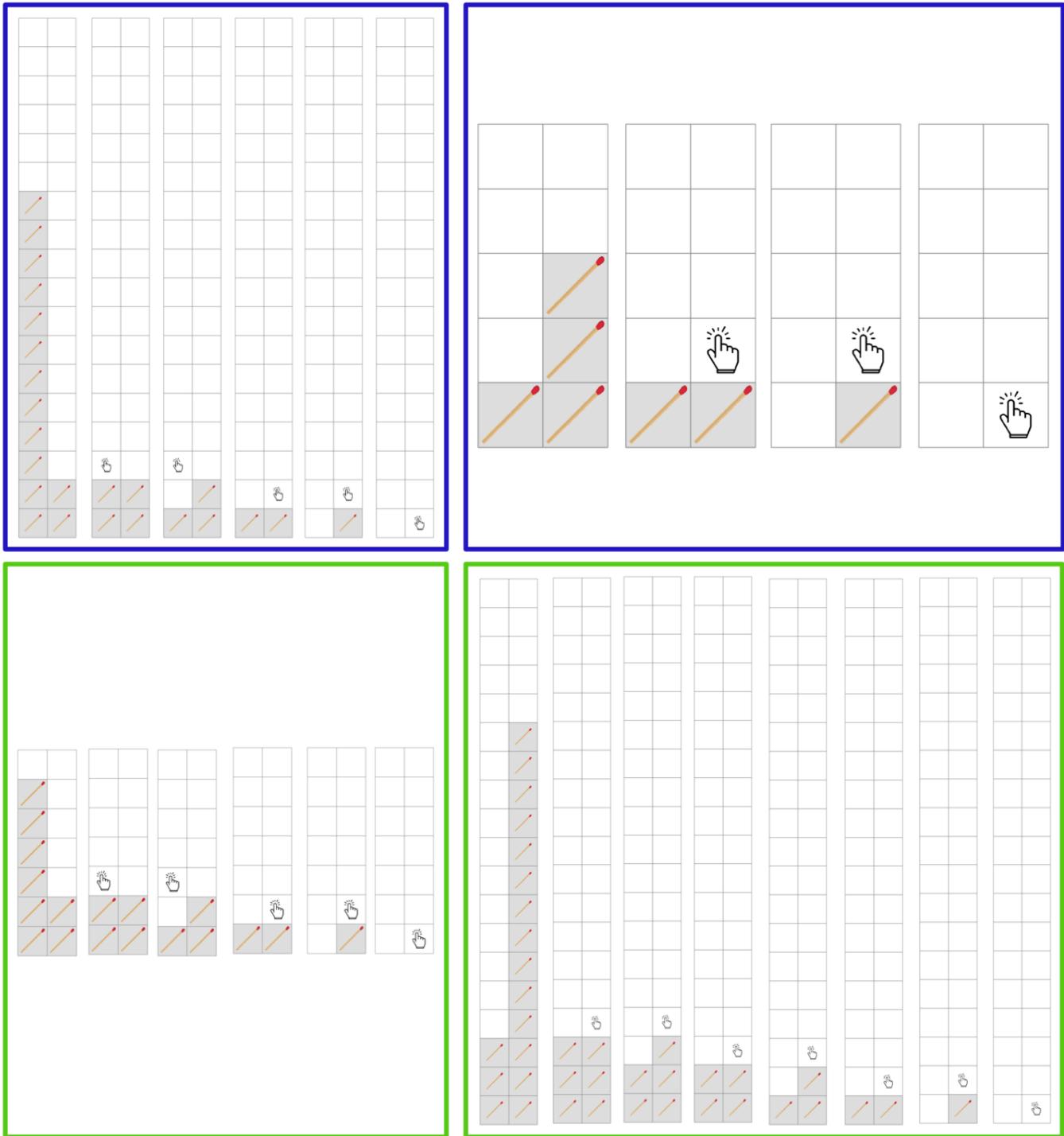


Figure 12: A LPP policy learned from two demonstrations of “Nim” (blue) generalizes perfectly to all test task instances (e.g. green).

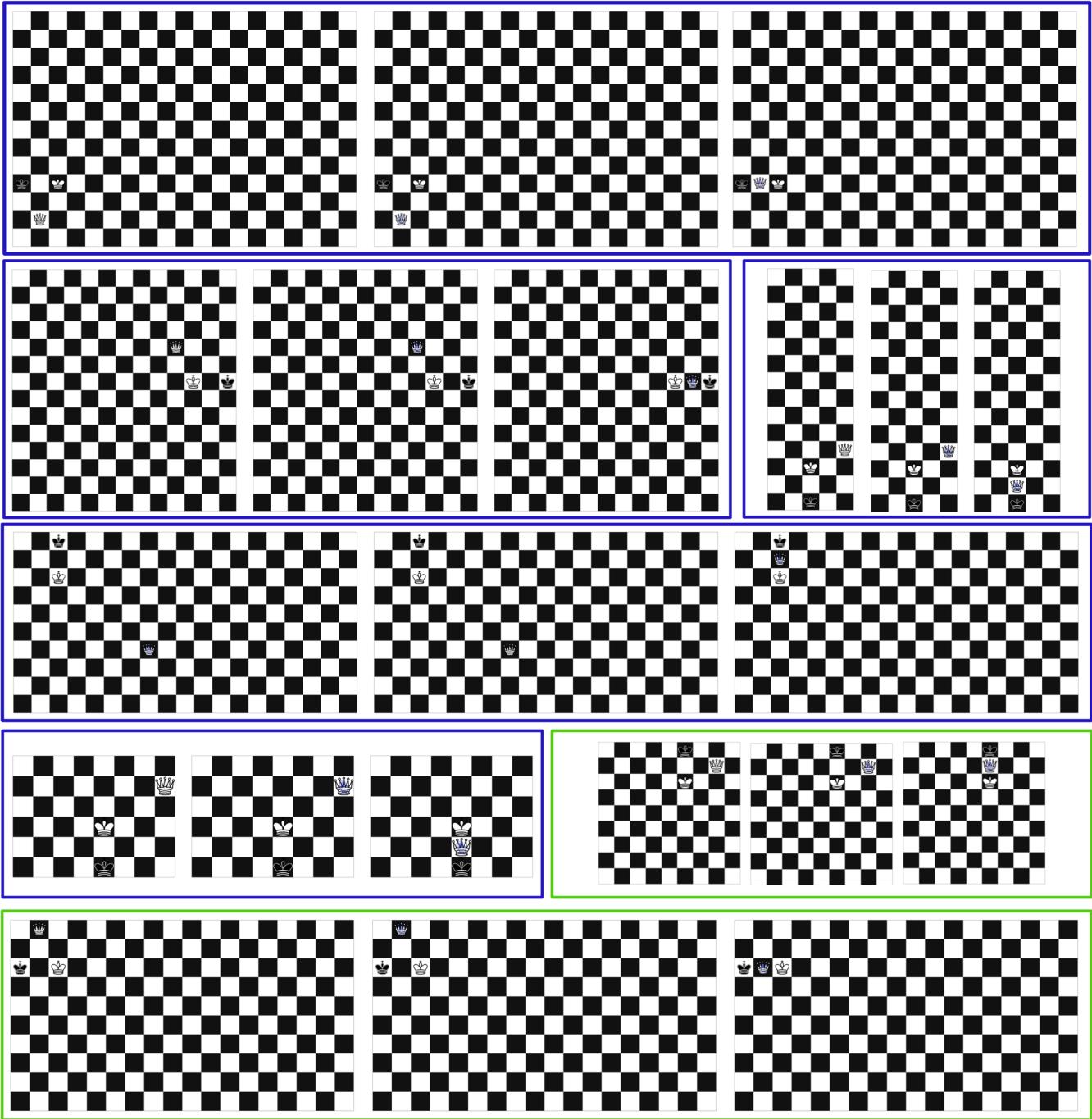


Figure 13: A LPP policy learned from five demonstrations of “Checkmate Tactic” (blue) generalizes perfectly to all test task instances (e.g. green).

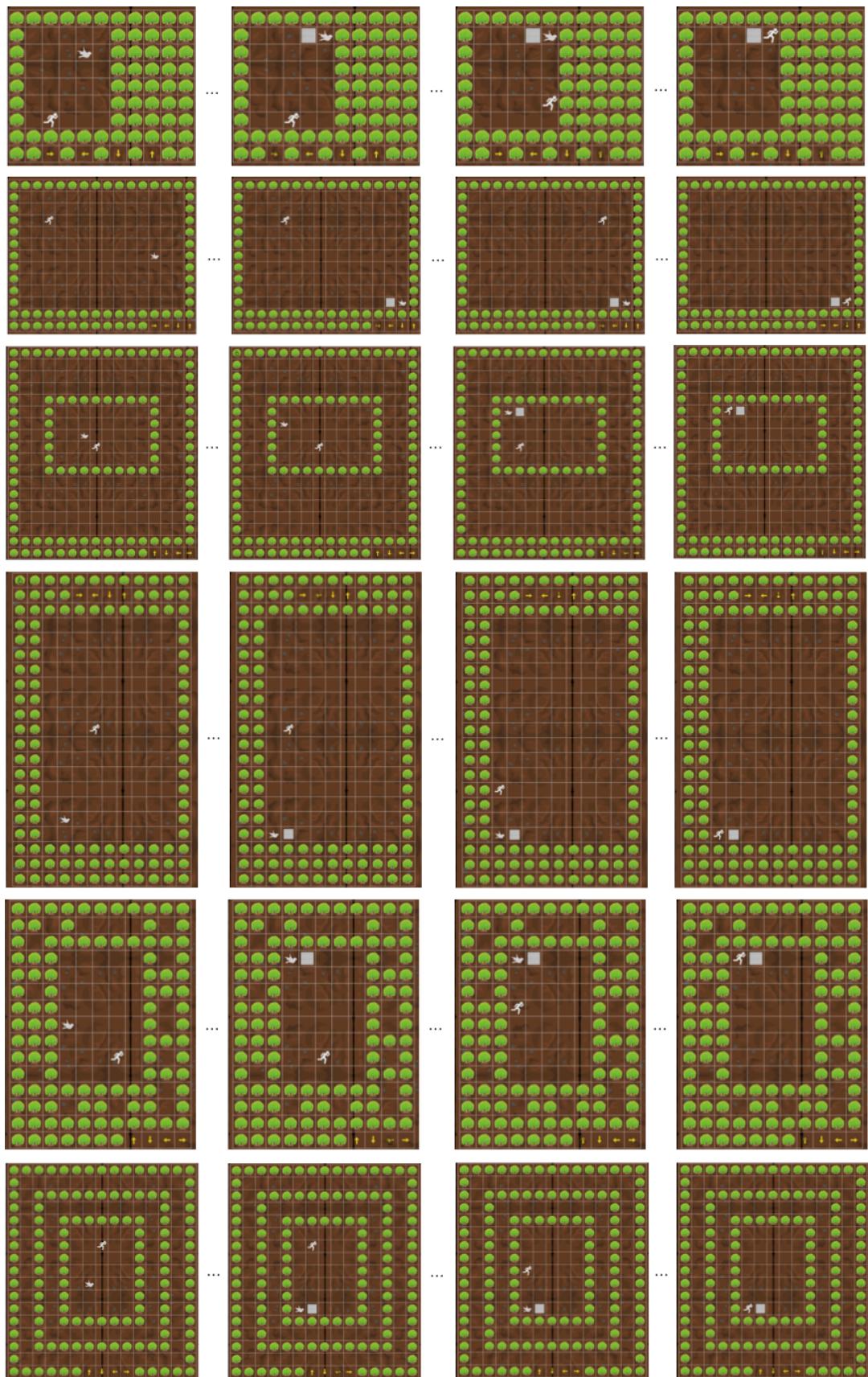


Figure 14: A LPP policy learned from four demonstrations of “Chase” (top) generalizes perfectly to all test task instances (e.g. bottom).

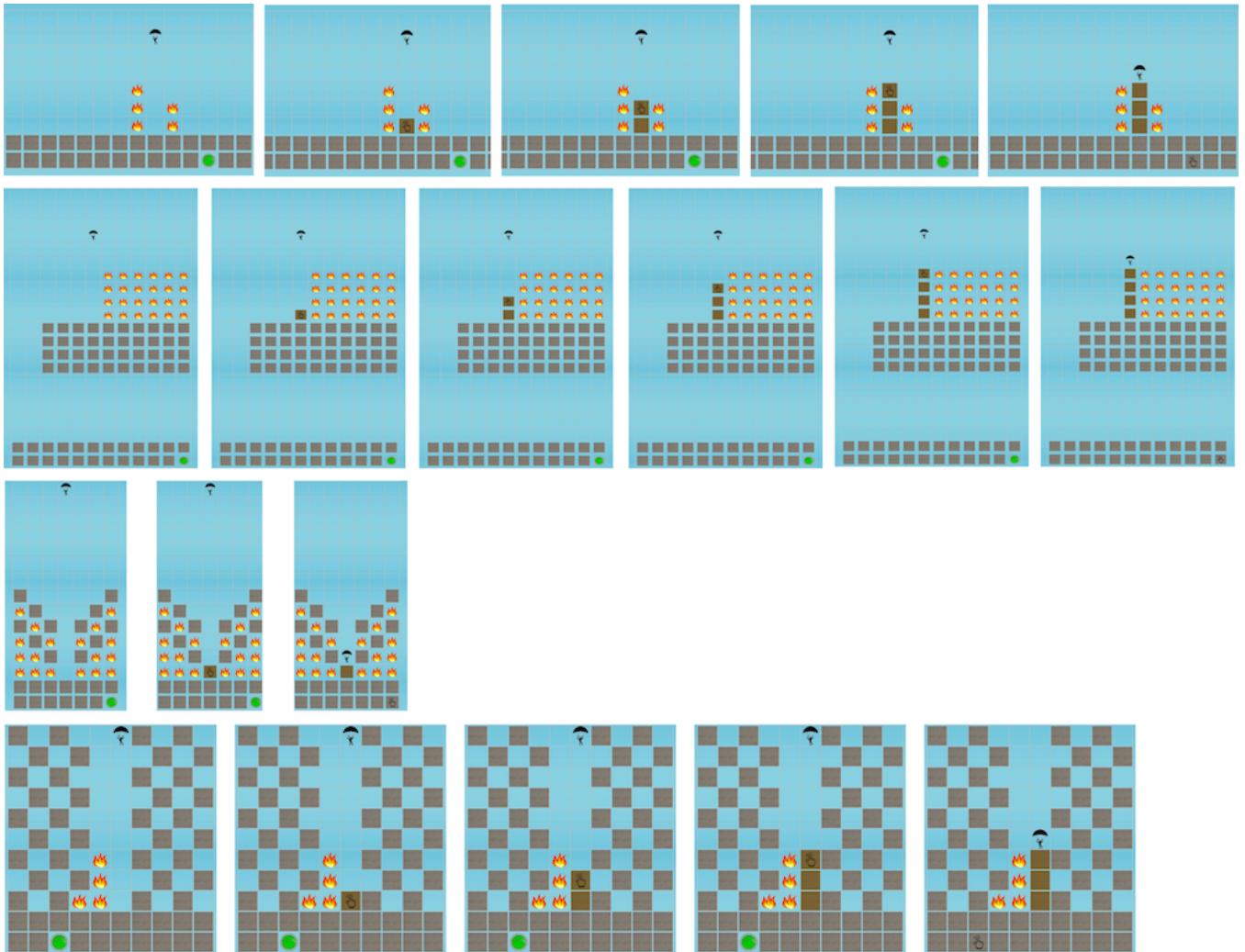


Figure 15: A LPP policy learned from three demonstrations of “Stop the Fall” (top) generalizes perfectly to all test task instances (e.g. bottom).

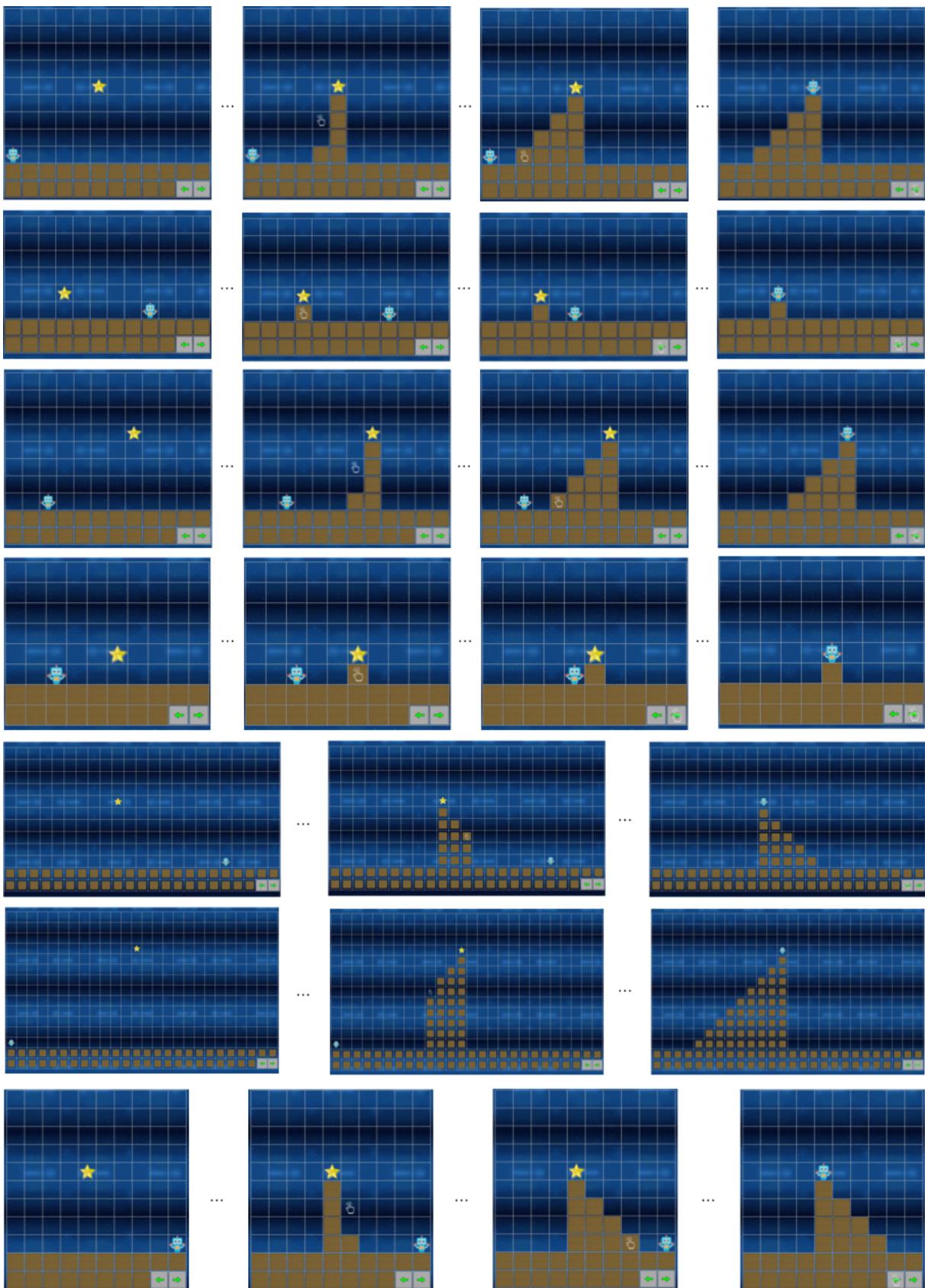


Figure 16: A LPP policy learned from five demonstrations of “Reach for the Star” (top) generalizes perfectly to all test task instances (e.g. bottom).