Premier-TACO is a Few-Shot Policy Learner: Pretraining Multitask Representation via Temporal Action-Driven Contrastive Loss

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce `Premier-TACO`, a novel multitask feature representation learning methodology aiming to enhance the efficiency of few-shot policy learning in sequential decision-making tasks. `Premier-TACO` pretrains a general feature representation using a small subset of relevant multitask offline datasets, capturing essential environmental dynamics. This representation can then be fine-tuned to specific tasks with few expert demonstrations. Building upon the recent temporal action contrastive learning (TACO) objective, which obtains the state of art performance in visual control tasks, `Premier-TACO` additionally employs a simple yet effective negative example sampling strategy. This key modification ensures computational efficiency and scalability for large-scale multitask offline pretraining. Experimental results from both Deepmind Control Suite and MetaWorld domains underscore the effectiveness of `Premier-TACO` for pretraining visual representation, facilitating efficient few-shot imitation learning of unseen tasks.

1 Introduction

Just as foundation models in language, like BERT [5] and GPT [22, 3], have revolutionized natural language processing by leveraging vast amounts of textual data to understand linguistic nuances, *pretrained foundation models* hold similar promise for sequential decision-making (SDM). In SDM, where decisions are influenced by a complex interplay of past actions, current states, and future possibilities, a pretrained foundation model can provide a rich, generalized understanding of decision sequences. This foundational knowledge, built upon diverse decision-making scenarios, can then be fine-tuned to specific tasks, much like how language models are adapted to specific linguistic tasks.

The following **challenges** are unique to sequential decision-making, setting it apart from existing vision and language pretraining paradigms. **(C1) Data Distribution Shift**: Training data usually consists of specific behavior-policy-generated trajectories. This leads to vastly different data distributions at various stages—pretraining, finetuning, and deployment—resulting in compromised performance [14]. **(C2) Task Heterogeneity**: Unlike language and vision tasks, which often share semantic features, decision-making tasks vary widely in configurations, transition dynamics, and state and action spaces. This makes it difficult to develop a universally applicable representation. **(C3) Data Quality and Supervision**: Effective representation learning often relies on high-quality data and expert guidance. However, these resources are either absent or too costly to obtain in many real-world decision-making tasks [2, 27]. Our **aspirational criteria** for foundation model for sequential decision-making encompass several key features: **(W1) Versatility** that allows the

Submitted to NeurIPS 2023 6th Robot Learning Workshop: Pretraining, Fine-Tuning, and Generalization with Large Scale Models. Do not distribute.

model to generalize across a wide array of tasks, even those not previously encountered, such as new embodiments viewed or observations from novel camera angles; **(W2) Efficiency** in adapting to downstream tasks, requiring minimal data through few-shot learning techniques; **(W3) Robustness** to pretraining data of fluctuating quality, ensuring a resilient foundation; and **(W4) Compatibility** with existing large pretrained models such as [20].

In this paper, rather than focusing on leveraging large computational vision datasets [20, 16, 15] that overlook control-relevant considerations and suffer from a domain gap between pre-training datasets and downstream visuo-motor tasks, we propose a novel control-centric objective function for pretraining. Our approach, called `Premier-TACO` (pretraining multitask representation via temporal action-driven contrastive loss), employs a temporal action-driven contrastive loss function for pretraining. Unlike TACO, which treats every data point in the batch as a potential negative example, `Premier-TACO` samples one negative example from a nearby window of the next state, yielding a negative example that is visually similar to the positive one. Consequently, the latent representation must encapsulate control-relevant information to differentiate between the positive and negative examples, rather than depending on irrelevant features such as visual appearance. This simple yet effective negative example sampling strategy incurs zero computational overhead, allowing for effortless scalability in multitask offline pretraining. Through extensive empirical evaluation, we demonstrate the **versatility** and **efficiency** of `Premier-TACO`' representations in terms of generalization to downstream tasks involving unseen embodiments and views, **robustness** to low-quailty data and **compatibility** for finetuneing a pretrained visual encoder such as R3M [20], resulting in an average performance improvement of approximately 50% across the evaluated tasks.

2 Preliminary

TACO: Temporal Action Driven Contrastive Learning Objective Temporal Action-driven Contrastive Learning (TACO) [40] is a reinforcement learning algorithm proposed for addressing the representation learning problem in visual continuous control. It aims to maximize the mutual information between representations of current states paired with action sequences and representations of the corresponding future states:

$$\mathbb{J}_{\text{TACO}} = \mathcal{I}(Z_{t+K}; [Z_t, U_t, ..., U_{t+K-1}]) \tag{1}$$

Here, $Z_t = \phi(X_t)$ and $U_t = \psi(A_t)$ represents latent state and action variables. Theoretically, it could be shown that maximization of this mutual information objective leads to state and action representations that are capable of representing the optimal value functions. Empirically, TACO estimates the lower bound of the mutual information objective by the InfoNCE loss, and it achieves the state of the art performance for both online and offline visual continuous control, demonstrating the effectiveness of temporal contrastive learning for representation learning in sequential decision making problems.

3 Method

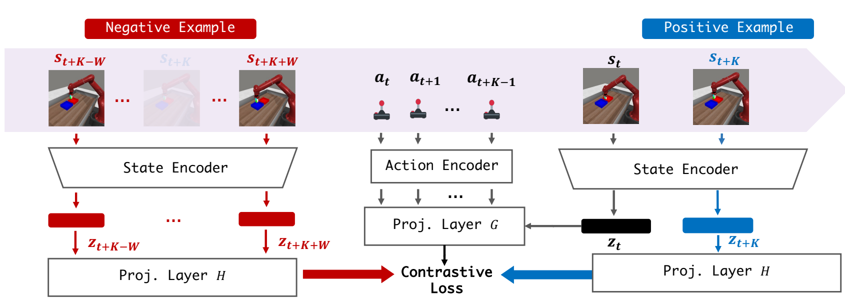

Figure 1: An illustration of `Premier-TACO` contrastive loss design. The two 'State Encoder's are identical, as are the two 'Proj. Layer H's. One negative example is sampled from the neighbors of framework s_{t+K}.

We introduce `Premier-TACO`, a generalized pre-training approach specifically formulated to tackle the multi-task pre-training problem, enhancing sample efficiency and generalization ability for downstream tasks. Building upon the success of temporal contrastive loss, exemplified by `TACO` [40], in acquiring latent state representations that encapsulate individual task dynamics, our aim is to foster representation learning that effectively captures the intrinsic dynamics spanning a diverse set of tasks found in offline datasets. Our overarching objective is to ensure that these learned representations exhibit the versatility to generalize across unseen tasks that share the underlying dynamic structures.

Nevertheless, when adapted for multitask offline pre-training, the online learning objective of TACO [40] poses a notable challenge. Specifically, TACO's mechanism, which utilizes the InfoNCE [32] loss, categorizes all subsequent states s_{t+k} in the batch as negative examples. While this methodology has proven effective in single-task reinforcement learning scenarios, it encounters difficulties when extended to a multitask context. During multitask offline pretraining, image observations within a batch can come from different tasks with vastly different visual appearances, rendering the contrastive InfoNCE loss significantly less effective.

Offline Pretraining Objective. We propose a straightforward yet highly effective mechanism for selecting challenging negative examples. Instead of treating all the remaining examples in the batch as negatives, `Premier-TACO` selects the negative example from a window centered at state s_{t+k} within the same episode.

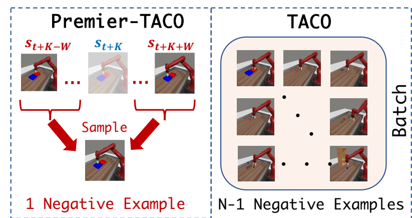

Figure 2: Difference between `Premier-TACO` and TACO for sampling negative examples

This approach is both computationally efficient and more statistically powerful due to negative examples which are challenging to distinguish from similar positive examples forcing the model to capture temporal dynamics differentiating between positive and negative examples. Specifically, given a batch of state and action sequence transitions $\{(s_t^{(i)}, [a_t^{(i)}, ..., a_{t+K-1}^{(i)}], s_{t+K}^{(i)})\}_{i=1}^N$, let $z_t^{(i)} = \phi(s_t^{(i)})$, $u_t^{(i)} = \psi(a_t^{(i)})$ be latent state and latent action embeddings respectively. Furthermore, let $\widetilde{s_{t+K}^{(i)}}$ be a negative example uniformly sampled from the window of size W centered at s_{t+K}: $(s_{t+K-W}, ..., s_{t+K-1}, s_{t+K+1}, ..., s_{t+K+W})$ with $\widetilde{z_t^{(i)}} = \phi(\widetilde{s_t^{(i)}})$ a negative latent state. Given these, define $g_t^{(i)} = G_\theta(z_t^{(i)}, u_t^{(i)}, ..., u_{t+K-1}^{(i)})$, $h_t^{(i)} = H_\theta(z_{t+K}^{(i)})$, and $\widetilde{h_t^{(i)}} = H_\theta(\widetilde{z_{t+K}^{(i)}})$ as embeddings of future predicted and actual latent states. We optimize:

$$\mathcal{J}_{\texttt{Premier-TACO}}(\phi, \psi, G_\theta, H_\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{{g_t^{(i)}}^\top h_{t+K}^{(i)}}{{g_t^{(i)}}^\top h_{t+K}^{(i)} + {\widetilde{g_t^{(i)}}}^\top h_{t+K}^{(i)}}. \tag{2}$$

4 Experiment

In our empirical evaluations, we consider two benchmarks, Deepmind Control Suite [31] for locomotion control as well as MetaWorld [37] for robotic manipulation tasks. The visualization of pretrain and test tasks on different domains is shown in Figure 4.

Setup and Baselines. The detailed introduction of pretrained tasks for `Premier-TACO` and baselines in our comparison can be found in Appendix C.1.

Pretrained feature representation by `Premier-TACO` facilitates effective few-shot adaptation to unseen tasks. We measure the performance of pretrained visual representations for few-shot imitation learning of unseen downstream tasks in both DMC and MetaWorld. Note that we only use $\frac{1}{5}$ of the number of expert trajectories used in [16] and $\frac{1}{10}$ of those used in [29]. In Table 1, we present the results for Deepmind Control Suite. The results for MetaWorld are provided in Table 2 of Appendix C. As shown here, pretrained representation of `Premier-TACO` significantly improves the few-shot imitation learning performance compared with Learn-from-scratch, with a **101%** improvement on Deepmind Control Suite and **74%** improvement on MetaWorld, respectively. Moreover, it also outperforms all the baselines across all tasks by a large margin.

	DMControl Tasks	Models LfS	SMART	Best PVRs	TD3+BC	Inverse	CURL	ATC	SPR	Premier-TACO
Seen Embodiments	Finger Spin	34.8±3.4	44.2 ± 8.2	38.4 ± 9.3	68.8 ± 7.1	33.4±8.4	35.1±9.6	51.1±9.4	55.9±6.2	**75.2 ± 0.6**
	Hopper Hop	8.0 ± 1.3	14.2 ± 3.9	23.2 ± 4.9	49.1 ± 4.3	48.3±5.2	28.7±5.2	34.9±3.9	52.3±7.8	**75.3 ± 4.6**
	Walker Walk	30.4±2.9	54.1 ± 5.2	32.6 ± 8.7	65.8 ± 2.0	64.4±5.6	37.3±7.9	44.6±5.0	72.9±1.5	**88.0 ± 0.8**
	Humanoid Walk	15.1±1.3	18.4 ± 3.9	30.1 ± 7.5	34.9 ± 8.5	41.9±8.4	19.4±2.8	35.1±3.1	30.1±6.2	**51.4 ± 4.9**
	Dog Trot	52.7±3.5	59.7 ± 5.2	73.5 ± 6.4	82.3 ± 4.4	85.3±2.1	71.9±2.2	84.3±0.5	79.9±3.8	**93.9 ± 5.4**
Unseen Embodiments	Cup Catch	56.8±5.6	66.8 ± 6.2	93.7 ± 1.8	97.1 ± 1.7	96.7±2.6	96.7±2.6	96.2±1.4	96.9±3.1	**98.9 ± 0.1**
	Reacher Hard	34.6±4.1	52.1 ± 3.8	64.9 ± 5.8	59.6 ± 9.9	61.7±4.6	50.4±4.6	56.9±9.8	62.5±7.8	**81.3 ± 1.8**
	Cheetah Run	25.1±2.9	41.1 ± 7.2	39.5 ± 9.7	50.9 ± 2.6	51.5±5.5	36.8±5.4	30.1±1.0	40.2±9.6	**65.7 ± 1.1**
	Quadruped Walk	61.1±5.7	45.4 ± 4.3	63.2 ± 4.0	76.6 ± 7.4	82.4±6.7	72.8±8.9	81.9±5.6	65.6±4.0	**83.2 ± 5.7**
	Quadruped Run	45.0±2.9	27.9 ± 5.3	64.0 ± 2.4	48.2 ± 5.2	52.1±1.8	55.1±5.4	2.6 ± 3.6	68.2±3.2	**76.8 ± 7.5**
Mean Performance		38.2	42.9	52.3	63.3	61.7	50.4	52.7	62.4	**79.0**

Table 1: [(W1) Versatility (W2) Efficiency] Few-shot Behavior Cloning (BC) for unseen task of DMC. Performance (Agent Reward / Expert Reward) of baselines and Premier-TACO on 10 unseen tasks on Deepmind Control Suite. **Bold** numbers indicate the best results. Agent Policies are evaluated every 1000 gradient steps for a total of 100000 gradient steps and we report the average performance over the 3 best epochs over the course of learning. Premier-TACO outperforms all the baselines, showcasing its superior efficacy in generalizing to unseen tasks with seen or **unseen embodiments**.

Premier-TACO pre-trained representation enables knowledge sharing across different embodiments. Ideally, a resilient and generalizable state feature representation ought not only to encapsulate universally applicable features for a given embodiment across a variety of tasks, but also to exhibit the capability to generalize across distinct embodiments. Here, we evaluate the few-shot behavior cloning performance of Premier-TACO pre-trained encoder from **DMC-6** on four tasks featuring unseen embodiments: Cup Catch, Cheetah Run, and Quadruped Walk. In comparison to Learn-from-scratch, as shown in Table 1, Premier-TACO pre-trained representation realizes an **82%** performance gain, demonstrating the robust generalizability of our pre-trained feature representations.

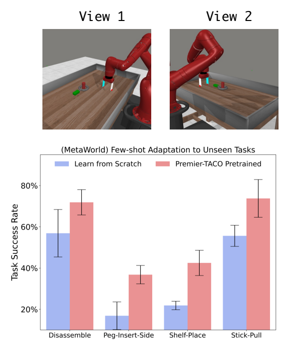

Figure 3: [(W1) Versatility] MetaWorld: Few-shot adaptation to unseen tasks from an unseen camera view.

Premier-TACO Pretrained Representation is also generalizable to unseen tasks with camera views. Beyond generalizing to unseen embodiments, an ideal robust visual representation should possess the capacity to adapt to unfamiliar tasks under novel camera views. In Figure 3, we evaluate the five-shot learning performance of our model on four previously unseen tasks in MetaWorld with a new view. In particular, during pretraining, the data from MetaWorld are generated using the same view as employed in [10, 26]. Then for downstream policy learning, the agent is given five expert trajectories under a different corner camera view, as depicted in the figure. Notably, Premier-TACO also achieves a substantial performance enhancement, thereby underscoring the robust generalizability of our pretrained visual representation.

Robustness to low-quality pretraining data. To further study the resilience of Premier-TACO, we employ low-quality data to train Premier-TACO representations in Appendix C.3.

Compatibility: Pretrained visual encoder finetuning with Premier-TACO. To further validate the compatibility of our Premier-TACO approach, we compared the results of R3M with no fine-tuning, in-domain fine-tuning [9], and fine-tuning using our method on selected Deepmind Control Suite and MetaWorld pretraining tasks. Results in Appendix C.4 unequivocally demonstrate that direct fine-tuning on in-domain tasks leads to a performance decline across multiple tasks. However, leveraging the Premier-TACO learning objective for fine-tuning substantially enhances the performance of R3M. This not only underscores the role of our method in bridging the domain gap and capturing essential control features but also highlights its robust generalization capabilities. Furthermore, these findings strongly suggest that our Premier-TACO approach is highly adaptable to a wide range of multi-task pretraining scenarios, irrespective of the model's size or the size of the pretrained data.

Ablation Studies. Ablation experiments for batch sizes and window sizes are in Appendix D.

A Problem Setting

A.1 Multitask Offline Pretraining

We consider a collection of tasks $\left\{\mathcal{T}_i : (\mathcal{X}, \mathcal{A}_i, \mathcal{P}_i, \mathcal{R}_i, \gamma)\right\}_{i=1}^{N}$ with the same dimensionality in observation space $\mathcal{X}$. Let $\phi : \mathcal{X} \rightarrow \mathcal{Z}$ be a representation function of the agent's observation, which is either randomly initialized or pre-trained already on a large-scale vision dataset such as ImageNet [4] or Ego4D [7]. Assuming that the agent is given a multitask offline dataset $\{(x_i, a_i, x_i', r_i)\}$ of a subset of K tasks $\{\mathcal{T}_{n_j}\}_{j=1}^{K}$. The objective is to pretrain a generalizable state representation ϕ or a motor policy π so that when facing an unseen downstream task, it could quickly adapt with few expert demonstrations, using the pretrained representation.

Below we summarize the pretraining and finetuning setups.

Pretraining: The agent get access to a multitask offline dataset, which could be highly suboptimal. The goal is to learn a generalizable shared state representation from pixel inputs.

Adaptation: Adapt to unseen downstream task from few expert demonstration with imitation learning.

B Related Work

Pretraining Visual Representations. Existing works apply self-supervised pre-training from rich vision data to build foundation models. However, applying this approach to sequential decision-making tasks is challenging. Recent works have explored large-scale pre-training with offline data in the context of reinforcement learning. Efforts such as R3M [20], VIP [15], MVP [34], PIE-G [38], and VC-1 [16] highlight this direction. However, there's a notable gap between the datasets used for pre-training and the actual downstream tasks. In fact, a recent study [9] found that models trained from scratch can often perform better than those using pre-trained representations, suggesting the limitation of these approachs. It's important to acknowledge that these pre-trained representations are not control-relevant, and they lack explicit learning of a latent world model. In contrast to these prior approaches, our pretrained representations learn to capture the control-relevant features with an effective temporal contrastive learning objective.

For control tasks, several pretraining frameworks have emerged to model state-action interactions from high-dimensional observations by leveraging causal attention mechanisms. SMART [29] introduces a self-supervised and control-centric objective to train transformer-based models for multitask decision-making, although it requires additional fine-tuning with large number of demonstrations during downstream time. As an improvement, DualMind [33] pretrains representations using 45 tasks for general-purpose decision-making without task-specific fine-tuning. Besides, some methods [25, 18, 35, 30] first learn a general representation by exploring the environment online, and then use this representation to train the policy on downstream tasks. In comparison, our approach is notably more efficient and doesn't require training with such an extensive task set. Nevertheless, we provide empirical evidence demonstrating that our method can effectively handle multi-task pretraining.

Contrastive Representation for Visual RL Contrastive learning is a self-supervised technique that leverages similarity constraints between data to learn effective representations (embeddings), and it has demonstrated remarkable success across various domains. In the context of visual reinforcement learning (RL), contrastive learning plays a pivotal role in training robust state representations from raw visual inputs, thereby enhancing sample efficiency. CURL [13] extracts high-level features by utilizing InfoNCE[32] to maximize agreement between augmented observations, although it does not explicitly consider temporal relationships between states. Several approaches, such as CPC [11], ST-DIM [1], and ATC [28] , introduce temporal dynamics into the contrastive loss. They do so by maximizing mutual information between states with short temporal intervals, facilitating the capture of temporal dependencies. DRIML [17] proposes a policy-dependent auxiliary objective that enhances agreement between representations of consecutive states, specifically considering the first action of the action sequence. Recent advancements by [12, 39] incorporate actions into the contrastive loss, emphasizing behavioral similarity. TACO [40] takes a step further by learning both state and action representations. It optimizes the mutual information between the representations of current states paired with action sequences and the representations of corresponding future states. In our approach, we build upon the efficient extension of TACO, harnessing the full potential of state and action representations for downstream tasks. On the theory side, the Homer algorithm [19] uses a binary temporal contrastive objective reminiscent of the approach used here, which differs by abstracting actions as well states, using an ancillary embedding, removing leveling from the construction, and of course extensive empirical validation.

C Experiments

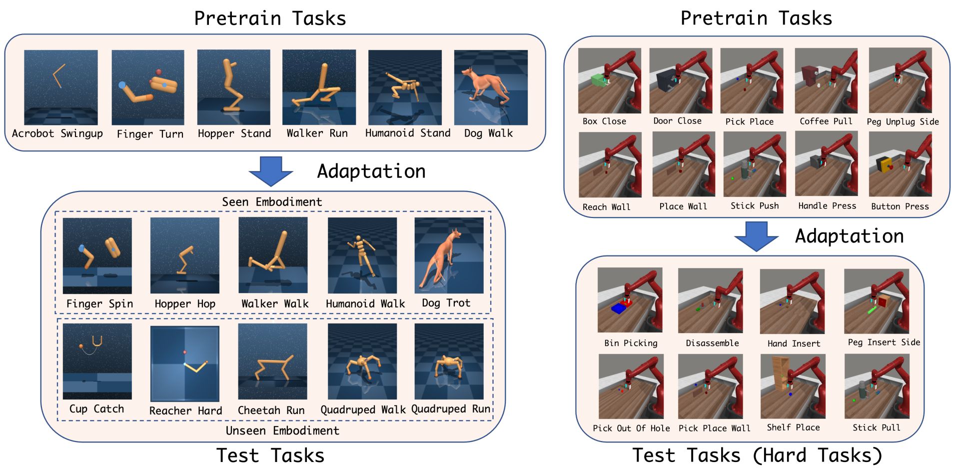

Figure 4: Pretrain and Test Tasks split for Deepmind Control Suite and MetaWorld. The left figures are Deepmind Control Suite tasks and the right figures MetaWorld tasks.

C.1 Experiment Setup

Deepmind Control Suite (DMC) [31]: We consider a selection of 16 challenging tasks from Deepmind Control Suite. Note that compared with prior works such as [16, 29], we consider much harder tasks, including ones from the humanoid and dog domains, which feature intricate kinematics, skinning weights and collision geometry. For pretraining, we select six tasks (**DMC-6**), including Acrobot Swingup, Finger Turn Hard, Hopper Stand, Walker Run, Humanoid Walk, and Dog Stand. We generate an exploratory dataset for each task by sampling trajectories generated in exploratory stages of a DrQ-v2 [36] learning agent. In particular, we sample 1000 trajectories from the online replay buffer of DrQ-v2 once it reaches the convergence performance. This ensures the diversity of the pretraining data, but in practice, such a high-quality dataset could be hard to obtain. So, later in the experiments, we will also relax this assumption and consider pretrained trajectories that are sampled from uniformly random actions.

MetaWorld [37]: We select a set of 10 tasks for pretraining, which encompasses a variety of motion patterns of the Sawyer robotic arm and interaction with different objects. To collect an exploratory dataset for pretraining, we execute the scripted policy with Gaussian noise of a standard deviation of 0.3 added to the action. By adding such a noise, the success rate of collected policies on average is only around 20% across ten pretrained tasks.

Baselines. We compare `Premier-TACO` with the following representation pretraining baselines:

- ▷ Learn from Scratch: Behavior Cloning with randomly initialized shallow ConvNet encoder. Different from [20, 16], which use a randomly initialized ResNet for evaluation, we find that using a shallow network with an input image size of 84×84 on both Deepmind Control Suite and MetaWorld yields superior performance. Additionally, we also include data augmentation into behavior cloning following [9].
- ▷ Policy Pretraining: We first train a multitask policy by TD3+BC [6] on the pretraining dataset. While numerous alternative offline RL algorithms exist, we choose TD3+BC as a representative

MetaWorld					Models				
Unseen Tasks	LfS	SMART	Best PVRs	TD3+BC	Inverse	CURL	ATC	SPR	Premier-TACO
Bin Picking	62.5 ± 12.5	71.3 ± 9.6	60.2 ± 4.3	50.6 ± 3.7	55.0 ± 7.9	45.6 ± 5.6	55.6 ± 7.8	67.9 ± 6.4	$\mathbf{78.5 \pm 7.2}$
Disassemble	56.3 ± 6.5	52.9 ± 4.5	70.4 ± 8.9	56.9 ± 11.5	53.8 ± 8.1	66.2 ± 8.3	45.6 ± 9.8	48.8 ± 5.4	$\mathbf{86.7 \pm 8.9}$
Hand Insert	34.7 ± 7.5	34.1 ± 5.2	35.5 ± 2.3	46.2 ± 5.2	50.0 ± 3.5	49.4 ± 7.6	51.2 ± 1.3	52.4 ± 5.2	$\mathbf{75.0 \pm 7.1}$
Peg Insert Side	28.7 ± 2.0	20.9 ± 3.6	48.2 ± 3.6	30.0 ± 6.1	33.1 ± 6.2	28.1 ± 3.7	31.8 ± 4.8	39.2 ± 7.4	$\mathbf{62.7 \pm 4.7}$
Pick Out Of Hole	53.7 ± 6.7	65.9 ± 7.8	66.3 ± 7.2	46.9 ± 7.4	50.6 ± 5.1	43.1 ± 6.2	54.4 ± 8.5	55.3 ± 6.8	$\mathbf{72.7 \pm 7.25}$
Pick Place Wall	40.5 ± 4.5	62.8 ± 5.9	63.2 ± 9.8	63.8 ± 12.4	71.3 ± 11.3	73.8 ± 11.9	68.7 ± 5.5	72.3 ± 7.5	$\mathbf{80.2 \pm 8.2}$
Shelf Place	26.3 ± 4.1	57.9 ± 4.5	32.4 ± 6.5	45.0 ± 7.7	36.9 ± 6.7	35.0 ± 10.8	35.6 ± 10.7	38.0 ± 6.5	$\mathbf{70.4 \pm 8.1}$
Stick Pull	46.3 ± 7.2	65.8 ± 8.2	52.4 ± 5.6	72.3 ± 11.9	57.5 ± 9.5	43.1 ± 15.2	72.5 ± 8.9	68.5 ± 9.4	$\mathbf{80.0 \pm 8.1}$
Mean	43.6	53.9	53.6	51.5	51.0	48.3	51.9	55.3	**75.8**

Table 2: [(W1) Versatility (W2) Efficiency] Five-shot Behavior Cloning (BC) for unseen task of MetaWorld. Success rate of `Premier-TACO` and baselines across 8 hard unseen tasks on MetaWorld. Results are aggregated over 4 random seeds. **Bold** numbers indicate the best results.

due to its simplicity and great empirical performance. After pretraining, we take the pretrained ConvNet encoder and drop the policy MLP layers.

- ▷ Pretrained Visual Representations (PVRs): We evaluate the state-of-the-art frozen pretrained visual representations including PVR [21], MVP [34], R3M [20] and VC-1 [16], and report the best performance of these PVRs models for each task.
- ▷ Control Transformer: SMART [29] is a self-supervised representation pretraining framework which utilizes a maksed prediction objective for pretraining representation under Decision Transformer architecture, and then use the pretrained representation to learn policies for downstream tasks.
- ▷ Inverse Dynamics Model: We pretrain an inverse dynamics model to predict actions and use the pretrained representation for downstream task.
- ▷ Contrastive/Self-supervised Learning Objectives: CURL [13], ATC [28], and SPR [23, 24]. CURL and ATC are two approaches that apply contrastive learning into sequential decision making problems. While CURL treats augmented states as positive pairs, it neglects the temporal dependency of MDP. In comparison, ATC takes the temporal structure into consideration. The positive example of ATC is an augmented view of a temporally nearby state. SPR applies BYOL objective [8] into sequential decision making problems by pretraining state representations that are self-predictive of future states.

Number of Demonstrations and Evaluation Metric. For DMC, we use **20 expert trajectories** for imitation learning except for the two hardest tasks, Humanoid Walk and Dog Trot, for which we use 100 trajectories instead. We record the performance of the agent by calculating the ratio of $\frac{\text{Agent Reward}}{\text{Expert Reward}}$, where Expert Reward is the episode reward of the expert policy used to collect demonstration trajectories. For MetaWorld, we use **5 expert trajectories** for all eight downstream tasks, and we use task success rate as the performance metric.

C.2 Adaptation to Unseen Tasks

The results of adaptation to unseen tasks in MetaWorld are shown in Table 2.

C.3 Robustness to Low-quality Data

`Premier-TACO` **Pre-trained Representation is resilient to low-quality data.** We evaluate the resilience of `Premier-TACO` by employing randomly collected trajectory data from Deepmind Control Suite for pretraining and compare it with `Premier-TACO` representations pretrained using an exploratory dataset and the learn-from-scratch approach. As illustrated in Figure 5, across all downstream tasks, even when using randomly pretrained data, the `Premier-TACO` pretrained model still maintains a significant advantage over learning-from-scratch. When compared with representations pretrained using exploratory data, there are only small disparities in a few individual tasks, while they remain comparable in most other tasks. This strongly indicates the robustness of Premier-TACO to low-quality data. Even without the use of expert control data, our method is capable of extracting valuable information.

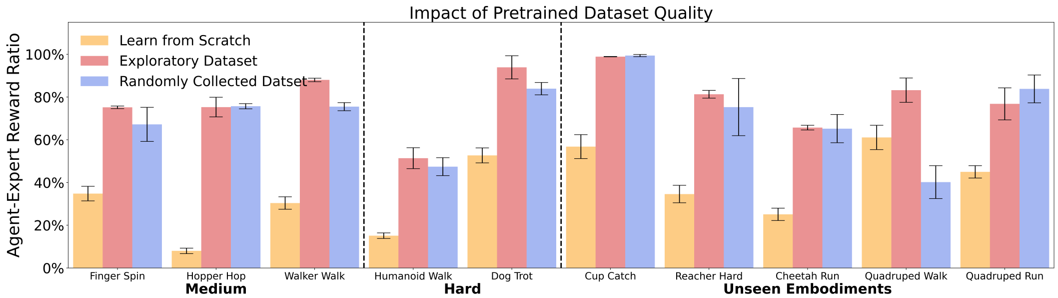

Figure 5: [(W3) Robustness] Premier-TACO pretrained with exploratory dataset vs. Premier-TACO pretrained with randomly collected dataset

C.4 Finetuning on pretrained visual representations

We conduct fine-tuning on pretrained visual representations using in-domain control trajectories following the Premier-TACO framework. Importantly, our findings deviate from the observations made in prior works [9, 16], where fine-tuning of R3M [20] on in-domain demonstration data using the task-centric behavior cloning objective, resulted in performance degradation. We speculate that two main factors contribute to this phenomenon. First, a domain gap exists between out-of-domain pretraining data and in-domain fine-tuning data. Second, fine-tuning with few-shot learning can lead to overfitting for large pretrained models. Comparisons among R3M [20], R3M with in-domain finetuning [9] and R3M finetuned with Premier-TACO in Deepmind Control Suite and MetaWorld are presented in Figure 6 and 7.

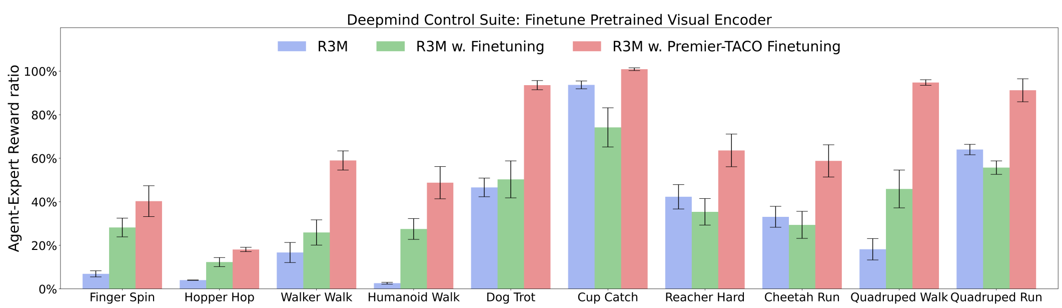

Figure 6: [(W4) Compatibility] Finetune R3M [20], a generalized Pretrained Visual Encoder with Premier-TACO learning objective vs. R3M with in-domain finetuning in Deepmind Control Suite and MetaWorld.

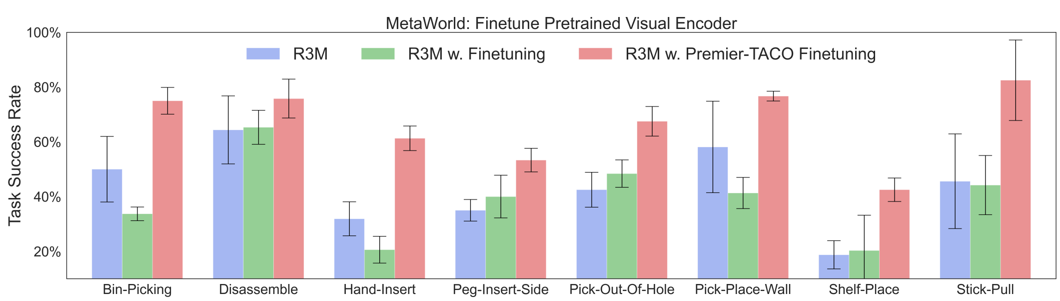

Figure 7: [(W4) Compatibility] Finetune R3M [20], a generalized Pretrained Visual Encoder with Premier-TACO learning objective vs. R3M with in-domain finetuning in Deepmind Control Suite and MetaWorld.

D Ablation Studies

D.1 Batch Size

Compared with TACO, the negative example sampling strategy employed in `Premier-TACO` allows us to sample harder negative examples within the same episode as the positive example. We expect `Premier-TACO` to work much better with small batch sizes, compared with TACO where the negative examples from a given batch could be coming from various tasks and thus the batch size required would scale up linearly with the number of pretraining tasks. In ours previous experimental results, `Premier-TACO` is pretrained with a batch size of 4096, a standard batch size used in contrastive learning literature. Here, to empirically verify the effects of different choices of the pretraining batch size, we train `Premier-TACO` with batch sizes other than 4096, and compare with the performance of TACO using a batch size of 4096. Figure 8 displays the average performance of few-shot imitation learning across all ten tasks in the DeepMind Control Suite. As depicted in the figure, our model markedly surpasses TACO, maintaining this superiority even with a batch size of 512, and exhibits performance saturation beyond a batch size of 4096. This observation substantiates that the negative example sampling strategy employed by `Premier-TACO` is indeed the key for the success of multitask offline pretraining.

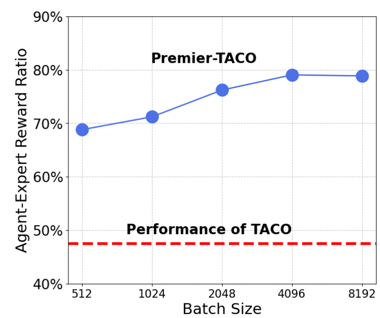

Figure 8: Averaged performance of `Premier-TACO` on 10 Deepmind Control Suite Tasks across different batch sizes.

D.2 Window Size

In `Premier-TACO`, the window size W determines the hardness of the negative example. A smaller window size results in negative examples that are more challenging to distinguish from positive examples, though they may become excessively difficult to differentiate in the latent space. Conversely, a larger window size makes distinguishing relatively straightforward, thereby mitigating the impacts of negative sampling. In the preceding experiments, a consistent window size of 5 was applied across all trials on both the DeepMind Control Suite and MetaWorld. Here we empirically evaluate the effects of varying window sizes on the average performance of our model across ten DeepMind Control Tasks, as depicted in Figure X. Notably, our observations reveal that performance is comparable when the window size is set to 3, 5, or 7, whereas excessively small ($W = 1$) or large ($W = 9$) window sizes lead to worse performance.

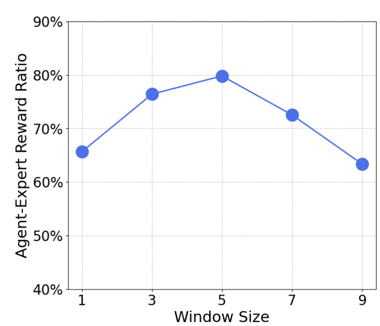

Figure 9: Averaged performance of `Premier-TACO` on 10 Deepmind Control Suite Tasks across different window sizes

E Implementation Details

Dataset For six pretraining tasks of the Deepmind Control Suite, we train visual RL agents for individual tasks with DrQ-v2 [36] until convergence, and we store all the historical interaction steps in a separate buffer. Then, we sample 200 trajectories from the buffer for all tasks except for Humanoid Stand and Dog Walk. Since these two tasks are significantly harder, we use 1000 pretraining trajectories instead. Each episode in the Deepmind Control Suite consists of 500 time steps. In terms of the randomly collected dataset, we sample trajectories by taking actions with each dimension independently sampled from a uniform distribution $\mathcal{U}(-1., 1.)$ For MetaWorld, we collect 1000 trajectories for each task, where each episode consists of 200 time steps. We add a Gaussian noise of standard deviation 0.3 to the provided scripted policy.

Pretraining For the shallow convolutional network, we follow the same architecture as in **(author?)** [36] and add a layer normalization on top of the output of the ConvNet encoder. We set the feature dimension of the ConNet encoder to be 100. In total, this encoder has around 3.95 million parameters.

```
class Encoder(nn.Module):
    def __init__(self):
        super().__init__()
        self.repr_dim = 32 * 35 * 35

        self.convnet = nn.Sequential(nn.Conv2d(84, 32, 3, stride=2),
                        nn.ReLU(), nn.Conv2d(32, 32, 3, stride=1),
                        nn.ReLU(), nn.Conv2d(32, 32, 3, stride=1),
                        nn.ReLU(), nn.Conv2d(32, 32, 3, stride=1),
                        nn.ReLU())
        self.trunk = nn.Sequential(nn.Linear(self.repr_dim,
    feature_dim),
                        nn.LayerNorm(feature_dim), nn.Tanh())

    def forward(self, obs):
        obs = obs / 255.0 - 0.5
        h = self.convnet(obs).view(h.shape[0], -1)
        return self.trunk(h)
```

Listing 1: Shallow Convolutional Network Architecture Used in `Premier-TACO`

For `Premier-TACO` loss, the number of timesteps K is set to be 3 throughout the experiments, and the window size W is fixed to be 5. The Action Encoder is a two-layer MLP with input size being the action space dimensionality, hidden size being 64, and output size being the same as the dimensionality of the action space. The projection layer G is a two-layer MLP with input size being feature dimension plus the number of timesteps times the dimensionality of the action space. Its hidden size is 1024. In terms of the projection layer H, it is also a two-layer MLP with input and output size both being the feature dimension and hidden size being 1024. Throughout the experiments, we set the batch size to be 4096 and the learning rate to be 1e-4. For the contrastive/self-supervised based baselines, CURL, ATC, and SPR, we use the same batch size of 4096 as `Premier-TACO`. For Multitask TD3+BC and Inverse dynamics modeling baselines, we use a batch size of 1024.

Imitation Learning A batch size of 128 and a learning rate of 1e-4 are used. During behavior cloning, we finetune the Shallow ConvNet Encoder. However, when we applied `Premier-TACO` for the large pre-trained ResNet/ViT model, we keep the model weights frozen.

In total, we take 100,000 gradient steps and conduct evaluations for every 1000 steps. For evaluations within the DeepMind Control Suite, we utilize the trained policy to execute 20 episodes, subsequently recording the mean episode reward. In the case of MetaWorld, we execute 50 episodes and document the success rate of the trained policy. We report the average of the highest three episode rewards/success rates from the 100 evaluated checkpoints.

Computational Resources For our experiments, we use 8 NVIDIA RTX A6000 with PyTorch Distributed DataParallel for pretraining visual representations, and we use NVIDIA RTX2080Ti for downstream imitation learning.

References

[1] Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, and R Devon Hjelm. Unsupervised state representation learning in atari. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 5

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale, 2023. 1

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 5

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1

[6] Scott Fujimoto and Shixiang (Shane) Gu. A minimalist approach to offline reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 20132–20145. Curran Associates, Inc., 2021. 6

[7] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo

Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022. 5

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 7

[9] Nicklas Hansen, Zhecheng Yuan, Yanjie Ze, Tongzhou Mu, Aravind Rajeswaran, Hao Su, Huazhe Xu, and Xiaolong Wang. On pre-training for visuo-motor control: Revisiting a learning-from-scratch baseline. In *CoRL 2022 Workshop on Pre-training Robot Learning*, 2022. 4, 5, 6, 8

[10] Nicklas A Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8387–8406. PMLR, 17–23 Jul 2022. 4

[11] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR, 13–18 Jul 2020. 5

[12] Minbeom Kim, Kyeongha Rho, Yong-duk Kim, and Kyomin Jung. Action-driven contrastive representation for reinforcement learning. *PLOS ONE*, 17(3):1–14, 03 2022. 5

[13] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5639–5650. PMLR, 13–18 Jul 2020. 5, 7

[14] Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *5th Annual Conference on Robot Learning*, 2021. 1

[15] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 5

[16] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence?, 2023. 2, 3, 5, 6, 7, 8

[17] Bogdan Mazoure, Remi Tachet des Combes, Thang Long Doan, Philip Bachman, and R Devon Hjelm. Deep reinforcement and infomax learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3686–3698. Curran Associates, Inc., 2020. 5

[18] Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, and Deepak Pathak. Discovering and achieving goals via world models. *Advances in Neural Information Processing Systems*, 34:24379–24391, 2021. 5

[19] Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. *CoRR*, abs/1911.05815, 2019. 6

[20] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *6th Annual Conference on Robot Learning*, 2022. 2, 5, 6, 7, 8

[21] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17359–17371. PMLR, 17–23 Jul 2022. 7

[22] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 1

[23] Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations*, 2021. 7

[24] Max Schwarzer, Nitarshan Rajkumar, Michael Noukhovitch, Ankesh Anand, Laurent Charlin, R Devon Hjelm, Philip Bachman, and Aaron Courville. Pretraining representations for data-efficient reinforcement learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 7

[25] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*, pages 8583–8592. PMLR, 2020. 5

[26] Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked world models for visual control. In *CoRL*, volume 205 of *Proceedings of Machine Learning Research*, pages 1332–1344. PMLR, 2022. 4

[27] Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. In *International Conference on Machine Learning*, pages 9870–9879. PMLR, 2021. 1

[28] Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9870–9879. PMLR, 18–24 Jul 2021. 5, 7

[29] Yanchao Sun, Shuang Ma, Ratnesh Madaan, Rogerio Bonatti, Furong Huang, and Ashish Kapoor. SMART: Self-supervised multi-task pretraining with control transformers. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 5, 6, 7

[30] Yanchao Sun, Ruijie Zheng, Xiyao Wang, Andrew E Cohen, and Furong Huang. Transfer RL across observation feature spaces via model-based regularization. In *International Conference on Learning Representations*, 2022. 5

[31] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite, 2018. 3, 6

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 3, 5

[33] Yao Wei, Yanchao Sun, Ruijie Zheng, Sai Vemprala, Rogerio Bonatti, Shuhang Chen, Ratnesh Madaan, Zhongjie Ba, Ashish Kapoor, and Shuang Ma. Is imitation all you need? generalized decision-making with dual-phase training. *arXiv preprint arXiv:2307.07909*, 2023. 5

[34] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control, 2022. 5, 7

[35] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, pages 11920–11931. PMLR, 2021. 5

[36] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *International Conference on Learning Representations*, 2022. 6, 9, 10

[37] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2019. 3, 6

[38] Zhecheng Yuan, Zhengrong Xue, Bo Yuan, Xueqian Wang, YI WU, Yang Gao, and Huazhe Xu. Pre-trained image encoder for generalizable visual reinforcement learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 13022–13037. Curran Associates, Inc., 2022. 5

[39] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. 5

[40] Ruijie Zheng, Xiyao Wang, Yanchao Sun, Shuang Ma, Jieyu Zhao, Huazhe Xu, Hal Daumé III, and Furong Huang. TACO: Temporal latent action-driven contrastive loss for visual reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3, 5