Learning Navigation Subroutines by Watching Videos

Ashish Kumar\textsuperscript{1} Saurabh Gupta\textsuperscript{2,3} Jitendra Malik\textsuperscript{1,2}
\textsuperscript{1}UC Berkeley \textsuperscript{2}Facebook AI Research \textsuperscript{3}UIUC
ashish.kumar@berkeley.edu, saurabhg@illinois.edu, malik@eecs.berkeley.edu

Abstract

Hierarchies are an effective way to boost sample efficiency in reinforcement learning, and computational efficiency in classical planning. However, acquiring hierarchies via hand-design (as in classical planning) is suboptimal, while acquiring them via end-to-end reward based training (as in reinforcement learning) is unstable and still prohibitively expensive. In this paper, we pursue an alternate paradigm for acquiring such hierarchical abstractions (or visuo-motor subroutines), via use of passive first-person observation data. We use an inverse model trained on small amounts of interaction data to pseudo-label the passive first person videos with agent actions. Visuo-motor subroutines are acquired from these pseudo-labeled videos by learning a latent intent-conditioned policy that predicts the inferred pseudo-actions from the corresponding image observations. We demonstrate our proposed approach in context of navigation, and show that we can successfully learn consistent and diverse visuo-motor subroutines from passive first-person videos. We demonstrate the utility of our acquired visuo-motor subroutines by using them as is for exploration, and as sub-policies in a hierarchical RL framework for reaching point goals and semantic goals. We also demonstrate behavior of our subroutines in the real world, by deploying them on a real robotic platform. Project website with videos, code and data: https://ashishkumar1993.github.io/subroutines/.

1. Introduction

Every morning, when you decide to get a cup of coffee from the kitchen, you think of going down the hallway, turning left into the corridor and then entering the room on the right. Instead of deciding the exact muscle torques, you plan at this higher level of abstraction by composing these reusable lower level visuo-motor subroutines to reach your goal. In this paper, we focus on learning such visuo-motor subroutines for computational agents. Behavior of our learned subroutines as executed on a robotic platform is shown in Figure 1.

These visuo-motor subroutines are classically known as operators in STRIPS planning [14] or more recently as options in RL [30]. They enable hierarchical planning which mitigates the known issue of high computational cost in classical planning and high sample complexity in reinforcement learning. But how can AI-agents acquire these visuo-motor routines? One way is to manually design them as done in classical robotics and STRIPS. Alternately, we can learn these visuo-motor routines through interaction by training a hierarchical agent in an end-to-end manner typically done via reward based reinforcement learning. Since these methods jointly learn all levels of hierarchy, they are unstable to train and still require a large number of active interaction samples.

To address these challenges, we propose to exploit passive observations of humans performing these tasks. This is a major theme in learning based robotics, drawing inspiration from the fact that in human society we learn to perform many tasks not solely through our own experimentation but by observing parents or other experts. To imitate an expert at a visuomotor task, ideally we need to know both the perceptual input to the expert and the action taken. This creates an interesting dilemma depending on whether we capture first person (ego-centric) video of a human performing...
a task or a third-person video. Third-person videos have the benefit of having action information but don’t have the perceptual input. So if the routine is dependent on responding to specific visual inputs, we are at a loss. The opposite challenge occurs when we have first person video of a person performing a task (e.g. a head mounted GoPro camera); then we have the perceptual input the agent gets at every moment in time, but typically we don’t know which action was executed. In this paper, we will address this case. An overview of our approach is shown in Figure 2.

We start with first person navigation videos from agents $R_1...R_n$, without corresponding action labels. People constantly upload these kinds of videos online making it freely available. Given these videos, we want our robot $S$ to learn subroutines from these videos. This happens in two phases.

In the first phase, overviewed in Figure 3, we generate pseudo-action labels for these videos by running an inverse model on every consecutive pair of images. This inverse model is learned by the agent $S$ using self supervision on random exploration data. An interesting thing to note is that action space of $R_1...R_n$ might be different from the action space of our agent $S$. Hence, these pseudo-action labels are not the actual action taken, but an action imagined by the agent $S$ to make transition between the observations in the reference video in the agent $S$’s action space as closely as possible.

In the second phase, shown in Figure 4, we start with these pseudo-labeled videos and train a forward prediction model that takes the image as input and predicts the corresponding action taken in the reference video. However, this is a fundamentally ambiguous task, for example, an agent $R_k$ in the reference video that is facing a T-junction could have gone either left or right. To disambiguate this, we allow another network to look at the entire sequence of actions of $R_k$ in the reference video and encode the behavior as a one-hot latent intent vector that is additionally used to make the forward prediction. We additionally train an affordance model to predict which subroutines can be invoked for a given input image from our repertoire of learned subroutines. We do this by predicting the inferred one-hot encoding of the trajectory from the first image.

We show that subroutines and affordance models learned in this manner, are useful for downstream navigation tasks. They can be composed together in novel environments as is for effective exploration and can be used to initialize hierarchies RL policies for downstream navigation tasks such as point and semantic navigation. We have even transferred our learned subroutines on a real robot and have found them to be consistent in their behavior over random starting locations.

We survey related work in Section 2. We describe our approach for acquiring subroutines and affordance models in Section 3 and present a qualitative evaluation in Sec-
tion 4. We demonstrate and evaluate use of our learned subroutines for downstream tasks in Section 5.

2. Related Work

Our work on learning subroutines for navigation is related to efforts for learning policies for navigation, skill learning using imitation learning, hierarchical RL and affordance learning from videos. We survey these lines of work and relate them to VMSR.

Classical Navigation. Classical approaches to navigation employ geometric reasoning to solve the task. They typically operate on an occupancy map (built a priori or incrementally) [12, 32] and use path planning [8, 22, 25] to determine sequence of actions that conveys the robot to desired goal locations. While most works optimize in the base action space of the agent, few works employ hand-crafted motion primitives to speed up planning [18]. In contrast, we are studying how to learn such motion primitives (or visuo-motor subroutines) automatically. To enable learning of motion primitives Dynamic Motion Primitives (DMPs) propose a framework for specifying atomic actions [29]. They can potentially be learned given a demonstration [20]. However, which demonstrations constitute atomic units of action are manually decided. In contrast, these atomic behaviors emerge as a consequence of our algorithm.

Learned Navigation. Recent learning-based-efforts use reinforcement learning (or behavior cloning) to learn policies for solving specific navigation tasks [17, 27, 35]. While these works learn to leverage high-level semantics and exhibit behavior such as walking down hallways, exit rooms through doors, they still directly operate in the base action space of the robot. Learned skills are task and environment specific, and it takes a large number of interaction samples to even solve the same task in a new environment [27, 35]. To address this some works ignore task-specific environmental rewards and instead uses intrinsic rewards such as diversity [16] or prediction error [28]. However, neither of these approaches distill out composable skills from learned policies to solve novel tasks. Moreover, as all training signal is derived from interaction with the environment, skill acquisition is extremely expensive. In fact, our experiments show that our use of passive videos for learning skills is more sample efficient and results in better performance than such purely interactive approaches.

Learning from State-Action Trajectories. Several works in robotics use learning from demonstration [2, 7]. Here, a human operates the robot to demonstrate how to solve tasks, often in the desired test environment. Thus, learning has access to both the observations as well as the ground truth actions to solve the task. Works like [19] extend these formulations to work with trajectory collections that have multiple modes. However, this line of work relies on ground truth action labels from an expert. In contrast, we only assume observation data (without paired actions), and study performance at novel tasks in novel environments.

Learning from State Only Trajectories. Contemporary works [4, 11, 33] study the problem of learning from state only trajectories, similar to our work here. However, all of these works only study the scenario where the agent solves the task in exactly the same environment that they have state-only demonstrations for. In contrast, we do not assume access to environments for which we have videos for, making learning more challenging and rendering these past techniques ineffective. Additionally, our goal is to learn subroutines that work in previously unseen environments, which goes beyond the focus of these works.

Sub-policies and Options in Hierarchical RL. Hierarchical RL or the idea of decomposing a task into sub-tasks that can be accomplished using sub-policies, options or macros is an active area of research [5, 10, 30], with a number of recent papers (such as [26, 34]). These works acquire these sub-policies in a top-down manner while interacting with the environment to solve a reward based task. Our approach on the other hand investigates a bottom-up development of subroutines and can learn from relatively inexpensive unlabelled passive data. Our learned subroutines can be used to initialize any of these top-down HRL methods to accelerate learning.

Affordance Learning from Videos. Researchers have studied affordance learning from Internet videos. Fouhey et al. [15] leveraged YouTube videos to learn about affordances (pixels that afford walking, sitting, etc.), by observing where people walk, sit, etc. While this is a great first step, it does not exactly tell us about actions that a robot should execute. In contrast, we are learning subroutines directly in the action space of the specific robot at hand, allowing immediate deployment.

3. Affordances and Subroutines from Videos

Our agent $S$ is equipped with an onboard RGB camera, and an action space $A_{x,θ}$. These actions are a) stay in place, b) turn left by $θ$, c) turn right by $θ$, and d) move forward by distance $x$. The goal of our work is to learn visuo-motor subroutines for $S$ from first person navigation videos of agents $R_1, R_2, \ldots, R_n$.

To learn these subroutines, we assume that the robot has access to a set of training environments $E_{train}$, where it can execute actions and record image observations from its on-board cameras. We additionally assume access to a large dataset $D$ of first-person (or ego-centric) navigation videos.

Figure 2 shows an overview of our approach. We use $E_{train}$ to learn an inverse model (Section 3.1). This inverse model is then used to generate imagined actions corresponding to the video sequences in the dataset $D$. We call these action labels pseudo labels, the subsequent dataset obtained as $\tilde{D}$ (Section 3.2). This pseudo-labeled dataset is
then used to mine visuo-motor routines (Section 3.3) from \( \mathcal{D} \) which lacks any demarcation or clustering information on these routines. We additionally learn an affordance model (Section 3.4) to predict which subroutines can be invoked given an image. We call our proposed method of learning visuo-motor subroutines and affordance model VMSR. We next describe each of these components.

### 3.1. Inverse Model

We first build a one-step inverse model \( \psi \) for the agent, \textit{i.e.} given a pair of consecutive image observations \( o_t \) and \( o_{t+1} \), we train a model to predict the action \( a_t \) that was executed to get from \( o_t \) to \( o_{t+1} \):

\[
\hat{a}_t = \psi(o_t, o_{t+1}). \tag{1}
\]

The agent \( S \) collects data \{\ldots, \( o_t, a_t, o_{t+1}, a_{t+1} \ldots \} \) to train \( \psi \) by sampling \( a_t \) uniformly from \{\text{left, right, forward} \} and executing it in \( \mathcal{E}_{\text{train}} \) conveying it from image \( o_t \) to \( o_{t+1} \). We sample triples \( \{o_t, a_t, o_{t+1}\} \) to train \( \psi \).

Function \( \psi \) is implemented via a multi-layer perceptron on top of feature embeddings computed via convolutional neural networks on the images, \textit{i.e.}, \( \psi(o_t, o_{t+1}) = \text{MLP}(\phi(o_t), \phi(o_{t+1})) \). All parameters in \( \psi \) are optimized via back-propagation to minimize the cross-entropy loss between the true action \( a_t \) and the action \( \hat{a}_t \) predicted by \( \psi \).

### 3.2. Pseudo-Labeling Video Data

We then use this learned inverse model \( \psi \), to pseudo-label the dataset \( \mathcal{D} \) of passive first-person videos, which contains sequences of images \{\( o_1, o_2, \ldots, o_T \)\}. Given a pair of observations \( o_t \) and \( o_{t+1} \), we use the inverse model to estimate \( \hat{a} = \psi(o_t, o_{t+1}) \) which would take the agent from \( o_t \) to \( o_{t+1} \) as closely as possible. This generates a pseudo labeled dataset \( \hat{\mathcal{D}} \), that contains image action sequences, \{\( o_1, \hat{a}_1, o_2, \hat{a}_2, \ldots, o_T \)\}. Note that this pseudo labeling is done purely by looking at pairs of consecutive images.

### 3.3. Learning Visuo-Motor Subroutines

Next, we use the generated pseudo-labeled dataset \( \hat{\mathcal{D}} \) to learn subroutines. Our subroutines are realized via a recurrent neural network \( \pi \), which predicts \( \hat{a}_t \) as:

\[
\hat{a}_t, h_{t+1} = \pi(o_t, z, h_t) \tag{2}
\]

where state \( h_t \) and \( h_{t+1} \) are the current and updated hidden states respectively, \( o_t \) is the current observation, and \( z \) is a one hot vector which specifies the subroutine to invoke. \( \pi \) then executes the subroutine in a closed loop, executing the predicted action \( \hat{a}_t \) and receiving the next observation \( o_{t+1} \) as it proceeds.

\( \pi \) is learned on the pseudo-labeled dataset \( \hat{\mathcal{D}} \). While we do have supervision for what action to execute (pseudo-labels \( \hat{a}_t \)), we don’t have the underlying subroutine id \( z \) that a given trajectory corresponds to. We overcome this challenge, by adding another network \( f \) (1D CNN followed by a multi-layer perceptron) which looks at the entire future trajectory and predicts the subroutine-id \( z \), which in turn is consumed by policy \( \pi \) to predict actions. Both the networks are jointly trained to maximize the likelihood of the pseudo-labeled action sequence. Thus, given an image action sequence \{\( p_1, \hat{a}_1, p_2, \hat{a}_2, \ldots, p_T \)\}:

\[
e = f(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_T) \tag{3}
\]

\[
z \sim \text{softmax}(e) \tag{4}
\]

\[
\hat{a}_t, h_{t+1} = \pi(p_t, z, h_t) \quad \forall t \in \{1 \ldots T-1\} \tag{5}
\]

\( \pi \) and \( f \) are optimized together to maximize the likelihood of the pseudo-labeled action sequence. The subroutine id \( z \) is sampled from the action embedding \( e \) through a Gumbel-Softmax distribution \[21\]. This allows estimating gradients for parameters of \( f \) despite the sampling. Figure 4 shows the network diagram.

### 3.4. Learning Affordance Model

To enable the subroutines to be used efficiently in downstream tasks, we train another network \( \alpha \) to predict the sub-
routine id given the first image of the video sequence:

$$\tilde{z} = \alpha(o_1).$$ (6)

Intuitively, given an image of a hallway with two doors, one on the left and other on the right, $\alpha$ should learn to assign high probabilities to both go into left door and go into right door subroutines, whereas when there are no doors, it should simply peak on the subroutine go down a hallway. Also note that the $\alpha$ learns to predict the possible subroutines considering the next $T$ steps, where $T$ is the length of the subroutines. As $\alpha$ predicts which subroutines can be invoked in what situations, we call it an affordance model.

4. Learning Affordances and Subroutines

4.1. Experimental Setup

Our experiments involve use of environments (where the agent can actively interact with the environment) $\mathcal{E}_{\text{train}}$ and $\mathcal{E}_{\text{test}}$, and a dataset of first-person videos $\mathcal{D}$. We describe choices for the environment, agent and this video dataset:

**Environments** $\mathcal{E}_{\text{train}}$ and $\mathcal{E}_{\text{test}}$: We model these environments using a visually realistic simulator derived from scans of real world indoor environments from the Stanford Building Parser Dataset [3] (SBPD) and the Matterport 3D Dataset [9] (MP3D). These scans have been used to study navigation tasks in [17, 24, 31], and we adapt publicly available simulation code from Gupta et al. [17]. We split these environments into four sets: $\mathcal{E}_{\text{train}}$, $\mathcal{E}_{\text{video}}$, $\mathcal{E}_{\text{val}}$ and $\mathcal{E}_{\text{test}}$. $\mathcal{E}_{\text{train}}$ is used to train subroutines, $\mathcal{E}_{\text{val}}$ is used for development of policies for down-stream tasks, and $\mathcal{E}_{\text{test}}$ is used for evaluating performance of our policies on down-stream tasks. Use of $\mathcal{E}_{\text{video}}$ is described below. We make sure there is no overlap in any pair of these splits. Finally, we also do some tests in the real world on a real robot (iCreate2 platform equipped with a RGB camera).

**Agent Model:** Our agent is modeled as a cylinder that has 4 actions: a) stay in place, b,c) rotate left or right by $\theta$ (= 30$^\circ$), and d) move forward by $x$ (= 40cm). The robot is equipped with a RGB camera mounted at a height $h$ (= 120cm) from the ground and at an elevation $\phi$ (= $-5^\circ$) from the horizontal.

**Dataset $\mathcal{D}$:** We consider MP3D Walks dataset for first-person videos. **MP3D Walks Dataset** is auto-generated using the $\mathcal{E}_{\text{video}}$ environments, by rendering out images along the shortest path between pairs of random points. As we don’t assume access to underlying actions, we throw them out and only keep the sequence of images. We additionally ensure that the videos also come from agents with different parameters ($\theta, x, h, \phi$) than our agent (which operates in $\mathcal{E}_{\text{train}}$ (for inverse model training) and $\mathcal{E}_{\text{test}}$ (for down-stream tasks)), see Table A1 for specifics. As we can control the diversity and scale of this dataset, we can use it for studying performance trade-offs of the design choices of VMSR. **MP3D Walks Dataset** consists of around 217K clips of 40 steps each.

4.2. Training Details

We provide details of various stages of subroutine learning, and present visualizations for the test time behavior of the learned subroutines.

**Inverse Model Training.** The agent is initialized at 1.5K different locations spread over the 4 environments in $\mathcal{E}_{\text{train}}$. It is allowed to execute actions randomly for 30 steps. The collected data (45K interaction samples) is used to train the inverse model. We use cross-entropy loss between the actual action and the predicted action. We use Adam [23] with 64 batch size and a 0.001 learning rate.

**Video Pseudo-Labeling.** We then pseudo-label videos in $\mathcal{D}$ using the learned inverse model to obtain dataset $\hat{\mathcal{D}}$ as described in Section 3.2. This pseudo-labeled dataset $\hat{\mathcal{D}}$ is

![Subroutine Consistency over Different Starting Locations](image-url)

Figure 5: Subroutine Consistency over Different Starting Locations: In this figure we show roll-outs for subroutine different locations in the environment. Rows show the same subroutine, while columns show different subroutines unrolled from the same location. Subroutines show consistent behavior, that is different between different subroutines. SubR1 always wants to turn right, SubR2 turns left, and SubR3 & SubR0 have a higher preference for going straight (with occasional left/right turning). Visualizations show top view, however policies only use first person views.
Figure 6: Affordance Model Visualization: Top row shows images that the affordance model assigns high probability for being feasible for subroutine 1 (that goes rightwards), while images in the bottom row show images that are assigned high probability for subroutine 2 (that goes leftward). These high scoring images, indeed afford the predicted subroutines.

Figure 7: Multi-modality in Affordance Predictions: We visualize the entropy of the distribution output by the affordance model as we walk the agent along different trajectories in the test environment. A larger circle denotes a higher entropy, that is, more subroutines can be invoked at that location. We observe that the affordance model predicts more subroutines to be possible as the agent approaches hallway intersections, or room entrances. This multi-modality collapses as the agent crosses the decision junction and proceeds further.

used to learn subroutines $\pi(., z)$ as described in Section 3.3 and the affordance model as described in Section 3.4.

**Subroutine Training:** We slice each of the 217K videos into clips of length 10 steps with a sliding window of 5. This gives us a total of 2.2M clips to train our subroutines. We experiment with 4 subroutines (i.e. the $z$ vector is 4-dimensional). These subroutines are sampled from the predicted action clusters using Gumbel-Softmax distribution [21]. We use the straight-through estimator [6] to estimate gradients across this non-differentiability. This model is trained by minimizing the cross-entropy loss between the actions output by the policy ($\tilde{a}$) and the pseudo-labels ($\hat{a}$) obtained from the inverse model.

**Affordance Training:** To generate the data for the affordance model, we use the inferred subroutine id $z$ for a given action sequence of length 10 as the affordance label for the first image in the sequence and train to minimize cross-entropy loss over the inferred $z$ label.

4.3. Results

**Behavior of Subroutines:** We unroll different subroutines from different locations in the test environment $E_{test}$, and visualize the trajectories followed by each of them in the top view. Trajectories shown in each individual row of Figure 5 demonstrate that a specific subroutine does the same thing even when initialized at different locations. SubR1 always turns right, SubR2 always turns left. Rollouts shown in different columns of Figure 5 show that different subroutines show diverse behaviors when started from the same location. This shows the consistency of each of our subroutines and the diversity across subroutines.

**Affordance Prediction:** To understand the affordance prediction model, we include images from $E_{test}$ where the affordance prediction model makes high prediction in Figure 5. Top row shows images that cause a high prediction for SubR1, while bottom row shows images that excite SubR2. Figure 7 visualizes the entropy in the prediction of the affordance model as a function of the location in the environment. Locations near locations where multiple subroutines are possible have higher entropy.

**Real Robot Deployment:** We deployed our subroutines (learned in simulation) in the real world on a real robot (iCreate2 platform equipped with a RGB camera) and show robustness and diversity in Figure 8 and diversity in Figure 9.

5. Using Affordances and Subroutines

Learned subroutines and affordance models can be transferred to downstream navigation tasks. Our subroutines and the affordance model can be used as is in conjunction with each other to tackle tasks like exploration of novel envi-
environments. We can simply compose our subroutines via affordance model to generate exploration behavior. Furthermore, this decomposition into subroutines and affordance models, very naturally fits into hierarchical reinforcement learning frameworks [5]. Our subroutines are analogous to sub-policies, while the affordance model is analogous to the meta-controller. We study three downstream tasks with our learned subroutines ($\pi$) and the affordance model ($\alpha$):

**Exploration Task:** Subroutines can be sequentially composed together using the affordance model. This simple strategy of using these models as is leads to effective exploration of the environment. Comparisons against hand-crafted baselines, and other learning-based baselines are favorable to our approach on a number of metrics like sample complexity of active interactions, coverage, distance from initial location and collision rates.

**PointGoal and AreaGoal Tasks:** Our learned navigation subroutines can be used to initialize hierarchical RL policies for studying PointGoal (go to a specified point) and AreaGoal (go to the washroom) tasks with sparse and dense rewards. Initialization with our learned subroutines makes training significantly 4× more sample efficient than alternate initialization schemes for hierarchical RL.

### 5.1. Exploration via Subroutines and Affordances

The exploration task requires the agent to explore a novel environment efficiently, *i.e.*, given a starting location how well can an agent visit different parts of the environment. This can be done in a zero-shot manner by simply compos-

---

**Figure 8:** Robustness and Diversity of Subroutines on Real Robot: Learned Subroutines when deployed on a real robot demonstrate a robust and consistent behavior over perturbations to starting location as shown in top row for SubR2 and bottom row for SubR1.

**Figure 9:** Consistency of Subroutines on Real Robot: Learned Subroutines when deployed on a real robot demonstrate consistent behavior over different starting locations, as demonstrated for SubR2.

**Figure 10:** Coverage Visualization: We show coverage of the overall space after sampling 20 roll-outs from 11 different locations in the test environment $E_{\text{test}}$. Note that VMSR covers more of the environment. It is able to come out of rooms and different roll-outs go towards different areas. Curiosity, diversity and Random policies spend most of their time inside rooms. Policies that are biased to move forward do come out, but do not show diverse behavior. Visualizations show top view, however policies only use first person views.

### Table 1: Exploration Metrics

| Method                        | # Samples | ADT (m) | Maximum Distance (m) | Collision Rate (%) |
|-------------------------------|-----------|---------|----------------------|--------------------|
| Random                        | 0         | 18.09   | 7.5                  | 65                 |
| Forward Bias Policy           | 0         | 15.25   | 13.11                | 82                 |
| Always Forward, Rotate on Collision | 0      | 14.89   | 13.31                | 72                 |
| Skills from Diversity [13]    | 10M       | 17.63   | 7.85                 | 67                 |
| Skills from Curiosity [28]    | 10M       | 17.68   | 7.87                 | 64                 |
| VMSR (Exploration via Subroutines) | 45K     | 7.73    | 27.78                | 12                 |

# Samples: Number of environment interactions used for training. Lower is better.

ADT or Average Distance to Trajectory: Given executed trajectories from a given starting location, we compute the mean geodesic distance of points in the environment to the closest point on the trajectory. If we wanted to visit a point in the environment, this metric measures how much we will need to go off the trajectory to get to this point, in expectation. We report the average over all starting locations. Lower ADT is better.

Maximum Distance From Start: We measure how far the executed trajectories convey the agent. For each trajectory, we measure the maximum of the geodesic distance from the starting location to all points on the trajectory. We report the average maximum geodesic distance over trials. Higher Maximum Distance is better.

Collision Rate: We measure how often predicted forward actions lead to collisions with obstacles in the environment. Good methods should lead to fewer collisions. Lower collision rate is better.
ing the learned subroutines via the affordance model.

**Exploration via Subroutines**: Given the visual observation from the current location, we repeat the following two steps: a) we use the affordance model \( \alpha \) to sample the subroutine \( z \) to execute, b) we execute the sampled subroutine for 10 steps. If we have learned meaningful subroutines, they will exhibit coherent navigation behavior to cover large distances in the environment, rather than hitting against obstacles or random walking around the starting location. Consistency and diversity of subroutines will enable an effective affordance model to appropriately compose subroutines for diverse exploration behaviour.

**Task Setup**: We randomly initialize the agent at 100 different locations in the novel test environment \( \mathcal{E}_{\text{test}} \). For each location, we do 5 random executions (from a randomly chosen initial orientation) of length 408 each (square root of navigable area in the floor).

**Metrics**: We measure different aspects of the exploratory navigation behavior via the following metrics: a) Number of environment interactions used for training, b) Average distance of points in the environment to trajectory (ADT), c) Maximum distance traveled by policies in a single run, and d) Collision rate among forward actions. More details about the metrics can be found in Table 1. These metrics serve to measure different aspects of skillful navigation. We emphasize that VMSR is not trained to optimize for any of these metrics. Yet, as we describe next, they perform well on these metrics and outperform a number of baselines.

**Baselines.** We compare with the three hand-crafted baselines: a) Random policy (randomly execute one of the 4 actions), b) Forward bias policy (biased to more frequently execute forward action), and c) Always forward but rotate on collision policy. We also compare to the state-of-the-art unsupervised RL-based skill learning method d) DIAYN [13], and e) Curiosity [28]. These learning based techniques were trained with comparable networks (ResNet 18 models pre-trained on ImageNet) for over 10M samples. More details about these baselines are in Section A1.

**Results.** Table 1 shows that VMSR compares favorably to all baselines on all metrics outperforming on all three metrics. Figure 10 overlay trajectories executed by different policies onto the map (only used for visualization). Indeed, our trajectories cover the map much better than other methods.

It is particularly striking that all of this was just learned from a total of 45K interactions with the environment. Successful learning from first-person videos allowed the agent to execute coherent trajectories, even though it had only ever executed random actions. It also successfully learned the bias towards forward actions in navigation and the notion of obstacle avoidance leading to a high maximum distance, and a low collision rate. It outperforms hand-crafted baselines that were designed using these insights in mind.

Furthermore, it outperforms state-of-the-art learning based techniques for learning skills [13, 28], that were trained on multiple orders of magnitude more interaction sampled (45K vs. 10 million). These past works have only been shown to perform well for low-dimensional state-spaces and simple game environments. It is not surprising that they completely breakdown when used with high-dimensional inputs such as realistic real world images. Further explanation of their performance is in Section A1.
5.2. PointGoal and AreaGoal via HRL

We next investigate how we can use VMSR to solve goal-driven tasks. We do this by setting up hierarchical RL policies based on our learned subroutines and affordance models. Our experiments show that use of VMSR in hierarchical RL in this manner leads to significant improvements in sample complexity for training HRL agents for goal-driven tasks of PointGoal and AreaGoal, as compared to not using a hierarchy, using a hierarchy derived using state-of-the-art skill learning method [13, 28], and alternate ways of initializing the hierarchy.

Task Setup: We setup two goal driven navigation tasks, PointGoal and AreaGoal as defined in [1]. For PointGoal task, the agent is required to reach a given goal location (specified as a relative offset from robot’s current location). For the AreaGoal task, the agent is required to go to the washroom. We study both tasks in sparse and dense reward settings. RL and HRL policies are developed on the validation environment $E_{\text{val}}$, and finally trained on the test environment $E_{\text{test}}$ set with 3 random seeds to assess sample efficiency for learning.

Metrics: We use train time sample complexity to compare different methods.

Comparisons: VMSR can serve as initializations for hierarchical policies. We compare with the following alternates for initializing the meta-controller and the sub-policies: a) Random Initialization, b) ImageNet Initialization, and c) Initialization from skills via DIAYN [13] pre-training. (c) doesn’t provide an affordance model, so we initialize the meta-controller image CNN with the sub-policy CNN.

We can also compare VMSR initialization to initialization obtained from Curiosity [28]. Pathak et al. [28] use a monolithic policy (i.e., without any handle to control what they do), and study a AreaGoal task. Thus, for a fair comparison, we limit the comparison to the AreaGoal task and use a monolithic RL policy instead of the hierarchical policy. We report three training plots: a) Random Initialization, b) Initialization from Curiosity [28], and c) VMSR, where we obtain a monolithic policy using a version of VMSR with a single subroutine.

Training rewards are plotted in Figure 11 and Figure 12. Figure 11 shows the comparison among hierarchical policies. We observe upto $4 \times$ faster training when initialized with VMSR and affordance models when compared to the next best baseline which is ImageNet initialization. Improvements are generally larger for the harder case of sparser rewards. DIAYN [13] based initialization entirely fails, as it collapses to a trivial policy (more details in Section A1). Even among non hierarchical policies (Figure 12), initializing with VMSR (VMSR (1 SubR)) performs best (AreaGoal Tasks), outperforming random initialization and initialization from curiosity policy [28].

Acknowledgements: Authors would like to thank Allan Jabri, Shiry Ginosar and Devendra Singh Chaplot for feedback on the manuscript.

References

[1] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, and A. Zamir. On evaluation of embodied navigation agents. arXiv preprint arXiv:1807.06757, 2018.
[2] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. Robotics and Autonomous systems, 2009.
[3] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3D semantic parsing of large-scale indoor spaces. In CVPR, 2016.
[4] Y. Aytar, T. Pfaff, D. Budden, T. L. Paine, Z. Wang, and N. de Freitas. Playing hard exploration games by watching youtube. arXiv preprint arXiv:1805.11592, 2018.
[5] A. G. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. Discrete event dynamic systems, 2003.
[6] Y. Bengio, N. Léonard, and A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432, 2013.
[7] A. Billard, S. Calinon, R. Dillmann, and S. Schaal. Robot programming by demonstration. In Springer Handbook of robotics. 2008.
[8] J. Canny. The complexity of robot motion planning. MIT press, 1988.
[9] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In 3DV, 2017.
[10] P. Dayan and G. E. Hinton. Feudal reinforcement learning. In NIPS, 1993.
[11] A. D. Edwards, H. Sahni, Y. Schroeker, and C. L. Isbell. Imitating latent policies from observation. arXiv preprint arXiv:1805.07914, 2018.
[12] A. Elles. Using occupancy grids for mobile robot perception and navigation. Computer, 1989.
[13] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is all you need: Learning skills without a reward function. In ICLR, 2019.
[14] R. E. Fikes and N. J. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. Artificial intelligence, 1971.
[15] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single view geometry. IJCV, 2014.
[16] J. Fu, J. Co-Reyes, and S. Levine. EX2: Exploration with exemplar models for deep reinforcement learning. In NIPS, 2017.
[17] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik. Cognitive mapping and planning for visual navigation. In CVPR, 2017.
[18] K. Hauser, T. Bretl, K. Harada, and J.-C. Latombe. Using motion primitives in probabilistic sample-based planning for humanoid robots. In Algorithmic Foundation of Robotics. 2008. 3
[19] K. Hauser, Y. Chebotar, S. Schaal, G. Sukhatme, and J. J. Lim. Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets. In NIPS, 2017. 3
[20] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal. Dynamical movement primitives: learning attractor models for motor behaviors. Neural computation, 25(2):328–373, 2013. 3
[21] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144, 2016. 4, 6
[22] L. E. Kavraki, P. Svestka, J.-C. Latombe, and M. H. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. RA, 1996. 3
[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 5
[24] A. Kumar*, S. Gupta*, D. Fouhey, S. Levine, and J. Malik. Visual memory for robust path following. In Advances in Neural Information Processing Systems, 2018. 5
[25] S. M. LaValle. Planning Algorithms. Cambridge University Press, Cambridge, U.K., 2006. Available at http://planning.cs.uiuc.edu/. 3
[26] A. Levy, R. Platt, and K. Saenko. Hierarchical actor-critic. arXiv preprint arXiv:1712.00948, 2017. 3
[27] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu, et al. Learning to navigate in complex environments. In ICLR, 2017. 3
[28] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In ICML, 2017. 3, 7, 8, 9, 11
[29] S. Schaal. Dynamic movement primitives-a framework for motor control in humans and humanoid robotics. In Adaptive motion of animals and machines, pages 261–280. Springer, 2006. 3
[30] R. S. Sutton, D. Precup, and S. Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. Artificial intelligence, 1999. 1, 3
[31] T. Swedish and R. Raskar. Deep visual teach and repeat on path networks. In CVPR, 2018. 5
[32] S. Thrun, W. Burgard, and D. Fox. Probabilistic robotics. MIT press, 2005. 3
[33] F. Torabi, G. Warnell, and P. Stone. Behavioral cloning from observation. In IJCAI, 2018. 3
[34] A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. arXiv preprint arXiv:1703.01161, 2017. 3
[35] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In ICRA, 2017. 3
A1. Baselines for Exploration

1. **Random Policy**: We randomly sample an action from the four possible actions (stay, left, right, forward) at every step.

2. **Forward Bias Policy**: Since motion is typically dominated by forward motion, we compare to another policy that samples the forward action more preferably. We use the distribution of actions in the MP3D Walks Dataset, probabilities for stop, turn left, turn right and forward were [0.0, 0.17, 0.17, 0.60] respectively.

3. **Always Forward, Rotate on Collision**: This baseline repeats the following procedure: rotate by a random angle sampled from \((-\pi, \pi]\), move straight till collision.

4. **Diversity Policy (DIAYN) [13]**: We use the state-of-the-art RL-based unsupervised skill learning algorithm from Eysenbach et al. [13] to learn 4 diverse skills on \(E_{\text{train}}\) environments. We test the learned skills for exploration by randomly sampling a skill, and then executing it for 10 steps, where we sample actions from the probabilities output by the selected skill. Policy architecture is same as those for our subroutines, discriminator is based of a ResNet 18 model. Both models are initialized from ImageNet. Policy is trained for over 10 million interaction, best performance occurs at around 1M interaction samples.

5. **Curiosity Policy [28]**: We train a curiosity-based agent that seeks regions of space where its forward model has high prediction error [28]. Policy architecture is same as that for our subroutines, except that it does not take in the latent vector \(z\), and initialized from ImageNet. Forward model is learned in the \(\text{conv}_5\) average pooled feature space of a fixed Resnet 18 model pre-trained on ImageNet. Trajectories are executed by sampling from the action probabilities output by the policy. Once again, policy is trained for over 10 million interaction, best performance occurs at around 1M interaction samples.

**Curiosity Model**: Pathak et al. [28] proposed use of prediction error of a forward model as an intrinsic reward for learning skills using RL. We were surprised at the rather poor performance for the curiosity model. We found that the model converges to the policy of simply rotating in-place. Such a degenerate solution makes sense as rotating in-place has higher prediction error than staying in-place and moving forward. In-place rotations cause new parts of the environment to become visible which makes for a harder prediction task. Staying in-place and moving forward cause only minor changes to the image or no changes at all. Thus, the curiosity model rightly learns to simply rotate in-place. We saw this same behavior across different runs with different hyper-parameters and different architectures: policies will collapse to outputting just the rotation actions. Entropy based regularization is used to prevent such a collapse. We used such regularization and cross-validated various choices for the trade-offs in loss between entropy regularization and policy gradient loss, but didn’t find it to alleviate this issue. We selected the best model for the task of exploration across different runs and different number of training iterations. This selected model ended up being a heavily regularized model that would pick actions almost uniformly at random, as that would get higher performance than simply rotating in-place. As both extremes (taking actions randomly, or picking only the rotate in-place action) are trivial solutions, the curiosity model starts to ignore the image and consequently performs on-par with uninitialized models for reinforcement learning tasks.

**Diversity Model**: The diversity model from Eysenbach et al. [13] seeks to classify states with the skill id that was used to get to it (see Algorithm 1 in [13]). While this works well for the environments studied in [13], it breaks down for visual navigation. This is because, the same state can be reached via different skills depending on the starting state. This causes the skill classifiers \(q\) to only perform at chance. Consequently, the reward for the skill policies is uniform, causing the policies to collapse (all actions produce the same reward, and hence no learning happens). We observed this empirically in our experiments as well: accuracy for state classification was at chance (25% for four skills), and the reward stayed constant. Best performing policy (based on validation for exploration metrics) always predicted the following probabilities for different actions for different skills: \([0.246, 0.232, 0.237, 0.285]\) (for stop, left, right, forward respectively). As this can be done without looking at the image, the policy learns to ignores the image. Thus, the model perform on-par with uninitialized models for hierarchical reinforcement learning experiments.

A2. RL Experimental Setup

We use \(E_{\text{test}}\) for RL experiments. We use A2C to train all our algorithms on Point Goal task and Area Goal task.

- **Area Goal**: The task is to find the nearest washroom. \(E_{\text{test}}\) contains 2 washroom, and we start the agent 10-23 steps away from the nearest washroom. We randomly start the agent at a different location for every episode.

- **Point Goal**: We specify the goal coordinates relative to the start position, and randomly sample the start and the goal locations every episode. The goal is 10-17 steps away from the start location.
Figure A1: Dependence on active environment interaction samples, length of reference videos and number of subroutines specified: Column 1 and 2: We plot the exploration metrics against the number of self supervision interaction samples. There are two orthogonal ways of achieving this – increasing the number of restarts while keeping each episode length fixed (Col 1) and increasing the length of each self supervision episode while keeping the number of restarts fixed (Col 2). We see that visual diversity improves performance on Max Dist metric, but saturates at 45K interaction samples (1500 restarts with 30 steps each). Performance roughly remains the same as we increase the episode length. Column 3: We change the number of subroutines learned on the x-axis and compare the use of affordance model for sampling subroutines to randomly sampling subroutines. Affordance model shows improvement in collision rate over random sampling, indicating that the affordance model better respects the constraints of the physical space. We don’t see an improvement in the exploration metric or max distance metric. Column 4: We observe improvements as we increase the path length of the reference trajectories. Longer trajectories presumably allow VMSR to learn more complex subroutines.

Table A1: Split of environments between different sets used in the paper. These environments are from Stanford Building Parser Dataset (SBPD) [3] and Matterport 3D Dataset (MP3D) [9]. We fix a step size ($x$) and rotation angle ($\theta$) for each area by randomly sampling from the list. For elevation angle and height of the robot, we resample a value from the given ranges for every video.