Learning and Retrieval from Prior Data for Skill-based Imitation Learning

Soroush Nasiriany[1], **Tian Gao**[1,2], **Ajay Mandlekar**[3], **Yuke Zhu**[1]

[1]The University of Texas at Austin, [2]IIIS, Tsinghua, [3]NVIDIA Research

Abstract: Imitation learning offers a promising path for robots to learn general-purpose behaviors, but traditionally has exhibited limited scalability due to high data supervision requirements and brittle generalization. Inspired by recent advances in multi-task imitation learning, we investigate the use of prior data from previous tasks to facilitate learning novel tasks in a robust, data-efficient manner. To make effective use of the prior data, the robot must internalize knowledge from past experiences and contextualize this knowledge in novel tasks. To that end, we develop a skill-based imitation learning framework that extracts temporally extended sensorimotor skills from prior data and subsequently learns a policy for the target task that invokes these learned skills. We identify several key design choices that significantly improve performance on novel tasks, namely representation learning objectives to enable more predictable skill representations and a retrieval-based data augmentation mechanism to increase the scope of supervision for policy training. On a collection of simulated and real-world manipulation domains, we demonstrate that our method significantly outperforms existing imitation learning and offline reinforcement learning approaches. Videos and code are available at https://ut-austin-rpl.github.io/sailor

Keywords: Imitation Learning, Skill Learning, Robot Manipulation

1 Introduction

A long-standing dream in robotics is to enable robots to perform general-purpose tasks in diverse environments. In recent years, imitation learning has led to great progress towards this goal. Recent work has demonstrated that well-engineered behavioral cloning methods can achieve competitive performance on diverse manipulation tasks [1, 2, 3]. Despite this promise, these imitation learning algorithms have been confined to relatively small-scale domains. For real-world problems, current imitation learning algorithms often demand a large number of task demonstrations, which can be difficult and costly to obtain.

Realizing these limitations, a series of recent work seeks to use prior interaction data to improve the sample efficiency of imitation learning on new tasks. Such prior data come in various forms, including task-agnostic exploratory "play" data [4] or demonstrations previously collected for different tasks [5]. An open question is how best to extract knowledge from these large prior datasets and use this knowledge to facilitate learning novel tasks. One promising approach is skill-based imitation learning [6, 7], which aims to learn a latent space of short-horizon sub-trajectories from the prior data (called *skill learning*) and subsequently learn a policy to invoke the skills to solve a specific downstream task (called *policy learning*). This approach offers several appealing advantages: First, the policy benefits from the temporal abstraction encapsulated by the skills, allowing the policy to focus on higher-level reasoning about *what* behavior to perform rather than *how* to execute that behavior; and second, by reasoning about the target task in relation to skills trained on prior data, the robot implicitly distills knowledge from the rich diverse interactions of the prior data into the policy. Even so, existing skill-based imitation learning approaches bring modest improvements over simple behavioral cloning methods. This work aims at identifying the underlying limitations of existing approaches and designing new methods to address these limitations.

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

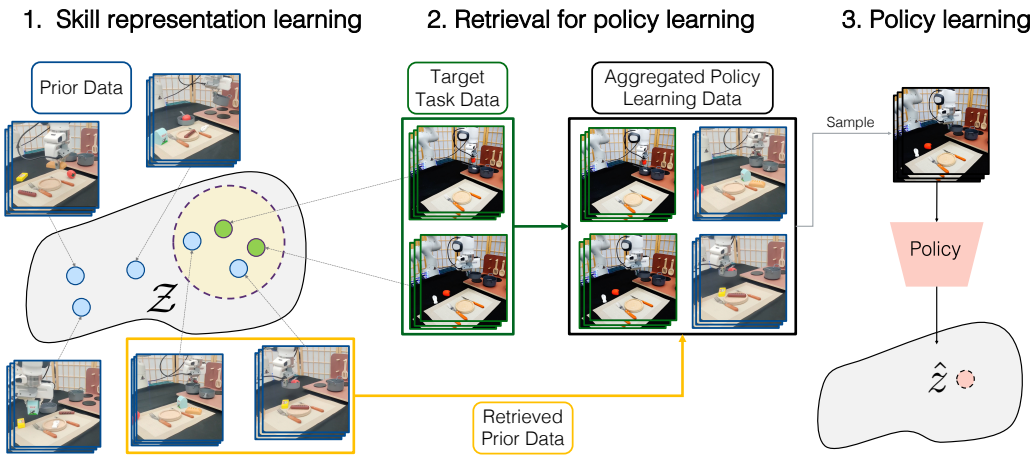

Figure 1: **Overview.** We present a skill-based imitation learning framework that uses prior data to effectively learn novel tasks. First, we learn a latent skill model on the prior data, with objectives to ensure a predictable skill representation. Given target task demonstrations, we use this latent space to retrieve similar behaviors from the prior data, expanding supervision for the policy. We then train a policy which outputs latent skills.

We argue that a practical approach should encompass two key properties. First, the latent skill space should serve as a *predictable* representation for downstream policy learning, allowing the policy to accurately infer which appropriate skill to use in new situations. While prior methods [8, 9, 6, 7] have widely adopted variational autoencoders to transform sub-trajectories into embeddings, we empirically show that this learning objective alone is insufficient for constructing a well structured latent space. As a result, the policy tends to execute irrelevant skills. Second, the robot should take advantage of the prior data to learn both the skills and the policy. A major limitation of existing skill-based imitation learning approaches [6, 7] is that they primarily focus on using prior data for skill learning but not policy learning. In these approaches, policies learned on a small number of target task demonstrations are prone to severe overfitting and covariate shift.

To address these limitations, we introduce **S**kill-**A**ugmented **I**mitation **L**earning with pri**o**r **R**etrieval (SAILOR). To improve the predictability of our skill representations, SAILOR uses an auxiliary temporal predictability objective to estimate the temporal distance between sub-trajectories in the demonstration sequences. To use prior data for policy learning, we introduce a novel retrieval-based data augmentation procedure that selectively retrieves data relevant to the target task. Specifically, we consider a subset of sub-trajectories from the prior dataset that have high latent skill similarity with sub-trajectories in the target task demonstrations (see Fig. 1). We evaluate on a wide variety of manipulation tasks in simulation and the real world, and show that SAILOR significantly outperforms state-of-the-art imitation learning and offline reinforcement learning approaches. Through comprehensive analysis, we highlight the roles that prior data, our representation learning objective, and retrieval-based data augmentation have in data-efficient learning of robust manipulation policies.

2 Related Work

Learning from Prior Data. There is a large body of work on learning manipulation tasks using human demonstrations [1, 10, 11, 12, 13, 14]. While promising, most of these works learn tasks independently, without re-using knowledge from prior tasks. As a result, they have high data requirements and exhibit brittle generalization in complex long-horizon tasks. To address these limitations, several lines of work have investigated leveraging large offline prior datasets to facilitate learning downstream robotic tasks. These prior datasets include task-agnostic play data [4, 15], demonstrations for related tasks [5], self-supervised agent-generated data [16], or a combination of these [17]. Alternatively to these, large video datasets are an appealing choice [18, 19, 20]. Recent work has leveraged these datasets to learn pre-trained visual representations for downstream control tasks [21, 22]. Yet despite their appeal, after experimenting with one such approach, R3M [21], we found that it can sometimes hinder downstream performance. One hypothesis is that the prior data on which these methods are trained on exhibit significant domain shift compared to downstream tasks, limiting transfer. In this work, we instead consider large multi-task robotic datasets.

Multi-Task Imitation Learning. Multi-task imitation learning methods offer a promising way to learn from diverse robotic data. These include task-conditioned [5], language-conditioned [23, 4, 24, 25], and skill-based [6, 7] imitation learning. While most task-conditioned and language-conditioned approaches seek to learn a single policy that performs well across a series of tasks, we focus on learning task-specific policies by utilizing a large multi-task prior dataset to learn a useful skill representation space, and supervising the task-specific policy using learned latent skills. These multi-task imitation approaches could be complementary to our skill representation learning pipeline, ie., we can consider incorporating language and task ID supervision into our approach.

Skill-based Imitation Learning. We adopt skill-based imitation learning as the underlying framework for our method. The objective is to learn temporally abstract representations of sensory-motor data (termed *skills*) to enable more effective imitation learning. A long line of work has proposed learning skill representations by segmenting demonstrations into trajectory segments termed *sub-trajectories*. Among these include work on segmenting demonstrations into variable-length sub-trajectories, either in an unsupervised manner [26, 27, 28, 29, 30, 31, 32] or relying on additional supervision [33, 34, 35, 36]. Recent work has shown promise for encoding fixed length sub-trajectories without any additional supervision [8, 9, 6, 7] using variational autoencoder-based approaches. We adopt this setting due its relative simplicity and scalability. In contrast to these prior approaches (see Sec. 3), we enforce two key properties—a temporal predictability objective in our learned skill representation and a retrieval-based mechanism to improve task-specific policy learning.

3 Problem Formulation

Our goal is to leverage prior data to effectively learn novel target tasks in a data-efficient manner. Formally, we consider a target task as a Markov Decision Process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, p, p_0, \gamma)$ representing the state space, action space, reward function, transition probability, initial state distribution, and discount factor. Our objective is to learn a policy that maximizes the discounted sum of rewards for the task. To learn the policy, we assume access to a small offline dataset $\mathcal{D}_{\text{target}}$ collected for the target task and a large offline dataset $\mathcal{D}_{\text{prior}}$ either from previous related tasks or task-agnostic interactions. These datasets consist of variable-length trajectories in the form $\{o_0, a_0, o_1, \cdots, o_{T_i}\}$ with $o_i \in \mathcal{O}$ denoting the observations and $a_i \in \mathcal{A}$ denoting actions. We highlight that $\mathcal{D}_{\text{prior}}$ and $\mathcal{D}_{\text{target}}$ may have significant differences, i.e., the two datasets may come from different environments, be collected by different human demonstrators, and demonstrate different tasks.

Skill-based Imitation Learning. We employ a skill-based imitation learning framework [6, 7] which consists of two stages: skill learning and policy learning. In the skill learning phase, we learn a skill embedding space $\mathcal{Z} \subset \mathbb{R}^d$ of fixed-length sub-trajectories $\tau = \{o_0, a_0, o_1, \cdots, o_{H-1}, a_{H-1}, o_H\}$ in $\mathcal{D}_{\text{prior}}$. These skill embeddings serve as an abstract representation of the agent's behavior and can be invoked to solve a range of downstream tasks. A number of representation learning methods can be employed to learn the skill embeddings, spanning reconstruction-based methods and contrastive learning. In the subsequent policy learning phase, we are given $\mathcal{D}_{\text{target}}$ and our goal is to learn a policy π for the target task $\mathcal{T}$. The policy now emits skill embeddings $z \in \mathcal{Z}$ and during execution we decode z into a sequence of actions $\{a_0, a_1, \cdots, a_{H-1}\}$ via a skill decoder model p_ψ. To learn the policy, prior work has proposed parametric [6] and semi-parametric [7] polices that first parse segments of $\mathcal{D}_{\text{target}}$ into skills and subsequently learn to map observations in $\mathcal{D}_{\text{target}}$ to these skill embeddings. Note that we use distinct terms—*skill* and *policy*—to distinguish the role of these two components. Skills represent short-horizon behaviors that can be re-used across many tasks, while the policy solves a specific long-horizon target task.

4 Skill-based Imitation Learning with Retrieval

In this section, we describe our skill-based imitation learning approach that can leverage prior multi-task data to efficiently learn novel target tasks with a small amount of task-specific demonstrations. As discussed in Sec. 3, this consists of two phases—a task-agnostic skill learning phase, where a latent skill space is learned using the prior data, and a task-specific policy learning phase where task-specific data is used to learn a policy using the skills as supervision. Compared to prior methods, our approach makes two key considerations—we (1) ensure that the learned skill space is a *predictable*

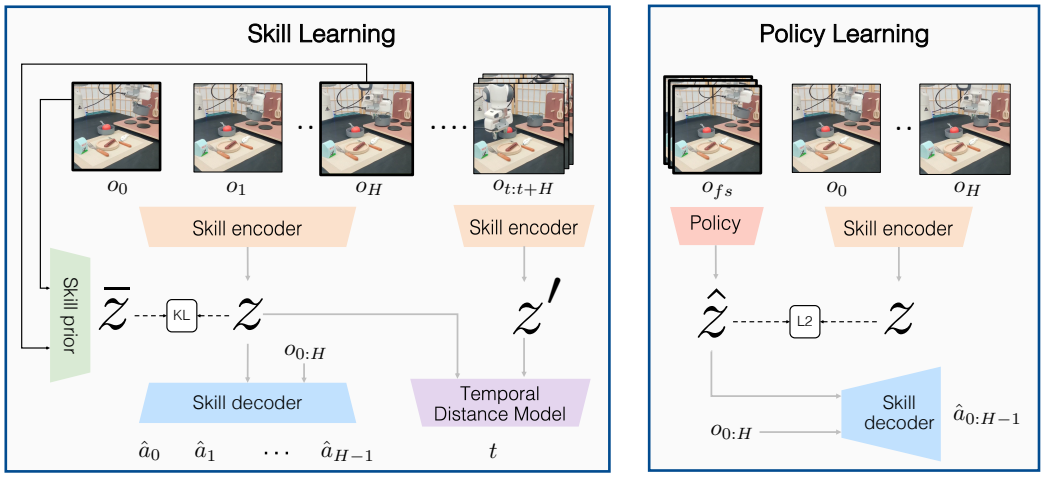

Figure 2: **Model Overview.** Our method consists of a skill learning and policy learning phase. (Left) In the skill learning phase, we learn a latent skill representation of sub-trajectories via a variational autoencoder. We include an additional temporal predictability term to learn a more consistent latent representation. (Right) In the policy learning phase, we train the policy to predict the latent skill given a history of observations preceding the sub-trajectory. To execute the policy, we decode the predicted latent using the skill decoder.

representation for downstream policy learning, and (2) improve the efficacy of task-specific policy learning by *retrieving* task-relevant datapoints from the prior dataset. See Fig. 2 for our model overview.

4.1 Learning a Predictable Representation of Skills

We learn a skill representation by encoding sub-trajectories with a variational autoencoder (VAE) [37], and we further introduce an auxiliary objective to shape the representation. Denoting a given sub-trajectory as τ, we employ a long short-term memory (LSTM) [38] encoder q_ϕ that encodes τ into a Gaussian distribution over latent skills. Our decoder is an LSTM network p_ψ that, for each timestep t, decodes a latent z and the given observation o_t into the reconstructed action $\hat{a}_t$. We additionally employ a learned prior p_θ to encourage sub-trajectories with similar starting and ending observations to have similar latent representations [15]. Our VAE loss objective is then

$$\mathcal{L}_{\text{VAE}}(\phi, \psi, \theta) = -\mathbb{E}_{z \sim q_\phi(z|\tau)} \left[\sum_{t=0}^{H-1} \log p_\psi(a_t | z, o_t) \right] + \beta \cdot D_{KL}(q_\phi(z|\tau) || p_\theta(z|o_0, o_H)), \quad (1)$$

where β controls the effect of the KL divergence term [39][1].

It is important to highlight that action reconstruction is not the sole objective of our skill learning model—learning a consistent and predictable representation of behavior is critical for downstream policy learning, as shown by recent work [40, 41]. While the KL divergence term is one step towards this objective (by encouraging skills to be predictable given partial information from the sub-trajectories), in this work we introduce an additional *temporal predictability* term that encourages the learned latent space to predict the temporal difference between two sub-trajectories. Specifically, given two-sub-trajectories τ_1 and τ_2 from the same trajectory separated by t timesteps, we learn a model m_ω to predict t given the corresponding skill mean embeddings of the trajectories:

$$\mathcal{L}_{\text{TP}}(\omega, \phi) = \Big(m_\omega \Big(\mu(q_\phi(z|\tau_1)), \mu(q_\phi(z|\tau_2)) \Big) - t \Big)^2, \quad (2)$$

where μ denotes taking the mean of the distribution. We back-propagate $\mathcal{L}_{\text{TP}}$ through the skill encoder model, allowing the this term to shape the learned skill representation. Note that this is just one way to encourage temporal predictability, other objectives are also readily compatible with our method, such as time-contrastive networks [42]. Our overall objective is a weighted combination of the VAE and temporal predictability objectives:

$$\mathcal{L}_{\text{Skill}}(\phi, \psi, \theta, \omega) = \mathcal{L}_{\text{VAE}}(\phi, \psi, \theta) + \alpha \mathcal{L}_{\text{TP}}(\omega, \phi). \quad (3)$$

[1]In practice we use a deterministic VAE decoder and we compute the reconstruction loss using ℓ_2 distance.

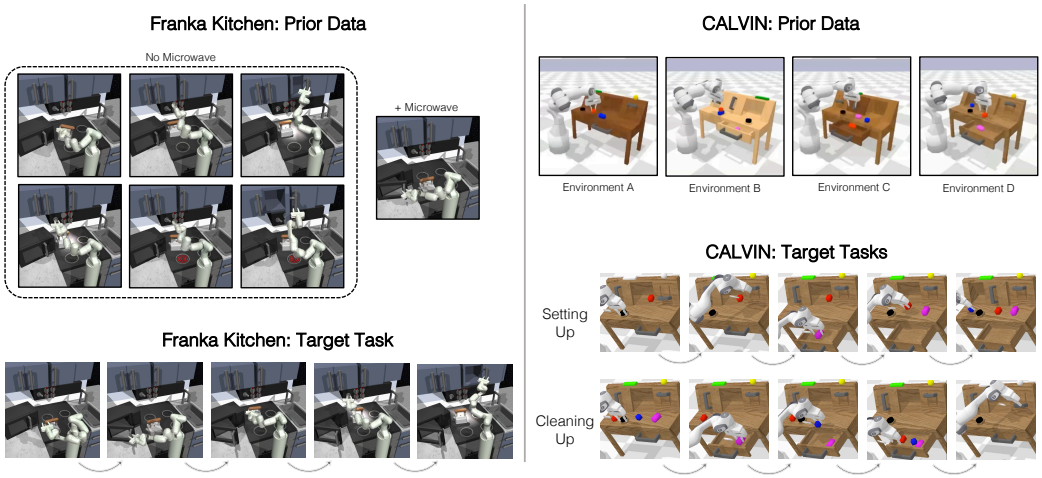

Figure 3: **Simulated Tasks.** We perform extensive evaluations on two simulation domains. (Left) Franka Kitchen: our target task involves a specific permutation of four subtasks and we consider two prior datasets: demonstrations involving all subtasks and demonstrations involving all subtasks except opening the microwave. (Right) CALVIN: we adopt the play dataset of Mees et al. [4] as our prior data and perform evaluations on two target tasks: setting up the playroom environment and, conversely, cleaning up the environment.

Refer to Algorithm 1 for a detailed summary of our skill learning algorithm.

4.2 Retrieval-based Policy Learning

In the policy learning phase, we employ an LSTM policy that outputs the skill z to execute next. We train the policy on a dataset $\mathcal{D}_{\text{policy}} = \{(o^i_{fs}, z_i = \mu(q_\phi(\tau_i))\}$, where τ_i is an H-length sub-trajectory, z_i is the mean encoding of that sub-trajectory, and o^i_{fs} is the frame-stacked history of F observations preceding the sub-trajectory. We train the policy to predict z_i from o^i_{fs} using a standard behavioral cloning loss. During execution, we roll out the LSTM skill decoder p_ψ in a closed-loop manner, i.e., at each timestep we observe o_t and execute the next action $a_t = \mu(p_\psi(z, o_t))$. After rolling out the skill for H timesteps we repeat the process by sampling a new skill from the policy.

A common approach to skill-based policy learning is training the policy with all H-length sub-trajectories in $\mathcal{D}_{\text{target}}$. However, this limits the amount of supervision, especially when $\mathcal{D}_{\text{target}}$ is small. On the other hand, naïvely training on all sub-trajectories in $\mathcal{D}_{\text{target}}$ and $\mathcal{D}_{\text{prior}}$ can hurt performance [43] due to divergent and conflicting behaviors between the prior and target datasets. We thus introduce a retrieval-based mechanism to train on sub-trajectories in $\mathcal{D}_{\text{prior}}$ that have high similarity with those in $\mathcal{D}_{\text{target}}$. While many similarity metrics are suitable, in this work we measure similarity with respect to the skill embedding space—intuitively, sub-trajectories with similar skill embeddings demonstrate similar behaviors. First we obtain skill embeddings of randomly sampled sub-trajectories in $\mathcal{D}_{\text{prior}}$ and $\mathcal{D}_{\text{target}}$:

$$Z_{\text{prior}} = \{\mu(q_\phi(\tau_i))\}, \tau_i \sim \mathcal{D}_{\text{prior}}; \qquad Z_{\text{target}} = \{\mu(q_\phi(\tau_j))\}, \tau_j \sim \mathcal{D}_{\text{target}}. \tag{4}$$

We then calculate the pairwise ℓ_2 distances between the prior and target dataset skill embeddings, *i.e.*, `D[i][j]` $= \|Z^i_{\text{prior}} - Z^j_{\text{target}}\|_2$. Next, for each prior dataset skill embedding $z_i \in Z_{\text{prior}}$, we find the closest corresponding target dataset skill, `D_min[i] = min(D[i][:])`. Finally, we retrieve the top-n sub-trajectories in $\mathcal{D}_{\text{prior}}$ with the smallest distance `argsort(D_min)[:n]`, resulting in the retrieval dataset $\mathcal{D}_{\text{ret}}$. We train the policy using the aggregated set of sub-trajectories in $\mathcal{D}_{\text{ret}}$ and $\mathcal{D}_{\text{target}}$. We use a behavioral cloning loss split across two terms: one for $\mathcal{D}_{\text{target}}$, and one for $\mathcal{D}_{\text{ret}}$ weighted by a factor γ to control the effect of the retrieval data relative to the target dataset. At the same time as training the policy we additionally fine-tune the skill model on $\mathcal{D}_{\text{target}}$. We summarize the retrieval, policy learning, and skill fine-tuning steps in Algorithm 2.

Dataset	BC-RNN	BC-RNN (FT)	BC-RNN (R3M)	IQL	IQL (UDS)	FIST	SAILOR (ours)
Kitchen-All	32.3 ± 9.0	93.0 ± 2.3	83.3 ± 2.1	65.0 ± 7.1	78.3 ± 7.0	68.7 ± 2.5	$\mathbf{95.7 \pm 1.7}$
Kitchen-No Microwave		77.7 ± 7.6			70.3 ± 7.3	83.0 ± 0.8	$\mathbf{95.0 \pm 2.2}$
CALVIN-Setting Up	33.0 ± 2.2	72.8 ± 5.4	26.0 ± 1.4	7.7 ± 1.7	16.3 ± 2.6	3.7 ± 2.4	$\mathbf{77.3 \pm 3.1}$
CALVIN-Cleaning Up	41.0 ± 1.6	61.8 ± 9.1	28.0 ± 2.2	14.7 ± 2.1	20.0 ± 3.7	9.3 ± 5.0	$\mathbf{88.0 \pm 5.1}$

Table 1: **Quantitative evaluation on two simulation domains.** We evaluate our method SAILOR against a set of six baselines and report the mean task success rate and standard deviation over three seeds (exception: six seeds for BC-RNN (FT) due to high variance). Note: for the kitchen tasks we report one number for baselines that do not involve prior data. We see that SAILOR significantly outperforms the baselines on all tasks.

5 Experiments

5.1 Simulated Experiment Setup

We perform empirical evaluations on two simulated robot manipulation domains (see Fig. 3):

Franka Kitchen [44]: A simulated kitchen environment involving different sub-tasks, such as opening cabinets, moving a kettle, and turning on a stove. This environment comes with a large dataset of approximately 600 demonstrations performing various permutations of seven subtasks. In this dataset, a subset of 18 demonstrations correspond to $\mathcal{D}_{\text{target}}$ and demonstrate a specific permutation of subtasks: opening the microwave, followed by moving the kettle, flipping on the light switch, and opening the sliding cabinet. We consider two prior datasets $\mathcal{D}_{\text{prior}}$: (1) using all demonstrations except the ones corresponding to the target task (`Kitchen-All`); and (2) using all demonstrations except those that involve interacting with the microwave (`Kitchen-No Microwave`). These prior datasets have 584 and 235 demonstrations, respectively.

CALVIN [4]: A simulated tabletop playroom environment accompanies by a large dataset of task-agnostic "play" data with 2.3M transitions. The play data encompass diverse behaviors, such as opening and closing drawers, turning on and off the lights, and picking, placing, and pushing blocks. We use all play data as $\mathcal{D}_{\text{prior}}$ to solve two target tasks. The first target task involves setting up the playroom environment in multiple stages (`CALVIN-Setting Up`). Specifically, the robot must turn on the lights, and retrieve three blocks and place them on the table. The second target task in contrast involves cleaning up the playroom environment (`CALVIN-Cleaning Up`). Specifically, the robot must open the drawer, place all three blocks into the drawer, close the drawer, and turn off the lights. For each task, we collect 30 demonstrations, which amounts to about half an hour of data collection.

The CALVIN domain is substantially more challenging than the Franka Kitchen domain, as the target tasks have a longer horizon and involve a greater number of objects. Also, in contrast to Franka Kitchen, the prior and target datasets in CALVIN are collected by *different human demonstrators* who exhibit different styles of teleoperation.

We learn vision-based policies for both domains. We refer readers to Appendix B for a detailed discussion of our tasks and datasets, and Appendix C.5 for environment implementation details.

5.2 Quantitative Analysis

We evaluate our method SAILOR against state-of-the-art imitation learning and offline reinforcement learning algorithms:

BC-RNN: behavioral cloning on $\mathcal{D}_{\text{target}}$ without prior data. We adopt the LSTM-based BC-RNN implementation in robomimic [1], which has shown superior performance over other behavioral cloning approaches.

BC-RNN (FT): BC-RNN variant that leverages prior data. We first pre-train BC-RNN on $\mathcal{D}_{\text{prior}}$ and subsequently fine-tune on $\mathcal{D}_{\text{target}}$. This baseline aims to examine the effectiveness of supervised pre-training on interaction data for imitation learning.

BC-RNN (R3M): behavioral cloning on $\mathcal{D}_{\text{target}}$ using a frozen R3M visual representation [21] pre-trained on the large-scale Ego4D video dataset [20]. This baseline intends to examine the effectiveness of using visual representations trained on natural images and videos.

Dataset	Ours	No TP	No Retrieval	All Retrieval	No Prior Data
CALVIN-Setting Up	$\mathbf{77.3 \pm 3.1}$	68.0 ± 3.7	65.0 ± 2.2	64.3 ± 10.8	50.7 ± 6.5
CALVIN-Cleaning Up	$\mathbf{88.0 \pm 5.1}$	74.7 ± 7.9	70.0 ± 0.8	65.3 ± 7.3	60.7 ± 3.1

Table 2: **Ablation Results.** We find that temporal predictability and retrieval are critical to skill-based imitation learning, as without these components the agent struggles against a naive BC-RNN (FT) baseline. In addition, we validate that prior data plays a large role in the performance of our method.

IQL: Implicit Q-Learning [45], a recent offline reinforcement learning method with state-of-the-art performance on the D4RL dataset [46]; trained on $\mathcal{D}_{\text{target}}$.

IQL (UDS): Implicit Q-Learning with Unlabeled Data Sharing [47, 48], which is a variant of IQL trained jointly on $\mathcal{D}_{\text{prior}}$ and $\mathcal{D}_{\text{target}}$, where the transitions in $\mathcal{D}_{\text{prior}}$ are labeled with the minimum reward (0 for our tasks). Singh et al. [47] and Yu et al. [48] show that this simple data augmentation procedure can effectively leverage prior data without additional rewards annotation.

FIST: Few-shot Imitation with Skill Transition Models [7], an analogue of our method that employs a semi-parametric policy to select the skill to execute next. This baseline uses the same underlying skill model as our method but a different policy learning scheme and does not involve retrieval.

Refer to Appendix C for additional implementation details on our model architecture, pseudocode, and evaluation protocols, and Appendix D for hyperparameter details. We report performance of all methods in Table 1. SAILOR greatly outperforms the baselines with an average task success rate of 89.0%. Notably it outperforms the most competitive baseline BC-RNN (FT) by 12.7%. BC-RNN performs poorly as it fails to learn an effective policy from a small number of demonstrations. In comparison, BC-RNN (R3M) shows significant improvements on the Franka Kitchen tasks, but performs worse on the CALVIN tasks. We hypothesize that this is due to the limited generalization ability of the pre-trained visual representations. The offline reinforcement learning baselines show more promising results on the Franka kitchen tasks but struggle on the more challenging CALVIN tasks. Finally, FIST significantly under-performs our method. As FIST uses the same underlying skill model as our method, we attribute the limitations of FIST to its semi-parametric policy.

5.3 Ablation Study

We perform an extensive ablation study to understand the effects of various modeling choices on our method. First, we study the effect of the temporal predictability term in Eq. (2) on downstream task performance by removing it from Eq. (3) (`No TP`). Next we study the role of retrieval by training the policy solely on sub-trajectories in $\mathcal{D}_{\text{target}}$ (`No Retrieval`). We also study the opposite case—training the policy on all of $\mathcal{D}_{\text{prior}}$ and $\mathcal{D}_{\text{target}}$ (`All Retrieval`). Finally, we study the role of prior data on our method by training the skill and policy solely on $\mathcal{D}_{\text{target}}$ (`No Prior Data`). We also report ablations in Appendix A on the size of the prior, retrieval, and target datasets, in addition to the choice of retrieval method.

We present results in Table 2 for the more challenging CALVIN tasks. First, we find that both the temporal predictability objective and the retrieval mechanism have a significant impact on the final performance. It is worth noting that removing these components makes our model degenerate into the skill-based imitation learning setting of OPAL [6]. In fact, these two ablations perform worse than the naïve BC-RNN (FT) baseline for the CALVIN-Cleaning Up task, indicating that both an effective skill representation and the retrieval mechanism play a critical role in skill-based imitation learning. The All Retrieval ablation also performs suboptimally—qualitatively we observe that the robot often was "distracted" and performed behaviors unrelated to the target task, likely due to the multimodal distributions in the prior data. Finally, the No Prior Data ablation validates the role of prior data in learning effective skills and policy. Comparing the No Retrieval ablation and the No Prior Data ablation, we see a $10 - 15\%$ gap in performance, and this is attributed to the fact that the skills in the No Retrieval ablation are additionally trained on the prior data. Despite the loss of performance in the No Prior Data ablation, we still see that it outperforms the BC-RNN baseline by a significant margin. We attribute this to the temporal abstraction afforded by our skill-based learning framework. In sum, our ablation studies suggest that effective skill abstractions, coupled with mechanisms that effectively leverage the prior data, allow us to achieve strong results.

Figure 4: **Real World Tasks.** On the left, we illustrate the set of objects we use for collecting the play dataset. On the right shows two of three target tasks, setting up breakfast and cooking.

5.4 Real World Experiments

Finally, we showcase the efficacy of our method in the real world with a kitchen environment involving eight food items, receptacles, a stove, and a serving area (see Fig. 4). We first collect a play dataset of exploratory interactions involving the food items and receptacles. We consider three target tasks: (1) `Real-Breakfast`: setting up a breakfast table by placing the bread, butter, and milk in the serving area; (2) `Real-Cook`: cooking a meal by placing the fish, sausage, and tomato into the frying pan; (3) `Real-Cook-Pan`: a variant of the Real-Cook task involving placing the pan onto the stove. We collect 30 demonstrations; refer to Appendix B.3 for detailed descriptions on the tasks datasets. We evaluate SAILOR against the most competitive baseline, BC-RNN (FT) (see Appendix C.3 for our evaluation protocol). We find that while on `Real-Breakfast` both methods achieve a success rate of 76.7%, on `Real-Cook` our method significantly outperforms BC-RNN (FT) with a success rate of 73.3% vs. 23.3%, and similarly for `Real-Cook-Pan` (76.7% vs. 46.7%). To see the value of prior data we ran the No Prior ablation for the `Real-Cook` task, achieving 53.3% success rate compared to our 73.3%. Interestingly we see that the No Prior ablation largely outperforms the BC-RNN (FT) baseline on this task (23.3%) which had access to additional prior data. Overall, we observe that BC-RNN (FT) often failed to correctly grasp objects. One hypothesis for this result is that pre-training stage biases the policy to learn the multi-modal behaviors in the prior dataset, preventing the policy from learning specialized target task behaviors during the fine-tuning phase.

6 Limitations

While our method shows significant promise, it leaves limitations that we hope to address in future work. First, acquiring large amounts of multi-task prior data is difficult and costly. To amortize the high cost, large prior multi-task datasets should be useful in a diverse range of downstream tasks, rather than a handful. In this work, we evaluate our method in a limited set of target tasks and leave it for future work to scale up the variety of tasks. We hope (and believe) that the need for large robotic datasets will be addressed in the coming years [10, 23]. Second, our method is more computationally expensive than the BC-RNN baseline [1], due to the higher number of losses and networks used. Third, our experiments focus on domains and datasets where the prior data and target tasks are reasonably close to each other. Notably our experiments do not evaluate generalization to unseen objects between the prior and target datasets. It would be interesting to investigate methods that are tolerant to much larger domain shifts between prior and target task data.

7 Conclusion

We present SAILOR, a skill-based imitation learning framework for robot manipulation. Our method uses prior data to construct a latent space of predictable and consistent skill representations. It uses these latent skills as the temporal abstraction to learn policies for vision-based manipulation. Key to its effectiveness is our newly designed representation learning objectives and retrieval-based data augmentation procedure. We demonstrate that our method can solve long-horizon manipulation tasks in simulation and on physical hardware. It brings forth a data-efficient way of programming robots with new behaviors using a small number of target task demonstrations. For future work, we plan to address the limitations we discussed in the previous section and investigate the effectiveness of this approach with various forms of prior data at different scales.

Acknowledgments

We would like to thank Jake Grigsby, Huihan Liu, and Zhenyu Jiang for providing feedback on this manuscript. We would also like to thank Yifeng Zhu for real robot infrastructure support. We acknowledge the support of the National Science Foundation (1955523, 2145283), the Office of Naval Research (N00014-22-1-2204), and Amazon.

References

[1] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning*, 2021.

[2] P. Florence, C. Lynch, A. Zeng, O. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson. Implicit behavioral cloning. In *Conference on Robot Learning*, 2021.

[3] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, and J. Lee. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, 2020.

[4] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard. Calvin - a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7327–7334, 2022.

[5] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. In *Robotics: Science and Systems (RSS)*, 2022.

[6] A. Ajay, A. Kumar, P. Agrawal, S. Levine, and O. Nachum. Opal: Offline primitive discovery for accelerating offline reinforcement learning. In *International Conference on Learning Representations*, 2021.

[7] K. Hakhamaneshi, R. Zhao, A. Zhan, P. Abbeel, and M. Laskin. Hierarchical few-shot imitation with skill transition models. In *International Conference on Learning Representations*, 2021.

[8] K. Pertsch, Y. Lee, and J. J. Lim. Accelerating reinforcement learning with learned skill priors. In *Conference on Robot Learning*, 2020.

[9] K. Pertsch, Y. Lee, Y. Wu, and J. J. Lim. Demonstration-guided reinforcement learning with learned skills. In *Conference on Robot Learning*, 2021.

[10] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, S. Savarese, and L. Fei-Fei. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, 2018.

[11] J. Wong, A. Tung, A. Kurenkov, A. Mandlekar, L. Fei-Fei, S. Savarese, and R. Martín-Martín. Error-aware imitation learning from teleoperation data for mobile manipulation. In *Conference on Robot Learning*, 2021.

[12] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[13] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In *Robotics: Science and Systems (RSS)*, 2018.

[14] A. Mandlekar, F. Ramos, B. Boots, S. Savarese, L. Fei-Fei, A. Garg, and D. Fox. Iris: Implicit reinforcement without interaction at scale for learning control from offline robot manipulation data. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[15] C. Lynch, M. Khansari, T. Xiao, V. Kumar, J. Tompson, S. Levine, and P. Sermanet. Learning latent plans from play. In *Conference on Robot Learning*, 2019.

[16] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn. Robonet: Large-scale multi-robot learning. In *Conference on Robot Learning*, 2019.

[17] S. Cabi, S. Gómez Colmenarejo, A. Novikov, K. Konyushkova, S. Reed, R. Jeong, K. Zolna, Y. Aytar, D. Budden, M. Vecerik, et al. Scaling data-driven robotics with reward sketching and batch reinforcement learning. In *Robotics: Science and Systems (RSS)*, 2020.

[18] R. Goyal, S. E. Kahou, V. Michalski, J. Materzyńska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense. *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[19] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[20] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022.

[21] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation, 2022. URL https://arxiv.org/abs/2203.12601.

[22] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik. Masked visual pre-training for motor control, 2022. URL https://arxiv.org/abs/2203.06173.

[23] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, 2021.

[24] M. Shridhar, L. Manuelli, and D. Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, 2021.

[25] C. Lynch and P. Sermanet. Language conditioned imitation learning over unstructured data. In *Robotics: Science and Systems (RSS)*, 2021.

[26] G. Konidaris, S. Kuindersma, R. Grupen, and A. Barto. Robot learning from demonstration by constructing skill trees. *International Journal of Robotics Research*, 2012.

[27] S. Niekum, S. Osentoski, G. Konidaris, and A. G. Barto. Learning and generalization of complex tasks from unstructured demonstrations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[28] S. Krishnan, R. Fox, I. Stoica, and K. Goldberg. DDCO: Discovery of deep continuous options for robot learning from demonstrations. In *Conference on Robot Learning*, 2017.

[29] T. Shankar and A. Gupta. Learning robot skills with temporal variational inference. In *International Conference on Machine Learning*, 2020.

[30] T. Shankar, S. Tulsiani, L. Pinto, and A. Gupta. Discovering motor programs by recomposing demonstrations. In *International Conference on Learning Representations*, 2020.

[31] T. Kipf, Y. Li, H. Dai, V. Zambaldi, A. Sanchez-Gonzalez, E. Grefenstette, P. Kohli, and P. Battaglia. Compile: Compositional imitation learning and execution. In *International Conference on Machine Learning*, 2019.

[32] D. Tanneberg, K. Ploeger, E. Rueckert, and J. Peters. Skid raw: Skill discovery from raw trajectories. *IEEE Robotics and Automation Letters (RA-L)*, 2021.

[33] K. Shiarlis, M. Wulfmeier, S. Salter, S. Whiteson, and I. Posner. Taco: Learning task decomposition via temporal alignment for control. In *International Conference on Machine Learning*, 2018.

[34] Z. Su, O. Kroemer, G. E. Loeb, G. S. Sukhatme, and S. Schaal. Learning manipulation graphs from demonstrations using multimodal sensory signals. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[35] Y. Zhu, P. Stone, and Y. Zhu. Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation. *IEEE Robotics and Automation Letters (RA-L)*, 2022.

[36] S. Belkhale and D. Sadigh. Plato: Predicting latent affordances through object-centric play, 2022.

[37] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[38] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[39] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. β-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[40] A. Allshire, R. Martín-Martín, C. Lin, S. Manuel, S. Savarese, and A. Garg. Laser: Learning a latent action space for efficient reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[41] M. Yang, S. Levine, and O. Nachum. Trail: Near-optimal imitation learning with suboptimal data. In *International Conference on Learning Representations*, 2022.

[42] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, and S. Levine. Time-contrastive networks: Self-supervised learning from video. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[43] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020.

[44] A. Gupta, V. Kumar, C. Lynch, S. Levine, and K. Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. In *Conference on Robot Learning*, 2021.

[45] I. Kostrikov, A. Nair, and S. Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022.

[46] J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020. URL https://arxiv.org/abs/2004.07219.

[47] A. Singh, A. Yu, J. Yang, J. Zhang, A. Kumar, and S. Levine. Cog: Connecting new skills to past experience with offline reinforcement learning. 2020.

[48] T. Yu, A. Kumar, Y. Chebotar, K. Hausman, C. Finn, and S. Levine. How to leverage unlabeled data in offline reinforcement learning. In *International Conference on Machine Learning*, 2022.

[49] O. Khatib. Inertial properties in robotic manipulation: An object-level framework. *International Journal of Robotics Research*, 1995.

[50] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, 2016.

[51] C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, and P. Abbeel. Deep spatial autoencoders for visuomotor learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[52] A. Mandlekar, D. Xu, R. Martín-Martín, S. Savarese, and L. Fei-Fei. Learning to generalize across long-horizon tasks from human demonstrations. In *Robotics: Science and Systems (RSS)*, 2020.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[54] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.