Abstract

We present a framework for learning hierarchical policies from demonstrations, using sparse natural language annotations to guide the discovery of reusable skills for autonomous decision-making. We formulate a generative model of action sequences in which goals generate sequences of high-level subtask descriptions, and these descriptions generate sequences of low-level actions. We describe how to train this model using primarily unannotated demonstrations by parsing demonstrations into sequences of named high-level subtasks, using only a small number of seed annotations to ground language in action. In trained models, the space of natural language commands indexes a combinatorial library of skills; agents can use these skills to plan by generating high-level instruction sequences tailored to novel goals. We evaluate this approach in the ALFRED household simulation environment, providing natural language annotations for only 10% of demonstrations. It completes more than twice as many tasks as a standard approach to learning from demonstrations, matching the performance of instruction sequencing to a sequence of low-level actions. In trained models, the space of natural language commands indexes a combinatorial library of skills; agents can use these skills to plan by generating high-level instruction sequences tailored to novel goals. We evaluate this approach in the ALFRED household simulation environment, providing natural language annotations for only 10% of demonstrations. It completes more than twice as many tasks as a standard approach to learning from demonstrations, matching the performance of instruction sequencing to a sequence of low-level actions.

1 Introduction

Building autonomous agents that integrate high-level reasoning with low-level perception and control is a long-standing challenge in artificial intelligence (Fikes et al., 1972; Newell, 1973; Sacerdoti, 1973; Brockett, 1993). Fig. 1 shows an example: to accomplish a task such as cooking an egg, an agent must first find the egg, then grasp it, then locate a stove or microwave, at each step reasoning about both these subtasks and complex, unstructured sensor data. Hierarchical planning models (e.g. Sutton et al., 1999)—which first reason about abstract states and actions, then ground these in concrete control decisions—play a key role in most existing agent architectures. But training effective hierarchical models for general environments and goals remains difficult. Standard techniques either require detailed formal task specifications, limiting their applicability in complex and hard-to-formalize environments, or are restricted to extremely simple high-level actions, limiting their expressive power (Bacon et al., 2017; Sutton et al., 1999; Dietterich, 1999; Kaelbling and Lozano-Pérez, 2011).

Several recent papers have proposed to overcome these limitations using richer forms of supervision—especially language—as a scaffold for hierarchical policy learning. In latent language policies (LLPs; Andreas et al., 2018; Shiarlis et al., 2018),
controllers first map from high-level goals to sequences of natural language instructions, then use instruction following models to translate those instructions into actions. But applications of language-based policy learning have remained quite limited in scope.

Current LLP training approaches treat language as a latent variable only during prediction, and require fully supervised (and often impractically large) datasets that align goal specifications with instructions and instructions with low-level actions. As a result, all existing work on language-based policy learning has focused on very short time horizons (Andreas et al., 2018), restricted language (Hu et al., 2019; Jacob et al., 2021) or synthetic training data (Shu et al., 2018; Jiang et al., 2019).

In this paper, we show that it is possible to train language-based hierarchical policies that outperform state-of-the-art baselines using only minimal natural language supervision. We introduce a procedure for weakly and partially supervised training of LLPs using ungrounded text corpora, unlabeled demonstrations, and a small set of annotations linking the two. To do so, we model training demonstrations as generated by latent high-level plans: we describe a deep, structured latent variable model in which goals generate subtask descriptions and subtask descriptions generate actions. We show how to learn in this model by performing inference in the infinite, combinatorial space of latent plans while using a comparatively small set of annotated demonstrations to seed the learning process.

Using an extremely reduced version of the ALFRED household robotics dataset (Shridhar et al., 2020)—with 10% of labeled training instructions, no alignments during training, and no instructions at all during evaluation—our approach matches or outperforms existing approaches that use ground-truth instructions and alignments during both training and evaluation. It correctly segments and labels subtasks in unlabeled demonstrations, including subtasks that involve novel compositions of actions and objects. Additional experiments show that pretraining on large (ungrounded) text corpora (Raffel et al., 2020) contributes to this success, demonstrating one mechanism by which background knowledge encoded in language can benefit tasks that do not involve language as an input or an output.

Indeed, our results show that relatively little information about language grounding is needed for effective learning of language-based policies—a rich model of natural language text, a large number of demonstrations, and a small number of annotations suffice for learning compositional libraries of skills and effective policies for deploying them.

2 Preliminaries

We consider learning problems in which agents must perform multi-step tasks (like cooking an egg; Fig. 1) in interactive environments. We formalize these problems as undiscounted, episodic, partially observed Markov decision processes (POMDPs) defined by a tuple \((\mathcal{S}, \mathcal{A}, T, \Omega, O)\), where \(\mathcal{S}\) is a set of states, \(\mathcal{A}\) is a set of actions, \(T : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}\) is an (unknown) state transition function, \(\Omega\) is a set of observations, and \(O : \mathcal{S} \rightarrow \Omega\) is an (unknown) observation function.\(^2\) We assume that observations include a distinguished goal specification \(g\) that remains constant throughout an episode; given a dataset \(D\) of consisting of goals \(g\) and demonstrations \(d\) (i.e. \(D = \{(d_1, g_1), (d_2, g_2), \ldots\}\); \(d = [(a_1, a_2), (a_2, a_3), \ldots]; o \in \Omega, a \in \mathcal{A}\)), we aim to learn a goal-conditional policy \(\pi(a_t | a_{t-1}, o_{t-1}, g) \Rightarrow o_{t-1}, o_1, \ldots, o_t, g\) that generalizes demonstrated behaviors to novel goals and states.

For tasks like the ones depicted in Fig. 1, this learning problem requires agents to accomplish multiple subgoals (like finding an egg or operating an appliance) in a feasible sequence. As in past work, we address this challenge by focusing on hierarchical policy representations that plan over temporal abstractions of low-level action sequences. We consider a generic class of hierarchical policies that first predict a sequence of subtask specifications \(\tau\) from a distribution \(\pi^C(\tau_i | \tau_{i-1}, g)\) (the controller), then from each \(\tau\) generate a sequence of actions \(a_1 \ldots a_n\) from a distribution \(\pi^E(a_i | a_{i-1}, o_{i-1}, \tau)\) (the executor).\(^3\) At each timestep, \(\pi^E\) may either generate an action from \(\mathcal{A}\); or a special termination signal \(\text{STOP}\); after \(\text{STOP}\) is selected, control is returned to \(\pi^C\) and a new \(\tau\) is generated. This process is visualized in Fig. 2(a). Trajectories generated by hierarchical policies themselves have hierarchical structure: each subtask specification \(\tau\) generates a segment of a trajectory (delimited by a \(\text{STOP}\) action) that

\(^2\)For notational convenience, we assume without loss of generality that \(T\) and \(O\) are deterministic.

\(^3\)In past work, \(\pi^E\) often conditions on the current observation as well as goal and history of past subtask specifications; we found that this extra information was not needed for the tasks studied here.
Figure 2: (a) When a hierarchical policy is deployed, $\pi^C$ generates a sequence of subtask specifications, and $\pi^E$ translates each of these to a low-level action sequence ending in STOP. At training time, this hierarchical structure is not available, and must be inferred to train our model. To do so, we assign each action $a_i$ an auxiliary alignment variable $\alpha_i$ identifying the subtask that produced it. Alignments divide an action sequence into a sequence of segments $s$ containing actions aligned to the same subtask. Automatically segmenting training demonstrations makes it possible to learn modular, reusable policies for individual subtasks without direct supervision. (b) Overview of the proposed learning algorithm (SL)$^5$, which alternates between segmenting (by aligning) actions to fixed subtask specifications; labeling segments given fixed alignments, and updating model parameters.

3 Approach

Overview We train hierarchical policies on unannotated action sequences by inferring latent natural language descriptions of the subtasks they accomplish (Fig. 2(b)). We present a learning algorithm that jointly partitions these action sequences into smaller segments exhibiting reusable, task-general skills, labels each segment with a description, trains $\pi^C$ to generate subtask descriptions from goals, and trains $\pi^E$ to generate actions from subtask descriptions.

Formally, we assume access to two kinds of training data: a large collection of unannotated demonstrations $D = \{(d_1, g_1), (d_2, g_2), \ldots\}$ and a smaller collection of annotated demonstrations $D^{ann} = \{(d_1, g_1, \tau_1), (d_2, g_2, \tau_2), \ldots\}$, where each $\tau$ consists of a sequence of natural language instructions $[\tau_1, \tau_2, \ldots]$ corresponding to the subtask sequence that should be generated by $\pi^C$. We assume that even these annotated trajectories leave much of the structure depicted in Fig. 2(a) unspecified, containing no explicit segmentations or stop markers. Training $\pi^E$ thus requires inferring the correspondence between actions and annotations on $D^{ann}$ while inferring annotations themselves on $D$.

Training objective To begin, it will be convenient to have an explicit expression for the probability of a demonstration given a policy $(\pi^C, \pi^E)$. To do so, we first observe that the hierarchical generation procedure depicted in Fig. 2(a) produces a latent alignment between each action and the subtask $\tau$ that generated it. We denote these alignments $\alpha_i$, writing $\alpha_i = j$ to indicate that $a_i$ was generated from $\tau_j$. Because $\pi^E$ executes subtasks in sequence, alignments are monotonic, satisfying $\alpha_i = \alpha_{i-1}$ or
\( \alpha_i = \alpha_{i-1} + 1 \). Let \( \text{seg}(\alpha) \) denote the segmentation associated with \( \alpha \), the sequence of sequences of action indices \([i : \alpha_i = 1], [i : \alpha_i = 2], \ldots \) aligned to the same instruction (see Fig. 2(a)).

Then, for a fixed policy and POMDP, we may write the joint probability of a demonstration, goal, annotation, and alignment as:

\[
p(d, g, \tau, \alpha) \propto \prod_{s \in \text{seg}(\alpha)} \left[ \pi^C(\tau_s \mid \tau_{s-1}, g) \right. \\
\times \left( \prod_{i \in 1 \ldots |s|} \pi^E(a_i \mid a_{s_{i-1}}, o_{s_i}, \tau_{a_i}) \right) \\
\times \pi^E(\text{STOP} \mid a_s, o_s) \right]. \tag{1}
\]

Here : \( s - 1 \) (in a slight abuse of notation) denotes all segments preceding \( s \), and \( s_i \) is the index of the \( i \)th action in \( s \). The constant of proportionality in Eq. (1) depends only on terms involving \( T(s' \mid s, a), O(a \mid s) \) and \( p(g) \), all independent of \( \pi^C \) or \( \pi^E \); Eq. (1) thus describes the component of the data likelihood under the agent’s control (Ziebart et al., 2013).

With this definition, and given \( \mathcal{D} \) and \( \mathcal{D}^{\text{ann}} \) as defined above, we may train a latent language policy using partial natural language annotations via ordinary maximum likelihood estimation, imputing the missing segmentations and labels in the training set jointly with the parameters of \( \pi^C \) and \( \pi^E \) (which we denote \( \hat{\theta} \)) in the combined annotated and unannotated likelihoods:

\[
\arg \max_{\hat{\tau}, \hat{\alpha}, \hat{\theta}} \mathcal{L}(\hat{\tau}, \hat{\alpha}, \hat{\theta}) + \mathcal{L}^{\text{ann}}(\hat{\alpha}, \hat{\theta}) \tag{2}
\]

where

\[
\mathcal{L}(\hat{\tau}, \hat{\alpha}, \hat{\theta}) = \sum_{(d,g) \in \mathcal{D}} \log p(d, g, \hat{\tau}, \hat{\alpha}) \tag{3}
\]

\[
\mathcal{L}^{\text{ann}}(\hat{\alpha}, \hat{\theta}) = \sum_{(d,g,\tau) \in \mathcal{D}^{\text{ann}}} \log p(d, g, \tau, \hat{\alpha}) \tag{4}
\]

and where we have suppressed the dependence of \( p(d, g, \tau, \alpha) \) on \( \hat{\theta} \) for clarity. This objective involves continuous parameters \( \hat{\theta} \), discrete alignments \( \hat{\alpha} \), and discrete labelings \( \hat{\tau} \). We optimize it via block coordinate ascent on each of these components in turn: alternating between re-segmenting demonstrations, re-labeling those without ground-truth labels, and updating parameters. The full learning algorithm, which we refer to as (SL)³ (semi-supervised skill learning with latent language), is shown in Algorithm 1, with each step of the optimization procedure described below.

**Segmentation:** \( \arg \max_{\alpha} \mathcal{L}(\hat{\tau}, \alpha, \hat{\theta}) + \mathcal{L}^{\text{ann}}(\alpha, \hat{\theta}) \)

The segmentation step associates each low-level action with a high-level subtask by finding the highest scoring alignment sequence \( \alpha \) for each demonstration in \( \mathcal{D} \) and \( \mathcal{D}^{\text{ann}} \). While the number of possible alignments for a single demonstration is exponential in demonstration length, the fact that \( \pi^E \) depends only on the current subtask implies the following recurrence relation:

\[
\begin{align*}
\max_{\alpha_{1:n}} p(d_{1:n}, g, \tau_{1:m}, \alpha_{1:n}) \\
= \max_{i} \left( \max_{\alpha_{1:i}} p(d_{1:i}, g, \tau_{1:i-1}, \alpha_{1:i}) \right. \\
\times p(d_{i+1:n}, g, \tau_{m}, \alpha_{i+1:n} = m) \right) \tag{5}
\end{align*}
\]

This means that the highest-scoring segmentation can be computed by with an algorithm that recursively identifies the highest-scoring alignment to each prefix of the instruction sequence at each action (Algorithm 2), a process requiring \( O(|d||\tau|) \) space and \( O(|d||\tau|) \) time. The structure of this dynamic program is identical to the forward algorithm for hidden semi-Markov models (HSSMs), which are widely used in NLP for tasks like language generation and word alignment (Wiseman et al., 2018). Indeed, Algorithm 2 can be derived immediately from Eq. (5) by interpreting \( p(d, g, \tau, \alpha) \) as the output distribution for an HSMM in which emissions are actions, hidden states are alignments, the emission distribution is \( \pi^E \) and the transition distribution is the deterministic distribution with \( p(\alpha + 1 | \alpha) = 1 \).

This segmentation procedure is extremely noisy before an initial executor policy has been trained. Thus, during the first iteration of training, we estimate a segmentation by fitting a 3-state hidden Markov model to training action sequences using the Baum–Welch algorithm (Baum et al., 1970), and mark state transitions segment boundaries. Details about the initialization step and the algorithm can be found in Appendix B.

**Labeling:** \( \arg \max_{\hat{\tau}} \mathcal{L}(\hat{\tau}, \hat{\alpha}, \hat{\theta}) \)

Inference of latent, language-based plan descriptions in unannotated demonstrations involves an intractable search over string-valued \( \tau \). To approximate this search tractably, we used a learned, amortized inference procedure (Wainwright and Jordan, 2008; Hoffman et al., 2013; Kingma and Welling, 2014) to impute descriptions given fixed segmentations. During each parameter update step...
Algorithm 1: (SL)\textsuperscript{3}: Semi-Supervised Skill Learning with Latent Language

**Input:** Unannotated demonstrations
\[ D = \{(d_1, g_1), (d_2, g_2), \ldots \}; \]
Annotated demonstrations
\[ D^{ann} = \{(d_1, g_1, \tau_1), (d_2, g_2, \tau_2), \ldots \} \]

**Output:** Inferred alignments \( \hat{\alpha} \), labels \( \hat{\tau} \), and parameters \( \theta \) for \( \pi^C \) and \( \pi^E \).

// Initialization

Initialize policy parameters \( \theta \) using a pretrained language model (Raffel et al., 2020).

Initialize inference network parameters \( \eta \leftarrow \arg \max_\eta \sum_{d \in D^{ann}} \sum_{a} \log q_\eta(\tau | a, o_a) \).

for iteration \( t \leftarrow 1 \ldots T \) do

// Segmentation

if \( t = 1 \) then

 Initialize \( \hat{\alpha} \) using the Baum–Welch algorithm (Baum et al., 1970)

else

\( \hat{\alpha} \leftarrow \arg \max_\alpha L(\hat{\tau}, \hat{\alpha}, \hat{\theta}) + L^{ann}(\hat{\alpha}, \hat{\theta}) \) [Algorithm 2].

end

// Labeling

\( \hat{\tau} \leftarrow \arg \max_\tau L(\hat{\tau}, \hat{\alpha}, \hat{\theta}) \)

// Parameter Update

\( \hat{\theta} \leftarrow \arg \max_\theta L(\hat{\tau}, \hat{\alpha}, \hat{\theta}) + L^{ann}(\hat{\alpha}, \hat{\theta}) \)

\( \hat{\eta} \leftarrow \arg \max_\eta \sum_{d} \sum_{\tau} \log q_\eta(\hat{\tau} | a, o_a) \)

end

Algorithm 2: Dynamic program for segmentation

**Input:** Demonstration \( d = [(o_1, a_1), \ldots, (o_n, a_n)] \);
Task specifications \( \tau = [(\tau_1, \ldots, \tau_m)] \);
Executor \( \pi^E(a | o, \tau) = \prod_i \pi^E(.) (a_i | a_{i-1}, o_i, \tau) \)

**Output:** Maximum a posteriori alignments \( \alpha \).

\[ \text{scores} \in \text{an } n \times m \text{ matrix of zeros} \]
// scores[i, j] holds the log-probability of the // highest-scoring sequence whose final action i is // aligned to subtask j.

for \( i \leftarrow 1 \ldots n \) do

for \( j \leftarrow 1 \ldots |\tau| \) do

\[ \text{scores}[i, j] \leftarrow -\infty \]

for \( k \leftarrow 1 \ldots i-1 \) do

\[ \text{scores}[i, j] \leftarrow \max \{ \]

\[ \text{scores}[i, j], \]

\[ \text{scores}[k, j-1] \]

\[ + \log \pi^E(a_{k+1:i} | o_{k+1:i}, \tau_j) \} \]

end

end

end

The optimal alignment sequence may be obtained from scores via back-tracing (Rabiner, 1989).

(described below), we train an inference model \( q_\eta(\tau | a_{\hat{\theta}(i)}, a_{\hat{\theta}(i+1)}, g) \) to approximate the posterior distribution over descriptions for a given segment given a goal, the segment’s actions, and the actions from the subsequent segment.\(^4\) Then, during the labeling step, we label complete demonstrations by choosing the highest-scoring instruction for each trajectory independently:

\[
\arg \max_\tau \log p(d, g, \tau, \alpha) \approx \left[ \arg \max_\tau q(\tau | a_{\hat{\theta}(i)}, a_{\hat{\theta}(i+1)}, g) \bigg| s(i) \in \text{seg}(\alpha) \right]
\]

(6)

Labeling is performed only for demonstrations in \( D \), leaving the labels for \( D^{ann} \) fixed during training.

**Param update:** \( \arg \max_\theta L(\hat{\tau}, \hat{\alpha}, \hat{\theta}) + L^{ann}(\hat{\alpha}, \hat{\theta}) \)

This is the simplest of the three update steps: given fixed instructions and alignments, and \( \pi^E, \pi^C \) parameterized as neural networks, this objective is differentiable end-to-end. In each iteration, we train these to convergence (optimization details are described in Section 4 and ??). During the parameter update step, we also fit parameters \( \eta \) of the proposal model to maximize the likelihood \( \sum_d \sum_{\tau} \log q_\eta(\hat{\tau} | a, o_a) \) with respect to the current segmentations \( s \) and labels \( \hat{\tau} \).

As goals, subtask indicators, and actions may all be encoded as natural language strings, \( \pi^C \) and \( \pi^E \) may be implemented as conditional language models. As described below, we initialize both policies with models pretrained on a text corpora.

4In our experiments, conditioning on observations or longer context did not improve the accuracy of this model.

4 Experimental Setup

Our experiments aim to answer two questions. First, does the latent-language policy representation described in Section 3 improve downstream performance on complex tasks? Second, how many natural language annotations are needed to train an effective latent language policy given an initial dataset of unannotated demonstrations?

**Environment** We investigate these questions in the ALFRED environment of Shridhar et al. (2020). ALFRED consists of a set of interactive simulated households containing a total of 120 rooms, accompanied by a dataset of 8,055 expert task demonstrations for an embodied agent annotated with 25,743 English-language instructions. Observations \( o \) are
bitmap images from a forward-facing camera, and actions \( a \) are drawn from a set of 7 low-level navigation and manipulation primitives. Manipulation actions (7 of the 12) additionally require predicting a mask over the visual input to select an object for interaction. See Shridhar et al. (2020) for details.

While the ALFRED environment is typically used to evaluate instruction following models, which map from detailed, step-by-step natural language descriptions to action sequences (Shridhar et al., 2020; Singh et al., 2020; Corona et al., 2021), our experiments focus on an autonomous evaluation in which agents are given goals (but not fine-grained instructions) at test time. Several previous studies have also considered the autonomous evaluation for ALFRED, but all have used extremely fine-grained supervision at training time, including full supervision of symbolic plan representations and their alignments to demonstrations (Blukis et al., 2021). In contrast, our approach supports learning from partial, language-based annotations without segmentations or alignments, and this data condition is the main focus of our evaluation.

**Modeling details** \( \pi^C \) and \( \pi^E \) are implemented as sequence-to-sequence transformer networks (Vaswani et al., 2017). \( \pi^C \), which maps from text-based goal specifications to text-based instruction sequences, is initialized with a pre-trained T5-small language model (Raffel et al., 2020). \( \pi^E \), which maps from (textual) instructions and (image-based) observations to (textual) actions and (image-based) object selection masks is also initialized with T5-small; to incorporate visual input, this model first embeds observations using a pretrained ResNet18 model (He et al., 2016) and transforms these linearly to the same dimensionality as the word embedding layer. Details about the architecture of \( \pi^C \) and \( \pi^E \) may be found in Appendix C.

**Baselines and comparisons** We compare the performance of (SL)\(^3 \) to several baselines:

**seq2seq**: A standard (non-hierarchical) goal-conditioned policy, trained on the \((g, d)\) pairs in \( D \cup D^{\text{ann}} \) to maximize \( \sum_{a, o, g} \log \pi(a \mid o, g) \), with \( \pi \) parameterized similar to \( \pi^E \).

**seq2seq2seq**: A supervised hierarchical policy with the same architectures for \( \pi^C \) and \( \pi^E \) as in (SL)\(^3 \), but with \( \pi^C \) trained to generate subtask sequences by maximizing \( \sum_{g, \tau} \log \pi^C(\tau \mid g) \) and \( \pi^E \) trained to maximize \( \sum_{a, o, \tau, g} \log \pi^E(a \mid o, \tau, g) \) using only \( D^{\text{ann}} \). Because \( \pi^E \) maps from complete task sequences to complete low-level action sequences, training of this model involves no explicit alignment or segmentation steps.

**no-pretrain, no-latent**: Ablations of the full (SL)\(^3 \) model in which \( \pi^C \) and \( \pi^E \) are, respectively, randomly initialized or updated only on \( L^{\text{ann}}(\alpha, \hat{\theta}) \) during the parameter update phase.

In addition to these baselines, we contextualize our approach by comparing it to several state-of-the-art models for the instruction following task in ALFRED: S+ (Shridhar et al., 2020), MOCA (Singh et al., 2020), and Modular (Corona et al., 2021). Like seq2seq, these are neural sequence-to-sequence models trained to map instructions to actions; they incorporate several standard modeling improvements from the instruction following literature, including progress monitoring (Ma et al., 2019), pretrained object recognizers (Singh et al., 2020), and independently parameterized policies for different subtasks (Andreas et al., 2016). This last group of models is trained with considerably stronger supervision than (SL)\(^3 \): instructions and alignments during training, and ground truth instructions during evaluation.

**Evaluation** Following Shridhar et al. (2020), our main evaluation Table 1 computes the online, subtask-level accuracy of each policy. Given a ground-truth (goal, demonstration) pair from the test set with ground-truth subtask boundaries, we perform forced decoding (“teacher forcing”) of the policy up to the beginning of each segment boundary, then allow the policy to take actions autonomously for up to 20 timesteps. If it completes the subtask within this window, the subtask is marked as successful. Online evaluation requires live interaction with a simulator, and is somewhat slow; for data-efficiency experiments involving a large number of policy variants Fig. 4, we instead use an offline evaluation in which we measure the fraction of subtasks in which a policy’s predicted actions (ignoring object selection masks) exactly match the ground truth action sequence.

In ALFRED, navigation in the autonomous condition requires exploration, but no exploration is demonstrated, and techniques other than imitation learning are required for this specific skill. To reflect this, we replace all annotations containing detailed navigation instructions go to the glass on the table to your left with generic ones find a glass. Additionally, while we report results for the navigation subtask, we do not include them when reporting av-
Table 1: Online subtask success rate for (SL)³, baselines, and instruction following models, grouped by subtask category. “Train: g + τ | Test: g” means models in this section were trained with both goals g and annotated subtask descriptions τ, but observed only goals during evaluation. With annotations for 10% of data and no alignment supervision, (SL)³ outperforms non-hierarchical baselines and is competitive with models that receive substantially more information during training and evaluation. *Navigation (GoTo) is omitted from this average; see text for discussion.

<table>
<thead>
<tr>
<th>Model</th>
<th>Avg</th>
<th>Clean</th>
<th>Cool</th>
<th>Hear</th>
<th>Pick</th>
<th>Put</th>
<th>Slice</th>
<th>Toggle</th>
</tr>
</thead>
<tbody>
<tr>
<td>(SL)³ (10%)</td>
<td>56</td>
<td>56</td>
<td>75</td>
<td>74</td>
<td>50</td>
<td>48</td>
<td>54</td>
<td>32</td>
</tr>
<tr>
<td>(SL)³ (100%)</td>
<td>58</td>
<td>68</td>
<td>82</td>
<td>75</td>
<td>50</td>
<td>45</td>
<td>55</td>
<td>32</td>
</tr>
<tr>
<td>seq2seq</td>
<td>27</td>
<td>16</td>
<td>33</td>
<td>64</td>
<td>20</td>
<td>15</td>
<td>25</td>
<td>13</td>
</tr>
<tr>
<td>seq2seqseq</td>
<td>42</td>
<td>15</td>
<td>69</td>
<td>58</td>
<td>29</td>
<td>42</td>
<td>50</td>
<td>32</td>
</tr>
</tbody>
</table>

Train: g + τ | Test: g

S+ | 49 | 21 | 94 | 88 | 20 | 51 | 14 | 54 | 21 |
MOCA | 49 | 71 | 38 | 86 | 44 | 39 | 55 | 11 | 32 |
Mod | 63 | 67 | 94 | 85 | 28 | 55 | 39 | 73 | 14 |

Table 2: Ablation experiments. Providing ground-truth language plans can be inferred from unlabeled demonstrations using a small set of seed annotations. Table 2 provides details. Language model pretraining improves automatically decision-making. Ablation experiments in Table 2 provide details. Language model pretraining improves automatically decision-making.

<table>
<thead>
<tr>
<th>Model</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>(SL)³ (10%)</td>
<td>56</td>
</tr>
<tr>
<td>(SL)³ (100%)</td>
<td>58</td>
</tr>
<tr>
<td>(SL)³ (ground-truth α)</td>
<td>65</td>
</tr>
<tr>
<td>no-pt</td>
<td>49</td>
</tr>
<tr>
<td>no-latent</td>
<td>52</td>
</tr>
</tbody>
</table>

Figure 3: Example of an inferred segmentation and labeling for an unannotated trajectory. The trajectory is parsed into a sequence of 10 segments and q₀ assigns high scoring natural-language labels to the segmented actions. These are consistent with the objects, receptacles, and sub-tasks. The overall sequence of latent-language skills is a good plan for the high-level goal.

5 Results

Table 1 compares (SL)³ with flat and hierarchical imitation learning baselines. The table includes two versions of the model: a 100% model trained with full instruction supervision (|D| = 0, |Dₜₙₐₜ₉| = 21000) and a 10% model trained with only a small fraction of labeled demonstrations (|D| = 19000, |Dₜₙₐₜ₉| = 2000). Baselines are always trained with 100% of natural language annotations, meaning the 10% version of (SL)³ has access to strictly less information than all models it is compared to. Results are shown in Table 1. We find:

(SL)³ improves on flat policies: In both the 10% and 100% conditions, it improves over the subtask completion rate of the seq2seq (goals-to-actions) model by 29%. Indeed, it is competitive with several recent state-of-the-art instruction following models, outperforming S+ and Mod and approaching the performance of MOCA and performing particularly well on Picking subtasks. Conditioning on detailed instructions is not needed for good task performance in ALFRED—while decomposing goals hierarchically appears to be extremely helpful, hierarchical policies can be trained to generate high-quality plans directly from goals.

Language-based policies can be trained with sparse natural language annotations: Performance of (SL)³ trained with 10% and 100% natural language annotations is similar (and in both cases superior to seq2seq and seq2seqseq trained on 100% of data). Fig. 4 shows more detailed supervision curves. Ablation experiments in Table 2 show that inference of latent training plans is important for this result: with no inference of latent instructions (i.e. training only on annotated demonstrations), performance drops from 56% to 52%. Fig. 3 shows an example of the structure inferred for an unannotated trajectory: the model inserts reasonable segment boundaries and accurately labels each step. Ultimately, relatively little language is needed to train effective hierarchical models: language plans can be inferred from unlabeled demonstrations using a small set of seed annotations.

Language model pretraining improves automated decision-making. Ablation experiments in Table 2 provide details. Language model pretraining...
ing of $\pi^C$ and $\pi^E$ (on ungrounded text) is crucial for good performance in the low-data regime: with 10% of annotations, models trained from scratch complete 49% of tasks (vs 56% for pretrained models). We hypothesize that this result can be partly attributable to the fact that pretrained language models already encode a great deal of information about the common-sense structure of plans, e.g. the fact that *slicing a tomato* first requires *finding a knife*. Such models are thus well-positioned to adapt to “planning” problems that require modeling relations between natural language strings. These experiments point to a potentially broad role for pretrained language models in tasks that do not involve language as an input or an output.

### 6 Related Work

Our approach draws on a large body of research at the intersection of natural language processing, representation learning, and autonomous control.

**Language-based supervision and representation** The use of natural language annotations to scaffold learning, especially in computer vision and program synthesis applications, has been the subject of a number of previous studies (Branavan et al., 2009; Frome et al., 2013; Andreas et al., 2018; Wong et al., 2021). Here, we use language to support policy learning, specifically by using natural language instructions to discover compositional subtask abstractions that can support autonomous control. Our approach is closely related to previous work on learning sketch libraries from policy sketches (Andreas et al., 2017; Shiarlis et al., 2018); instead of the fixed skill inventory used by policy sketches, (SL)$^3$ learns an open-ended, compositional library of behaviors indexed by natural language strings.

**Hierarchical policies** Hierarchical policy learning and temporal abstraction have been major areas of focus since the earliest research on reinforcement learning and imitation learning (McGovern and Barto, 2001; Konidaris et al., 2012; Daniel et al., 2012). Past work typically relies on direct supervision or manual specification of the space of high-level skills (Sutton et al., 1999; Kulkarni et al., 2016) or fully unsupervised skill discovery (Dietterich, 1999; Bacon et al., 2017). Our approach uses policy architectures from this literature, but aims to provide a mechanism for supervision that allows fine-grained control over the space of learned skills (as in fully supervised approaches) while requiring only small amounts of easy-to-gather human supervision.

**Language and interaction** Outside of language-based supervision, problems at the intersection of language and control include *instruction following* (Chen and Mooney, 2011; Branavan et al., 2009; Tellex et al., 2011; Anderson et al., 2018; Misra et al., 2017), *embodied question answering* (Das et al., 2018; Gordon et al., 2018) and dialog tasks (Tellex et al., 2020). As in our work, representations of language learned from large text corpora facilitate grounded language learning (Shridhar et al., 2021), and interaction with the environment can in turn improve the accuracy of language generation (Zellers et al., 2021); future work might extend our framework for semi-supervised inference of plan descriptions to these settings as well.

### 7 Conclusion

We have presented (SL)$^3$, an algorithm for learning hierarchical policies from demonstrations sparsely annotated with natural language descriptions. Using these annotations, (SL)$^3$ infers the latent structure of unannotated demonstrations, automatically segmenting them into subtasks and labeling each subtask with a compositional description. In the learnt hierarchical policy, natural language serves as an abstract representation of subgoals and plans.

While our evaluation has focused on household robotics tasks, the hierarchical structure inferred by (SL)$^3$ is present in a variety of learning problems, including image understanding, program synthesis, and language generation. In all those domains, generalized versions of (SL)$^3$ might offer a framework for building high-quality models using only a small amount of rich natural language supervision.
References


P. Bacon, Jean Harb, and Doina Precup. 2017. The option-critic architecture. In AAAI.


Thomas G Dietterich. 1999. Hierarchical reinforcement learning with the MAXQ value function decomposition.


Hengyuan Hu, Denis Yarats, Quocheng Gong, Yuandong Tian, and M. Lewis. 2019. Hierarchical decision making by generating and following natural language instructions. In NeurIPS.


Yiding Jiang, S. Gu, K. Murphy, and Chelsea Finn. 2019. Language as an abstraction for hierarchical deep reinforcement learning. In NeurIPS.


Dipendra Kumar Misra, J. Langford, and Yoav Artzi. 2017. Mapping instructions and visual observations to actions with reinforcement learning. In EMNLP.


A Out-of-distribution Generalization

One of the advantages of language-based skill representations over categorical representations is open-endedness: \((\text{SL})^3\) does not require pre-specification of a fixed inventory of goals or actions. As a simple demonstration of this potential for extensibility, we design goal prompts consisting of novel object names, verbs and skill combinations not seen at training time, and test the model’s ability to generalize to out-of-distribution samples across the three categories. Some roll-outs can be seen in Fig. 6. We observe the following:

Novel sub-task combinations  We qualitatively evaluate the ability of the model to generalize systematically to novel subtask combinations and sub-task ordering not encountered at training time. Examples are shown in Fig. 6. For example, we present the model with the goal slice a heated apple; in the training corpus, objects are only heated after being sliced. It can be seen in Fig. 6 that the model able correctly orders the two subtasks. The model additionally generalizes to new combinations of tasks such as clean and cool an apple.

Novel objects and verbs  The trained model also exhibits some success at generalizing novel object categories such as carrot and mask. In the carrot example, an incorrect Find the lettuce example is generated at the first step, but subsequent subtasks refer to a carrot (and apply the correct actions to it). The model also generalizes to new but related verbs such as scrub but fails at ones like squash that are unrelated to training goals.

Limitations  One shortcoming of this approach is that affordances and constraints are incompletely modelled. Given a (physically unrealizable) goal clean the bowl and then slice it, the model cannot detect the impossible goal and instead generates a plan involving slicing the bowl. Another shortcoming of the model is the ability to generalize to goals that may involve considerably larger number of subgoals than goals seen at training time. For plans that involve very long sequences of skills (slice then clean then heat…) the generated plan skips some subtasks Fig. 6.

B Initialization: Segmentation Step

The training data contains no STOP actions, so \(\pi^E\) cannot be initialized by training on \(D^\text{ann}\). Using a randomly initialized \(\pi^E\) during the segmentation step results in extremely low-quality segmentations. Instead, we obtain an initial set of segmentations via unsupervised learning on low-level action sequences.

In particular, we obtain initial segmentations using the Baum–Welch algorithm for unsupervised estimation of hidden Markov models (Baum et al., 1970). We replace string-valued latent variables produced by \(\pi^C\) with a discrete set of hidden states (in our experiments, we found that three hidden states sufficed). Transition and emission distributions, along with maximum a posteriori sequence labels, are obtained by running the expectation-maximization algorithm on state sequences. We then insert segment boundaries (and an implicit STOP action) at every transition between two distinct hidden states. Evaluated against ground-truth segmentations from the ALFRED training set, this produces an action-level accuracy of \(87.9\%\). The detailed algorithm can be found in Baum et al. (1970).

C Model Architecture: Details

The controller policy \(\pi^C\) is a fine-tuned T5-small model. The executor policy \(\pi^E\) decodes the low-level sequence of actions conditioned on the first-person visual observations of the agent. We use the same architecture across the remaining baselines too. Fig. 5 depicts the architecture of the image-conditioned T5 model. In addition to task specifications, we convert low-level actions to templated commands: for example, put (cup, table) becomes put the cup on the table. These are parsed to select actions to send to the ALFRED simulator. During training, both models are optimized using the AdamW algorithm (Loshchilov and Hutter, 2019) with a learning rate of \(1e-4\), weight decay of 0.01, and \(\epsilon = 1e-8\). We use a MaskRCNN model to generate action masks, selecting the predicted mask labeled with the class of the object name generated by the action decoder. The same model architecture is used across all baselines.

D Role of trajectory length

We conclude with an additional set of ablation experiments aimed at clarifying what aspects of the demonstrated trajectories (\(\text{SL})^3\) is better able to model than baselines. We begin by observing that most actions in our data are associated with navigation, with sequences of object manipulation actions (like those depicted in Fig. 3) constitut-
Figure 5: Model architecture for $\pi^E$, seq2seq and seq2seq2seq: Language parametrized sub-task/goal is input to the encoder and actions templated in natural language are generated sequentially token-wise. The predictions made are conditioned on the visual field of view of the agent at every time step along with the token generated the previous time step. At the end of every low-level action (when ‘.’ is generated) the action is executed. For manipulation actions, the mask corresponding to the object predicted is selected from the predictions of a MaskRCNN model on the visual state. Navigation actions do not operate over objects. Once the action is taken, the environment returns the updated visual state and the policy continues to be unrolled until termination (STOP).

Figure 6: Generalization of $\pi^C$ in out-of-distribution (OOD) settings including novel a) sub-task orders b) objects c) verbs. OOD generalization enabled by means of representing plans in natural language overcomes the issue of having to pre-specify the inventory of objects and actions specific to environments. d) Failures: The model fails to predict actions based on the true affordances of objects and cannot generate arbitrarily long plans.
ing only about 20% of each trajectory. We construct an alternative version of the dataset in which all navigation subtasks are replaced with a single TeleportTo action. This modification reduces average trajectory length from 50 actions to 9. In this case, (SL)$^3$ and seq2seq2seq perform comparably well (55.6% success rate and 56.7% success rate respectively), and only slightly better than seq2seq (53.6% success rate). Thus, while (SL)$^3$ (and all baselines) perform quite poorly at navigation skills, identifying these skills and modeling their conditional independence from other trajectory components seems to be crucial for effective learning of other skills in the long-horizon setting. Hierarchical policies are still useful for modeling these shorter plans, but by a smaller margin than for long demonstrations.